III MACI 2011

9 al 11 de mayo de 2011
Bahía Blanca, Argentina

Anales del III Congreso de Matemática Aplicada, Computacional e Industrial

L.R. Castro, M.C. Maciel y
S.M. Castro, Eds.

III MACI 2011

III Congreso de Matemática Aplicada, Computacional e Industrial

Bahía Blanca, Buenos Aires, Argentina, 9-11 de mayo de 2011

LILIANA RAQUEL CASTRO
MARÍA CRISTINA MACIEL
SILVIA MABEL CASTRO
Editores

Departamento de Matemática
UNIVERSIDAD NACIONAL DEL SUR

asamaci
ASOCIACIÓN ARGENTINA DE MATEMÁTICA APLICADA, COMPUTACIONAL E INDUSTRIAL

AR-SIAM
SIAM Society for Industrial and Applied Mathematics
SECCIÓN ARGENTINA

Patrocinadores

CONICET

Departamento de Matemática UNS

Departamento de Ingeniería Química UNS

KB Engineering S.R.L.

siam Society for Industrial and Applied Mathematics

AGENCIA

BANCOPATAGONIA

asamaci ASOCIACIÓN ARGENTINA DE MATEMÁTICA APLICADA, COMPUTACIONAL E INDUSTRIAL

Auspiciantes

UNSAM UNIVERSIDAD NACIONAL DE SAN MARTÍN ESCUELA DE CIENCIA Y TECNOLOGÍA

COPROTUR Bahía Blanca

UNRC

UNIVERSIDAD NACIONAL DEL COMAHUE 1972

UNIVERSIDAD AUSTRAL Facultad de Ciencias Empresariales

FCEIA FACULTAD DE CIENCIAS EXACTAS, INGENIERÍA Y AGRIMENSURA

UTN FACULTAD REGIONAL BAHÍA BLANCA

FaEA Facultad de Economía y Administración

ASOCIACIÓN ARGENTINA DE MECÁNICA COMPUTACIONAL AMCA 25 AÑOS 1985-2010

OPTYCON

UNION MATEMATICA ARGENTINA FUNDADA EN 1936

PREFACIO

El segundo volumen de la colección MACI, recientemente creada, incluye los trabajos presentados en el Tercer Congreso de Matemática Aplicada, Computacional e Industrial, III MACI 2011, que tuvo lugar en ámbitos de la Universidad Nacional del Sur en la ciudad de Bahía Blanca, del 9 al 11 de mayo de 2011.

Este congreso de la Asociación Argentina de Matemática Aplicada Computacional e Industrial, ASAMACI, y de la sección Argentina de Society for Industrial and Applied Mathematics, AR-SIAM, se realizó por primera vez en Córdoba en 2007 junto con el ENIEF de la Asociación Argentina de Mecánica Computacional, AMCA. El II MACI 2009 tuvo lugar en Rosario en diciembre de 2009.

Los congresos MACI están dirigidos a Matemáticos, Ingenieros, Físicos, Biólogos, Economistas y otros profesionales interesados en aplicaciones de la matemática y para los cuales ésta tiene un rol significativo en el desarrollo de sus investigaciones.

En esta oportunidad se presentaron 208 trabajos, de los cuales se aceptaron 197 contribuciones, se dictaron cuatro cursos para estudiantes avanzados y cuatro conferencias plenarias a cargo de renombrados científicos. Cabe remarcar que la conferencia inaugural estuvo a cargo del reconocido matemático Prof. Gilbert Strang. También se contó con la participación de los investigadores Dres. Luis Caffarelli, Ricardo Sánchez Peña, José Mario Martiínez, Sergio Preidikman y Gabriel Soto. La conferencia de clausura estuvo a cargo del presidente de la Asociación Física Argentina, Dr. Francisco Tamarit.

La organización del III MACI 2011 fue responsabilidad del Comité Organizador Local y contó con la invalorable colaboración de los coordinadores de sesión, quienes tuvieron la responsabilidad de hacer evaluar los trabajos correspondientes. El proceso de referato de los artículos presentados en este Congreso fue anónimo, lo que permitió que muchos de los manuscritos fueran sustancialmente mejorados.

El III MACI 2011 contó con el financiamiento de instituciones nacionales como la Agencia de Promición Científica y Tecnológica, el Consejo Nacional de Investigaciones Científica y Técnicas, y de los Departamentos de Matemática y de Ingeniería Química de la Universidad Nacional del Sur. Por parte del sector privado brindaron apoyo económico KB Engineering y el Banco Patagonia. También se contó con el importante aval de la Society for Industrial and Applied Mathematics, SIAM.

La Comisión Directiva de ASAMACI y la Comisión Organizadora del III MACI 2011 agradecen profundamente el esfuerzo realizado por los coordinadores de sesión, los revisores de artículos, los miembros del Comité Organizador Local y el Dr. Carlos Zuppa, quien desinteresadamente se ocupó de la construcción y mantenimiento del sitio web de este congreso. Por último, agradece a todas las Universidades y Agrupaciones Científicas que auspiciaron este evento así como a la Municipalidad de Bahía Blanca.

L.R. Castro, M.C. Maciel y S.M. Castro (Eds.)
Bahía Blanca, mayo de 2011

III MACI 2011
TERCER CONGRESO DE MATEMÁTICA APLICADA, COMPUTACIONAL E INDUSTRIAL

9 al 11 de mayo de 2011, Bahía Blanca, ARGENTINA

Comité científico

Carlos D'Attellis, Univ. Favaloro - UNSAM, Buenos Aires

Pablo Jacovkis, UBA, Buenos Aires

María Cristina Maciel, UNS, Bahía Blanca

Sergio Preidikman, CONICET - UNC, Córdoba

Diana Rubio, UNSAM, Buenos Aires

Rubén Spies, IMAL (CONICET - UNL), Santa Fe

Juan Santos, CONICET-UNLP, La Plata

Domingo Tarzia, CONICET - UA, Rosario

Cristina Turner, CONICET - UNC, Córdoba

Comité organizador local

María Cristina Maciel

Liliana Raquel Castro

Víctor Cortínez

María Soledad Díaz

Jorge Moiola

Ricardo Pignol

Fernando Tohmé

Marta Cecilia Vidal

Colaboradores

Marcela Álvarez, Liliana Boscardín, Flavia Edith Buffo, Gabriel Aníbal Carrizo, Silvia Mabel Castro, María Gabriela Eberle, Graciela Paolini, Diana Salgado, Adriana Beatriz Verdiell.

Coordinadores de sesiones científicas

Biomatemática: Carlos D'Attellis - Gabriel Soto

Economía Matemática: Fernando Tohmé - Alejandro Neme

Ecuaciones Diferenciales y Aplicaciones: Julián Fernández Bonder - Cristina Turner

Finanzas Cuantitativas: Elsa Cortina - Rodolfo Oviedo

Fundamentos de Métodos Numéricos y Aplicaciones: Gabriel Acosta - Pedro Morin

Matemática Discreta y Aplicaciones: Marisa Gutiérrez

Matemática Industrial y Aplicaciones: Javier Etcheverry - Adrián Will

Mecánica Computacional: Sergio Idelsohn - Alejandro Limache

Modelos Matemáticos Interdisciplinarios: Pablo Jacovkis

Optimización: Teoría y Aplicaciones: Roberto Andreani - Laura Schuverdt

Probabilidad, Estadística y Procesos Estocásticos: Beatriz Marrón - Elina Mancinelli

Problemas de Frontera Libre y Aplicaciones: Adriana Briozzo - Claudia Lederman

Problemas Inversos y Aplicaciones: Karina Temperini - Diana Rubio

Problemas Matemáticos en Mecánica del Continuo: Sergio Elaskar - Sergio Preidikman

Procesamiento de Señales e Imágenes: Liliana Castro - Eduardo Serrano

Sistemas Dinámicos: María Inés Troparevsky - Ernesto Kofman

Teoría de Control Óptimo y Aplicaciones: Laura Aragone - Pablo Lotito

Transferencia de Calor y Materia: Luis Villa - Eduardo Santillán Marcus

Pósteres de Estudiantes de Grado: Graciela Sottosanto - Omar Faure

Pósteres de Estudiantes de Posgrado: Graciela Sottosanto - Omar Faure

Cursos

Silvia Castro: *Visualización y Matemática*

Mario Martínez: *Optimización con restricciones para problemas de gran tamaño*

Sergio Preidikman: *Dinámica no-lineal y Caos: Conceptos y Aplicaciones*

Gabriel Soto: *Modelos biofísicos de las neuronas*

Conferencias plenarias

Luis Caffarelli: *Optimal control for Levi processes*

Ricardo Sánchez Peña: *Identificación y control: la brecha entre la teoría y la práctica*

Francisco Tamarit: *Memoria asociativa con redes complejas*

Gilbert Strang: *Matrix factorizations, old and new*

Índice

DINÁMICA DE ADICCIÓN AL TABAQUISMO CON POBLACIÓN CONSTANTE

Nini Johana Fiallo Rendon, Leonardo Duvan Restrepo Alape y Anibal Muñoz Loaiza

Facultad de Educación, Programa de Licenciatura en Matemáticas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia, ninis-428@hotmail.com, anibalml@hotmail.com, www.uniquindio.edu.co

Resumen: Se construye un modelo para la dinámica de adicción al tabaquismo con base en un sistema de ecuaciones diferenciales no lineales que interpreta la dinámica, realizando el análisis de estabilidad local en términos del umbral de adicción aplicando el criterio de la traza y el determinante y el criterio de Routh - Hurwitz. En relación a la estabilidad global se aplica el método directo de Lyapunov para el caso de la solución libre de adicción al tabaquismo. Los resultados se concluyen en teoremas de estabilidad, estos análisis se complementan simulando el sistema con MATLAB utilizando valores hipotéticos para los parámetros y de acuerdo al umbral sociológico de adicción.

Palabras clave: *Modelo dinámico, Tabaquismo, Estabilidad local, Estabilidad global, Función de Lyapunov.*

1. INTRODUCCIÓN

El tabaquismo es una droga presentada en diferentes formas de consumo derivada de la planta de nombre científico Nicotiana tabacu y causante de diferentes y dolorosas enfermedades. Se define como ”Intoxicación crónica producida por el abuso del tabaco”. El tabaquismo no es un hábito nocivo, es una drogodependencia. Nosotros lo definiríamos como enfermedad crónica producida por el consumo reiterado de cualquier producto cuya materia prima deriva de la planta del tabaco. El tabaco es el máximo responsable de enfermedades invalidantes e incapacidades laborales, con una grave repercusión familiar, social y económica. Es el causante de 5.000.000 de muertes anuales y del 90 % de todos los cánceres de pulmón. La adicción al tabaco tiene un doble componente: uno es la nicotina y otro es el hábito de comportamiento, es decir, las distintas situaciones a las que se encuentra sometido el fumador. En relación a estudios en drogadición citamos (1), (3), (4).

2. DESCRIPCIÓN DEL MODELO

En esta sección se formula y realiza el análisis de estabilidad local y global con base en el umbral sociológico de adicción al tabaquismo. Los supuestos del modelo son: grupo de riesgo de personas al consumo de tabaquismo desde la edad promedio de 10 años, principio de acción de masas y mortalidad por consumo de tabaquismo en el caso de que sea crónico y las variables y parámetros son: S : número promedio de personas mayores de 10 años susceptibles a ser fumadores, I_a : número promedio de personas mayores de 10 años fumadores, I_p : número promedio de personas mayores de 10 años fumadores pasivos, C : número promedio de personas mayores de 10 años fumadores crónicos, N : flujo de personas que cumplen la edad continuamente y que ingresan al grupo de riesgo de la población susceptible en un tiempo t respectivamente; α : tasa de personas que dejan de fumar y vuelven hacer susceptibles, β : coeficiente de encuentros efectivos, para personas que se vuelven fumadoras, δ : fuerza de infección de fumadores pasivos que adquieren el habito de fumar, μ : tasa de mortalidad natural, σ : flujo de personas que cumplen la edad continuamente y que ingresan al grupo de fumadores pasivos, θ : tasa de fumadores que recaen a fumadores crónicos.

Las ecuaciones diferenciales que gobiernan la dinámica son:

$$\frac{dS}{dt} = \mu N + \alpha I_a - \beta \frac{I_a}{N} S - \sigma \frac{I_a}{N} S - \mu S \tag{1}$$

$$\frac{dI_a}{dt} = \beta \frac{I_a}{N} S + \delta I_p - (\alpha + \theta + \mu) I_a \tag{2}$$

$$\frac{dI_p}{dt} = \sigma \frac{I_a}{N} S - (\delta + \mu) I_p \tag{3}$$

$$\frac{dC}{dt} = \theta I_a - \mu C \tag{4}$$

2.1. ANÁLISIS DE ESTABILIDAD

Se inicia el análisis (2), (5), reduciendo el sistema, puesto que la población total es constante $N = S + I_a + I_p + C$, luego $C = N - S - I_a - I_p$, quedando el sistema (1)-(4) reducido a un sistema de dimension tres:

$$\frac{dS}{dt} = \mu N + \alpha I_a - \beta \frac{I_a}{N} S - \sigma \frac{I_a}{N} S - \mu S \tag{5}$$

$$\frac{dI_a}{dt} = \beta \frac{I_a}{N} S + \delta I_p - \eta I_a \tag{6}$$

$$\frac{dI_p}{dt} = \sigma \frac{I_a}{N} S - \omega I_p \tag{7}$$

donde, $\eta = \alpha + \theta + \mu$ y $\omega = \delta + \mu$. Además $(\mu, \alpha, \beta, \sigma, \delta, \theta) > 0$ y sus condiciones iniciales son: $S(0) = S_0$, $I_a(0) = I_{a0}$, $I_p(0) = I_{p0}$ y la región de invarianza de sentido Sociológico de adicción al tabaquismo es,

$$\Omega = \{(S, I_a, I_p) \in \mathbb{R}^3_+ : S + I_a + I_p \leq N\}$$

Umbral de adicción al tabaquismo Θ_0:
El Θ_0, umbral de adicción al tabaquismo, se define como el número promedio de fumadores activos que un fumador activo puede provocar durante su tiempo promedio de fumador en una población susceptible:

$$\Theta_0 = \frac{\beta(\delta + \mu) + \sigma\delta}{(\alpha + \theta + \mu)(\delta + \mu)} = \frac{\beta}{\alpha + \theta + \mu} + \sigma \left(\frac{\delta}{\delta + \mu}\right)\left(\frac{1}{\alpha + \theta + \mu}\right) = \hat{\Theta}_0 + \tilde{\Theta}_0$$

donde, $\hat{\Theta}_0$: es el número promedio de casos de adicción al tabaquismo inducidos por un fumador activo durante el tiempo de fumador en una población susceptible y $\tilde{\Theta}_0$: la incidencia de nuevos casos por adicción pasiva con los fumadores activos.

Soluciones estacionarias:
Las soluciones estacionarias se obtienen haciendo $\frac{dx}{dt} = 0$, $\frac{dI_a}{dt} = 0$, $\frac{dI_p}{dt} = 0$ en el sistema reducido y resolviendo el sistema algebraico no lineal resultante para S, I_a y I_p obteniendo la solución estacionaria trivial (libre de adicción), $E_0 = (N, 0, 0)$ y la solución estacionaria no trivial (en presencia de tabaquismo): $E_1 = (\hat{S}_1, \hat{I}_{a1}, \hat{I}_{p1})$. Donde,

$$\hat{S}_1 = \frac{\omega\eta N}{\beta\omega + \delta\sigma} \quad , \quad \hat{I}_{a1} = \frac{\mu N \omega \eta (\Theta_0 - 1)}{\omega(\beta + \sigma)(\theta + \mu) + \sigma\alpha\mu} \quad , \quad \hat{I}_{p1} = \frac{\sigma\mu\eta^2\omega N(\Theta_0 - 1)}{(\beta\omega + \delta\sigma)[\sigma\alpha\mu + \omega(\beta + \sigma)(\theta + \mu)]}$$

donde, $\eta = \alpha + \theta + \mu$ y $\omega = \delta + \mu$ y la cual tiene sentido sociológico cuando $\Theta_0 \geq 1$. La matriz de estabilidad:

$$J(E) = \begin{pmatrix} -\beta\frac{\hat{I}_a}{N} - \sigma\frac{\hat{I}_a}{N} - \mu & \alpha - \beta\frac{\hat{S}}{N} - \sigma\frac{\hat{S}}{N} & 0 \\ \beta\frac{\hat{I}_a}{N} & \beta\frac{\hat{S}}{N} - \eta & \delta \\ \sigma\frac{\hat{I}_a}{N} & \sigma\frac{\hat{S}}{N} & -\omega \end{pmatrix} \tag{8}$$

evaluada en la solución estacionaria trivial conduce a la ecuación característica $|J(E_0) - \lambda I| = (-\mu - \lambda)\left[(\beta - \eta - \lambda)(-\omega - \lambda)\right] - \sigma\delta = 0$. En la cual, un valor propio es $\lambda_1 = -\mu$ y los otros dos valores propios son las raíces de la ecuación:

$$\lambda^2 + \left[\omega - \eta(\Theta_0^a - 1)\right]\lambda - \omega\eta(\Theta_0 - 1) = 0$$

aplicando el criterio de la traza y el determinante tenemos: $\omega - \eta(\Theta_0^a - 1) < 0$ y $\omega\eta(\Theta_0 - 1) > 0$ cuando $\Theta_0 < 1$ y $\Theta_0^a < \Theta_0$. Concluimos en el siguiente teorema:

Teorema 1 *Si $\Theta_0 < 1$, la solución estacionaria trivial $E_0 = (N, 0, 0)$ del sistema (5)-(7) es local y asintóticamente estable.*

Se demostró que para $\Theta_0 < 1$ la solución estacionaria trivial es local y asintóticamente estable en consecuencia formulamos y demostramos el siguiente teorema de estabilidad global:

Teorema 2 *Si $\Theta_0 < 1$, la existencia de estabilidad local implica estabilidad global.*

Prueba. Definimos la función de Liapunov (definida positiva): $V = (\delta+\mu)I_a+\delta I_p$ derivando con respecto a t, obtenemos: $\frac{dV}{dt} = (\delta + \mu)\frac{dI_a}{dt} + \delta\frac{dI_p}{dt}$. Sustituyendo las funciones correspondientes a $\frac{dI_a}{dt}$ y $\frac{dI_p}{dt}$ resulta:

$$\frac{dV}{dt} = (\delta + \mu)\left[\beta\frac{I_a}{N}S - (\alpha + \theta + \mu)I_a + \delta I_p\right] + \delta\left[\sigma\frac{I_a}{N}S - (\delta + \mu)I_p\right]$$

como $S = N - (I_a + I_p)$, tenemos:

$$\frac{dV}{dt} = I_a(\delta + \mu)(\alpha + \theta + \mu)\left[\Theta_0 - 1\right] - \left[\beta(\delta + \mu) + \delta\sigma\right]\frac{I_a}{N}(I_a + I_p)$$

donde, $\Theta_0 = \frac{(\delta+\mu)\beta+\sigma\delta}{(\delta+\mu)(\alpha+\theta+\mu)}$.

En la región de sentido biológico Ω tenemos: $\frac{dV}{dt} < 0$ cuando $\Theta_0 < 1$ y así queda demostrado el teorema. Se observa que en la región Ω y para $\Theta_0 \leq 1$: $\frac{dV}{dt} \leq 0$. Así, $\frac{dV}{dt} = 0$ implica que:

$$I_a(\delta + \mu)(\alpha + \theta + \mu)\left[\Theta_0 - 1\right] - \left[\beta(\delta + \mu) + \delta\sigma\right]\frac{I_a}{N}(I_a + I_p) = 0$$

Por tanto, si $\Theta_0 < 1$ entonces $I_a = 0$ y si $\Theta_0 = 1$ entonces $I_a = 0$. Luego, el conjunto $\{E_1\}$ es un conjunto invariante dentro del conjunto:

$$\left\{(S, I_a, I_p) : \frac{dV}{dt}(S, I_a, I_p) = 0\right\}$$

De acuerdo al teorema de conjunto invariante: toda trayectoria en Ω tiende a E_1 cuando t se incrementa y si E_1 es localmente estable entonces es global y asintóticamente estable. □

Para analizar la estabilidad de la solución estacionaria no trivial, hacemos en la matriz jacobiana general las sustituciones:

$$a = \beta\frac{\hat{I}_a}{N} + \sigma\frac{\hat{I}_a}{N} \quad , \quad b = \beta\frac{\hat{I}_a}{N} \quad , \quad c = \sigma\frac{\hat{I}_a}{N} \quad , \quad d = \beta\frac{\hat{S}}{N} + \sigma\frac{\hat{S}}{N} \quad , \quad e = \beta\frac{\hat{S}}{N} \quad , \quad r = \sigma\frac{\hat{S}}{N}$$

obteniendo la ecuación característica $|J(E_1) - \lambda I| = 0$:

$$(a + \mu + \lambda)(e - \eta - \lambda)(\omega + \lambda) + \delta c(\alpha - d) + r\delta(a + \mu + \lambda) + b(\alpha - d)(\omega + \lambda) = 0$$

que tiene la forma $\lambda^3 + A\lambda^2 + B\lambda + D = 0$ donde,

$A = \omega + \eta + a + \mu - e$
$B = (\eta - e)(a + \mu) + (\eta + a + \mu - e)\omega - r\delta + b(d - \alpha)$
$D = (a + \mu)(\eta - e)\omega + \delta c(d - \alpha) - r\delta(a + \mu) + b(d - \alpha)$
Aplicando el criterio de estabilidad de Routh-Hurwitz, si se cumplen las desigualdades: $A > 0$, $D > 0$ y $AB > D$ las raíces de la ecuación característica tienen parte real negativa y por lo tanto, la solución estacionaria no trivial es estable.

Teorema 3 *Si $\Theta_0 > 1$ entonces la solución estacionaria no trivial del sistema (5)-(7) es local y asintóticamente estable.*

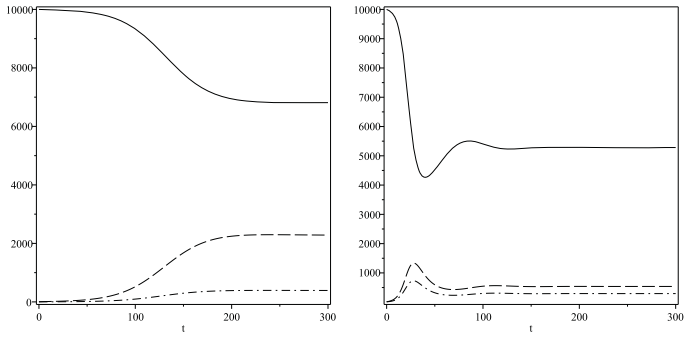


Figura 1: comportamiento en el tiempo de personas susceptibles $S\ (-)$, fumadores activos $I_a\ (-.-.-)$ y fumadores pasivos $I_p\ (- - --)$ con $\Theta_0 \approx 1.47$ y $\Theta_0 \approx 1.89$.

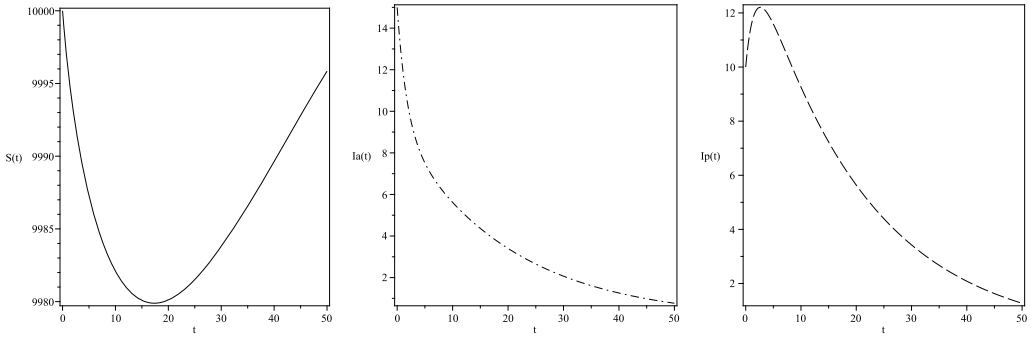


Figura 2: comportamiento en el tiempo de personas susceptibles $S\ (-)$, fumadores activos $I_a\ (-.-.-)$ y los fumadores pasivos $I_p\ (- - --)$ con $\Theta_0 \approx 0.81$.

3. Resultados Numéricos

Las simulaciones del modelo se hicieron utilizando el programa MAPLE con condiciones iniciales y valores hipotéticos de los parámetros.

Agradecimientos

Al Programa de matemáticas y Facultad de Educación, Universidad del Quindío.

Referencias

[1] A.B. Gumel, O. Sharomi, *Curtailing smoking dynamics: A mathematical modeling approach*, Applied Mathematics and Computation, 19(2008)475-499.

[2] L.A. Muñoz, T.H. Colorado, A.O.M. Garcia, *Modelos Biomatemáticos I*, ISBN 978-958-44-4079-2, Editado por Ediciones Elizcom(2008).

[3] D. Winklera, J. Caulkinsb, D. Behrensa, G. Traglera, *Estimating the relative efficiency of various forms of prevention at dufferent stages of drug epidemic*, Socio-Economic Planning Sciencies 38(2004) 43-560.

[4] G. Mulone, B. Straughan, *A note on heroin epidemics*, Mathematical Biosciencies, 218(2009) 138-141.

[5] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer - Verlag New York, Inc., (1900).

MODELADO ESPACIO - TEMPORAL DE CRECIMIENTO POBLACIONAL DEL *Aedes aegypti*

Carlos Alberto Abello Muñoz[1,2], Anibal Muñoz Loaiza[2] y Hernán Darío Toro Zapata[2]

[1] *Maestría en Enseñanza de las Matemáticas, Universidad Tecnológica de Pereira, Pereira, Risaralda, Colombia,*
[1,2] *Facultad de Educación, Programa de Matemáticas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia; caabello@uniquindio.edu.co, anibalml@hotmail.com, hdtoro@uniquindio.edu.co, www.uniquindio.edu.co*

Resumen: Se presenta el avance de la construcción y análisis de un modelo determinista con base en ecuaciones diferenciales parciales incluyendo difusión en una coordenada para la dinámica de crecimiento poblacional del mosquito *Aedes aegypti* transmisor del virus del dengue. Se analiza el modelo propuesto por la técnica de ondas viajeras, transformando las EDP en un sistema de ecuaciones diferenciales no lineales acopladas al cual se le realiza el análisis de estabilidad local y simulaciones con MAPLE utilizando valores hipotéticos para los parámetros. Por ultimo se describe una extension del modelo incluyendo difusión en dos coordenadas espaciales, para posterior estudio numérico.

Palabras clave: *Modelado Matemático, Ondas viajeras, Difusión, Estabilidad local, Aedes aegypti, Dengue.*

1. INTRODUCCIÓN

El dengue es una enfermedad viral transmitida al hombre por mosquitos vectores *Aedes aegypti* el más importante en la transmisión de la enfermedad. El dengue uno de los problemas de mayor impacto en la salud pública en la actualidad. De ahí Instituciones como la Organización Mundial de la Salud (OMS), Organización Panamericana de la Salud (OPS) y otras orientan sus esfuerzos a controlar los brotes de mosquitos desde diferentes enfoques como los estudios experimentales de laboratorio y de campo, la implementación de estrategias de control integrado del vector, las campañas de educación de la población. La ausencia de vacunas o medicamentos contra el virus nos conduce a determinar, por lo pronto que el control debe basarse en la disminución y eventual eliminación de los mosquitos vectores [3] y de ahí la importancia de modelar la dinámica espacio - temporal de crecimiento poblacional del mosquito *A. aegypti*. Controlando las poblaciones vectoriales se disminuye la incidencia del dengue clásico y los brotes epidémicos de dengue hemorrágico, enfermedades de gran impacto en la salud humana en países tropicales y en general en el mundo , por los cambios climáticos que se presentan actualmente. De ahí que este trabajo se oriente a resolver el problema del modelado matemático de la dinámica espacio - temporal de crecimiento del mosquito que contribuye a establecer estrategias de control más apropiadas disminuyendo el impacto económico y social del dengue.

2. EL MODELO

Se modela con base en ecuaciones diferenciales parciales la dinámica de crecimiento poblacional del *A. aegypti* considerando su ciclo de vida, el estado maduro y un estado inmaduro que incluye el huevo, la larva y la pupa, dispersion del mosquito maduro, tasa de muerte natural, tasa de desarrollo del estado inmaduro al estado maduro, tasa de ovoposición y una fracción que regula el crecimiento poblacional del estado inmaduro dependiente de la capacidad de carga. Por simplicidad matemática se considera un espacio uno dimensional.

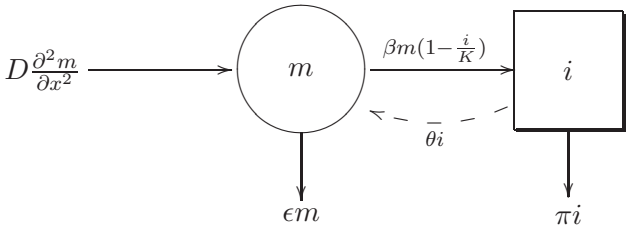


Diagrama N 1: Dinámica de dispersión denso - dependiente.

$$\frac{\partial m}{\partial t} = D\frac{\partial^2 m}{\partial x^2} + \theta i - \epsilon m \tag{1}$$

$$\frac{\partial i}{\partial t} = \beta m(1 - \frac{i}{K}) - (\pi + \theta)i \tag{2}$$

donde, $\theta, \epsilon, \beta, \pi \geq 0$. Las variables y parámetros del modelo propuesto son: $m(t, x)$: número de mosquitos maduros, $i(t, x)$: número de estados inmaduros (huevos, larvas, pupas) en un tiempo t en el lugar x respectivamente. $D = \alpha$: coeficiente de dispersión constante del mosquito, θ : tasa de estados inmaduros que desarrollan a estado maduro, ϵ : tasa de mortalidad natural de los mosquitos maduros, β : tasa de oviposición por mosquito, K : capacidad de carga de los estados inmaduros y π : tasa de mortalidad natural de los estados inmaduros.

3. Análisis del Modelo

Para establecer la existencia de ondas viajeras del sistema, se asume que tiene solución de la forma:

$$m(x, t) = M(x + \sigma t) \quad , \quad i(x, t) = I(x + \sigma t)$$

donde las funciones M, I son funciones de la variable de onda viajera $z = x + \sigma t$ y el parámetro de onda σ es positivo. Luego,

$$\frac{\partial m}{\partial t} = \sigma\frac{\partial M}{\partial z} \quad , \quad \frac{\partial i}{\partial t} = \sigma\frac{\partial I}{\partial z} \quad , \quad \frac{\partial^2 m}{\partial x^2} = \frac{\partial^2 M}{\partial z^2}$$

Sustituyendo en las ecuaciones (1), (2) obtenemos el sistema:

$$\sigma\frac{dM}{dz} = \alpha\frac{d^2 M}{\partial z^2} + \theta I - \epsilon M \tag{3}$$

$$\sigma\frac{dI}{dz} = \beta M(1 - \frac{I}{K}) - (\pi + \theta)I \tag{4}$$

Por motivaciones biológicas, se requiere que las ondas viajeras M y I sea positivas y satisfagan las condiciones de frontera:

$$M(-\infty) = 1 \quad , \quad M(+\infty) = \bar{M} \quad ; \quad I(-\infty) = 0 \quad , \quad I(+\infty) = \bar{I}$$

Haciendo, $\frac{dM}{dz} = W$ y $\frac{d^2 M}{dz^2} = \frac{dW}{dz}$ en las ecuaciones (3), (4) obtenemos el siguiente sistema de ecuaciones diferenciales ordinarias no lineales de ondas viajeras:

$$\frac{dM}{dz} = W \tag{5}$$

$$\alpha\frac{dW}{dz} = \theta I - \epsilon M - \sigma W \tag{6}$$

$$\sigma\frac{dI}{dz} = (\pi + \theta)I - \beta M\left(1 - \frac{I}{K}\right) \tag{7}$$

Tomando, $\rho_1 = \frac{\theta}{\alpha}$, $\rho_2 = \frac{\epsilon}{\alpha}$, $\rho_3 = \frac{\sigma}{\alpha}$, $\phi = \frac{\pi+\theta}{\sigma}$, $\psi = \frac{\beta}{\sigma}$ en las ecuaciones (6)-(7) tenemos el sistema:

$$\frac{dM}{dz} = W = f_1(M, W, I) \tag{8}$$

$$\frac{dW}{dz} = \rho_1 I - \rho_2 M - \rho_3 W = f_2(M, W, I) \tag{9}$$

$$\frac{dI}{dz} = \phi I - \psi M\left(1 - \frac{I}{K}\right) = f_3(M, W, I) \tag{10}$$

Las soluciones estacionarias se obtienen haciendo $\frac{dM}{dz} = 0$, $\frac{dW}{dz} = 0$, $\frac{dI}{dz} = 0$ y resolviendo el sistema algebraico no lineal para las variables de ondas viajeras M, W y I, obteniendo la solución estacionaria trivial $E_0 = (0, 0, 0)$ y la solución estacionaria no trivial:

$$E_1 = (\bar{M}_1, \bar{W}_1, \bar{I}_1) = \left(\frac{\rho_1 K}{\rho_2}(1-\eta), 0, K(1-\eta) \right)$$

la cual es positiva y tiene sentido biológico cuando $\eta = \frac{\rho_2 \phi}{\rho_1 \psi} < 1$.

Linealizamos el sistema no lineal considerando pequeñas desviaciones a nivel local: $M = \bar{M} + p, W = \bar{I} + q, I = \bar{I} + s$ de las soluciones estacionarias y expandiendo las funciones f_1, f_2, f_3 en series de Taylor, obteniendo la matriz de coeficientes del sistema lineal o matriz jacobiana:

$$J(\bar{M}, \bar{W}, \bar{I}) = \begin{pmatrix} 0 & -1 & 0 \\ -\rho_2 & -\rho_3 & \rho_1 \\ -\psi\left(1 - \frac{\bar{I}}{K}\right) & 0 & \phi + \frac{\psi}{K}\bar{M} \end{pmatrix}$$

La ecuación característica correspondiente $|J(\bar{M}, \bar{W}, \bar{I}) - \lambda I| = \lambda^3 + A\lambda^2 + B\lambda + C = 0$ donde,
$A = \rho_3 - \left(\phi + \frac{\psi}{K}\bar{M}\right), \quad B = -\rho_2 + \rho_3\left(\phi + \frac{\psi}{K}\bar{M}\right), \quad C = \rho_2\left(\phi + \frac{\psi}{K}\bar{M}\right) - \rho_1\psi\left(1 - \frac{\bar{I}}{K}\right).$

En el caso de la solución estacionaria trivial los coeficientes de la ecuación característica son:

$$A = \rho_3 - \phi \quad , \quad B = \rho_3\phi - \rho_2 \quad , \quad C = \rho_2\phi - \rho_1\psi = \rho_1\psi(\eta - 1)$$

Aplicando el criterio de Routh-Hurwitz, la ecuación característica tiene valores propios (raíces) con parte real negativa si se cumplen las desigualdades: $A > 0, C > 0, AB > C$. Se cumple que $C = \rho_1\psi(\eta - 1) < 0$ cuando $\eta < 1$. Por lo tanto, la solución estacionaria trivial es inestable.

Teorema 1 *Si $\eta < 1$ entonces la solución estacionaria trivial $(0, 0, 0)$ del sistema (3) a (5) es inestable.*

Para el caso de la solución estacionaria no trivial concluimos la siguiente conjetura:

Conjetura 1 *Si $\eta > 1$ entonces la solución estacionaria no trivial del sistema (3) a (5) es inestable.*

Cuando $\eta = 1$ se obtiene un valor propio cero, caso critico donde pueden existir orbitas periódicas. Acorde con estos resultados no existen dos trayectorias que unan las dos soluciones estacionarias y en consecuencia no existe un bifurcación homoclínica lo cual implica que no existen ondas viajeras. Concluimos la siguiente conjetura:

Conjetura 2 *Si $\eta < 1$ y $\eta > 1$ las soluciones estacionarias del sistema (1)-(2) son inestables entonces no existe bifurcación homoclínica que implica que no existen ondas viajeras.*

4. Descripción de un modelo de simulación con doble difusión

Se extiende el modelo anterior a un espacio bidimensional para la dinámica de crecimiento poblacional del mosquito *A. aegypti*. En este caso las variables y parámetros del modelo son: $m(t, x, y)$: número de mosquitos maduros, $i(t, x, y)$: número de estados inmaduros (huevos, larvas, pupas) en un tiempo t en el punto (x, y) respectivamente; $D = \alpha$: coeficiente de dispersión constante, θ : tasa de estados inmaduros que desarrollan a estado maduro, ϵ : tasa de mortalidad natural de los mosquitos maduros, β : tasa de oviposición por mosquito, K : capacidad de carga de los estados inmaduros, π : tasa de mortalidad natural de los estados inmaduros y las ecuaciones diferenciales parciales no lineales que interpretan la dinámica son:

$$\frac{\partial m(x,y,t)}{\partial t} = \nabla^2 m(x,y,t) + \theta i(x,y,t) - \epsilon m(x,y,t) \tag{11}$$

$$\frac{\partial i(x,y,t)}{\partial t} = \beta m(x,y,t)\left(1 - \frac{i(x,y,t)}{K}\right) - (\theta + \pi) i(x,y,t) \tag{12}$$

donde $\alpha, \theta, \epsilon, \beta, \pi > 0$. El análisis de este modelo de simulación es ahora tema de estudio.

5. Resultados numéricos del modelo unidimensional

se resolvió utilizando el programa MATLAB con las condiciones iniciales, condiciones terminales y valores hipotéticos de los parámetros.

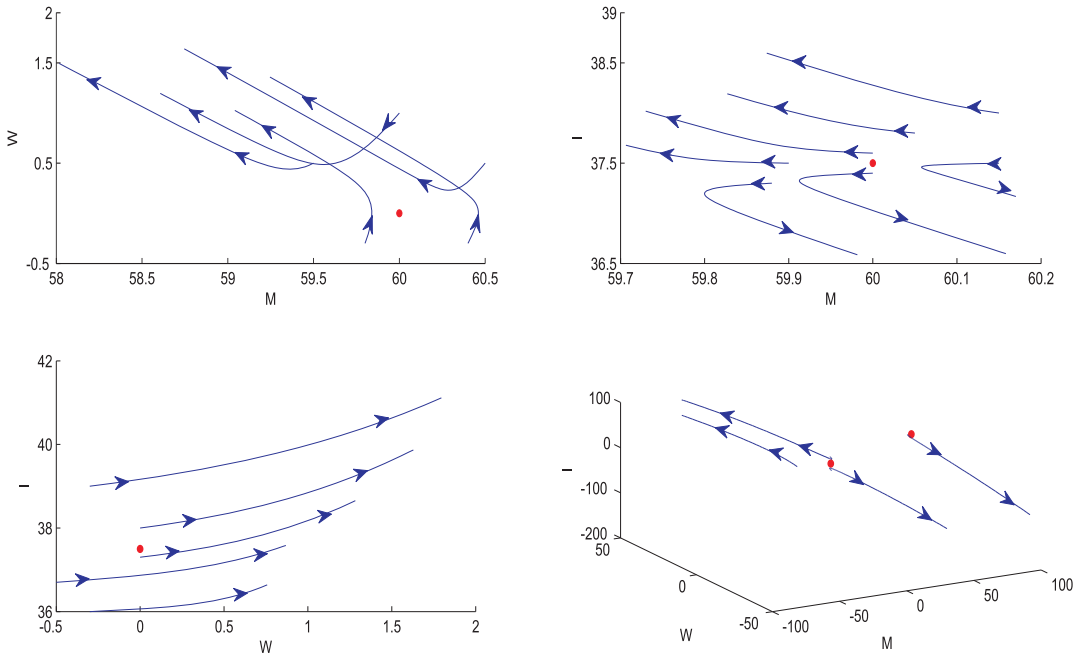


Figura 1: Comportamiento de las variables M, W, I y espacio de fase.

Agradecimientos

Al Programa de Licenciatura en Matemáticas, Facultad de Educación, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío.

Referencias

[1] Busenberg S. and Cooke K. *Vertically transmitted diseases*, Berlin: Springer - Verlag (1988).

[2] Lin G., Hong Y.. *Travelling Wave fronts in a vector disease model with delay*, Applied Mathematical Modelling 32(2008) 2831-2838.

[3] Secretaría de Salud. *Manual para la vigilancia epidemiológica del dengue*, (1984).

[4] Takahashi L. T., Maidana N. A. and Ferreira Jr. W. C. *O Aedes e sua onda*, Biomatemática XIII : 33 - 44, (2003).

[5] World Health Organization. *Manual on environmental management for mosquito control*, Geneva, (1982).

[6] Zhang J.. *Existence of travelling waves in modified vector - disease model*, Applied Mathematical Modelling 33(2009) 626-632.

UN PROBLEMA DE FRONTERA LIBRE PARA EL CRECIMIENTO Y TRATAMIENTO DE TUMORES

Damián A. Knopoff[♭,†], Germán A. Torres[♭,†] y Cristina V. Turner[♭,†]

[♭]*Grupo de Análisis Numérico y Computación, FaMAF, Universidad Nacional de Córdoba, Medina Allende s/n, 5000 Córdoba, Argentina*
[†]*Centro de Investigaciones y Estudios en Matemática - CONICET*

Resumen: En este trabajo se presenta un modelo matemático para el crecimiento de tumores con quimioterapia. Dicho modelo está planteado como un problema de frontera libre, constituida por el borde del tumor. Se tiene un sistema de ecuaciones diferenciales parciales para el número de células tumorales, la concentración de nutrientes, la velocidad del flujo de células y la concentración de droga, siendo el dominio de definición del sistema el propio tumor. En primera instancia, se considera que la forma del tumor es esférica, con simetría radial, con lo cual las variables independientes son el radio y el tiempo.

Palabras clave: *crecimiento tumoral, quimioterapia, frontera libre*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Para analizar el crecimiento de un tumor sólido, se ha considerado un modelo que considera un continuo de células vivas que, a través de cambios de volumen locales originados por nacimiento y muerte celular, crean movimientos descritos por un campo de velocidades. En este modelo se asume que las células nacen y mueren a una tasa que depende de la concentración local de nutrientes, es decir, la muerte celular es un proceso gradual que no ocurre instantáneamente ([1], [2]).

Se considera al tumor como una masa de células alteradas genéticamente que evoluciona, cambiando su tamaño en el tiempo. La forma en que se aborda su crecimiento es a través de los llamados problemas de frontera libre, donde el crecimiento o regresión del tumor se debe a la proliferación de nuevas células o a la muerte de células de acuerdo al nivel de la concentración de nutrientes.

2. ECUACIONES DEL MODELO

Se asume que el tumor es esférico y simétrico, que ocupa la región $\{r \leq S(t)\}$, donde su frontera es una incógnita $r = S(t)$. La concentración de nutriente $c(r,t)$ satisface la ecuación de reacción–difusión. Las otras variables del modelo son la concentración de células vivas $n(r,t)$, la concentración de droga $w(r,t)$ y la velocidad del flujo de células $v(r,t)$. Entonces el crecimiento del tumor esférico simétrico puede ser modelizado, tal como se propone en [2], a través de las siguientes ecuaciones:

$$\frac{\partial n}{\partial t} + \frac{1}{r^2}\frac{\partial (r^2 v n)}{\partial r} = f_1(n,c,w), \tag{1}$$

$$\frac{\partial c}{\partial t} + \frac{1}{r^2}\frac{\partial (r^2 v c)}{\partial r} - \frac{D}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial c}{\partial r}) = f_2(n,c,w), \tag{2}$$

$$\frac{1}{r^2}\frac{\partial (r^2 v)}{\partial r} = f_3(c,n,w), \tag{3}$$

$$\frac{\partial w}{\partial t} + \frac{1}{r^2}\frac{\partial (r^2 v w)}{\partial r} - \frac{D_w}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial w}{\partial r}) = f_4(n,c,w) \tag{4}$$

donde D es el coeficiente de difusión de nutrientes, D_w es el coeficiente de difusión de droga y las f_i son funciones que se derivan de cuestiones de cinética química y biológica, que involucran otros parámetros tales como tasa de mitosis y muerte celular, o efectividad de la droga.

El dominio de definición de las ecuaciones diferenciales es $0 \leq r \leq S(t)$, $t > 0$.

Las condiciones iniciales son: $n(r,0) = n_i(r)$, $c(r,0) = c_i(r)$, $S(0) = S_i$ y $w(r,0) = 0$. Asimismo, las condiciones de borde son, en $r = 0$, $\frac{\partial c}{\partial r} = 0$, $\frac{\partial w}{\partial r} = 0$ y $v = 0$. En tanto que en la frontera libre $S(t)$ se tiene $\frac{dS(t)}{dt} = v(S(t),t)$, $c = c_0(t)$ y $w = w_0(t)$.

3. Resolución numérica del sistema

- **Adimensionalización:** Como primer paso se adimensionalizó el sistema , con lo cual se da una nueva escala a las variables. Por supuesto, esto modifica coeficientes en las EDPs y las condiciones de borde e iniciales.
- **Dominio espacial fijo:** Haciendo el cambio de variables $y = r/S(t)$, se fija el dominio espacial al intervalo $[0,1]$.
- **Crecimiento del tumor avascular sin tratamiento:** En primera instancia se usa el modelo sin droga, es decir considerando ausencia total de droga, hasta un tiempo adimensional de 200 (aproximadamente 25 hs), durante el cual el tumor ha crecido hasta alcanzar un radio adimensional de ≈ 200.
- **Aplicación de la quimioterapia** A partir de allí, se resuelve el sistema completo de ecuaciones, usando un esquema de diferencias finitas (para avanzar en el tiempo con la ecuación para n) y la resolución de un problema con valores en la frontera (para las ecuaciones restantes) usando el paquete **bvp4c** de MatLab.

4. Resultados

En primera instancia se resolvió el sistema del modo descrito en la sección anterior empleando valores de parámetros obtenidos de la bibliografía. De este modo se obtuvieron resultados como el que se muestra en la figura 1,en la cual se observa la concentración de células vivas en función del radio y del tiempo.

Asimismo, manipulando parámetros a priori desconocidos, como por ejemplo uno que determina el *grado de efectividad de la droga* (denotado por α), se resolvió el sistema tomando distintos valores del mismo. En la figura 2 se muestra cómo evoluciona el radio del tumor para distintos valores de dicho parámetro.

Cabe recalcar que precisamente el deconocimiento de estos parámetros es motivo del actual trabajo, que consiste en recuperarlos resolviendo el problema inverso correspondiente ([5]).

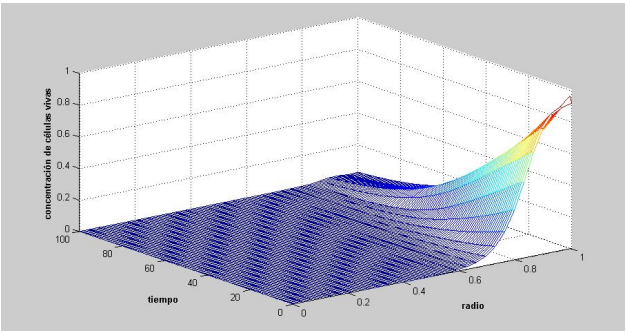


Figura 1: Concentración de células vivas en función del radio y del tiempo

Referencias

[1] J.P. Ward, J.R. King, *Mathematical modelling of avascular tumour growth*, IMA J. Math. Appl. Med. Biol. 14 (1997) 39.

[2] J.P. Ward, J.R. King, *Mathematical modelling of drug transport in tumour multicell spheroids and monolayer cultures*, Math. Biosciences 181 (2003) 177-207.

[3] C. Turner, A. Barrea, *A numerical analysis of a model for growth tumor*, Applied Mathematics and Computation, 167 (2005), 345-354.

[4] H. Byrne, *Dissecting cancer through mathematics: from the cell to the animal model*, Nature Reviews (2010).

[5] J. Agnelli, A. Barrea, C. Turner, *Tumor location and parameter estimation by thermography*, Mathematical and Computer Modelling. In press, doi:10.1016/j.mcm.2010.04.003, (2010).

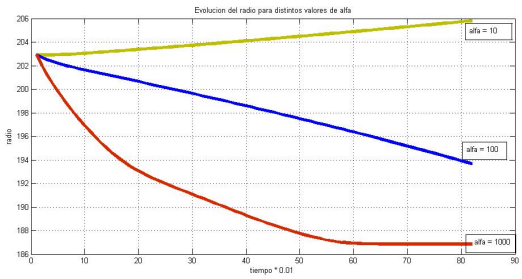


Figura 2: Evolución del radio del tumor para distintos valores de efectividad de la droga

DEPURACIÓN ERITROCITARIA : MODELO MATEMÁTICO

Gustavo W. Vega† ,
†*Escuela de Ciencia y Técnica, Universidad Nacional de San Martín, 25 de Mayo y Francia, San Martín, Provincia de Buenos Aires, Argentina, gwvega@fibertel.com.ar*

Resumen: La permanencia del eritrocito en circulación es regulada por señales que determinan su depuración por el sistema mononuclear fagocítico luego de unos 120 días. No obstante la evidencia acumulada, aún no se han establecido con precisión los mecanismos que expliquen los procesos de senescencia y depuración eritrocitaria. En este trabajo se desarrolla un modelo matemático que intenta reproducir algunos de los aspectos reconocidos de la depuración eritrocitaria, con miras a mejorar la comprensión de problemas de significación clínica.

Mathematical modeling ,red blood cell life span, senescent erythrocyte
AMS Subjects Classification: 92C-02

INTRODUCCIÓN

El eritrocito (ER) es la célula sanguínea especializada en el transporte de oxígeno. Una vez liberado desde la médula ósea a circulación, progresa a través de varios estadios: el reticulocito, inmaduro y de breve duración, el ER maduro y finalmente, un ER senescente, terminal, probablemente de corta duración. Luego de permanecer unos 120 días en circulación, son depurados por el sistema mononuclear fagocítico (SMF). Por otro lado, aproximadamente 3×10^9 nuevos ERs se producen en la médula ósea por kilogramo de peso corporal cada día. Cada conjunto de ERs que es liberado al mismo tiempo, atraviesa simultáneamente los mismos estadios [1] y constituye una cohorte.

En algunas especies, (por ejemplo, el ratón) la depuración eritrocitaria incluye células jóvenes y envejecidas, pero en el hombre, la permanencia en circulación ("esperanza de vida") es similar para toda la cohorte, lo que sugiere un mecanismo de remoción a través de señales específicas y no por efecto del azar. La dependencia estricta del mismo con la edad del ERs, llevó a postular que dichas señales aparecerían abruptamente hacia el final del ciclo eritrocitario y determinarían una rápida eliminación [2]. Entre las señales propuestas, se incluyen: cambios enzimáticos y limitación en la capacidad energética, alteraciones en el balance de calcio, cambios en la carga de superficie, injuria oxidativa, anticuerpos autólogos contra la membrana y asimetría en la composición fosfolipídica de la misma. Sin embargo, parámetros tales como la superficie del ER, el contenido de hemoglobina o de enzimas, muestran cambios progresivos con la edad [3]. Del mismo modo, la disparidad entre la curva experimental de sobrevida eritrocitaria y las predichas suponiendo una esperanza de vida única (eliminación edad-dependiente) [4], indicarían la necesidad de utilizar modelos más complejos. No obstante la evidencia acumulada, no se han establecido aún claramente los mecanismos que expliquen los procesos de senescencia y depuración del ER [5].

El estudio de la depuración eritrocitaria reviste utilidad tanto en investigación como en clínica. En particular, contribuye a definir la fisiopatología de las enfermedades hemolíticas y probablemente sea un elemento a considerar en el control de la diabetes mellitus [2,6].

Este trabajo, que continúa el presentado en II MACI 2009 [7], tiene por objetivo desarrollar un modelo matemático que describa las características conocidas más relevantes de la depuración eritrocitaria.

DESARROLLO

Sea una cohorte de ERs, es decir, un conjunto de ERs de la misma edad. Su estancia en circulación se simula mediante la sucesión ininterrumpida de ciclos durante los cuales sufren una influencia no conocida que les produce una alteración con probabilidad p, que puede revertir por un mecanismo específico con probabilidad r o simplemente mantenerse sin cambio alguno con probabilidad q.

El proceso de eliminación es tarea del SMF y ocurre con probabilidad d que depende del número Δ de alteraciones (u es la probabilidad de permanencia). Se define la capacidad límite de alteraciones Ω como el número de alteraciones que una vez superado, determina la segura eliminación del ER. Asimismo, se define la capacidad límite de reversión Λ como la cantidad máxima de reversiones que pueden acumularse. En el siguiente desarrollo p, q, r y u son funciones de Δ. La asignación de probabilidades es,

$$p(1 \le \Delta \le \Omega + 1) = p_\Delta$$
$$r(1 \le \Delta \le \Omega) = r_\Delta \qquad r(\Delta = 0) = 0$$
$$q(0 \le \Delta \le \Omega) = q_\Delta$$

$$p_{\Delta+1} + q_\Delta + r_\Delta = 1 \tag{1.0}$$

$$d(\Delta) = 1 - u(\Delta)$$
$$u(0 \le \Delta \le \Omega) = u_\Delta \qquad u(\Delta \ge \Omega + 1) = 0 \tag{1.1}$$

Estas suposiciones intentan combinar la evidencia, ya citada, de una depuración que se verifica en un período relativamente breve, con otra que, por el contrario, muestra modificaciones progresivas de ciertos parámetros con la edad.

En una cohorte que inicialmente tiene un número n_0 de ERs, sean n_c y m_c el número de ERs que permanecen y que han sido eliminados respectivamente hasta el ciclo c inclusive,

$$n_c = n_0 - m_c \tag{1.2}$$

La relación $V_c = n_c / n_0$ representa la fracción de viables, es decir , la proporción de ERs que permanece en circulación con $\Delta \le \Omega + 1$. La distribución de probabilidades para las fracciones viables (no depuradas) se construye de acuerdo a (se ilustra para los primeros dos ciclos, en el tercero se indica el resultado):

$c = 1$

$p_1(d_1 + u_1) + q_0(d_0 + u_0)$ de la cual sólo se consideran los términos con u (viables)

$p_1 u_1 + q_0 u_0$

$c = 2$

$$p_1 u_1 (p_2 + q_1 + r_1) + q_0 u_0 (p_1 + q_0) = p_2 p_1 u_1 + p_1\{q_1 u_1 + q_0 u_0\} + q_0^2 u_0 + (p_1 r_1) u_1$$

$$p_2 p_1 u_1 (d_2 + u_2) + p_1\{q_1 u_1 + q_0 u_0\}(d_1 + u_1) + q_0^2 u_0 (d_0 + u_0) + (p_1 r_1) u_1 (d_0 + u_0)$$

$$p_2 p_1 u_1 u_2 + p_1\{q_1 u_1^2 + q_0 u_0 u_1\} + q_0^2 u_0^2 + (p_1 r_1) u_0 u_1$$

$c = 3$

$$p_2 p_1 u_1 u_2 (p_3 + q_2 + r_2) + p_1\{q_1 u_1^2 + q_0 u_0 u_1\}(p_2 + q_1 + r_1) + q_0^2 u_0^2 (p_1 + q_0) + (p_1 r_1) u_0 u_1 (p_1 + q_0)$$

$$p_3 p_2 p_1 u_1 u_2 u_3 + p_2 p_1 (q_2 u_1 u_2^2 + q_1 u_1^2 u_2 + q_0 u_0 u_1 u_2) + p_1 (q_1^2 u_1^3 + q_1 q_0 u_1^2 u_0 + q_0^2 u_0^2 u_1)$$
$$+ q_0^3 u_0^3 + p_1\{(p_2 r_2) u_1^2 u_2 + (p_1 r_1) u_1^2 u_0\} + (p_1 r_1)(q_1 u_1^2 u_0 + 2 q_0 u_0^2 u_1)$$

……

y siguientes.

El resultado es una expresión polinomial donde cada término corresponde a una fracción de viables con un cierto número de alteraciones. El total de viables para el ciclo c es:

$$n_c = n_0 \sum_{k=K}^{k=c+1} \sum_{j=0}^{j=j} V_{c,j,k} \tag{1.3}$$

donde:

c es el número de ciclo,

j es el número de reversiones, siendo $0 \le j \le \Lambda$

k es el número de término, siendo $1 \le k \le c + 1$

$K = 1 \iff c < \Omega + 1$

$K = c + 1 - \Omega \iff c \ge \Omega + 1$, reflejando que de acuerdo con las hipótesis, una vez que $c \ge \Omega + 1$, sólo los términos $c + 1 - \Omega \le k \le c + 1$ representan fracciones viables.

A su vez, cada término se puede representar como:

$$V_{c,j,k} = P_{c+1-k} X_{c,j,k} \tag{1.4}$$

donde:

$$P_{c+1-k} = \prod_{\Delta=1}^{\Delta=c+1-k} p_\Delta \tag{1.5}$$

$$X_{c,j,k} = Q_{c,j,k} \cdot R_{c,j,k} \cdot U_{c,k}$$

El coeficiente $X_{c,j,k}$ puede representarse como el producto escalar de tres vectores: $Q_{c,j,k}$, $R_{c,j,k}$, $U_{c,j,k}$, cada uno función de q_Δ, $p_\Delta \cdot r_\Delta$ y u_Δ respectivamente.

Desde luego, toda reversión es precedida por una alteración. Por lo tanto, el primer término con una reversión aparece en $c = 2$, el primero con dos reversiones en $c = 4$, y así sucesivamente. Los ciclos

impares no aumentan el número de reversión *j*, sino que incluyen más términos con *j* equivalente al ciclo inmediato anterior. El número *j* correspondiente al ciclo *c* resulta entonces,

$$j = c/2 \quad (c \text{ par}) \qquad j = 1/2\,(c-1) \quad (c \text{ impar}) \tag{1.6}$$

Es factible calcular cualquier término del ciclo c en base a términos del ciclo $c-1$ empleando la siguiente expresión:

$$V_{c=c,\,j=j,k=k} = P_{c+1-k}U_{c,k}(X_{c-1,\,j,k} + q_{c+1-k}X_{c-1,\,j,k-1} + p_{c+2-k}r_{c+2-k}X_{c-1,\,j-1,k-2}) \tag{1.7}$$

siendo $X_{c,j,k} = 0$ en tanto,

$$k < 2j+1 \quad \vee \quad k > c+1 \quad \vee \quad k < c+1-\Omega \quad \Leftrightarrow \quad c \geq \Omega+1$$
$$\vee \quad j < c/2 \quad (c \text{ par}) \qquad j < 1/2\,(c-1) \quad (c \text{ impar})$$

Por otro lado, cuando $c > 2\Lambda$, la ecuación (1.7) se modifica para aquellos términos con $j = \Lambda$, puesto que, por hipótesis, no es posible añadir nuevas reversiones en los siguientes ciclos. Luego,

$$V_{c=c>2\Lambda,\,j=\Lambda,k=k} = P_{c+1-k}U_{c,k}\{X_{c-1,\Lambda,k} + (1-p_{c+2-k})X_{c-1,\Lambda,k-1}\} \tag{1.8}$$

Finalmente, se contempla la alternativa de que la capacidad de reversión sea afectada por algún proceso que opera en el tiempo, y que determina su disminución. En tal caso, el valor de *j* se establece según:

$1 \leq c \leq 2\Lambda_{máx}$		$j = c/2$ (c par) $\quad j = 1/2\,(c-1)$ (c impar)
$2\Lambda_{máx} < c < C_{lím}$	$\Lambda = \Lambda(c)$	$0 \leq j \leq \Lambda(c)$
$c \geq C_{lím}$	$\Lambda = 0$	$j = 0$

donde
$\Lambda_{máx}$ es la intersección entre las curvas $\Lambda(c)$ y $j(c)$
$\Lambda(C_{lím}) = 0$

Mediante un algoritmo desarrollado en el entorno de *Mathematica*® se calcula la distribución de fracciones de ERs correspondiente a cada alteración, la fracción de viables y la media de alteraciones en intervalos escogidos.

DISCUSIÓN

En el presente trabajo se desarrolla un modelo matemático que incorpora aspectos reconocidos de la depuración eritrocitaria. En [7], se introdujo las suposición de un límite Ω para el número de alteraciones previas a la depuración con $p_\Delta = constante$ y $u(\Delta < \Omega+1) = 1$. Con ello se intentaba asimilar la acumulación progresiva de cambios en los ERs con la depuración abrupta, edad-dependiente, característica de este proceso en el hombre. El total de ERs viables en función del ciclo era descripto aproximadamente por una función gama que asemeja las curvas experimentales. Sin embargo, la permanencia del ER es suficientemente prolongada y el modelo requiere que Ω sea un número grande (si la alteración es un suceso frecuente) o que *p* sea pequeña . Por otro lado, la remoción depende de la capacidad de reconocer los ERs alterados y en tal sentido, sería razonable esperar que la depuración (consecutiva al reconocimiento) se verifique con una probabilidad que varíe gradualmente de cero a uno conforme aumenta el valor de Δ dentro de un intervalo.

Es propia de la naturaleza del funcionamiento los sistemas vivos, la capacidad de revertir cambios que surgen de la interacción con el entorno y que interfieren con alguno de sus mecanismos. Sobre esta base, se incorpora la capacidad de reversión Λ. Como resultado, la probabilidad p_Δ se disocia del límite Ω, dado que, aún cuando el evento de una alteración tuviese probabilidad apreciable, la reversión de las mismas permite la coexistencia con valores pequeños de Ω.

La aceleración del proceso depurativo es una alternativa aquí considerada. Podría ocurrir en tanto la presencia de alteraciones precipita la aparición de otras nuevas. Los fenómenos de cooperatividad observados en bioquímica [8], serían un ejemplo del caso. Otra variante conceptualmente diferente consiste en suponer que es el propio mecanismo responsable de la reparación el que se comprometería luego de un cierto número de ciclos, por cualquier causa endógena o exógena que determine su atenuación. En una

célula carente de núcleo como el ER, tal hipótesis cobraría sentido dada la incapacidad para la recuperación o renovación.

En la próxima etapa, el modelo se aplicará en el estudio de problemas de significación clínica tal como la glicosilación de la hemoglobina, en donde la eliminación de eritrocitos jugaría un papel relevante.

AGRADECIMIENTOS

Agradezco al Dr. E. Serrano y al Ing. G. La Mura por sus sugerencias.

REFERENCIAS

[1] J. KEENER, AND J. SNEYD, *Mathematical Physiology*, pp. 490-495, Springer-Verlag New York Inc., 1998.

[2] R.S. FRANCO, *The measurement and importance of red cell survival*, Am.J.Hematol., Vol 84 (2009), pp.109-114.

[3] S.C. GIFFORD, J. DERGANC, S. SHEVKOPLYAS, T. YOSHIDA, AND M.W. BITENSKY, *A detailed study of time-dependent changes in human red blood cells: from reticulocyte maduration to erythrocyte senescence*, Br. J. Haematol., Vol 135 (2006), pp.395-404.

[4] C.J. LINDSELL, R.S. FRANCO, E.S. SMITH, C.H. JOINER, AND R.M. COHEN, *A method for the continuous calculation of the age of labeled red blood cells*, Am.J.Hematol., Vol 83 (2008), pp.454-457.

[5] J.P. GREER, J. FOERSTER, G.M. RODGERS, F. PARASKEVAS AND B. GLADER, *Wintrobe's Clinical Hematology 12th Ed.*, Vol I, pp. 156-159, Lippincott Williams & Wilkins, 2009.

[6] R.M. COHEN, R.S. FRANCO, P.K. KHERA, E.P. SMITH, C.J. LINDSELL, P.J. CIRAOLO, M.B. PALASCAK, AND C.J. JOINER *Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA1c*, Blood, Vol. 112 (2008), pp. 4284-4291.

[7] G.W. VEGA, *Modelo de depuración eritrocitaria*, presentado en II Congreso de Matemática Aplicada computacional e industrial 2009, 14-16/12/2009, Rosario, Argentina.

[8] J. M. BERG, J.L. TYMOCZKO, AND L. STRYER, *Biochemistry Fifth Edition*, Part I, Chapter 10, pp. 402-443, W.H. FREEMAN New York, 2002.

OPTIMIZACIÓN DE REDES METABÓLICAS

Cecilia I. Paulo, Jimena Di Maggio, Vanina Estrada, M. Soledad Diaz

Planta Piloto de Ingeniería Química (PLAPIQUI), Universidad Nacional del Sur-CONICET, Camino La carrindanga Km 7, Bahía Blanca 8000, Argentina, {cpaulo, jdimaggio, vestrada, sdiaz}@plapiqui.edu.ar

Resumen: en el presente trabajo se formularon modelos de redes metabólicas para distintos grupos de microorganismos. Por un lado se implementó un modelo con enfoque cinético para describir las principales rutas del metabolismo del carbono en Escherichia coli, obteniéndose un sistema de ecuaciones diferencial algebraico no lineal. Por otra parte se planteó un modelo lineal mixto entero (MILP) basado en el análisis de flujos metabólicos para la optimización de la producción de etanol en Synechocystis PCC 6803. La utilización de estos modelos permitió profundizar el conocimiento del comportamiento de dichos microorganismos, constituyéndose en una herramienta de creciente utilidad para la compresión de los mecanismos de regulación y la optimización de la producción de ciertos metabolitos de interés.

Palabras claves: *optimización, redes metabólicas.*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

En los últimos años ha sido posible medir no solo concentraciones de metabolitos extracelulares, sino también intracelulares, así como el nivel de proteínas y sus actividades. El avance en las técnicas experimentales y el consecuente aumento en la cantidad de datos disponibles sobre la dinámica del funcionamiento de células ha abierto el camino para la construcción de modelos de redes metabólicas, tanto en estado estacionario como dinámico, lo cual permite, a su vez, la predicción del comportamiento de microorganismos y constituye una herramienta fundamental en ingeniería metabólica.

En este trabajo se formulan modelos dinámicos para el metabolismo central de carbono en *Escherichia coli* K-12 W3110 [1]. En primer lugar, se ha llevado a cabo un estudio de sensitividad global sobre esta red dinámica, con el objetivo de determinar los principales parámetros a estimar en base a datos experimentales y se ha formulado el problema de estimación de parámetros sujeto al sistema diferencial algebraico. Por otra parte, se han planteado modelos de balance de flujo estacionario para una red metabólica de Synechocystis PCC 6803, para la maximización simultánea de biomasa y producción de etanol.

2. MODELADO MATEMÁTICO DE REDES METABÓLICAS

Los modelos dinámicos de redes metabólicas proveen perfiles temporales para la concentración de metabolitos involucrados en la red. En estos modelos, que se derivan de balances dinámicos alrededor de cada metabolito, las velocidades de reacción se calculan en función de parámetros cinéticos a través de cinéticas tipo Michaelis Menten u otras. El empleo de datos in vivo es esencial para el ajuste de los parámetros, ya que éstos son en general diferentes de los obtenidos de datos in vitro. Las cinéticas de enzimas in vivo se derivan de las concentraciones de metabolitos bajo una perturbación tipo pulso. En este trabajo, se ha formulado un modelo dinámico para el camino Embden-Meyerhof-Parnas, el ciclo de las pentosas fosfato y el sistema fosfotransferasa en *Escherichia coli* K-12 W3110, basado en el modelo propuesto por Chassagnole et al. [1]. Entre las modificaciones introducidas, se encuentra la cinética de la enzima fosfofructokinasa ([4], [5]). El modelo resultante posee 18 ecuaciones diferenciales que corresponden a los balances de masa glucosa extracelular y metabolitos intracelulares, 30 ecuaciones cinéticas y siete ecuaciones algebraicas adicionales para las concentraciones de co-metabolitos. Las Ecs. 1 y 2 muestran balances para glucosa extracelular y metabolitos intracelulares, respectivamente.

$$\frac{dC_{glc}^{ext}}{dt} = D(C_{glc}^{a\lim} - C_{glc}^{ext}) + f^{pulso} - \frac{C_X r_{PTS}}{\rho_X} \tag{1}$$

$$\frac{dC_i}{dt} = \sum_{j \in S_{enz}} \nu_{ij} r_j - \mu C_i \quad i = 1,.., NC_i \tag{2}$$

Donde μ es la velocidad específica de crecimiento del microorganismo, D es la tasa de dilución, C_i es la concentración del metabolito i, r_j corresponde a la expresión cinética para la velocidad de reacción de la enzima j, ν_{ij} hace referencia al coeficiente estequiométrico del metabolito i en la reacción j, C_{glc}^{alim} y C_{glc}^{ext} corresponden a la concentración de glucosa en la alimentación al reactor y en el medio extracelular, respectivamente, por último los parámetros C_x y ρ_X hacen referencia a la concentración y densidad de biomasa, respectivamente.

2.1. ANÁLISIS DE SENSITIVIDAD GLOBAL EN REDES METABÓLICAS

Como primer paso, se ha realizado un análisis de sensitividad global sobre el modelo de la red metabólica, mediante técnicas basadas en varianza para identificar los parámetros más influyentes, y cuáles de estos parámetros impactan en mayor medida en las salidas del modelo [5]. El análisis de sensitividad global provee información en las salidas del sistema cuando se explora simultáneamente el espacio completo de variación de los parámetros, muestreando desde la función de distribución asociada a cada parámetro de entrada y realizando repetidas simulaciones del modelo. También se pueden identificar interacciones entre los parámetros. El método es apropiado para el análisis del modelo altamente no lineal formulado para la red metabólica, ya que no se requiere realizar la hipótesis de linealidad o aditividad. Los índices de sensitividad para cada parámetro se calculan siguiendo la aproximación de Sobol´ [8], que emplea métodos de simulación Monte Carlo para el cálculo de perfiles temporales de las varianzas condicionales con respecto a los parámetros de entrada. Posteriormente se calculan los perfiles temporales de los índices de sensitividad de primer orden. Las simulaciones Monte Carlo fueron implementadas en g-PROMS [9] mientras que el cálculo de las varianzas y de los índices de sensitividad se realizó en Fortran 90. La Fig. 1a muestra el perfil temporal de los índices de primer orden (S_y) para la concentración de 6-fosfogluconato (C_{6pg}), mientras que la figura 1b muestra los perfiles de S_y para la concentración de piruvato (C_{pyr}).

2.2. ESTIMACIÓN DINÁMICA DE PARÁMETROS

Basado en los resultados del análisis de sensitividad y en datos experimentales de concentraciones de metabolitos intracelulares tomados con una frecuencia de 4-5 por segundo [2], se formuló el problema de estimación de parámetros en g-PROMS [9] como un problema de estimación de parámetros de máxima verosimilitud (Maximum Likelihood), tal como se muestra en la Ec. 6.

$$\min \phi = \frac{N}{2} \ln(2\pi) + \min_{\theta} \sum_{i=1}^{NE} \sum_{j=1}^{NV} \sum_{k=1}^{NM} \left[\ln(\sigma_{ijk}^2) + \frac{(C_{ijk}^M - C_{ijk})^2}{\sigma_{ijk}^2} \right]$$

$$s.t. \tag{6}$$

$$\frac{dC_{glucose}^{ext}}{dt} = D(C_{glucose}^{feed} - C_{glucose}^{ext}) + f_{pulse} - \frac{C_x r_{PTS}}{\rho_x}$$

$$\frac{dC_i}{dt} = \sum_{j \in S_{enz}} \nu_{ij} r_j - \mu C_i \quad i = 1,.., NC_i$$

$$0 \le C_i \le C_i^U, \quad p_j^L \le p_j \le p_j^U, \quad C_i = C_i^0$$

Los 11 parámetros estimados corresponden a velocidades máximas de reacción y constantes a mitad de saturación y de inhibición para algunas enzimas que participan en la red metabólica (r_k). Las figuras 2a y 2b muestran los perfiles temporales obtenidos por simulación junto a los datos experimentales para las concentraciones de fructosa-6-fosfato y fructosa-1,6.difosfato, respectivamente.

3. MAXIMIZACIÓN DE LA PRODUCCIÓN DE ETANOL MEDIANTE CIANOBACTERIAS

El potencial de las microalgas como fuente de energía renovable ha recibido considerable interés ya que la emisión de CO_2 por combustión es prácticamente la misma que se consume para su producción mediante fotosíntesis, a la vez que no compiten con tierras dedicadas a la producción de alimentos. Las algas pueden crecer en agua dulce o salada y el 80% del agua puede ser reciclada. Se han formulado sucesivos modelos de redes metabólicas para la cianobacteria Synechocystis PCC 6803, en estado estacionario, para la producción heterotrófica y autotrófica de etanol, con el objetivo de maximizar simultáneamente la producción de etanol y la biomasa. Para ello se ha considerado la posibilidad de adición de los genes de piruvato descarboxilasa (pdc) and alcohol deshidrogenasa II (adhII) de Zymomonas mobilis en Synechocystis sp. PCC6803 [3]. Se han asociado variables binarias a las reacciones potenciales a llevarse a cabo dentro de la red metabólica [7]. Se ha reformulado el problema de optimización de dos niveles en un solo nivel tal como fue propuesto por Burgard y Maranas [10], obteniéndose un problema de programación lineal mixto entera, con 3837 variables y 827 variables binarias el cual fue implementado en GAMS (Brooke et al., 2005) y resuelto con CPLEX.

La Fig. 3 muestra un esquema simplificado de los principales flujos metabólicos obtenidos para el crecimiento autotrófico (CO_2 como fuente de carbono y luz). Se puede observar que tanto la fotosíntesis como la fermentación constituyen vías activas en el microorganismo debido a la inserción de los genes anteriormente mencionados. El modelo propuesto, basado en la información genómica disponible, permite la determinación de posibles knockouts de genes para la maximización de la producción de etanol, así como de otros flujos metabólicos involucrados en la red.

4. CONCLUSIONES

Se han formulado modelos dinámicos y estacionarios para redes metabólicas de E. coli y Synechocystis sp., respectivamente. El ajuste de dichos modelos con datos experimentales permite su empleo como herramienta fundamental en la guía de experimentos en ingeniería metabólica. Se está trabajando en el desarrollo de modelos integrados de bioreactor y red metabólica [6].

5. FIGURAS

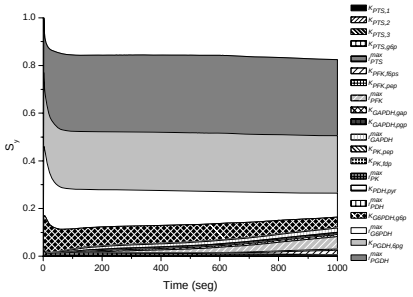


Figura 1a. Perfil de S_y para C_{6pg}

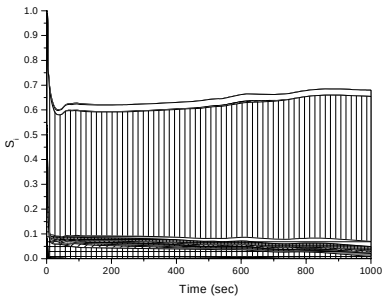


Figura 1b. Perfil de S_y para C_{pyr}

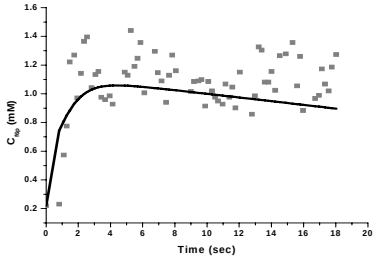


Figura 2a. Perfil temporal de C_{f6p} simulado y medido

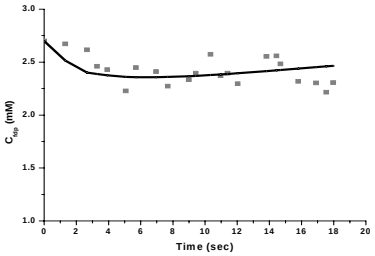


Figura 2b. Perfil temporal de C_{fdp} simulado y medido

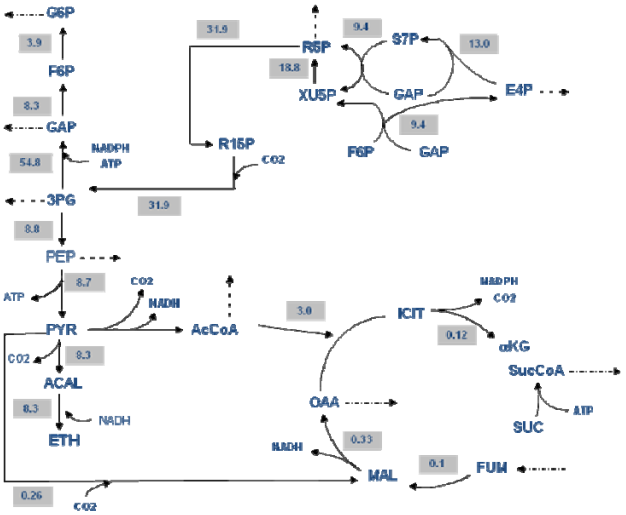


Figura 3. Principales flujos metabólicos en Synechocystis PCC 6803 para crecimiento autotrófico.

AGRADECIMIENTOS

Los autores agradecen el soporte económico de la Comisión Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional del Sur y de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCYT), Argentina.

REFERENCIAS

[1] C. CHASSAGNOLE, N. NOISOMMIT-RICCI, J. SCHMID, K. MAUCH, M. REUSS, *Dynamic Modeling of the Central Carbon Metabolism of Escherichia coli*, Biotechnology and Bioengineering, 79, 1 (2002), pp.54-73.

[2] D. DEGENRING, C. FROEMEL, G. DIKTA, R. TAKORS, *Sensitivity analysis for the reduction of complex metabolism models*, J Process Control, 14 (2004), pp.729-745.

[3] J. DEXTER, FU, P. (2009). *Metabolic engineering of cyanobacteria for ethanol production*, Energy and Environmental Sciences, DOI: 10.1039/b811937f.

[4] J. C. DIAZ RICCI, 1996, *Influence of Phosphoenolpyruvate on the Dynamic Behavior of Phosphofructokinase of Escherichia coli*, J Theoretical Biology, 178, 2 (1996), pp.145-150.

[5] J. DI MAGGIO, J.C. DIAZ RICCI, M.S. DIAZ, *Global Sensitivity Analysis in dynamic metabolic Networks*, Comp. & Chem. Eng., 34 (2010), pp.770–781.

[6] J. LEPPÄVUORI, M.S. DIAZ, L. BIEGLER, M. DOMACH, *Dynamic Unit Operation Model for Industrial Fermentation Processes*, (2010) AIChE Annual Meeting, Salt Lake City.

[7] C. PAULO, J. DI MAGGIO, V. ESTRADA, M.S. DIAZ, *An MILP Approach to the Optimization of Cyanobacteria Metabolic Network for Bioethanol Production*, (2010) AIChE Annual Meeting, Salt Lake City.

[8] I.M. SOBOL', *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, Math Comp in Simulation 55 (2001), pp.271-280.

[9] g-PROMS, http://www.psenterprise.com

[10] A. BURGARD, P. PHARKYA, C. MARANAS, *OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization*, Biotechnology and Bioengineering 84, 6 (2003), pp. 647-657.

SELECCIÓN DE MUESTRAS RELEVANTES EN ESPECTROSCOPÍA NIR PARA ANÁLISIS DE CAÑA DE AZÚCAR

Natalia Sorol‡, Jorge Gotay Sardiñas†, Jorge Bustos†, Adrián Will †

†Facultad de Ciencias Exactas y Tecnología – Fa.C.E.T., Universidad Nacional de Tucumán, Tucumán, Argentina
‡Estación Experimental Agroindustrial Obispo Colombres, Las Talitas, Tucumán, Argentina

Resumen: El análisis de muestra de jugo de caña de azúcar por espectroscopía NIR es una técnica importante por su rapidez, economía, y bajo impacto ambiental: Las pruebas tradicionales de laboratorio consumen tiempo, reactivos, y en particular emplean plomo como parte del proceso. La espectroscopía NIR es una metodología ampliamente difundida en otros países del mundo debido a estas ventajas, por lo que la EEAOC decidió implementarla e impulsar su uso en la industria local.
Este análisis se realiza a partir de una base de referencia de espectros medidos. La EEAOC dispone en la actualidad de una base de datos de 8500 muestras, y se agregan anualmente unas 1000, con características que en principio difieren año a año, ya que la calidad, composición y características del jugo de caña de azúcar dependen de varios factores. Esta base está resultando inmanejable para los equipos y software disponibles.
Se plantea el problema de desarrollar un método de clustering que permita seleccionar y controlar las muestras relevantes de la base de datos, manteniendo la calidad de los resultados que se obtienen con el equipo y mejorando la velocidad de procesamiento del mismo. Se resolvió el problema utilzando SVD y k-means++, sobre la base de 1970 muestras, con resultados satisfactorios.

Palabras claves: *Espectroscopía, Industria de Azúcar de Caña, Clustering, SVD, K-MEANS++*

1. INTRODUCCIÓN

La espectroscopia de infrarrojo cercano (NIRS) estudia la interacción de la radiación electromagnética con la materia, comprendiendo el segmento de luz de longitudes de ondas entre 800 y 2600 [nm]. Analiza la absorción de energía en dicha región por los grupos funcionales de las moléculas de la muestra produciendo su vibración, constituyendo de esa manera un conjunto de valores de absorbancia o transmitancia a diferentes longitudes de onda que dan lugar a un espectro característico para cada muestra. El mismo contiene información valiosa si se dispone de tratamientos matemáticos y estadísticos apropiados, como la quimiometría, que es la disciplina química que utiliza la matemática, la estadística y otros métodos, empleando la lógica formal para diseñar o seleccionar procedimientos de medida óptimos y proporcionar la información química relevante mediante el análisis de los datos (Massart 2008).
Esta metodología está ampliamente difundida en todo el mundo por su alto potencial para determinaciones cualitativas y cuantitativas en una gran variedad de áreas de aplicación. Permite realizar análisis rápidos de muchos componentes en una muestra con un mínimo de preparación; no es destructiva ni emplea reactivos químicos, disminuye el error del operador y requiere menos mano de obra que los métodos tradicionales empleados en el laboratorio.
Sin embargo, se debe tener presente que es un método secundario, lo cual significa que debe ser calibrado contra otras metodologías y que sus respuestas no presentarán mayor exactitud que la de los métodos primarios empleados (Rein 2007), además los equipos tienen un costo elevado, el operador de los mismos debe poseer alta competencia técnica y se requiere bastante tiempo para desarrollar una base de datos adecuada.
En Tucumán existe un creciente interés de empresas del sector azucarero en utilizarla para el control de calidad de la caña de azúcar, en el pago a productores, y en el proceso de elaboración. Por este motivo la Estación Experimental Agroindustrial Obispo Columbres (EEAOC) comenzó a implementar esta metodología desde el año 2005 en el Laboratorio de Investigaciones Azucareras. Desde ese año y hasta el 2009 se procesaron muestras de caña limpia, sin “trash”, en un trapiche de planta piloto, con un molino de tres rodillos y un nivel de extracción comprendido entre 57% y 65%. El jugo primario de 8548 muestras provenientes de las distintas variedades de caña evaluadas fue analizado en forma paralela por las técnicas tradicionales y por la metodología NIRS, para evaluar los dos parámetros estudiados: Pol % jugo y Brix Refractométrico.
Los métodos de referencia primarios empleados en nuestro laboratorio fueron:

- Pol % jugo: utilización de un polarímetro digital Optical Activity, modelo Polar 2001 (Chen, J. C. P. 1985. Cane Sugar Handbook. 11. ed. John Wiley & Sons, New York, USA.)
- Brix refractométrico (BrixR): mediante un refractómetro digital marca Leica, modelo AR600 (Chen, J. C. P. 1985. Cane Sugar Handbook. 11. ed. John Wiley & Sons, New York, USA.)

Todas las muestras fueron escaneadas en el rango de longitudes de onda de 400-2500 [nm] en un instrumento monocromador Foss NIRSystem 6500 (Silver Spring, Maryland, USA) para muestras líquidas (detector de transmitancia). Se emplearon los softwares ISIscan 2.21 para la obtención de los espectros, y WinISI III para los modelos de calibración correspondientes.
Las muestras evaluadas son representativas de diversos factores de importancia para la caña cosechada, como por ejemplo el grado de maduración, el tipo de suelo, la variedad y las condiciones climáticas, los cuales inciden en la obtención de un modelo de calibración más robusto.
La gran cantidad de muestras analizadas hace imposible o dificulta el desarrollo de los modelos de calibración por la elevada cantidad de información contenida en los complejos espectros NIR lo que hace necesario buscar una manera de clusterizar y seleccionar muestras representativas (evitando tendencias y muestras redundantes) para disminuir el tamaño de la base de datos sin perder información valiosa y llevar a cabo la calibración de la mejor manera posible. Es importante destacar que existen soluciones tecnológicas al problema, pero éstas requieren de inversiones poco factibles en la industria de nuestro país en la actualidad.

2. SELECCIÓN DE MUESTRAS RELEVANTES MEDIANTE SVD Y K-MEANS++

Se procesaron para este trabajo 1970 de las 8548 muestras disponibles correspondientes a muestras de caña de azúcar limpia de la zafra 2005-2006, analizadas por el espectoscopio NIR. De cada muestra se tienen los datos de laboratorio (Pol y Brix), y el espectro de transmitancia entre 400 y 2498 [nm], medidos cada 2 [nm], dando un total de 1050 frecuencias medidas por muestra. Siguiendo los procedimientos utilizados por el espectroscopio, se procedió a limpiar los datos de la siguiente manera:

1. Recorte del espectro a la ventana 1200 a 2300 [nm], que es la ventana recomendada para el análisis de muestras de caña de azúcar.
2. Aplicación de smoothing mediante un filtro de Savitzky-Golan, con span igual a 21 y grado 4. Esto está indicado en espectroscopía para reducir el ruido sin afectar la altura y amplitud de los picos del espectro.
3. Aplicación de la derivada segunda numérica indicado en espectroscopia para mejorar la resolución.
4. Centrado por la media (la normalización no está indicada, según Varmuza 2009).

Se calculó entonces el SVD de la matriz resultante con la finalidad de reducir el tamaño del problema (PCA Varmuza 2009, Elden 2007), y mejorar la velocidad de convergencia del algoritmo de k-means clustering. En nuestro caso, tanto el paso de aplicación del SVD como los procedimientos de limpieza resultan críticos dado que la matriz original de espectros es densa y el procedimiento se hace rápidamente impracticable.
Se procedió a elegir 12 valores singulares con lo que se explica el 99,90% de la varianza de la matriz (Figura 1). Esto reduce el problema al clustering a una matriz sparse 1970x12, lo cual es mucho más simple y rápido de realizar. Finalmente se aplicó el algoritmo de k-means++ (Arthur 2007) para encontrar los candidatos a centros iniciales y el algoritmo de k-means para encontrar los clusters.

3. RESULTADOS OBTENIDOS

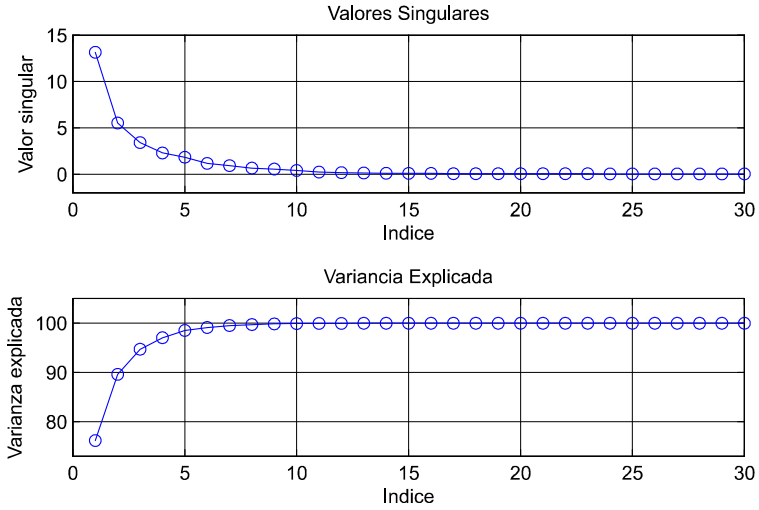


Figura 1

En el gráfico superior de la figura 1 se representan los 30 primeros valores singulares. En el gráfico inferior se representa el porcentaje de varianza explicada para los 30 primeros valores singulares. Para un índice determinado, tenemos el porcentaje de varianza explicada utilizando todos los valores singulares hasta el índice en cuestión. Observamos que utilizando solamente los 12 primeros índices se puede explicar la variación total en un 99%.

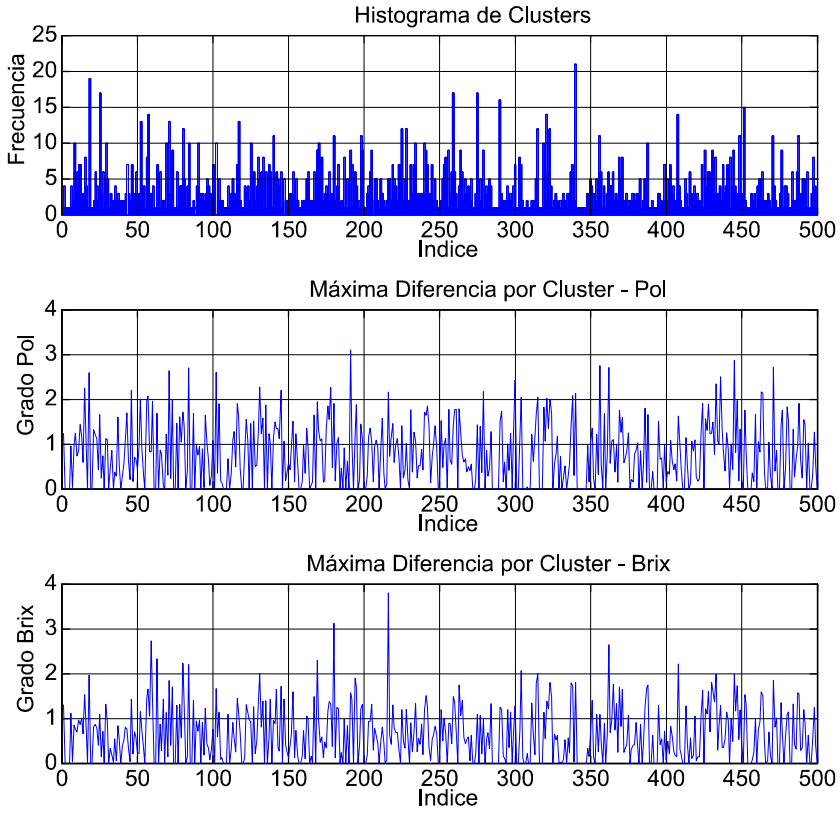


Figura 2

Se efectuaron pruebas para determinar el número k de clusters, y se encontró que k = 500 era el óptimo para el conjunto de datos, ya que con valores menores el error de Pol y Brix aumenta, y valores mayores no mejoran significativamente estos errores. Con este valor se encontró que en la mayoría de los casos los clusters son pequeños (entre 2 y 10 individuos), como se ve en el histograma de la figura 2, gráfico superior. En los dos gráficos siguientes se representan los errores máximos encontrados en cada cluster, de acuerdo a:

$$\max_{y \in C_i} \left(\left| \mathrm{Pol}(y) - \overline{\mathrm{Pol}(C_i)} \right| \right) \quad ; \quad \max_{y \in C_i} \left(\left| \mathrm{Brix}(y) - \overline{\mathrm{Brix}(C_i)} \right| \right)$$

donde C_i es el cluster i. O sea, la máxima diferencia entre el valor de Pol de un punto del cluster, y el valor de Pol promedio del cluster (y similarmente para los valores de Brix en el gráfico inferor). Se observa que el error de Brix es menor que el error de Pol.

4. CONCLUSIONES

El rápido aumento de la cantidad de muestras, unas 1000 muestras al año, hace imprescindible el diseño de un procedimiento externo que permita la selección de una cantidad razonable de muestras a permanecer en la base de datos del aparato. El presente trabajo representa un importante avance en la dirección correcta ya que permite la selección y preservación externa de muestras, lo que va a permitir el análisis e incorporación de muestras relevantes, y la eliminación de muestras redundantes del aparato, de manera de mantener una base interna de tamaño razonable. Más aún, la supervisión de los procedimientos de selección de muestras relevantes es crítica, dada la necesidad de conservar en la base de datos de referencia muestras de características especiales (heladas, deterioro por enfermedades específicas, etc.).

El presente trabajo es parte de una investigación en curso, que comprende varios aspectos:

- Los resultados en cuanto a precisión de Pol y Brix están de acuerdo a los estándares internacionales, pero para aplicaciones específicas sería deseable obtener resultados más precisos (con menos variación de Pol y Brix para muestras dentro de un mismo cluster).
- Se utilizó sólo una parte de los datos disponibles debido a que el problema consiste en desarrollar una metodología de trabajo, y no en resolver este problema en particular.
- Es necesario agilizar y mejorar el proceso, tanto de cálculo del SVD como de clusterización, para que resulte posible y práctico escalar a tamaños superiores. La presencia de una base de datos de referencia representativa y de tamaño razonable, agiliza en gran medida el trabajo de laboratorio.
- Resulta imprescindible responder este problema y otros similares a nivel laboratorio, ya que para la extensión de la tecnología a la industria es menester asegurar una baja inversión a largo plazo y brindar una metodología de trabajo establecida.

Resulta claro que existen soluciones tecnológicas al problema, pero requieren de importantes inversiones, poco realizables en la industria de nuestro país en la actualidad. El desafío consiste entonces en encontrar soluciones que puedan llevarse a cabo con los equipos disponibles, permitiendo lograr resultados similares a los que se obtendrían con tecnologías más modernas.

AGRADECIMIENTOS

Se agradece a la Estación Experimental Agroindustrial Obispo Columbres por el uso de los datos. Los autores están soportados por el proyecto CIUNT 26/E457 de la Universidad Nacional de Tucumán. Se agradece a la Dra. Silvia Zossi, responsable del laboratorio de Investigaciones Azucareras de la EEAOC, y a la Mg. Ma. Marcela Lazarte por su colaboración.

REFERENCIAS

[1] L. ELDEN, "MATRIX METHODS IN PATTERN RECOGNITION," *SIAM SERIES ON FUNDAMENTALS OF ALGORITHMS.* SIAM, PHILADELPHIA, 2007. ISBN 978-0-898716-26-9
[2] K. VARMUZA & P. FILMZOSER: ·"INTRODUCTION TO MULTIVARIATE STATISTICS IN CHEMOMETRICS", CRC PRESS, 2008. ISBN 978-1-4200-5947-2
[3] D. MASSART: "CHEMOMETRICS: A TEXT BOOK", ELSEVIER, 1988
[4] P. REIN: "CANE SUGAR ENGINEERING". ED. BARTENS, BERLIN, GERMANY, 2007
[5] D. ARTHUR & S. VASSILVITSKII: "K-MEANS++: THE ADVANTAGES OF CAREFUL SEEDING", *PROCEEDINGS OF THE EIGHTEENTH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS SODA 2007*, SIAM, PHILADELPHIA, 2007

DYNAMICAL AND STATISTICAL ANALYSIS OF SPIKING NEURONS

Yudy Carolina Daza Caro and Inés Samengo

Centro Atómico Bariloche and Instituto Balseiro, (8400) San Carlos de Bariloche, Río Negro, Argentina

Abstract: Single neurons are mathematically described by systems of differential equations. When driven with strong constant currents, these systems have a limit-cycle attractor, corresponding to regular spiking. In realistic conditions, however, neurons receive fluctuating input currents through their dendritic afferents. In this work, we characterize the statistical properties of the stochastic input currents that are most effective in inducing spiking. We simulate a Hodgkin-Huxley neuron, and find that for supra-threshold stimulation, neurons preferentially spike when the stochastic input resonates with the frequency of the limit cycle. This analysis allows us to link the dynamical properties of single neurons with their coding characteristics.

Keywords: *neuron, dynamical systems, stochastic, Hodgkin-Huxley*
2000 AMS Subject Classification: 92B20

1 INTRODUCTION

Neurons fire or not, depending on the spatiotemporal characteristics of the input signal arriving through their dendrites. Some stimulations patterns induce spiking, whereas others do not. One of the fundamental goals of computational neuroscience is to understand what spikes mean in terms of the input: When a spike is observed, which presynaptic stimulus is likely to have occurred?

The mathematical modeling of neural dynamics provides key theoretical tools to answer this question. A neuron in the nervous system is represented by a system of differential equations describing how the voltage of the cell evolves in time, as well as some other internal variables representing biophysical properties of the molecules regulating charge exchange in the cell membrane. The most widely used dynamical model describing neuronal dynamics was developed by Nobel-prize winners Alan Lloyd Hodgkin and Andrew Huxley in 1952 [1], to describe the voltage fluctuations of the giant axon of the squid. The equations read

$$\begin{aligned} C\frac{\mathrm{d}V}{\mathrm{d}t} &= -g_{\mathrm{L}}(V-V_{\mathrm{L}}) - g_{\mathrm{K}}n^4(V-V_{\mathrm{K}}) - g_{\mathrm{Na}}m^3h(V-V_{\mathrm{Na}}) + I(t), \\ \frac{\mathrm{d}n}{\mathrm{d}t} &= -\left[n-n_\infty(V)\right]/\tau_{\mathrm{n}}(V), \\ \frac{\mathrm{d}m}{\mathrm{d}t} &= -\left[m-m_\infty(V)\right]/\tau_{\mathrm{m}}(V), \\ \frac{\mathrm{d}h}{\mathrm{d}t} &= -\left[h-h_\infty(V)\right]/\tau_{\mathrm{h}}(V). \end{aligned} \tag{1}$$

Here, V represents the voltage of the neuron, while the other variables are the activating (n and m) and inactivating (h) probability of each membrane-channel subunit. The parameters appearing in Eq. 1 can be found in the original publication [1]. The variables representing the model neuron, hence, evolve in a 4-dimensional space.

In the system 1, $I(t)$ appears as an additive term in the first equation. This term represents the total weighted sum of all the external currents arriving to the neuron through presynaptic afferents. Cortical neurons typically receive asynchronous signals from other tens of thousands of neurons. Their weighted sum therefore is typically a fluctuating signal. It makes sense, hence, to model $I(t) = I_0 + \xi(t)$, where I_0 is the mean input, and $\xi(t)$ is Gaussian white noise of zero mean and variance σ^2.

2 RESULTS

In order to understand the coding properties of Hodgkin-Huxley neurons, we first describe their dynamical properties in the absence of noise ($\sigma = 0$). For $I_0 > 9\mu$ A / cm^2, the dynamical system 1 has a single limit-cycle attractor. The periodic circulation along the limit cycle represents repetitive spiking, when all

the neuronal variables vary periodically with a well-defined frequency that depends on the strength of I_0. As soon as noise is incorporated $\sigma > 0$, periodicity is disrupted, and a certain degree of irregularity is observed in the spiking times. Some noise segments favour spiking, whereas others disrupt it. In the present framework, the fundamental question of computational neuroscience becomes: Which noise patterns favour spiking? One way to answer this question, is to calculate the average stimulus preceding a spike: The so-called spike-triggered average (STA). In Fig. 1 we depict the STA of a Hodgkin-Huxley neuron, for different noise levels. The most noticeable characteristic of the curves in Fig. 1 is their damped periodicity. In all

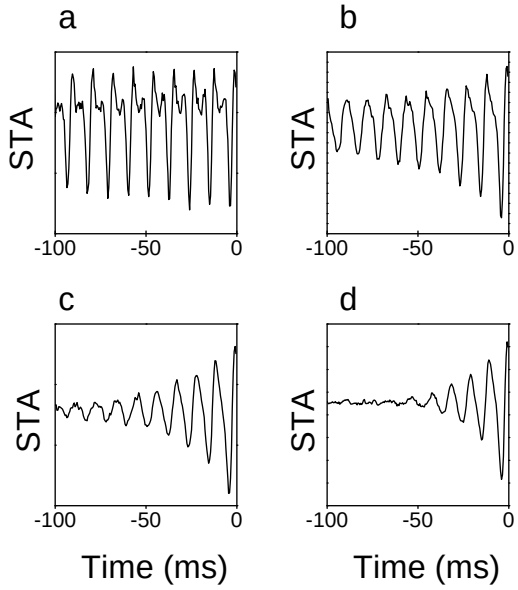


Figure 1: Spike-triggered average (STA) of a Hodkin-Huxley neuron, for different noise levels. In all cases, $I_0 = 29.6$ μ A/cm^2. (a) $\sigma = 1$ μA/cm^2 ms$^{1/2}$, (b) $\sigma = 2.6$ μA/cm^2 ms$^{1/2}$, (c) $\sigma = 5$ μA/cm^2 ms$^{1/2}$, (d) $\sigma = 10$ μA/cm^2 ms$^{1/2}$.

cases, the average stimulus preceding spiking exhibits marked oscillations. The damping constant depends on the noise level, in such a way that high noise is linked with high damping. We also notice that the shape of the curves varies with σ. Whereas high noise is associated to fairly smooth sinusoidal-like oscillations (panel d), low noise discloses a more complex waveform (panel a). To further characterize the changes in shape, in Fig. 2 we show the Fourier transform of the plots depicted in Fig. 1. We see that high-noise STAs contain a single peak in their spectrum, whereas low-noise STAs exhibit multiple peaks, all of them higher harmonics of the fundamental frequency. The fundamental frequency remains unchanged, as the noise level increases.

These results imply that Hodgkin-Huxley neurons, when stimulated with a supra-threshold I_0 and noise, are mainly selective to noise stretches that have large power in one basic frequency, and sometimes, also in its harmonics. These neurons operate, hence, as frequency detectors, or *resonators* [2]. The natural question that follows, then, is what determines the fundamental frequency of the STA. In Fig. 3, we show this frequency as a function of the firing rate of the unperturbed system (that is, for $\sigma = 0$), for several modified neurons. Data points of different colour represents different model neurons, obtained by modifying the value of the maximal conductance to sodium $g_{\rm Na}$ in Eqs. 1. Black circles represent the standard model. Different points sharing the same colour represent one model neuron driven with different values of I_0, and thus associated to different firing rates.

There we observe a perfect match beetween the fundamental frequency of the oscillations in the STA and the firing rate. Hence, noisy Hodgkin-Huxley neurons are selective to stimulus stretches containing high power in the frequency corresponding to the limit cycle of the system.

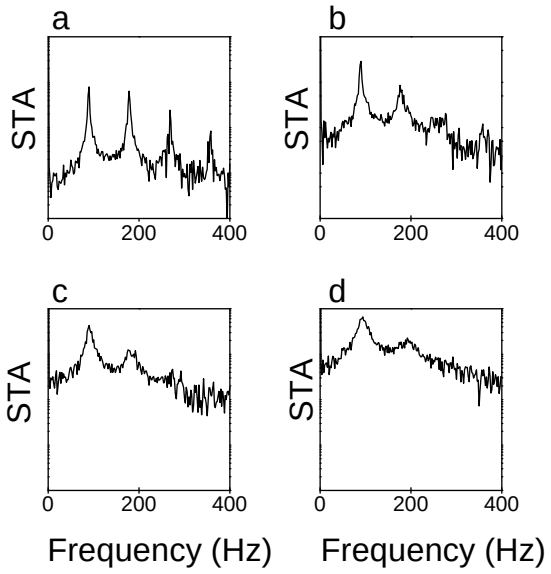


Figure 2: Fourier transform of the STA traces displayed in Fig. 1 (same parameters).

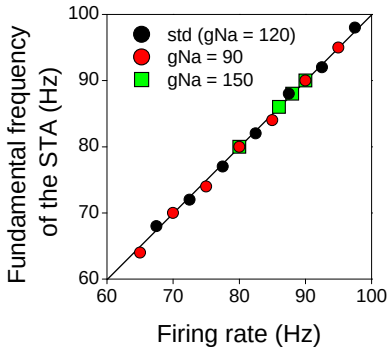


Figure 3: Fundamental frequency of the STA, as a function of the firing rate of the unperturbed stimulus (that is, $\sigma = 0$). Different colours represent different model neurons. The diagonal is the identity line.

3 Conclusion

The mathematical analysis of a Hodgkin-Huxley neuron has allowed us to connect the dynamical and the computational properties of the neuron. The dynamics of the system determine the frequency of the limit cycle, for a given I_0. When noise is added to the input current, the neuron becomes selective to those particular noise stretches that contain high power around the frequency of the limit cycle. These neurons, hence, are frequency detectors that resonate with the fundamental frequency of the limit cycle.

Acknowledgments

This work was supported by CONICET, CNEA, ANPCyT and Universidad Nacional de Cuyo.

References

[1] A. L. Hodgkin, and A. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. 117 (1952) pp 500-544.

[2] E. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. The MIT Press, 2007.

NEURAL DYNAMICS IN THE PRESENCE OF NOISY INPUTS

Soledad Gonzalo Cogno and Inés Samengo

Centro Atómico Bariloche and Instituto Balseiro, (8400) San Carlos de Bariloche, Río Negro, Argentina

Abstract: The mathematical description of neuron dynamics is based in systems of differential equations representing the internal biophysical parameters of the cell. The input current arriving to the dendrites of the neuron is modelled as a control parameter in the dynamical system. When driven with constant stimuli, neurons rapidly approach a limit cycle and fire periodically. In realistic conditions, though, neurons are subject to noise. Thus, the limit-cycle dynamics is perturbed by stochastic fluctuations. The classical theory of oscillatory dynamics predicts that white-noise input currents do not alter, in average, the natural frequency of oscillators. Here we present numerical simulations of a Hodkin-Huxley neuron, and demonstrate that the firing frequency is indeed altered by noise. We develop the theoretical tools to understand this phenomenon, and show that the firing frequency of the perturbed system can be derived from the unperturbed case, using averaging techniques.

Keywords: *neural dynamics, stochastic, Hodgkin-Huxley*
2000 AMS Subject Classification: 92B20

1 INTRODUCTION

The computational power of the nervous system relies on the oscillator properties of its neurons and networks. Single neurons, when driven with a constant current, set on a limit cycle and fire periodically. In realistic conditions, though, neurons are hardly ever driven by constant currents. The massive synaptic bombardment received through their presynaptic afferents actually resembles a stochastic drive. Therefore, in modelling studies, mathematical methods based on stochastic dynamical systems prove to be remarkably useful. In such a context, the total input current entering a neuron is usually represented as $I(t) = I_0 + \xi(t)$, where $\xi(t)$ is Gaussian white noise of zero mean and variance σ^2.

The most widely used dynamical model describing neuronal dynamics was developed by Nobel-prize winners Alan Lloyd Hodgkin and Andrew Huxley in 1952 [1], to describe the voltage fluctuations of the giant axon of the squid. The equations read

$$\begin{aligned} C\frac{\mathrm{d}V}{\mathrm{d}t} &= -g_{\mathrm{L}}(V - V_{\mathrm{L}}) - g_{\mathrm{K}}n^4(V - V_{\mathrm{K}}) - g_{\mathrm{Na}}m^3h(V - V_{\mathrm{Na}}) + I(t), \\ \frac{\mathrm{d}n}{\mathrm{d}t} &= -\left[n - n_\infty(V)\right]/\tau_{\mathrm{n}}(V), \\ \frac{\mathrm{d}m}{\mathrm{d}t} &= -\left[m - m_\infty(V)\right]/\tau_{\mathrm{m}}(V), \\ \frac{\mathrm{d}h}{\mathrm{d}t} &= -\left[h - h_\infty(V)\right]/\tau_{\mathrm{h}}(V). \end{aligned} \tag{1}$$

Here, V represents the voltage of the neuron, while the other variables are the activating (n and m) and inactivating (h) probability of each channel subunit. The parameters appearing in Eq. 1 can be found in the original publication [1]. The system, hence, evolves in a 4-dimensional space. If I_0 is above 9.8 μ A / cm^2, there is a single limit-cycle attractor. In these conditions, a phase description of the perturbed system is possible, as developed by Yoshiki Kuramoto [2]. In the phase description, the state of the system is represented by a phase ϕ, specifying its position on the limit cycle (see Fig. 1). The original dynamical system of Eq. 1 may thus be reduced to

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 2\pi\omega_0 + \Delta(\phi)\xi_{\|}(t), \tag{2}$$

where $\Delta(\phi)$ is the phase-resetting curve of the system, ω_0 is the natural frequency, and $\xi_{\|}(t)$ is the component of the noise tangential to the limit cycle.

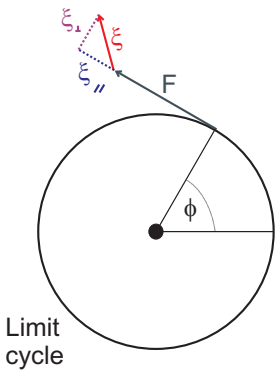


Figure 1: Phase representation of the Hodgkin-Huxley neural model. The force **F** drives the system along the limit cycle. The noise $\xi(t)$ constitutes a perturbation. In classical phase models, only the parallel components is relevant.

Integrating Eq. 2 and averaging over different realizations of the noise one can easily show that if $\xi(t)$ has zero mean, the perturbed system has the same average frequency as the unperturbed one [2]. Therefore, the first-order perturbation phase description predicts no frequency change, due to the presence of a stochastic component in the driving current.

2 RESULTS

In Fig. 2A, the firing rate of a Hodgkin-Huxley neuron is displayed as a function of the magnitude of the injected noise. This curve was obtained by integrating the stochastic system of Eqs. 1 for 5 seconds, and counting the number of spikes. In spite of the predictions of the phase model, the firing rate varies as a function of the noise intensity. For later need, in Fig. 2B we show the unperturbed firing rate f_0 as a function

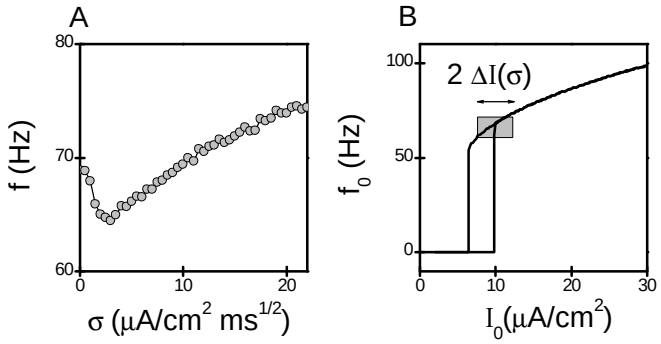


Figure 2: Firing rate of the Hodgkin-Huxley model neuron, as a function of the input strength. A: The amplitude of the stochastic input is varied, with a constant drive of $I_0 = 10$. B: The constant input is varied, with no noise ($\sigma = 0$).

of the intensity of a constant input.

To explain these results, we must go beyond the phase approximation. We propose that the noise not only advances or delays the phase of the limit cycle, but that it also affects its natural frequency. Our hypothesis is that in the presence of noise, the natural frequency of the limit cycle is no longer the original unperturbed frequency corresponding to the mean input current ($f_0(I_0)$ in Fig. 2B). In the noisy case, this frequency is replaced by an effective frequency $\bar{f}(\Delta I)$ obtained as an average of f_0 in an interval around I_0 of size ΔI, such that

$$\bar{f}(I_0, \sigma) = \frac{1}{2\Delta I} \int_{-\Delta I}^{+\Delta I} \tilde{f}_0(I_0 + x + \mu x^2)\, \mathrm{d}x, \tag{3}$$

where $\tilde{f}_0$ is a weighted average of the two hysteresis branches of f_0 (see Fig. 2B), and μ is a parameter scaling the degree up to which the neuron responds to the energy of the input current ($\propto x^2$), as compared

to the raw input current ($\propto x$). The effective frequency $\bar{f}(\Delta I)$, hence, can be obtained from the original unperturbed frequency $f_0(I_0)$, by assuming that the presence of noise is equivalent to a random sample of input currents around I_0, extending to higher or lower values within an interval of size $2\Delta I$. In Fig. 2B, the integration interval ΔI is highlighted in grey. Biologically, this quantity represents the effective current variations produced by the noise.

In order to test the validity of this hypothesis, we calculate the effective frequency of Eq. 3 for several values of ΔI. The results are displayed in Fig. 3A. The dependence of the effective frequency $\bar{f}$ as a

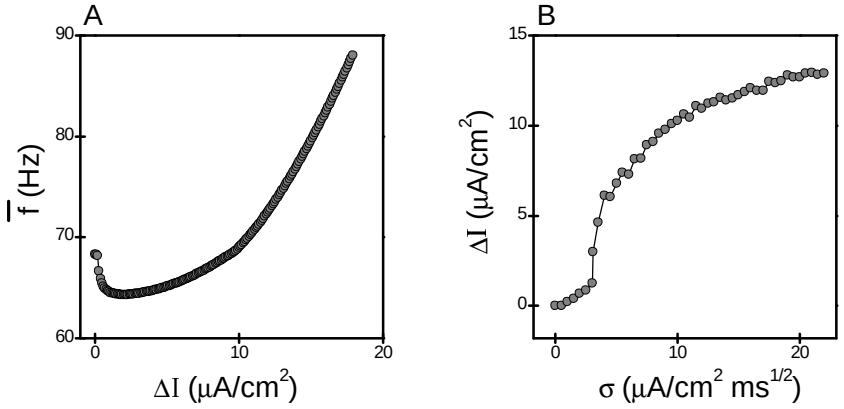


Figure 3: Test of the hypothesis of the effective frequency. A: Effective firing rate $\bar{f}$ as a function of the width of the averaging interval ΔI. B: Averaging interval ΔI needed to obtain an effective frequency $\bar{f}$ equal to the real frequency f, as a function of the noise level σ.

function of the integration interval ΔI is similar to the dependence of the true firing rate f as a function of σ (compare Figs. 3A and 2A). Both of them decrease initially, reach a minimum, and later start rising. Hence, one can qualitatively explain the variations in firing rate produced by input noise in terms of an effective frequency. In order for the interpretation to be valid also quantitatively, we need to establish a correspondence between the input noise σ and the effective integration interval ΔI. This correspondence is shown in Fig. 3B. For each input noise σ, there is an effective integration interval ΔI, such that the effective frequency $\bar{f}_0(\Delta I)$ coincides with the real frequency $f(\sigma)$. Notice that in the present theory, all the noise components participate in the determination of ΔI, not only the the tangential ones, as in classical phase models.

3 Conclusion

Neuron models can be analyzed as dynamical systems driven by stochastic currents. Numerical simulations of such systems show that the amount of input noise determines the firing frequency of a cell. This result contradicts previous analytical derivations. We therefore present an alternative analytical derivation, where the stochastic input is interpreted as a blurred effective signal. The firing frequency of the stochastic system is then derived from the firing frequency of the unperturbed system, by averaging the unperturbed frequency in the blurred interval.

Acknowledgments

This work was supported by CONICET, CNEA, ANPCyT and Universidad Nacional de Cuyo.

References

[1] A. L. Hodgkin, and A. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. 117 (1952) pp 500-544.

[2] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Dover Publications (2003).

RECRUITMENT DIFFUSION ADVECTION REACTION MODELS FOR ANTARCTIC MARINE FISHERIES: GEOGRAPHIC MODELING, ERROR ESTIMATION AND ADAPTIVITY

Nadia S. Alescio[♭,†], Liliana B. Taborda[*], Esteban R. Barrera-Oro[♭,†], Marta B. Bergallo[*], Enrique R. Marschoff[†] and Carlos-E. Neuman Meira[*]

[*] *Mathematics Department (FIQ), Universidad Nacional del Litoral, Santiago del Estero 2829, 3000 Santa Fe, Argentina, bergallo@fiq.unl.edu.ar*

[†] *Instituto Antártico Argentino, Cerrito 1248, 1010 Buenos Aires, República Argentina, marschoff@dna.gov.ar*

[♭] *CONICET and Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", División Ictiología, Ángel Gallardo 470, 1405 Buenos Aires, Argentina*

Abstract: We study the problem of modeling fish populations (*Notothenia coriiceps*, *Notothenia rossii*, *Gobionotothen gibberifrons*, and other *Nototheniidæ*) with emphasis on the changes of their levels (dependent on biomass) and their spatial localization along several stages of the process. The models are based on recruitment equations and control and surveillance modules that are aimed to the sustainable management of fisheries based on these populations. The recruitment diffusion advection reaction equations are the basic framework of the model. The "reaction" part of the model, the recruitment and growth parts as well as the boundary conditions are non linear, so the problem is mathematically difficult and really dependent of the model specification. Further non linearities appear in the form of age (length) thresholds. We consider the case of a single isolated species and that of two interacting species by means of a model able to be generalized as new data becomes available, specially on species interactions. The age distributions and a complex density dependence are considered using an ecological brake and an age dependent reproductive capacity. Model results were compared with fish samples collected through an overall period of 25 years at Potter Cove, South Shetland Islands, Antarctica, after the impact of the fishery in the area in 1978-80. Fisheries in the Antarctic region are currently managed by the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) which has prohibited fishing in the area under study since 1991.

Keywords: *Parabolic problems, Border singularities,* A posteriori *error estimation, Mixed boundary conditions, ALBERTA, Antarctic marine fisheries,* Nototheniidaæ*, CCAMLR.*
2000 AMS Subject Classification: 65N30 - 92D25

1 INTRODUCTION

We study the problem of modeling fish populations of *Notothenia coriiceps*, *Notothenia rossii*, and *Gobionotothen gibberifrons* and other *Nototheniidæ*, with emphasis on the changes of their levels (dependent on biomass) and their spatial localization (we note that the geographic dimension and specification is not common in this type of models) along several stages of the process. The models are based on recruitment equations and control and surveillance modules that are aimed to the sustainable management of fisheries based on these populations. The recruitment diffusion advection reaction equations are the basic framework of the model. The "reaction" part of the model, the recruitment, and growth parts as well as the boundary conditions are non linear, so the problem is mathematically difficult and extremely dependent of the model specification. Further non linearities appear in the form of age (length) thresholds. For simplicity we consider the case of each of the species (accumulated number of individuals) and state some equations for the model of two interacting species (the fish and the forage for the fishes) by means of a model able to be generalized as new data becomes available, specially on species interactions. The age distributions and a complex density dependence are considered using an ecological brake and an age dependent reproductive capacity. The captures in the case of one or two species and the surveillance and control strategies were studied based on prediction and bilinear-quadratic models that have proved to be adequate in other contexts and very useful here.

Model results were compared (see [1]) with fish samples collected through an overall period of 25 years at Potter Cove ([62^o14'S, 58^o42'W] is a cove indenting the southwest side of King George Island to the east of Barton Peninsula, in the South Shetland Islands, Antarctica), after the impact of the fishery in the area

in 1978-80. The sharp decline in abundance of *N. rossii* reported for the period 1983–1992 is consistent with the increase in mean size observed between 1983 and 1987 and the duration of the inshore phase of the species, which is known to last for 6–7 years. In the following years, until 1992, the decreasing abundance is consistent with the entrance of low strength cohorts with the consequent reduction of mean size. This interpretation is supported by the length distributions observed between 1982/3 and 1985/6 where the modal age changes from 2/3 to 6/7 years. After 1991/2 the densities, mean sizes and abundances do not depend on a single forcing event but on several interacting factors. The length data of the others species are studied as well and used for the validation of the model. Fisheries in the Antarctic are currently managed by the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) which has prohibited fishing in the area under study since 1991.

One of the basic objectives of this work is to state the basis and conceptual framework for the modeling of the evolution of antarctic fish populations. Our models differ conceptually with those actually in use in the CCAMLR for similar purposes of this paper respect to the management of the fisheries. In section 2 we state the main equations of the model. In section 3 several considerations about the management of the fishery are developed. The last section is for conclusions.

2 Recruitment Diffusion Advection Reaction Models Distributed in ℓ and in (z, φ)

In principle and with greater generality than that of this article, we consider the variables
$u = u(t, \ell, z, \psi, \theta) = u(t, z, \psi, \theta)$ biomass of "forage" (fish forage)
$v = v(t, \ell, z, \psi, \theta)$ quantity of specimens of fish (interesting for us, those we seek to study, model and control)
$w = w(t, \ell, z, \psi, \theta) = w(t, z, \psi, \theta)$ biomass of predators (our fishes are preys of these, not considered in this work). The variables are ℓ, the length of the specimen of fish, z the depth of the sea, ψ the distance to the shoreline and θ the angle in the polar coordinates (ψ, θ) surrounding the island, but in this work for reasons of simplicity (although the main issues are treated indeed) we drop the variables z and θ from all dependences. We also drop, as stated, the variable w: so we reduce the problem to one equation for variable v and complement the model with a simplified global equation for u, "subordinated" to the equation for v. The variable t is time.

For variable v we consider

$$\hat{V} = \int_{\ell_0}^{\ell_\infty} \tilde{v}(t, \ell) \mathrm{w}_m(\ell) d\ell \qquad \text{with} \qquad \tilde{v}(t, \ell) = \int_{\psi_{min}}^{\psi_{max}} v(t, \ell, \psi) d\psi \tag{1}$$

and

$$\bar{\ell}(t) = \int_{\ell_0}^{\ell_\infty} \breve{v}(t, \ell) \ell d\ell \qquad \text{with} \qquad \breve{v}(t, \ell) = \int_{\psi_{min}}^{20000} v(t, \ell, \psi) d\psi \tag{2}$$

where $\hat{V}$ is the total biomass, $\tilde{v}$ is the distribution of lengths, $\bar{\ell}$ is the mean length and $\breve{v}$ is the distribution of lengths in the zone near the shoreline (20 km). In this context the mass-length relationship is $\mathrm{w}_m(\ell) = a_w \ell^{b_w}$, where a_w and b_w are parameters for each species ($b_w \simeq 3.0$), and the length-age relationship is $\ell_a = \ell_{max}(1 - \mathrm{e}^{-ka})$ where k is a parameter with a value for each species.

2.1 The model

It may be stated an evolution equation for v and other for u (w is not considered in this article)
For u we use the known model of Schaeffer-Rosenzweig-MacArthur and variants

$$du/dt = a_g u(1 - u/K_u) - \alpha \hat{v}(t) \qquad \text{with} \qquad \hat{v}(t) = \int_{\psi_{min}}^{\psi_{max}} \int_{\ell_0}^{\ell_\infty} v(t, \ell, \psi) d\ell d\psi \tag{3}$$

where a_g is a growth parameter, K_u is a parameter that represents an ecological brake, $\hat{v}$ is the total biomass obtained by integration of the distributed variable v (from the main equation) and α is a parameter associated to the predation process.

For the main variable v the equation reads

$$\begin{aligned}\partial v/\partial t - D_\ell \partial^2 v/\partial \ell^2 - D_\psi \partial^2 (v/(\varepsilon_\psi \tilde{K}(t,\psi)))/\partial \psi^2 + \text{div}(\vec{b}(t,\ell,\psi)(v/(\varepsilon_\psi \tilde{K}(t,\psi))) \\ +\partial(k_{al}(\ell_\infty - \ell)v)/\partial \ell + r(v)v - \beta u v + \mu v + E(t,\ell,\psi)v = 0\end{aligned} \tag{4}$$

where $g(\ell) = k_{al}(\ell_\infty - \ell)$ is a growth rate in terms of ℓ, ε_ψ is a calibration factor to deal with the differences in scales, $\tilde{K}(t,\psi)$ is the ecological break, D_ψ is the diffusion coefficient in the ψ variable (distance to the shoreline), D_ℓ is the diffusion coefficient in the ℓ variable, $\vec{b}$ is a field of velocities, k_{al} is a constant that connects the ages distribution with the lengths distribution, ℓ_∞ is the maximum length, $r(v)$ is a function that models the growth of species, β is a parameter that stands for the shortage in growth due to lack of forage, μ is the mortality constant and in the function $E(t,\ell,\psi)$ we condense all control and sustainable management that we can perform over the protected fishery, this function is called "fishing effort". Note that

$$D_\ell \partial^2 v/\partial \ell^2 + D_\psi \partial^2 (v/(\varepsilon_\psi \tilde{K}(t,\psi)))/\partial \psi^2 \quad \text{"must be"} \quad \text{div}(D(\ell,\psi)\ \text{grad}(v)) + l.o.t. \tag{5}$$

The accompanying boundary condition is

$$v(t,\ell_0,\psi) = \int_{\ell_J}^{\ell_\infty} \tilde{r}(t,\ell,\psi)d\ell \qquad \text{with} \qquad \tilde{r}(t,\ell,\psi) = S(\ell)\left(1 - \int_{\ell_J}^{\ell_\infty} v(t,\ell,\psi)d\ell/\tilde{K}(t,\psi)\right)v(t,\ell,\psi) \tag{6}$$

where $S(\ell)$ is a ℓ-dependent fertility coefficient and the parentheses stands for a recruitment ecological brake.

2.2 Insertion in a Common Framework

These equations can be inmersed in the system of equations proposed in a companion paper (see [2], see also [3] and [4]) that will be presented in this Congress. We refer to that paper for the general equations, the error estimators, the adaptive algorithms and the stability considerations of the linearization procedures.

Our equation (Eq. (4)) is able to take the form of the equations (2) and (3) of [2] with the necessary modifications in the reaction part.

It is necessary to add another equation to deal with the evolution of variable u (see Eq. 3 and this is done with a discretization *ad hoc* that solves the problem without creating instability or affecting convergence.

The error estimators of section 5. and 6. of article [2] apply *cæteris paribus* (it is necessary to add an estimator for the error in variable u, but this is done as usual) provided the necessary modifications associated to the reaction part of the equations.

The same adaptive algorithms of [2] are used in the software framework ALBERTA. So we perform refinement and coarsening when needed in order to regularize the distribution of errors.

We have obtained convergence and stability with this methodology of attack.

3 Management of the fishery

We suppose that the complete fishery (zone from 0 m to 20000 m, the cove) is depleted at instant 0 and observe the growth of the species in time. The observables are the evolution of mean lengths of the juvenile fraction of the population not directly affected by the directed fishery and the relative abundances of the considered species For this results we have statistical data obtained from the real historic inshore population. So we suppose that the fishery is in a steady state and, in order to simulate the ultimate depredating action performed in years 1979-80, almost all fishes from the band ℓ_J–ℓ_{max} were extracted.

We represent the evolution of mean lengths *a posteriori* of the dramatic extractive fishing, *i.e.* $\bar{\ell} = \int_0^{20000} \int v\ell d\ell d\psi$ and the evolution in t of the relative abundance of the considered fishes.

Note that the growth pattern of *N. rossii* (positive) is qualitatively different from those of *G. gibberiforms* (negative), but both species are really depleted.

The time series of abundances of *N. rossii* and *G. gibberifrons* relative to *N. coriiceps* at Potter Cove are presented in Fig. 1 (taken from [1]). They were constructed with the methods and data from [1]. The

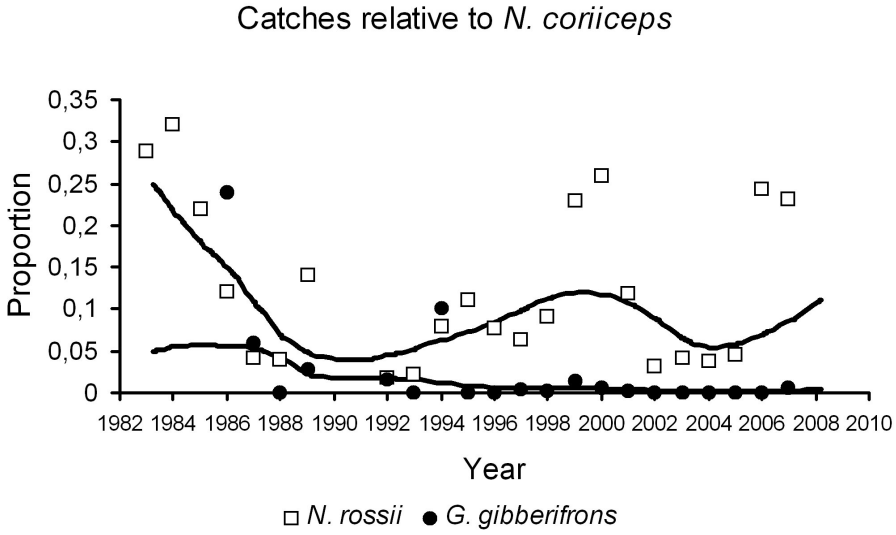


Figure 1: Expected catches of *Notothenia rossii* and *Gobionotothen gibberifrons* relative to the catches of *Notothenia coriiceps* as functions of catch date (Epanenchikov kernel, bandwidth = 0.15), together with the observed split-year mean values.

relative abundance (RA) of each species is calculated as $RA(b) = Nb/(Nc + Nb)$ where Nb is the number of specimens of the species considered (*N. rossii* and *G. gibberiforms*), and Nc is the number of specimens of *N. coriiceps*, producing the values of relative abundance for heach haul and species considered. The results of our simulations are compared with those of Fig. 1.

4 Conclusions

In this work adaptive FEM with *a posteriori* error detection in boundary singular systems was addressed, applied to the modelisation of the growth of antarctic living marine species of fish. Convergent non-linear (Picard linearized) systems associated to the growth models were devised and good responses of adaptivity, error estimation, convergence and stability were observed. The methods proposed for the control (and fishing strategies) of the studied cases are oriented to the sustainable management of these very affected fisheries.

We can observe that the action of no fishing at all these species (what is that what actually must be done as CCAMLR dictates, and is done, except for the common illegal fishing very difficult to control) could not be, perhaps, sufficient. It would be necessary (for reasons on dependence of density for the growth) to take proactive actions (reduction of quantity and quality of predators or other) but this is ecologically infeasible. Our model differs conceptually from those actually in use in the CCAMLR.

References

[1] Barrera-Oro, E.R. and Marschoff, E.R.: "Information on the status of fjord *Notothenia rosii*, *Gobionotothen gibberiforms* and *Nothotenia coriiceps* in the lower South Shetland Islands, derived from the 2000-2006 monitoring program at Potter Cove". CCAMLR Science, **14**, 83–87, (2007).

[2] Bergallo, M.B. and Neuman, C.E.: "Error Estimates for Nonlinear Adaptive Parabolic Finite Element Method in Catalytic Reactor Modeling", *Maci III*, (2011).

[3] Bergallo, M.B., Neuman, C.E. and Sonzogni, V.E.: "Errores *A Posteriori* y Mejoramiento de la Valuación de Opciones Financieras Dependientes del Camino", *Mecánica Computacional*, **25**, 1051–1069, (2006).

[4] Neuman, C.E.: "Modeling Catalytic Reactors as Elliptic Problems with Essential Border Singularities", *Mecánica Computacional*, **23**, 2845–2862, (2004).

MODELO MATEMÁTICO PARA EL TRANSPORTE DE BIOTOXINAS EN UNA RED TRÓFICA MARINA

Daniel Arbeláez Alvarado, Jorge M. Ruiz Vera y Hernán Estrada Bustos

Maestría en Matemática Aplicada, Departamento de Matemáticas, Universidad Nacional de Colombia – Sede Bogotá, Cr.30 No. 45 – 03, Bogotá D. C., Colombia, darbelaeza@unal.edu.co, jmruizv@unal.edu.co, hestradab@unal.edu.co

Resumen: Algunos casos de intoxicación ocasionados por el consumo de peces tóxicos se han reportado en los últimos años en varias regiones insulares de países tropicales. Problemas gastrointestinales, cardiovasculares y neurológicos están entre los síntomas más comunes, pero también la presencia de esta enfermedad afecta negativamente la economía de países del Caribe y el Pacífico que dependen de la pesca y el turismo. Con el propósito de entender el transporte de biotoxinas en la red trófica marina, proponemos un modelo depredador – presa con respuesta funcional Holling Tipo I basado en dos sistemas de ecuaciones diferenciales, el cual es analizado bajo los enfoques cualitativo y numérico.

Palabras clave: *ciguatera, modelo matemático, transporte de biotoxinas, ecuaciones diferenciales.*

1. INTRODUCCIÓN

En las regiones insulares de países tropicales del mundo se presentan con frecuencia brotes de ciguatera, una intoxicación en el ser humano producida por el consumo de peces contaminados con biotoxinas [1][2]. Los microorganismos responsables de la producción de las ciguatoxinas (biotoxinas de la ciguatera) pertenecen al grupo de los dinoflagelados, y su asociación con las macroalgas en arrecifes de coral da origen a un proceso de transporte de ciguatoxinas a través de la red trófica marina [3]. En los arrecifes viven numerosas especies de peces herbívoros que se alimentan de las macroalgas; al consumirlas, introducen en sus tejidos la ciguatoxina producida por los dinoflagelados, que nuevamente es transportada cuando peces de niveles tróficos superiores hacen presa de los herbívoros para alimentarse. Finalmente, el ser humano adquiere la ciguatoxina cuando consume peces que la han acumulado en sus tejidos, sean estos herbívoros o piscívoros.

En la actualidad, la ciguatera afecta a cerca de 50.000 personas en todo el mundo al año, siendo en algunos casos mortal [4]. Los reportes de ciguatera se remontan cerca de cinco siglos atrás, pero sigue siendo una enfermedad poco conocida; es considerada un problema de salud pública y un fenómeno natural que afecta la economía de las comunidades isleñas que viven del turismo [1], razón por la cual es necesario profundizar en su conocimiento. Para avanzar en esta dirección, proponemos un modelo matemático que permite entender la dinámica del transporte de ciguatoxina desde las microalgas hasta los peces piscívoros. El modelo consiste en dos sistemas de ecuaciones diferenciales ordinarias que describen la dinámica de las poblaciones involucradas y el fenómeno de transporte de ciguatoxina a través de una red trófica marina.

2. SUPUESTOS DEL MODELO MATEMÁTICO

Los supuestos del modelo provienen de los conocimientos sobre la ecología de la enfermedad. Estos son:

A) La ciguatoxina se acumula a través de la cadena trófica marina [5]. Gracias a la asociación de los dinoflagelados con las macroalgas, las ciguatoxinas pasan a los hérbívoros que las consumen y por depredación, luego se transfieren a los peces piscívoros.

B) Las poblaciones de algas crecen logísticamente en ausencia de depredación.
C) La ciguatera es visualmente indetectable en los peces [2], por lo cual se asume que la presencia de ciguatoxina en ellos no afecta su tasa de reproducción ni su capacidad de defensa frente a depredadores.
D) La ciguatoxina es degradada natural [5] y constantemente por el organismo de los peces.
E) Por simplicidad, se considera una respuesta funcional Holling Tipo I para ambos sistemas.

3. FORMULACIÓN DEL MODELO MATEMÁTICO

El modelo incluye dos sistemas de ecuaciones diferenciales ordinarias no lineales. El primero modela la interacción alga – herbívoro – piscívoro, y el segundo (que está acoplado al primero) describe el transporte de ciguatoxina entre las tres poblaciones. Ambos consideran respuesta funcional Holling Tipo I.

3.1 SISTEMA DEPREDADOR – PRESA PARA LAS POBLACIONES

El primer sistema es:

$$\left.\begin{aligned} dA/dt &= C_1A(1 - C_2A/C_1) - C_3AH \\ dH/dt &= C_4AH - C_5HP \\ dP/dt &= C_6HP - C_7P \end{aligned}\right\} \quad (1)$$

donde:

A: biomasa de algas.
H: biomasa de peces herbívoros.
P: biomasa de peces piscívoros.
C_1: tasa intrínseca de natalidad en algas.
C_2: coeficiente de competencia en algas.
C_3: coeficiente de depredación de herbívoros sobre algas.
C_4: factor de conversión de algas a herbívoros.
C_5: coeficiente de depredación de piscívoros sobre herbívoros.
C_6: factor de conversión de herbívoros a piscívoros.
C_7: coeficiente de la tasa de mortalidad de piscívoros.

Se definen nuevas variables para el escalamiento apropiado de las ecuaciones:

$$a = AC_2/C_1, \quad h = HC_2/C_1, \quad p = PC_2/C_1, \quad \tau = \eta t$$

y se obtiene así el sistema depredador – presa escalado:

$$\begin{aligned} da/dt &= a[k_1(1 - a) - k_1k_2h] \\ dh/dt &= h[C_4k_1k_2a - k_1k_3p] \\ dp/dt &= p[C_6k_1k_3h - k_4] \end{aligned}$$

donde los coeficientes k_1, k_2, k_3 y k_4 son cocientes entre los parámetros ya definidos. Las soluciones de equilibrio del sistema (1) son:

$(a, h, p) = (0,0,0)$: las tres poblaciones están extintas.
$(a, h, p) = (1,0,0)$: herbívoros y piscívoros extintos; la biomasa se concentra en las macroalgas.
$(a, h, p) = (0, h_1, 0)$: macroalgas y piscívoros extintos.
$(a, h, p) = (a_2, h_2, p_2)$: coexisten macroalgas, herbívoros y piscívoros. Punto de equilibrio de interés.
donde: $a_2, h_1, h_2, p_2 \neq 0$

3.1.1 CONDICIÓN DE EXISTENCIA DE $(a, h, p) = (a_2, h_2, p_2)$, donde: $a_2, h_2, p_2 \neq 0$

El punto de equilibrio interior existe siempre que:

$$k_2k_4 < C_6k_1k_3$$

y está dado por:

$$a = 1 - k_2k_4/C_6k_1k_3, \quad h = k_4/C_6k_1k_3, \quad p = (C_4k_2/k_3)(1 - k_2k_4/C_6k_1k_3).$$

3.1.2 ESTABILIDAD

Teorema 1 *El punto de equilibrio interior* $(a, h, p) = (a_2, h_2, p_2)$ *es localmente asintóticamente estable siempre que exista.*

Prueba. Después de linealizar el sistema depredador - presa obtenemos que el polinomio característico del sistema linealizado es:

$$p(\lambda) = \lambda^3 + d_1\lambda^2 + d_2\lambda + d_3,$$

donde:

$$d_1 = k_1 a, \qquad d_2 = C_6 k_1^2 k_3^2((1-a)/k_2)(C_4 k_2 a/k_3) + C_4 {k_1}^2 {k_2}^2 a(1-a)/k_2,$$
$$d_3 = C_6 {k_1}^3 {k_3}^2 a((1-a)/k_2) C_4 {k_1}^2 {k_2}^2 a.$$

Es fácil demostrar que para el punto de equilibrio (a_2, h_2, p_2) tenemos que:

$$d_1 > 0, \;\; d_3 > 0 \;\; y \;\; d_1 d_2 > d_3,$$

luego, por el criterio de Routh-Hurwitz el punto de equilibrio es localmente asintóticamente estable. □

3.2 SISTEMA PARA EL TRANSPORTE DE CIGUATOXINAS

El segundo sistema es:

$$\left.\begin{aligned} dX/dt &= -\eta(X/A)H \\ dY/dt &= \eta(X/A)H - \alpha(Y/H)P - \beta Y \\ dZ/dt &= \alpha(Y/H)P - \mu Z \end{aligned}\right\} \quad (2)$$

donde:

X: cantidad de ciguatoxina en algas.
Y: cantidad de ciguatoxina en peces herbívoros.
Z: cantidad de ciguatoxina en peces piscívoros.
β: coeficiente de degradación natural de la ciguatoxina en herbívoros.
η: coeficiente de transporte de la ciguatoxina de algas a herbívoros, debido a la depredación.
α: coeficiente de transporte de la ciguatoxina de herbívoros a piscívoros, debido a la depredación.
μ: coeficiente de degradación natural de la ciguatoxina en piscívoros.

Después de realizar un escalamiento apropiado del sistema (2), y considerando que las poblaciones de algas, herbívoros y piscívoros son constantes por la marcada diferencia de escalas temporales, se obtiene una única solución de equilibrio: $(x, y, z) = (0,0,0)$: es un punto de equilibrio estable. En ausencia de brotes consecutivos, la cantidad de toxina en las tres poblaciones tiende a desaparecer en el largo plazo.

4. RESULTADOS NUMÉRICOS

Se presentan dos posibles escenarios de acumulación de ciguatoxina en los tejidos de peces herbívoros y piscívoros (Figuras 1 y 2). Dadas las escalas de tiempo diferentes, que involucran la evolución de las poblaciones y la cantidad de ciguatoxina, empleamos el método de Gear para solucionar numéricamente las ecuaciones.

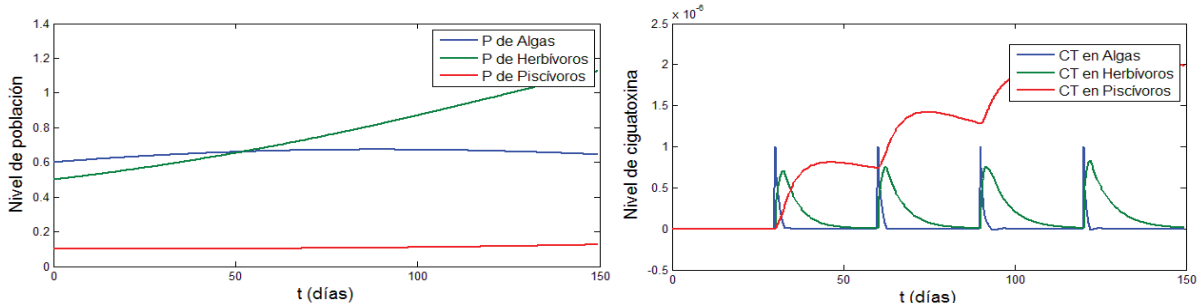


Figura 1. Nivel de las poblaciones y de la ciguatoxina (CT) con brotes periódicos de un mes. Condiciones iniciales: $a_0 = 0.6, \quad h_0 = 0.5, \; p_0 = 0.1, \; x(t_0) = 1\text{x}10^{-6}, \; y(t_0) = z(t_0) = 0.$

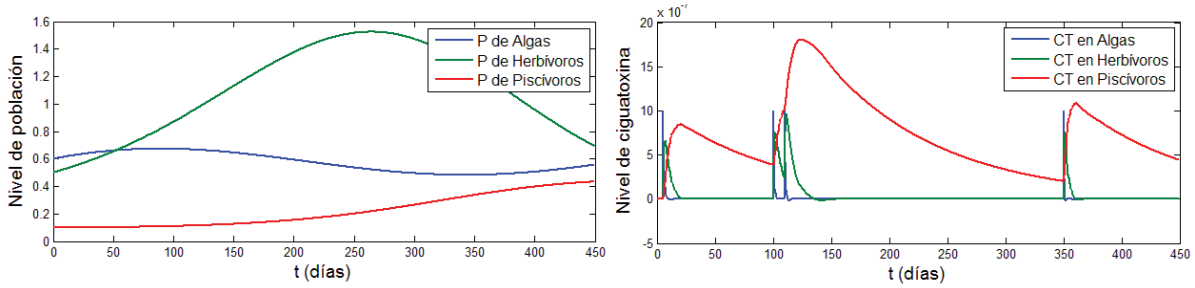


Figura 2. Nivel de las poblaciones y de la ciguatoxina (CT) con brotes cercanos y alejados en el tiempo. Condiciones iniciales: $a_0 = 0.6, \quad h_0 = 0.5, \; p_0 = 0.1, \; x(t_0) = 1\text{x}10^{-6}, \; y(t_0) = z(t_0) = 0.$

5. CONCLUSIONES

Es notable la diferencia en las escalas de tiempo de ambos fenómenos. El transporte de ciguatoxina entre los ensamblajes ecológicos ocurre a una escala de tiempo más corta que la dinámica de los ensamblajes.

La acumulación de ciguatoxina en los tejidos, en especial en peces piscívoros, constituye el mayor peligro potencial para la población nativa y los turistas en las islas tropicales. Cuando los brotes de ciguatoxina se presentan muy seguidos uno del otro (Figura 2) se requiere de medidas que disminuyan el consumo de pescado por más tiempo, para que la toxicidad se reduzca por degradación natural a niveles inocuos.

AGRADECIMIENTOS

A la Universidad Nacional de Colombia, Sede Bogotá y Sede Caribe, por su apoyo para la realización de este trabajo. Al profesor José Ernesto Mancera por sus valiosos comentarios.

REFERENCIAS

[1] G. ARENCIBIA, J. E. MANCERA & G. DELGADO, *La ciguatera un riesgo potencial para la salud humana: preguntas frecuentes*, Universidad Nacional de Colombia – Sede Caribe, 2009.

[2] FAO, *Biotoxinas marinas. Estudio FAO: alimentación y nutrición*, Organización de la Naciones Unidas para la Agricultura y la Alimentación, 2005.

[3] L. LEHANE & R.J. LEWIS, *Review Ciguatera: recent advances but the risk remains*. Int. J. Food Microbiol. 61 (2000), pp.91-25.

[4] R.J. LEWIS, *The changing face of ciguatera*, *Toxicon.,* 39 (2001), pp. 97-106.

[5] J.C. DE FOUW, H.P. VAN EGMOND AND G.J.A. SPEIJERS, *Ciguatera fish poisoning: a review. RIVM Report No.388802021* (available at www.rivm.nl/bibliotheek/rapporten/388802021.html), 2001.

JUEGOS DE FAMILIAS BALANCEADAS

Roberto P. Arribillaga[b]

[b] *Instituto de Matemática Aplicada de San Luis(IMASL) CONICET-Univ. Nacional de San Luis. Av. Ejército de los Andes 950 5700 San Luis, ARGENTINA rarribi@unsl.edu.ar*

Resumen: Los juegos cooperativos en los cuales solo se pueden formar determinadas coaliciones, llamadas coaliciones básicas, han sido ampliamente estudiados, con distintas perspectivas, desde Kaneko y Wooders (1982)[6] asta la actualidad por Aguilera-Escalante (2010)[1] (entre otros). Todos esos enfoques consideran que los jugadores (o la coaliciones) se organizaran por medio de particiones formadas a partir de coaliciones básicas. En el presente trabajo, estudiamos, como las familias balanceadas (que se pueden ver como un generalización de las particiones) es un buen concepto que representa una forma de organizar la participación de los jugadores en más de una coalición y como este tipo de organización nos deja siempre en juegos con core no vacío

Palabras clave: *familias balanceadas, juegos de partición, core*
2000 AMS Subject Classification: 91A12

1. JUEGOS DE FAMILIAS BALANCEADAS CON PAGOS LATERALES

Sea N un conjunto finito de jugadores $N = \{1, 2, ..., n\}$ y se π una clase de coaliciones satisfaciendo que $\{i\} \in \pi$ para todo $i \in N$. A una coalición $T \in \pi$ la llamaremos una coalición básica. Al par (N, π) lo llamaremos *un problema de organización en coaliciones básicas (POCB).*

Antes de comenzar con la descripción del modelo, veamos como la noción de familia balanceada es una idea adecuada para organizar la participacón de los jugadores una más de una coalición (esto se podría leer en algún sentido como una asignación muchos a muchos).

Ejemplo 1 *1. Consideremos el problema en el cual tenemos un conjunto de tres jugadores $N = \{1, 2, 3\}$ y el conjunto de coaliciones básicas esta dado por los singles y las parejas esto es $\pi = \{T \subset N : 1 \leq |T| \leq 2\}$ (esto podría ser un juego de asignación unilateral ó bilateral) y debemos decidir como podemos organizar los jugadores para que cooperen en su accionar (hay una amplia literatura que intenta dar solució a este problema). Intentaremos mostrar como el concepto de familia balanceada puede jugar un rol principal en estos contextos:*

Podemos pensar que la coalición N se va organizar por medio de una partición tomada de π, por ejemplo por medio de $\{1, 3\}; \{2\}$ o bien por medio de $\{1\}; \{2\}; \{3\}$. En esta dirección tenemos "the central assignment game"de Kaneko (1982)[5] y un generalización de estos en los juegos de partición de Kaneko y Wooders (1982)[6]. Cuyo modelo se sigue estudiando actualmente desde distintas perspectivas por ejemplo en Aguilera-Escalante(2010)[1]

Proponemos que la coalición N también se podría organizar de la siguiente manera: se forman las coaliciones $\{1, 2\}; \{1, 3\}; \{2, 3\}$ solo la mitad de "tiempo"(o con la mitad de las dotaciones iniciales, etc). Esto es que cada una de las parejas se forma solo $1/2$ del "tiempo"de modo que cada jugador tiene todo su "tiempo"distribuido en las coaliciones que participa (que es este caso son dos). Bajo esta perspectiva se podra presentar Sotomayor (2009)[9].

2. Consideremos ahora $N = \{1, 2, 3, 4\}$ y π definida como en el ejemplo anterior.
Las organizasiones posibles que nosotros proponemos para N serían:

a) *Todas las particiones de N tomadas de π. Esto es lo propuesto en Kaneko y Wooders (1982)[6] cada jugador participa todo el "tiempo.[en] una sola coalición.*

b) *Que los jugadores pueden participar en más de una coalición de una manera ordenada. Por ejemplo:*
$\{1, 2\}, \{1, 3\}, \{2, 3\}, \{4\}$. *Donde las tres primeras coaliciones se formarán la mitad del "tiempoz la coalición formada por el 4 se formará por el total del "tiempo". (Aquí los jugadores $1, 2, 3$ participan en dos coaliciones y el 4 en solo una).*

Ahora si consideramos la coalición $S = \{1, 3, 4\} \subset N$, se puede considerar las mismas (o similares) posibilidades

Esta propuesta [1] de la organización de los jugadores es la clásica noción de familia balanceada que definimos formalmente a continuación.

Definicin 1 *Dada una coalición $S \subset N$*

1. *Una familia no vacía de coaliciones $\beta_S \subset \mathcal{P}(\mathcal{S})$ es balanceada, para S, si existe un conjunto de números reales positivo $\Lambda = (\lambda_T)_{T \in \beta_S}$ tales que $\sum_{\substack{T \in \beta_S \\ T \ni i}} \lambda_T = 1$, para todo $i \in S$. Los números $(\lambda_T)_{T \in \beta_S}$ son llamados pesos de balance para β_S.*

2. *Si β_S es una familia balanceada (para S) tal que $\beta_S \subset \pi$ diremos que β_S es una $\pi-$familia balanceada. (Notar que una $\pi-$familia balanceada es una familia balanceada formada solo por las coaliciones básicas)*

3. *Si la familia balanceada β_S no contiene propiamente otra familia balanceada decimos que es una familia balanceada minimal.*

Nota 1 *Al conjunto de todas las $\pi-$ familias balanceadas minimales para S lo denotaremos por $\mathcal{B}_S$*

Miremos ahora un POCB (N, π) en el cual tenemos definida sobre π un función real $\bar{v}$. Donde $\bar{v}(T)$ es la utilidad que puede obtener la coalición T si se reúne, para cada $T \in \pi$. A la terna $(N, \pi, \bar{v})$, la llamaremos *POCB con utilidades transferibles.*

Definicin 2 *Dado (N, π) a un juego (N, v) con pagos laterales, lo llamaremos un juego de familia balanceada con pagos laterales, asociado a (N, π), si existe una función real $\bar{v}$ definida sobre π, tal que $v(S) = \max_{\beta_S \in \mathcal{B}_S} \sum_{T \in \beta_S} \lambda_T \bar{v}(T)$, donde λ_T son los pesos de balances asociados con β_S*

En adelante siempre que escribamos (N, v) entenderemos un juego de familia balanceada con pagos laterales, asociado a $(N, \pi, \bar{v})$, para alguna $\bar{v}$

Nota 2 *La defición de Kaneko y Wooders (1982)[6] de juegos de partición es similar a la presentada en anteriormente, pero en aquel trabajo se define el valor de una coalición S como $v^*(S) = \max_{p_S \in P_S} \sum_{T \in p_S} \bar{v}(T)$, donde p_S es una $\pi-$ partición de S y P_S es el conjunto de todas las $\pi-$particiones de S.*

1.1. EL CORE

Dado un juego (N, v) con utilidades transferibles el core del juego esta definido por:
$C(N, v) = \{x \in \mathbf{R}^n : \sum_{i \in N} x_i = v(N)$ y $\sum_{i \in S} x_i \geq v(S)$ para todo $S \subset N\}$.

El teorema de Bondareba-Shapley establece que un juego (N, v) tiene core no vacío si y solo si el juego es balanceado ($\sum_{T \in \beta} \lambda_T \bar{v}(T) \leq v(N)$ para toda familia balanceada β de N)

Teorema 1 *Todos los juegos de familias balanceadas tienen core no vacío.*

Algunas observaciones:

[1] Existen algunos trabajo actuales que considerán relevante este tipo de organización (solo para la gran coalición) entre ellos Cesco (2009)[4], Zaho (2008-2010)[10]; Arribillaga (2010)[2]. Pero no bajo la mirada de que tenemos un colección π arbitraria de coaliciones básicas, sino cuando $\pi = 2^N \setminus \{\varnothing\}$

1. Cuando Shapley y Shubik (1972)[8] introducen los juegos de asignación prueban que si $\pi = \{T \subset N : 1 \leq |T| \leq 2\}\}$, y si (N, v^*) es el juego de partición asociado a $(N, \pi, \bar{v})$ (en el sentido de Kaneko y Wooders (1982)[6]), entonces el core es no vácio para cualquier $\bar{v}$

2. Kaneko y Wooders (1982)[6] muestran condiciones necesarias y suficiente fuertes sobre π para la existencia de core del juego de partición (N, v^*) asociado a $(N, \pi, \bar{v})$ [2].Una larga y actual literatura se dedica a estudiar estas condiciones

3. El teorema anterior estable que si permitimos la organización de los jugadores por medio de familias balanceadas, lo cual nos lleva a un juego de familias balanceadas minimales (N, v), asociado a $(N, \pi, \bar{v})$, siempre tenemos core sin condiciones sobre la π y la $\bar{v}$.

Lema 1 *Si (N, v) es un juego asociado a $(N, \pi, \bar{v})$. Entonces*
$C(N, v) = \{x \in \mathbf{R}^n : \sum_{i \in N} x_i = v(N)$ *y* $\sum_{i \in T} x_i \geq v(T)$ *para todo* $T \in \pi\} =$
$\{x \in \mathbf{R}^n : \sum_{i \in N} x_i = v(N)$ *y* $\sum_{i \in T} x_i \geq \bar{v}(T)$ *para todo* $T \in \pi\}$

Observación

1. Como en Shapley y Shubik (1972)[8] (para los juegos de asiganación) y Kaneko y Wooders (1982)[6] (para los juegos de particiones) el lema anterior muestra que, en los juegos de familias balanceadas, dada una preimputación una coalición cualquiera $S \subset N$ la podrá objetar (según v) si y solo si la preimputación es objetada por una coalición básica $T \in \pi$ (según $\bar{v}$).

2. El lema anterior nos da un criterio muy eficiente para calcular puntos en el core, lo cual en la mayoría de los casos suele ser bastante costoso.

3. Con el resultado anterior podemos tenemos que en respecto al core, cuando este es no vaco en los juegos de particiones, no hay variacón si se utilizamos particiones ó familias balanceadas para organizar los jugadores, como lo estable el siguiente teorema.

Corolario 1 *Sea (N, v) es un juego de familia balanceada y (N, v^*) es un juego de partición asociados a $(N, \pi, \bar{v})$. Si $C(N, v^*) \neq \varnothing$, entonces $C(N, v) = C(N, v^*)$*

2. Juegos de Familias Balanceadas con Pagos Laterales Parciales.

Tomemos ahora un POCB con utilidades transferibles $(N, \pi, \bar{v})$, bajo el supuesto que solo es factible pagos laterales dentro de las coaliciones básicas (esto es el valor $\bar{v}(T)$ se puede repartir de cualquier modo entre los miembros de T para cualquier T en π), pero si una coalición S se organiza bajo la $\pi-$familia balanceada $\beta_S = \{T_1, T_2, ...T_k\}$, no podrá haber transferencia de utilidad desde T_i a T_j si $i \neq j$. Muchos problemas se modelan con estos supuestos, por ejemplo Sotomayor (2009)[9] entre otros. En estas condiciones dado $(N, \pi, \bar{v})$ podemos definir un juego de familias balanceadas con pagos laterales parciales (N, V) del siguiente modo.

Definicin 3 *Dado (N, π) a un juego (N, V) con utilidades no transferible, lo llamaremos un juego de familia balanceada con pagos laterales parciales, asociado a (N, π), si existe una función real $\bar{v}$ definida sobre π, tal que*

$$V(S) = \bigcup_{\beta_S \in \mathcal{B}_S} \left\{ x \in \mathbf{R}^n \,/\, x = \sum_{T \in \beta_S} y^T : y^T(T) \leq \lambda_T \bar{v}(T) \text{ para todo } T \in \beta_S \text{ y } y_i^T = 0 \text{ si } i \notin T \right\}.$$

donde λ_T son los pesos de balances asociados con β_S.

[2]La siguiente frase es tomada de aquel paper .ºbviously, the conditions stated for the nonemptiness of the cores of partitioning games are extremely restrictive and, without some very special structure on the collection of basic coalitions, we would not expect these conditions to be met".

En adelante siempre que escribamos (N, V) entenderemos un juego de familia balanceada con pagos laterales parciales, asociado a $(N, \pi, \bar{v})$

Nota 3 *1. Hemos supuesto que si una coalicón $T \in \pi$ tiene una valor (o utilidad) dada por $\bar{v}(T)$, si esta coalición se forma solo por λ_T del "tiempo"su valor será $\lambda_T \bar{v}(T)$.*

2. En los juegos de partición de Kaneko y Wooders (1982)[6] tenemos una definicón similar reemplazando $\mathcal{B}_S$ por P_S.

Ejemplo 2 *Dado $(N, \pi, \bar{v})$ consideremos en un ejemplo la diferencia entre los juegos asociados de: familia balanceadas con pagos laterales (N, v) y familias balanceadas con pagos laterales parciales (N, V). Sea $N = \{1, 2, 3\}$ y sea $\pi = \{S \subset N : 1 \leq |S| \leq 2\}$*
$\bar{v}(\{i\}) = 0$ para $i = 1, 2, 3$; $\bar{v}(\{1, 2\}) = 6$; $\bar{v}(\{2, 3\}) = 4$; $\bar{v}(\{1, 3\}) = 10$.

1. Tenemos que $x = (1, 2, 0) + (0, 1, 1) + (1, 0, 4) = (2, 3, 5) \in V(N)$, pero $z = (10, 0, 0) \notin V(N)$ (donde $\sum_{i \in N} x_i = \sum_{i \in N} z_i$), esto resalta la imposibilidad de realizar pagos laterales fuera de las coaliciones bsicas.

2. $\sum_{i \in N} x_i = 10 \leq v(N)$. En general si $x \in V(N)$, entonces $x(N) \leq v(N)$.

2.1. EL CORE

Recordemos la definición de Core en un juego con utilidades no transferibles introducida por Aumann (1961)[3]. Dada la preimputación $x \in V(N)$, decimos que una coalición $S \subset N$ bloquea x si existe $z \in V(S)$ tal que $z_i > x_i$ para toda $i \in S$.
El core del juego esta definido por:
$C(N, V) = \{x \in \mathbf{R}^n : x \in V(N)$ y no existe $S \subset N$ que bloquea a $x\}$.
[3]

Teorema 2 *Si (N, v) y (N, V) son juegos asociados a $(N, \pi, \bar{v})$, entonces $C(N, v) = C(N, V)$.*

Observación

El teorema anterior estable que si permitimos la organización de los jugadores por medio de familias balanceadas, el hecho de permitir o no transferencia de utilidad entre las coaliciones que no son básicas no modifica el conjunto de puntos propuestos por el core.

Corolario 2 *Todos los juegos de familias balanceadas con pagos laterales parciales tienen core no vacío.*

REFERENCIAS

[1] N. AGUILERA, M. ESCALANTE, *A Polyhedral approach to hte stability of a family of Coalitions*, Discrete Applied Mathematics 158 (2010) pp. 379-396.

[2] R. ARRIBILLAGA, *The Balanced Core in NTU games: Axiomatic Characterization*, mimeo Universidad Nacional de San Luis (2010).

[3] R. AUMANN, *The Core of a Cooperative Game without Side Payments*, Transactions of the American Mathematical Society 98 (1961), pp.539-552.

[4] J. CESCO, *The M-Core: Definiton and Axiomatic Characterization*, Working paper(2008).

[5] M. KANEKO, *The Central Assignment Game and the Assigmnet Markets*, Journal of Mathematical Economics 10 (1982) pp. 205 - 232.

[6] M. KANEKO, M. WOODERS, *Core of Partitioning Games*, Mathematical Social Sciences 3 (1982) pp. 313-327.

[7] H. SCARF, *The Core of an N Person Game*, Econometrica Vol. 35 No. 1 (1967), pp.50-69.

[8] L. SHAPLEY, M. SHUBIK, *The assignment game I: the core*, Int. J. Game Theory 11 (1972) pp. 111130.

[9] M. SOTOMAYOR, *Correlating New Cooperative and Competitive Solution Concepts in the Time-Sharing Assignment Game*, mimeo Universidad de San Pablo (2009).

[10] J. ZAHO, *The maximal payo and coalition formation in coalitional games*, Fundazione Eni Enrico Mattei, Nota di Lavoro 27.2008.

[3] Scarf(1964)[7] muestra que si un juego (N, V) es balanceado ($\bigcap_{T \in \beta} V_T \leq V(N)$ para toda familia balanceada β de N)[4], entonces tiene core no vacío.

¿BUENA GESTIÓN O BUENA SUERTE?

Marcelo A. Fernández

Departamento de Economía, Universidad de San Andrés, Vito Dumas 284, Victoria, Argentina, fernandezm@udesa.edu.ar, www.udesa.edu.ar

Resumen:

En ausencia de mecanismos de *enforcement* de las promesas electorales, y en presencia de *payoff dominant shocks*, el voto retrospectivo (votar de acuerdo a los actos del político de turno) resulta óptimo. Sin embargo, la presencia de estos shocks socava el poder del voto retrospectivo en resolver los problemas de agencia en un sistema democrático de elección periódica. Incumbentes indisciplinados cuyos incentivos se encuentran desalineados con aquellos de los votantes pueden resultar reelectos, en tanto incumbentes que comparten los incentivos con los ciudadanos (y que actúan en consecuencia) pueden no resultar reelectos. Estos shocks dificultan la identificación entre buena suerte y buena administración. Lo que es más, cuanto mayor es la probabilidad de suceso de estos shocks, menor es el control que tienen los votantes sobre los políticos, pudiendo esto explicar diferencias de *accountability cross-country* lograda por la misma regla.

Palabras clave: *applications of game theory, informational economics, political science*
2000 AMS Subject Classification: 91A80 - 91B44 - 91F10

1. INTRODUCCIÓN

El objetivo del presente trabajo es estudiar la capacidad del voto retrospectivo en resolver la problemática presentada por la relación principal-agente entre los votantes y el gobernante, en un sistema democrático de elección periódica, cuando está sujeto a la presencia de *payoff dominant shocks*.

Entendemos que existe una relacion principal-agente, puesto que los ciudadanos (principal) están delegando la autoridad de decisión respecto de alguna dimensión relevante en un gobernante (agente), el cual posee una ventaja informativa respecto de la acción que toma (*moral hazard*) como de los incentivos personales que posee (*adverse selection*).

Los votantes están interesados en utilizar un mecanismo que, por un lado, incentive al gobernante a esforzarse en el sentido socialmente deseado y por otro lado, que evite seleccionar candidatos con características indeseables.

El análisis del voto retrospectivo, es decir votar de acuerdo con los actos realizados por el incumbente, nace dentro de la literatura de *political economy* en contraposición al voto prospectivo, es decir votar de acuerdo con las promesas de campaña. En efecto, el voto retrospectivo resulta óptimo cuando no existe un mecanismo de *enforcement* que obligue a cumplir las promesas electorales, dado que éstas se convierten en no creíbles.

El tratamiento clásico (ver [4]) reconoce en el voto retrospestivo un grupo de características deseables: siempre reelige al incumbente que tiene los incentivos alineados con la sociedad, mientras que aquellos políticos cuyos incentivos no están alineados, o bien son disciplinados o no son reelectos.

En este trabajo analizamos la validez de las mismas en un especificación distinta del modelo. En particular vamos a incorporar al modelo la existencia de *payoff dominant shocks*, es decir que existe la posibilidad (sin embargo no la certeza) que sucedan shocks que dominan completamente el resultado, sin importar el esfuerzo realizado por el gobernante o su característica inobservable.

Para ello, siguiendo a la literatura (ver [3], [4]), vamos a plantear una economía de dos períodos[1], donde los votantes eligen en cada período a un único político para tomar una acción antes de observar el estado del mundo. El puede tomar una acción que (intente) aislar a la economía de cualquier posible shock (negativo), o en cambio puede decidir no tomar acción alguna y dejar que, de suceder un shock negativo, el mismo tenga su efecto sobre la economía.

[1] La economía reducida a dos períodos permite incorporar una visión intertemporal por parte de los jugadores y al mismo tiempo considerar una regla ampliamente observada como lo es la posibilidad de una única reelección.

Al momento de la elección, los votantes deben optar entre reelegir al incumbente o tomar un candidato aleatoriamente de un pool de políticos.

Consideramos que los votantes son homogéneos, es decir, todos tienen la misma estructura de *payoffs*, estando interesados por una única dimensión, por ejemplo, sirve a forma de ilustración pensar en performance económica.

Por su parte, asumimos que hay diversos tipos de políticos, los cuales pueden ser descompuestos en dos grandes grupos, congruentes y disonantes. La distinción entre congruentes y disonantes, nace de que los primeros comparten los objetivos de los votantes en tanto los últimos están interesados en extraer rentas de su posición privilegiada. Dentro de éstos tenemos toda una familia, los cuales se distinguen por distintos grados de ambición o *rent-seeking*. Debemos recordar que los votantes no conocen sino que deben inferir esta característica según los actos realizados.

2. SETUP

Modelo de dos períodos: $t \in \{1, 2\}$; β factor de descuento intertemporal común.

Votantes homogéneos.

2 tipos de políticos: $\theta \in \{congruente; disonante\}$ =$\{c; d\}$. Un político tomado aletoriamente del pool es congruente con probabilidad τ.

Acciones disponibles para los votantes es reelegir al incumbente $w(.) = 1$, o no hacerlo $w(.) = 0$ y tomar uno al azar del pool de candidatos.

Acciones disponibles del político (inobservables para el votante):
$e_t \in \{N, A\} = \{No\ aislar; (Tratar\ de)\ Aislar\} = \{0; 1\}$

El estado del mundo (desconocido por el político a la hora de tomar la acción, desconocido por los votantes a la hora de decidir si lo reeligen o no): $s_t \in \{low, middle, high\} = \{l, m, h\}$ que suceden con probabilidad $(p, 1 - p - q, q)$

El político en el gobierno puede elegir entre tratar de aislar a la economía o no hacerlo, sin embargo su éxito en hacerlo no está asegurado. Los estados l, h son *payoff dominant*, es decir, el resultado para los votantes es independiente del esfuerzo realizado por el político. En consecuencia, la probabilidad de ocurrencia de un *payoff dominant shock* es $(p + q)$.

La Tabla 1 resume la estructura de *payoffs* de etapa del juego, [2] [3]

	Estado del Mundo		
Incumbente Congruente	Low	Middle	High
N	(Δ; E+Δ)	(0; E)	(0; E)
A	(Δ; E+Δ)	(Δ; E+Δ)	(0; E)

	Estado del Mundo		
Incumbente Disonante	Low	Middle	High
N	(Δ; E+r)	(0; E+r)	(0; E+r)
A	(Δ; E-δ)	(Δ; E-δ)	(0; E-δ)

Tabla 1: Estructura de Payoffs de Etapa: la primer componente payoff observado por los votantes (output y) y la segunda componente es el payoff percibido por el incumbente

Donde E es la "renta de Ego" de mantener el cargo; δ representa un costo para el político disonante en tanto tiene realizar un esfuerzo [4]. Naturalmente, asumimos que $E - \delta > 0$. r es la extracción de renta que puede hacer un político disonante de su posición privilegiada. Tomamos a $r \in (0, R]$ como una variable

[2] Cada tabla refleja el tipo de político en cargo, reflejando así cuán alineados con los votantes están sus incentivos.

[3] Cuando la economía se ve afectada por un shock demasiado bajo o demasiado alto, el resultado observado por los votantes no es modificado por el esfuerzo que realizó el político. Sin embargo, sí importa el esfuerzo/accionar del político en el caso de que la economía se vea afectada por un shock intermedio. A forma de ejemplo consideremos el esfuerzo o accionar del político como obras de infraestructura de drenaje de una ciudad y el estado del mundo como la cantidad de precipitación. Los individuos carecen de pluviometros y no fiscalizan (ni tienen capacidad para juzgar la eficacia) personalmente el cumplimiento de las obras públicas. Si apenas garúa, la vivienda de los individuos no se inunda, independientemente de si se han realizado o no obras de drenaje. Si la cantidad de precipitación es extrema, ningún tamaño de obra podrá evitar de que se inunden. Sin embargo, para un nivel de precipitaciones intermedio, la cantidad/calidad de las obras realizadas, es decir el esfuerzo o acción realizado por el político, sí importa.

[4] Aunque en este contexto basta la presencia de r para hacer la distinción entre políticos congruentes y disonantes.

aleatoria i.i.d. de una distribución $G(r)$ - función de densidad acumulada - con media μ, lo cual nos permite hablar de distintos grados de *rent seeking* y $R > \beta(\mu + E)$, lo cual nos asegura que los políticos "malos" eligen algo que no es lo mejor para los votantes una parte positiva del tiempo.

3. TIMING

1. Un político es tomado aleatoriamente del pool (con lo cual también Naturaleza determina el tipo del político). Si el político que toma es disonante, entonces Naturaleza toma un r_1 de la distribución $G(r)$.
2. El político elige su acción.
3. La Naturaleza determina el estado del mundo en el primer período.
4. Los individuos ven el *payoff* del primer período. (Recordamos que los votantes ven los *payoffs* pero desconocen tanto el estado del mundo, como el tipo de político y la acción que éste tomó).
5. Los votantes deciden si reeligen al incumbente, o lo desechan y toman otro aleatoriamente del pool de candidatos.
6. Si eligieron un candidato nuevo del pool, entonces Naturaleza determina r_2, de lo contrario $r_2 = r_1$.
7. El político toma la acción correspondiente al período 2.
8. La naturaleza determina el estado del mundo en el segundo período.
9. Se observan los *payoffs* del período 2.
10. El juego finaliza.

4. EQUILIBRIO

Definición 1 $\{e(c); e(d); w(y)\}$ *es un Perfect Bayesian Equilibrium si se cumplen las siguientes condiciones:*

$$e(c) = argmax_{e\in\{0,1\}}\{E + p\Delta + (1-p-q)e\Delta + \beta[\Pi(c,e)w(\Delta) + (1-\Pi(c,e))w(0)][E + (1-q)\Delta]\}$$

$$e(d) = argmax_{e\in\{0,1\}}\{E + r - e(\delta + r) + \beta[\Pi(d,e)w(\Delta) + (1-\Pi(d,e))w(0)][E + r]\}$$

$$w(y) = argmax_{w\in\{0,1\}}\{w[p + z(y)(1-p-q)] + (1-w)[p + \tau(1-p-q)]\}$$

donde $z(y)$ está dado por la regla de Bayes:

$$z(\Delta) = prob(\theta = C \mid y = \Delta) = \frac{\Pi(c, e(c))}{\Pi(c, e(c)) + \Pi(d, e(d))}$$

$$z(0) = prob(\theta = C \mid y = 0) = \frac{[1 - \Pi(c, e(c))]}{[1 - \Pi(c, e(c))] + [1 - \Pi(d, e(d))]}$$

y la probabilidad de obtener un output alto (Δ) condicional en que el incumbente sea de tipo θ es:

$$\Pi(\theta, e(\theta)) = p + (1-p-q)e(\theta)$$

Proposición 1 *Si* $\beta(E+r)(1-p-q)-\delta \geq r$ *y si* $z(\Delta)=1/2 \geq \tau$*, entonces una estrategia*

$$\{e(c)=1, e(d)=1, w(\Delta)=1; w(0)=0\}$$

es Perfect Bayesian Equilibrium.

Proposición 2 *Si* $\beta(E+r)(1-p-q)-\delta < r$ *y si* $z(\Delta)=\frac{1-q}{1-q+p} \geq \tau$*, entonces una estrategia*

$$\{e(c)=1, e(d)=0, w(\Delta)=1; w(0)=0\}$$

es Perfect Bayesian Equilibrium.

Corolario 1 *Cuanto mayor es la probabilidad de ocurrencia de un payoff dominant shock, menor es la probabilidad de disciplinar a los políticos disonantes (dada una distribución de los mismos).* [5]

5. CONCLUSIONES

De acuerdo con el tratamiento clásico, el voto retrospectivo siempre reelige al incumbente que tiene los incentivos alineados con la sociedad, mientras que aquellos políticos cuyos incentivos no están alineados, o bien son disciplinados o no son reelectos.

Al incorporar la posibilidad de *payoff dominant shocks*, si bien la regla de voto retrospectivo continúa siendo óptima, sufre sin embargo de algunas desventajas no observadas previamente. En particular, políticos cuyos incentivos están alineados pueden no ser reelectos, políticos disonantes e indisciplinados pueden ser reelectos. Adicionalmente la familia de candidatos que logra disciplinar la regla de voto es estrictamente menor. Es decir, la capacidad de resolver los problemas de agencia (tanto *moral hazard* como *adverse selection*) se ve disminuida por la presencia de estos shocks, ya que los mismos representan una nueva dimensión donde los candidatos esconden su verdadero tipo y esfuerzo, dificultando así la distinción entre buena (mala) suerte y buena (mala) administración.

Como corolario, una economía más susceptible de ser afectada por un *payoff dominant shock* tiene menor control sobre el problema de agencia tienen los votantes sobre los políticos, lo cual explicaría diferencias de *accountability cross-country* de la misma regla.

La inclusión de estos shocks rompe con el *mapping* perfecto entre acciones inobservables del incumbente y payoffs observados por los votantes (que se hallaba presente en el tratamiento clásico), y en consecuencia debilita el poder del voto retrospectivo en resolver los problemas de agencia entre votantes y políticos.

AGRADECIMIENTOS

Quisiera expresar un agradecimiento especial hacia el Lic. Pablo Fajfar por su apoyo y orientación durante la presente investigación. Todo posible error remanente es exclusivamente mío.

REFERENCIAS

[1] R. AUMANN, AND S. HART, *Long cheap talk*, Econometrica, 71 (2003), pp.1619-1660.

[2] J. BANKS, AND R. SUNDARAM, *Adverse selection and moral hazard in repeated elections* in Proceedings of the Seventh International Symposium in Economic Theory and Econometrics, Cambridge University Press, (1993), 295-311.

[3] J.C. BERGANZA, *Two roles for elections: disciplining the incumbent and selecting a competent candidate*, Public Choice, 105 (2000), pp.165-193.

[4] T. BESLEY, *Principled Agents? The Political Economy of Good Government*, Oxford University Press, 2006.

[5] Podemos reescribir la condición de disciplina sobre los políticos disonantes despejando r de la siguiente expresión: $\beta(E+r)(1-p-q)-\delta \geq r$, y obteniendo así: $\frac{\beta E[1-(p+q)]-\delta}{1-\beta[1-(p+q)]} \geq r$. De esta manera podemos definir un umbral $\overline{r}=\left(\frac{\beta E[1-(p+q)]-\delta}{1-\beta[1-(p+q)]}\right)$ como el político más ambicioso (rentista) que la regla logra disciplinar. Asimismo podemos definir $\lambda = G(\overline{r})$ como la probabilidad de que un político disonante se comporte en el mejor interés de los votantes. En efecto esta probabilidad cuenta cual es la cantidad de tipos de político disonante que logra disciplinar la regla, dada la distribución de tipos.

Para probar el corolario basta observar: $\frac{\partial \overline{r}}{\partial(p+q)} = \frac{-\beta(E-\delta)}{[1-\beta(1-(p+q))]^2} < 0 \square$

UN MÉTODO PARA OBTENER ESTABILIDAD EN LOS MODELOS DE ASIGNACIÓN CON RESTRICCIÓN DE CAPACIDAD.

Mabel Marí[♭]

[♭]*Departamento de Matemática, Universidad Nacional de San Juan, Ignacio de la Roza 230 Oeste, San Juan, Argentina, mabelmari@speedy.com.ar*

Resumen: En este trabajo se considera un modelo de asignación especial, en el cual intervienen dos tipos de agentes complementarios y una institución, la cual tiene preferencias sobre las posibles asignaciones. Esta institución tiene para contratar un conjunto de pares de trabajadores, y tiene una cuota q que es el número máximo de pares a contratar.

Se muestra un algoritmo, mediante el cual partiendo de una asignación arbitraria de este modelo, ésta converge a una asignación q-estable.

Palabras clave: *Matching, estabilidad, cuota*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Los *modelos de asignación bilateral* son utilizados para estudiar problemas de mercados cuyo rasgo distintivo es que los agentes involucrados, están en conjuntos disjuntos con características diferentes (por ejemplo, directores y estudiantes). El problema fundamental del modelo de asignación consiste en asignar a cada agente de un conjunto al menos un agente del otro lado. Los agentes tienen preferencias sobre sus compañeros potenciales. La estabilidad ha sido considerada la propiedad principal. Un matching se llama estable si todos los agentes son asignados a un compañero aceptable y no hay un par de trabajadores no asignados que se prefieran mutuamente a las asignaciones respectivas en el matching.

Una variante del modelo de asignación es el *modelo de asignación con restricción de capacidad* que consiste en asignar cada trabajador, de un lado del mercado, con un trabajador, sobre el otro lado, tal que los pares contratados por la institución son, a lo sumo q. La propiedad de estabilidad en este modelo depende de las preferencias de los agentes y de la institución, por ello se define *q-estabilidad*. El *conjunto de matchings q-estables* puede ser vacío. Sin embargo, con *preferencias responsive* de la institución, la existencia del conjunto de matchings q-estables está garantizada y se obtuvo una caracterización del mismo. Estos resultados han sido obtenidos conjuntamente con Femenia, Neme y Oviedo.

En este trabajo se da un algoritmo mediante el cual partiendo de una asignación arbitraria del modelo, ésta converge a una asignación q-estable.

2. NOTACIÓN Y RESULTADOS PREVIOS

Consideramos los conjuntos complementarios D y E, notamos con P_d, P_e las preferencias de los agentes $D(E)$ con respecto a los agentes de $E(D)$; el prefil de prefrencias con $\mathbf{P}$ y el modelo de asignación con $M = (D, E, \mathbf{P})$.

Definición 2.1 *Un **matching o asignación** es una función $\mu : D \cup E \to D \cup E \cup \{\emptyset\}$ tal que, para todo $d \in D$ y $e \in E$ satisface:*
1 - $\mu(d) \in E$ o bien $\mu(d) = \emptyset$.
2 - $\mu(e) \in D$ o bien $\mu(e) = \emptyset$.
3 - $\mu(d) = e$ si y sólo si $\mu(e) = d$.

Simbolizamos con $\mathcal{M}$ el conjunto de todas las posibles matching en el modelo $M = (D, E, \mathbf{P})$.

Definición 2.2 *Un matching μ es **bloqueado por un agente** f si $\emptyset \, P_f \, \mu(f)$.*

Decimos que un matching es *individualmente racional* si no está bloqueada por ningún agente.

Definición 2.3 *Un matching μ esta **bloqueado por un par de trabajajadores** (d, e) si $d\, P_e\, \mu(e)$ y $e\, P_d\, \mu(d)$.*

Definición 2.4 *Un matching μ es* **estable** *si no está bloqueada por un agente o por un par de agentes.*

$S(M)$ es el conjunto de matchings estables. Gale y Shapley (1962) probarón que $S(M) \neq \emptyset$. La institución

U tiene una preferencia R_U sobre el conjunto de pares que están trabajando para ella. Ésta puede elegir algunos matchings de $\mathcal{M}$ de acuerdo a sus preferencias P_U y a su cuota q. Este nuevo modelo de matching lo denotamos por $M_U^q = (M; R_U, q)$.

Definición 2.5 *Un matching μ es* aceptable *para U si $\mu \in \mathcal{M}_q$ y $\mu R_U \mu^\emptyset$ ($\mu^\emptyset(f) = \emptyset$ para todo $f \in D \cup E$, con $\mathcal{M}_q = \{\mu \in \mathcal{M} : \#\mu \leq q\}$.*

Definición 2.6 *Un matching $\mu \in \mathcal{M}$ es $q-$**individualmente racional** si $\#\mu \leq q$, $\mu R_U \mu^\emptyset$ y para todo $f \in D \cup E$ tal que $\mu(f) \neq \emptyset$ se verifica $\mu(f)\, R_f \emptyset$.*

Definición 2.7 *Un matching μ está $q-$**bloqueado** por el par (d, e) si $eP_d\mu(d)$, $dP_e\mu(e)$ y se verifica:*
a) $\mu(d) \in E$ y $\mu(e) \in D$, o
b) $\mu_{(d,e)}$ es $q-$individualmente racional y $\mu_{(d,e)} R_U \mu$.

Definición 2.8 *Un matching μ es $q-$**estable** si es $q-$individualmente racional y no está $q-$bloqueado por un par de agentes.*

En el modelo M_U^q denotamos con $S(M_U^q)$ el conjunto de matchings $q-$estables.

Teorema 2.1 *Si M_U^q es un modelo de asignación con restricción de capacidad con preferencias R_U responsive entonces $S(M_U^q) \neq \emptyset$.*

Sea $F \in \{D, E\}$ y $F^c \in \{D, E\}$ tal que $\{F, F^c\} = \{D, E\}$, y denotamos con $f \in F$ a un agente genérico. Sea $F' \subseteq F$, y $P_{|F'}$ la restricción de P_{F^C} a F'. Dado $M = (F, F^C, \mathbf{P})$, notamos $M_{F'} = (F', F^C, P_{|F'}, P_{F^C})$ la restricción de M a F'. Para simplificar, sea $M_{F'} = (F', F^C, \mathbf{P})$, donde $\mathbf{P} = (P_{|F'}, P_{F^C})$.

Para cada t, podemos definir el subconjunto $F^t \subseteq F$ tal que $|F^t| = t$, y si $f \in F^t$ y $f' \notin F^t$ tenemos que $f \succ_F f'$ ($\succ_F$ preferencia individual de U que genera R_U responsive).

Sea $N = \{(t^*, s^*) : t^* \in \{1, 2, ..., n\},\ s^* \in \{1, 2, ..., m\}\}$, para cada $(t^*, s^*) \in N$, denotamos $M^{(t^*, s^*)} = \left(D^{t^*}, E^{s^*}, \mathbf{P}\right)$ (modelo reducido).

Dado $M_U = (M, R_U)$, $(t^*, s^*) \in N$, y q, se definen los subconjuntos de matchings estables:

$$T_q(M^{(t^*, s^*)}) = \begin{cases} S(M^{(t^*, s^*)}) & \text{si } \#\mu = q, \text{ para cada } \mu \in S(M^{(t^*, s^*)}) \\ \emptyset & \text{en otro caso.} \end{cases}$$

donde $T_q(M) = \left\{\mu \in \mathcal{M} : \text{existe} \quad (t^*,\, s^*) \in N \text{ y } \mu \in T_q(M^{(t^*, s^*)})\right\}$

y

$T_{<q}(M^{(t^*, s^*)}) = \{\mu \in S(M^{(t^*, s^*)}) : \#\mu < q,\ d \text{ y } e \text{ no son mutuamente aceptables, para todo } (d, e) \in D \backslash \mu(E^{s^*}) \times E \backslash \mu(D^{t^*})\}$ con

$T_{<q}(M) = \left\{\mu \in \mathcal{M} : \text{existe} \quad (t^*,\, s^*) \in N \text{ y } \mu \in_{<q} (M^{(t^*, s^*)})\right\}$.

El resultado que nos caracteriza el conjunto es q-estables es:

Teorema 2.2 *Sea M_U^q. Si R_U responsive entonces $S(M_U^q) = T_q(M) \cup T_{<q}(M)$.*

3. UN ALGORITMO PARA CALCULAR UN q-ESTABLE EN M_U^q.

Para el modelo de asignación bilateral $M = (D, E, \mathbf{P})$ Roth y Vande-Vate(1990) demuestran que partiendo de una asignación μ que no es estable, siempre podemos encontrar un camino mediante el cual, satisfaciendo los pares bloqueantes escogidos de tal forma que, en el par bloqueante, el compañero de uno de los agentes sea el mas preferido, de todos los que con él forman un par bloqueante se puede obtener una asignación estable. A la asignación encontrada lo representaremos por $\gamma_{(n\,,m)}(\mu) = \mu_k$.

Definición 3.1 *Sean μ una asignación del modelo M y $\succ_D$ las preferncias individuales de la institución. La asignación μ' es una* ***truncación*** *de μ respecto de d_t, si*

$$\mu'(d) = \begin{cases} \mu(d) & \text{si} \quad d \succeq_D d_t \\ \emptyset & \text{en otro caso.} \end{cases}$$

En forma simétrica se define la truncación de una asignación para los agentes de E.
Si $\#\mu^{'} = q$, diremos que $\mu^{'}$ es una q***-truncación*** de μ con respecto a d_t.

Lema 3.1 *Sea $\mu \in S(M)$ tal que $\#\mu > q$ y $\overline{\mu}$ la q-truncación de μ con respecto a $d_{\bar{t}}$. Si $\overline{\mu} \notin S(M_U^q)$, entonces $\#\gamma_{(\bar{t}\,,m)}(\overline{\mu}) \geq q$.*

Sea μ_1 una asignación arbitraria en el modelo M_U^q, queremos encontrar un procedimiento mediante el cual podamos hallar un nueva asignación que pertenezca a $S(M_U^q)$. En el caso que esta asignación no sea estable en el modelo aplicamos el Algoritmo de Roth-Vande Vate y encontramos $\gamma_{(n\,,m)}(\mu_1) \in S(M)$, si su cardinalidad es menor o igual a q, el proceso termina. En caso contrario realizamos una q-truncación a $\gamma_{(n\,,m)}(\mu_1)$ con respecto a $d_{\bar{t}}$, a la cual llamaremos μ_2, si esta truncación es estable, encontramos la asignación buscada. En otro caso, aplicamos el algoritmo de Roth-Vande Vate y se vuelve a repetir el proceso anterior, que debe concluir pues el conjunto D que vamos considerando es, en cada paso, menor.

Describimos a continuación los pasos a seguir:

1. Si $\mu_1 \in S(M_U^q)$, entonces el algoritmo para.
2. Si $\mu_1 \notin S(M_U^q)$, entonces:
   aplicar el algoritmo de Roth-Vande Vate y encontramos $\mu_2 \in S(M)$.
3. $i := 2$.
4. Calcular $h = \#\mu_i = \max\limits_{1 \leq k \leq n} \{\mu_i(d_k) \neq \emptyset\}$,

   4.1 Si $\#\mu_i \leq q$, el algoritmo para.

   4.2 Si $\#\mu_i > q$, sea d_t tal que

   a) $\mu(d_t) \neq \emptyset$,

   b) $\#\{d \in D : d \succeq_D d_t \text{ y } \mu(d) \neq \emptyset\} = q$

   y obtenemos la q-truncación

   4.2.1 $\mu_{i+1}(d) = \begin{cases} \mu_i(d) & \text{si} \quad d \succeq_D d_t \\ \emptyset & \text{en otro caso.} \end{cases}$

   4.2.2 Si $i + 1 = 3$, sea $s : e_s = \min\limits_{\succ_E} \{\mu_3(d) : d \succeq_D d_t\}$.

   4.2.3 Si $\mu_{i+1} \in S(M^{(t,s)})$, entonces el algoritmo para.

   4.2.4 Si $\mu_{i+1} \notin S(M^{(t,s)})$, sea $i := i + 2$, aplicar en $M^{(t,s)}$ el algoritmo de Roth-Vande Vate y encontrar $\gamma_{(t\,,s)}(\mu_{i+1}) = \mu_{i+2} \in S(M^{(t,s)})$.

5. Ir al paso 4.

Nota En los pasos 4.2, 4.2.1 y 4.2.2 truncamos el matching μ_i y obtenemos el menor modelo reducido $M^{(t,s)}$ en el cual la q-truncación es una asignación. Además la condición 4.2.2 solo se aplica cuando $i = 2$ y al repetir el paso 4 de este algoritmo, en el modelo $M^{(t,s)}$, s se mantiene constante y solo va cambiando t, que en cada pasada es menor.

Teorema 3.1 *Sea μ una asignación de M_U^q. Entonces existe una sucesión finita de asignaciones $\mu_1, \mu_2, ..., \mu_k$, obtenidos por el algoritmo anterior, tal que $\mu_k \in S(M_U^q)$.*

REFERENCIAS

[1] D. FEMENIA, M. MARÍ, A. NEME AND J. OVIEDO (2008). "Stable solutions on matchings models with quota restrictions ". (En prensa)

[2] D. GALE AND L. SHAPLEY (1962). "College admissions and the stability of marriage", *American Mathematical Monthly,* **69**, 9-15.

[3] D. GALE AND M. SOTOMAYOR (1985). "Some remarks on the stable matching problem", *American Mathematical Monthly,* **11**, 223-32.

[4] A. ROTH AND M. SOTOMAYOR (1990). *Two-sided Matching: A Study in Game-Theoretic Modeling and Analysis.* Cambridge University Press, Cambridge, England. [Econometrica Society Monographs No. 18].

[5] ROTH, ALVIN AND JOHN H. VANDE VATE. (1990).*Rondon paths to stability in two-sided matching.* Econometrica, 58, 1475-1480.

SEQUENTIAL ENTRY IN ONE-TO-ONE MATCHING MARKET

Beatriz Millán[♭]

[♭]*Instituto de Matemática Aplicada-San Luis (UNSL-CONICET) and Instituto de Ciencias Básicas (UNSJ), bmillan@unsl.edu.ar*

Abstract: We study in one-to-one matching market a process of sequential entry, in which participants enter in the market one at a time, in some arbitrary given order. We identified a large family of orders (optimal orders) which converges to the optimal matching. We also show that the last agent who enters the market will receive his/her optimal outcome under stable matchings.

Keywords: *Stable matching, sequential entry, optimal matching*

1 INTRODUCTION

We study sequential entry in marriage model. The marriage model describes a two-sided, one-to-one matching market without money where the two sides of the market for instance are workers and firms (job matching) or medical students and hospitals. We use the common terminology in the literature and refer to one side of the market as men and to the other as women. An outcome of a marriage market is called a matching, which can simply be described by a collection of single agents and married pairs (consisting of one man and one woman). Loosely speaking, a matching is stable if all agents have acceptable spouses and there is no couple whose members both like each other better than their current spouses.

Originally, Gale and Shapley [4] proved that the set of stable matching is not empty. Their constructive proof identifies an algorithm which starts with the empty matching and generates a sequence of matching where blocking pairs are matched at each iteration. In the original Gale-Shapley algorithm men propose simultaneously at each iteration, McVitie and Wilson [7] observed that it can be modified by letting at each iteration only one man propose to the woman he prefers most among those who have not yet turned him down.

Kunth [5] demonstrated that starting with an arbitrary matching and iteratively satisfying blocking pairs will not necessarily lead to a stable matching. The example raises the question of whether it is possible to obtain an algorithm which will start with an arbitrary matching, rather than the empty one, will match a blocking pair at each iteration, and will always terminate with a stable matching. Knuth's original problem was answered negatively by Tamura [11] and Tan and Su [12]. However, Roth and Vande Vate [9] designed an algorithm which starts with an arbitrary matching and reaches a stable one by iteratively matching blocking pairs. Their algorithm introduces the players successively into the system and lets them iteratively satisfy blocking pairs at each stage by a decentralized system.

Many authors considered incremental algorithms similar to Roth and Vande Vates, such as Jimpeng Ma [7],Yosef Blum and Uriel RothBlum [3] and Peter Biró, Katarína Cechlárová and Tamás Fleiner [1].

Following Roth and Vande Vates we suppose that agents sequentially enter the market. We study a process of sequential entry, in which participants enter in the market one at a time, in some arbitrary given order. When an agent enters the market the Deferred Acceptance algorithm with input matchings is applied. At the end this process generates a stable matching. We identified a large family of orders (optimal orders) which converges to the optimal matching. We also show that the last agent who enters the market will receive his/her optimal outcome under stable matchings.

2 PRELIMINARES

First we introduce the marriage market. In this market there are two finite and disjoint sets of agents, the set M of n "men" and the set $W$of p "women". We refer to $V = M \cup W$ as the set of agents and we denote a generic agent by v. Each $m \in M$ has a strict preference relation $P(m)$ over the set

$W \cup \{m\}$, and each $w \in W$ has a strict preference relation $P(w)$ over the set $M \cup \{w\}$.[1] The preference relation of a man m, for instance, can be represented by an ordered list of elements in $W \cup \{m\}$, $P(m) = w_1, w_3, m, w_2, \cdots, w_k$ indicates that m prefers w_1 to w_3 and he prefers remaining single to any other woman. Preferences profile is $(n+p)$-tupla of preference relations and it is represented by $P = (P(m_1), \cdots, , P(m_n), P(w_1), \cdots, P(w_p))$. We write $wP(m)w'$ if w is preferred to w' under the preference relation $P(m)$, and $wR(m)w'$ if $wP(m)w'$ or $w = w'$. Similarly we write $mP(w)m'$ and $mR(w)m'$. A woman w is acceptable to a man m if $wP(m)m$. Analogously, m is acceptable to w if $mP(w)w$. The marriage market is fully described by the triplet (M, W, P).

A matching μ is a one-to-one correspondence from F to itself, such that for each $m \in M$ and for each $w \in W$ we have $\mu(m) = w$ if and only if $\mu(w) = m$, $\mu(m) \notin W$ then $\mu(m) = m$, and similarly $\mu(w) = w$ if $\mu(w) \notin M$. If $\mu(m) = w$, then man m and woman w are matched to one another. We say that agent v is single if $\mu(v) = v$. Given a matching μ, we call $\mu(v)$ the outcome for v under μ.

A matching μ is individually rational if $\mu(v)R(v)v$ for all $v \in V$. A blocking pair for a matching μ is a pair $(m, w) \in M \times W$ such that $mP(w)\mu(w)$ and $wP(m)\mu(m)$.

Definition 1 *A matching μ is stable if it is individually rational and there are no blocking pairs for it.*

We denote the set of stable matchings by $S(M, W, P)$. Gale and Shapley [4] proved that a stable matching must exist. They further proved the existence of a stable matching μ_M which is optimal for all men in the sense that no other stable matching μ exists that gives any man m an outcome $\mu(m)$ that he prefers to $\mu_M(m)$ and a (possibly different) stable matching μ_W that is optimal for all women.

Deferred acceptance algorithm with arbitrary input

We will describe the Deferred Acceptance (DA) algorithm with arbitrary input matchings following Blum, Roth and Rothblum [2]. The algorithm starts with an arbitrary matching, selects a single man m and its most preferred woman w (if any) then it checks whether they form a blocking pair. If they do, this will be a maximal blocking pair (w is the most preferred woman for m among those with whom m forms a blocking pair), and when this blocking pair is satisfied a new matching is formed. This process is then iterated until there is no single man who is part of a blocking pair.

Formally, the algorithm is described as follows.

Input

Let μ be an arbitrary matching.

Initial stage

(0) (i) For all $m \in M : A_m^0 = \{w \in W : wP(m)m\} \setminus \{\mu(m)\}$.
(ii) $\mu^0 = \mu$; $i = 1$.

Iteration stage

(1) If there is no $m \in M$ such that $\mu^{i-1}(m) = m$ and $A_m^{i-1} \neq \emptyset$, stop with output μ^{i-1}.

(2) Let m be such that $\mu^{i-1}(m) = m$ and $A_m^{i-1} \neq \emptyset$ and set w be the most preferred woman for m in A_m^{i-1}.

(3) i) if $\mu^{i-1}(w)P(w)m$, then $\mu^i = \mu^{i-1}$.
ii) otherwise,
if $\mu^{i-1}(w) = w$, then $\mu^i(w) = m$ and $\mu^i(f) = \mu^{i-1}(f)$
for all $f \in F \setminus \{w, m\}$, and
if $\mu^{i-1}(w) = m^*$, then $\mu^i(w) = m$, $\mu^i(m^*) = m^*$ and

[1] $P(m)$ is a complete, antireflexive, and transitive binary relation on $W \cup \{m\}$.

$$\mu^i(f) = \mu^{i-1}(f) \text{ for all } f \in F \setminus \{w, m, m^*\} .$$

(4) $A^i_m = A^{i-1}_m \setminus \{w\}$ and for all $m' \neq m, A^i_{m'} = A^{i-1}_{m'}$

(5) $i = i + 1$, go to (1).

We denote $DA_M(\mu)$ the output of the DA algorithm with input μ and the men proposing. By exchanging the roles of men and women we get an alternative algorithm to which we refer to as the female version of DA algorithm (DA_W). When the input is the empty matching, the above formulation corresponds to the McVitie-Wilson [7] version of the deferred acceptance algorithm where at each step at most one pair is satisfied.

3 CYCLES OF $P(\mu)$

Irving and Leather [6] described an algorithm to compute every stable matching.[2] The description of this algorithm is based on the concept of cycle. It is easy to describe a cycle in terms of certain "reduced"preference lists obtained from the original preference lists by eliminating some unachievable agents .

If μ is any stable matching of the model, we will call reduced preferences profile and will denote $P(\mu)$ the profile which is obtained through the following process.

For all $m \in M$ and $w \in W$:

Step 1: Remove from $m's$ list of acceptable women all w who are more preferred than $\mu(m)$. Remove from $w's$ list of acceptable men all m who are more preferred than $\mu_W(w)$.

Step 2: Remove from $m's$ list of acceptable women all w who are less preferred than $\mu_W(m)$. Remove from $w's$ list of acceptable men all m less preferred than $\mu(w)$.

Step 3: After steps 1 and 2, if m is not acceptable to w (i.e., m is not on $w's$ preferences list as now modified),then remove w from $m's$ list of acceptable women. And similarly, if w is not acceptable to m then remove m from the $w's$ list of acceptable men.

Definition 2 *A set of men $\{a_1, \cdots, a_r\}$ is a cycle for the reduced preferences profile $P(\mu)$, if*

i) The second woman in $P(\mu)(a_i)$ is $\mu(a_{i+1})$ for all $i = 1, \cdots, r-1$ (i.e., the first woman in $P(\mu)(a_{i+1})$

ii) The second woman in $P(\mu)(a_r)$ is $\mu(a_1)$ (i.e., the first woman in $P(\mu)(a_1)$

We denote a cycle by $\sigma = (a_1, \cdots, a_r)$ and we say that a_i generates cycle σ for any $i = 1, \cdots, r$. We denote $\Sigma(\mu)$ the set of cycles for $P(\mu)$.

4 OPTIMAL SEQUENTIAL ENTRY

In this section we study a process of sequential entry. We assume that participants enter in the market one at a time, in some arbitrary given order. At the beginning, we assume stability and the Deferred Acceptance algorithm with input matchings is applied after the entry of each agent. At the end this process generates a stable matching. We identified a large family of orders (optimal orders) which converges to the optimal matching. We also show that the last agent who enters the market will receive his/her optimal outcome under stable matchings. We finally observe that not all stable matchings can be obtained under the process of sequential entry.

An order over the agents is a biyective function from $\{n \leq |V|\}$ to V. We denote by Γ the set of all orders. For $\gamma \in \Gamma$, $\gamma(i)$ represents the ith agent that enters in the market and v enters in the market before v' if $\gamma^{-1}(v) < \gamma^{-1}(v')$. We consider the kth reduced market $(M(k), W(k), P_k)$ where $M(k)$ is the set of men that enters in the market until the kth position, i.e. $M(k) = \{m \in M : \gamma^{-1}(m) \leq k\}$. Similarly $W(k)$ is the set of women that enters in the market until the kth position.

The process of sequential entry is described as follows:

[2] See Roth and Sotomayor [9], for a full description of the algorithm.

Let $\gamma \in \Gamma$.

Initial stage

(0) Let $(M(1), W(1), P_1)$ be the fisrt reduced market and $\mu_1(\gamma(1)) = \gamma(1)$.

Iteration stage

For $k = 2$ until $|V|$ do:

(1) Let $(M(k), W(k), P_k)$ be the kth reduced market.

(2) $\mu'_{k-1}(v) = \begin{cases} \mu_{k-1}(v) \text{ if } v \neq \gamma(k) \\ v \text{ if } v = \gamma(k) \end{cases}$ for all $v \in M(k) \cup W(k)$.

(3) If $\gamma(k) \in M$, then $\mu_k = DA_{M(k)}\left(\mu'_{k-1}\right)$,
if not $\mu_k = DA_{W(k)}\left(\mu'_{k-1}\right)$.

Given any order γ we denote μ_γ the matching that is obtained by the application of the process of sequential entry wiht input γ.

Theorem 1 *Let $\gamma \in \Gamma$. Then μ_γ is stable.*

The following theorem shows that the last agent who enters the market receives his/her optimal outcome under stable matchings.

Theorem 2 *Let $\gamma \in \Gamma$ and $\gamma^{-1}(|V|) = v$. Then $\mu_\gamma(v)$ is the optimal outcome for v under stable matchings.*

For $\gamma \in \Gamma$, we define $A = \{n \leq |V| : \text{there exists } \sigma \in \Sigma(\mu_M) \text{ satisfying } \sigma \subseteq \{\gamma(i) : i \leq n\}\}$ and let $s = minA$.

Definition 3 *Let $\gamma \in \Gamma$. γ is an optimal order for M if:*

(i) For each m such that $\gamma^{-1}(m) \leq s$ then $\gamma^{-1}(\mu_M(m)) < s$.

(ii) For each m such that $\gamma^{-1}(m) > s$ then $\gamma^{-1}(\mu_M(m)) < \gamma^{-1}(m)$.

Loosely speaking the order γ is optimal for M if a man m enters in the market before the first cycle for $P(\mu_M)$ is completed , then $\mu_M(m)$ should also enters before the first cycle is completed, and if a man m enters in the market after the first cycle for $P(\mu_M)$ is completed then $\mu_M(m)$ should enter in the market before him.

The following theorem shows that an optimal order for M leads to the optimal stable matching μ_M.

Theorem 3 *Let γ be an optimal order for M. Then $\mu_\gamma = \mu_M$.*

REFERENCES

[1] P. BIRÓ, K. CECHLÁROVÁ, AND T. FLEINER , *The dynamics of stable matching markets and half-matchings for the stable marriage and roommates problems*, mimeo, Safarik University,2006.

[2] Y. BLUM, A. ROTH, AND U. ROTHBLUM , *Vacancy chains and equilibration in senior-level labor markets*, J. Econom. Theory, 76 (1997), pp.362-411.

[3] Y. BLUM AND U. ROTHBLUM , *Timing is everything and marital bliss*, J. Econom. Theory, 103 (1997), pp.429-443.

[4] D. GALE AND L. SHAPLEY , *Collage admissions and stability of marriage*,Amer. Math. Monthly 69 (1962), pp.9-15.

[5] D. KNUTH , *Marriage Stables*,Les Presses de l'Universite Montreal. (1976).

[6] R.W. IRVING AND P. LEATHER , *The complexity of counting stable marriages* ,SIAM Journal on Computing, 15 (1986), pp.655-667.

[7] D. MACVITIE AND L. WILSON, *Stable marriage assignment for unequal sets* ,BIT, 10 (1970), pp.295-309.

[8] J. MA, *On randomised matching mechanism* ,Econom. Theory, 8 (2002), pp.377-381.

[9] A. ROTH AND M. SOTOMAYOR , *Two-sided Matching: A Study in Game-Theoretic Modeling and Analysis.*, Vol. 18 of Econometric Society Monographs. Cambridge University Press, Cambridge England, 1990.

[10] A. ROTH AND J. VANDE VATE, *Random paths to stability in two sided matching* ,Econometrica, 58 (1992), pp.1475-1480.

[11] A. TAMURA , *A property of the divorce digraph for a stable marriage* , Research Report B-234, Tokyo Institute of Technology, 1990.

[12] J TAN AND W SU , *On the divorce digraph of the stable marriage problem* , unpublished manuscript.

EXTREMAL MATRICES OF THE CONSTRAINED TRANSPORTATION PROBLEM

Ezio Marchi[♭], Jorge Oviedo[♭] and Pablo Tarazaga[†]

[♭]*Instituto de Matemática Aplicada de San Luis. Universidad Nacional de San Luis and CONICET. Ejército de los Andes 950. 5700, San Luis, Argentina. E-mails: joviedo@unsl.edu.ar*

[†] *Department of Mathematics and Statistics, Texas A & M University, Corpus Christi, TX 78412, USA, tarazaga@sci.tamucc.edu*

Abstract: In this paper we characterize the extremal matrices of the constrained transportation problem. We also introduce an algorithm to compute all the extremals of that convex set, generalizing one presented by Jurkat and Ryser for the transportation problem.

Keywords: *Transportation problem, convexity, extremality.*

1 INTRODUCTION

The classic transportation problem is given as follows: we have m supply depots whose availability of a unique good is giving by s_i units, for $i = 1, ..., m$. This commodity needs to be delivered to n demand depots that require r_j units of the good for $j = 1, ..., n$. We denote by x_{ij} the number of unit transported from the supply depot i to the demand depot j. A natural constraint is x_{ij} 0 for $i = 1, ..., m$ and $j = 1, ..., n$.

We will assume here that the total supply equal the total demand, in other words

$$\sum_{i=1}^{m} s_i = \sum_{j=1}^{n} r_j$$

The problem can be written as follows

$$\sum_{j=1}^{n} x_{ij} = s_i \qquad i = 1, ..., m, \tag{1}$$

$$\sum_{i=1}^{m} x_{ij} = r_j \qquad j = 1, ..., m, \tag{2}$$

$$x_{ij} \geq 0 \qquad i = 1, ..., m; \quad j = 1, ..., n. \tag{3}$$

A matrix X whose entries are x_{ij} and satisfies the restrictions given above is called a transportation plan. It is clear that the set of all transportation plans form a convex set (See [1]). This set will be denoted by T.

An important part of a transportation problem is the cost function that establishes the cost of a transportation plan based on the cost of sending a unit of the commodity from the supply depot i to the demand depot j denoted by γ_{ij} . The total cost of a given transportation plan is given by

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} x_{ij}$$

Jurkat and Ryser studied the extremal points of the convex set T and characterize the extremal transportation plans algorithmically in [2].

We will consider in this paper the problem described by the equation $(1) - (3)$but adding a new set of constraints given as follows:

$$x_{ij} \leq c_{ij} \quad i = 1, ...m; \quad j = 1, ..., n. \tag{4}$$

where this entries c_{ij} set the maximum amount of units that can be transported between the supply depot i and the demand depot j. In order to make the problem feasible we need that

$$\sum_{i=1}^{m} c_{ij} \geq r_j \qquad \text{and} \qquad \sum_{j=1}^{n} c_{ij} \geq s_i$$

for $i = 1, ..., m$ and $j = 1, ..., n$.

We will deal in this paper with the structure of the convex set described by the equations $(1) - (4)$ that we will denote by CT for constrained transportation problem. For this reason we will not refer anymore in this paper to the objective function.

The structure of this paper is as follows: next section introduce the concept of cycles and give a characterization of the extreme points. The following section characterize algorithmically all the extremal points using a generalization of the algorithm used by Jurkat and Ryser in [2].

2 Cycles and Extremality

Given a transportation plan x_{ij}, we say that x_{ij} contain a cycle if there is a set of positions $(i_1, j_1), (i_1, j_2), (i_2, j_2), ..., (i_r$ for which we have that

$$0 < x_{ij} < c_{ij}.$$

The cycle is simple if each row and column contain either 0 or 2 positions of the cycle. We say that X contains a cycle.

Given a cycle, we can define a matrix P of size $m \times n$ as follows: $p_{ij} = 0$ for every entry (i, j) that does not belong to the cycle. For the position in the cycle we define $p(i_1, j_1) = 1$, $p(i_1, j_2) = -1$, $p(i_2, j_1) = 1$, and so on. It is clear that the row and column sums of the matrix P are zero.

Theorem 1 *Given X in CT, X is extremal if and only if X does not contain a cycle.*

Proof. First we prove that if X is extremal then X does not contain a cycle, proving the contrapositive. Assume that X contains a cycle P, then there exists $\epsilon > 0$ small enough such that $X + \epsilon P$ and $X - \epsilon P$ in CT. Now we have that

$$X = \frac{1}{2}(X + \varepsilon P) + \frac{1}{2}(X - \varepsilon P),$$

which implies that X is not extremal.

To prove the reciprocal we use again the contrapositive. If X is not extremal, there exist matrices A and B in CT, both satisfying $X \neq A$, $X \neq B$ and

$$X = \frac{1}{2}A + \frac{1}{2}B.$$

Since A and B are distinct from X, there exist at least a position (i_1, j_1) where

$$0 \leq a_{i_1 j_1} < x_{i_1 j_1} < b_{i_1 j_1} \leq c_{i_1 j_1}.$$

Note that the role of A and B are interchangeable in the inequalities above.

Now we will present an argument that allow us to build a cycle. First observe that for any position (i, j) where A or B are distinct from X, one of the following inequalities happen

$$0 < a_{ij} < x_{ij} < b_{ij} \quad \text{or} \quad 0 < b_{ij} < x_{ij} < a_{ij}$$

since $x_{ij} = \frac{1}{2}a_{ij} + \frac{1}{2}b_{ij}$ and the fact that all the entries of these matrices are nonnegative. Moreover if any entry of X is zero or equal to the corresponding maximum capacity then the matrices A and B have to be equal to X in that entry. Let us define this position as (i_1, j_1) and assume $0 < a_{i_1 j_1} < x_{i_1 j_1} < b_{i_1 j_1}$.

The nonnegativity of the entries of the matrices X, A and B and the fact that all of them have equal row sum for the row i_1, implies that there exists j_2 such that $x_{i_1 j_2} < a_{i_1 j_2}$. But now we have that $x_{i_1 j_2} = \frac{1}{2}a_{i_1 j_2} + \frac{1}{2}b_{i_1 j_2}$, then

$$\begin{aligned} 2x_{i_1 j_2} &= a_{i_1 j_2} + b_{i_1 j_2} \\ 2x_{i_1 j_2} &> x_{i_1 j_2} + b_{i_1 j_2} \\ x_{i_1 j_2} &> b_{i_1 j_2} \end{aligned}$$

which implies

$$b_{i_1 j_2} < x_{i_1 j_2} < a_{i_1 j_2}$$

Now we add the position (i_1, j_2) to build the cycle. The same argument is valid for the column j_2 and so on and because the process is finite, we build a cycle. The nonzero positions in this cycle correspond to position where A and B are distinct to X.

□

3 Algorithmic Characterization of Extremal Matrices

We will introduce a class of matrices in CT using an inductive process. First we need some notation. We denote by S the supply vector $S = (s_1, s_2, ..., s_m)^t$ and the demand vector is defined as $R = (r_1, r_2, ..., r_n)^t$.

The class of matrices that we define below is denoted by $\mathcal{E}(S, R, C)$, where C is the maximum capacities matrices.

We describe now the class as follows: in order to generate a transportation plan X, select a position (i, j) and define the corresponding entry

$$x_{i,j} = min(s_i, r_j, c_{ij})$$

If the minimum is s_i, then fill the rest of the row i with zeros and repeat the same step for the class $\mathcal{E}(S_1, R_1, C_1)$ where $S_1 = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_m)^t$, $R_1 = (r_1, ..., r_{j-1}, r_j - s_i, r_{j+1}, ..., r_n)^t$ and $C_1 = C$. Note that row i of the A matrix is now fixed.

If the minimum is r_j , then fill the the rest of the column j with zeros and repeat the same step for the class $\mathcal{E}(S_1, R_1, C_1)$ where $R_1 = (r_1, ..., r_{j-1}, r_{j+1}, ..., r_m)^t$, $S_1 = (s_1, ..., s_{i-1}, s_i - r_j, s_{i+1}, ..., s_n)^t$ and $C_1 = C$. Note that column j of the A matrix is determined.

If the minimum is c_{ij} , then we consider the class $\mathcal{E}(S_1, R_1, C_1)$, where $S_1 = (s_1, ..., s_{i-1}, s_i - c_{ij}, s_{i+1}, ..., s_n)^t$, $R_1 = (r_1, ..., r_{j-1}, r_j - c_{ij}, r_{j+1}, ..., r_n)^t$ and finally $(C_1)_{st} = c_{st}$ for $(s, t) \neq (i, j)$ and $(C_1)_{ij} = 0$.

Now we will characterize the class of matrices $\mathcal{E}(S, R, C)$. For a line in a matrix we will understand a row or a column.

Lemma 1 *A transportation plan X is in $\mathcal{E}(S, R, C)$ if and only if every submatrix of X contains a line with at most one element different from zero and different from the corresponding maximum capacity.*

Proof. The proof is by induction on the number of lines. It is easy to see that the assertion is trivial for matrices with a single line.

First we prove that the condition is necessary. Given $X \in \mathcal{E}(S, R, C)$, let us denote by X_1 the matrix resulting after eliminating the first line in the construction process. Remember X_1 is in $\mathcal{E}(S_1, R_1, C_1)$ and now take an arbitrary submatrix X_2 of X. If X_2 contain at least part of the deleted line, the result follows. If not X_2 is a submatrix of X_1 and the result follows from the induction hypothesis.

We prove now that the condition is sufficient. Since X satisfies the necessary condition, there is a line with only one element say xij different from zero and the corresponding maximum capacity, in other words

$$0 < x_{ij} < c_{ij}.$$

Then in that line all the other elements are or zero or they are equal to the maximum capacity corresponding to that position. Then we choose first any position in the line where the matrix entry is equal to the maximum capacities (in any order) and then we choose the position where the matrix entry is different from zero and the maximum capacity. At that moment the line sum is achieved and the rest of the line position set to zero. Then the reminding submatrix is in $\mathcal{E}(S_1, R_1, C_1)$ and we can repeat the process and generate the matrix using our inductive process and then the matrix X is in $\mathcal{E}(S, R, C)$.

□

Now we are in condition of establish and prove our main result.

Theorem 2 *A matrix X in CT is an extremal if and only if X is in the class E(S,R,C).*

Proof. It is clear that the theorem is true if X has only one line. Let us prove that the condition is sufficient by induction on the number of lines.

Take $X \in \mathcal{E}(S, R, C)$ and suppose that is not extremal, then there exists matrices A and B in CT such that

$$X = \frac{1}{2}A + \frac{1}{2}B.$$

Since $A \in E(S, R, C)$ then there exist a line with only one element satisfying

$$0 < x_{ij} < c_{ij}.$$

Note that the rest of the entries in the line are zero or equal to the corresponding capacity. Then these entries must be equal to the entries of the matrices A and B. But the same happen for the entries of the three matrices in the position (i, j) since the line has fixed sum. Then X, A and B coincide in that line.

Now consider X_1, A_1 and B_1 as the original matrices, from which the common line is deleted. Observe that still is valid

$$X_1 = \frac{1}{2}A_1 + \frac{1}{2}B_1$$

and these matrices, because the induction hypothesis have to satisfy that $X_1 = A_1 = B_1$. But this together with the fact that X, A and B already coincide in the deleted line imply that $X = A = B$ which complete the proof of the sufficiency.

Now we prove that the condition is necessary using the contrapositive. If we assume that $X \notin \mathcal{E}(S, R, C)$ then any line has at least two entries whose values are strictly between zero and the corresponding maximum capacity. This property implies that we can generate a cycle of position with this property. This process is similar to the one generated in Theorem 1. We can choose an arbitrary row i_0 to start. Then this row has two entries say (i_0, j_0) and (i_0, j_1) satisfying

$$0 < x_{i_0,j_0} < c_{i_0,j_0} \qquad \text{and} \qquad 0 < x_{i_0,j_1} < c_{i_0,j_1}.$$

But now the column j_1 has another entry with that property, say (i_1, j_1). Then we continue the construction with row i_1 and so on. Because the matrix has a finite number of rows an column we are able to close this path generating a cycle.

Now we define a matrix P having values 1 and -1 alternatively on the cycle and zeros otherwise. Clearly the row and column sums of P are zeros which implies that for $\epsilon > 0$

and small the matrices $X + \epsilon P$ and $X - \epsilon P$ are in CT. Then we have

$$X = \frac{1}{2}(X + \epsilon P) + \frac{1}{2}(X - \epsilon P)$$

which implies that X is not extremal, which complete the proof.

□

A very important result is a straightforward consequence of this theorem concerning the values that the entries of the extremal matrices take when the component of the vectors s and r and the entries of the matrix C are integers.

Theorem 3 *If the vectors s and r and the matrix C have integer components and entries, then every extremal matrix has integer entries.*

References

[1] R. A. Brualdi, Combinatorial Matrix Classes, Encyclopedia of Mathematics and its Applications, Vol. 108, Cambridge University Press 2006.

[2] W. B. Jurkat and H. J. Ryser [1967], Term Rank and Permanent of Nonnegative Matrices, Journal of Algebra 5, 342-357.

A MODEL OF STRATEGIC, PRIVATE INCOME TRANSFERS

Maximiliano Miranda Zanetti

CONICET-UNS (Departamento de Economía). mmiranda@uns.edu.ar

Abstract: This paper analyses a model in which two agents, characterized by distinct labor productivities, interact under a framework of income inter-dependence. The main result of this interaction is the possibility of exchange of voluntary, welfare-enhancing income transfers. We characterize possible equilibria and show the existence of pure-strategy Nash equilibria for the underlying game.

KEYWORDS: Nash equilibrium, income redistribution, private transfers

1 INTRODUCTION

Income distribution constitutes a sub-area of Economics full of normative or prescriptive considerations. In particular, several distinct views suggest alternative posits with differing implications. (See, for instance, [6].)

In particular, income redistribution ([1]) is a very sensitive issue within this area. There are three basic approaches to redistribution concerns:

1. The view that redistribution can be characterized as a merit good, or a type of public activity or regulation that should be carried out in the best interest of society as a whole.
2. The approach of considering redistribution as a public good (i.e. a public activity with outstanding spill-over effects over the whole society).
3. In a somewhat limited view of the previous interpretation, the conception of redistribution as an act justified on the basis of the external effects caused by considerations of altruism, generosity, self-sacrifice, etc.

It is in this last line that we formulate a model of interaction between two agents facing external effects on income levels.

2 OUR MODEL

2.1 PREVIOUS RELATED WORKS

The amount of work about the modelling of income redistribution under frameworks of agent interaction (as in strategic games) is certainly not overwhelming. Some works ([4, **?**, 7, 5, 2, 3]) do model redistribution problems under Game Theory frameworks.

Withal, we will present a model different to previous formulations. We will develop a framework under which two agents interact; the scheme is original, as is characterized by the following simutaneous singular properties:

1. Agents derive utility from consumption of income and leisure.
2. Income is endogenously determined within the framework, considering a set of possible productivity levels for each agent.
3. Both own income and the other agent's wealth influence each individual's utility.

4. Utility formulation is essentially the same for any individual (agents are basically distinguished by their relative productivity, but the same utility formula applies for both consumers).

5. All factors comprised in the utility formula interact smoothly (the expression is analogous to a generalized Cobb-Douglas formulation).

2.2 THE MODEL

2.2.1 Preliminaries

We posit a model of interaction between two agents, characterized by their labor productivity, θ_1 and θ_2. Each agent has a utility formulation given by

$$u_i = \ln c_i + \ln\left(y_i - t_i + t_{-i}\right) + \alpha \ln\left(y_{-i} - t_{-i} + t_i\right)$$

for each individual indexed $i = 1, 2$ and the remaining agent $-i = \{1; 2\} \setminus i$. We allow for the possibility that agent i give an amount of income t_i to the other individual.

Both own personal total income ($y_i^d = y_i - t_i + t_{-i}$) and that corresponding to the other agent ($y_{-i}^d = y_{-i} - t_{-i} + t_i$) enter into the utility formulation; notice that the remaining individual income y_{-i}^d enters into the formula weighted by parameter α, $0 < \alpha < 1$. Variables c_i show the level of time leisure consumed. Amounts y_i represent the portion of income derived from work sold at the labor market with $y_i = \theta_i \cdot (1 - c_i)$ (for simplicity, we assume that the total amount of time to spend on labor and leisure is the unit).

Transfers are deemed to fall into the interval $[0; y_i]$.

Labor productivities are constants (publicly known by both agents) drawn from some interval of possible values $\left[\underline{\theta}; \overline{\theta}\right]$.

2.2.2 Time structure

We will assume that individuals (knowing both productivity levels) make a simultaneous choice of leisure and income levels c_i, y_i. On a later stage, agents get to know the rival's decision and are then faced with the optimal choice of transfer amounts.

3 RESULTS

3.1 EQUILIBRIA

An elegant approach to the problem resolution is to solve the game by means of *backward induction*. We first consider the optimal response of agent i in terms of transfers t_i given levels $\bar{c}_i, \bar{c}_{-i}, \bar{y}_i, \bar{y}_{-i}$ of leisure and (labor) income. The best strategy at this stage is given by

Lemma 1 *Optimal transfers*

For all possible income combinations y_i,y_{-i}, the best amount of transfers is given by the formula

$$t_i\left(y_i, y_{-i}\right) = \max\left(\frac{\alpha y_i - y_{-i}}{1 + \alpha}, 0\right) \tag{1}$$

Corollary 1 *It is the agent with the highest income the only one that will give away resources, and he will do so only if his chosen income times α is greater than the rival's labor earnings.*

Given this optimal sub-game response, we turn to the first-stage best response. Given best-response transfers given by the formula (1), and taking into account the labor market restriction, we are able to re-formulate the strategic problem in terms of incomes alone. We can then define the indirect utility for each individual in terms of income: $v_i\left(y_i, y_{-i}\right)$.

First-order conditions must be considered for each possible alternative: mainly the three possible stances as transfer receptor, transfer donor and as an AUTARKIC AGENT (this last qualification refers to a situation in which both agents have set income levels fulfilling the condition $\alpha y_{-i} \leq y_i \leq \dfrac{y_{-i}}{\alpha}$, in which case both $t_1$and t_2 amount to zero).

Considering the possible first-order conditions, we can arrive to the following characterization of possible Nash equilibria:

Theorem 1 *Equilibrium characterization*

Assume without loss of generality $\theta_i \leq \theta_{-i}$. If $(\hat{y}_i, \hat{y}_{-i})$ is a Nash equilibrium for the game, then it takes one of the following forms:

- $(\hat{y}_i, \hat{y}_{-i}) = \left(\hat{y}_i^{A_0}, \hat{y}_{-i}^{C_0}\right) = \left(0, \frac{1+\alpha}{2+\alpha}\theta_{-i}\right)$
- $(\hat{y}_i, \hat{y}_{-i}) = \left(\hat{y}_i^{A_{int}}, \hat{y}_{-i}^{C_{int}}\right) = \left(\frac{(2+\alpha)\theta_i - \theta_{-i}}{3+\alpha}, \frac{(2+\alpha)\theta_{-i} - \theta_i}{3+\alpha}\right)$
- $(\hat{y}_i, \hat{y}_{-i}) = \left(\hat{y}_i^{B}, \hat{y}_{-i}^{B}\right) = \left(\frac{1}{2}\theta_i, \frac{1}{2}\theta_{-i}\right)$

and such forms could only be achieved respectively under the following (necessary) conditions:

- $\theta_{-i} \geq (2+\alpha)\,\theta_i$
- $\dfrac{2\theta_i}{1+\alpha} < \theta_{-i} < (2+\alpha)\,\theta_i$
- $\theta_{-i} \leq \dfrac{\theta_i}{\alpha}$

Corollary 2 *Let $(\hat{y}_i, \hat{y}_{-i})$ be a Nash equilibrium for the game. Then, $\hat{y}_i \leq \hat{y}_{-i} \iff \theta_i \leq \theta_{-i}$.*

The most important technical result of this analysis is given by the next theorem, which presents a positive result on existence of equilibria for the game:

Theorem 2 *Existence of equilibria*

For every possible combination of productivities θ_1, θ_2 and interdependence weight α, there exists (at least) one pure-strategy Nash equilibrium for the game.

It is indeed the case that, under certain values of the parameters, equilibrium multiplicity may arise. However, such cases are comparatively rare over the parameter space. It is possible to calculate necessary and sufficient conditions for the prevalence of unicity for each type of equilibrium forms given by theorem 1, as well as sufficient conditions for each possible multiplicity case; those conditions, however, are cumbersome and do not provide much insight or intution onto the equilibrium prevalence issue.

3.2 EFFICIENCY

Having characterized the equilibria for the game (and proved their existence), we tackle the issue of efficiency considerations.

We first notice that equilibria (considering retracement to the first-stage income option) are Pareto-inefficient:

Theorem 3 *Inefficiency of equilibria*

Every Nash equilibrium $(\hat{y}_1, \hat{y}_2)$ *for the retracted game is Pareto-inefficient: It is possible to find a pair of incomes* $\tilde{y}_1, \tilde{y}_2$ *such that for each agent, the corresponding indirect utility* $v_i\left(y_i, y_{-i}\right)$ *is greater with the new pair of incomes, for each individual.*

It is also possible to find a similar (but not so dramatic) result for the second stage of the game:

Theorem 4 *Inefficiency of individually optimal transfers*

Given any set of incomes$(\hat{y}_i, \hat{y}_{-i})$*, the Pareto-efficient transfers must satisfy*

$$t_i\left(y_i, y_{-i}\right) - t_{-i}\left(y_i, y_{-i}\right) = \frac{y_i - y_{-i}}{2} \tag{2}$$

which is of course not satisfied by the formula (1)*.*

4 CONCLUDING REMARKS

We have presented an interaction framework between two agents, under which arises the possible existence of welfare-enhancing voluntary income transfers.

These transfers will more likely occur, and be of greater amount, the greater the difference between the innate capabilities of the agents (here the parameter α acts as an amplifier [or compressor] of the differences, being greater the redistribution the greater the value of the parameter).

Thus, these trasfers act as a means of reducing the natural differences accounted between individuals. This redistributive system acts by means of private, voluntary income flows occurring under a Nash equilibrium of the underlying game.

According to the findings of section 3.2, there possibly exist better allotments in terms of efficiency. Notice however that it is not clear how such assignments may arise, and surely such social improvements require compelling action by a central (and possibly less-than-perfectly informed) authority.

REFERENCES

[1] Boadway R. y Keen M. (2000). "Redistribution". *Handbook of Income Distribution*. Elsevier Science B.V.

[2] Clotfelter, C.T. (1985). *Federal Tax Policy and Charitable Giving*. Chicago Press, Chicago.

[3] Duncan, B. (1999). "Modeling charitable contributions of time and money". *Journal of Public Economics* 72.

[4] Nakayama, M. (1980). "Nash Equilibria and Pareto Optimal Income Redistribution". *Econometrica* 48.

[5] Roberts, R.D. (1984). "A Positive Model of Private Charity and Public Transfers". *Journal of Political Economy* 92.

[6] Sen, A. (2000). "Social justice and the distribution of income". *Handbook of Income Distribution*. Elsevier Science B.V.

[7] Sugden, R. (1982). "On the Economics of Philanthropy". *The Economic Journal* 92.

[8] Warr, P.G. (1982). "Pareto Optimal Redistribution and Private Charity". *Journal of Public Economics* 19.

MATCHING WITH CONTRACTS: CALCULATION OF ALL STABLE ALLOCATIONS

Eliana Beatriz Pepa Risma[♭]

[♭]*Instituto de Matemática Aplicada San Luis, Universidad Nacional de San Luis, Ejército de los Andes 950, CP 5700 San Luis, Argentina, ebpepa@unsl.edu.ar*

Abstract: Here we aim to increase the knowledge about the set of stable allocations introduced by Hatfield and Milgrom at their model of matching with contracts many-to-one (2005) which included, as special cases, the college admission problem, the Kelso-crawford labor market matching model, and ascending package auctions. For this purpose, we present an algorithm to calcule the full set of stable allocations requesting only Substitutability for the hospitals preferences, condition that ensures the existence of a stable allocation. Doing a few natural changes, our algorithm works in the more general context of matching with contracts many to many.

Keywords: *matching, contracts, all stable allocations*
2000 AMS Subject Classification: C78

1 PRELIMINARIES

Hatfield and Milgrom 2005 considered a model of matching with contracts which included, as special cases, the college admission problem, the Kelso-crawford labor market matching model, and ascending package auctions. Below we describe that model. here are two disjoint sets of agents: the set of n hospitals H and the set of m doctors D, and the finite set of contracts is denoted by X . We assume that each contract $x \in X$ is bilateral, so that it is associated with one doctor $x_D \in D$ and one hospital $x_H \in H$. Each doctor can sign only one contract and each hospital can sign multiple contracts. Given $Y \subset X$ a subset of contracts, for each $d \in D$ and $h \in H$ we denote: $Y|_d = \{d \in D : x_D = d\}$ and $Y|_d = \{h \in H : x_H = h\}$ Each doctor $d \in D$ has a strict preference relation P_d over the elements of $X|_d \cup \{\emptyset\}$ (here $\emptyset$ represents the null contract) and each hospital $h \in H$ has a strict preference relation P_h over the set of all the subsets of $X|_h$ including the null contract $\emptyset$. Preference profiles are (m+n)-tuples of preference relations and they are represented by $P = (P_{d_1}, ..., P_{d_m}; P_{h_1}, ..., P_{h_n})$ We denote a model of contracts (X,P) and for the rest of this paper we will work with a fixed contracts model (X,P).

Given a set $Y \subset X$, for each $d \in D$ and $h \in H$ we denote d's most-preferred contract in Y and h's most-preferred subset of Y (choice sets) in the following way: $C_d(Y) = max_{P_d}[Y|_d \cup \emptyset]$ and $C_h(Y) = max_{P_h}\{Z \subset Y|_h : (x, x' \in Z; \quad x \ unequal \ to \ x' \ implicates \ x_D \ unequal \ to \ x'_D)\}$

In add, we denote: $C_D(Y) = \bigcup_{d \in D} C_d(Y)$, $C_H(Y) = \bigcup_{h \in H} C_h(Y)$, $R_D(Y) = Y - C_D(Y)$ and $R_H(Y) = Y - C_H(Y)$

The next concept emphasizes the fact that each doctor can sign at most one contract.

Definition 1 *A subset of contracts $Y \subset X$ is an allocation if $x, x' \in Y$ implicates that x_D is different to x'_D.*

The allocations such that there is no a different allocation strictly preferred by some hospital and weakly preferred by all of the doctors hired by that hospital, and such that no doctor strictly prefers to reject his contract are stable in the sense that there is no coalition that can improve by deviating from it. Next we formalize the definiton of stable allocation.

Definition 2 *An allocation $Y \subset X$ is stable if*

i) $C_D(Y) = C_H(Y) = Y$ and

ii) there exist no $h \in H$ and set of contracts Z$\subset C_h(Y)$ such that $Z = C_h(Y \cap Z) \subset C_D(Y \cap Z)$. The set of stable allocations for the model (X,P) will be denoted S(X,P)..

The substitutability condition states that if a contract is chosen by an agent from some set of available contracts, then that contract will still be chosen from any smaller set that includes it.

Definition 3 *A preferences ordering P_i of agent i satisfies Substitutability if for any subsets $Z \subset Y \subset X$we have that $R_i(Z) \subset R_i(Y)$.*

From now on, we will suppose that the preferences of all agents $i \in D \cap H$ satisfy substitutability (for doctors' side, this condition is satisfied trivially).

2 COMPUTATION OF S(X,P)

In this section, we exhibit an algorithm to compute the full set of stable allocation S(X,P) for a profile of preferences P where the preferences of all agents satisfy Substitutability. Note that in this many to one market the preferences of the doctors satisfy Substitutability trivially. This result is an adaptation to this context of the result obtained by Ruth Martinez et al (2003) for matching many to many without contracts. We remark that the extension of our algorithm to the more general market of matching with contracts many-to-many is almost trivial.

Definition 4 *Given the original profil of preferences P and the contract $x \in X$, the profil of preferences $P^x$is the x-truncation of P if:*

1 All sets containing x are unacceptable to x_H in accordance with $P^x_{x_H}$, that is, $x \in S$ implicates $\emptyset P^x_{x_H} S$.

2 Preferences P_{x_H} and $P^x_{x_H}$ coincide on all sets that do not contain x, that is , $x \in S_1 \cap S_2$ implicates $S_1 P^x_{x_H} S_2$ if and only if $S_1 P_{x_H} S_2$.

3 Preferences P_{x_H} and $P^x_{x_H}$ coincide on all pair of sets containing x, that is, $x \in S_1 \cap S_2$ implicates $S_1 P^x_{x_H} S_2$ if and only if $S_1 P_{x_H} S_2$.

4 Set artificially made unacceptable in $P^x_{x_H}$ is preferred to the original unacceptable sets, that is, if S_1 and S_2 are such that $x \in S_1$ and $S_1 P_{x_H} \emptyset P_{x_H} S_2$ then $S_1 P^x_{x_H} S_2$.

Conditions 3 and 4 are irrelevant for stability but given the profile of preferences P and $x \in X$,they guarantee the uniqueness of the x-truncation of P

5 All $i \in D \cup H - \{x_H\}$, $P^x_i = P_i$.

We will denote O^x_H and O^x_D the optimal allocations under P^x for the hospitals and for the doctors respectively, they can be calculated using iterated applications of the function F like in theorem 3 in Hatfield and Milgrom (2005) using the profile of preferences P^x in place of P. Let $S(X, P^x)$ be the set of all stable allocations under the profil of preferences P^x

The algorithm consists of applying sussesively the following procedure: First, using the original profile as imput, compute the stable allocations O_H and O_D following the theorem 3 in Hatfield and Milgrom (2005). Second, for each $x \in O_H - O_D$ obtain the x-truncation P^x and using this new profile of preferences calcule the (new) optimal for hospitals (O^x_H in this case). Third, for $x \in O_H - O_D$ it could be that O^x_H does not belong to S(X,P). But as we will see, if O^x_H satisfies $O^x_H|_{x_D} P_{x_D} O_H|_{x_D}$, the stability of O^x_H relative to the original profile P is guaranteed. In such case, we keep O^x_H and proceed again from the very beginning using this modified profile as imput. The algorithm stops when there is no contract belonging to the intersection between the complement of O_D and the optimal for hospitals (relative to the present truncated preference profile).

2.1 FORMAL DEFINITION OF THE ALGORITHM

At the following formal definition there is a prescindible step which is used only in order to do agiler the algorithm.

BEGINING Set T(X,P):=P , S(X,P):=$\{O_H\}$ and k:=0
REPEAT
Step 1: Define $V(T^k(X,P)) = \{P^{x_1 \ldots x_K x} : x \in O_H^{x_1 \ldots x_K} - O_H \quad and \quad P^{x1 \ldots x_K} \in T^k(X,P)\}$
Step 2: if $V(T^k(X,P)) = \emptyset$ set $T^k(X,P) = \emptyset$ and $S^{k+1}(X,P) = S^k(X,P)$.
ELSE, for each truncation $P^{x_1 \ldots x_k x} \in V(T^k(X,P))$ obtain the allocation $O_H^{x_1 \ldots x_k x}$, it there exist due to lemma 2 which is stated below.
Step 3: Define $T^*(T^k(X,P)) = \{P^{x_1 \ldots x_k x} \in V(T^k(X,P)) : O_H^{x_1 \ldots x_k x}|_{x_D} P_{x_D} O_H^{x_1 \ldots x_k}|_{x_D}\}$
Arrange the set $T^*(T^k(X,P))$ in an arbitrary way and let $<^{k+1}$ denote this ordering.
Step 4: Define $W(T^k(X,P)) = \{P^{x_1 \ldots x_k x} \in T^*(T^k(X,P)) : \quad for \quad all \quad P^{x_1 \ldots x_K x'} \in T^*(T^k(X,P))$ such that $P^{x_1 \ldots x_K x} <^{k+1} P^{x_1 \ldots x_K x'} \quad and \quad x' \in O_H^{x_1 \ldots x_k x}\}$
Set $T^{k+1}(X,P) := W(T^k(X,P))$
$S^{k+1}(X,P) := S^k(X,P) \cup \{O_H^{x_1 \ldots x_k x} : P^{x_1 \ldots x_k x} \in T^{k+1}(X,P)\}$
k:=k+1
UNTIL $T^k(X,P)$ is empty.
END.

2.2 MAIN THEOREM

In order to demonstrate effectiveness of the algorithm we set the theorem 1 at the end of this section. Before, we enounce some lemmas which yield to prove the theorem. Their proofs are omitted here because of the lack of space.

Lemma 1 *Given $A \subset X$, $C^x_{x_H}(A) = C_{x_H}(A - \{x\})$.*

Lemma 2 *If $x'_H s$ original preferences P_{x_H} satisfy Substitutability, then the modified preferences $P^x_{x_H}$ also satisfy Substitutability.*

Let $S^x(X,P)$ be the set of stable allocations (respect to the profile P^x) satisfying the following property: $S^x(X,P) = \{X \in S(X,P^x) : X|_{x_D} P_{x_D} O_H|_{x_D}\}$. Lemma 2.3 proves $S^x(X,P) \subset S(X,P)$. So, the property $X|_{x_D} P_{x_D} O_H|_{x_D}$ is sufficient to guarantee the stability respect to the original profile of an allocation X which is stable respect to P^x.

Lemma 3 *Given $x \in O_H - O_D$. If $X' \in S^x(X,P)$ is an allocation, then $X' \in S(X,P)$.*

Lemma 4 is proved using lemma 3.

Lemma 4 *Let P^x be a x-truncation such that $O^x_H|_{x_D} P_{x_D} O_H|_{x_D}$. Then, $X' \in S(X,P^x)$ implicates $X' \in S(X,P)$.*

Lemma 5 *Let X' be an allocation such that $X' \in S(X,P)$ and $C^x_i(X') = X'|_i \quad$ for all $i \in D \cup H$. Then, $X' \in S(X,P^x)$.*

The following is a corollary of lemma 5

Corollary 1 *Let P^x and $P^{x'}$ be two truncations such that $O^x_H \in S(X,P)$ and $O^{x'}_H \in S(X,P)$. If x does not belong to $O^{x'}_H$, then $O^{x'}_H \in S(X,P^x)$.*

Lemma 6 *If $x \in O_H \cap O_D$, then $x \in X'$ for all $X' \in S(X, P)$.*

Lemma 7 *Let $X' \in S(X, P)$ such that X' is different to O_H. Then $X' \in S(X, P^x)$ for some P^x with $x \in O_H - O_D$ such that x does not belong to X'.*

Next there is a corollary of lemma 7

Corollary 2 *Let $X' \in S(X, P)$ be an allocation different to O_H. Then, there exists a sucession of contracts $x_1, ..., x_n$ such that $X' = O_H^{x_1 ... x_n} \in S(X, P^{x1 ... x_n})$.*

Theorem 1 *Let K be the step where the algorithm stops, that is, $T^K(X, P) = \emptyset$. If the preferences of all agents satisfy Substitutability in the original profile P then $S^K(X, P) = S(X, P)$.*

Proof. First, due to lemma 4, $S(X, P) \subset S(X, P)$. Applying iteratively the lemma 4 at the successive stages, we obtain $S^K(X, P) \subset S(X, P)$. Suppose $X' \in S(X, P)$, then because of corollary 2, there exists $k \in \{0, ..., K\} = K$ such that $X' \in S^k(X, P)$. Consequently, $S(X, P) \subset S^K(X, P)$. Therefore, $S(X, P) = S^K(X, P)$. □

3 REFERENCES

[1] Matching with contracts- John William Hatfield, Paul R. Milgrom (2005)

[2] An Algorithm to Compute the Full Set of Many to Many Stable Matching- Ruth Martinez, Jordi Mass, Alejandro Neme, Jorge Oviedo (2003)

MODELO GENERALIZADO CON RESTRICCIÓN DE CAPACIDAD

Alicia Pedrosa[♭]

[♭]*Departamento de Matemática, Universidad Nacional de San Juan, Ignacio de la Roza 230 Oeste, San Juan, Argentina, aliciaepfa@yahoo.com.ar*

Resumen: En este trabajo se presenta el modelo de asignación generalizado con restricción de capacidad $((A,T),q)$, que consiste en un conjunto de agentes A, la tabla de preferencias de los agentes T, una empresa E que ofrece alojamientos, la cual dispone de q habitaciones. El propósito es encontrar asignaciones estables de parejas de compañeros de cuarto o ciclos de compañeros de cuarto de medio tiempo o agentes sin asignar. En este modelo analizamos la existencia de medio matchings estables con cuota q, que en este contexto denominaremos medio matchings $q-$estables. Bajo preferencias responsive de la institución se muestra que el conjunto de medio matchings estables para el modelo (A,T) es no vacío y se dan luego algunas caracterizaciones de este conjunto con número de habitaciones igual o menor que q.

Palabras clave: *medio matching, restricción de cuota, q- estables*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Dado un conjunto de agentes A el **problema de compañeros de cuarto** consiste en agrupar los agentes en pares donde cada $x \in A$ tiene un orden de preferencias $\succ_x$ sobre los agentes de $A' \subseteq A$

Una preferencia es un orden completo, reflexivo y transitivo.

Describimos, para un problema dado, el conjunto de listas de preferencia de los agentes de A mediante una tabla T, denominada 'tabla de preferencias'.

Se considera a la tabla T, simétrica, es decir si un agente m está en la lista del agente i entonces i está en la lista del agente m.

Notación 1.1 *Dado un conjunto de agentes A y una tabla T de preferencia de los agentes de A sobre A. Con (A,T) indicamos a un modelo de compañeros de cuarto.*

Escribimos $r(a,b)=k$ para indicar que la persona b está en la posición k en la lista de preferencia de a.

$r(a,b) < r(a,c)$ si y sólo si $b \succ_a c$, esto es, a prefiere b a c

$r(a,b) \leq r(a,c)$ si y sólo si $b \succeq_a c$, es decir, a prefiere b tanto como a c

Nos interesa asignar parejas de compañeros de cuarto, para ello consideramos todas las posibles asignaciones entre elementos del conjunto A.

Llamaremos matching a una asignación de compañeros de cuarto.

Definición 1.1 *Sea (A,T) un modelo de compañeros de cuarto.*

Un **matching** *μ es una asignación uno a uno de A sobre A tal que*

- *$\mu(x) \in A$ para todo $x \in A$*
- *$\mu(x) = y$ si y sólo si $\mu(y) = x$, para todo $\{x,y\} \subset A$.*

Definición 1.2 *Dos agentes $\{x,y\}$* **bloquean** *un matching μ si no están asignados como compañeros de cuarto en el matching μ y ambos se prefieren estrictamente a los compañeros de cuarto asignados por μ.*

Es decir, $\{x,y\} \subset A, x \neq y$, bloquea a μ si $r(x,y) < r(x,\mu(x))$ y $r(y,x) < r(y,\mu(y))$

Definición 1.3 *Un matching es* **estable** *si y sólo si no posee pares de agentes que lo bloqueen.*

El propósito es encontrar asignaciones estables para este problema, es decir encontrar asignaciones o matchings que no estén bloqueados por un par de agentes, pero el conjunto de los matchings estables en el problema de compañeros de cuarto puede ser vacío.

El problema de existencia de matchings estables en el problema general de compañeros de cuarto puede ser tratado utilizando el algoritmo de Gusfield e Irving [2], el que mediante un proceso de eliminación de entradas de las listas de preferencia y de las llamadas rotaciones expuestas, permite determinar con certeza la existencia o no existencia de matching estable para el problema dado.

Consideremos una relación de preferencia cualquiera donde no necesariamente el número de agentes en A es par.

Imaginemos que los agentes del ejemplo anterior son jugadores de tenis, donde a cada uno se le asigna un compañero para jugar una hora. Tan [3] mostró que si los agentes pueden asignarse para jugar media hora, existe una solución estable en el sentido que ningún par de jugadores asignados desearía incrementar ese período.

Tenemos entonces que considerar que cada agente puede tener un compañero, dos *compañeros de medio tiempo*, un compañero de medio tiempo o quedar sin asignar.

Definición 1.4

En este nuevo conjunto tenemos asignación de parejas, de ciclos de compañeros de medio tiempo y agentes sin asignar, y podemos definir estabilidad de la siguiente forma:

Definición 1.5 *Sea (A, T) un modelo de compañeros de cuarto. Sea μ un medio matching asociado a este modelo,*

$\{a, b\} \subseteq A$ es un **bloqueo** *para μ si y sólo si: $r(a, b) < r(a, \mu_1(a))$ y $r(b, a) < r(b, \mu_1(b))$*

Es decir, a prefiere a b a su peor asignación $\mu_1(a)$ ($b \succ_a \mu_1(a)$) y b prefiere a a a su peor asignación $\mu_1(b)$ ($a \succ_b \mu_1(b)$)

Definición 1.6 *Un medio matching μ es* **estable** *si no es bloqueado por dos agentes.*

Dado un modelo de compañeros de cuarto, aplicando el algoritmo de Irving modificado por Tan, se pueden encontrar todos los medio matching estables existentes.

El propósito es encontrar, cuando sea posible, los matchings estables para una relación de preferencia dada, para lo cual debería lograrse que todos los ciclos que definen el medio matching estable asociado a ese modelo tengan cardinalidad par. Como esto no es factible en general nos proponemos encontrar los medio matchings estables.

Queremos analizar la conveniencia de elegir un medio matching estable a otro, para ello estudiamos una variable más, el "número de habitaciones asignadas a un medio matching estable".

Podemos establecer que el número de habitaciones necesarias para un medio matching estable está dado por

$\#_h(\mu) = \#P_2(\mu) + \sum_{B \in \mathcal{B}} \frac{\#B}{2}$, siendo $P_2(\mu)$ el conjunto de parejas asignadas por μ y $\mathcal{B}$ el conjunto de ciclos impares definidos por μ.

Es claro que dos medio matchings estables para un modelo de compañeros de cuarto requieren el mismo número de habitaciones.

Se pretende analizar el Modelo de asignación generalizado con restricción de capacidad (que generaliza a Femenía y otros [1]). Para ello se considera (A, T) un modelo de compañeros de cuarto, suponemos que una empresa E ofrece alojamientos, que tiene preferencias $\succ_E$ sobre los agentes de A y dispone de q habitaciones.

Tenemos así un modelo de compañeros de cuarto con restricción de capacidad, que denotamos por $((A, T), q)$.

μ será aceptado por E si $\#_h(\mu) \leq q$

Analizamos el comportamiento de los medio matchings en un modelo generalizado con restricción de capacidad con la finalidad de establecer la existencia de medio matchings estables con cuota q, que en este contexto denominaremos medio matchings $q-$estables. Para ello debemos definir las nociones de $q-$bloqueo y de $q-$estabilidad.

Definimos

Definición 1.7 *Sea $((A,T),q)$ un modelo de compañeros de cuarto con restricción de cuota q. Sea μ un medio matching asociado a este modelo. Sea $\succ_E$ la preferencia de la empresa E sobre los agentes de A. Sea P_E las preferencias de E sobre $\mathcal{P}_{\mathcal{M}}$.*

*$\{x,y\} \subseteq S$ es un $q-$***bloqueo** *para μ si y sólo si*

i) $\mu_2(x) \neq x$, $\mu_2(y) \neq y$, $r(x,y) < r(x,\mu_1(x))$ y $r(y,x) < r(y,\mu_1(y))$ *ó*

ii) $\mu_2(y) = y$, $r(x,y) < r(x,\mu_1(x))$, $r(y,x) < r(y,y)$ y $\mu_{\{x,y\}} P_E \mu$ y $\#_h\left(\mu_{\{x,y\}}\right) \leq q$

Es decir que, para establecer si los agentes $\{x,y\}$ constituyen un $q-$bloqueo para μ necesitamos realizar dos análisis: en primer lugar si los agentes están asignados o no. Si están asignados es porque su mejor opción no son ellos mismos ($\mu_2(x) \neq x$, $\mu_2(y) \neq y$). Luego analizamos si prefieren estar asignados entre ellos que estar con quien les fuera asignado en el medio matching μ.

El segundo análisis corresponde al caso en que uno de ellos es single (por ejemplo y). Se prefieren mutuamente (x prefiere a y antes que al agente que tiene asignado), a la empresa le interesa tal asignación $\left(\mu_{\{x,y\}} P_E \mu\right)$ y el número de habitaciones de este nuevo medio matching no excede a q $\left(\#_h\left(\mu_{\{x,y\}}\right) \leq q\right)$.

Definición 1.8 *Dado el modelo de compañeros de cuarto con restricción de capacidad $((A,T),q)$. Un medio matching se dice $q-$***estable** *si no posee pares $q-$bloqueantes.*

Notación 1.2 *Indicamos con $\mathcal{S}(A,T)$ al conjunto de los medio matchings estables para el modelo (A,T) y con $\mathcal{S}_q(A,T)$ al conjunto de los medio matchings $q-$estables, es decir estables para el modelo $((A,T),q)$.*

En el modelo $((A,T),q)$ podemos demostrar mediante un ejemplo que el conjunto $\mathcal{S}_q(A,T)$ puede ser vacío. Por tal motivo nos restringiremos al modelo enel cual la preferencia de la empresa es responsive.

Las preferencias de la empresa R_E se llaman responsive respecto a las preferencias $\succ_E$ sobre los agentes individuales si para cualquier par de subconjuntos de agentes que difieren en sólo un agente la empresa prefiere al que contiene al agente más preferido (y es indiferente entre ellos si es indiferente entre los agentes).

Sabemos por Tan que el conjunto de medio matchings estables para el modelo (A,T) es no vacío, probamos ahora que el conjunto de medio matchings estables para el modelo con restricción de capacidad $((A,T),q)$ es también no vacío.

Es decir, $\mathcal{S}_q(A,T) \neq \emptyset$.

Teorema 1.1 *El conjunto de medio matchings $q-$estables en el modelo (A,T) es no vacío.*

Analizaremos los medio matchings $q-$estables asociados al modelo (A,T) en dos partes, la primera cuando tienen número de habitaciones igual a q y la segunda cuando tienen número de habitaciones menor que q.

Denotemos con $x_1, x_2, \cdots, x_n$ a los agentes del conjunto A.

Sin pérdida de generalidad y con la intención de simplificar la notación, supongamos que $x_1 \succ_E x_2 \succ_E x_3 \succ_E \cdots \succ_E x_n$, de tal forma que $x_i \succ_E x_j$ para todo $i < j$, es decir que los agentes identificados con el subíndice menor son más preferidos por E que los de subíndice mayor.

Consideremos los conjuntos $A^i = A - \{x \in A : x_i \succ_E x, i > 1\} = \{x_1, \cdots, x_i\}$, conjuntos que se obtienen suprimiendo de A los agentes menos preferidos por la empresa E, y sean T^i las tablas que se obtienen de la tabla T suprimiendo a los agentes que no pertenecen a A^i.

Para $i \in I\!N$ simbolizamos con (A^i, T^i) el modelo reducido de (A, T) y con $\mathcal{S}(A^i, T^i)$ al conjunto de los medio matchings estables en el modelo reducido.

Como para todo modelo (A, T) el conjunto $\mathcal{S}(A, T) \neq \emptyset$ podemos afirmar que para todo $i \in J, \mathcal{S}(A^i, T^i) \neq \emptyset$.

En el modelo $((A, T), q)$ consideremos entonces sólo los conjuntos $\mathcal{S}(A^i, T^i)$ cuyos medio matchings tienen número de habitaciones igual a q

Para cada $i \in J$ definimos el conjunto

$M_q(A^i, T^i) = \begin{cases} \mathcal{S}(A^i, T^i) & \text{si para } \mu \in \mathcal{S}(A^i, T^i), \#_h(\mu) = q \\ \emptyset & \text{en otro caso} \end{cases}$

Definimos $M_q(A, T) = \bigcup_{i \in J} M_q(A^i, T^i)$

Mostramos que todo medio matching que está en $M_q(A, T)$ es estable para el modelo (S, T). Es decir

Teorema 1.2 $M_q(A, T) \subseteq \mathcal{S}_q(A, T)$

Consideramos ahora el conjunto $M_{<q}(A^i, T^i) = \{\mu \in \mathcal{S}(A^i, T^i) : \#_h(\mu) < q$ y todo $x, y \in A : \mu_2(x) = x, \mu_2(y) = y$ no son mutuamente aceptables$\}$

Observemos que $M_{<q}(A^i, T^i) = \mathcal{S}(A^i, T^i)$ ó $M_{<q}(A^i, T^i) = \emptyset$

A partir de estos conjuntos definimos un nuevo conjunto:

$M_{<q}(A, T) = \bigcup_{i \in J} M_{<q}(A^i, T^i)$ y probamos que

Teorema 1.3 *Si* $((A, T), q, \succ_E^*)$ *es un modelo de compañeros de cuarto con restricción de cuota, entonces* $M_{<q}(A, T) \subseteq \mathcal{S}_q(A, T)$

Teorema 1.4 $\mathcal{S}(A, T) = M_q(A, T) \cup M_{<q}(A, T)$

REFERENCIAS

[1] D. FEMENIA, M. MARÍ, A. NEME AND J. OVIEDO (2008). "Stable solutions on matchings models with quota restrictions ". (En prensa)

[2] GUSFIELD D. AND IRVING R., *The Stable Marriage Problem: Structure and Algorithms.* The MIT Press. Massachusetts Institute of Technology. Cambridge. Massachusetts 02142. 1989

[3] TAN J., *A necessary and sufficient condition for the existence of a complete stable matching.* Journal of Algorithms 12.154-178(1991).

A NON-CONSTRUCTIVE PROOF OF THE EXISTENCE OF STABLE MATCHINGS IN THE MARRIAGE MODEL

Juan Carlos Cesco

Abstract: In this note we present a non-constructive proof of the existence of stable matchings in the marriage model which uses a game theoretic approach. To this end, we develop a theory of hedonic partitioning games. Our approach differs from that used by Sotomayor (1996) based upon fixed point theory.

Keywords: *Marriage model, stable matchings, hedonic games*
2000 AMS Subject Classification: 91A12

1 INTRODUCTION

The seminal paper of Gale and Shapley [2] was the starting point of modern matching theory. There, the marriage model is developed and the central notion of stable matching is introduced. A proof of its existence for any marriage model is carried out by designing a computational procedure, the deferred acceptance algorithm, which is proven to converge to a stable matching. It was until 1996 when Sotomayor [6] presented the first non-constructive proof about the existence of stable matchings in the marriage model. She used a fixed point approach to get a very simple proof. In this note, from another point of view, which takes advantages of the relationships existing between two models of coalition formation, namely, matching models and hedonic games, we present another non-constructive proof of the existence of stable matchings for the marriage model. Toward this end, we study properties of a subclass of hedonic games with a restricted family of coalitions which captures, in a hedonic framework, the relevant characteristics of the partitioning games studied by Kaneko and Wooders [4] in the context of games with and without transferable utility.

2 PARTITIONING HEDONIC GAMES

We start with a finite set $N = \{1, ..., n\}$ whose elements are going to be called the players, while a subset of them will be a coalition. Given any family $\mathcal{B}$ of coalitions, and a player $i \in N$, let us denote by $\mathcal{B}(i)$ the subfamily of those coalitions in $\mathcal{B}$ containing player i. A family of coalitions $\mathcal{A}$ such that $\{i\} \in \mathcal{A}$ for all $i \in N$ is called a family of basic coalitions (Kaneko and Wooders [4]). A hedonic game with $\mathcal{B}$ as its family of basic coalitions is a 3-tuple $(N, \succeq; \mathcal{A})$, where N is the set of players and $\succeq = (\succeq_i)_{i \in N}$ is a preference profile with $\succeq_i$ being a reflexive, complete and transitive binary relation on $\mathcal{A}(i)$ for each $i \in N$. An individual preference is strict if $S \succeq_i T$ and $S \neq T$ implies that not $T \succeq_i S$). For each $i \in N, \succeq_i$ will stand for the strict preference relation related to $\succeq_i$ ($S \succ_i T$ iff $S \succeq_i T$ but not $T \succeq_i S$). $\mathcal{P}^{\mathcal{A}}(N)$ will denote the family of partitions of N having all its elements in $\mathcal{A}$. Given $\pi = \{\pi_1, ..., \pi_p\} \in \mathcal{P}^{\mathcal{A}}(N)$ and $i \in N, \pi(i)$ will denote the unique set in π containing player i.

Given a hedonic game $(N, \succeq; \mathcal{A})$ and $\pi \in \mathcal{P}^{\mathcal{A}}(N)$, we say that $T \in \mathcal{A}$ blocks π if for each $i \in T, T \succ_i \pi(i)$. The core $C(N, \succeq; \mathcal{A})$ of $(N, \succeq; \mathcal{A})$ is the set of partitions in $\mathcal{P}^{\mathcal{A}}(N)$ blocked by no coalition $T \in \mathcal{A}$.

A family of non-empty coalitions $\mathcal{B} \subseteq \mathcal{N}$ is called balanced if there exists a set of positive real numbers (the balancing weights) $(\lambda_S)_{S \in \mathcal{B}}$ satisfying $\sum_{S \in \mathcal{B}(i)} \lambda_S = 1$, for all $i \in N$. $\mathcal{B}$ is minimal balanced if there is no proper balanced subfamily of it. In this case, the set of balanced weights is unique. Let us call a family of basic coalitions $\mathcal{A}$ partitionable (Kaneko and Wooders [4]) if the only minimal subfamilies that it contains are partitions.

A family $\mathcal{I} = (\mathcal{I}(A))_{A \in \mathcal{A}}$ is called an $\mathcal{A}$-distribution, or simply a distribution (Iehlé [i]) if, for each non-empty coalition $A \in \mathcal{A}$, $\phi \neq \mathcal{I}(A) \subseteq A$ Given a distribution $\mathcal{I}$, a family $\mathcal{B} \subseteq \mathcal{A}$ is $\mathcal{I}$-balanced if the family $(\mathcal{I}(B))_{B \in \mathcal{B}}$ is balanced.

Definition 1 *$(N, \succeq; \mathcal{A})$ is ordinally balanced if for each balanced family $\mathcal{B} \subseteq \mathcal{A}$ there exists a partition π, whose elements belong to $\mathcal{A}$, such that, for each $i \in N, \pi(i) \succeq_i B$ for some $B \in \mathcal{B}(i)$.*

Definition 2 *$(N, \succeq; \mathcal{A})$ is pivotally balanced with respect to an $\mathcal{A}$-distribution $\mathcal{I}$, if for each $\mathcal{I}$-balanced family $\mathcal{B}$, there exists a partition π whose elements belong to $\mathcal{A}$ such that, for each $i \in N, \pi(i) \succeq_i B$ for some $B \in \mathcal{B}(i)$. The game is pivotally balanced if it is pivotally balanced with respect to some distribution $\mathcal{I}$.*

Note 1 *Ordinal balancedness was first introduced by Bogomolnaia and Jackson [1], while the notion of pivotal balancedness was first employed by Iehlé [3], both notions stated for the case in which $\mathcal{A}$ is the whole family of non-empty coalitions.*

The first part of the following theorem is a sufficient condition for the existence of core-partitions for hedonic games with coalitional restrictions which parallels the first part of Theorem 1 in Bogomolnaia and Jackson [1], while the second part parallels the characterization given in Theorem 3 of Iehlé [3], and whose proofs are carried out in a similar way.

Theorem 1 *Let $(N, \succeq; \mathcal{A})$ be a hedonic game with $\mathcal{A}$ as its family of admissible coalitions. a) If the game is ordinally balanced, and has strict individual preferences, then $C(N, \succeq; \mathcal{A})$ is non-empty. b) $C(N, \succeq; \mathcal{A})$ is non-empty if and only if the game is pivotally balanced.*

Note 2 *Ordinal balancedness implies pivotal balancedness with respect to the distribution $\mathcal{I} = (\chi_A)_{A \in \mathcal{A}}$, being χ_A the indicator vector of the coalition A.*

Definition 3 *A basic partitioning hedonic game is a hedonic game $(N, \succeq; \mathcal{A})$ where the family of admissible coalitions is partitionable.*

Theorem 2 *Let a partitionable essential family of coalitions $\mathcal{A}$ be given. Then, every basic partitioning hedonic game $(N, \succeq; \mathcal{A})$ has non-empty core.*

Proof. The proof follows from the fact that the basic partitioning hedonic game $(N, \succeq; \mathcal{A})$ is ordinally balanced. To see this, let $\mathcal{B}$ be a balanced family of coalitions and because $\mathcal{A}$ is partitionable, $\mathcal{B}$ contains a partition π. Then, since for each $i \in N$ it holds that $\pi(i) \succeq_i \pi(i)$, we conclude that the game is ordinally balanced. Thus, by part b) of Theorem 1 we conclude that its core is non-empty. □

Note 3 *We point out that being the individual preferences in the game not necessarily strict, the non-emptiness of the core is guaranteed by part b) rather than by part a) of Theorem 1.*

3 Existence of stable matchings

As a simple but important consequence of Theorem 2, we derive a new proof about the existence of stable matchings in the marriage model of Gale and Shapley [2]. To do this, we first associate a basic partitioning hedonic game to each matching problem. Then, we will show that the core of the game is related to the set of stable matchings. Finally, we will use the fact that the family of admissible coalitions in the game is partitionable to get our result.

The marriage model consists of two finite sets of agents, the sets M of 'men' and the set W of 'women'. It is assumed that each man $m \in M$ is endowed with a preference $\succeq^m$over the set $W \cup \{m\}$, and that each woman w has a preference $\succeq^w$on the set $M \cup \{w\}$. Individual preferences are assumed to be reflexive, complete and transitive on their corresponding domains. Let us denote by $(M, W, \succeq^M, \succeq^W)$ a marriage problem, where $\succeq^M = (\succeq^m)_{m \in M}$ and $\succeq^W = (\succeq^w)_{w \in W}$ are the preference profile corresponding to the men and women.

A matching is a function $\mu : M \cup W \to M \cup W$ satisfying:

a) For each $m \in M$,if $\mu(m) \neq m$, then $\mu(m) \in W$.

b) For each $w \in W$,if $\mu(w) \neq w$, then $\mu(w) \in M$.

c) $\mu(\mu(k)) = k$ for all $k \in M \cup W$.

A matching μ is stable if $\mu(k) \succeq^k k$ for all $k \in M \cup W$ (individual stability) and if there is no pair $m \in M, w \in W$ such that $\mu(m) \neq w, \mu(w) \neq m$, and $w \succ^m \mu(m)$ and $m \succ^w \mu(w)$ (pairwise stability).

The pair (m, w) is called a blocking pair. Given a matching problem $(M, W, \succeq^M, \succeq^W)$, let us consider the family $\mathcal{A} = \{S \subseteq M \cup W : |S \cap M| = 1 \text{ or } |S \cap W| = 1\}$. Clearly $\mathcal{A}$ is basic. Moreover, it is a partitionable family as follows from the result of Kaneko and Wooders [4] about the existence of core-points for every assignment game (Shapley and Shubik [5]).

With each matching problem $(M, W, \succeq^M, \succeq^W)$ we associate a basic partitioning hedonic game $(N, \hat{\succeq}; \mathcal{A})$ where $N = M \cup W$, and for each $i \in N$, $\hat{\succeq}_i$ is defined on $\mathcal{A}(i)$ as follows. If $i = m$ for some $m \in M$, $S \hat{\succeq}_i T$ if and only if

$S \cap W \succeq^m T \cap W$ if $|S| = |T| = 2$,
$S \cap W \succeq^m m$ if $|S| = 2$ and $T = 1$,
$m \succeq^m T \cap W$ if $|S| = 1$ and $T = 2$.

If $i = w$ for some $w \in W$, $S \hat{\succeq}_i T$ if and only if

$S \cap M \succeq^m T \cap M$ if $|S| = |T| = 2$,
$S \cap M \succeq^m w$ if $|S| = 2$ and $T = 1$,
$w \succeq^m T \cap M$ if $|S| = 1$ and $T = 2$.

For any $i \in N$, we also declare that $S \succeq_i T$ when $|S| = |T| = 1$.

With each partition π in the game $(N, \hat{\succeq}; \mathcal{A})$, we associate the matching μ^π, where, for each $m \in M$, $\mu^\pi(m) = \pi(m) \cap W$ if $|\pi(w)| = 2$ and $\mu^\pi(m) = m$ if $|\pi(m)| = 1$. Similarly, for each $w \in W, \mu^\pi(w) = \pi(w) \cap W$ if $|\pi(w)| = 2$ and $\mu^\pi(w) = w$ if $|\pi(w)| = 1$.

Now, we are ready to state the following result.

Theorem 3 *Let $(M, W, \succeq^M, \succeq^W)$ be a marriage problem. Then, its set of stable matchings is non-empty.*

Proof. From Theorem 1 we get that $C(N, \hat{\succeq}; \mathcal{A}) \neq \phi$. We claim μ^π is a stable matching for each core-partition π. To see this, let $m \in M$. If $\mu^\pi(m) \neq m$, then $|\pi(m)| = 2$ and since π is a core partition, m can not be strictly preferred to $\pi(m)$. Thus, $\pi(m) \hat{\succeq}_m m$, and according to the definition $\hat{\succeq}_m$this implies that $\pi(m) \cap W = \mu^\pi(m) \succeq^m m$. A similar argument shows that, for each $w \in W, \mu^\pi(w) \succeq^w w$ for any $w \in W$ such that $\mu^\pi(w) \neq w$. Then, μ^π is individually stable.

On the other hand, let us assume that there is a blocking pair (m, w) to μ^π. We claim that the coalition $S = \{m, w\}$ objects the partition π, leading to a contradiction. Indeed, from $w \succ^m \mu^\pi(m)$ we get that $S \cap W \succ^m \mu^\pi(m) \cap W$ or, equivalently, that $S \hat{\succ}_m \pi(m)$ when $\mu^\pi(m) \neq m$ ($|\pi(m)| = 2$), and we also get that $S \hat{\succ}^m \pi(m)$ when $\pi(m) = m$ ($|\pi(m)| = 1$). In a similar way we obtain that $m \succ^w \mu^\pi(w)$ implies that $S \hat{\succ}_w \pi(w)$ showing that S blocks π. □

REFERENCES

[1] BOGOMOLNAIA AND A., JACKSON, M. *The stability of hedonic coalition structures.* Games and Economic Behavior 38 (2002), 201-230.

[2] D. GALE AND L. SHAPLEY. *College admissions and the stability of marriage.* American Mathematical Monthly 69 (1962), 9-15.

[3] V. IEHLÉ. emphThe core-partition of a hedonic game. Mathematical Social Sciences 54 (2007), 176-185.

[4] M. KANEKO AND M. WOODERS. *Cores of partitioning games.* Mathematical Social Sciences 3 (1982), 313-327.

[5] L. SHAPLEY AND M. SHUBIK. *The assignment game I: The core.* International Journal of Game Theory 1 (1972), 111-130.

[6] M. SOTOMAYOR. *A non-constructive elementary proof of the existence of stable marriages.* Games and Economic Behavior 13 (1996), 135–137.

EL IMPACTO DE CONOCER EL NÚMERO DE OFERENTES SOBRE LOS RESULTADOS DE UNA SUBASTA.

Andrés Fioriti†

†Grupo de Investigación "Información en Economía: Teoría y Evidencia", Universidad Nacional del Sur, Avenida Alem 1245, 8000 Bahía Blanca, Argentina, elpepi@gmail.com, www.uns.edu.ar

Resumen: En una subasta la cantidad de rivales a las que un oferente se enfrenta puede ser desconocida. El presente trabajo busca encontrar las funciones de oferta de equilibrio tanto en una subasta de primer precio (SPP) como en una subasta de segundo precio (SSP), donde las valuaciones son privadas e independientes. Se demostrará que los oferentes prefieren que se revele cuántos rivales se enfrentan en una SPP, que son indiferentes entre saber o no la cantidad de rivales en una SSP, y que prefieren la SSP a la SPP. Mientras que los vendedores, si son menos pesimistas que los oferentes, prefieren no revelar la cantidad de oferentes en una SPP y optar por la SPP frente a la SSP.

Palabras claves: *Subastas, Incertidumbre, Cantidad de oferentes.*

1. INTRODUCCIÓN

Información asimétrica y competencia imperfecta son los dos pilares fundamentales de la teoría de subastas. En los modelos se asume que cierta información es de conocimiento común: todos los oferentes saben con cuántos rivales, m, se enfrentan y m es un número fijo. Usando como paradigma m fijo el análisis se simplifica y produce importantes resultados en lo correspondiente al ranking de los ingresos generados por las distintas subastas y en el diseño de las subastas óptimas. Pero este supuesto pocas veces se cumple en la realidad.

En una subasta inglesa un oferente, usualmente, no puede identificar a sus rivales. Los otros oferentes pueden estar actuando en representación de algún jefe, siendo que este puede ser el vendedor, por lo cual no todas las personas presentes serían oferentes activos. Dichos agentes se pueden comunicar con el subastador mediante una mueca, un gesto particular con la cabeza, mover un lápiz, o incluso mirar al subastador a los ojos, siendo todos estos hechos perfectamente legales.

Si la subasta se desarrolla en sobres cerrados, entonces las razones para justificar que un oferente conozca cuántos rivales enfrenta son aún menores, dado que los competidores no se reúnen en un mismo lugar para realizar su oferta.

La pregunta es si los resultados de la teoría de subastas, principalmente en las de primer y segundo precio (SPP y SSP respectivamente), son sensibles al supuesto de que cada oferente conoce exactamente a cuántos rivales enfrenta. La respuesta pareciera ser que sí.

Milgrom y Weber [1] demostraron que, en varias circunstancias, está en el interés del subastador revelar la información que posee. Pero más tarde McAfee y McMillan [2] demostraron que, si los oferentes son adversos al riesgo, el ingreso esperado por el vendedor en una SPP es mayor si los oferentes desconocen cuántos rivales enfrentan que si conocieran tal información. También llegaron a la conclusión de que, si los oferentes son neutrales al riesgo, la subasta óptima y directa que sea compatible con los incentivos es la misma independientemente que los oferentes conozcan o no cuánta competencia enfrentan.

2. EL MODELO

Siguiendo el planteo de Levin y Ozdenoren [3] se asume que existe un bien indivisible para vender. Se consideran SPP y SSP en las cuales existe un precio de reserva, r, que es de conocimiento común. Los competidores ofrecen en forma simultánea. Aquellas ofertas que no igualen o superen el precio de reserva r no son aceptadas. El intervalo en el cual se encuentran las valuaciones de los oferentes es $[0,1]$, siendo v_i la valuación del oferente i y $r < 1$. Las valuaciones de los oferentes son conocidas en forma privada, y es de

común conocimiento que las valuaciones son extracciones independientes de la distribución $F(\cdot)$ en el intervalo $[0,1]$ con una función de densidad $f(\cdot)$ positiva.

En la subasta hay como máximo $\mathbf{m}$ potenciales oferentes indexados según $i = 1, \dots, m$. En el modelo se asume que los oferentes no conocen la cantidad activa de competidores que participan en la subasta. Los oferentes son maximizadores maximin de la utilidad esperada (MMUE), lo cual significa que los oferentes tienen creencias a priori sobre cuántos rivales enfrentan y su utilidad es el mínimo de la utilidad esperada a medida que las creencias varían.

Formalmente sea P un conjunto cerrado y convexo sobre las medidas del subconjunto $\{1, \dots, m\}$. P representa la creencia de un potencial oferente acerca de la cantidad de competidores en una subasta, incluyéndose a sí mismo, si él se transforma en un oferente activo. Esto es, para cada $p \in P$ y $k \leq m$, y para algún oferente activo, $p(k)$ denota la probabilidad de que hay k oferentes activos en la subasta. La probabilidad de que un oferente potencial pase a ser un oferente activo no depende de la valuación que tenga, por lo cual se asume que P es independiente de las valuaciones e igual para todos los oferentes.

Se analizarán dos tipos de subastas con esta modalidad. En primer lugar se hará un análisis de la SPP, en la cual el oferente que más ofrece recibe el objeto y paga exactamente lo que ofreció al vendedor. En segundo lugar se analizará la SSP, en la cual el oferente que más ofrece recibe el objeto y paga exactamente lo que ofreció el segundo que más ofreció al vendedor.

3. SUBASTA DE PRIMER PRECIO (SPP)

El número de oferentes activos en una subasta es $n \in \{1, \dots, m\}$ y se los puede enumerar como $1, \dots, n$. La oferta del individuo j es $b_j \in [0, \infty)$. Sea z_n^i la mayor oferta contra la cual compite el agente i, donde $z_1^i = 0$ y $z_n^i = \max\left\{b_j | j \neq i, j \in \{1, \dots, n\}\right\}$ para $n \geq 2$. t representa la cantidad de oferentes cuya oferta iguala a la oferta de i, incluyendo la oferta de i, formalmente $t = \#\left\{j | b_j = b_i, j \in \{1, \dots, n\}\right\}$. La función de pagos del oferente i en una SPP está dada por:

$$\pi_i(v_i, b_i, z_n^i) = \begin{cases} \frac{1}{t}(v_i - b_i) & \text{si } b_i \geq z_n^i \text{ y } b_i \geq r \\ 0 & \text{si } b_i < z_n \text{ o } b_i < r \end{cases}$$

La estrategia del oferente j se denota como $s_j: [0,1] \to [0, \infty)$, por ende si el oferente j es un participante activo de la subasta con una valuación v_j, su oferta será $s_j(v_j)$. Siendo $H(\cdot|n)$ la función de distribución de z_n^i, para cada $p \in P$ $H^p(\cdot) = \sum_{n=1}^m p(n)H(\cdot|n)$. La función de utilidad del individuo i, siendo que tiene preferencias MMUE y sin tener en cuenta los empates en las ofertas, se puede expresar como:

$$U_i\left(v_i, b_i, \{s_j\}_{j \neq i}\right) = \begin{cases} \min_{p \in P}[(v_i - b_i)H^p(b_i)] & \text{si } b_i \geq r \\ 0 & \text{si } b_i < r \end{cases}$$

En el presente modelo el vector de las estrategias de oferta, $\{s_i\}_{i=1}^m$, es un vector de equilibrio para todo i y $v_i \in [0,1]$ si:

$$U_i\left(v_i, s_i(v_i), \{s_j\}_{j \neq i}\right) \geq U_i\left(v_i, b_i, \{s_j\}_{j \neq i}\right) \quad \text{para todo } b_i \in [0, \infty)$$

Dado que el objetivo es encontrar un equilibrio simétrico en una SPP, y suponiendo que el individuo i sabe que todos los otros oferentes están usando una función de oferta $s(\cdot)$ simétrica, estrictamente creciente y diferenciable, para cada $b \in [0, \infty)$:

$$H^p(\cdot) = \sum_{n=1}^m p(n)H(b|n) = H^p(\cdot) = \sum_{n=1}^m p(n)F^{n-1}(s^{-1}(b))$$

De donde se puede reescribir la función de utilidad de i como:

$$U_i\left(v_i, b_i, \{s_j\}_{j\neq i}\right) = \begin{cases} (v_i - b_i)\min_{p\in P}\sum_{n=1}^{m} p(n)F^{n-1}(s^{-1}(b_i)) & \text{si } b_i \geq r \\ 0 & \text{si } b_i < r \end{cases}$$

Estableciendo $G^{min}(v) = \min_{p\in P}\sum_{n=1}^{m} p(n)F^{n-1}(v)$ se puede reescribir como:

$$U_i\left(v_i, b_i, \{s_j\}_{j\neq i}\right) = \begin{cases} (v_i - b_i)\min_{p\in P}G^{min}\left(s^{-1}(b_i)\right) & \text{si } b_i \geq r \\ 0 & \text{si } b_i < r \end{cases}$$

Proposición 1 En una SPP la única función de oferta simétrica de equilibrio está dada por:

$$s^{min}(v) = v - \frac{\int_r^v G^{min}(t)\,dt}{G^{min}(v)} \text{ para todo } v \in [0,1]$$

Donde $s^{min}(v)$ representa lo mínimo que debe ofertar un individuo cuya valuación del bien es v.

4. Subasta de Segundo Precio (SSP)

En este caso es trivial ver que en una SSP el equilibrio usual, en el cual cada oferente oferta su verdadera valuación siempre y cuando ésta esté por encima del precio de reserva es el único equilibrio posible, dado que es una estrategia dominante, aún cuando los oferentes desconocen cuánta competencia enfrentan como consecuencia de que para establecer la mejor estrategia un oferente solo debe analizar su valuación del bien independientemente de lo que opinen el resto de los oferentes.

5. Estimación Individual de probabilidades

Los oferentes que participan en la subasta asignan subjetivamente, y a priori, las distintas probabilidades sobre la cantidad de agentes que participan en la subasta. A modo de ejemplo se puede decir que un agente, siendo que hay 4 oferentes potenciales (incluyéndose a sí mismo), puede establecer $p(1) = 0.1$, $p(2) = 0.1$, $p(3) = 0.3$ y $p(4) = 0.5$, de lo cual se deduce que el escenario más probable para él es que en la subasta haya 4 oferentes activos.

Dentro de este marco, conjuntamente con el comportamiento de los agentes como MMUE, es que tiene sentido realizar estática comparativa en el comportamiento de los oferentes. Sin embargo se debe hacer la salvedad de que en el caso de la SSP realizar análisis de estática comparativa en el comportamiento de los oferentes con respecto a su grado de aversión al riesgo es trivial porque la estrategia de que cada oferente ofrezca su propia valuación es débilmente dominante, y cualquier cambio en el grado de aversión al riesgo no generará modificaciones en el comportamiento.

En el caso de oferentes que se comportan según MMUE se dice que uno es más pesimista que otro si su conjunto de probabilidades, con respecto a la cantidad de oferentes que enfrenta en la subasta, es de mayor magnitud que el conjunto de probabilidades del otro individuo. Formalmente, un individuo i, con una estimación individual de probabilidades P, se dice que es más adverso al riesgo que otro individuo j, cuya estimación de probabilidades es $\hat{P}$, si $\hat{P} \subset P$.

La consecuencia directa de que un individuo sea más adverso al riesgo que otro es que el primero va a ofertar más agresivamente, porque el temor a perder la subasta la genera una desutilidad mayor que la que le genera ofertar un valor más próximo a su valuación del bien.

6. Conclusiones

Desde el lado de los oferentes la incertidumbre sobre la cantidad de participantes activos en una subasta los conduce a inferir la cantidad de rivales que enfrentan asignando una estimación de probabilidades a las distintas cantidades de competidores en base a la cantidad total de competidores potenciales. La modelización MMUE permite distinguir el grado de aversión al riesgo en base a las probabilidades que cada agente asigna a los distintos escenarios posibles, siendo el más adverso el que cree que hay mayor cantidad de competidores en la subasta.

La incertidumbre sobre la cantidad de rivales, sumado a la aversión al riesgo, genera que los oferentes se comporten más agresivamente que si conocieran a cuántos se enfrentan, y por lo cual su utilidad disminuye (como en la mayoría de los casos en los cuales tiene lugar información asimétrica).

Pero estos hechos sólo tienen lugar en el caso de una SPP, dado que en una SSP sigue siendo una estrategia dominante ofertar acorde a la valuación personal que cada uno tiene del bien independientemente de conocer o no la cantidad de rivales. Es por esta modificación en el comportamiento de los agentes en la SPP cuando tiene lugar la incertidumbre y por permanecer invariable su conducta en la SSP que los oferentes prefieren la SSP por sobre la SPP, dado que no deben enfrentar la incertidumbre y por ende no disminuye su utilidad.

Además, la SPP y la SSP, de no medir la incertidumbre, serían equivalentes según el principio de equivalencia del ingreso, por lo cual se puede inferir que los agentes prefieren que el subastador les revele a cuántos rivales enfrentan a que no se los diga, dado que así aumentaría su utilidad.

A modo de resumen se puede decir que los oferentes prefieren la SSP a la SPP en el caso de incertidumbre sobre la cantidad de competidores y que prefieren que el subastador les diga a cuántos se enfrentan a no saberlo y por ende tener mayor incertidumbre.

Desde el punto de vista del vendedor, o subastador, la situación es la opuesta. Dado que la SPP, con incertidumbre sobre la cantidad de competidores, genera que los oferentes se comporten más agresivamente y así ofrecen más, hace que la estrategia óptima sea no informar a los agentes sobre la cantidad de rivales, porque así aumentan su ingreso/beneficio. Además, como la incertidumbre solo modifica el comportamiento en la SPP pero no en la SSP, prefieren la SPP porque así se aseguran un mayor beneficio.

A modo de resumen se puede decir que los vendedores prefieren la SPP a la SSP en el caso de incertidumbre sobre la cantidad de competidores y que prefieren no revelar cuántos competidores activos hay en la subasta y que los oferentes tengan mayor incertidumbre para que se comporten más agresivamente.

REFERENCIAS

[1] P. MILGROM, AND R.J. WEBER, *A theory of auctions and competitive bidding*, Econometrica 50 (1984), pp.1089-1122.

[2] R.P. MCAFEE, AND J. MCMILLAN, *Auctions with a Stochastic Number of Bidders*, Journal of Economic Theory 43 (1987), pp.1-19.

[3] D. LEVIN, AND E. OZDENOREN, *Auctions with uncertain numbers of bidders*, Journal of Economic Theory 118 (2004), pp.229-251.

[4] V. KRISHNA, *Auction Theory*, Academic Press (2002), San Diego, California.

[5] J. RILEY AND W. SAMUELSON, *Optimal Auctions*, American Economic Review 71 (1981), pp.381-392.

[6] R. HARSTAD, J. KAGEL AND D. LEVIN, *Equilibrium Bid Functions for Auctions with an Uncertain Number of Bidders*, Economic Letters 33 (1990), pp.35-40.

REALLOCATION IN MIXED OWNERSHIP ECONOMIES WITH SINGLE-PEAKED PREFERENCES

Agustín G. Bonifacio†

†*Instituto de Matemática Aplicada San Luis, Universidad Nacional de San Luis and CONICET,*
abonifacio@unsl.edu.ar

Abstract: In the context of a mixed ownership economy consisting of agents with single-peaked preferences and individual endowments together with an outside obligation with the rest of the world, we study rules to reallocate the resources available to the agents involved. We show that the adaptation to our model of the *uniform reallocation rule* fulfills nice properties (obligation monotonicity, efficiency, consistency and replacement monotonicity) and that is immune to strategic behavior in a very general way: it is a *bribe-proof* rule. A rule is bribe-proof if no group of agents can compensate one of its subgroups to misrepresent their characteristics (either preferences or endowments, or both at the same time) and, after an appropriate redistribution of their shares, each obtain a weakly preferred share and all agents in the misrepresenting subgroup obtain a strictly preferred share. Bribe-proofness includes as special cases known and widely studied non-manipulation properties such as strategy-proofness, withholding-proofness and borrowing-proofness.

Keywords: *mixed ownership economies, uniform reallocation rule, bribe-proofness.*
2000 AMS Subject Classification: 91B54 - 91B32

1 INTRODUCTION

In this paper we study the problem of reallocating a perfectly divisible good among a group of agents with initial endowments and an outside obligation with the "rest of the world". We restraint our analysis to the case where agent's preferences are single-peaked: each agent has an optimal amount of the good, below which and above which preference is decreasing.

In the situation that involves only individual endowments to be reallocated, a solution concept that satisfies many appealing properties is the *uniform reallocation rule*. This rule has been extensively studied by Klaus [3], and Klaus, Peters and Storcken [4] & [5]. In these works, several characterizations of the rule are presented and relations among different properties established.

Here we enlarge the domain over which solutions are defined by specifying an economy not only with agent's preferences and endowments but considering also an amount that represent net trades with the outside world. We call such economy a *mixed ownership economy*, following Thomson [9]. In this context, a solution provides for each economy a way of redistributing the endowments and accommodating the net trade amount as a function of agent's preferences. The principal reason for this extension is to formulate in a natural way an important property known as *consistency*. A rule is consistent when the alternative chosen by the rule for an economy is in agreement with the alternative chosen for each of the reduced economies that result when some agents have received the net trade given by the rule and left. Although the property of consistency has been studied in the model that considers initial endowments (for example, in Klaus *et al* [4] a related property called bilateral consistency is studied), the problem in such setting is to define a reduced economy properly, since we need to match the net trade left by the agents who leave with the resources available to the agents of the reduced economy. With our model the way to make the connection is quite natural: through an adjustment of the outside obligation.

Another properties we discuss are *efficiency*, *replacement monotonicity* and a resource monotonicity property we call *(one-sided) obligation monotonicity*. Efficiency is the usual Pareto optimality criterium. Replacement monotonicity requires that, if an agent receives a larger amount after changing his preference relation, then all the other agents should receive smaller amounts. Obligation monotonicity describes the change of the solution when a variation in the outside obligation is considered. It says that by decreasing (increasing) the obligation in case of excess demand (supply) no individual is better off than before. We show that any obligation monotonic, efficient and consistent rule is replacement monotonic.

In this general setting of mixed ownership economies, we are particularly interested in the straightforward generalization of the uniform reallocation scheme, that we will continue to call the uniform reallocation rule[1]. All the properties presented above are shown to be fulfilled by the uniform reallocation rule.

We also investigate the strategic aspect of the problem by considering manipulability issues. Allocation rules can often be manipulated by agents misrepresenting their preferences. A *strategy-proof* rule does not allow for such a behavior. An agent can also benefit by manipulating the resources he controls. The manipulation through endowments may be perform in two ways. In the first, an agent could, by withholding some of his endowment prior to the operation of the rule, and after adding the resources he withheld to the consumption that the rule assigns to him, end up better off than if he had not withheld. A rule that cannot be manipulated in such fashion is called *withholding-proof*. In the second, suppose that prior to the operation of the chosen rule, an agent borrows resources to enlarge his endowment. The rule is then applied, the agent receives his assigned consumption, and returns what he had borrowed. The end result may be an outcome that he prefers to the one that he would have been assigned had he not borrowed. A rule not subject to this kind of manipulation fulfills the property of *borrowing-proofness*.

Another source of manipulation could arise in situations in which an agent borrows from one of his fellow traders. He should of course provide the lender the incentive to do so: after the borrower has returned what he borrowed, the lender should be better off lending than not doing it at all. A rule immune to this behavior is called *borrowing-from-fellow-trader-proof*.

Our objetive is to study an interesting property, called *bribe-proofness* (introduced by Massó and Neme [7] in the context of a social endowment instead of individual ones plus and outside obligation), that includes all the aforementioned types of manipulation as particular cases. If a rule is bribe-proof, then no group of agents can compensate one of its subgroups to misrepresent their characteristics (either preferences or endowments, or both at the same time) and, after an appropriate redistribution of their shares, each obtain a weakly preferred share and all agents in the misrepresenting subgroup obtain a strictly preferred share. We show that the uniform reallocation rule is bribe-proof, and discuss some consequences of this fact.

2 PRELIMINARIES

Let $N = \{1, \dots, n\}$ be the set of agents. Each $i \in N$ is characterized by an endowment $\omega_i \in \mathbb{R}_+$ of the good and a preference relation R_i defined over $\mathbb{R}$. Call P_i to the strict preference associated with R_i. We suppose that agents' preferences are *single-peaked*, i.e., each R_i has a unique maximum $\tau_i = \tau(R_i) \in \mathbb{R}_+$ (its "peak") such that, for any pair $z_i, z_i' \in \mathbb{R}$, we have $z_i P_i z_i'$ as long as either $z_i' < z_i < \tau_i$ or $\tau_i < z_i < z_i'$ holds. Denote by $\mathcal{R}$ the domain of single-peaked preferences defined on $\mathbb{R}$. Let $T \in \mathbb{R}$ be an *outside obligation* to or from the rest of the world. A *mixed ownership economy* will consist of a profile of preferences $R \in \mathcal{R}^N$, an initial endowment vector $\omega \in \mathbb{R}_+^N$ and an (outside) obligation $T \in \mathbb{R}$ with $\sum_{j \in N} \omega_j + T \geq 0$ and will be denoted by (R, ω, T). Call $\mathcal{E}^N$ the domain of mixed ownership economies with at most N agents. An allocation $z = (z_1, \dots, z_n) \in \mathbb{R}_+^n$ is *feasible* if $\sum_{j \in N} z_j = \sum_{j \in N} \omega_j + T$. Denote by $Z(R, \omega, T)$ the set of feasible allocations for economy (R, ω, T). A *rule* is a function that associates to each mixed ownership economy (R, ω, T) an element of $Z(R, \omega, T)$.

We now define formally the properties a rule should satisfy:

Definition 1 A rule f is **efficient** if $f(R, \omega, T) = z$ and there is no other feasible allocation y such that $y_i R_i z_i$ for all i and $y_j P_j z_j$ for some j.

For the definition of consistency, we need to define beforehand what a reduced economy is. Given a mixed ownership economy (R, ω, T), an allocation $z \in Z(R, \omega, T)$ and a set $S \subset N$, the associated *reduced economy* $(R, \omega, T')_{S,z}$ is the economy formed by the members of S with the same preferences and endowments as in the original economy and outside obligation $T' := T + \sum_{j \in N \setminus S} (\omega_j - z_j)$.

Definition 2 A rule is **consistent** if it recommends to each reduced economy its corresponding reduced allocation, i.e., if f is a consistent rule, (R, ω, T) is a mixed ownership economy and $f(R, \omega, T) = z$, then $\forall S \subset N$ we have $f((R, \omega, T)_{S,z}) = z_S$, where $z_S := (z_j)_{j \in S} \in \mathbb{R}_+^S$.

[1] Thomson [9] calls this rule "generalized uniform rule".

The two monotonicity requirements are the following:

Definition 3 A rule f is **replacement monotonic** if for all $(R,\omega,T)\in\mathcal{E}^N$, and $R_i'\in\mathcal{R}$, we have that $[f_i(R,\omega,T)\leq f_i(R_i',R_{-i},\omega,T)]$ implies $[f_j(R,\omega,T)\geq f_j(R_i',R_{-i},\omega,T)\ \forall j\neq i]$.

Definition 4 We say that a rule f is **(one-sided) obligation monotonic** if, given $(R,\omega,T)\in\mathcal{E}^N$ and $T'\in\mathbb{R}$ with $T'\leq T$ then: (i) $\sum_{j\in N}\omega_j+T\leq\sum_{j\in N}\tau_j$ implies $f_i(R,\omega,T)R_if_i(R,\omega,T')\ \forall i\in N$ and (ii) $\sum_{j\in N}\omega_j+T'\geq\sum_{j\in N}\tau_j$ implies $f_i(R,\omega,T')R_if_i(R,\omega,T)\ \forall i\in N$.

Finally we give the formal statement of bribe-proofness:

Definition 5 A rule f is **bribe-proof** if for all $(R,\omega,T)\in\mathcal{E}^N$ and all $V\subseteq S\subseteq N$, there are no $(R_i',\omega_i')_{i\in V}\in\mathcal{R}^V\times\mathbb{R}_+^V$ and $(\ell_j)_{j\in S}\in\mathbb{R}^S$ such that:

1. $\sum_{j\in S}\ell_j=\sum_{j\in S}f_j((R_i',\omega_i')_V,(R,\omega)_{-V},T)+\sum_{i\in V}(\omega_i-\omega_i')$,
2. $\ell_jR_jf_j(R,\omega,T)$ for all $j\in S$ and
3. $\ell_iP_if_i(R,\omega,T)$ for all $i\in V$.

It is easy to check that the following properties of immunity to manipulation are indeed particular instances of bribe-proofness:

1. **Strategy-proofness**. For any $R_i'\in\mathcal{R}$, $f_i(R,\omega,T)R_if_i(R_i',R_{-i},\omega,T)$.
2. **Withholding-proofness**. If $\omega_i'\leq\omega_i$ then $f_i(R,\omega,T)R_i\left(f_i(R,\omega_i',\omega_{-i},T)+(\omega_i-\omega_i')\right)$.
3. **Borrowing-proofness**. If $\omega_i\leq\omega_i'$ then $f_i(R,\omega,T)R_i\left(f_i(R,\omega_i',\omega_{-i},T)-(\omega_i'-\omega_i)\right)$.
4. **Borrowing-from-fellow-trader-proofness**. For each pair $\{i,j\}\subset N$ there is no $b>0$ such that $f_k(R,\omega_i+b,\omega_j-b,\omega_{-\{i,j\}},T)P_kf_k(R,\omega,T)$ for $k=i,j$.

3 Results

Our first result is a general one that relates the properties of obligation monotonicity, efficiency and consistency with that of replacement monotonicity:

Proposition 1 *Any obligation monotonic, efficient and consistent rule is replacement monotonic.*

Next, we define the (generalized) uniform reallocation rule. It is essentially the same definition as in the case of economies without the outside obligation. The only difference, due to the outside obligation agents must face, is that the feasibility scalar that determine the outcome need not be a nonnegative real number:

Definition 6 Given a mixed ownership economy $(R,\omega,T)\in\mathcal{E}^N$, the **uniform reallocation rule** u is defined as follows:

$$u_i(R,\omega,T)=\begin{cases}\min\{\tau_i,\max\{\omega_i+\lambda,0\}\} & \text{if } \sum_{j\in N}\tau_j\geq\sum_{j\in N}\omega_j+T\\ \max\{\tau_i,\omega_i-\lambda\} & \text{if } \sum_{j\in N}\tau_j\leq\sum_{j\in N}\omega_j+T\end{cases}$$

where $\lambda\in\mathbb{R}$ solves $\sum_{j\in N}u_j(R,\omega,T)=\sum_{j\in N}\omega_j+T$.

As mentioned before, the uniform reallocation rule fulfills all the nice properties discussed earlier. Our first result in this respect is the following:

Proposition 2 *The uniform reallocation rule is a efficient, consistent and obligation monotonic rule.*

The previous propositions allow us to state that the rule also satisfies replacement monotonicity.

Corollary 1 *The uniform reallocation rule is replacement monotonic.*

Analyzing the strategic aspect of the rule we obtain the following result

Proposition 3 *The uniform reallocation rule is a bribe proof rule.*

Needless to say, as the uniform rule is bribe-proof, it is also strategy-proof, withholding-proof, borrowing-proof and borrow-from-fellow-trader-proof.

Note 1 When $T = 0$, as noted by Dagan [2], the uniform reallocation rule coincides with the Walrasian rule adapted to the context of single-peaked preferences, the so-called *Walrasian equilibrium with slack* (this solution concept was introduced, among others, by Mas-Colell [6] to deal with satiated preferences in the general equilibrium model). As bribe-proofness implies borrowing-proofness, we obtained a generalization of the Walrasian mechanism that fulfills borrowing-proofness in a (very) restricted preference domain. Identifying such domain restrictions is pointed out by Thomson [8] as an interesting endeavor.

REFERENCES

[1] BARBERÀ, S., M. JACKSON AND A. NEME, *Strategy-Proof Allotment Rules*, Games and Economic Behavior, 18 (1997), pp. 1-21.

[2] DAGAN, N., *Consistent Solutions in Exchange Economies: A Characterization of the Price Mechanism*, Universitat Pompeu Fabra Economics Working Paper #141, 1995.

[3] KLAUS, B., *The Characterisation of the Uniform Reallocation Rule Without Pareto Optimality* in Parthasarathy, T. et al (Eds.) Game Theoretical Applications to Economics and Operations Research (Kluwer), (1997), pp. 239-255.

[4] KLAUS, B., H. PETERS AND T. STORCKEN, *Reallocation of an Infinitely Divisible Good*, Economic Theory, 10 (1997), pp. 305-333.

[5] KLAUS, B., H. PETERS AND T. STORCKEN, *Strategy-Proof Division with Single-Peaked Preferences and Individual Endowments*, Social Choice and Welfare, 15 (1998), pp. 297-311.

[6] MAS-COLELL, A., *Equilibrium Theory with Possibly Satiated Preferences* in Majumdar, M. (Ed.), Equilibrium and Dynamics: Essays in Honor of David Gale (Macmillan), (1992), pp. 201-213.

[7] MASSÓ, J. AND A. NEME, *Bribe-proof Rules in the Division Problem*, Games and Economic Behavior, 61 (2007), pp. 331-343.

[8] THOMSON, W., *Borrowing-proofness*, Rochester Center for Economic Research Working Paper #545, 2008.

[9] THOMSON, W., *Consistent Allocation Rules*, Mimeo, 2010.

[10] THOMSON, W., *Fair Allocation Rules*, Rochester Center for Economic Research Working Paper #539, 2007.

[11] THOMSON, W., *The Replacement Principle in Economies with Single-Peaked Preferences*, Journal of Economic Theory, 76 (1997), pp. 145-168.

EL MODELO DE ASIGNACIÓN VARIOS A UNO CON RESTRICCIÓN DE CAPACIDAD

Delfina Femenia[þ]

[þ]*Departamento de Matemática, Facultad de Filosofía, Humanidades y Artes. Universidad Nacional de San Juan.*
delfinafemenia@speedy.com.ar

Resumen: En este trabajo se presentan una variantes del modelo de asignación varios a uno, en el cual intervienen dos tipos de agentes complementarios (trabajadores del tipo I y trabajadores del tipo II) y una institución, la cual quiere contratar trabajadores para realizar determinadas tareas y cada una de ellas pueden realizarla un trabajador del tipo I con varios trabajadores del tipo II. La institución tiene preferencias sobre los posibles matchings (asignaciones) y tiene una cuota q, que es el número máximo de trabajadores del tipo II que puede contratar. En este modelo se extiende, en un camino natural, el concepto de estabilidad, se define el concepto de q-etabilidad y se estudia la existencia de asignaciones q-estables.

Palabras clave: *matching, restricción de cuota, q_E-estables.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Los *modelos de asignación varios a uno* son utilizados para estudiar problemas de mercados cuyo rasgo distintivo es que los agentes involucrados, desde el comienzo, están en conjuntos disjuntos con características diferentes (por ejemplo, diretores y estudiantes). Cada agente de un conjunto tiene preferencias sobre los agentes del otro conjunto. La naturaleza del problema que se estudia aquí, consiste en asignar a cada agente de un conjunto (conjunto de directores D), varios agentes del otro conjunto (conjunto de estudiantes E). Los agentes que no son asignados con algún agente del otro conjunto son llamados *singles*. El Çollege admissions problem.[es] el nombre dado por Gale and Shapley (1961) al modelo mas simple de estos modelos. En el caso que se asigna a cada agente de un conjunto, a lo sumo un agente del otro conjunto tal modelo es llamado *modelo de asignación uno a uno*

Una variante del modelo de asignación uno a uno fué presentada por Femenia, Marí, Neme and Oviedo [1], el *modelo de asignación uno a uno con restricción de capacidad*, consiste en asignar cada trabajador, de un lado del mercado, con un trabajador, sobre el otro lado, tal que los pares de trabajadores contratados por una institución U son, a lo sumo q. La institución tiene preferencias R_U sobre pares de trabajadores que puede contratar, por ello, la institución deberá elegir a lo sumo q pares de trabajadores acorde con su orden de preferencia.

En este trabajo se presenta *el modelo de asignación varios a uno con restricción de capacidad*, que consiste en asignar cada trabajador, de un lado del mercado ($d \in D$), con varios trabajadores, sobre el otro lado ($e \in E$), tal que los trabajadores de E son a lo sumo q. Es decir, la institución deberá elegir a lo sumo q trabajadores del conjunto E acorde con su orden de preferencia.

2. NOTACIÓN Y RESULTAVOS PREVIOS

Consideramos los conjuntos complementarios D y E y notamos con P_d (P_e) las preferencias de los agentes de D (E) sobre los agentes E (D). Con M indicamos el modelo de asignación uno-uno, con M_U^q el modelo de asignación uno-uno con restricción de cuota con y con $\bar{M}$ el modelo de asignación varios-uno.

Definición 2.1 *Un **matching**, en $\bar{M}$, es una función $\mu : D \cup E \longrightarrow 2^{D \cup E}$ tal que, para todo $d \in D$ y $e \in E$ satisface:*

1. $\mu(e) \subseteq D$ *y* $\#\mu(e) = 1$ *o bien* $\mu(e) = \emptyset$.
2. $\mu(d) \subseteq E$.

3. $\mu(e) = d$ *si y solo si* $e \in \mu(d)$.

Definición 2.2 *Dado* M*, el par de agentes* (d, e) ***bloquea*** *a* μ *si:*

$$eP_d\mu\,(d)\,,\, dP_e\mu\,(e)$$

Definición 2.3 *Dado* M*, un matching* μ *es* ***individualmente racional*** *si para todo agente* $f \in D \cup E$ $\mu(f)P_f\emptyset$

Definición 2.4 *Dado* $\bar{M}$*, el de agentes* (d, e) ***bloquea*** *a* μ *si:*

$e \notin \mu(d)$, $dP_e\mu(e)$, $e \in Ch(\mu(d) \cup \{e\}, P_d)$

Definición 2.5 *Dado* $\bar{M}$*, un matching* μ *es* ***individualmente racional*** *si para todo agente* $e \in E$, $\mu(e)P_e\emptyset$ *y para todo* $d \in D$, $\mu(d) = Ch(\mu(d), P_d)$.

Definición 2.6 *Dado* M_U^q, μ *es* q_E***-individualmente racional*** *si* $\#_E\mu \leq q_E$, $\mu R_U \mu^{\emptyset}$, *para todo* $e \in E$, $\mu(e)P_e\emptyset$ *y para todo* $d \in D$, $\mu(d) = Ch(\mu(d), P_d)$.

Definición 2.7 *Dado* M_U^q*, el par de agentes* (d, e) $q-$***bloquea*** *a* μ *si:*

1. $eP_d\mu\,(d)\,,\, dP_e\mu\,(e)$ *y*
2. *se verifica:*

   a) $\mu(d) \in E$ *y* $\mu(e) \in D$

   o

   b) $\mu_{(d,e)}$ *es* $q-$*individualmente racional y* $\mu_{(d,e)}R_U\mu$.

Un matching es **estable (**q**-estable)** si es individualmente racional (q-individualmente racional) y no está bloqueado (q-bloqueado) por algún par de agente. Notamos el conjunto de los estables de M con $S(M)$, el conjunto de los estables de M_U^q con $S(M_U^q)$ y el conjunto de los estables de $\bar{M}$ con $S(\bar{M})$.

Algunos resultados, respecto a la estabilidad, en estos modelos son:

Teorema 1 *[2]. Si* M *es un modelo de asignación uno a uno, entonces* $S(M) \neq \emptyset$

Teorema 2 *[1]. Si* M_U^q *es un modelo de asignación uno a uno con restriccción de capacidad con preferencia* R_U *responsive, entonces* $S(M_U^q) \neq \emptyset$ *para* P_U *responsive*

Teorema 3 *[1]. Si* M_U^q *es un modelo de asignación uno a uno con restriccción de capacidad,*$S(M_U^q) = T_q(M) \cup T_{<q}(M)$

Teorema 4 *[4]. Si* $\bar{M}$ *es un modelo de asignación varios a uno con preferencias* P_d *responsive, entonces* $S(\bar{M}) \neq \emptyset$

3. EL MODELO DE ASIGNACIÓN VARIOS-UNO CON RESTRICCIÓN DE CUOTA SOBRE E

Notamos tal modelo con $\bar{M}_U^{q_E}$. En este modelo se presentan dos nociones de bloqueo (una general y otra restringida)

Definición 3.1 *Dado $\bar{M}_U^{q_E}$, el par de agentes (d,e) q_E-**bloquea-G** a μ si:*

1. $e \notin \mu(d)$, $dP_e\mu(e)$, $e \in Ch(\mu(d) \cup \{e\}, P_d)$ *y*
2. *se verifica:*

   a) $\mu(d) \neq \emptyset$ *y* $\mu(e) \in D$

   o

   b) *existe* $E' \subseteq \mu(d) \cup \{e\}$ *tal que* $e \in E'$, $\mu_{(d,e)}^{E'}$ *es* q_E-*individualmente racional y* $\mu_{(d,e)}^{E'} R_U \mu$.

Definición 3.2 *Dado $\bar{M}_U^{q_E}$, el par de agentes (d,e) q_E-**bloquea-R** a μ si:*

1. $e \notin \mu(d)$, $dP_e\mu(e)$, $e \in Ch(\mu(d) \cup \{e\}, P_d)$ *y*
2. *se verifica:*

   a) $\mu(d) = q_d$ *y* $\mu(e) \in D$

   o

   b) *existe* $E' \subseteq \mu(d) \cup \{e\}$ *tal que* $e \in E'$, $\mu_{(d,e)}^{E'}$ *es* q_E-*individualmente racional y* $\mu_{(d,e)}^{E'} R_U \mu$.

Siendo $\mu_{(d,e)}^{E'}(f) = \begin{cases} \mu(f) & \text{si} \quad f \notin \{d,e,\mu(e)\} \cup \mu(d) \\ Ch(E', P_f) & \text{si} \quad f = d \\ d & \text{si} \quad f = e \\ Ch(\mu(f) \backslash \{e\}, P_f) & \text{si} \quad f = \mu(e) \\ \emptyset & \text{en otro caso.} \end{cases}$

Se estudia la estabilidad en este modelo, bajo las condiciones de preferencias P_d responsive y preferencias R_U responsive. Bajo tales condiciones se asegura la existencia del conjunto $S_G(\bar{M}_U^{q_E})$:

Teorema 5 *Si $\bar{M}_U^{q_E}$ es un modelo de asignación varios a uno con restricción de capacidad, con preferencias R_U responsive, entonces $S_G(\bar{M}_U^{q_E}) \neq \emptyset$.*

En forma equivalente al uno-uno con restricción de cuota, se definen los conjuntos:

$$T_{q_E}(\bar{M}) = \bigcup_{(t_1,t_2) \in N} T_{q_E}(\bar{M}^{(t_1,t_2)}).$$

siendo

$$T_{q_E}(\bar{M}^{(t_1,t_2)}) = \begin{cases} S(\bar{M}^{(t_1,t_2)}) & \text{si } \#\mu = q_E, \text{ para } \mu \in S(\bar{M}^{(t_1,t_2)}) \\ \emptyset & \text{en otro caso.} \end{cases}$$

y

$$T_{<q_E}(\bar{M}) = \bigcup_{(t_1,t_2) \in N} T_{<q_E}(\bar{M}^{(t_1,t_2)}).$$

siendo

$T_{<q_E}(\bar{M}^{(t_1,t_2)}) = \{S(\bar{M}^{(t_1,t_2)}), \#_E\mu < q_E,\ e \notin Ch(\mu(d) \cup e, P_d) \quad \text{para todo}$

$(d,e) \in D \times E \backslash \mu(D^{t_1})\}.$

Utilizando técnicas de demostración análogas a las utilizadas en el modelo M_U^q se obtiene una caracterización del conjunto $S_G(\bar{M}_U^{q_E})$:

Teorema 6 *Si $\bar{M}_U^{q_E}$ es un modelo de asignación varios a uno con restricción de capacidad, con preferencias $R_U responsive, entonces S_G(\bar{M}_U^{q_E}) = T_{q_E}(\bar{M}) \cup T_{<q_E}(\bar{M})$*

Considerando el conjunto de matchings en $\bar{M}$, asociados a matchings de $S(M_U^q)$, $S^*(M_U^q) = \{\mu \in \bar{\mathcal{M}} : \mu' \in S(M_U^q)\}$ se relaciona los conjuntos $S_G(\bar{M}_U^{q_E})$ y $S_R(\bar{M}_U^{q_E})$, mediante el siguiente resultado:

Teorema 7 *Si $\bar{M}_U^{q_E}$ es un modelo de asignación varios a uno con restricción de capacidad, con preferencias $R_U responsive, entonces S_G(\bar{M}_U^{q_E}) \subseteq S^*(M_U^q) \subseteq S_R(\bar{M}_U^{q_E})$.*

Se presentan ejemplos donde $S_G(\bar{M}_U^{q_E}) \neq S^*(M_U^q)$ y $S^*(M_U^q) \neq S_R(\bar{M}_U^{q_E})$.

REFERENCIAS

[1] D. FEMENIA, M. MARÍ, A. NEME AND J. OVIEDO (2008). "Stable solutions on matchings models with quota restrictions ". (En prensa)

[2] D. GALE AND L. SHAPLEY (1962). "College admissions and the stability of marriage", *American Mathematical Monthly,* **69**, 9-15.

[3] D. GALE AND M. SOTOMAYOR (1985). "Some remarks on the stable matching problem", *American Mathematical Monthly,* **11**, 223-32.

[4] A. ROTH AND M. SOTOMAYOR (1990). *Two-sided Matching: A Study in Game-Theoretic Modeling and Analysis.* Cambridge University Press, Cambridge, England. [Econometrica Society Monographs No. 18].

ON THE RELATIONSHIP BETWEEN COMPLETENESS AND AWARENESS IN POSSIBILITY MODELS

Esteban J. Peralta*, Fernando A. Tohmé**

*Economics Department, Universidad de San Andrés, Vito Dumas 284, Buenos Aires, Argentina, peraltaej@hotmail.com
**Economics Department, Universidad Nacional del Sur, 12 de Octubre y San Juan, Buenos Aires, Argentina, ftohme@criba.edu.ar

Abstract: [1] introduced the notion of *possibility models* and showed that, under very weak conditions, a *complete* model does not exist; i.e., we can always find a possibility set that it is not "represented" within the model. Accordingly, this paper wants to go one step further and argue that, defined in a suitable way, the existence of some non-empty event about which agents are unaware of, is both necessary and sufficient for a possibility model to be incomplete.

Keywords: *incomplete possibility models, knowledge, language, unawareness*
2000 AMS Subjects Classification: 00A06 – 03E20

1 INTRODUCTION

One of the fundamental questions that we have to address as analysts concerns what agents know about the model itself. This is not only an interesting question from a philosophical viewpoint, but also a crucial one if we want to correctly define a model. After all, the analyst must define the "world" at which an agent's beliefs lie. Yet, a minute of reflection leads us to realize that a complete and exhaustive description of a world should include those beliefs as well. After all, a state of the world is such only if the agents "think" that it is a possible state of the world [7]. But then, a circular argument arises, since every world contains a description of the beliefs that constitutes it. Of course, Game Theory, being a representation of mutual externality situations, is not an exception since the beliefs of any agent about everything that is relevant for her must include a belief about each other's beliefs, a belief about each other belief about her beliefs, and so on. The literature has been looking for different ways of representing such sequences. This search began with the work of [4], where he introduced the concept of *type structure* and argued that it is a model that describes the hierarchies of beliefs. However, such a description must be "appropriate", not only in terms of tractability, but also in the sense that it must describe the same information than the process itself. This is the idea of *completeness* introduced by [1]. That is, a model is complete if every possible belief is "represented" and, therefore, there is no loss of generality when using it. Unfortunately, [1] and [2] show that it is not always true that every possible belief is actually "represented" in any model the analyst could use. Thus, there might be some "potential" beliefs that are not held by the agents. Since beliefs refer to worlds, there might be some worlds that are, in some sense, not *conceived* by the agents. The epistemic game literature has offered a specific knowledge-related meaning for the term *conception*; namely, *awareness*. In particular, it is said that an agent is aware of some event if she knows the event or she knows that she does not know it. Thus, *not to conceive* an event is interpreted as the complement of *being aware of it*. That is, not to know and not to know that it is not known.

Unawareness appears to be really important in several economic situations, especially through the players' recognition of the possibility of being unaware of *something* **[**5**].** This seems natural, since agents usually face unforeseen contingencies about their environment (think, for instance, in contract theory). In any case, what seems to be really underlying such phenomena is that as far as an agent might be unaware of something, there seems to be certain "limits" in the way she can "think". Accordingly, it seems very likely that *completeness* and *awareness* could be related in some way. After all, completeness seems to be related to some *lack of conception*. While completeness and unawareness have been extensively studied in the literature (see [3], [5] and [6], among others), to the best of our knowledge, no paper has addressed the question of their possible connection. Thus, *if*, *how* and *to what extent* they are related are the main concerns of the present paper. The rest of the paper is organized as follows. The next section deals with a brief description of the framework we will use throughout the paper; *possibility models*. Section 3 introduces the notion of completeness proposed by [1] and [2]. Accordingly, by introducing a formal definition of unawareness, section 4 shows that under certain non-trivial conditions, a possibility model is incomplete if and only if we find a set of states of the world about which the agents are unaware of.

2 POSSIBILITY MODELS

Possibility models, first introduced by [1], are a particular interactive model of the implicit kind. In some sense, they are the non-probabilistic analogues of *type models*. To see them formally, we follow the treatment given by [2]. For simplicity, we assume a two-player situation ($I = \{a, b\}$).

Definition 1: A *Possibility Model* is a structure $\mathrm{P} = \langle R^a, R^b, P^a, P^b \rangle$

R^a and R^b are non-empty (universe) sets and P^a and P^b are proper subsets *of* $R^a \times R^b$ and $R^b \times R^a$, respectively. Members of R^a (resp. R^b), denoted by x (resp. y), are called *states for* a (resp. b)[1]. Members of $R^a \times R^b$, denoted by $w = (x, y)$, are called *states of the world*. Thus, we will let $\Omega = R^a \times R^b$ be the set of states of the world and will call every $E \subseteq \Omega$ an event. Usually, events are the primitive object of uncertainty. P^a and P^b are called *possibility* relations and relate states of one agent with states of the other. Thus, for any x, $P^a(x, y)$ denote the *possibility set* of x (likewise for b). We suppose the relations are serial; i.e., $\forall x \ \mathrm{P}^a(x, y) \neq \varnothing$ and $\forall y \ \mathrm{P}^b(y, x) \neq \varnothing$ hold. Thus, every state always considers some states to be possible. We will say that x *knows* a set $Z \subseteq R^b$ if $\mathrm{Pr}oj_{R^b} P^a(x, y) = Z$; i.e., if the set of states that x considers possible coincide with Z [1]. Therefore, note that since x (resp. y) knows the set of states that x (resp. y) considers possible, every state for a (resp. b) knows a unique subset of R^b (resp. R^a). Moreover, while every x (resp. y) knows a non-empty subset of R^b (resp. R^a), some x (resp. y) knows a proper subset of R^b (resp. R^a). We suppose players "know their own state". That is, for any event $E \subseteq \Omega$ and state x for a, the set of states in E that a does not rule out is $\{(x', y') \in E : x' = x\}$. So, let $E_x = \{y' \in R^b : (x, y') \in E\}$. We will say, given a knowledge operator $K^a : 2^\Omega \to 2^\Omega$ for a, that a knows E in (x, y) if $\mathrm{Pr}oj_{R^b} P^a(x, y) = E_x$. Therefore, we let $K^a(E) = \{(x, y) : \mathrm{Pr}oj_{R^b} P^a(x, y) = E_x\}$ be the set of states of the world in which a knows E (Likewise for b). This is standard in the literature. Moreover, note that there are states of the world in which a does not know Ω. That is, $K^a(\Omega) \subset \Omega$. (likewise for b).

3 COMPLETE POSSIBILITY MODELS

Unfortunately, [1], [2] and [7] showed that, under very weak conditions, a complete possibility model does not exist. Among these conditions, the more important is perhaps what [1] and [2] called the *language* of the players. Briefly, the language can be interpreted as the "form" or "shape" of the elements that will be object of uncertainty. In some sense, it "tells" the players the "way" in which they can think about everything that is relevant for them; i.e., it is their "domain of thought". As [2] point out (page. 8):*"Given a belief model, the next step is to specify a language used by the players to think about beliefs. We'll then be able to talk about the completeness of a model, which is relative to a language"*. That is, we need to specify *how* the players think before we can say whether everything they can think of is present [B-98]. This is the sense in which completeness is relative to a language.

Definition 2: Let P be a possibility model and let L be a language for P. P is *Complete for* L if each non-empty set $E \in L^b$ is known by some $x \in R^a$ and each non-empty set $H \in L^a$ is known by some $y \in R^b$.

Given a set X, the power-set is denoted by $\Upsilon(X)$, and the cardinality by $|X|$. Unfortunately, [2] shows that, given the power-set language, a complete possibility model does not exist.

Proposition 1 ([2]): *No possibility model* P *is complete for a language* L *such that* $L^a = \Upsilon(R^a)$ *and* $L^b = \Upsilon(R^b)$.

4 UNAWARENESS AND COMPLETENESS

The standard approach to reasoning about knowledge implicitly assumes that agents are aware of all that is relevant for them. However, it is usual that in real life situations agents have to deal with unforeseen contingencies. Following [6], we will relate unawareness to knowledge. Formally, we let $U^a : 2^\Omega \to 2^\Omega$ be the unawareness operator defined by $U^a(E) \equiv \neg K^a(E) \bigcap \neg K^a \neg K^a(E)$ such that $U^a(E)$ represents the set of states of the world at which a is unaware of E. Let $\neg U^a(E) = A^a(E)$ for every $E \subseteq \Omega$ (likewise for b). As usual, the possibility model satisfies the following axioms (and likewise with a and b reversed):

$(KU)\ \forall E \subseteq \Omega, K^a U^a(E) = \varnothing$ (*KU Introspection*)

$(AU)\ \forall E \subseteq \Omega, U^a(E) \subseteq U^a U^a(E)$ (*AU Introspection*)

$(UK)\ \forall E \subseteq \Omega, U^a(E) \subseteq U^a K^b(E)$ (*UK Introspection*)

$(S)\ \forall E \subseteq \Omega, U^a(E) = U^a(\neg E)$ (*Symmetry*)

$(SR)\ \forall E \subseteq \Omega, A^a(E) = A^a K^a(E)$ (*Self-Reflection*)

These axioms are now standard in the literature. The next results show that incompleteness and unawareness are indeed related in a non-trivial way. To this end, we first show that if a possibility model is incomplete (for a given language), there must be events about which agents are unaware of. Then, we show that, under suitable conditions, the converse is also true.

Proposition 2: *Suppose* P *is an incomplete possibility model for some language* L *such that* $L^a \subseteq \Upsilon(R^a)$ *and* $L^b \subseteq \Upsilon(R^b)$. *Then, there are sets* $\varnothing \neq Z \subseteq \Omega$ *and* $\varnothing \neq H \subseteq \Omega$ *such that* $U^a(Z) \neq \varnothing$ *or* $U^b(H) \neq \varnothing$.

Proof: Since P is incomplete, we must have a language L with $L^a \subseteq \Upsilon(R^a)$ and $L^b \subseteq \Upsilon(R^b)$, such that there exits a set $\varnothing \neq A \in L^b \subseteq \Upsilon(R^b)$ that no x knows or a set $\varnothing \neq B \in L^a \subseteq \Upsilon(R^a)$ that no y knows. Suppose the former. Pick an arbitrary x such that $\{x\} \times A = Z$. By construction, $Z_x = A$ and $Z_{x'} = \varnothing$ for every $x' \neq x$. Thus, it follows that $\neg K^a(Z) = \Omega$. Since P^a is a proper subset of $R^a \times R^b$, there must be (at least) one state of the world w such that $w \in \Omega / K^a(\Omega)$. □

It is really worth to note that this claim does not restrict attention to any particular language. Thus, the result is very general in that it holds for every language defined as a non-empty subset of $\Upsilon(R^a)$ and $\Upsilon(R^b)$ relative to which a possibility model is incomplete[1]. Of course, from proposition 1 follow the following corollary:

Corollary 1: *Fix a possibility model* P *with language* L *such that* $L^a = \Upsilon(R^a)$ *and* $L^b = \Upsilon(R^b)$. *Then, there are sets* $\varnothing \neq Z \subseteq \Omega$ *and* $\varnothing \neq H \subseteq \Omega$ *such that* $U^a(Z) \neq \varnothing$ *or* $U^b(H) \neq \varnothing$.

Proof: By proposition 1, P is incomplete for L. Then, proposition **2** ensures that there must exist sets $\varnothing \neq Z \subseteq \Omega$ and $\varnothing \neq H \subseteq \Omega$ such that $U^a(Z) \neq \varnothing$ or $U^b(H) \neq \varnothing$. □

Proposition 3: *Fix a possibility model* P *that satisfies KU introspection and suppose there are sets* $\varnothing \neq Z \subseteq \Omega$ *and* $\varnothing \neq H \subseteq \Omega$ *such that* $U^a(Z) \neq \varnothing$ *or* $U^b(H) \neq \varnothing$. *Then, there must exist a language* L *with* $L^a \subseteq \Upsilon(R^a)$ *and* $L^b \subseteq \Upsilon(R^b)$ *such that* P *is incomplete for* L.

Proof: Suppose that $U^a(Z) \neq \varnothing$ for some set $\varnothing \neq Z \subseteq \Omega$. Let $U^a(Z) = A$. By KU introspection, $K^a A = \varnothing$. Then, for every $x \in R^a$, $\mathrm{Proj}_{R^b} P^a(x, y) \neq A_x$. Since $A \neq \varnothing$, $\varnothing \neq A_x$. Since $A_x \subseteq R^b$, there must be a language L such that $A_x \in L^b$ but no x knows A_x. The case $U^b(H) \neq \varnothing$ is treated similarly. □

ACKNOWLEDGMENT

We thank Martin Alfaro, Marcelo Auday, Gustavo Bodanza, Enrique Kawamura, Rodrigo Moro, Ana Reynoso, Ignacio Viglizzo and Federico Weinschelbaum for helpful discussions and Adam Brandenburger and Eleonora Cresto for their excellent suggestions.

REFERENCES

[1] A. Brandenburger, *On the existence of a "complete" possibility structure*, working paper 98-039, (1998), Harvard Business School.

[2] A. Brandenburger, and J. Keisler, *An impossibility theorem on beliefs in games*, Studia Logica, Vol. 84, No. 2, (2006), pp. 211-240.

[3] E. Dekel, B. Lipman, and A. Rustichini, *Standard State-Space Models Preclude Unawareness*, Econometrica, Vol. 66, No. 1, (1998), pp. 159-173.

[4] J. Harsanyi, *Games of Incomplete Information Played by 'Bayesian' Players, I-III,"*, Management Science, Vol. 14, (1967/68), pp. 159-182, 320-334, 486-502.

[5] J. Li, *Information Structures with Unawareness*, working paper, (2006), University of Pennsylvania.

[6] S. Modica, and A. Rustichini, *Awareness and partitional information structures*, Theory and Decision 37 (1994), pp. 107-124.

[7] F. Tohmé, *Existence and definability of states of the world*, Mathematical Social Sciences, Vol. 49 (2005), pp. 81-100.

SOME EXACT SOLUTIONS THROUGH SYMMETRY ANALYSIS FOR THE VAKNENKO EQUATIONS

M.L. Gandarias[♭] and M. S. Bruzon[†]

[♭,†]*Departamento de Matemáticas, Universidad de Cádiz, PO.BOX 40, 11510 Puerto Real, Cádiz, Spain, marialuz.gandarias@uca.es, matematicas.casem@uca.es*

Abstract: In this paper we apply the classical and the nonclassical method to the integrable Vaknenko equation. The nonclassical method applied to the associated potential system yields solutions which are neither solutions arising from nonclassical symmetries of the Vaknenko equation nor solutions arising from potential symmetries. Some of these solutions have an interesting behaviour such as "nonlinear superposition".

Keywords: *symmetries, solutions, solitons*
2000 AMS Subject Classification: 35D05-70G65

1 INTRODUCTION

The Ostrovsky equation

$$(u_t + (u^2)_x - \beta u_{xxx})_x = \gamma u \tag{1}$$

was proposed by Ostrovsky to describe long internal waves in a rotating ocean. Here $x \in \mathcal{R}$ and $\gamma > 0$, where β and γ are dispersion coefficients. For long waves, for wich high-frecuency dispersion is negligible, $\beta = 0$ (1) becomes the so called ROE. This equation has been considered by many authors (see [7] and references therein). Vakhnenko and Parkes [9] proved that the ROE can be transformed to the new integrable equation

$$uu_{xxt} - u_x u_{xt} + u^2 u_t = 0 \tag{2}$$

which is known as Vaknenko equation (VE). In a recent paper [8] the $tanh$ and $sine$-$cosine$ methods were used to construct exact periodic and soliton solutions of nonlinear evolution equations such as (2).

There have been several generalizations of the classical Lie group method for symmetry reductions. Bluman and Cole developed the nonclassical method to study the symmetry reductions of the heat equation. The basic idea of the method is that the partial differential equation (PDE) is augmented with the invariance surface condition

$$\Phi \equiv \xi u_x + \tau u_t - \phi = 0, \tag{3}$$

which is associated with the vector field

$$\mathbf{v} = \xi \partial_x + \tau \partial_t + \phi \partial_u. \tag{4}$$

By requiring that both (2) and (3) are invariant under the transformation with infinitesimal generator (4) one obtains an overdetermined nonlinear system of equations for the infinitesimals $\xi(x,t,u)$, $\tau(x,t,u)$ and $\phi(x,t,u)$.

In [2] Bluman introduced a method to find a new class of symmetries for a PDE. Suppose a given scalar PDE of second order

$$F(x,t,u,u_x,u_t,u_{xx},u_{xt},u_{tt}) = 0, \tag{5}$$

where the subscripts denote the partial derivatives of u, can be written as a conservation law

$$\frac{D}{Dt} f(x,t,u,u_x,u_t) - \frac{D}{Dx} g(x,t,u,u_x,u_t) = 0, \tag{6}$$

for some functions f and g of the indicated arguments. Here $\frac{D}{Dx}$ and $\frac{D}{Dt}$ are total derivative operators. Through the conservation law (6) one can introduce an auxiliary potential variable v and form an auxiliary potential system

$$\begin{aligned} v_x &= f(x,t,u,u_x,u_t), \\ v_t &= g(x,t,u,u_x,u_t). \end{aligned} \tag{7}$$

Any Lie group of point transformations

$$\mathbf{w} = \xi(x,t,u,v)\partial_x + \tau(x,t,u,v)\partial_t + \phi(x,t,u,v)\partial_u + \psi(x,t,u,v)\partial_v, \tag{8}$$

admitted by (7) yields a nonlocal symmetry *potential symmetry* of the given PDE (6) if and only if the following condition is satisfied

$$\xi_v^2 + \tau_v^2 + \phi_v^2 \neq 0. \tag{9}$$

In [3] the nonclassical method is applied to the associated potential system (7). Any Lie group of point transformations (8) admitted by (7) yields a nonlocal symmetry *potential symmetry* of the given PDE (6) if the condition (9) is satisfied. In [4] a modification for this method was proposed. By using this method some nonclassical potential symmetries for the Burgers equation were derived in [5]. In this work we apply the classical and the nonclassical method to the VE as well as to the natural associated potential system. The nonclassical symmetry reductions obtained for

$$\begin{aligned} v_x &= u \\ v_t &= -\frac{u_{tx}}{u} \end{aligned} \tag{10}$$

generate a wide variety of interesting analytical solutions for (2).

2 Classical symmetries

The classical Lie method applied to (2) yields the following generators:

$$\mathbf{v}_1 = \frac{\partial}{\partial x}, \quad \mathbf{v}_2 = \alpha(t)\frac{\partial}{\partial t}, \quad \mathbf{v}_3 = x\frac{\partial}{\partial x} - 2u\frac{\partial}{\partial u},$$

here $\alpha(t)$ is an arbitrary function Solving (3) we obtain two canonical symmetry reductions:

Reduction 1 By using the generator $\mathbf{v}_2 + \mathbf{v}_3$ we can set $\alpha(t) = \beta(t)/\beta'(t)$. Hence we obtain the symmetry reduction

$$z = \frac{x}{\beta(t)}, \qquad u = \frac{h(z)}{\beta^2(t)}, \tag{11}$$

where $h(z)$ satisfies

$$-zhh''' + zh'h'' - zh^2h' - 4hh'' + 3(h')^2 - 2h^3 = 0. \tag{12}$$

Reduction 2 By using the generator $\mathbf{v}_1 + \mathbf{v}_2$ we obtain the similarity variables and similarity solution

$$z = x - \beta(t), \qquad u = h(z), \tag{13}$$

where $h(z)$ satisfies

$$hh''' - h'h'' + h^2h' = 0. \tag{14}$$

We observe that dividing by h^2 and integrating once with respect to z equation (14) becomes

$$h'' + h^2 = kh. \tag{15}$$

Equation (15) admits solutions in terms of the Jacobi elliptic functions.

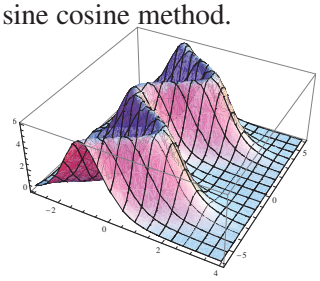
From solutions of (14) and using (13) we obtain the following solutions for (2) in terms of the Jacobi elliptic function sn:

$$\begin{aligned} u &= \mathrm{a}_2\,\mathrm{sn}^{\,2}\left(k(x-\beta(t)),p\right) + \frac{\mathrm{b}_2}{\mathrm{sn}^{\,2}\left(k\,(x-\beta(t)),p\right)} + \mathrm{a}_0, \\ u &= \frac{\mathrm{b}_2}{\mathrm{sn}^{\,2}\left(k\,(x-\beta(t)),p\right)} + \mathrm{a}_0, \\ u &= \mathrm{a}_2\,\mathrm{sn}^{\,2}\left(k\,(x-\beta(t)),p\right) + \mathrm{a}_0, \end{aligned} \tag{16}$$

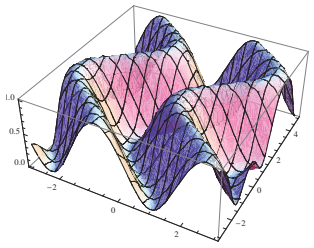
and the corresponding solutions in terms of the degenerate trigonometric and hyperbolic functions

$$\begin{aligned} u &= -6k^2\text{sech}^2(k(x-\beta(t))\text{cosech}^2(k(x-\beta(t)), \\ u &= 4k^2 - 6k^2 cosec^2(k(x-\beta(t)), \\ u &= 6k^2\text{sech}^2(k(x-\beta(t)). \end{aligned} \tag{17}$$

Some particular cases of these solutions, with $\beta(t) = \lambda t$, were derived in [8] by using the tanh method and the sine cosine method.



$u = 6\text{Sech}^2(x + \cos(t))$



$u = Sn(x + \cos(t), 1/4)^2.$

3 NONCLASSICAL SYMMETRIES FOR THE VE

To obtain nonclassical symmetries of (2), we apply the algorithm described in [7] for calculating the determining equations. We can establish that for $\tau \neq 0$ the nonclassical method applied to (2) gives rise to a set of twenty determining equations for the infinitesimals. From these equations we obtain

$$\xi = \xi(x,t), \quad \phi = \alpha(x,t)u + \beta(x,t,)u^{1/3}.$$

We can distinguish the following cases:

Case 1. $\xi = 0$. In this case we obtain

$$\phi = \alpha(x,t)u + \beta(x,t)u^{1/3}$$

with $\alpha(x,t)$ $\beta(x,t)$ arbitrary functions. These generators are too general to be practical.

Case 2. $\xi \neq 0$. in this case $\beta = 0$, $\alpha = -2\xi_x$ where ξ must satisfy owing subcases:

Subcase i. $\xi_x \neq 0$ $f_1 \neq 0$ then

$$\xi = f_1(t)(x + k_1) \quad \tau = 1, \quad \phi = -2f_1(t)u$$

Subcase ii. $\xi_x = 0$

$$\xi = f_2(t) \quad \tau = 1, \quad \phi = 0.$$

In both cases that the infinitesimals are equivalent to the infinitesimals obtained using the classical method. The nonclassical method applied to (10) gives rise to

$$\begin{aligned} \phi &= -u^2\,\xi_v + (\psi_v - \xi_x)\,u + \psi_x, \\ \xi &= \alpha(x,t,v)e^{\beta(x,t,v)}. \end{aligned}$$

We can distinguish the following cases:

Case 1. $\xi_u \neq 0$ or $\xi_x \neq 0$. The infinitesimals can be obtained using the classical method.

Case 2.$\xi_u = 0$, $\xi_x = 0$, $\xi = \xi(t)$, $\psi = \psi(x,t)$ where $\xi(t)$ and $\psi(x,t)$ must satisfy the following equations:

$$\begin{aligned} &\xi_t\psi + \xi^2\psi_x - \xi\psi_t = 0, \\ &\xi_t\psi_{xx} - \xi\psi\psi_x - \xi\psi_{txx} = 0. \end{aligned}$$

Solving these equations as well as the characteristic equation yields

$$\xi = \delta'(t), \quad \psi = k\delta'(t)F(w),$$

where $w = x + \delta(t)$, $\delta(t)$ is an arbitrary differentiable function and $F(w)$ satisfies

$$F_{ww} + \frac{k}{2}F^2 = k_1,$$

with k_1 an arbitrary constant. Solving the characteristic equation we obtain the nonclassical reduction

$$v(x,t) = f(w) + h(z), \quad w = x + \delta(t), \quad z = x - \delta(t), \tag{18}$$

where $\delta(t)$ is an arbitrary function and $F(w) = f_w$ and $H(z) = h_z$ satisfy

$$F_{ww} + F^2 = \mu_1, \quad H_{zz} + H^2 = \mu_1, \tag{19}$$

respectively, where μ_1 is an arbitrary constant. This equation is equivalent to the Weierstrass elliptic function equation. Setting $\mu_1 = 1$ the solution given by Mathematica is

$$f_w = 6^{1/3} WeierstrassP[(w + C[1])/6^{1/3}, \{0, C[2]\}].$$

Integrating once with respect to w

$$f = -6^{2/3} WeierstrassZeta[(w + C[1])/6^{1/3}, \{0, C[2]\}].$$

The solution $v = f(x+t) + h(x-t)$ and $u = f'(x+t) + h'(x-t)$ exhibit a "nonlinear superposition". We remark that this "decoupling" of the nonclassical symmetry reduction solution into a function of $x + \delta(t)$ and a function of $x - \delta(t)$ is unusual.

4 Conclusions

In this paper we have applied Lie classical method and the the nonclassical method to the integrable VE which has been proved [9] equivalent to the ROE. For the VE the nonclassical method applied to the associated potential system yields solutions which are neither solutions arising from nonclassical symmetries of (2) nor solutions arising from potential symmetries. Some particular cases of the solutions we obtained by Lie classical method, were derived in [8] by using the tanh method and the sine cosine method. Some of the solutions arising from the nonclassical reductions of the potential system (10) present a "nonlinear superposition" and we remark that this "decoupling" of the nonclassical symmetry reduction solution into a function of $x + \delta(t)$ and a function of $x - \delta(t)$ is unusual.

Acknowledgements: The authors acknowledge the financial support from Junta de Andalucía group FQM–201 and from project MTM2009-11875.

References

[1] G.W. BLUMAN, and S. KUMEI, *Symmetries and differential equations,* Springer-Verlag, 1989.

[2] G.W. BLUMAN, G.J. REID, AND S.KUMEI, *New classes of symmetries for partial differential equations.* J. Math. Phys., 29 (1988), pp.806-811.

[3] G.W. BLUMAN, Z. YAN, *Nonclassical potential solutions of partial differential equations*. Eur. J. Appl. Math. 16 (2005), pp.239-261.

[4] M.L. GANDARIAS *New Potential symmetries*. CRM Proceedings and Lecture Notes ed. American Math. Society. Publ., Providence RI. 25 (2000), pp.161-165.

[5] M.L. GANDARIAS AND M.S. BRUZON *Noclassical potential symmetries for the Burgers equation.* Nonlinear Analysis. 71 (2009), pp.1826-1834.

[6] L.A. OSTROVSKY, *Nonlinear internal waves in a rotating ocean.* Okeanologiya. 18 (1978), pp.181-191.

[7] Y.A. STEPANYANTS, *On stationary solutions of the reduced Ostrovsky equation: Periodic waves, compactons and compound solitons.* Chaos Solitons and Fractals. 28 (2006), pp.193-204.

[8] E.YUSUFOGLU AND A. BEKIR, *The tanh and the sinecosine methods for exact solutions of the MBBM and the Vakhnenko equations.* Chaos Solitons and Fractals. 32 (2008), pp.1126-1133.

[9] V.O. VAKHNENKO AND E.J. PARKES, *The two loop soliton solution of the Vakhnenko equation.* Nonlinearity. 11 (1998), pp.1457-1464.

AN INITIAL-BOUNDARY VALUE PROBLEM FOR THE ONE-DIMENSIONAL NON-CLASSICAL HEAT EQUATION IN A SLAB

Natalia N. SALVA ([ab]) - Domingo A. TARZIA ([ac]) - Luis T. VILLA ([ad])

([a]) CONICET, Argentina.
([b]) TEMADI, Centro Atómico Bariloche, Av. Bustillo 9500, 8400 Bariloche, Argentina, natalia@cab.cnea.gov.ar.
([c]) Depto. de Matemática, Universidad Austral, Paraguay 1950, S2000FZF Rosario, Argentina, DTarzia@austral.edu.ar.
([d]) Facultad de Ingeniería, Universidad Nacional de Salta, Buenos Aires 144, 4400 Salta, Argentina, villal@unsa.edu.ar.

Abstract: A nonlinear problem for the one-dimensional heat equation in a bounded and homogeneous medium with temperature data on the boundary $x=0$ and $x=1$ is studied. It is considered a non- classical heat conduction problem because a uniform spatial heat source depending on the heat flux (or the temperature) on the boundary $x=0$ is taken into account. Existence and uniqueness for the solution are proved under suitable assumptions on the data. Comparisons results and asymptotic behavior for the solution regarding some particular cases for the heat source, initial, and boundary data are also obtained.

Keywords: Non-classical heat equation, Volterra integral equations, Uniform heat source.
2000 AMS Subject Classification: 35C15, 35K55, 45D05, 80A20.

1. INTRODUCTION

In this paper, we will consider initial and boundary value problems (IBVP), for the one-dimensional non-classical heat equation motivated by some phenomena regarding the design of thermal regulation devices that provides a heater or cooler effect [1, 5, 7, 6, 8, 9]. We first study a IBVP with Dirichlet boundary conditions and a heat source that depends on the heat flux at the fixed face $x=0$, and afterwards, we study a similar problem but with Neumann boundary conditions and a heat source that depends on the temperature on the fixed face $x=0$. We obtain in both cases existence of a solution through a system of second kind Volterra integral equations.

A heat conduction problem of the first type but for a semi-infinite material was analyzed in [8,9], where results on existence, uniqueness and asymptotic behavior for the solution were obtained. In other respects, a class of heat conduction problems characterized by a uniform heat source given as a multivalued function from $\mathbb{R}$ into itself was studied in [7] with results regarding existence, uniqueness and asymptotic behavior for the solution. Other references on the subject are [5,6]. Recently, free boundary problems (Stefan problems) for the non-classical heat equation have been given in [2-4], where some explicit solutions are also given.

2. PROBLEM (P1) - EXISTENCE AND UNIQUENESS

We study the following IBVP for the heat equation in the slab [0,1] (Problem (P1)):

$$(P1)\begin{cases} u_t - u_{xx} = -F(u_x(0,t),t), \quad (x,t) \in \Omega \equiv \{(x,t): 0<x<1, \ 0<t<T\} & (2.1)\\ u(0,t) = f(t), \ 0< t <T & (2.2)\\ u(1,t) = g(t), \ 0< t <T & (2.3)\\ u(x,0) = h(x), \ 0 \le x \le 1 & (2.4)\end{cases}$$

where the unknown function $u = u(x,t)$ denotes the temperature profile for an homogeneous medium occupying the spatial region $0<x<1$, the boundary data f and g are real functions defined on $\mathbb{R}^+$, the initial temperature $h(x)$ is a real function defined on [0,1], and F is a given function of two real variables, which is related to the evolution of the heat flux $u_x(0,t)$ in this case.

For data $h=h(x)$, $g = g(t)$, $f = f(t)$ and F in problem (1.1)-(1.4) we shall consider the following assumptions:

(HA) g and f are continuously differentiable functions on $\mathbb{R}^+$;

(HB) h is a function $C^1[0,1]$, which verifies compatibility conditions: $h(0) = f(0)$, $h(1) = g(1)$;

(HC) The function $F = F(V,t)$ verifies:

(HC1) It is defined and continuous on the region $D = \mathbb{R}\times[0,T]$;

(HC2) For each $M>0$ and for $|V| \le M$, it is uniformly Hölder continuous in variable t for each compact subset of $(0,T]$;

(HC3) For each bounded set B of D, there exists a bounded positive function $L_o = L_o(t)$, defined for $0< t \le T$, such that

$$\left|F(V_2,t)-F(V_1,t)\right| \le L_O(t)\left|V_2-V_1\right|, \forall (V_2,t),(V_1,t) \in B \quad ;$$

(HC4) It is bounded for bounded V for all $t \ge 0$;

(HD) $F(0,t)=0$, $0< t \le T$.

(HE) $V\,F(V,t) > 0 \quad , \quad \forall\, V \neq 0, \quad \forall\, t > 0$;

(HF) $f(t) \equiv 0 \quad \forall t > 0 \quad , \quad g(t) \equiv u_{1_o} > 0 \quad \forall t > 0 \quad , \quad h'(x) > 0 \ \forall\, x \in [0,1] \quad , \quad h(1) \le u_{1_o}$.

(HG) $f(t) \equiv 0 \quad \forall t > 0 \quad , \quad g(t) \equiv 0 \quad \forall t > 0 \quad , \quad h(x) > 0 \ \forall\, x \in [0,1]$

Theorem 1. *Under the assumptions (HA) to (HD), the solution u to the problem (P1) has the expression*

$$\begin{aligned} u(x,t) = {} & \int_0^1 \left[\theta(x-\xi,t)-\theta(x+\xi,t)\right]h(\xi)d\xi - 2\int_0^t \theta_x(x,t-\tau)f(\tau)d\tau + 2\int_0^t \theta_x(x-1,t-\tau)g(\tau)d\tau \\ & -\int_0^t \left\{\int_0^1 \left[\theta(x-\xi,t-\tau)-\theta(x+\xi,t-\tau)\right]d\xi\right\} F(V(\tau),\tau)d\tau \end{aligned} \qquad (2.5)$$

where $V=V(t)$, defined by $V(t) = u_x(0,t)$ for all $t>0$, must satisfy the following second kind Volterra integral equation:

$$V(t) = 2\int_0^1 \theta(\xi,t)h'(\xi)d\xi - 2\int_0^t \theta(0,t-\tau)\,\dot{f}(\tau)d\tau + 2\int_0^t \theta(-1,t-\tau)\,\dot{g}(\tau)d\tau - \int_0^t \overline{K}(t-\tau)F(V(\tau),\tau)d\tau \qquad (2.6)$$

and the functions $\theta = \theta(x,t)$, $K = K(x,t)$ and $\overline{K} = \overline{K}(t)$ are defined in the following way:

$$\theta(x,t) = K(x,t) + \sum_{j=1}^{\infty}\left[K(x+2j,t)+K(x-2j,t)\right],\ K(x,t) = \frac{1}{2\sqrt{\pi t}}e^{-\frac{x^2}{4t}},\ \overline{K}(t) = 2\left(\theta(0,t)-\theta(1,t)\right),\quad t>0. \qquad (2.7)$$

Theorem 2 *Under assumptions (HA), (HB), (HC1) and (HC4), there exists at least one solution $V(t) \in C^o(\mathbb{R}^+)$ to the integral equation (2.6), therefore we have at least one solution for the IBVP (2.1)-(2.4).*

Theorem 3 *Under the assumptions (HA) to (HD), there exists a unique solution to the problem (P1). Moreover, there exists a maximal time $\beta > 0$, such that the unique solution to (2.1) – (2.4) can be extended to the interval $0 \le t \le \beta$.*

3. PROPERTIES OF THE SOLUTION TO PROBLEM (P1)

Theorem 4 *Under assumptions (HA) to (HD), the solution u to problem (P1) is bounded in terms of the initial and boundary data h, f and g.*

Lemma 5 *Let* $u_0(x,t)$ *be the solution to (2.1)-(2.4) with null heat source (i.e.* $F \equiv 0$*). Under the assumptions (HD), (HE) and (HF), we have that:*

a) $0 \le u(x,t) \le u_o(x,t), \quad \forall\, x \in [0,1], \forall t > 0.$ *b)* $\lim_{t \to +\infty} u(x,t) = 0, \forall x \in [0,1].$

Now we will consider the continuous dependence of the functions *V=V(t)* and *u=u(x,t)* given by (2.5) and (2.6) respectively upon the data *f, g , h* and *F*. Let us denote by $V_i=V_i(t)$ (*i=1,2*) the solution to (2.6) and $u_i=u_i(x,t)$ given by (2.5) respectively for the data f_i, g_i, h_i and F (*i=1,2*) in problem (P1).

Theorem 6 *Considering problem (P1) under the assumptions (HA) to (HD), we obtain that* $V_2 - V_1$ *is bounded in terms of the initial and boundary data h, f and g. Therefore, the difference of the solutions* $u_2 - u_1$ *is also bounded in terms of the initial and boundary data h, f and g.*

Theorem 7 *Let* $u_i=u_i(x,t)$, $V_i=V_i(t)$ *(i=1,2) be the functions given by (2.5) and (2.6) for the data f, g, h and* F_i *(i=1,2) in problem (P1). Under the assumptions (HA) to (HD), we obtain the following estimation:*

$$\left|u_2(x,t) - u_1(x,t)\right| \le M_o \left\|F_2 - F_1\right\|_{t,M} \left[t + \frac{2\left\|L_2\right\|_\infty}{\sqrt{\pi}} \sqrt{t}\; e^{\left\|L_2\right\|_\infty \frac{2\sqrt{t}}{\sqrt{\pi}}} \right] \qquad (3.1)$$

where $\left\|F_1 - F_2\right\|_{t,M} = \sup_{\substack{\|z\|_\infty \le M \\ 0<\tau\le t}} \left|F_1(z(\tau),\tau) - F_2(z(\tau),\tau)\right|$, *and* M_o *is a positive constant which verifies the inequality*

$$\int_o^1 \left|\theta(x-\xi,t-\tau) - \theta(x+\xi,t-\tau)\right| d\xi \le M_o,\ 0 < \tau < t \le T, \quad 0 \le x \le 1. \qquad (3.2)$$

4. PROBLEM (P2) - EXISTENCE AND UNIQUENESS

Now, we will consider a new non-classical initial-boundary value problem (P2) for the heat equation in the slab [0,1], which is related to the previous problem (P1), given by:

$$(P2)\begin{cases} u_t - u_{xx} = -F(u(0,t),t), & (x,t) \in \Omega \equiv \{(x,t):\ 0<x<1,\ 0<t\le T\} & (4.1)\\ u_x(0,t) = f(t), & 0< t \le T & (4.2)\\ u_x(1,t) = g(t), & 0< t \le T & (4.3)\\ u(x,0) = h(x), & 0 \le x \le 1. & (4.4) \end{cases}$$

Theorem 8 *Under the assumptions (HA) to (HD), the solution u to the problem (P2) has the expression*

$$\begin{aligned} u(x,t) = {} & \int_0^1 \left[\theta(x-\xi,t) + \theta(x+\xi,t)\right] h(\xi)\, d\xi - 2\int_0^t \theta(x,t-\tau) f(\tau)\, d\tau + 2\int_0^t \theta(x-1,t-\tau) g(\tau)\, d\tau \\ & - \int_0^t \left\{ \int_0^1 \left[\theta(x-\xi,t-\tau) + \theta(x+\xi,t-\tau)\right] d\xi \right\} F(V(\tau),\tau)\, d\tau \end{aligned} \qquad (4.5)$$

where V=V(t), defined by V(t)=u(0,t), must satisfy the following second kind Volterra integral equation:

$$V(t) = 2\int_0^1 \theta(\xi,t)h(\xi)d\xi - 2\int_0^t \theta(0,t-\tau)f(\tau)d\tau + 2\int_0^t \theta(-1,t-\tau)g(\tau)d\tau - 2\int_0^t\int_0^1 \theta(\xi,t-\tau)d\xi F(V(\tau),\tau)d\tau. \qquad (4.6)$$

Theorem 9 *Under the assumptions (HA) to (HD), there exists a unique solution to the problem (P2). Moreover, there exists a maximal time $\beta > T > 0$, such that the unique solution to (4.1) – (4.4) can be extended to the interval $0 \le t \le \beta$.*

5. PROPERTIES OF THE SOLUTION TO PROBLEM (P2)

Theorem 10 *Under the assumptions (HA) to (HD), the solution u to problem (P2) is bounded in terms of the initial and boundary data h, f and g.*

Theorem 11 *Let us define V_i and u_i (i=1,2) as in Theorem 6, with respect to the problem (P2). Under the assumptions (HA) to (HD), we obtain that $V_2 - V_1$ is bounded in terms of the initial and boundary data h, f and g. Therefore, the difference of the solutions $u_2 - u_1$ is also bounded in terms of the initial and boundary data h, f and g.*

Theorem 12 *Let us define V_i and u_i (i=1,2) as in Theorem 7, with respect to the problem (P2).Under the assumptions (HA) to (HD), then we obtain the following estimation:*

$$|u_2(x,t) - u_1(x,t)| \le M_1 \|F_2 - F_1\|_{t,M}\, t\left[1 + \|L_2\|_t\, C_3\, \exp(C_3 \|L_2\|_t)\right]. \qquad (4.7)$$

where M_1 is a positive constant which verifies the inequality

$$\int_o^1 |\theta(x-\xi,t-\tau) + \theta(x+\xi,t-\tau)|d\xi \le M_1,\ 0 < \tau < t \le T,\ 0 \le x \le 1. \qquad (4.8)$$

Theorem 13 *Under the hypotheses (HG) and (HE), we have that $0 < u(x,t) < \|h\|_\infty$, $\forall x \in [0,1]$, $\forall\, t \ge 0$.*

ACKNOWLEDGEMENTS

This paper was partially sponsored by the project PIP No. 0460 of CONICET - UA (Rosario, Argentina), and Grant FA9550-10-1-0023.

REFERENCES

[1] L.R. BERRONE , D.A. TARZIA, L.T.VILLA, *Asymptotic behavior of a Non –classical Heat Conduction Problem for a Semi-infinite Material*, Mathematical Methods in the Applied Sciences, 23 (2000), pp. 1161-1177.

[2] A.C. BRIOZZO, D.A. TARZIA, *Existence and uniqueness of a one-phase Stefan problem for a non-classical heat equation with temperature boundary condition at the fixed face*, Electronic Journal of Differential Equations, 2006 (2006) No. 21, pp1-16.

[3] A.C. BRIOZZO, D.A. TARZIA, *A one-phase Stefan problem for a non-classical heat equation with a heat flux condition on the fixed face*, Applied Mathematics and Computation, 182 (2006), pp. 809-819.

[4] A.C. BRIOZZO, D.A. TARZIA, *Exact solutions for non-classical Stefan problems*, International Journal of Differential Equations, 2010 (2010), Article ID 868059, pp. 1-19.

[5] K. GLASHOFF, J. SPREKELS, *The regulation of temperature by thermostats and set-valued integral equations*, J. Integral Eq., 4 (1982), pp. 95-112.

[6] N. KENMOCHI, *Heat conduction with a class of automatic heat source controls*, Pitman Research Notes in Mathematics Series #186 (1990), pp. 471-474.

[7] N. KENMOCHI, M. PRIMICERIO, *One-dimensional heat conduction with a class of automatic heat source controls*, IMA J. Appl. Math., 40 (1998), pp. 205-216.

[8] D.A. TARZIA, L.T VILLA, *Some nonlinear heat conduction problems for a semi-infinite strip with a non-uniform heat source*, Rev. Un. Mat. Argentina, 41 (1998), pp. 99-114.

[9] L.T. VILLA, *Problemas de control para una ecuación unidimensional del calor*, Rev. Un. Mat. Argentina, 32 (1986), pp. 163-169.

SOLUCIONES EXACTAS PARA UNA ECUACIÓN MODIFICADA DE BENNEY-LIN

M. S. Bruzón[♭] y M.L. Gandarias[†]

[♭,†]*Departamento de Matemáticas, Universidad de Cádiz, Spain, matematicas.casem@uca.es, marialuz.gandarias@uca.es, www.uca.es*

Resumen: En este trabajo realizamos una clasificación de las simetrías de una ecuación modificada de Benney-Lin utilizando el método clásico de Lie. Determinamos todas las ecuaciones reducidas y obtenemos soluciones exactas tipo ondas viajeras.

Palabras clave: *simetrías, solución, kink*
2000 AMS Subject Classification: 35D05 - 70G65

1. INTRODUCCIÓN

Muchos fenómenos de las ciencias e ingenierías son descritos por ecuaciones en derivadas parciales (EDPs) y no existe un método general para resolver este tipo de ecuaciones. Uno de los métodos más eficiente para obtener soluciones exactas de EDPs es el método de las transformaciones puntuales o método clásico de Lie [1, 8]. La idea básica del método aplicado a una EDP con dos variables independientes consiste en buscar una transformación de forma que reduzca la EDP a ecuaciones diferenciales ordinarias (EDOs). Obtener soluciones de estas ecuaciones y, deshaciendo la transformación, obtener soluciones de la ecuación original.

En los últimos años se ha avanzado mucho en el desarrollo de métodos y sus aplicaciones para encontrar soluciones exactas de ecuaciones diferenciales ordinarias no lineales [2, 3, 7, 9].

En este trabajo, consideramos una ecuación modificada de Benney-Lin,

$$u_t + uu_x + \psi u_{xxx} + \beta u_{xx} + \alpha u_{xxxx} + \eta u_{xxxxx} = 0, \tag{1}$$

donde $\beta > 0$ y $\psi, \alpha, \eta \in \mathbb{R}$. La ecuación (1) describe, en problemas de dinámica de fluidos, la evolución de ondas largas. Cuando $\alpha, \beta = 0$ se reduce a la ecuación de Kamara (o ecuación de Korteweg-de Vries de quinto orden) que modeliza la evolución de las ondas sometidas a tensión superficial. Cuando $\eta = 0$ es una ecuación generalizada de Kuramoto-Sivashinsky [5].

Nosotros aplicamos el método clásico de Lie a la ecuación (1) y obtenemos un sistema de ecuaciones lineales para los infinitesimales, donde las soluciones del sistema dependen de η. Una vez determinadas los infinitesimales, estudiamos los sistemas óptimos unidimensionales de subálgebras. A partir de las soluciones de similaridad, reducimos la ecuación (1) a ecuaciones diferenciales ordinarias. Estudiamos ciertas soluciones que tienen semejanzas con estructuras coherentes.

2. SIMETRÍAS CLÁSICAS. REDUCCIONES

Aplicamos el método clásico de Lie a la Ec. (1) e imponemos que la Ec. (1) sea invariante bajo un grupo uniparamétrico de transformaciones con generador infinitesimal,

$$v = \xi(x,t,u)\frac{\partial}{\partial x} + \tau(x,t,u)\frac{\partial}{\partial t} + \phi(x,t,u)\frac{\partial}{\partial u}. \tag{2}$$

El criterio de invarianza implica que la prolongación quinta del campo vectorial v actuando sobre la Ec. (1) debe ser cero en los puntos donde dicha ecuación se satisfaga, es decir,

$$pr^5(v)(\Delta) = 0,$$

donde ([8], Teor. 2.36)

$$pr^5(v) = v + \sum_J \phi^J(x,t,u^{(5)})\frac{\partial}{\partial u_J}$$

siendo

$$\phi^J(x,t,u^{(5)}) = D_J(\phi - \xi u_x - \tau u_t) + \xi u_{Jx} + \tau u_{Jt},$$

con $J = (j_1, \ldots, j_k)$, $1 \leq j_k \leq 2$ y $1 \leq k \leq 5$. Obtenemos un sistema de 45 ecuaciones determinantes. A continuación resolvemos el sistema dependiendo del valor de la constante η:
Si $\eta \neq 0$

$$\xi = \frac{1}{5}\tau_t x + \alpha(t), \qquad \tau = \tau(t), \qquad \phi = \left(\delta(t) - \frac{\alpha}{25\eta}\tau_t x\right) u + \omega(x,t), \tag{3}$$

donde τ, α, δ y ω están relacionadas por las siguientes condiciones,

$$\begin{aligned}
\tfrac{5\eta\psi - 2\alpha^2}{25\,\eta}\,\tau_t &= 0,\\
\tfrac{\alpha\psi - 5\beta\eta}{25\,\eta}\,\tau_t &= 0,\\
(\alpha\,\tau_t\,u + 5\,\eta\,\tau_{tt})\;x + (-25\,\eta\,\delta - 20\,\eta\,\tau_t)\;u + 25\,\eta\,\alpha_t - 25\,\eta\,\omega + 2\,\beta\alpha\,\tau_t &= 0,\\
\alpha\,\tau_{tt}\,u\,x + \alpha\,\tau_t\,u^2 + (-25\,\eta\,\delta_t - 25\,\eta\,\omega_x)\;u - 25\,\eta^2\,\omega_{xxxxx} - 25\,\alpha\,\eta\,\omega_{xxxx}&\\
-25\,\eta\,\omega_{xxx} - 25\,\beta\,\eta\,\omega_{xx} - 25\,\eta\,\omega_t &= 0.
\end{aligned} \tag{4}$$

Del sistema (4) deducimos que

$$\xi = k_1 t + k_2, \qquad \tau = k_3, \qquad \phi = k_1. \tag{5}$$

De estos infinitesimales obtenemos que la ecuación es invariante bajo el grupo de las traslaciones con respecto al espacio y al tiempo

$$\mathbf{v_1} = \partial_x, \qquad \mathbf{v_2} = \partial_t$$

y bajo el grupo

$$\mathbf{v_3} = t\partial_x + \partial_u.$$

Si $\eta = 0$ la ecuación (1) es una ecuación generalizada de Kuramoto-Sivashinski. Las simetrías clásicas, estudiadas por Bruzón, Gandarias y Camacho en [4], coinciden con las obtenidas para $\eta \neq 0$ en (5).
REDUCCIÓN 1. Consideramos el generador $\lambda\mathbf{v_1} + \mathbf{v_2}$. Sustituyendo los infinitesimales del generador en la ecuación característica

$$\xi u_x + \tau u_t = \phi \tag{6}$$

deducimos la transformación

$$u(x,t) = h(z), \qquad z(x,t) = x - \lambda t \tag{7}$$

que reduce la Ec. (1) a la EDO

$$\eta h^{(5)} + \alpha h^{(4)} + \psi h^{(3)} + \beta h'' + hh' - \lambda h' = 0. \tag{8}$$

Integrando Ec. (8) con respecto a z obtenemos

$$\eta h^{(4)} + \alpha h^{(3)} + \psi h'' + \beta h' + \frac{1}{2}h^2 - \lambda h + k_1 = 0, \tag{9}$$

donde k_1 es una constante de integración. La ecuación (9) no admite simetrías de Lie. Nosotros aplicamos el método de la ecuación más simple a la ecuación (9) y obtenemos soluciones tipo ondas viajeras de la ecuación de la forma

$$\begin{aligned}
h(z) = \; & -\frac{280e^{z+c_1}}{92807\left(e^z - e^{c_1}\right)^4}\left(-\left(-12489 + \sqrt{102079263}\right)e^{2z}\right.\\
& \left. + \left(12489 + \sqrt{102079263}\right)e^{2c_1} + 14316e^{z+c_1}\right).
\end{aligned}$$

REDUCCIÓN 2. Consideramos el generador $\mathbf{v_2} + \mathbf{v_3}$. Sustituyendo los infinitesimales del generador en la ecuación característica (6) obtenemos la transformación

$$u(x,t) = t + h(z), \qquad z(x,t) = x - \frac{t^2}{2} \tag{10}$$

que reduce la Ec. (1) a la EDO

$$\eta h^{(5)} + \alpha h^{(4)} + \psi h^{(3)} + \beta h'' + hh' + 1 = 0. \tag{11}$$

Integrando Ec. (11) con respecto a z obtenemos

$$\eta h^{(4)} + \alpha h^{(3)} + \psi h'' + \beta h' + \frac{1}{2}h^2 + (k_1 + 1)z = 0, \tag{12}$$

donde k_1 es una constante de integración. Esta ecuación no admite simetrías de Lie. Para $k_1 = -1$ la ecuación (12) se reduce a la ecuación (9).

REDUCCIÓN 3. Consideramos el generador $\mathbf{v_3}$. Sustituyendo los infinitesimales del generador en la ecuación característica (6) obtenemos la reducción

$$u(x,t) = \frac{x}{t} + h(z), \qquad z(x,t) = t \tag{13}$$

que reduce la Ec. (1) a la EDO

$$h' + \frac{1}{z}h = 0. \tag{14}$$

La solución de Ec. (14) es $h = \dfrac{c_1}{z}$.

3. Soluciones tipo ondas viajeras

En esta sección obtenemos un tipo de soluciones ondas viajeras de la ecuación (1) denominados kinks. Los kinks son estructuras coherentes no lineales en una dimensión espacial. Estas soluciones son ondas viajeras con la particularidad de que su energía se localiza alrededor del centro del kink. Para su obtención, utilizamos el método de la transformación de Cole-Hopf y consideramos soluciones de la ecuación (1) de la forma

$$u(x,t) = R\frac{f_x}{f}, \tag{15}$$

donde la función auxiliar f está dada por

$$f = 1 + k\exp(\mu x - \lambda t). \tag{16}$$

Sustituyendo (15) y (16) en la ecuación (1), encontramos que obtenemos solución para la ecuación (1) cuando $\eta = \alpha = \delta = 0$, es decir para la ecuación de Burgers, y es

$$u(x,t) = \frac{2k\mu\beta\exp(\mu(x+\beta\mu t))}{1 + k\exp(\mu(x+\beta\mu t))}. \tag{17}$$

En la Figura 1 mostramos la solución (17) para $\mu = \frac{1}{4}$, $\beta = 2$ y $k = 1$, en la que podemos observar cómo la solución describe un kink.

4. Conclusiones

En este trabajo, haciendo uso del método clásico de Lie se ha realizado una clasificación completa de las simetrías de Lie de una ecuación modificada de Benney-Lin. Hemos construido todas las soluciones invariantes a partir del sistema óptimo de subálgebras, así como todas las ecuaciones diferenciales ordinarias a las que se reducen la ecuación (1). Se ha aplicado el método de la ecuación más simple a una de las ecuaciones obtenidas. Se han obtenido soluciones tipo ondas viajeras de la ecuación. Estas soluciones compactas - kinks - interactúan entre sí sólo a cortas distancias, porque no poseen colas infinitas.

Agradecimientos

Este artículo está subvencionado por el proyecto MTM2009-11875 de DGICYT y por el Grupo PAI FQM-201 de la Junta de Andalucía.

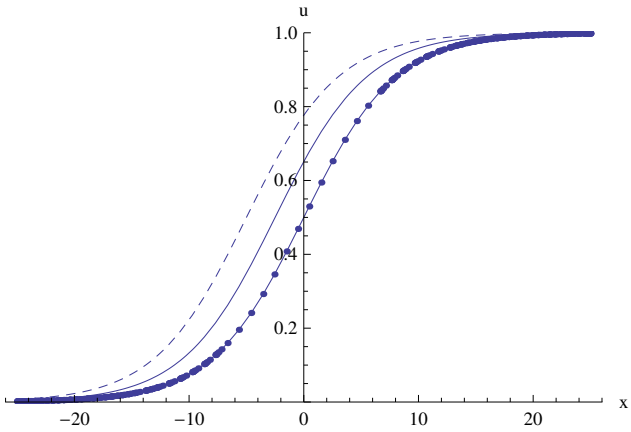


Figura 1: Solución (17) para $\mu = \frac{1}{4}$, $\beta = 2$ y $k = 1$

REFERENCIAS

[1] G.W. BLUMAN, AND S. KUMEI, *Symmetries and differential equations*, Springer-Verlag, 1989.

[2] M.S. BRUZÓN, M.L. GANDARIAS, AND J.C. CAMACHO, *Symmetry for a Family of BBM Equations*. J. Nonl. Math. Phys. 15 (2008), pp.81-90.

[3] M.S. BRUZÓN, *Exact Solutions for a Generalized Boussinesq Equation*. Theor. Math. Phys. 160 (2009), pp.894-904.

[4] M.S. BRUZÓN, M.L. GANDARIAS AND J.C. CAMACHO, *Classical and nonclassical symmetries for a Kuramoto-Sivashinsky equation with dispersive effects*. Math. Meth. Appl. Sci. 30 (2007), pp.2091-2100.

[5] W. CHEN, AND J. LI, *On the low regularity of the Benney-Lin equation*. J. Math. Anal. Appl. 339 (2008), pp.1134-1147.

[6] N.A. KUDRYASHOV, *Simplest equation method to look for exact solutions of nonlinear differential equations* Chaos, Solitons and Fractals 24 (2005), pp.1217-1231.

[7] N.A. KUDRYASHOV, AND N.B. LOGUINOVA, *Extended simplest equation method for nonlinear differential equations*. Appl. Math. and Comput. 205 (2008), pp.396-402.

[8] P.J. OLVER, *Applications of Lie groups to differential equations*, Springer-Verlag, 1986.

[9] M. WANG, X. LI, AND J. ZHANG, *The* $\left(\frac{G'}{G}\right)$*-expansion method and travelling wave solutions of nonlinear evolution equations in mathematical physics*. Phys. Lett. A. 372 (2008), pp.417-423.

SOLUCIÓN LOCAL Y GLOBAL DEL PROBLEMA DE CAUCHY ASOSIADO A UNA PERTURBACIÓN NO LOCAL DE LA ECUACIÓN DE BENJAMÍN-ONO PERIÓDICA

Darwin Peña González[♭]

[♭]*Grupo de Matemática Aplicada, Universidad Autónoma del Caribe, Barranquilla, Colombia, dpena@uac.edu.co, www.uac.edu.co*

Resumen: La propuesta de este trabajo es el análisis de la ecuación BENJAMÍN-ONO agregandole una cantidad disipativa y otra de inestabilidad, y haciendo uso de técnicas clásicas probaremos que (4) está local y globalmente bien planteado en $H^s(\mathbb{T})$ para $s > \frac{1}{2}$.

Palabras clave: *Problema de Cauchy, Ecuación BENJAMÍN-ONO, Local y Globalmente bien puesto.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

La ecuación Korteweg-de Vries (KdV)

$$\begin{cases} \partial_t u(t,x) + u(t,x)\,\partial_x u(t,x) + \partial_x^3 u(t,x) - \eta\left(\mathcal{H}\partial_t u(t,x) + \mathcal{H}\partial_x^3 u(t,x)\right) = 0 \\ u(\cdot,0) = \psi(\cdot) \end{cases} \tag{1}$$

ha sido objeto de intensa investigación en diversas áreas de la física y la matemática tales como en teoría de fluidos [6], sistemas completamente integrables[1], en ecuaciones diferenciales parciales[7] por mencionar algunas de ellas. El parámetro η es un número positivo arbitrario y $\mathcal{H}$ denota la transformada de Hilbert. Problemas de este tipo han surgido en el estudio y el análisis de algunos problemas físicos [1]. El estudio de ellos ha originado nuevas ideas en diferentes áreas de la Físico-Matemática (Mathematical Physics)[7].

Alvarez en [2] estudia el problema de Cauchy asociado a una perturbación no local de la ecuación (KdV):

$$\partial_t u(t,x) + u(t,x)\,\partial_x u(t,x) + \partial_x^3 u(t,x) = 0, t > 0, x \in \mathbb{R}. \tag{2}$$

La ecuación (2) es tratada por S. Qian, H.H. Chen, and Y.C. Lee, *A turbulence model whit stochastic soliton motion*, Departement o Physics and astronomy. The laboratory for plasma and fusion energy studies, University of Maryland, College Park, Maryland 20742 [10]. (Received 19 may 1988; accepted for publication 20 september 1989) hace un estudio de un modelo de turbulencia con soluciones estocásticas, este análisis lo hace a partir de la ecuación BENJAMÍN-ONO (3) la cual describe la interface entre dos fluidos de diferentes densidades y de profundidad infinita.

$$\partial_t u = -2u\partial_x u - \mathcal{H}\partial_x^2 \tag{3}$$

El proposito de este trabajo es abordar el problema de valor inicial

$$\begin{cases} \partial_t u = \nu\partial_x^2 u - 2u\partial_x u + \mu\mathcal{H}\partial_x u - \beta\mathcal{H}\partial_x^2 u \\ u(t,0) = \phi(t), \quad \phi \in C([0,T]; H^s(\mathbb{T})) \end{cases} \tag{4}$$

En (4) las expresiones $\partial_x^2 u$ y $\mathcal{H}\partial_x u$ generan disipación e inestabilidad respectivamente y la expresión $\mathcal{H}\partial_x^2 u$ es el término dispersivo, además $\mathbb{T} \cong \dfrac{\mathbb{R}}{2\pi\mathbb{Z}}$ y $\mathcal{H}$ denota la transformada de Hilbert como:

$$\widehat{\mathcal{H}f}(k) = -\frac{i}{2} sgn(k)\,\widehat{f}(k),$$

El buen planteamiento que haremos del problema (4) sera en espacios de Sobolev periódicos $H^s(\mathbb{T})$ donde $\nu > 0, \mu > 0, \beta \in \mathbb{R}$ y haciendo uso de técnicas clásicas probaremos que (4) está localmente bien planteado en $H^s(\mathbb{T})$ para $s > \frac{1}{2}$ y que globalmente tambien exite.

2. BUEN PLANTEAMIENTO LOCAL EN $H^s(\mathbb{T})$ PARA $s > \frac{1}{2}$.

Teorema 1 *Si $s > \frac{1}{2}$, el problema (4) es equivalente a la ecuación integral*

$$u(t) = \underbrace{e^{-tA}\phi - \int_0^t e^{-(t-\tau)A}\left(\partial_x u^2(\tau)\right) d\tau}_{\Psi(u)} \tag{5}$$

Más precisamente, si $u \in C([0,T]; H^s(\mathbb{T}))$ con $s > \frac{1}{2}$, es una solución de (4) entonces u satisface (5). Reciprocamente, si $u \in C([0,T]; H^s(\mathbb{T}))$, con $s > \frac{1}{2}$, es solución de (5) entonces $u \in C^1\left([0,\mathbb{T}]; H^{s-1}(\mathbb{T})\right)$ y satisface (4) con derivada dada por

$$\lim_{h\to 0}\left\| \frac{u(t+h)-u(t)}{h} - Bu(t) + \left(\partial_x u^2\right)(t)\right\|_{s-2} = 0 \tag{6}$$

Prueba. La primera parte de la prueba es consecuencia del método de variación de parametros y de observar que el término no lineal $\partial_x u^2$ tiene sentido, pues puede considerarse como una distribución periodica, ya que u^2 es continua y se anula en infinito para $s > \frac{1}{2}$, en virtud del Lema de Sovolev. La segunda parte se obtiene por reemplazar la u de la ecuación integral (5) en (6) □

Teorema 2 *Sea $\phi \in H^s(\mathbb{T})$,con $s > \frac{1}{2}$. Entonces, existe $T = T(\|\phi\|_s) > 0$ y una única solución $u \in C([0,T]; H^s(\mathbb{T}))$*

Prueba. La idea de la prueba es aplicar el teorema de contracción de Banach a la función definida por el miembro derecho de (5) en el espacio adecuado. Con este fin definimos lo siguiente:
$\Phi_T(M) = \left\{u \in C([0,T], H^s(\mathbb{T})) \diagup \left\|u(t) - e^{-tA}\phi\right\|_s \leq M\right\}$ el cual es cerrado en $C([0,T], H^s(\mathbb{T}))$ y por lo tanto es completo. Y se muestra que que $\Psi(u) \in C([0,T]; H^s(\mathbb{T}))$ es decir

$$\left\|\Psi(u(t)) - \Psi\left(u\left(t^{\prime}\right)\right)\right\|_s \underset{t^{\prime}\to t}{\to} 0, \text{ con } t \in [0,T] \tag{7}$$

□

Teorema 3 *El problema (4) es localmente bien puesto en $H^s(\mathbb{T})$, para $s > \frac{1}{2}$. Más precisamente, para $\phi \in H^s(\mathbb{T})$, existen $T > 0$ y una única solución $u \in C([0,T]; H^s(\mathbb{T}))$ que satisface (4) y tal que $u \in C^1\left([0,T]; H^{s-1}(\mathbb{T})\right)$. Además, la aplicación de $\phi \in H^s(\mathbb{T}) \mapsto u \in C([0,T]; H^s(\mathbb{T}))$ es continua en el siguiente sentido: Sean $\phi_n \in H^s(\mathbb{T})$, $n = 1,2,\ldots,$ tales que $\phi_n \to \phi$ y sean $u_n \in C([0,T_n]; H^s(\mathbb{T}))$ soluciones de (4) con $u_n(0) = \phi_n$. Entonces, las soluciones u_n pueden ser extendidas si es necesario al intervalo $[0,T]$ para n suficientemente grande y*

$$\lim_{n\to\infty} \sup_{[0,T]} \|u(t) - u_n(t)\|_s = 0 \tag{8}$$

3. EL PROBLEMA GLOBAL EN $H^1(\mathbb{T})$ Y $H^2(\mathbb{T})$

Proposición 1 *Sean $\phi \in H^2(T)$ y $u \in C\left([0,T]; H^2(T)\right)$ la solución de (4) con $u(0) = \phi$. Entonces*

$$\|u(t)\|_0 \leq \|\phi\|_0 \tag{9}$$

$$\|\partial_x u(t)\|_0 \leq \left\|\phi^{\prime}\right\|_0 e^{C\|\phi\|_0^4 T} \tag{10}$$

$$\left\|\partial_x^2 u(t)\right\|_0 \leq \left\|\phi^{\prime\prime}\right\|_0 e^{C\|\phi\|_0^4 T} \tag{11}$$

Teorema 4 *Sea $\phi \in H^s(\mathbb{T})$, con $s = 1$ o $s = 2$. Entonces el problema de valor inicial (1) es globalmente bien puesto en $H^s(\mathbb{T})$.*

Prueba. Este resultado es consecuencia directa de la Proposición (1), y del Teorema (3) de la buena colocación local de (4). □ □

4. EL PROBLEMA GLOBAL EN $H^s(\mathbb{T})$ PARA $s \geq 2$

Teorema 5 *Si* $s \geq 2$, *el problema de valor inicial (4) es globalmente bien planteado.*

Prueba. Este resultado es consecuencia directa de

$$\|u\|_s^2 \leq \|\phi\|_s^2 e^{K(\phi)t} \leq \|\phi\|_s^2 e^{K(\phi)T} \tag{12}$$

y del Teorema (3) de la buena colocación local de (4). □

5. EL PROBLEMA GLOBAL EN $H^s(\mathbb{T})$ PARA $1 \leq s \leq 2$

Teorema 6 *Si* $1 < s < 2$, *el problema de valor inicial (4) es globalmente bien planteado.*

Prueba. Solo resta por obtener una estimativa a priori de la $\|u\|_s$ de la solución (4) con $1 \leq s \leq 2$. Sabemos que (4) es equivalente a la ecuación integral

$$u(t) = e^{-tA}\phi - \int_0^t e^{-(t-\tau)A}\partial_x u^2(\tau)\, d\tau$$

Con $A = -\left(-\mu\mathcal{H}\partial_x u + \beta\mathcal{H}\partial_x^2 u + \nu\partial_x^2 u\right)$ y $\phi \in H^s(\mathbb{T})$. Sea $\lambda \in (0,1)$. La desigualdad

$$\left\|e^{-tA}\phi\right\|_{s+\lambda} \leq C_\lambda \left(1 + \left(\frac{\lambda}{t}\right)^{\frac{\lambda}{2}}\right) \|\phi\|_s \tag{13}$$

con $s = 0$ y $1 + \lambda$, junto con (9) y (10) implica que:

$$\|u\|_{1+\lambda} \leq \|\phi\|_{1+\lambda} + K(\phi)\left[T + T^{\frac{1-\lambda}{2}}\right] \tag{14}$$

donde $K(\phi) = C_\lambda\left(\|\phi\|_0^2 + \|\phi'\|_0^2 e^{C\|\phi\|_0^4 T}\right)$ que se obtiene a partir de (9) y (10). Por lo tanto, la estimativa (14) muestra que la norma $\|u\|_{1+\lambda}$ permanece acotada. □

AGRADECIMIENTOS

A Guillermo Rodriguez Blanco, Doctor Universidad Nacional de Bogotá, Colombia, a la Universidad Autónoma del Caribe, a mis Amigos y Familiares y muy especialmente a Elvira mi compañera.

REFERENCIAS

[1] Bao-Feng Feng, T. Kawahara, *Multi-hump stationary waves for a Korteweg-de Vries equation with nonlocal perturbations*, Physica D, 137, (2000), pp. 237-246.

[2] Borys Y. Alvarez S., *On the Cauchy problem for a nonlocal perturbation of the KdV equation*, Tesis Doctoral, IMPA, 2002.

[3] R. J. Iório, Jr.,*On the Cauchy porblem for the Benjamin-Ono Equation*, Comm. PDE, 11, (1986), pp. 1031-1081.

[4] R. J. Iório, Jr.,*The Benjamin-Ono Equation in Weighted Sobolev Spaces*, J. Math. Anal. Appl., Vol.157, No. 2, (1991), pp. 577-590.

[5] R. J. Iório, Jr.,*KdV, BO and Friends in Weighted Sobolev Spaces, Function Analytic Methods for Partial Differential Equations*, Springer-Verlag, vol. 1450(1990) pp. 105-121.

[6] R. J. Iório, Jr., Valèria de Magalhães Iório, *Fourier Analysis and Partial Differential Equations*, Cambridge studies in avanced mathematics, 70, (2001).

[7] T. Kato, *Nonstationary Flows of Viscous and Ideal Fluids in* $\mathbb{R}^3$, Journal of functional Analysis, 9 (1972), pp. 296-305.

[8] T. Kato, *On the Cauchy problem for the (Generalized) Korteweg-de Vries Equation*, Studies in Applied Mathematics, Advances in Mathematics Supplementary Studies, Vol. 8,(1983), pp. 93-128.

[9] T. Kato and H. Fujita, *On the non-stationary Navier-Stokes system*, Rend. Sem. Mat. Univ. Padova, 32(1962), pp.243-260.

[10] S. Qian, H.H. Chen, and Y.C. Lee, *A turbulence model whit stochastic soliton motion*, (Received 19 may 1988; accepted for publication 20 september 1989).

[11] N. Alon, R. A. Duke, H. Lefmann, V. Rodl, and R. Yuster. *The algorithmic aspects of the regularity lemma*. J. Algorithms, 16(1): 80-109, 1994.

[12] T. Roger. *Infinite-dimensional dynamical systems in mechanics and physics*. Springer-Verlag. p(50), 1988.

[13] Henry, *Geometric theory of semilinear parabolic equation, Lectures Notes in Mathematics*, vol. 840, Springer (1957).

EXISTENCIA Y UNICIDAD DE SOLUCIÓN GLOBAL PARA LA ECUACIÓN DEL CALOR NO-CLÁSICA PARA UN SEMI-ESPACIO N-DIMENSIONAL

Mahdi Boukrouche * , Domingo A. Tarzia †‡
* PRES Lyon University, University of Saint-Etienne, Laboratory of Mathematics, LaMUSE EA-3989,
23 rue Paul Michelon, 42023 Saint-Etienne, France.
E-mail: Mahdi.Boukrouche@univ-st-etienne.fr
† Departamento de Matemática, Facultad de Ciencias Empresariales, Univ. Austral,
Paraguay 1950, S2000FZF Rosario, Argentina.
‡ CONICET, Argentina.
E-mail: DTarzia@austral.edu.ar

Resumen: Sea *D* un semi-espacio n-dimensional con frontera *S*. Se considera la ecuación del calor no-clásica en el dominio *D* para la cual la fuente de energía interna depende del flujo de calor sobre la frontera *S*. El problema está motivado por la modelización de la regulación de la temperatura en el medio. Utilizando la función de Green para el dominio *D* se encuentra para la solución una representación integral en función del flujo de calor *V* sobre *S* que es una incógnita suplementaria del problema. Se obtiene que *V* debe satisfacer una ecuación integral de Volterra de segunda especie en el tiempo *t* y con un parámetro en $\mathbb{R}^{n-1}$. Bajo ciertas condiciones sobre los datos del problema se demuestra que existe una única solución local que puede extenderse globalmente en el tiempo. Se generalizan resultados obtenidos en Berrone-Tarzia-Villa, Math. Meth. Appl. Sci., 23 (2000), pp. 1161-1177 y Tarzia-Villa, Rev. Unión Mat. Argentina, 41 (1998), pp. 99-114 para el caso unidimensional.

Palabras clave: *Ecuación del calor no-clásica n-dimensional, Ecuación integral de Volterra, Existencia y unicidad de solución, Representación integral de la solución..*
2000 AMS Subjects Classification: 35C15, 35K05, 35K20, 35K60, 80A20.

1. INTRODUCCIÓN

En los trabajos [1, 17, 18] se han considerado el siguiente problema no lineal para la ecuación del calor no-clásica para un material semi-infinito:

$$\begin{cases} i)\, u_t - u_{xx} = -F\left(u_x(0,t)\right), \quad x>0, \quad t>0 \\ ii)\, u(0,t) = g(t), \quad t>0 \\ iii)\, u(x,0) = h(x), \quad x>0 \end{cases} . \tag{1}$$

Tales problemas están motivados por la modelización de la regulación de la temperatura en un medio isótropo, con una fuente no uniforme la cual provee un enfriamiento o calentamiento del sistema dependiendo de las propiedades de *F* con relación al desarrollo del flujo del calor en el borde $x=0$ [5, 7], por ejemplo, si se supone:

$$V\ F(V,t) > 0, \quad \forall V \neq 0, \quad F(0) = 0, \tag{2}$$

entonces la fuente enfría cuando $u_x(0,t) > 0$ y calienta cuando $u_x(0,t) < 0$. Algunas otras referencias en el tema son [6, 9-12].

En el presente trabajo se considera un caso n-dimensional. A los efectos de facilitar la notación se denota un punto de $\mathbb{R}^n$ de la siguiente manera:

$$(x_1, x_2, ..., x_n) = (x, y), \quad \text{con} \quad x = x_1 \in \mathbb{R}, \quad y = (x_2, ..., x_n) \in \mathbb{R}^{n-1} .$$

Sea *D* un semi-espacio n-dimensional con frontera *S*, definidos por:

$$D = \mathbb{R}^+ \times \mathbb{R}^{n-1} = \left\{(x,y) \in \mathbb{R}^n \,/\, x = x_1 > 0,\ y = (x_2, ..., x_n) \in \mathbb{R}^{n-1}\right\} , \tag{3}$$

$$S = \partial D = \{0\} \times \mathbb{R}^{n-1} = \left\{(x,y) \in \mathbb{R}^n \,/\, x = 0,\ y \in \mathbb{R}^{n-1}\right\} . \tag{4}$$

El objetivo del presente trabajo es el de estudiar el siguiente problema para la ecuación del calor no-clásica en el dominio *D* para la cual la fuente de energía interna depende del flujo de calor sobre la frontera *S* : Hallar la temperatura u=u(x,y,t) de manera que satisfaga las siguientes condiciones:

$$\begin{cases} i)\, u_t - \Delta u = -F\left(u_x\left(0,y,t\right)\right), \quad x>0, \quad y\in\mathbb{R}^{n-1}, \quad t>0 \\ ii)\, u\left(0,y,t\right)=0, \quad y\in\mathbb{R}^{n-1}, \quad t>0; \quad iii)\, u\left(x,y,0\right)=h\left(x,y\right), \quad x>0, \quad y\in\mathbb{R}^{n-1} \end{cases} . \tag{5}$$

En la Sección II se dan las soluciones básicas para la ecuación del calor n-dimensional que serán utilizadas en la Sección III para demostrar que, bajo ciertas condiciones sobre los datos *F* y *h* del problema (5), existe una única solución local que puede extenderse globalmente en el tiempo. Se generalizan resultados obtenidos en [1, 17] para el caso unidimensional.

2. SOLUCIONES BÁSICAS PARA LA ECUACIÓN DEL CALOR N-DIMENSIONAL

La solución del siguiente problema de Cauchy para la ecuación del calor n-dimensional:

$$\begin{cases} i)\, u_t - \Delta u = 0, \quad (x,y)\in\mathbb{R}^n, \quad t>0 \\ ii)\, u\left(x,y,0\right)=h\left(x,y\right), \quad (x,y)\in\mathbb{R}^n \end{cases} \tag{6}$$

es conocida como la fórmula de Poisson dada por la expresión [8,13]:

$$u\left(x,y,t\right)=\int_{\mathbb{R}^n} K\left(x,y,t;\xi,\eta,0\right)h\left(\xi,\eta\right)d\xi\, d\eta \tag{7}$$

donde *K* es la solución fundamental de la ecuación del calor n-dimensional definida por:

$$K\left(x,y,t;\xi,\eta,\tau\right)=\frac{1}{\left(2\sqrt{\pi\left(t-\tau\right)}\right)^n}\exp\left(-\frac{\left(x-\xi\right)^2+\left\|y-\eta\right\|^2}{4\left(t-\tau\right)}\right), (x,y)\in\mathbb{R}^n, (\xi,\eta)\in\mathbb{R}^n, t>\tau\,, \tag{8}$$

con

$$\begin{cases} (\xi_1,\xi_2,.....,\xi_n)=(\xi,\eta), \quad \text{con} \quad \xi=\xi_1\in\mathbb{R}, \quad \eta=(\xi_2,.....,\xi_n)\in\mathbb{R}^{n-1} \\ \left\|y-\eta\right\|=\sqrt{\sum_{i=2}^{n}\left(x_i-\xi_i\right)^2}: \quad \text{norma en } \mathbb{R}^{n-1} \end{cases} . \tag{9}$$

Lema 1 *La solución del problema:*

$$\begin{cases} i)\, u_t - \Delta u = 0, \quad x>0, \quad y\in\mathbb{R}^{n-1}, \quad t>0 \\ ii)\, u\left(0,y,t\right)=0, \quad y\in\mathbb{R}^{n-1}, \quad t>0 \\ iii)\, u\left(x,y,0\right)=h\left(x,y\right), \quad x>0, \quad y\in\mathbb{R}^{n-1} \end{cases} , \tag{10}$$

está dada por la siguiente expresión:

$$u\left(x,y,t\right)=\int_D G_1\left(x,y,t;\xi,\eta,0\right)h\left(\xi,\eta\right)d\xi\, d\eta \tag{11}$$

donde G_1 es la función de Green para la ecuación del calor n-dimensional con condición de frontera de tipo Dirichlet nula, dada por la siguiente expresión:

$$G_1\left(x,y,t;\xi,\eta,\tau\right)=K\left(x,y,t;\xi,\eta,\tau\right)-K\left(-x,y,t;\xi,\eta,\tau\right)=\frac{1}{\left(2\sqrt{\pi\left(t-\tau\right)}\right)^{n-1}}\exp\left(-\frac{\left\|y-\eta\right\|^2}{4\left(t-\tau\right)}\right)G\left(x,t;\xi,\tau\right) \tag{12}$$

siendo G la función de Green para el caso unidimensional.

Lema 2 *La solución del problema:*

$$\begin{cases} i)\, u_t - \Delta u = 0, \quad x>0, \quad y\in\mathbb{R}^{n-1}, \quad t>0 \\ ii)\, u_x\left(0,y,t\right)=0, \quad y\in\mathbb{R}^{n-1}, \quad t>0 \\ iii)\, u\left(x,y,0\right)=h\left(x,y\right), \quad x>0, \quad y\in\mathbb{R}^{n-1} \end{cases} , \tag{13}$$

está dada por la siguiente expresión:

$$u(x,y,t)=\int_D N_1(x,y,t;\xi,\eta,0)\,h(\xi,\eta)\,d\xi\,d\eta \tag{14}$$

donde N_1 es la función de Green para la ecuación del calor n-dimensional con condición de frontera de tipo Neumann nula, dada por la siguiente expresión:

$$N_1(x,y,t;\xi,\eta,\tau)=K(x,y,t;\xi,\eta,\tau)+K(-x,y,t;\xi,\eta,\tau)=\frac{1}{\left(2\sqrt{\pi(t-\tau)}\right)^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)N(x,t;\xi,\tau) \tag{15}$$

siendo N la función de Neumann para el caso unidimensional.

Lema 3 *Las funciones G_1 y N_1 tienen las siguientes propiedades fundamentales:*

$$\int_{\mathbb{R}^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)d\eta=\left(2\sqrt{\pi(t-\tau)}\right)^{n-1},\quad \int_{\mathbb{R}^{n-1}}\|y-\eta\|^2\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)d\eta=\frac{(n-1)(t-\tau)}{\sqrt{2}}\left(2\sqrt{\pi(t-\tau)}\right)^{n-1} \tag{16}$$

$$G_1(0,y,t;\xi,\eta,\tau)=0,\quad N_{1x}(0,y,t;\xi,\eta,\tau)=0 \tag{17}$$

$$\int_{\mathbb{R}^{n-1}}G_1(x,y,t;\xi,\eta,\tau)\,d\eta=G(x,t;\xi,\tau),\quad \int_{\mathbb{R}^{n-1}}N_1(x,y,t;\xi,\eta,\tau)\,d\eta=N(x,t;\xi,\tau) \tag{18}$$

$$\int_{\mathbb{R}^{n-1}}G_{1x}(0,y,t;\xi,\eta,\tau)\,d\eta=G_x(0,t;\xi,\tau)=\frac{\xi}{2\sqrt{\pi}\,(t-\tau)^{3/2}}\exp\left(-\frac{(x-\xi)^2}{4(t-\tau)}\right) \tag{19}$$

$$\int_0^{+\infty}N_1(x,y,t;\xi,\eta,\tau)\,d\xi=\frac{1}{\left(2\sqrt{\pi(t-\tau)}\right)^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right) \tag{20}$$

$$\int_0^{+\infty}G_1(x,y,t;\xi,\eta,\tau)\,d\xi=\frac{1}{\left(2\sqrt{\pi(t-\tau)}\right)^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)erf\left(\frac{x}{2\sqrt{t-\tau}}\right) \tag{21}$$

$$\int_0^{+\infty}G_{1x}(0,y,t;\xi,\eta,\tau)\,d\xi=\frac{1}{\left(2\sqrt{\pi(t-\tau)}\right)^{n}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)\ . \tag{22}$$

3. EXISTENCIA Y UNICIDAD DE SOLUCIÓN DEL PROBLEMA (5)

Teorema 4 *La representación integral de la solución del problema (5) está dada por la siguiente expresión:*

$$\begin{aligned}u(x,t)=&\int_D G_1(x,y,t;\xi,\eta,0)\,h(\xi,\eta)\,d\xi\,d\eta\\&-\int_0^t\frac{erf\left(\dfrac{x}{2\sqrt{t-\tau}}\right)}{\left(2\sqrt{\pi(t-\tau)}\right)^{n-1}}\left[\int_{\mathbb{R}^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)F(V(\eta,\tau))d\eta\right]d\tau\end{aligned} \tag{23}$$

donde la función $V=V(y,t)$, definida por $V(y,t)=u_x(0,y,t)$, $y\in\mathbb{R}^{n-1}$, $t>0$ (flujo de calor en la superficie $x=0$), satisface la ecuación integral de Volterra siguiente:

$$\begin{aligned}V(y,t)=&\int_D G_{1x}(0,y,t;\xi,\eta,0)\,h(\xi,\eta)\,d\xi\,d\eta\\&-2\int_0^t\frac{1}{\left(2\sqrt{\pi(t-\tau)}\right)^{n}}\left[\int_{\mathbb{R}^{n-1}}\exp\left(-\frac{\|y-\eta\|^2}{4(t-\tau)}\right)F(V(\eta,\tau))d\eta\right]d\tau\end{aligned} \tag{24}$$

en la variable $t>0$ siendo $y\in\mathbb{R}^{n-1}$ un parámetro.

Teorema 5 *Bajo las hipótesis*

$$h \in C^0(D), \quad F \in C^0(\mathbb{R}) \text{ y localmente Lipschitz en } \mathbb{R} \tag{25}$$

existe una única solución local que se puede extender globalmente en el tiempo.

Prueba. Con los datos (25) se verifican las hipótesis (H1)-(H3) y (H5)-(H8) para poder aplicar los Teoremas 1.2 [14, pp. 91] y 1.3 [14, pp. 97] a la ecuación integral (24) con lo cual se pueden obtener los resultados de la tesis. □

Observación 1 *Para los casos particulares*:

$$h(x,y) = h(x), \quad F\left(V\left(y,t\right)\right) = F\left(V\left(t\right)\right), \quad \forall y \in \mathbb{R}^{n-1} \tag{26}$$

se reencuentran resultados obtenidos en [1, 17, 18] para el caso unidimensional.

Observación 2 *Recientemente se han obtenido resultados para el problema de Stefan para la ecuación del calor no- clásica unidimensional para un material semi-infinito en [2-4, 16]. También se ha estudiado la ecuación del calor no- clásica unidimensional para un material finito en [15].*

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por los Proyectos PIP Nº 0460 de CONICET-UA y ANPCyT PICTO Austral No. 73, Rosario, Argentina.

REFERENCIAS

[1] A.C. BERRONE, D.A. TARZIA, AND L.T. VILLA, *Asymptotic behavior of a non-classical heat conduction problem for a semi-infinite material*, Math. Meth. Appl. Sci., 23 (2000), pp. 1161-1177.

[2] A.C. BRIOZZO, AND D.A. TARZIA, *Existence and uniqueness for one-phase Stefan problem of a non-classical heat equation with temperature boundary condition at a fixed face*, Electron. J. Diff. Eq., 2006 No. 21 (2006), pp. 1-16.

[3] A.C. BRIOZZO, AND D.A. TARZIA, *Exact solutions for nonclassical Stefan problems*, Int. J. Diff. Eq., Vol. 2010, Article ID 868059, pp. 1-19.

[4] A.C. BRIOZZO, AND D.A. TARZIA, *A Stefan problem for a non-classical heat equation with a convective condition*, Appl. Math. Comput., 217 (2010), pp. 4051-4060.

[5] J.R. CANNON, *The one-dimensional heat equation*, Addison-Wesley, Menlo Park, California, 1984.

[6] J.R. CANNON, AND H.M. YIN, *A class of non-linear non-classical parabolic equations*, J. Diff. Eq., 79 (1989), pp. 266-288.

[7] H.S. CARSLAW, AND C.J. JAEGER, *Conduction of heat in solids*, Clarendon Press, Oxford, 1959.

[8] A. FRIEDMAN, *Partial differential equations of parabolic type*, Prentice Hall (1964).

[9] K. GLASHOFF, AND J. SPREKELS, *An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats*, SIAM J. Math. Anal., 12 (1981), pp. 477-486.

[10] K. GLASHOFF, AND J. SPREKELS, *The regulation of temperature by thermostats and set-valued integral equations*, J. Integral Eq., 4 (1982), pp. 95-112.

[11] N. KENMOCHI, *Heat conduction with a class of automatic heat source controls*, Pitman Research Notes in Mathematics Series, 186 (1990), pp. 471-474.

[12] N. KENMOCHI, AND M. PRIMICERIO, *One-dimensional heat conduction with a class of automatic heat source controls*, IMA J. Appl. Math., 40 (1988), pp. 205-216.

[13] O.A. LADYZENSKAJA, V.A. SOLONNIKOV, AND N.N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, American Math. Society, Providence (1968).

[14] R.K. MILLER, *Nonlinear Volterra integral equations*, W.A. Benjamin (1971).

[15] N.N. SALVA, D.A. TARZIA, AND L.T. VILLA, *An initial-boundary value problem for the one-dimensional non-classical heat equation in a slab,* en Congreso III MACI 2011, Bahía Blanca, 9-11 Mayo 2011.

[16] D.A. TARZIA, *A Stefan problem for a non-classical heat equation*. MAT - Serie A, 3 (2001), pp. 21-26.

[17] D.A. TARZIA, AND L.T. VILLA, *Some nonlinear heat conduction problems for a semi-infinite strip with a non-uniform heat source*, Rev. Unión Mat. Argentina, 41 (1998), pp. 99-114.

[18] L.T. VILLA, *Problemas de control para una ecuación unidimensional no homogenea del calor*, Rev. Unión Mat. Argentina, 32 (1986), pp. 163-169.

TEORÍA CUASILINEAL DE KATO

César Loza Rojas[♭] y Juan Montealegre Scott[†]

[♭]*Departamento de Matemática, Universidad Nacional San Luis Gonzaga, Av. Los Maestros s/n, Ciudad Universitaria, Ica, Perú, lozacr@gmail.com, lozacr@yahoo.com, www.unica.edu.pe*
[†]*Departamento de Ciencias, Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel, Lima, Perú, jmscott@pucp.edu.pe*

Resumen: En este trabajo analizaremos el problema de Cauchy local asociado a la ecuación de Korteweg-de Vries (KdV) en H^s con $s > 3/2$, $p \in \mathbb{N}$

$$\begin{cases} \partial_t u(x,t) + \partial_x^3 u(x,t) + u^p(x,t)\,\partial_x u(x,t) = 0 \\ u(0) = u_0. \end{cases} \tag{1}$$

para cuya solución usaremos la teoría cuasilineal de Kato, de manera que la ecuación

$$\begin{cases} \partial_t u + A(t,u)\,u = f(t,u) \\ u(0) = u_0 \end{cases} \tag{2}$$

posea solución única en el espacio de Sobolev $H^s,\ s > 3/2$.

Palabras clave: *Espacios de Sobolev, problema de Cauchy, teoría de Kato*
2000 AMS Subject Classification: 35Q53 - 37K05

1. ESPACIOS DE SOBOLEV

Teorema 1 *Sean $s \in \mathbb{R}$, $k, n \in \mathbb{N}$ con $s > \frac{n}{2} + k$, entonces $H^s(\mathbb{R}^n) \hookrightarrow C_\infty^k(\mathbb{R}^n)$. Además, en tal caso, si $u \in H^s(\mathbb{R}^n)$ tenemos*

$$\|\partial^\alpha u\|_{L^\infty} \leq \max_{|\alpha| \leq k} \|\partial^\alpha u\|_{L^\infty} \leq C_s \|u\|_{H^s}$$

para todo multi-índice α con $|\alpha| \leq k$.

2. TEORÍA CUASI-LINEAL DE KATO

Antes de aplicar la teoría de Kato, hacemos el cambio de variable

$$u(t) = e^{-t\partial_x^3} v(t).$$

Por tanto, obtenemos el problema

$$\begin{cases} \partial_t v(t) + A(t, v(t))\, v(t) = 0, \quad t \geq 0 \\ v(0) = u_0. \end{cases} \tag{3}$$

donde A es un operador lineal que depende de $(t, y) \in [0, +\infty[\times H^s(\mathbb{R})$.

Consideremos el problema de Cauchy para la ecuación de evolución cuasi-lineal de tipo hiperbólico

$$\begin{cases} \partial_t u + A(t,u)\,u = f(t,u) \\ u(0) = u_0 \end{cases} \tag{4}$$

en un espacio de Banach X. Asumimos que para cada t y u, $A(t,u)$ es un operador lineal en X que genera un semigrupo, $f : [0,T] \times X \to X$ es una función dada y $u : [0,T] \to X$.

Presentaremos dos teoremas locales debidos a Kato, para (4), uno de ellos es de existencia y unicidad y el otro sobre dependencia continua de la solución en el dato inicial. Estos resultados son locales.

Con este fin consideremos las siguientes condiciones referidas al problema de Cauchy (4):

(X) Sean X e Y dos espacios de Banach reflexivos tales que Y está contenido densamente y continuamente en X. Además, existe un isomorfismo $S : Y \to X$ y la norma de Y es escogida de forma que S sea una isometría.

(A1) $A(\cdot,\cdot)$ es un operador definido en $[0,T_0]\times W$ con valores en $\mathcal{G}(X:1,\omega)$, siendo W una bola abierta en Y y ω un número real. Es decir, para cada $(t,y)\in[0,T_0]\times W$, $-A(t,y)$ es el generador de un semigrupo fuertemente continuo en X tal que

$$\left\|e^{-sA(t,y)}\right\|_{\mathcal{L}(X)}\leq e^{\omega s},\quad s\geq 0,\, t\in[0,T_0],\, y\in W.$$

(A2) Para cada $(t,y)\in[0,T_0]\times W$, tenemos

$$SA(t,y)S^{-1}=A(t,y)+B(t,y), \tag{5}$$

donde $B(t,y)\in\mathcal{L}(X)$ y $\|B(t,y)\|_{\mathcal{L}(X)}\leq\lambda_B$ con $\lambda_B>0$ una constante.

(A3) Para cada $(t,y)\in[0,T_0]\times W$ tenemos que $A(t,y)\in\mathcal{L}(Y,X)$, en el sentido que $Y\subseteq\mathcal{D}(A(t,y))$ y $A(t,y)|_Y\in\mathcal{L}(Y,X)$. Además, para cada $y\in W$ la aplicación $t\in[0,T_0]\mapsto A(t,y)$ es fuertemente continua.

(A4) Para cada $t\in[0,T_0]$ la aplicación $y\in W\mapsto A(t,y)$ es Lipschitz continua en $\mathcal{L}(Y,X)$; es decir, existe $\mu_A>0$ tal que

$$\|A(t,y_1)-A(t,y_2)\|_{\mathcal{L}(X)}\leq\mu_A\|y_1-y_2\|_Y,$$

para todo $t\in[0,T_0]$ y $y_1,y_2\in W$.

(A5) Existe $\mu_B>0$ tal que

$$\|B(t,y_1)-B(t,y_2)\|_{\mathcal{L}(X)}\leq\mu_B\|y_1-y_2\|_Y, \tag{6}$$

para todo $t\in[0,T_0]$ y $y_1,y_2\in W$.

(f1) $f:[0,T]\times W\to Y$ es acotada

$$\|f(t,y)\|_Y\leq\lambda_3,\qquad t\in[0,T],\qquad y\in W.$$

Para cada $y\in W$, $t\longmapsto f(t,y)$ es continua de $[0,T]$ en X mientras que para cada $t\in[0,T]$ la aplicación $y\in W\longmapsto f(t,y)$ es lipschitz en X esto es,

$$\|f(t,y)-f(t,z)\|_X\leq\mu_2\|y-z\|_X$$

donde $\mu_2\geq 0$ una constante.

(f2) Existe $\mu_4>0$ tal que

$$\|f(t,y)-f(t,z)\|_Y\leq\mu_4\|y-z\|_Y$$

para todo $t\in[0,T]$, y,$z\in W$.

En consecuencia, para (4), tenemos el resultado principal

Teorema 2

1. *Sea $s>3/2$. Para cada uno $u_0\in H^s$, existe $T>0$, dependiendo sólo de $\|u_0\|_{H^s}$, y una solución única u para (4) tal que*

$$u\in C([0,T]:H^s)\cap C^1\left([0,T]:H^{s-3}\right). \tag{7}$$

2. *La aplicacin $u_0\in H^s\longmapsto u\in C([0,T]:H^s)$ es continua en la norma de H^s. Con mayor precisin, si $u_n\in H^s$, $n\in\mathbb{N}$, con $\|u_n-u\|_{H^s}\to 0$ y $T'<T$, la solucin u_n para $u_n(0)=u_{0,n}$ existe en $[0,T']$ para un n suficientemente grande y $\|u_n(t)-u(t)\|_{H^s}\to 0$ uniformemente en $t\in[0,T']$.*

REFERENCIAS

[1] R. ADAMS, *Sobolev space*, Academic Press. London. 1975.

[2] G. DARMOIS, *Evolution equation in a Banach space*. Thesis, University of California, (1974).

[3] R. IÓRIO, W. NUNES, *Introdução à Equações de Evolução Não Lineares*. 18 ^{0} Colóquio Brasileiro de Matemática, IMPA/CNPq, (1991).

[4] T. KATO, *Quasi linear equations of evolution with application to partial differential equations*. Lect. Note in Math., 448, 25-70, (1975).

[5] T. KATO, *On the Cauchy problem for the (generalized) KdV equations*. Studies in Applied Mathematics, Advances in Mathematics Supplementary Studies, 8, 93-128.

[6] F. LINARES, G. PONCE, *Introduction to Nonlinear Dispersive Equations*. Publicações matemáticas. IMPA. RJ,(2008).

[7] M. DOS SANTOS, *A versão de Kato-Lai de Galerkin e a equação de Korteweg-de Vries*. Tesis de Mestrado, IMPA (1987).

[8] J.C. SAUT, R. TEMAM, *Remarks on the Korteweg-de Vries equation.* Israel J. of Math. 24, 78-87, (1976).

MÉTODO DE DESCOMPOSICIÓN DE ADOMIAN: SOLUCIONES APROXIMADAS DE UN PROBLEMA DE VALORES INICIALES

Silvia Seminara† y María Inés Troparevsky†

†*Departamento de Matemática, Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850 C1063ACV, Ciudad Autónoma de Buenos Aires, Argentina.*
sseminara@sinectis.com.ar , mitropa@fi.uba.ar

Resumen: El *método de descomposición* de Adomian (ADM) fue introducido por George Adomian en los años '80 para resolver ecuaciones diferenciales no lineales. Desde entonces ha sido utilizado en la resolución de numerosas ecuaciones diferenciales, integrales, integro-diferenciales y funcionales generales. Sin embargo, la demostración de la convergencia del método no ha sido completada para el caso general, sino bajo ciertas hipótesis particulares. En este trabajo construimos de dos maneras distintas la sucesión de soluciones aproximadas que propone este método para una ecuación en derivadas parciales no lineal de primer orden, y comparamos estas aproximaciones con las que se obtienen siguiendo el procedimiento sugerido en la demostración del clásico Teorema de Cauchy-Kovalevskaya para PDE. En base a los resultados obtenidos, elaboramos algunas conclusiones.

Palabras claves: *inicial value problems, PDE, decompositon methods*
2000 AMS Subjects Classification: 35F20 – 35F25 – 49M27

1. INTRODUCCION

En los años '80 George Adomian ([2], [3]) presentó su *método de descomposición* (ADM) que propone resolver un problema de valores iniciales no lineal de la forma $u_t = Lu + N(u)$ para t>0, con $u(x,0) = f(x)$, invirtiendo el operador lineal L y desarrollando el operador no lineal N en serie de funciones alrededor de una solución inicial. Suponiendo la existencia de una solución analítica, ADM construye una fórmula de recurrencia que permite aproximar la solución en un entorno de la condición inicial. Desde su presentación el método ha sufrido diferentes adaptaciones y ha sido usado satisfactoriamente para resolver numerosas ecuaciones no lineales: diferenciales, integro-diferenciales, algebraicas y funcionales de distinto tipo (en [8] se pasa revista a muchas de las aplicaciones). La convergencia del método fue probada en distintos contextos y bajo diferentes hipótesis: en [4], a partir de teoremas de punto fijo en espacios de Banach; en [1], a partir del teorema de Cauchy-Kovalevskaya (CK); en [10], reordenando los términos de una serie de Taylor generalizada; en [9], bajo ciertas hipótesis de acotación de las derivadas del operador no lineal, etc. En este trabajo construimos soluciones aproximadas para una PDE no lineal de primer orden mediante el ADM, considerando distintas elecciones de operadores L y N. Observamos las diferencias existentes entre las aproximaciones obtenidas y las comparamos con el desarrollo en serie de potencias de la solución que se obtiene por el método sugerido en la demostración de CK. En base a los resultados, elaboramos conclusiones.

2. EL MÉTODO DE DESCOMPOSICIÓN DE ADOMIAN

Exponemos brevemente el ADM. Puede encontrarse un desarrollo detallado en [3].

Sea el problema de valores iniciales $\begin{cases} u_t = Lu + N(u) & \forall t > 0 \\ u(x,0) = f(x) \end{cases}$ (1), con $L : X \to Y$ un operador lineal entre los espacios de Banach X e Y (con $X \subseteq Y$), $N : X \to Y$ un funcional no lineal, $f \in X$ la condición inicial conocida y $u_0 \in X$ la solución del problema libre asociado, $\begin{cases} u_t = Lu & \forall t > 0 \\ u(x,0) = f(x) \end{cases}$. Si N es analítico alrededor de u_0, y suponiendo la existencia de una

solución analítica de (1), ADM construye una sucesión de soluciones aproximadas a partir de plantear los desarrollos $u(x,t)=\sum_{n=0}^{\infty}u_n(x,t)$ y $N(u)=\sum_{n=0}^{\infty}A_n(u_0,u_1,...,u_n)$ y reemplazarlos en la forma integral equivalente de (1), $u(x,t)=E(t)f(x)+\int_0^t E(t-s)N(u(x,s))ds \quad \forall t\geq 0$, donde $E(t)=e^{tL}$ es el operador fundamental asociado al problema libre. De esta forma se obtiene una ecuación de recurrencia para el cálculo de u_n:

$$\begin{cases} u_0(x,t)=E(t)f(x) & \text{(solución del problema libre)} \\ u_{n+1}(x,t)=\int_0^t E(t-s)A_n(u_0(x,s),u_1(x,s),...u_n(x,s))ds & \forall n\geq 0 \end{cases} \quad (3)$$

Los términos $A_n(u_0,u_1,...,u_n)$ que componen el desarrollo de $N(u)$ se denominan *polinomios de Adomian* (se trata de polinomios en las u_i y sus derivadas), y no son únicos. Una forma de obtener una familia de tales polinomios es a partir de la fórmula $A_n=\frac{1}{n!}\frac{\partial^n}{\partial\varepsilon^n}(N(u_0+\varepsilon u_1+\varepsilon^2u_2+...+\varepsilon^n u_n))\Big|_{\varepsilon=0}$ (4).

Adomian argumentó en [3] que el desarrollo de $N(u)$ propuesto es solamente un reordenamiento del desarrollo de Taylor generalizado de N alrededor de la solución libre u_0, por lo que la convergencia de la serie $\sum_{n=1}^{\infty}u_n$ hacia la solución del problema puede asegurarse si N se supone analítico. Bajo esa hipótesis, $\varphi_m=\sum_{n=1}^{m}u_n$ constituye una aproximación a la solución del problema. Señaló, además, que la convergencia es rápida en razón de que el desarrollo se realiza alrededor de una primera aproximación óptima ([3], pp. 9-11).

Debemos señalar que en muchos de los trabajos que utilizan este método no se hace mención alguna a la existencia de la solución que se pretende aproximar. Por otra parte, si bien (4) provee de una forma aparentemente simple de obtener los polinomios de Adomian, el cálculo puede tornarse por demás engorroso, según el operador *N* de que se trate. La convergencia está probada en el Teorema 3 de [1].

3. El Teorema de Cauchy-Kovalevskaya

Este teorema, fundamental en la teoría de las ecuaciones diferenciales asociadas a problemas de valores iniciales, fue probado en varias etapas, comenzando por la demostración de Cauchy de 1835 para ODE y finalizando en 1875 con la prueba para PDE desarrollada por Sonya Kovalevskaya ([5]). Puede enunciarse del siguiente modo ([7]):

Teorema:

Si $G,\varphi_0,\varphi_1,...,\varphi_{k-1}$ *son analíticas cerca del origen, existe un entorno del origen en el cual el problema* $\begin{cases} \partial_t^k u=G(x,t,(\partial_x^\alpha\partial_t^j u)_{|\alpha|+j\leq k,\ j<k}) \\ \partial_t^j u(x,0)=\varphi_j(x) \quad \text{para } 0\leq j<k \end{cases}$ *tiene una única solución analítica.*

Existe una demostración constructiva que consiste en transformar el PVI en un sistema de primer orden cuasi-lineal autónomo, con condiciones homogéneas, y calcular formalmente las sucesivas derivadas parciales de la función incógnita en el origen; la convergencia de la serie de Taylor construida se demuestra mediante el método de mayorantes.

Aunque en las primeras publicaciones del método de Adomian no se menciona este teorema relevante, es en el marco del mismo que puede suponerse la existencia de la solución analítica que ADM aproxima. Demostraciones posteriores de la convergencia del ADM sí hacen referencia a él ([1], [10]).

4. EJEMPLO DE APLICACIÓN

Emplearemos el desarrollo en serie de potencias según sugiere la demostración del teorema de CK, y el método de Adomian, para obtener soluciones aproximadas del problema de valores iniciales, no lineal, de primer orden $u_t = u(1+u_x)$ para t>0, con $u(x,0) = e^x$, que satisface las hipótesis del teorema así como las condiciones de convergencia del ADM (ver [1]).

Observación: *En general, puede concluirse la analiticidad alrededor del origen de la función G del Teorema* CK *a partir de la analiticidad de* $N: X \to Y$ *de* (1) *alrededor de* $u_0 \in X$ *si se considera un espacio funcional* $(X, \| \|)$ *adecuado.*

4.1 SOLUCIÓN APROXIMADA MEDIANTE CK

Sabiendo que el problema $\begin{cases} u_t = u(1+u_x) & \forall t > 0 \\ u(x,0) = e^x \end{cases}$ tiene solución analítica alrededor del origen, ya que $G(x,t,u,u_x) = u(1+u_x)$ y $\varphi_0(x) = e^x$ lo son, es posible utilizar la ecuación y la condición inicial para calcular los coeficientes del desarrollo de Taylor de u alrededor de $(0,0)$. Resultan $u(0,0) = 1, u_x(0,0) = u_{x\ldots x}(0,0) = 1$, $u_t(0,0) = u(1+u_x)\big|_{(0,0)} = 2$, $u_{xt}(0,0) = u_x(1+u_x) + uu_{xx}\big|_{(0,0)} = 3$, $u_{tt}(0,0) = u_t(1+u_x) + uu_{xt}\big|_{(0,0)} = 7$, etc. El polinomio de Taylor de u de orden 4 alrededor de (0,0) es, entonces, $p_4(x,t) = 1 + x + 2t + \frac{1}{2}x^2 + 3xt + \frac{7}{2}t^2 + \frac{1}{6}x^3 + 8xt^2 + \frac{5}{2}x^2t + 7t^3 + \frac{1}{24}x^4 + \frac{3}{2}x^3t + 10x^2t^2 + \frac{133}{6}xt^3 + \frac{47}{3}t^4$

4.2 MÉTODO DE ADOMIAN

4.2.1 CON TÉRMINO LINEAL NO NULO

El problema puede escribirse de la forma (1) con $Lu = u$, $N(u) = uu_x$ y $f(x) = e^x$.

El problema libre $\begin{cases} u_t = u & \forall t > 0 \\ u(x,0) = e^x \end{cases}$ tiene solución $u_0(x,t) = E(t)f(x) = e^{tI}e^x = e^{x+t}$.

Los polinomios de Adomian, mediante (4), resultan $A_0 = u_0 u_{0_x}$, $A_1 = u_{0_x}u_1 + u_0 u_{1_x}$, $A_2 = u_{0_x}u_2 + u_1 u_{1_x} + u_0 u_{2_x}$, $A_3 = u_{0_x}u_3 + u_1 u_{2_x} + u_{1_x}u_2 + u_0 u_{3_x}$, etc..

Empleando la fórmula de recurrencia (3) se obtienen $u_0 = e^{x+t}$, $u_1 = e^{2x+t}(e^t - 1)$, $u_2 = \frac{3}{2}e^{3x+t}(e^t-1)^2$, $u_3 = \frac{8}{3}e^{4x+t}(e^t-1)^3$, etc. Observando la regularidad de los términos, se propone la forma general $u_n = \frac{(n+1)^{n-1}}{n!}e^{(n+1)x+t}(e^t-1)^n$ cuya validez se prueba por inducción.

Para $n = 4$, por ejemplo, se obtiene la solución aproximada

$$\psi_4(x,t) = e^{x+t}[1 + e^x(e^t-1) + \tfrac{3}{2}e^{2x}(e^t-1)^2 + \tfrac{8}{3}e^{3x}(e^t-1)^3 + \tfrac{125}{24}e^{4x}(e^t-1)^4]$$

4.2.2 CON TÉRMINO LINEAL NULO

El problema también podría escribirse de la forma (1) con $Lu = 0$, $N(u) = u(1+u_x)$ y $f(x) = e^x$.

El problema libre ahora resulta $u_t = 0$ para t>0, con $u(x,0) = x$, y su solución es $u_0(x,t) = e^{t0}e^x = e^x$

Los polinomios de Adomian son: $A_0 = u_0(1+u_{0_x})$, $A_1 = u_1 + u_{0_x}u_1 + u_0 u_{1_x}$, $A_2 = u_2 + u_{0_x}u_2 + u_1 u_{1_x} + u_0 u_{2_x}$, $A_3 = u_3 + u_{0_x}u_3 + u_1 u_{2_x} + u_{1_x}u_2 + u_0 u_{3_x}$, etc.

Empleando la fórmula (3) se tiene: $u_0 = e^x$, $u_1 = (e^x + e^{2x})t$, $u_2 = (\frac{1}{2}e^x + \frac{3}{2}e^{2x} + \frac{3}{2}e^{3x})t^2$, $u_3 = (\frac{1}{6}e^x + \frac{7}{6}e^{2x} + 3e^{3x} + \frac{8}{3}e^{4x})t^3$, etc., y para $n = 4$ se obtiene la solución aproximada

$$\varphi_4(x,t) = e^x[1 + (1 + e^x)t + (\tfrac{1}{2} + \tfrac{3}{2}e^x + \tfrac{3}{2}e^{2x})t^2 + (\tfrac{1}{6} + \tfrac{7}{6}e^x + 3e^{2x} + \tfrac{8}{3}e^{3x})t^3 + \\ + (\tfrac{1}{24} + \tfrac{5}{8}e^x + \tfrac{25}{8}e^{2x} + \tfrac{20}{3}e^{3x} + \tfrac{125}{24}e^{4x})t^4]$$

5. CONCLUSIONES

Elegimos un dominio pequeño alrededor de (0,0), y medimos la bondad de cada una de las aproximaciones propuestas, para cuatro términos, observando la norma 2 del error relativo. En las Figuras 1 y 2 se grafican las aproximaciones (en color azul, la solución exacta que provee Mathematica®).
En la Figura 3 se representan los errores relativos para $(x,t) \in [0, 0.5] \times [0, 0.2]$ (en color azul, el que corresponde a la solución obtenida por el Teorema de CK; en verde, el obtenido con ADM con *L* nulo y en rojo, con *L* no nulo).

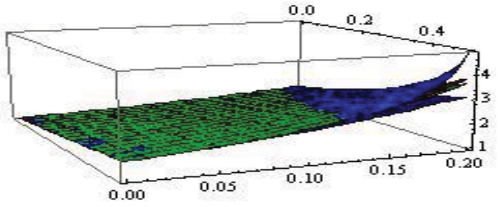


Figura 1

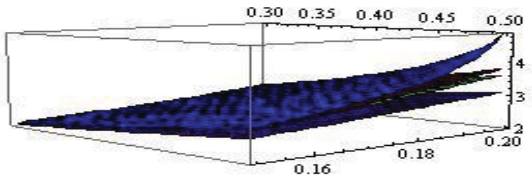


Figura 2

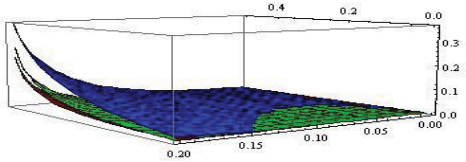


Figura 3

El menor error corresponde a la aproximación que propone ADM considerando el operador lineal *L* no nulo (sección 4.2.1). Esta ventaja se debe principalmente a que el término que contiene al operador *L* se "invierte" exactamente, restando solamente aproximar la parte de la solución correspondiente a *N*. ADM aproxima además la solución desarrollando *N* globalmente, en serie generalizada alrededor de u_0, y no alrededor de un punto, como en el caso de CK. Debe señalarse que es el teorema de CK el que garantiza la existencia y unicidad de soluciones analíticas del PVI que el método de descomposición de ADM intenta aproximar.
Todas las soluciones aproximadas son locales.

6. REFERENCIAS

[1] A. ABDELRAZEC and D. PELINOVSKY, *Convergence of the Adomian Decomposition Method for Initial-Value Problems*, Numerical Methods for Partial Differential Equations, n/a. doi: 10.1002/num.20549, 2009.
[2] G. ADOMIAN, *Stochastic Systems*, Academic Press, New York, NY, 1983.
[3] G. ADOMIAN, *Solving Frontier Problems of Physics: The Decomposition Method*, Kluwer Academic Publishers, Dordrecht, 1994.
[4] Y. CHERRUAULT, *Convergence of Adomian's Method*, Kybernetes 18, N°2 (1989), pp. 31-38.
[5] R. COOKE, *The Cauchy-Kovalevskaya Theorem*, conferencia dictada en Lisboa, 2001, disponible on line en http://www.cems.uvm.edu/~cooke/ckthm.pdf; consultado el 14 de noviembre de 2010.
[6] L. EVANS, *Partial Differential Equations*, Providence: American Mathematical Society, 1998.
[7] G. FOLLAND, *Introduction to partial differential equations*, N.J. : Princeton Univ. Press, 1995.
[8] M. J. FREITAS de SOUSA BASTO, *Adomian Decomposition Method, Nonlinear Equations and Spectral Solutions of Burgers Equation*, Tese, Facultade de Engenharia da Universidade de Porto, 2006.
[9] N. HIMOUN, K. ABBAOUI and Y. CHERRUAULT, *New results on Adomian method*, Kybernetes 32, Nº 4 (2003), pp. 523-539.
[10] R. RACH, *A new definition of the Adomian polynomials*, Kybernetes 37, N° 7 (2008), pp. 910-955.

ROGUE WAVES AND DISSIPATION

Constance Schober and Alvaro Islas

Department of Mathematics, University of Central Florida, Orlando, FL 32826, www.math.ucf.edu

Abstract: We investigate the effects of dissipation on the development of rogue waves and downshifting by adding nonlinear and linear damping terms to Gramstad's and Trulsen's new hamiltonian higher order nonlinear Schrödinger (HONLS) equation. Irreversible downshifting occurs when nonlinear damping is the dominant damping effect. Rogue waves do not develop after the downshifting becomes permanent. Thus permanent downshifting serves as an indicator that damping is sufficient to prevent the further development of rogue waves. Using the inverse spectral theory of the NLS equation, simulations of the damped HONLS equations for sea states characterized by JONSWAP spectrum consistently show that rogue wave events are well predicted by proximity to homoclinic data, as measured by the spectral splitting distance. The cut off distance decreases as the strength of the damping increases, indicating that for stronger damping the JONSWAP initial data must be closer to homoclinic data for rogue waves to occur.

Keywords: *rogue waves, downshifting, nonlinear Schrodinger equation*
2000 AMS Subject Classification: 35Q55

1 INTRODUCTION

Rogue waves are rare transient large amplitude waves whose heights are significantly larger than the background sea. Although the concept of rogue waves originated in the study of water waves, recently it has developed relevance in other applications, e.g. in optics, in Bose-Einstein condensates, and in superfluids. In water waves, the development of rogue waves in deep water is often attributed to the Benjamin-Feir (BF) instability and a nonlinear focusing of uncorrelated waves in a very localized region of the sea [7]. Given that the BF instability is described to leading order by the focusing NLS equation

$$iu_t + u_{xx} + 2|u|^2 u = 0, \tag{1}$$

Peregrine suggested that rogue waves in deep water may be related to the excitation of breather-type solutions of the NLS equation [7]. Subsequent studies showed homoclinic orbits of unstable plane wave solutions of the NLS equation can be used to effectively model rogue waves [4, 14]. However these solutions represent rogue waves in the simplest of cases since a typical background state is not uniform and the wave dynamics is only approximated by the NLS equation.

A more accurate description of the dynamics is obtained by retaining higher order terms in the asymptotic expansion for the surface wave displacement. A commonly used higher order NLS equation is the Dysthe equation [5]. Homoclinic orbits of the unstable Stokes wave have been shown to be robust to the higher order corrections in the Dysthe equation (see Fig. 1a). Laboratory experiments conducted in conjunction with numerical simulations of the Dysthe equation established that the generic long-time evolution of initial data near an unstable Stokes wave with two or more unstable modes is chaotic [2, 3]. Subsequent studies of the Dysthe equation showed that for a rather general class of such initial data, the modulational instability leads to high amplitude waves, structurally similar to the optimal phase modulated homoclinic solutions of the NLS equation, rising intermittently above a chaotic background [4, 14]. These earlier studies relating homoclinic solutions and rogue waves ignored the fact that the Dysthe equation is not Hamiltonian and neglected damping which, even when weak, can have a significant effect on the wave dynamics.

2 DAMPED HIGHER ORDER NONLINEAR SCHRÖDINGER MODELS

Recently, Gramstad and Trulsen used the Zakharov equation enhanced with the Krasitskii kernel to bring the Dysthe equation into Hamiltonian form, obtaining a new higher order nonlinear Schrödinger (HONLS) equation [6]. Here we investigate the effects of dissipation on the development of rogue waves and downshifting by adding nonlinear and linear damping to this new HONLS equation as follows:

$$iu_t + u_{xx} + 2|u|^2 u + i\Gamma u + i\epsilon\left(\frac{1}{2}u_{xxx} - 8|u|^2 u_x - 2ui(1+i\beta)\left[H\left(|u|^2\right)\right]_x\right) = 0, \tag{2}$$

where $H(f)$ represents the Hilbert transform of f. Periodic boundary conditions, $u(x,t) = u(x+L,t)$, are considered. Throughout this paper "HONLS equation" refers to (2) with $\Gamma = 0$ and $\beta = 0$. The Hamiltonian for the HONLS eq. is given by

$$\mathcal{H} = \int_0^L \left\{ -i|u_x|^2 + i|u|^4 - \frac{\epsilon}{4}\left(u_x u_{xx}^* - u_x^* u_{xx}\right) + 2\epsilon|u|^2\left(u^* u_x - u u_x^*\right) + i\epsilon|u|^2 H\left[\left(|u|^2\right)_x\right] \right\} dx. \quad (3)$$

Uniform linear damping occurs for $\Gamma > 0$ and $\beta = 0$ and has been used extensively to compare physcial wave-tank experiments with damped wave trains studies [15]. Localized nonlinear damping of the mean flow occurs for $\epsilon, \beta > 0$ and $\Gamma = 0$ [11, 10].

The mass or wave energy, E , and the momentum or total energy flux, P, are defined by $E = \int_0^L |u|^2\, dx$, and $P = i\int_0^L \left(u^* u_x - u u_x^*\right) dx$. Using the Fourier spectrum of $u(x,t)$, $\hat{u}_k$, two different choices of diagnostic frequencies can be used to measure downshifting. On the one hand, the dominant mode or spectral peak intuitively corresponds to the k for which $|\hat{u}_k|$ achieves its maximum and is denoted as k_{peak}. On the other hand, Uchiyama and Kawahara defined the wave number for the spectral center or mean frequency of the spectrum as: $k_m = -\frac{1}{2}P/E$. The wave train is understood to experience a permanent frequency downshift when the spectral center k_m decreases monotonically in "time" or there is a permanent downshift of the spectral peak k_{peak} [16].

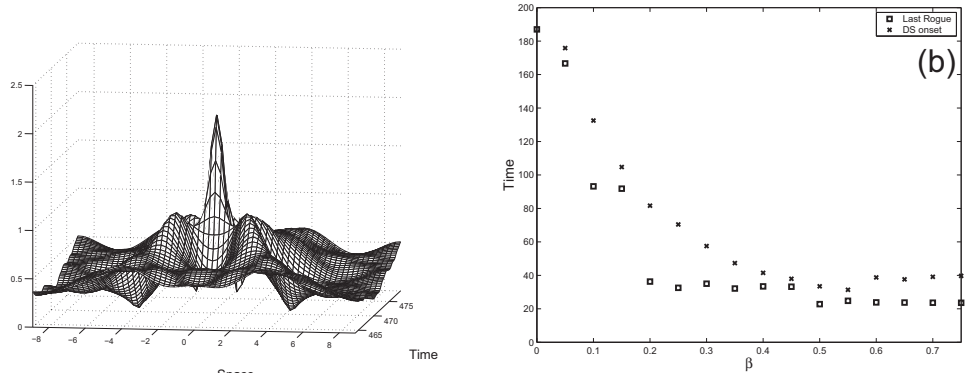


Figure 1: (a) Rogue wave solution of the Dysthe equation. (b) Damped HONLS eqn., $\epsilon = 0.05, \Gamma = 0$ and $0 < \beta < 0.75$. The time downshifting is irreversible (x) and the time the last rogue wave occurs (box) as a function of β, $u_0 = a(1 + 0.01\cos\mu x)$, averaged over $0.57 < a < 0.67$.

3 Relation of rogue waves and downshifting in the damped HONLS equation

Downshifting in the evolution of nearly uniform plane waves was first observed by Lake *et. al.* [12] and confirmed by other laboratory experiments, e.g. [8]. Lake's experiments showed that after the wavetrain becomes strongly modulated, it recurs as a nearly uniform wavetrain with the dominant frequency permanently downshifted. Our studies of rogue waves show that an ϵ-neighborhood of the unstable plane wave (the same regime in which Lake examined downshifting) is effectively a rogue wave regime for the HONLS equation since the likelihood of obtaining a rogue wave is extremely high [10]. Using this unstable modulated wavetrain initial data we examined the occurence of irreversible downshifting in our damped HONLS equation and what characterizes the damped wave train evolution on a short and long time scale.

We find that rogue waves may emerge in both the linear and nonlinear damped regimes that were not present without damping. Although damping decreases the growth rates of the individual modes, the modes may coalesce due to changes in their focusing times, thus resulting in larger waves in the damped regime. Even so, on average, the strength is smaller and fewer rogue waves occur when damping is present [10]

Since the nonlinear β-term is large only near the crest of the envelope, the damping is localized when the wavetrain is strongly modulated. Due to the BF instability irreversible downshifting occurs when the nonlinear damping is the dominant damping effect. In particular, when only nonlinear damping is present, permanent downshifting occurs for all values of the nonlinear damping parameter β, appearing abruptly

for larger values of β. Significantly, we find that rogue waves do not develop in the nonlinearly damped evolution after permanent downshifting occurs for any other pair of parameter values (a,β) considered in our experiments. This is summarized in Fig. 1b which compares, for $0 < \beta < 0.75$, the average time of the last rogue wave (box) with the average time at which downshifting is irreversible (x), where the averages are over the six simulations with initial data amplitude $0.57 < a < 0.67$. Thus permanent downshifting serves as an indicator that there has been sufficient cumulative damping to inhibit the further development of rogue waves (see Fig. 1b).

4 Damped random oceanic sea states

In this section we examine rogue waves in the presence of damping for sea states characterized by the Joint North Sea Wave Project (JONSWAP) spectrum. We consider initial data for the surface elevation to be of the form [13] $\eta(x,0) = \sum_{n=1}^{N} C_n \cos(k_n x - \phi_n)$, where the random phases ϕ_n are uniformly distributed on $(0, 2\pi)$. The spectral amplitudes, $C_n = \sqrt{2S_n/L}$, are obtained from the JONSWAP spectrum [9]. Earlier studies sought to relate the occurrence of rogue waves to the paarameters in the JONSWAP spectrum or to the Benjamin-Feir index [13]. These approaches were not sufficient to explain rogue wave generation as they did not take into account the phase information used to reconstruct η.

In a recent study we used the Floquet decomposition of an ensemble of JONSWAP initial data, which takes the phase information ϕ_n into account, to develop a novel criterion for predicting the occurrence and strength of rogue waves [9]. This criterion is based on the inverse spectral theory of the NLS equation which arises as the compatibility condition of a pair of linear systems:

$$\mathcal{L}^{(x)}\phi = \begin{pmatrix} D + i\lambda & -u \\ u^* & D - i\lambda \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = 0, \quad \mathcal{L}^{(t)}\phi = 0, \tag{4}$$

where D denotes the derivative with respect to x, λ is the spectral parameter and ϕ is the eigenfunction [17]. For periodic boundary conditions, $u(x+L,t) = u(x,t)$, the spectrum $\sigma(u) = \{\lambda \in \mathbb{C} \mid \Delta(u,\lambda) \in \mathbb{R}, \ |\Delta(u,\lambda)| \leq 2\}$, of the linear operator $\mathcal{L}^{(x)}$, is given in terms of the *Floquet discriminant*, $\Delta(u,\lambda) =$ Trace $[M(x+L;u,\lambda)]$, with $M(x;u,\lambda)$ a fundamental solution matrix of (4) [1]. Within the discrete spectrum ($\Delta(\lambda,u) = \pm 2$) one identifies *simple points* ($\Delta' \neq 0$) and double points ($\Delta' = 0$ and $\Delta'' \neq 0$). Typically, simple points and real double points are associated with stable degrees of freedom, while complex double points correspond to unstable modes and label the corresponding homoclinic orbits.

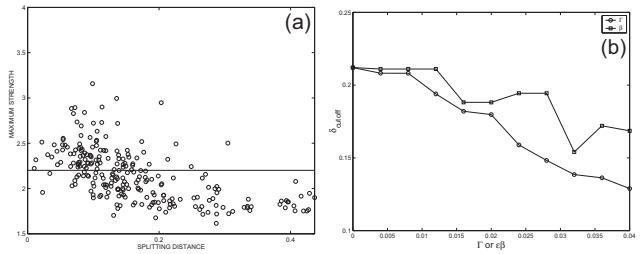


Figure 2: (a) Maximum strength vs. $\delta(\lambda_+,\lambda_-)$ for the HONLS equation ($\epsilon = 0.05$) with nonlinear damping $\beta = 0.5$, $\Gamma = 0$. (b) The damped HONLS equation ($\epsilon = 0.05$): δ_{cutoff}^{damped} as a function of Γ (circles) and $\epsilon\beta$ (boxes)

We used the "splitting" distance between two consecutive simple points, $\delta(\lambda_+,\lambda_-) = |\lambda_+ - \lambda_-|$, to measure the proximity in the spectral plane to homoclinic data, i.e. to complex double points and their corresponding instabilities. JONSWAP data was found to correspond to semi-stable N-phase solutions, i.e. perturbations of unstable N-phase solutions and their associated homoclinic orbits. For fixed α and γ the spectral distance δ varies significantly with the random phases [9]. Further, irrespective of the values of α and γ, in simulations of the NLS and HONLS equations we find that extreme waves develop for JONSWAP initial data that is "near" NLS homoclinic data, whereas the JONSWAP data that is "far" from NLS homoclinic data typically does not generate extreme waves. In [9, 14], thousands of simulations of both the NLS and HONLS equations using JONSWAP initial data consistently show that rogue wave events are well predicted by proximity to homoclinic data, as measured by δ.

Figure 2 shows the maximum wave strength vs. the splitting distance $\delta(\lambda_+, \lambda_-)$ for 250 random simulations of the damped HONLS equation ($\epsilon = 0.05$, $\beta = 0.5, \Gamma = 0$). JONSWAP initial data was used for different (γ, α) pairs with $\gamma = 1, 2, 4, 6, 8$ and $\alpha = 0.008, 0.012, 0.016, 0.02$ and randomly generated phases. Each circle represents the maximum strength attained during one simulation, $0 < t < 20$. Compared with earlier HONLS rresults, the strength and likelihood of rogue waves occuring in a given simulation is typically smaller when damping is present. A cutoff vaue of δ exists such that rogue waves typically do not occur if $\delta > \delta_{cutoff}$. We determine δ_{cutoff} by requiring that 95% of the rogue waves occur for $\delta < \delta_{cutoff}$. For example, the cutoff values for the linear ($\Gamma = 0.02$) and nonlinear ($\epsilon\beta = 0.02$) damped HONLS equation are $\delta_{cutoff}^{damped} \approx 0.16, 0.17$, respectively, which are less than the cutoff value for the undamped HONLS equation where, $\delta_{cutoff}^{undamped} \approx 0.2$. This implies that when damping is present the JONSWAP initial data must be closer to homoclinic data and instabilities to obtain rogue waves.

Figure 2 shows δ_{cutoff}^{damped} as a function of the damping parameters Γ (circles) and $\epsilon\beta$ (boxes). For each of the ten values of Γ and of $\epsilon\beta$, $0 < \Gamma, \epsilon\beta < 0.04$, 250 simulations of the damped HONLS equation were carried out. We find that δ_{cutoff}^{damped} is generally decreasing as the strength of the damping increases. The cutoff value in δ for the nonlinearly damped case is not monotonically decreasing as atypical cases exist where a small increase in damping may be offset by a coalescence of the modes due to changes in the focusing times. The overall decay in δ_{cutoff}^{damped} indicates that for rogue waves to occur the JONSWAP initial data must lie in a shrinking neighborhood of the homoclinic data. Thus the proximity to instabilities and homoclinic data is more essential for the development of rogue waves when damping is present.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation, Grant # NSF-DMS0608693.

REFERENCES

[1] ABLOWITZ, M. J. AND SEGUR, H., *Solitons and the Inverse Scattering Transform*, SIAM, 1981.

[2] ABLOWITZ, M. J., HAMMACK, J., HENDERSON, D., AND SCHOBER, C. M., *Modulated Periodic Stokes Waves in Deep Water*, Phys. Rev. Lett., 84 (2000), pp.887-890.

[3] ABLOWITZ, M. J., HAMMACK, J., HENDERSON, D., AND SCHOBER, C. M., *Long time dynamics of the modulational instability of deep water waves*, Physica D, 152–153 (2001), pp.416-433.

[4] CALINI, A. AND SCHOBER, C. M., *Homoclinic chaos increases the likelihood of rogue waves*, Phys. Lett. A, 298 (2002), pp.335-349.

[5] DYSTHE, K., *Note on a modification to the nonlinear Schrodinger equation for deep water*, Proc. R. Soc. Lon. Ser.-A, 369 (1979), pp.105-114.

[6] GRAMSTAD, O. AND TRULSEN, K., *Hamiltonian form of the modified nonlinear Schrödinger equation for gravity waves on arbitrary depth*, J. Fluid Mech., in press, 2011.

[7] HENDERSON, K. L., PEREGRINE, D. H., AND DOLD, J. W., *Unsteady water wave modulations: fully nonlinear solutions and comparison with the nonlinear Schrödinger equation*, Wave Motion, 29 (1999), pp.341-361.

[8] HUANG, N., LONG, S., AND SHEN, Z., *The mechanism for frequency downshift in nonlienar wave evolution*, Adv. Appl. Mech., 32 (1996), pp.59-117.

[9] ISLAS, A. AND SCHOBER, C. M., *Predicting rogue waves in random oceanic sea states*, Phys. Fluids, 17 (2005), pp.1-4.

[10] ISLAS, A. AND SCHOBER, C. M., *Rogue waves, dissipation, and downshifting*, Physica D, accepted, 2010.

[11] KATO, Y. AND OIKAWA, M., *Wave number down-shift in Modulated Wavetrain through a Nonlinear Damping effect*, J. Phys. Soc. Jpn, 64 (1995), pp.4660-4669.

[12] LAKE, B., YUEN, H., RUNGALDIER, H., AND FERGUSON, W., *Nonlinear deep-water waves:theory and experiment. Part 2. Evolution of a continuous wave train*, J. Fluid Mech., 83 (1977), pp.49-74.

[13] ONORATO, M., OSBORNE, A., SERIO, M. AND BERTONE, S., *Freak waves in random oceanic sea states*, Phys. Rev. Lett., 86 (2001), pp.5831-5834.

[14] SCHOBER, C. M., *Melnikov analysis and inverse spectral analysis of rogue waves in deep water*, Eur. J. Mech. B-Fluids, 25(5), 602–620, doi:10.1016/j.euromechflu.2006.02.005, 2006.

[15] SEGUR, H., HENDERSON, D., CARTER, J., HAMMACK, J., LI, C., PHEIFF, D., AND SOCHA, K., *Stabilizing the Benjamin-Feir Instability*, J. Fluid Mech., 539 (2005), pp.229-271.

[16] TRULSEN, K. AND DYSTHE, K., *Frequency down-shift in three-dimensional wave trains in a deep basin*, J. Fluid Mech., 352 (1997), pp.359-373.

[17] ZAKHAROV, V. E. AND SHABAT, A. B., *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Sov. Phys. JETP, 34 (1972), pp.62-69.

ECUACIONES DE EVOLUCIÓN PARA UN CASO SEMILINEAL DE MEMBRANAS ACOPLADAS

Peñas Galezo Ramiro†

†*Universidad del Atlántico, Barranquilla, Colombia, rpegazo@yahoomail.es*

Resumen: En este documento se desrrollan las ecuaciones de evolución que modela dos membranas elasticas acopladas con fuerzas que dependen del tiempo, la posición y velocidad. La existencia y unicidad de soluciones se prueban con la teoría de semigrupo de operadores fuertemente continuos.

Palabras clave: *ecuaciones de evolución, membranas acopladas, caso semilineal*
2000 AMS Subject Classification: 21A54 - 55P54

1. ECUACIONES DEL MODELO

El problema de evolución del que trata este trabajo corresponde a las ecuaciones que modela un sistema de dos membranas acopladas de masas m_1 y m_2. En estado de reposos las membranas ocupan regiones $\overline{\Omega_1}$ y $\overline{\Omega_2}$ los cuales son dominios en el plano donde $\partial\Omega_1$ y $\partial\Omega_2$ son de clase C^1. El desplazamiento de cada membrana a uno y otro lado de su estado de reposos se representará con $\mathbf{u}$, la cual es una función de dos variables (x e y). Sobre el plano de cada membrana se aplican tensiones por unidad de área $\mathbf{F}_1$, $\mathbf{F}_2$ y lateralmente sobre $\partial\Omega_1 \cup \partial\Omega_2$ actuan fuerzas por unidasd de longitud $\mathbf{f}_1$ y $\mathbf{f}_2$. Suponemos además que en una región $S \subset \partial\Omega_1 \cup \partial\Omega_2$ las membranas se encuentran fijas a un soporte.

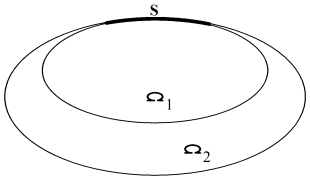


Figura 1: membranas acopladas

La formulación matemática requiere el principio del trabajo virtual ([1], 175-176), para membranas, .

$$\delta W_{ext} - \delta W_{int} = \int_\Omega \rho(x) \frac{\partial^2 \mathbf{u}(t,\mathbf{x})}{\partial t^2} \cdot \delta\mathbf{u}(\mathbf{x})\, dx \tag{1}$$

donde δW_{ext} corresponde al trabajo virtual de las fuerzas externas ([5], 61) definido por $\delta W_{ext} = \int_\Omega \mathbf{F} \cdot \delta\mathbf{u}\, dx + \int_{\partial\Omega} \mathbf{f} \cdot \delta\mathbf{u}\, ds$, y δW_{int} es el trabajo virtual de las fuerzas internas definido como $\delta W_{int} = \dfrac{d}{d\tau}\left[\dfrac{k}{2}\int_\Omega |\nabla(\mathbf{u}+\tau\delta\mathbf{u})|^2\, dx|_{\tau=0}\right]$, donde k es una constante que depende del material. En la ecuación (1) $\delta\mathbf{u}$ se conoce como desplazamiento virtual y $\rho(x)$ es la densidad del

material del que está compuesto la membrana que en adelante se asumira constante para cada material. También se limitará el problema al caso de ausencia de fuerzas laterales sobre las membranas.

Para los casos de estiramiento de membranas, los aportes del desplazamiento $\mathbf{u}$ en la dirección del plano de las membranas son despreciables ([4], pagina 687), así que en lugar de $\mathbf{u}_1$, $\mathbf{u}_2$ se tomarán campos escalares u_1, u_2, que denoten los desplazamientos verticales de las membranas, y definiremos los campos escalares F_1, F_2, en lugar de $\mathbf{F}_1$, $\mathbf{F}_2$, respectivamente. Es decir,

$$u_i: \begin{array}{l}[0,T]\times\Omega_i \to \mathbf{R} \\ (t,\mathbf{x}) \longmapsto u_i(t,\mathbf{x})\end{array}, \qquad F_i: \begin{array}{l}[0,T]\times\Omega_i \to \mathbf{R} \\ (t,\mathbf{x}) \longmapsto F_i(t,\mathbf{x},u_i(t,\mathbf{x}),\frac{\partial u_i}{\partial t}(t,\mathbf{x}))\end{array},$$

Debido a la configuración del problema se cuenta con las condiciones de frontera

$$\left.\begin{array}{c} u_1|_\Gamma = u_2|_\Gamma \text{ para todo } t\in[0,T], \text{ donde } \Gamma = \partial\Omega_1 \cap \partial\Omega_2 \\ u_1|_{\partial\Omega_1\cap S} = 0,\ u_2|_{\partial\Omega_1\cap S} = 0 \text{ para todo } t\in[0,T], \\ f_1 = 0,\ f_2 = 0 \end{array}\right\} \tag{2}$$

Las ecuaciones diferenciales que modelan el problema no homogeneo de membranas acopladas son

$$\left.\begin{array}{c} \rho_i \dfrac{\partial^2 u_i}{\partial t^2} - k_i \Delta u_i = F_i \text{ en } [0,T]\times\Omega_i,\ i=1,2 \\ k_i \nabla u_i \cdot v^i - f_i = 0 \text{ sobre } [0,T]\times\{\partial\Omega_i \setminus (\Gamma\cup S)\} \\ [k_1\nabla u_1 - k_2 \nabla u_2]\cdot v^1 - f_1 - f_2 = 0 \text{ sobre } [0,T]\times\Gamma\setminus S \end{array}\right\} \tag{3}$$

Los calculos han sido omitidos para brevedad del documento (ver [3])

Supondremos que las condiciones iniciales del problema son

$$u_i(0,\mathbf{x}) = g_i(\mathbf{x}); \frac{\partial u_i}{\partial t}(0,\mathbf{x}) = h_i(\mathbf{x}),\ \mathbf{x}\in\Omega_i\ i=1,2 \tag{4}$$

2. Formulación abstracta del problema

Sea $H^1(\Omega_i)$ el espacio de Sobolev $W^{1,2}(\Omega_i)$, y tomemos de $H^1(\Omega_i)$ aquellas funciones $u_i \in H^1(\Omega_i)$, $(i=1,2)$, cuya traza sobre $\partial\Omega_i \cap S$ es igual a cero[1], Se hace necesario definir los espacio $(V,(\cdot|\cdot)_V)$, $(H,(\cdot|\cdot)_H)$ como

$$\begin{aligned} V \ &: = \{(u_1,u_2) \mid u_i \in H^1(\Omega_i),\ \gamma_0 u_1|_\Gamma = \gamma_0 u_2|_\Gamma,\ \gamma_0 u_i|_{\partial\Omega_i\cap S} = 0, i=1,2\}, \\ &((u_1,u_2)|(v_1,v_2))_V = (u_1|v_1)_{H^1(\Omega_1)} + (u_2|v_2)_{H^1(\Omega_2)}, \end{aligned}$$

$$H := L^2(\Omega_1)\times L^2(\Omega_2);\ ((u_1,u_2)|(v_1,v_2))_H = (u_1|v_1)_{L^2(\Omega_1)} + (u_2|v_2)_{L^2(\Omega_2)},$$

Puede demostrarse que los espacios de Hilbert V y H son separables, V esta inmerso continuamente en H $(V \hookrightarrow H)$, y V es denso en H (Peñas [7] 30).

Multiplicando (3) por una función de prueba $w_i \in H^1(\Omega_i)$, e integrando, se tiene:

$$\int_{\Omega_i} \rho_i \frac{\partial^2 u_i}{\partial t^2} w_i dx - \int_{\Omega_i} k_i \Delta u_i w_i dx = \int_{\Omega_i} F_i w_i dx. \tag{5}$$

[1] $\gamma_0(u_1)|_{\partial\Omega_1\cap S} = 0$, $\gamma_0(u_2)|_{\partial\Omega_2\cap S} = 0$, donde $\gamma_0 : H^1(\Omega_i) \to L^2(\partial\Omega_i)$ es el operador traza en $H^1(\Omega_i)$.

y por la primera identidad de Green en el segundo término de (5) junto con (2) resulta

$$\sum_{i=1}^{2}\left(\rho_i\frac{\partial^2 u_i}{\partial t^2}|w_i\right)_{L^2(\Omega_i)}+\sum_{i=1}^{2}k_i\int_{\Omega_i}\nabla u_i\cdot\nabla w_i dx=((F_1,F_2))\,|\,(w_1,w_2))_H \tag{6}$$

Definamos para todo $(u_1,u_2),(w_1,w_2)\in V$

$$a\left((u_1,u_2)\,|\,(w_1,w_2)\right):=k_1\int_{\Omega_1}\nabla u_1\cdot\nabla w_1 dx+k_2\int_{\Omega_2}\nabla u_2\cdot\nabla w_2 dx$$

se tiene entonces que: $a\,(\cdot|\cdot)$ es una forma bilineal continua en V, coerciva ([7] página 34) y simétrica, la cual tiene una representación

$$a((u_1,u_2),(w_1,w_2))=(A(u_1,u_2)|(w_1,w_2))_H.$$

donde $-A$ es el generador infinitesimal de un semigrupo de operadores de clase $\mathcal{C}^0$ sobre V. ([2], 332).

Haciendo $u=(u_1,u_2)$, $w{:=}\,(w_1,w_2)$, $F=(F_1,F_2)$ de manera tal que si $u\in[0,T]\times V$, $u_i\in C^1(]0,T[)$,

$$\begin{aligned}u(t,\mathbf{x}) \;&:\; =(u_1(t,\mathbf{x}),u_2(t,\mathbf{x}));\quad w(\mathbf{x}):=(w_1(\mathbf{x}),w_2(\mathbf{x})),\\ &F(t,\mathbf{x},u(t,\mathbf{x}),\tfrac{\partial}{\partial t}u(t,\mathbf{x})):=\begin{pmatrix}F_1(t,\mathbf{x},u_1(t,\mathbf{x}),\frac{\partial u_1}{\partial t}(t,\mathbf{x}))\\ F_2(t,\mathbf{x},u_2(t,\mathbf{x}),\frac{\partial u_2}{\partial t}(t,\mathbf{x}))\end{pmatrix}^T,\end{aligned}$$

entonces (6) se puede representar como

$$\left(\frac{\partial^2}{\partial t^2}\Phi u|w\right)_H+a\,(u,w)=(F|w)_H\,,\ \text{donde } \Phi(u_1,u_2)=(\rho_1u_1,\rho_2u_2) \tag{7}$$

3. EXISTENCIA Y UNICIDAD DE SOLUCIONES DÉBILES.

Por razones técnicas se define el espacio $(\tilde{V},\|\cdot\|_{\tilde{V}})$ como

$$\begin{aligned}\tilde{V} \;&=\; \{(\tilde{u}_1,\tilde{u}_2)\in H^1(\Omega_1)\times H^1(\Omega_2):(\frac{1}{\sqrt{\rho_1}}\tilde{u}_1,\frac{1}{\sqrt{\rho_2}}\tilde{u}_2)\in V\},\\ \|(\tilde{u}_1,\tilde{u}_2)\|_{\tilde{V}} \;&=\; \left[\frac{1}{\rho_1}\|\tilde{u}_1\|^2_{H^1(\Omega_1)}+\frac{1}{\rho_2}\|\tilde{u}_2\|^2_{H^1(\Omega_2)}\right]^{\frac{1}{2}},\end{aligned}$$

y la forma bilineal $\tilde{a}$ sobre $\tilde{V}\times\tilde{V}$ definida así:

$$\tilde{a}((\tilde{u}_1,\tilde{u}_2),(\tilde{w}_1,\tilde{w}_2)):=a((\tfrac{1}{\sqrt{\rho_1}}\tilde{u}_1,\tfrac{1}{\sqrt{\rho_2}}\tilde{u}_2),(\tfrac{1}{\sqrt{\rho_1}}\tilde{w}_1,\tfrac{1}{\sqrt{\rho_2}}\tilde{w}_2)),$$

La forma bilineal $\tilde{a}$ permite que escribamos la ecuación (7) de una manera más conveniente: $\left(\frac{\partial^2\tilde{u}}{\partial t^2}|\tilde{w}\right)_H+\left(\tilde{A}\tilde{u}|\tilde{w}\right)=\left(\tilde{F}|\tilde{w}\right)_H$, donde $\tilde{A}$ es el generador infinitesimal que define la forma bilineal $\tilde{a}$, y donde $\tilde{F}(t,\cdot,\tilde{u}(t,\cdot),\partial_t\tilde{u}(t,\cdot))=\left(\ \frac{1}{\sqrt{\rho_1}}F_1(t,u_1,\frac{\partial}{\partial t}u_1),\quad \frac{1}{\sqrt{\rho_2}}F_2(t,u_2,\frac{\partial}{\partial t}u_2)\ \right)$.

La formulación variacional del problema de membranas acopladas implica *encontrar* $\tilde{u}$ *que satisfaga*

$$\frac{d^2}{dt^2}\tilde{u}(t) + \tilde{A}\tilde{u}(t) = \tilde{F}(t, \tilde{u}(t), \partial_t\tilde{u}(t)) \; en \; H. \tag{8}$$

Con el cambio de notación: $\mathbf{u} = (u, \partial_t u) \in V \times H$; $\tilde{\mathbf{F}}(t,\tilde{\mathbf{u}}) = \left(\mathbf{0}, \quad \tilde{F}(t, \cdot, \tilde{u}(t, \cdot), \partial_t\tilde{u}(t, \cdot))\right)$, nuestro problema de segundo orden se convierte en un problema de primer orden que consiste en encontrar $\tilde{\mathbf{u}} = \left(\tilde{u}, \frac{d}{dt}\tilde{u}\right) \in \tilde{V} \times H$, que satisfaga

$$\frac{d\tilde{\mathbf{u}}(t)}{dt} + \tilde{\mathbf{A}}\tilde{\mathbf{u}}(t) = \tilde{\mathbf{F}}(t, \tilde{\mathbf{u}}(t)), \; t > t_0 \\ \tilde{\mathbf{u}}(0) = \tilde{\mathbf{u}}_0 \quad , \text{ donde } \tilde{\mathbf{A}}\tilde{\mathbf{u}}(t) = \begin{pmatrix} -\partial_t\tilde{u} \\ \tilde{A}\tilde{u} \end{pmatrix}, \; \tilde{\mathbf{u}}(0) = \begin{pmatrix} \tilde{g}_1 & \tilde{g}_2 \\ \tilde{h}_1 & \tilde{h}_2 \end{pmatrix}$$

Lema 1 *Para que* $\tilde{\boldsymbol{F}}$ *:* $[0,T] \times \tilde{V} \times H \to \tilde{V} \times H$*, sea continuamente diferenciable sobre* $[0,T] \times \tilde{V} \times H$ *con derivadas parciales acotadas, son condiciones suficientes*

1. *Para todo* $t \in [0,T]$, $(i = 1, 2)$, $F_i(t, x, u_i(x), \frac{\partial}{\partial t}u_i(x))$ *es medible si* $u_i, \frac{\partial u_i}{\partial t} : \Omega_i \to \mathbf{R}$ *son medibles.*

2. *Existen las derivadas parciales de* $F_i\,(t, x, u, v)$ *con respecto a* t, u, v *para todo* $(t, x, u, v) \in [0,T] \times \Omega_i \times \mathbf{R} \times \mathbf{R}$*, y además son Lipschitz continua sobre* $[0,T] \times \Omega_i \times \mathbf{R} \times \mathbf{R}$.

3. F_i *sea lineal en la cuarta componente, así que* $F_i\,(t, x, u, v) = \partial_v F_i\,(t, x, u, v)\, v$ *para todo* $(t, x, u, v) \in [0,T] \times \Omega_i \times \mathbf{R} \times \mathbf{R}$.

Teorema 1 *Si* $F(t)$ *satisface las condiciones del lema anterior,* $(g_1, g_2) \in D(A)$*, y* $(h_1, h_2) \in V$*, entonces exite una única solución debil del problema de membranas acopladas. (Ver prueba en Pazy [6], pagina 187).*

4. REFERENCIAS

REFERENCIAS

[1] **DAUTRAY Robert, LIONS Jackes.** Mathematical Analisis and numerical methods for science and technology. Vol I. Springer Verlag. 2000.

[2] **DAUTRAY Robert, LIONS Jackes.** Mathematical Analisis and numerical methods for science and technology. Vol V. Springer Verlag. 2000.

[3] **HERNANDEZ, Jairo.** Modelos Matemáticos Para la Deformación de Placas y Membranas. Universidad del Valle. Barranquilla.1997

[4] **KREYSZIG, Erwin.** Matemáticas avanzadas para ingenieros. Volumen II. Mexico: Limusa. 1976.

[5] **NEČAS Jindřich-HLAVÁČEK Ivan.** Mathematical Theory Of Elastic And Elasto-Plastic Bodies. Elsevier Scientific Publishing Co. 1981.

[6] **PAZY A.** Semigroups of linear operators and aplications to partial differential equations. Springer Verlag. New York. 1983.

[7] **PEÑAS Ramiro**, Ecuaciones de evolución para membranas acopladas. Tesis de Maestría, Universidad Nacional. 2004.

ON CERTAIN ASPECTS OF A *Solid Combustion Model*

Alejandro Omón Arancibia[þ]

[þ]*Departamento de Ingeniería Matemática, Universidad de La Frontera, Temuco-Chile, aomon@ufro.cl*

Abstract: The stability of a *Solid Fuel Model*, which represents a thermal reaction of a solid material is studied. The model is given by two reaction diffusion equations, coupled in the reaction term, and with different boundary conditions. A strong bifurcation criteria for the steady problem is presented, the same as blow-up estimates for the unsteady case. Numerical trials are presented for the different regimes to be presented.

Keywords: *Solid Combustion Model, Blow-up, Nonlinear Eigenvalue Problem.*
2000 AMS Subject Classification: 35K45, 35J57, 35B34 and 35B32

1 INTRODUCTION

The one step irreversible reaction

$$\nu_F F + \nu_O O \to \lambda_P P \ , \tag{1}$$

models many phenomena in disciplines like Chemistry or Biology.

If now in reaction (1) F represents fuel, O an oxidant and P a final product, with ν_F, ν_P, λ_P being the stoichiometric coefficients, (1) models a combustion process.

By the *Method of Activation Energy Asymptotic*, and after some assumptions and simplifications on the Physics and Chemist behind the modeling of (1), it is obtained as a dynamical description of a solid combustion process, known in the literature as *Solid Fuel Model System* (SFM), the system of reaction diffusion equations given by

$$\frac{\partial \theta}{\partial t} - \Delta\theta = \delta\,(1-\epsilon c)^m\, e^{\theta/(1+\epsilon\theta)} \quad \text{in } \Omega\times]0,\infty[, \tag{2}$$

$$\frac{\partial c}{\partial t} - \beta\,\Delta c = \nu\,(1-\epsilon c)^m\, e^{\theta/(1+\epsilon\theta)} \quad \text{in } \Omega\times]0,\infty[, \tag{3}$$

$$\theta(0,x) = \theta_0(x) \quad \text{in } \Omega, \tag{4}$$

$$c(0,x) = c_0(x) \quad \text{in } \Omega, \tag{5}$$

$$\theta(t,x) = 0 \quad \text{on } \partial\Omega\times]0,\infty[, \tag{6}$$

$$\frac{\partial c}{\partial n}(t,x) = 0 \quad \text{on } \partial\Omega\times]0,\infty[, \tag{7}$$

where θ represents a temperature distribution, c corresponds to the amount of consumed solid fuel and $\nu = \epsilon\,\delta\,\Gamma$ are chemical parameters. Reference **[4]** provides a more accurate deduction of the system.

Although very extended in the literature, this system has not been studied in extensive. In fact, it has always been simplified and reduced to only one equation or to a one dimension space model, see for example **[10]**-**[5]**-**[1]**. Also, it can be observed in **[4]** that the existence of solutions for (SFM) is strongly limited to the geometry of the domain, in particular to its convexity for ensuring *Invariance Principle* for

the time evolution, see also **[3]**.

There will be presented new results that lift the geometrical restriction on the domain, which study different regimes for the evolution of (SFM) with respect to the stoichiometric coefficients, but also on the initial conditions on (SFM). It is observed that as almost all known results for (SFM) are based in the *Invariance Principle*, then neither the parameters nor the initial conditions play any role which physically has no sense.

To motivate the previous paragraph let us observe figures 1 and 2, which have the same initial condition for both dependent variables θ and c, and almost the same set of parameters except the one associated to the power term of the reaction. It can be proved that in figure 1 there is a global solutions that converges to a steady state, while in figure 2 there is blow-up. It is remarked the difference in time scale of the two figures, which also shows how fast the blow-up can be.

Also, interesting information about the steady solutions for (SFM) will be given, the same as some numerical trials that illustrate the presented results.

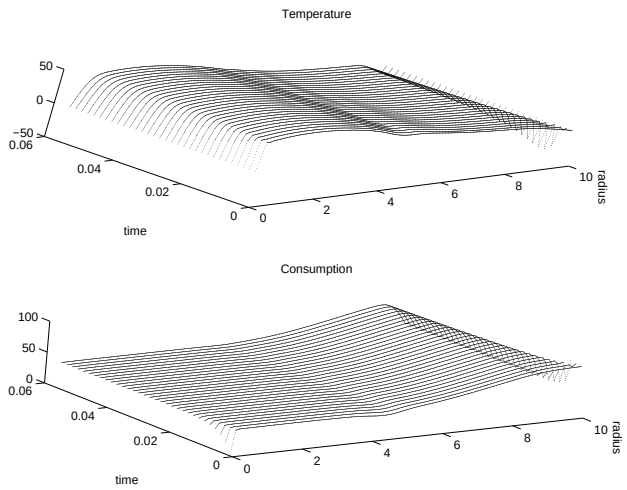


Figure 1: Parameters: $m = 3$, $\delta = 1$, $\gamma = 1$, $\beta = 1$ and $\epsilon = 0.025$.

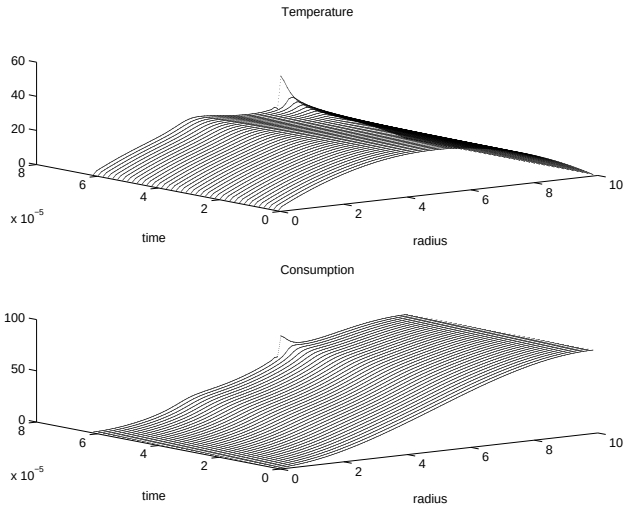


Figure 2: Parameters: $m = 2$, $\delta = 1$, $\gamma = 1$, $\beta = 1$ and $\epsilon = 0.025$.

ACKNOWLEDGMENTS

The author would like to thanks *Vice-rectoría de Investigación* and *Dirección de Investigación* both at Universidad de La Frontera, and the anonymous referee on reference [9].

REFERENCES

[1] M. AL-REFAI, *Existence, uniqueness and bounds for a problem in combustion theory*, J. Comp. Appl. Maths., vol. 167 (2004), pp. 255-269.

[2] M. AL-REFAI, *Bounds and critical parameters for a combustion problem*; J. Comp. Appl. Maths., vol. 167 (2004), pp. 255-269.

[3] J. BEBERNES, K.N. CHUE, AND W. FULKS, *Some applications of invariance for parabolic systems*, Indiana Univ. Math. J., vol. 28 (1979), pp. 269-277.

[4] J. BEBERNES, AND D. EBERLY, *Mathematical Problems from Combustion Theory*; Applied Mathematical Sciences vol. 83, Springer-Verlag (1989).

[5] J. BEBERNES, AND A. LACEY, *Finite-time blowup for a particular parabolic system*; SIAM J. Math. Anal. vol. 21-6 (1990).

[6] D. LUSS, AND N. AMUNDSON, *Uniqueness of the steady state solutions for chemical reaction occuring in a catalyst particle or in a tubular reactor with axial diffusion*, Chem. Eng. Sci., vol. 22 (1967), pp. 253-266.

[7] A. KAPILA, *Arrhenius systems: dynamics of jump due to slow passage through criticality*, SIAM J. App. Maths., vol. 41-1(1981), pp. 29-42.

[8] D. A. KASSOY, *Extremely rapid transient phenomena in Combustion, Ignition and explosion*, Asymptotic Methods and Singular Perturbations (R. O'Malley editor), Vol. X SIAM-AMS Proceedings (1976), pp. 61-72.

[9] A. OMÓN ARANCIBIA, AND A. ÁVILA, *Existence and blow-up for a parabolic system from Combustion theory*, to appear in J. Nonlinear Science.

[10] D.H. SATTINGER, *A nonlinear parabolic system in the theory of Combustion*; Quart. Appl. Maths., vol. 33 (1975), pp. 47-61.

EL VAN Y EL PUNTO MUERTO FINANCIERO DE UN PROYECTO DE INVERSIÓN CON UNA ECUACIÓN DE DEMANDA HIPERBÓLICA EN FUNCIÓN DE LA TASA DE DESCUENTO

Domingo A. Tarzia †

† Depto. de Matemática - CONICET, FCE, Universidad Austral, Paraguay 1950, S2000FZF Rosario, Argentina
DTarzia@austral.edu.ar

Resumen: Se estudia un proyecto de inversión con la existencia de dos variables independientes (la cantidad de unidades Q a vender en cada año que está relacionada con el precio P a través de una ecuación de demanda hiperbólica del tipo $PQ = C_0$ $(C_0 > 0)$ y la tasa de descuento r. Se obtiene la expresión explícita del *VAN* (valor actual neto) del proyecto de inversión en función de Q y r. También se determina explícitamente el punto muerto financiero $Q_f(r)$ en función de los parámetros restantes del problema y se estudia analíticamente su comportamiento respecto de la tasa de descuento r. Por último, se demuestra que el *VAN* será positivo si y solamente si $0 < Q < Q_f(r)$ con $0 < r < r_1$ donde r_1 es una tasa de descuento límite superior que se determina como la única solución de una ecuación que también se explicita bajo cierta hipótesis sobre el parámetro positivo C_0.

Palabras claves: *Valor actual neto, Punto muerto financiero, tasa de descuento, ecuación de demanda.*

2000 AMS Subjects Classification: 91B28
JEL Classification Codes: C02, C63, G10, G31

1. INTRODUCCIÓN

Se considera un proyecto de inversión simple en el cual se realiza solamente una inversión inicial I (flujo de fondo con signo negativo) y en los n años de duración del mismo se tendrán, en general, flujos de fondos de signo positivo**.**

Es muy importante la evaluación del proyecto de inversión para poder conocer si el mismo es o no es rentable**.** Existen varios criterios para la evaluación [10] como son: el *valor actual neto* (conocido como *VAN*), la *tasa interna de retorno* (conocida como *TIR*), el *período de recuperación de la inversión* (conocido como *PRI*) y la *rentabilidad inmediata* (conocido como RI). En este trabajo se utilizará el Valor Actual Neto o *VAN* como criterio de evaluación**.** El *VAN* es aquel que permite determinar la valoración de una inversión en función de la diferencia entre el valor actualizado de todos los cobros derivados de la inversión y todos los pagos actualizados originados por la misma a lo largo del plazo de la inversión realizada**.** En otras palabras, el *VAN* de un proyecto es igual a la sumatoria de los valores actuales (al momento cero) de todos los flujos de fondos (negativos y positivos) que genera el mismo proyecto**.** La inversión será aconsejable si su *VAN* es positivo [1, 2, 9,10, 14]. La importancia del criterio del *VAN* puede apreciarse en [3, 7, 11].

En el presente trabajo se estudia un proyecto de inversión con la existencia de dos variables independientes (la cantidad de unidades Q a vender en cada año que se encuentra relacionada con el precio P a través de una ecuación de demanda dada por la ecuación hipérbola (1)) y la tasa de descuento r que pueden hacer, según los valores que adopten, que el proyecto sea viable o no**.** Por ende, el VAN será una función de las variables Q y r. Es de mucha importancia encontrar el valor de la variable independiente Q que haga que el correspondiente *VAN* sea nulo para una dada tasa de descuento r. Se define como Punto Muerto Financiero (break even point) el valor de la variable independiente *Q* para el cual el *VAN* es nulo.

Planteo del problema, hipótesis y resultados obtenidos. Se tienen los siguientes parámetros:

- I : Inversión inicial que se realiza en el año cero (antes del comienzo del primer año del desarrollo del proyecto de inversión);

- n : cantidad de años de duración del proyecto de inversión $(2 \le n)$;
- A : Amortización anual**.** Es la parte anual de la inversión que permite bajar (mejorar) el pago de impuestos a las ganancias;
- Q : Cantidad de unidades del producto vendidas por año;
- P : Precio de venta unitario al que la Compañía vende cada producto;
- C_v : Costo variable por unidad para producir el producto;
- C_f : Costo fijo anual de la Compañía;
- t_{ig} : Tasa del impuesto a las ganancias (en tanto por uno);
- r : Tasa de descuento o costo de oportunidad (en tanto por uno).

En [4] se realiza un estudio del *VAN* de un proyecto de inversión en función de la tasa de descuento r ; se demuestra que el *VAN* de un proyecto de inversión en función de la tasa de descuento r es una función estrictamente decreciente y convexa. Dicho estudio fue ampliado adecuadamente en [12] realizando un análisis del punto muerto financiero, respecto de la variable Q , en función de la tasa de descuento r , además de un análisis de sensibilidad. En [13] se estudió cuando la empresa funciona con una ecuación de demanda lineal; en el presente trabajo se generalizará dicho estudio para una ecuación de demanda hiperbólica del tipo

$$PQ = C_0 \qquad \left(P = \frac{C_0}{Q} \right) \tag{1}$$

que relaciona la cantidad de unidades a vender Q con el precio unitario P , donde C_o es una constante positiva.

En el proyecto de inversión simple se considerarán las siguientes hipótesis de trabajo:

- Toda la inversión I se realiza de una sola vez y en el año 0;
- La inversión inicial se amortiza totalmente en n años, con lo cual la amortización anual está dada por: $A = I / n$
- En los n períodos de tiempo de duración del proyecto de inversión se realizan las mismas actividades;
- Se considera que la compañía vende un solo producto (podría producirse o comprarse para luego revenderse);

El objetivo del trabajo es el de obtener la expresión explícita del *VAN* del proyecto de inversión en función de la variable independiente Q para una dada tasa de descuento r . También se determinará explícitamente el punto muerto financiero Q_f en función de los parámetros restantes del problema (I , n , C_f , C_v , C_0 t_{ig} , r) y se estudiará analíticamente sus comportamientos respecto de la tasa de descuento r . Por último, se demuestra que el *VAN* será positivo si y solamente si $0 < Q < Q_f(r)$ con $0 < r < r_1$ donde r_1 es una tasa de descuento límite superior que se determina como la única solución de una ecuación que también se explicita bajo cierta hipótesis sobre la constante C_0 .

2. Proyecto de inversión dependiente de la variable cantidad Q

Como en cada año ($i = 1, 2, ..., n$) se realizan las mismas operaciones se supone que los parámetros Q , P , C_f , C_v , r , t_{ig} son constantes durante los n años de duración del proyecto de inversión. Para cada año t ($t = 1, 2, ..., n$) se tiene:

Ingresos (precio por cantidad)	$PQ = C_0$
Costos variables	$C_v Q$
Costos fijos	C_f

Amortización	$A = \dfrac{I}{n}$
Beneficios antes de impuestos (BAT)	$BAT_t = (P - C_v)Q - C_f - A$
Impuesto a las Ganancias (IG)	$IG_t = t_{ig}\left[(P - C_v)Q - C_f - A\right]$
Beneficio Neto $(BN = BAT - IG)$	$BN_t = (1 - t_{ig})\left[(P - C_v)Q - C_f - A\right]$
Flujo de Tesorería Neto $(F = BN + A)$	$F_t = (1 - t_{ig})\left[(P - C_v)Q - C_f - A\right] + A$
	$= -(1 - t_{ig})C_v Q + (C_0 - C_f)(1 - t_{ig}) + t_{ig}A$
Factor de descuento para el año t	$\dfrac{1}{(1+r)^t}$.

Teniendo en cuenta que la inversión I se realiza en el período 0 se tiene que el correspondiente VAN del proyecto de inversión viene dado por $-I$ más los valores actuales de todos los flujos de fondos F_t obtenidos en cada año t variando t desde 1 a n, es decir [15]:

$$(2) \qquad VAN(Q,r) = -I + \sum_{t=1}^{n} \frac{F_t}{(1+r)^t} = -I + f(r)[-(1-t_{ig})QC_v + (1-t_{ig})(C_0 - C_f) + t_{ig}A)]$$

$$= -m(r)Q + h(r)$$

expresión que resulta ser una función afín de la variable Q donde se han definido las funciones reales $f = f(r), h = h(r)$ y $m = m(r)$ de la siguiente manera:

$$(3) \qquad f(r) = \frac{1}{r}\left[1 - \frac{1}{(1+r)^n}\right], \quad r > 0,$$

$$(4) \qquad h = h(r) = -I + f(r)\left[(1-t_{ig})(C_0 - C_f) + t_{ig}A\right], \quad m(r) = (1-t_{ig})C_v f(r) > 0.$$

Lema 1 i) La función real $h = h(r)$ tiene las siguientes propiedades:

$$(5) \qquad h(0^+) = n(1-t_{ig})(C_0 - C_f - A), \qquad h(+\infty) = -I < 0,$$

$$(6) \qquad h'(r) = [(1-t_{ig})(C_0 - C_f) + t_{ig}A]f'(r).$$

ii) Si C_0 verifica la desigualdad

$$(7) \qquad C_0 > C_f + A$$

entonces $h(0^+) > 0$, $h'(r) < 0, \quad \forall r > 0$, y por lo tanto h es una función estrictamente decreciente.

iii) Si C_0 verifica la desigualdad (7) entonces h tiene un único cero $r_1 > 0$ que viene dado por la única solución de la ecuación

$$(8) \qquad f(r) = \frac{I}{(1-t_{ig})(C_0 - C_f) + t_{ig}A}, \quad r > 0.$$

Además, la función real $r_1 = r_1(C_0)$, que a cada parámetro C_0 le asigna la única solución de la ecuación (8), es una función estrictamente creciente de la variable C_0 con las propiedades siguientes:

(9) $$r_1(C_f + A) = 0, \qquad r_1(+\infty) = +\infty .$$

Teorema 2 i) Si $0 < C_0 \le C_f + A$ entonces $VAN(Q,r) < 0, \quad \forall Q, r > 0$ y por ende el proyecto de inversión no es viable.
ii) Si C_0 verifica la desigualdad (7) entonces existe un único punto muerto financiero que anula el VAN dado por la expresión:

(10) $$Q_f = Q_f(r) = \frac{1}{(1-t_{ig})C_v}[(1-t_{ig})(C_0 - C_f) + t_{ig}A - \frac{I}{f(r)}]$$

que está bien definido para $0 < r < r_1$, donde r_1 está dado por la única solución de la ecuación (8).

iii) El proyecto de inversión es viable cuando la cantidad de unidades a vender Q verifica las siguientes condiciones:

(11) $$VAN(Q,r) > 0 \quad \Leftrightarrow \quad 0 < Q < Q_f(r), 0 < r < r_1 .$$

Prueba. El resultado se deduce de la siguiente expresión del $VAN(Q,r)$ dado por:

(12) $$VAN(Q,r) = m(r)[Q_f(r) - Q], \quad Q > 0, \quad r > 0 \ .$$ □

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por PIP No. 0460 de CONICET-UA y PICTO AUSTRAL 2008 nº 73, Rosario, Argentina.

REFERENCIAS

[1] R. BAKER, AND R. FOX, *Capital investment appraisal: A new risk premium model,* Int. Transactions on Operations Research, 10 (2003), pp. 115-126.
[2] R. BREALEY AND S. MYERS, *Fundamentos de financiación empresarial*, Mc Graw- Hill, Madrid, 1993.
[3] K.J. CHUNG AND S.D. LIN, *An exact solution of cash flow for an integrated evaluation of investment in inventory and credit*, Production Planinning & Control, 9 (1998), pp. 360-365.
[4] M. FERNANDEZ BLANCO, *Dirección financiera de la empresa*, Pirámide, Madrid, 1991.
[5] M.M. HAJDASINSKI, Remarks in the context of the case for a generalized net present value formula, The Engineering Economist, 40 (1995), 201-210.
[6] M.M. HAJDASINSKI, *Compatibility, project ranking, and related issues*, The Engineering Economist, 42 (1997), 325-339.
[7] S.P. LAN, K.J. CHUNG, P. CHU AND P.F. KUO, *The formula approximation for the optimal cycletime of the net present value*, The Engineering Economist, 48 (2003), pp. 79-91.
[8] A. PIERRU, E. FEUILLET-MIDRIER, *Discount rate value and cash flow definition: a new relationship and its implications*, The Engineering Economist, 47 (2002), 60-74.
[9] S. REICHELSTEIN, *Providing managerial incentives: Cash flows versus accrual accounting*, J. Accounting Research, 38 (2000), pp. 243-269.
[10] N. SAPAG CHAIN, *Evaluación de proyectos de inversión en la empresa*, Prentice Hall, 2001.
[11] R.E. STANFORD, *Optimizing profits from a system of accounts receivable*, Management Sci., 35(1989), pp. 1227-1235.
[12] D.A. TARZIA, *El punto muerto financiero de un proyecto de inversión simple en función de la tasa de descuento*, Mecánica Computacional, 26 (2007), pp. 614-632. Ver también, *El punto muerto financiero de un proyecto de inversión en función de la tasa de descuento*, Tesis Maestría en Finanzas, Univ. Nac. de Rosario, Rosario, 2010.
[13] D.A. TARZIA, *El VAN y el punto muerto financiero de un proyecto de inversión con una ecuación de demanda en función de la tasa de descuento*, en Congreso II MACI 2009, Rosario, 14-16 Dic. 2009, MACI, 2 (2009), pp. 85-88.
[14] M. VANHOUCKE, E. DEMEULEMEESTER AND W. HERROELEN, *On maximizing the net present value of a project under renewable resource constraints*, Management Sci., 47 (2001), pp. 1113-1121.
[15] J.L. VILLALOBOS, *Matemáticas financieras*, Prentice-Hall, México, 2001.

HEDGING LATE FROST RISK WITH WEATHER DERIVATIVES

Elsa Cortina[♭] and Ignacio Sánchez[†]

[♭]*Instituto Argentino de Matemática (IAM-CONICET), Saavedra 15,3er. piso, 1083 Buenos Aires, Argentina, elsa.cortina@gmail.com, www.iam.conicet.gov.ar*
[†]*Posgrado en Finanzas, Universidad de San Andrés, 25 de mayo 586, 1002 Buenos Aires, Argentina, isanchez@udesa.edu.ar, www.udesa.edu.ar*

Abstract: The objective of this paper is to present an option depending on minimum temperature to hedge late frost risk. We also present a numerical application to fruit cultivation in the province of Mendoza, Argentina.

Keywords: *Weather derivatives, minimum temperature, agriculture.*
JEL Classification code: G13, G32

1 INTRODUCTION

Weather derivatives are financial contracts with payoffs that depend on a climatological underlying, such as temperature, rainfall or snowfall, and are mainly used to hedge potential weather risks. Since 1996, when the first contract was traded in USA, the market increased to $ 45.25 billion in 2006, according to the Weather Risk Management Association. Among the number of industries exposed to weather risk, agriculture is perhaps one of the bussiness activities most sensitive to climatological conditions. In this paper we propose an American binary option on minimum temperature to hedge the risk of late frost faced by fruit producers, and present a numerical application using data collected at meteorological stations in Mendoza, Argentina.

For extensive surveys on weather derivatives we refer to [6] and [10]. In [9] a comparison between traditional insurance policies and wheather derivatives in terms of risk management applications is discussed.

The experimental data are analyzed in Section 2. In Section 3 we construct the model that describes the minimum temperature behaviour. Section 4. is dedicated to the estimation of the parameters of the model. The option is presented in Section 5., together with a pricing numerical example using Monte Carlo simulations.

2 DATA ANALYSIS

Our dataset consists of 11 years of daily minimum temperature data collected at Tunuyán station, in the province of Mendoza, the main area of fruit cultivation in Argentina. As 37 data were missing, our first step was to interpolate these missing observations. Following [8], we use the method of Principal Component Analysis (PCA) described in [5] to reconstruct 26 of them. For the remaining 11 data we use linear interpolation, since the records from neighbouring meteorological stations required to apply PCA were not available.

Figure 1. shows that the temperature exhibits a clear seasonal behaviour and reverts to a long term periodic mean term. The normality tests performed indicate a high level of serial correlation, consistent with seasonality. However, the frequency distribution of the raw data justifies the adoption of a Normal distribution to model the stochastic component.

3 THE MODEL

The temperature path can be seen as a combination of a deterministic trend together with random shocks. As it is observed in Figure 1., there is strong evidence of a periodic component and mean reversion. We propose a model for the minimum temperature T_t^{min} similar to the first one-factor model introduced in [7]

$$T_t^{min} = f(t) + X_t, \tag{1}$$

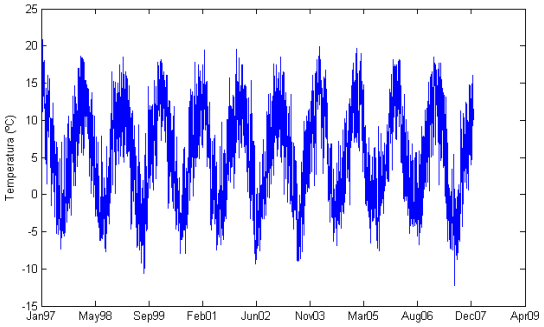


Figure 1: Daily minimum temperature - Tunuyán station

$$dX_t = -\kappa X_t dt + \sigma_t dW_t \tag{2}$$

where $f(t)$ is a totally predictable term, the dynamics of X_t is given by the SDE (2), $\kappa > 0$, $X(0) = x_0$, and dW_t is the increment to a standard Wiener process W_t; i.e., X_t follows a mean-reverting process, with a time dependent long-term reversion level, and reversion speed κ. Assuming that $f(t)$ satisfies the required regularity conditions, from (1) and (2) we derive the stochastic differential equation for the minimum temperature

$$dT_t^{min} = \kappa \left(g(t) - T_t^{min} \right) dt + \sigma_t dW_t, \tag{3}$$

where

$$g(t) \equiv \frac{1}{\kappa} \frac{\partial f(t)}{\partial t} - f(t) \tag{4}$$

For derivative pricing we would need to incorporate the market price of risk in the drif. Since there is not a wheather derivative market in Argentina to estimate the implicit market price of risk, our valuation results will be based on (3) and (4); i.e., under the restrictive assumption that the market price of risk equals zero.

4 Calibration of the Model

We model the function $f(t)$ in (1) as

$$f(t) = a_1 + a_2 t + a_3 \sin(\omega t) + a_4 \cos(\omega t), \quad \omega = \frac{2\pi}{365},$$

and by applying ordinary least squares to estimate its parameters we found $a_1 = 5.67$, $a_2 = -0.0002$, $a_3 = 2.00$, and $a_4 = 7.44$.

To estimate the rate of mean reversion κ, we need a first (and rough) estimation of the volatility. In [1] it is argued that the volatility varies throughout the months but remains approximately constant for each month; i.e.

$$\sigma_k = \frac{1}{N_k} \sum_{j=1}^{N_t - 1} \left(T_{j+1}^{min} - T_j^{min} \right)^2, \tag{5}$$

where the index $k = 1, 2, \ldots 12$ runs throughout the months of the year (t_1= January,t_2= February,...t_{12}= December), N_k is the number of days in the month t, and T_j is the minimum temperature registered on day j. We use these first estimates of the volatility given by (5) to estimate the monthly mean reversion rates (c.f.[1], [3]) as

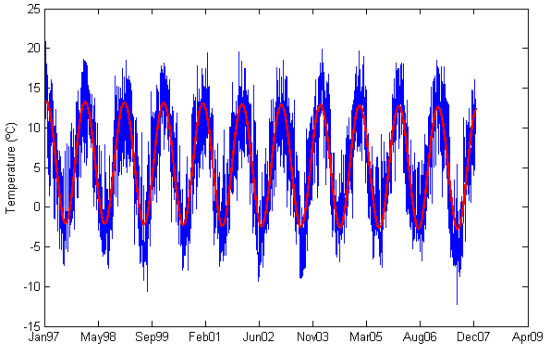


Figure 2: Minimum temperature and deterministic component

$$\kappa_n = -\log \frac{\sum_{i=1}^{n} \left[-\Delta T_{i-1} \Delta T_i\right] / \sigma_{i-1}^2}{\sum_{i=1}^{n} -\left(\Delta T_{i-1}\right)^2 / \sigma_{i-1}^2},$$

where $\Delta T_i = f(t_i) - T_i^{min}$, and $n = 136$ is the number of months in the sample, and we computed the estimate of the mean reversion rate κ as the average of κ_n. In a forthcoming paper, where more refined calculations will be presented, we use the set of κ_n estimates for the simulation of the minimum temperature paths. Tests performed on the deseasonalized returns confirm our initial assumption of normality. From the analysis of the volatility residues we obtain the following results: 1) The Durbin-Watson test equals 1,87, allowing us to assume that there is no first-order correlation; 2) we can reject at 5 $\%$ the null hypothesis of ARCH effects in the Engels test on heteroscedasticity, thus we assume that the volatility residues are i.i.d. Gaussian perturbations. Therefore, unlike the results in [4], where the the behaviour of the volatility is described by a Vasicek model, all our tests indicate that the monthly volatility is suitably modeled by a white noise

$$\sigma_t = \sigma_0 + \gamma \varepsilon_t, \tag{6}$$

These conclusions were also confirmed by spectrum analysis. The estimates computed for the trend and the volatility of volatility in (6) are $\sigma_0 = 2.91$, and $\gamma = 0.37$.

5 A Temperature Option

We propose an American option with binary payment (cash-or-nothing) defined by

$$\text{Payoff} = \mathcal{H}\left(K - T^{min}\right), \tag{7}$$

where $\mathcal{H}$ is the Heaviside function, and K is the critical minimum temperature for frost injuries; i.e., the payoff has value 1 when its argument is positive, and zero otherwise. It is always optimum to exercise the option as soon as K exceeds T^{min} (c.f. [11]). These contracts must be tailored to the individual location and species, since the ocurrence of frost damages depends on the phenological stages of the selected species and the strike is the critical temperature for each stage. Phenological data from the Department of Contingencies of Mendoza are availble for several fruit plants, and some of them are exhibited in Table 1.

	Beginning of bloom (bud)	Full bloom	Small green fruits	2 cm fruits
Grapevine	-1.1	-0.6	-1.1	
Cherry	-2.8	-2.2	-1.1	-3.0
Plum	-3.4	-2.2	-1.1	-2.0

Table 1: Phenological Stages Critical Temperatures (°C)

5.1 Numerical Example

We choose the apple cultivation for a numerical example and evaluate a contract with payoff given by (7), where the parameters are defined in terms of the phenological data shown in Table 2.:

- Two contract periods: form 13/09/10 to 23/09/10, and from 24/09/10 to 27/10/10.
- Two strikes: -3.9 °C, and -2.2 °C

In order to price the option we simulate 20,000 miminum temperature trajectories, and under the assumption of a risk-free rate = 0.095 the price obtained is 0.56.

	Beginning of bloom (bud)	Full bloom	Small green fruits
Critic temperature (°C)	-3.9	-2.2	-1.1
Period	13-Sep to 23-Sep	24-Sep to 27-Oct	28-Oct to 4-Nov

Table 2: Phenological stages in apple trees

Acknowledgments

The first author is a fellow of the Consejo Nacional de Investigaciones Científicas y Técnicas.

References

[1] P. Alaton,B. Djehiche, and D. Stillberger, *On modelling and pricing weather derivatives*, Applied Mathematical Finance, March, 2002.

[2] I. Basawa,and B. Prasaka Rao, *Statistical Inference for Stochastic Processes*, Academic Press, 1980.

[3] B. Bibby, and M. Srensen, *Martingale Estimation Functions for Discretely Observed Diffusion Processes*, Bernoulli, Vol. 1, Nr. 1 (1995), pp. 17-39.

[4] A. Bhowan, *Temperature Derivatives*, University of de Wiwatersrand, Working Paper, (2003), pp.25.

[5] C. L. Dunis, and V. Karalis, *Weather Derivatives Pricing and Filling Analysis for Missing Temperature Data*, University of de Wiwatersrand, Working Liverpool Business School & CIBEF, (2003).

[6] S. Jewson, a. Briz, and C. Ziehmann, *Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations*, Cambridge University Press,2005.

[7] J. Lucia,and E. Schwartz, *Electricity prices and power derivatives: Evidence from the Nordic Power Exchange*, Review of Derivatives Research,Vol. 5 (2002),pp 5-50.

[8] M. Mraoua, and D. Bari, *Temperature stochastic modeling and weather derivatives pricing: empirical study with Moroccan data*, Afrika Statistika,Vol.2, Nr. 1 (2007), pp.22-43 .

[9] J. Tindall, *Weather Derivatives: Pricing and Risk Management Applications*, The Institute of Actuaries of Australia, (2006), pp. 50.

[10] C. Turvey, *Weather Derivatives and Specific Events Risk* , Annual Meeting of the AAEA, (1999), pp. 12.

[11] P. Wilmott, J. Dewinne, and S. Howison, *Option Pricing: Mathematical models and computation*, Oxford Financial Press,1993.

VALUACIÓN DE LAS OPCIONES SOBRE FUTUROS ESTILO ARGENTINO

Gabriela S. Facciano† y Rodolfo Oviedo‡

†Mercado a Término de Buenos Aires S.A., Bouchard 454, 4º piso, Ciudad Autónoma de Buenos Aires, Argentina, gfacciano@matba.com.ar
‡Universidad Austral, Facultad de Ciencias Empresariales, Paraguay 1950, Rosario, Argentina, roviedo@austral.edu.ar

Resumen: Las opciones sobre futuros argentinas tienen un esquema particular de flujos de fondos durante su vida. No existen fórmulas cerradas para determinar su prima. Presentamos una solución que usa múltiples árboles binomiales. A continuación se presenta un método para hacer más exacta la valuación con el uso de dos tipos de opciones para las que sí existen fórmulas cerradas y que actuarán como cotas superior e inferior.

Palabras claves: *opciones, futuros, Argentina, MATBA, ROFEX*
2000 AMS Subjects Classification: 91G20. JEL Classification: G12, G13

1. INTRODUCCIÓN

Las opciones sobre futuros que actualmente se negocian en los mercados de futuros se pueden clasificar de acuerdo a la distribución en el tiempo de los flujos de fondos durante la vida de la opción. Las formas más conocidas de opciones son las opciones estilo tradicional y las estilo futuro. En la Argentina, las opciones sobre futuros tienen un esquema de pagos que es intermedio entre los dos anteriores. Estas serán el objeto de nuestro estudio.

Cuando se negocia una opción estilo argentino, la prima se paga al contado, más precisamente, antes de comenzar la rueda del día siguiente, tal como ocurre con las opciones estilo tradicional. Sin embargo, el comprador recibe, en el mismo día en que paga la prima, el valor intrínseco de la opción calculado al cierre del día de la compra. Desde ese momento hasta el vencimiento, cobra o paga diariamente todas subas y bajas del valor intrínseco. En definitiva, al vencer la opción, el comprador habrá recibido el valor intrínseco vigente al momento del vencimiento, aunque conformado por la suma de flujos de fondos repartidos entre el día de contratación y el de vencimiento.

2. VALUACIÓN MEDIANTE EL USO DE ÁRBOLES BINOMIALES

Para hallar el valor de las opciones estilo Argentino se emplea un conjunto de árboles binomiales que permitirá tener en cuenta las características especiales de las mismas. El tiempo hasta el vencimiento de la opción se divide en n períodos, con lo que se consideraran $n+1$ momentos $j = 0,1,...,n$, donde 0 representa el momento de valuación y n el vencimiento de la opción.

Quien compra una opción estilo Argentino recibe inmediatamente el valor intrínseco $VI(0)$ al momento 0 de comprar la opción. En adelante, se le liquidan diferencias diarias $VI(j)-VI(j-1)$ para $j = 0,1,...,n$ de acuerdo a la variación del valor intrínseco. Las diferencias acumuladas al vencimiento n serán iguales a $VI(n)-VI(0)$. Sumando el valor recibido el día de compra, $VI(0)$, resulta $VI(n)$. En definitiva, al vencimiento, el comprador habrá recibido un flujo de fondos acumulado de $VI(n)$, equivalente al valor intrínseco en ese momento.

Como al comprar la opción estilo argentino se recibe inmediatamente el valor intrínseco, esta cantidad pasa inmediatamente a formar parte de la prima. Entonces, sólo resta calcular el valor temporal, que no puede ser otra cosa que el valor presente de las diferencias diarias futuras. El árbol binomial que proponemos a continuación tiene por objetivo justamente calcular el valor temporal mencionado. Sumando el valor intrínseco y valor temporal, obtenemos la prima.

En la Figura 1 se presentan los tres esquemas de los árboles binomiales utilizados para la valuación. Las celdas negras se completan con números o fórmulas. Las celdas *blancas y grises* tienen valor *cero*.

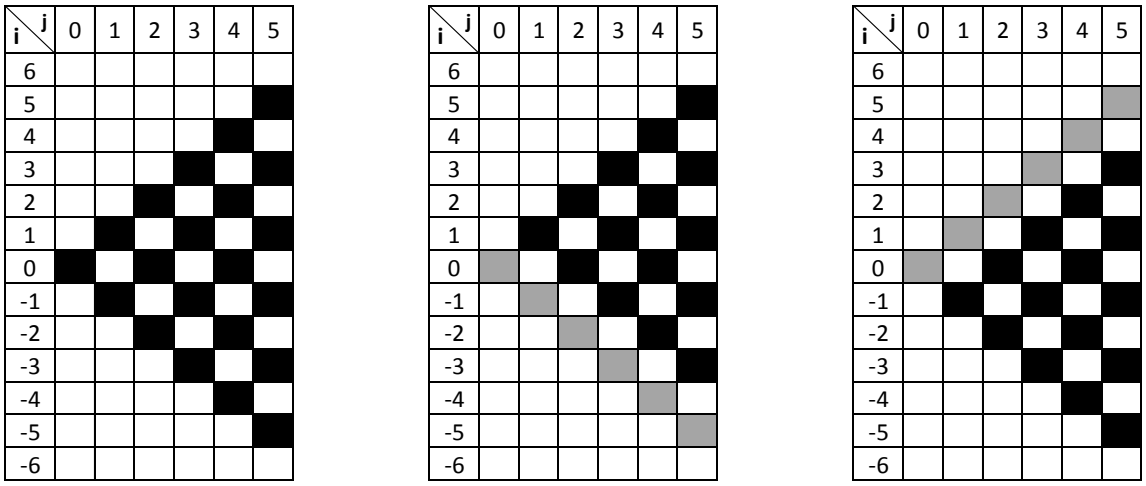


Figura 1: Esquemas de los árboles binomiales utilizados para la valuación.

El árbol de precios se construye según el modo tradicional, siguiendo el esquema (a) de la Figura 1. Los restantes árboles seguirán también este esquema, salvo cuando se indique expresamente que usan los esquemas (b) o (c).

$F(0;0)$ es el precio a futuro actual. Los precios de la diagonal superior, donde $i = j$, se obtienen haciendo: $F(i;j) = F(i-1;j-1)\times u$. El resto del árbol se puede calcular haciendo $F(i;j) = F(i+1;j-1)\times d$.

Se construye un árbol de valores intrínsecos con la fórmula $VI(i;j) = \max(0;(F(i;j)-K)\times Y)$, donde $Y = 1$ para un call e $Y = -1$ para un put.

El árbol de diferencias $U(i;j)$ para los casos de suba de F se arma rellenando las celdas negras del esquema (b) de la Figura 1 haciendo $U(i;j) = VI(i;j) - VI(i-1;j-1)$.

El árbol de diferencias $D(i;j)$ para los casos de baja de F se arma según el esquema (c) de la Figura 1, donde $D(i;j) = VI(i;j) - VI(i+1;j-1)$.

Para construir el árbol de probabilidades de llegar a cada nodo se vuelve al esquema (a) de la Figura 1, en donde $P(0;0) = 1$ y las restantes probabilidades se obtienen haciendo $P(i;j) = P(i-1;j-1)\times p + P(i+1;j-1)\times q$, donde $p = (1-d)/(u-d)$ y $q = 1-p$.

Valor esperado del flujo de fondos en cada nodo es: $VE(i;j) = D(i;j)\times P(i+1;j-1)\times q + U(i;j)\times P(i-1;j-1)\times p$.

Para cada momento $j = 1,...,n$, se suman los flujos de fondos esperados en cada uno de los nodos de la columna correspondiente a ese momento j, para obtener $VE(j) = \sum_i VE(i;j)$. Luego se descuenta esa suma a la tasa libre de riesgo hasta el momento 0 para obtener el valor presente $VP(VE(j))$ de los flujos de fondos en j.

Por último, se suman los valores presentes $VP(VE(j))$ para $j = 1,...,n$, y así se obtiene el valor temporal $VT(0;0)$ al momento de valuación: $VT(0;0) = VP(VE(1)) + ... + VP(VE(n))$. Finalmente, se suman

valor temporal $VT(0;0)$ y valor intrínseco $VI(0;0)$ para obtener el valor de la prima EA de una opción estilo argentino según el método binomial: $EA = VI(0;0) + VT(0;0)$.

Nótese que en ningún momento se plantea la posibilidad de ejercicio anticipado. Esto no es necesario porque, en todo momento, el comprador ya ha recibido el valor intrínseco; con lo cual, no tiene incentivo alguno para ejercer anticipadamente[1]. De ahí que una opción estilo futuro americana valga lo mismo de una europea.

3. VALUACIÓN ROBUSTA USANDO UNA COTA SUPERIOR Y OTRA INFERIOR

Para obtener mayor precisión en la valuación, usamos una variación de la técnica de *control variates* que utiliza una cota superior y una inferior, dadas respectivamente por la prima de una opción estilo futuro y de una estilo tradicional europea[2]. Como ya se indicó, el comprador de una opción estilo argentino cobra, al negociar una opción, el valor intrínseco de ese momento y, desde allí hasta el vencimiento, cobrará (pagará) diariamente los aumentos (disminuciones) del valor intrínseco. De esta forma, a cada momento, el saldo de los flujos de fondos (FF) igualará el valor intrínseco vigente. Por adquirir los FF anteriores, el comprador paga una prima de contado. Para describir las opciones estilo futuro de forma paralela a las estilo argentino, supondremos que el comprador también paga la prima de contado. Esta prima comprará los FF que se describen a continuación. Al negociar la opción, el comprador recibe el valor de la prima de mercado. Desde allí hasta el vencimiento, cobrará (pagará) los aumentos (disminuciones) de la prima de mercado. De esta forma, a cada momento, el saldo de los FF igualará la prima de mercado vigente, que al vencimiento coincidirá con el valor intrínseco. Es así que el saldo al vencimiento de los FF recibidos por el comprador de una opción estilo futuro es igual que el recibido por el comprador de una opción estilo argentino. Sin embargo, durante la vida de estas opciones, el saldo de los FF recibidos por el comprador de una opción estilo argentino será sólo el valor intrínseco de la opción, mientras que el saldo correspondiente a una opción estilo futuro igualará el valor completo de la prima de mercado, esto es, el valor intrínseco más el valor temporal[3]. Por lo explicado, en una opción estilo futuro hay FF que se reciben con anterioridad a los FF de una opción argentina. De ahí que la prima de una opción estilo futuro es una cota superior de la estilo argentino. La prima de una opción estilo tradicional es una cota inferior porque recién al vencimiento se obtiene la totalidad del valor intrínseco.

Llamamos EF y ET a la primas de una opción estilo futuro y una opción europea estilo tradicional, respectivamente, calculadas según los árboles binomiales utilizados más arriba. EF se obtiene empleando la última columna de los arboles binomiales de valor intrínseco y de probabilidades: $EF = \sum_i VI(i,n) \times P(i,n)$. Para obtener ET sólo hay que descontar el valor anterior

$$ET = EF \times e^{-r \times T}, \tag{1}$$

donde r es la tasa anual libre de riesgo compuesta instantáneamente y T es el tiempo hasta el vencimiento de la opción medido en años.

Llamamos $\overline{EF}$ y $\overline{ET}$ a las primas de las mismas opciones estilo futuro y tradicional dadas por las fórmulas de Lieu [4] y la de Black [1] respectivamente. La relación entre ambas está dada por

$$\overline{ET} = \overline{EF} \times e^{-r \times T} \tag{2}$$

---

[1] Esto está demostrado formalmente en Facciano y Oviedo [3]. En particular se demuestra que aun a quien previera una erosión del valor intrínseco de la opción le convendría operar futuros para cubrirse de esta eventualidad y dejar la opción sin ejercer.

[2] No se especifica si la opción estilo futuro es europea o americana porque, en estas últimas, el derecho de ejercicio anticipado tiene valor nulo. Véase una demostración general en Oviedo y Tarzia [5]. Este resultado ya había sido demostrado, para modelos particulares, por Lieu [4] y por Chen y Scott [2]. En adelante, nombraremos la opción estilo futuro sin aclarar que nos referimos a una europea.

[3] La positividad del valor temporal fue demostrada por los artículos mencionados en la nota 2, con los niveles de generalidad allí indicados.

Para obtener un valor $\overline{EA}$ más robusto que el valor EA obtenido por con el esquema binomial usaremos EA, EF, ET, $\overline{EF}$ y $\overline{ET}$. Aprovechando las relaciones funcionales (1) y (2) y su paralelismo, trabajamos con el logaritmo de las primas para reducir los cálculos.

Partimos calculando la proporción $Prop$ en que se aleja $\log EA$ de $\log EF$ hacia $\log ET$ según el esquema binomial:

$$Prop = \frac{\log EA - \log EF}{\log ET - \log EF} = \frac{\log(EA/EF)}{\log(ET/EF)} = \frac{\log(EA/EF)}{-rT} = \frac{\log(EF/EA)}{rT}\,, \tag{3}$$

donde la última igualdad resulta de usar (1). El valor robusto del logaritmo de la prima de la opción estilo argentino es $\log\overline{EA} = \log\overline{EF} + Prop \times (\log\overline{ET} - \log\overline{EF}) = \log\overline{EF} + Prop \times \log(\overline{ET}/\overline{EF})$ que, en virtud de (2), se reduce a $\log\overline{EA} = \log\overline{EF} - \text{Prop} \times r \times T$, de donde:

$$\overline{EA} = \overline{EF} \times e^{-r \times T \times \text{Prop}}\,. \tag{4}$$

Así como en (2) se descuenta la prima de una opción estilo futuro desde el vencimiento hasta el momento de valuación para obtener la prima de una opción estilo tradicional, en (4) se descuenta por una proporción del tiempo hasta el vencimiento para obtener la prima de una opción estilo argentino. Este resultado es intuitivo porque, a tenor de lo explicado al comienzo de esta sección, los FF de la opción argentina son anteriores a los de la estilo tradicional pero posteriores a los FF de la etilo futuro.

El siguiente paso de esta investigación es determinar empíricamente el valor aproximado de proporción (3) como función de cuán *in-* o *out-of-the-money* está la opción, de la volatilidad y del tiempo hasta el vencimiento. Ahora bien, si se desea mayor precisión en el cálculo de la prima, asumiendo el costo computacional de calcular los árboles binomiales explicados, basta con reemplazar (3) en (4) y simplificar para obtener

$$\overline{EA} = \overline{EF} \times \frac{EA}{EF}$$

Hemos obtenido una fórmula muy sencilla para calcular un valor robusto de la prima de una opción estilo argentino en la cual, gracias al paralelismo entre (1) y (2) y el haber usado una interpolación exponencial de la prima (lineal en el logaritmo de la prima), se usa sólo la cota superior.

AGRADECIMIENTOS

Rodolfo Oviedo agradece subsidios de ROFEX y del Fondo de Becas para Proyectos de Investigación de la Universidad Austral.

REFERENCIAS

[1] F. BLACK, *The pricing of commodity contracts*, Journal of Financial Economics 3, no. 2 (1976) pp. 167-179.

[2] R. CHEN, AND L. SCOTT, *Pricing interest rate futures options with futures style margining*, Journal of futures markets, Vol. 13, pp. 15-22.

[3] G. FACCIANO AND R. OVIEDO, *El ejercicio sólo al vencimiento de las opciones en Rofex es una estrategia óptima, incluso en mercados imperfectos*, Journal of Management for Value, Vol. 2, no. 2 (2007), pp. 20-48.

[4] D. LIEU, *Option princing with futures-style margining*, Journal of Futures Markets, Vol. 10 (1990) pp. 327-338.

[5] R. OVIEDO AND D. TARZIA, *Theory of rational futures-style option pricing*, available at SSRN: http://ssrn.com/abstract=1469552, (2009).

A COINTEGRATION APPROACH FOR GENERATING SYNTHETIC PRICES OF HIGH FREQUENCY TIME SERIES

P. Arce[♯,†], A. Cañete[♭], C. Fernández[♯,†], R. León[♯,†], O. Orellana[♯,‡], R. Plaza[♯,†,♮] and L. Salinas[♯,†,♮]

[♯]*CTI HPC and CCTVal, Basal Project FB0821, UTFSM, Casilla 110-V, Valparaíso, Chile*
[♭]*Pan Alpha Trading LLC, 205 Lexington Av., 8th floor, NY 10016, New York*
[†]*Informatics Department, UTFSM, Casilla 110-V, Valparaíso, Chile*
[‡]*Mathematics Department, UTFSM, Casilla 110-V, Valparaíso, Chile*
[♮]*FONDECYT Project 1100805*

Abstract: This paper presents a forecast model for financial time series. The model allows a quantitative estimation of the influence of a set of stocks on a particular stock by means of a cointegration approach. The estimation of the stock price is done by a linear combination of other stock prices weighted by coefficients which represent the relative influences on price determination of a stock. An experimental validation using historic last traded prices with a frequency of one value per minute is provided considering a five month period in 2010.

Keywords: *Cointegration, High Frequency Time Series, Equity Markets.*

1 INTRODUCTION

High frequency trading provides huge amounts of information, which modern information technologies and mathematical computational models have allowed to process. As a result, the development of quantitative methods for the analysis and forecasting of time series has become essential in high frequency data processing [10].

The price of a stock at time t reflects what is thought about its value at that time. It is very difficult to determine the price of a stock due to the fact that the related information comes from multiple sources [3]. This fact reveals the inefficient markets problem, i.e. markets that are incapable of transmitting all the information generated at the same time to all market actors involved. Despite this, it is possible to appreciate the existence of a strong inter-relationship between the prices of stocks that belong to the same financial sector, such as technology, energy, etc. It is therefore possible to propose that information related to an economic sector could contribute to estimate the behavior of some of the stocks belonging to that sector [3]. It would be interesting to measure the influence that one stock has on others belonging to the same sector. In that way, it would be possible to find structural dependences to create new analysis perspectives.

In this paper, a new procedure is presented in order to generate a synthetic stock price using the last transaction prices of a set of stocks. These stocks belong to the same industry and are used as input values for a cointegration approach. The objective is to find structural relationships among those stocks [1, 2]. More specifically, an optimization process is carried out over the cointegration vector space to express a stock price as a linear combination of other stock prices belonging to the same industry. To verify the suitability of this procedure, a set of tests is carried out using stocks belonging to the technology industry.

The remainder of this paper is organized as follows. In Section 2, we give a brief overview of cointegration and its relevance to finance. In Section 3, we present a model to estimate the price of a particular stock in terms of other stock prices belonging to the same industry. Numerical results are shown in Section 4. The last section contains the conclusions and comments regarding future work.

2 COINTEGRATION

Let $\{y_t\}$ be a multivariable time series with n components $y_t\,(i)$ for $i = 1, \ldots, n$, each of which is a stationary time series of order 1. The series $\{y_t\}$ is said to be cointegrated if a vector $\hat{a} = [\hat{a}(1), \ldots, \hat{a}(n)]^T \in \mathbb{R}^n$ exists such that the time series,

$$Z_t := \langle \hat{a}, y_t \rangle = \hat{a}\,(1)\, y_t\,(1) + \cdots + \hat{a}\,(n)\, y_t\,(n)\,, \tag{1}$$

is stationary of order 0. This vector $\hat{a}$, known as a cointegration vector, is not unique. Eventually, a multivariable series $\{y_t\}$ could be cointegrated by more than one cointegration vector [4] and it is then possible to define a cointegration space $\mathbb{A} \subseteq \mathbb{R}^n$ of dimension h as the vector subspace expanded by a set of h linearly independent cointegration vectors $\hat{a}_1, \ldots, \hat{a}_h$.

In financial terms, cointegration allows the identification of the existence of long-term structural interactions within a set of time series of stock prices, providing evidence to indicate whether they share structural features that allow their dynamic to be followed [8, 9]. These interactions show the degree of influence of a stock over another and how these influences can summarize the behavior of a stock by introducing new information from a set of stocks. Using this information, it is then possible to express a stock price in terms of other stock prices belonging to the same industry or financial indexes.

3 MODEL FOR GENERATING SYNTHETIC PRICES

The objective of this section is to obtain a synthetic price of a particular stock $y_t(k)$. In this paper we refer to the synthetic price of a stock $y_t(k)$, as the stock price induced by the prices of a given stock set $y_t(j)$, $j = 1 \ldots n, j \neq k$ belonging to the same market sector as $y_{t_i}(k)$, at the same given time t. Therefore, the synthetic price is not a prediction of the stock price $y_t(k)$ but a measure of the influence of the market sector information on the stock price $y_t(k)$. This new knowledge of the stock price will allow the development of high frequency trading strategies which will be considered in subsequent papers.

In particular, the objective is to express $y_t(k)$ as a linear combination of the other stock prices $y_t(j)$, with $j = 1, \ldots, n, j \neq k$. Thus, we want to find coefficients $\hat{a}_j \in \mathbb{R}$, with $j = 1, \ldots, n$, such that $Z_t = \sum_{j=1}^{n} \hat{a}(j) y_t(j) \approxeq 0$, where $\approxeq$ means that the time series Z_t has zero mean and approximately zero variance. We achieve this goal through a two-steps method. Whenever these coefficients $\hat{a}(j)$ exist, and the resulting time series Z_t is stationary of order 0, then the time series $\{y_t\}$ is, in particular cointegrated. Thus, in the first step we start finding vectors $\hat{a}_i \in \mathbb{R}^n$ making Z_t cointegrated, without requiring $Z_t \approxeq 0$. In the second step, assumming that we have found h linearly independent cointegrating vectors $\hat{a}_i \in \mathbb{A} \subseteq \mathbb{R}^n$, by means of an ad-hoc optimization method, we construct a linear combination:

$$\hat{a} = \eta(1)\hat{a}_1 + \ldots + \eta(h)\hat{a}_h, \tag{2}$$

where $\eta(i) \in \mathbb{R}$, such that,

$$Z_t = \langle \hat{a}, y_t \rangle = \sum_{j=1}^{n} \hat{a}_i(j) y_t(j) \approxeq 0. \tag{3}$$

3.1 AN OPTIMAL COINTEGRATION VECTOR

Let us assume that the series $\{y_t\}$ has a cointegration vector space $\mathbb{A} \subseteq \mathbb{R}^n$ of dimension $h \leq n$, (this fact can be verified using hypothesis tests, for more details see [9, 12]), and denote the vector basis for this space as $\hat{a}_i$ with $i = 1, \ldots, h$. Then, every vector in $\mathbb{A}$ can be expressed as a linear combination of the basis vectors as shown in equation (2).

The objective of this subsection is to present a strategy to find a vector $\eta = [\eta(1) \ldots \eta(h)]^T$ such that the cointregrating vector $\hat{a}$ has an associated time series Z_t with zero mean and variance. If a vector $\hat{a} \in \mathbb{A}$ with these properties is found, then the cointegration equation (3) holds.

For the following discussion we consider a set of T observations of the time series $\{y_t\}$. According to our model we first need to impose the condition of zero mean on the series Z_t. Therefore, we have to find a cointegration vector $\hat{a}$ satisfying:

$$E(Z_t) = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{a}, y_t \rangle = \left\langle \hat{a}, \frac{1}{T} \sum_{t=1}^{T} y_t \right\rangle = \langle \hat{a}, \bar{y} \rangle = 0,$$

where $\bar{y}$ is the vector with components $\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y(i)$, with $i = 1, \ldots, n$.

A simple way to obtain a time series Z_t with zero mean is to find a vector $\hat{a}$ orthogonal to the vector $\bar{y}$. Furthermore, to reduce the variance of the series $\{Z_t\}$, the following optimization problem is formulated:

$$\min_{\substack{\hat{a}\in\mathbb{A}^n \\ \langle\hat{a},\bar{y}\rangle=0}} \frac{1}{2}\sum_{t=1}^{T}\left(\langle\hat{a}, y(t)\rangle\right)^2, \tag{4}$$

where the orthogonality constraint $\langle\hat{a},\bar{y}\rangle = 0$ is included to impose the condition $E(Z_t) = 0$.

Since $\hat{a} \in \mathbb{A}$, replacing (2) in (4) we arrive to the following equivalent optimization problem for η:

$$\min_{\substack{\eta\in\mathbb{R}^h \\ \langle\eta,\alpha\rangle=0}} \frac{1}{2}\sum_{t=1}^{T}\left(\sum_{k=1}^{n}\langle\eta, y_t(k)\cdot w_k\rangle\right)^2, \tag{5}$$

where $\alpha \in \mathbb{R}^h$ with $\alpha(j) = \langle\hat{a}_j,\bar{y}\rangle$ for each $j = \{1,\ldots,h\}$ and $w_k = [\hat{a}_1(k),...,\hat{a}_h(k)]^T$ for each $k \in \{1,\ldots,n\}$, where $\hat{a}_j(k)$ is the k-th component of the j-th cointegration vector.

Introducing the notation $\Phi_t = \sum_{k=1}^{n} y_t(k)\cdot w_k$ it is possible to rewrite the expression in equation (5) as follows:

$$\sum_{t=1}^{T}\left(\langle\eta,\Phi_t\rangle\right)^2 = \eta(1)^2\sum_{t=1}^{T}\Phi_t(1)^2 + \ldots + \eta(h)^2\sum_{t=1}^{T}\Phi_t(h)^2 + 2\sum_{u=1}^{h-1}\sum_{v=u+1}^{h}\eta(u)\eta(v)\left[\sum_{t=1}^{T}\Phi_t(u)\Phi_t(v)\right].$$

We introduce the matrix:

$$\Xi = \left[\sum_{t=1}^{T}\Phi_t(u)\Phi_t(v)\right]_{u,v=\{1,\ldots,h\}}.$$

Using this matrix, we rewrite the optimization problem (5) as the following standard quadratic problem:

$$\min_{\substack{\eta\in\mathbb{R}^h \\ \langle\eta,\alpha\rangle=0}} \frac{1}{2}\langle\eta,\Xi\eta\rangle,$$

which is trivially solvable for $\eta = 0$. According to (3), to generate a synthetic price for the i-th stock, the i-th component of $\hat{a}$ cannot be zero or near to zero, therefore, this component is forced to be greater than 1. Expressing this constraint in terms of the variable η, the following optimization problem is obtained:

$$\min_{\substack{\eta\in\mathbb{R}^h \\ \langle\eta,\alpha\rangle=0 \\ \langle w_i,\eta\rangle\geq 1}} \frac{1}{2}\langle\eta,\Xi\eta\rangle.$$

4 Experimental Results

The model presented was applied to forecast the last traded price (LTP) of a set of stocks from the technology sector, using market data between February 25^{th} and May 5^{th}, 2010. The data used were consolidated trade and quotes directly consumed from the following venues: ARCA, BATS, NYSE and INET. To measure the error obtained by the model, the following standard quadratic measure was used,

$$\varphi(y_t,\overline{y}_t) = \sqrt{\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}_i}{y_i}\right)^2},$$

where y_t is the price time series (recorded by the stock market) and $\overline{y}_t$ the synthetic LTP time series generated by the model. The generation of synthetic prices involves two stages: training and testing. In the training stage a prediction model is generated choosing a set of cointegration vectors only considering a portion of the data. The training procedure was carried out for each cointegration vector. Then, the vector with the

least error is used in the testing stage. During the testing stage, a last transaction price estimation for the rest of the data is generated tick-by-tick by using the previously calculated cointegration vectors.

The data was split into two sets for training and testing. Table 1 shows the measured errors for each stock according to the split data. The results with the lowest percentage error are highlighted. The strategy using 90% of training data had greater accuracy according to the measured error.

Strategy (%)		Error (%)					
Training	Testing	CSCO	HPQ	IBM	INTC	MSFT	VZ
50	50	16,899	22,301	20,032	41,458	12,67	15,728
60	40	3,004	6,528	7,193	3,955	8,906	8,684
70	30	3,472	3,419	4,394	4,419	5,055	4,594
80	20	2,943	3,236	4,508	4,863	5,068	5,168
90	10	**1,509**	**1,714**	**1,936**	**1,933**	**1,278**	**2,143**

5 Conclusions and Future Work

This paper presents a forecasting model based on a cointegration approach of the last traded price of stocks belonging to the same financial sector. In this way, the model allows the structural relationships between the stocks and the sector dynamism to be defined.

As usual, the accuracy of the approximation increases with the number of data in the training set. The numerical results show that the ratio 90:10 produces the best result in terms of the measured error. This result reveals that it is possible to obtain a very accurate approximation of the price of a stock when only the prices of other stocks belonging to the same financial sector are known.

Future research will include the use of a moving time frame for training in order to obtain greater accuracy by introducing more information for each prediction and also to carry out future forecasting work using a higher sampling frequency.

References

[1] Clive W. J. Granger, *Some properties of time series data and their use in econometric model specification*, Journal of Econometrics, 16(1):121–130, 1981.

[2] Clive W. J. Granger, *Developments in the study of cointegrated economic variables*, Oxford Bulletin of Economics and Statistics, 48(3):213–28, 1986.

[3] Joel Hasbrouck, *Empirical Market Microstructure: The Institutions, Economics and Econometrics of Securities Trading*, Oxford University Press, 2007.

[4] Søren Johansen, *Statistical analysis of cointegration vectors*, Journal of Economic Dynamics and Control, 12(2-3):231–254, 1988.

[5] Sandrine Lardic & Valérie Mignon, *Oil prices and economic activity: An asymmetric cointegration approach*, Energy Economics, 30(3):847 – 855, 2008.

[6] Svetlana Maslyuk & Russell Smyth, *Cointegration between oil spot and future prices of the same and different grades in the presence of structural change*, Energy Policy, 37(5):1687 – 1693, 2009.

[7] Shiv N. Mehrotra & Shashi Kant, *Use of composite forest commodity price indices for cointegration analysis*, Journal of Forest Economics, 15(4):237 – 260, 2009.

[8] J.Y. Park, *Maximum likelihood estimation of simultaneous cointegration models*, Economics working papers, School of Economics and Management, University of Aarhus, 1990.

[9] P. C. B. Phillips & S. Ouliaris, *Asymptotic properties of residual based tests for cointegration*, Econometrica, 58(1):165–193, 1990.

[10] Stephen J. Taylor, *Modelling Financial Times Series*, World Scientific Publishing Company, 2007.

[11] Wang Yu, Guo Ju'e & Xi Youmin, *Study on the dynamic relationship between economic growth and China energy based on cointegration analysis and impulse response function*, China Population, Resources and Environment, 18(4):56 – 61, 2008.

[12] J.H. Stock & M.W. Watson, *Testing for Common Trends*, Journal of the American Statistical Association, 83(404):1097–1107, 1988.

A discrete inf-sup condition for a nonconforming finite element approximation of the Stokes equations in a domain with an external cusp

Ricardo G. Durán and Eduardo M. Garau

Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina, e-mail: rduran@dm.uba.ar, egarau@santafe-conicet.gov.ar, Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón I, (1428) Buenos Aires, Argentina.

Abstract: We consider the Stokes equations in a particular cuspidal domain in $\mathbb{R}^2$. Durán and López García [4] have proved the well-posedness of the problem in suitable weighted Sobolev spaces. In this article, we analyze the nonconforming finite element discretization for this problem given by Crouzeix and Raviart [3]. We establish a mild condition on the meshes which guarantees the stability of the discrete problems.

Keywords: *inf-sup condition, nonconforming finite elements, Stokes equations, cuspidal domain*

1 Introduction

Let us consider the Stokes problem in a bounded domain $\Omega \subset \mathbb{R}^2$ which consists in finding a velocity $\mathbf{u}$ and a preasure p such that

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = 0 & \text{on } \partial\Omega. \end{cases} \tag{1}$$

It is known that, if $\mathbf{f} \in H^{-1}(\Omega)^2$, there exists a unique solution $(\mathbf{u}, p) \in H_0^1(\Omega)^2 \times L_0^2(\Omega)$, provided Ω is Lipschitz (or more generally a John domain [1]), and moreover, there holds

$$\|\mathbf{u}\|_{H_0^1(\Omega)^2} + \|p\|_{L^2(\Omega)} \leq C \|\mathbf{f}\|_{H^{-1}(\Omega)^2},$$

where the constant $C > 0$ only depends on the domain Ω.

In this article we consider the domain with an external cusp given by

$$\Omega := \{(x_1, x_2) \in \mathbb{R}^2 \mid 0 < x_1 < 1, \quad 0 < x_2 < x_1^\gamma\},$$

where the constant $1 < \gamma < 3$ is fixed. It is known that the mentioned result does not hold for this domain. However, in the recent paper [4], the well-posedness of problem (1) was proved when we consider suitable weighted Sobolev spaces, as we explain briefly now.

Let $\mathbb{V}$ and $\mathbb{Q}$ be defined by

$$\mathbb{V} := \{\mathbf{v} \in H_0^1(\Omega)^2 \mid \operatorname{div} \mathbf{v} \in L^2(\Omega, \omega^{-1})\} \qquad \text{and} \qquad \mathbb{Q} := \left\{ q \in L^2(\Omega, \omega) \mid \int_\Omega q\omega = 0 \right\},$$

where the weight ω is given by $\omega(x) := |x|^{2(\gamma-1)}$, for $x \in \mathbb{R}^2$. The variational form of problem (1) consists in finding $(\mathbf{u}, p) \in \mathbb{V} \times \mathbb{Q}$ such that

$$\begin{cases} a(\mathbf{u}, \mathbf{v}) - b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v})_{L^2(\Omega)}, & \forall\, \mathbf{v} \in \mathbb{V} \\ b(\mathbf{u}, q) = 0, & \forall\, q \in \mathbb{Q} \end{cases} \tag{2}$$

where $a : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $b : \mathbb{V} \times \mathbb{Q} \to \mathbb{R}$ are the bilinear forms defined by

$$a(\mathbf{v}, \mathbf{w}) := \int_\Omega D\mathbf{v} : D\mathbf{w}, \qquad \text{and} \qquad b(\mathbf{v}, q) := \int_\Omega q \operatorname{div} \mathbf{v}.$$

Note that $\mathbb{V}$ and $\mathbb{Q}$ are Hilbert spaces, and the induced norms are given by

$$\|\mathbf{v}\|_{\mathbb{V}}^2 := \|D\mathbf{v}\|_{L^2(\Omega)}^2 + \|\operatorname{div}\mathbf{v}\|_{L^2(\Omega,\omega^{-1})}^2, \qquad \mathbf{v}\in\mathbb{V}, \qquad \text{and} \qquad \|q\|_{\mathbb{Q}} := \|q\|_{L^2(\Omega,\omega)}, \qquad q\in\mathbb{Q}.$$

On the one hand, it is easy to check that a and b are continuous, and that a is coercive on

$$\mathbb{V}_0 := \{\mathbf{v}\in\mathbb{V} \mid b(\mathbf{v},q)=0,\ \forall\, q\in\mathbb{Q}\} = \{\mathbf{v}\in\mathbb{V} \mid \operatorname{div}\mathbf{v}=0\}.$$

On the other hand, from the results given in [4], it follows that b satisfies the inf-sup condition

$$\inf_{0\neq q\in\mathbb{Q}} \sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{\mathbb{V}}\|q\|_{\mathbb{Q}}} =: \beta > 0, \tag{3}$$

where the constant β only depends on γ.

In consequence, applying the general abstract theory for saddle point problems [2], it follows that problem (2) has a unique solution $(\mathbf{u},p)\in\mathbb{V}\times\mathbb{Q}$ which satisfies

$$\|\mathbf{u}\|_{\mathbb{V}} + \|p\|_{\mathbb{Q}} \le C\|\mathbf{f}\|_{H^{-1}(\Omega)^2},$$

where the constant $C>0$ only depends on γ.

2 Finite element discretization

Let $\mathcal{T}_h$ be a conforming triangulation such that

$$\Omega_h := \bigcup_{T\in\mathcal{T}_h} T \supset \Omega.$$

We assume that $T\cap\Omega\neq\emptyset$ for all $T\in\mathcal{T}_h$.

It is easy to prove that if $\mathbf{v}\in\mathbb{V}$, and we define $\mathbf{v}\equiv 0$ in $\Omega_h\setminus\Omega$, then $\mathbf{v}\in H^1(\Omega_h)$.

Let $\mathbb{V}_h$ and $\mathbb{Q}_h$ be the approximation spaces given by

$$\mathbb{V}_h := \Big\{\mathbf{v}_h\in L^2(\Omega_h)^2 \mid \mathbf{v}_{h|_T}\in(\mathcal{P}_1)^2,\ \mathbf{v}_h \text{ continuous at the midpoints of sides of } \mathcal{T}_h \\ \text{and } \mathbf{v}_h \text{ vanishes at the midpoints lying on } \partial\Omega_h\Big\}$$

and

$$\mathbb{Q}_h := \left\{q_h\in L^2(\Omega_h) \mid q_{h|_T}\in\mathcal{P}_0,\ \int_\Omega q_h\omega = 0\right\}.$$

Let $a_h:\mathbb{V}_h\times\mathbb{V}_h\to\mathbb{R}$ and $b_h:\mathbb{V}_h\times\mathbb{Q}_h\to\mathbb{R}$ be the bilinear forms defined by

$$a_h(\mathbf{v}_h,\mathbf{w}_h) := \int_{\Omega_h} D_h\mathbf{v}_h : D_h\mathbf{w}_h, \qquad \text{and} \qquad b_h(\mathbf{v}_h,q_h) := \int_{\Omega_h} q_h \operatorname{div}_h\mathbf{v}_h,$$

where for $\mathbf{v}_h\in\mathbb{V}_h$, we define the piecewise constant fields D_h and div_h by

$$(D_h\mathbf{v}_h)_{|_T} := D\mathbf{v}_h, \qquad \text{and} \qquad (\operatorname{div}_h\mathbf{v}_h)_{|_T} := \operatorname{div}\mathbf{v}_h, \qquad \forall\, T\in\mathcal{T}_h.$$

Thus, the discrete version of problem (2) consists in finding $(\mathbf{u}_h,p_h)\in\mathbb{V}_h\times\mathbb{Q}_h$ such that

$$\begin{cases} a_h(\mathbf{u}_h,\mathbf{v}_h) - b_h(\mathbf{v}_h,p_h) = (\mathbf{f},\mathbf{v}_h)_{L^2(\Omega)}, & \forall\,\mathbf{v}_h\in\mathbb{V}_h \\ b_h(\mathbf{u}_h,q_h) = 0, & \forall\, q_h\in\mathbb{Q}_h. \end{cases} \tag{4}$$

We consider the following norms in these discrete spaces:

$$\|\mathbf{v}_h\|_{\mathbb{V}_h}^2 := \|D_h\mathbf{v}_h\|_{L^2(\Omega)}^2 + \|\operatorname{div}_h\mathbf{v}_h\|_{L^2(\Omega,\omega^{-1})}^2, \quad \mathbf{v}_h\in\mathbb{V}_h, \qquad \text{and} \qquad \|q_h\|_{\mathbb{Q}_h} := \|q_h\|_{L^2(\Omega,\omega)}, \quad q_h\in\mathbb{Q}_h.$$

The main goal of this article is to prove the well-posedness of problem (4). Taking into account the general abstract theory for saddle point problems [2], since a_h and b_h are continuous, it is sufficient to prove that

$$a_h \text{ is coercive on } \mathbb{V}_0^h := \{\mathbf{v}_h \in \mathbb{V}_h \mid b_h(\mathbf{v}_h, q_h) = 0,\ \forall\, q_h \in \mathbb{Q}_h\}, \tag{5}$$

and that b_h satisfies the discrete inf-sup condition

$$\inf_{0\neq q_h\in\mathbb{Q}_h} \sup_{0\neq \mathbf{v}_h\in\mathbb{V}_h} \frac{b_h(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{\mathbb{V}_h}\|q_h\|_{\mathbb{Q}_h}} =: \beta_h > 0. \tag{6}$$

As an immediate consequence of the next lemma, (5) holds.

Lemma 1 *Let $\mathbf{v}_h \in \mathbb{V}_h$. If $b_h(\mathbf{v}_h, q_h) = 0$ for all $q_h \in \mathbb{Q}_h$, then $\operatorname{div}_h \mathbf{v}_h \equiv 0$. In other words,*

$$\mathbb{V}_0^h = \{\mathbf{v}_h \in \mathbb{V}_h \mid \operatorname{div}_h \mathbf{v}_h = 0\}.$$

Proof. Let $\mathbf{v}_h \in \mathbb{V}_h$ such that $b_h(\mathbf{v}_h, q_h) = 0$ for all $q_h \in \mathbb{Q}_h$. Let q_h be such that $q_{h|_T} := \frac{\int_T \operatorname{div}_h \mathbf{v}_h}{\int_{T\cap\Omega} \omega}$. Thus,

$$\int_\Omega q_h\omega = \sum_{T\in\mathcal{T}_h} q_{h|_T} \int_{T\cap\Omega} \omega = \sum_{T\in\mathcal{T}_h} \int_T \operatorname{div}_h \mathbf{v}_h = \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathbf{v}_h \cdot \mathbf{n} = 0,$$

we conclude that $q_h \in \mathbb{Q}_h$. Finally, since

$$0 = b_h(\mathbf{v}_h, q_h) = \int_{\Omega_h} q_h \operatorname{div}_h \mathbf{v}_h = \sum_{T\in\mathcal{T}_h} q_{h|_T} \int_T \operatorname{div}_h \mathbf{v}_h = \sum_{T\in\mathcal{T}_h} \frac{(\operatorname{div}_h \mathbf{v}_h)^2_{|_T}|T|^2}{\int_{T\cap\Omega}\omega},$$

it follows that $\operatorname{div}_h \mathbf{v}_h \equiv 0$. □

In order to prove the discrete inf-sup condition (6), we introduce the operator $\Pi_h : \mathbb{V} \to \mathbb{V}_h$, where $\Pi_h\mathbf{v}$ for $\mathbf{v} \in \mathbb{V}$ is defined by

$$\Pi_h\mathbf{v}(m_\ell) := \fint_\ell \mathbf{v}, \qquad \forall\, \ell \in \Sigma_{\mathcal{T}_h}.$$

Here, m_ℓ denotes the midpoint of the edge ℓ, and $\Sigma_{\mathcal{T}_h}$ the set of edges of $\mathcal{T}_h$.

Note that, for each $\mathbf{v} \in \mathbb{V}$, there hold

$$\int_\ell \Pi_h\mathbf{v} = \int_\ell \mathbf{v}, \qquad \forall\, \ell \in \Sigma_{\mathcal{T}_h},$$

and

$$(D_h\Pi_h\mathbf{v})_{|_T} = \fint_T D\mathbf{v}, \qquad \text{and} \qquad (\operatorname{div}_h \Pi_h\mathbf{v})_{|_T} = \fint_T \operatorname{div}\mathbf{v}, \qquad \forall\, T \in \mathcal{T}_h.$$

Let $C_h > 0$ be given by

$$C_h := \max_{T\in\mathcal{T}_h} C_T, \qquad \text{where } C_T := \fint_{T\cap\Omega} \omega \fint_{T\cap\Omega} \omega^{-1}. \tag{7}$$

The following properties of Π_h are the key to prove the discrete inf-sup condition (6).

Lemma 2 *There hold*

(i) $b_h(\Pi_h\mathbf{v}, q_h) = b(\mathbf{v}, q_h), \qquad \forall\, \mathbf{v} \in \mathbb{V},\ q_h \in \mathbb{Q}_h.$

(ii) $\|\Pi_h\mathbf{v}\|_{\mathbb{V}_h} \le \max(1, \sqrt{C_h})\|\mathbf{v}\|_{\mathbb{V}}, \qquad \forall\, \mathbf{v} \in \mathbb{V}.$

Proof.

(i) Let $\mathbf{v} \in \mathbb{V}$ and $q_h \in \mathbb{Q}_h$. Then,

$$b_h(\Pi_h \mathbf{v}, q_h) = \sum_{T \in \mathcal{T}_h} \int_T q_h \operatorname{div}_h \Pi_h \mathbf{v} = \sum_{T \in \mathcal{T}_h} (q_h)_{|_T} (\operatorname{div}_h \Pi_h \mathbf{v})_{|_T} |T| = \sum_{T \in \mathcal{T}_h} (q_h)_{|_T} \int_T \operatorname{div} \mathbf{v} = b(\mathbf{v}, q_h).$$

(ii) Let $\mathbf{v} \in \mathbb{V}$. We have that

$$\left\| \frac{\partial (\Pi_h \mathbf{v})_i}{\partial x_j} \right\|_T^2 = \left(\frac{\partial (\Pi_h \mathbf{v})_i}{\partial x_j} \right)_{|_T}^2 |T| = \left(\int_T \frac{\partial \mathbf{v}_i}{\partial x_j} \right)^2 |T|^{-1} \leq |T|^{-1} |T| \int_T \left(\frac{\partial \mathbf{v}_i}{\partial x_j} \right)^2 = \left\| \frac{\partial \mathbf{v}_i}{\partial x_j} \right\|_T^2,$$

and therefore,

$$\|D_h \Pi_h \mathbf{v}\|_{L^2(\Omega)}^2 \leq \|D\mathbf{v}\|_{L^2(\Omega)}^2. \tag{8}$$

On the other hand,

$$\begin{aligned} \|\operatorname{div}_h \Pi_h \mathbf{v}\|_{L^2(T \cap \Omega, \omega^{-1})}^2 &= (\operatorname{div}_h \Pi_h \mathbf{v})_{|_T}^2 \int_{T \cap \Omega} \omega^{-1} = \frac{1}{|T|^2} \left(\int_{T \cap \Omega} \operatorname{div} \mathbf{v} \right)^2 \int_{T \cap \Omega} \omega^{-1} \\ &= \frac{1}{|T|^2} \left(\int_{T \cap \Omega} \operatorname{div} \mathbf{v} \omega^{-1} \omega \right)^2 \int_{T \cap \Omega} \omega^{-1} \leq \frac{1}{|T \cap \Omega|^2} \int_{T \cap \Omega} (\operatorname{div} \mathbf{v} \omega^{-1})^2 \omega \int_{T \cap \Omega} \omega \int_{T \cap \Omega} \omega^{-1} \\ &= C_T \int_{T \cap \Omega} (\operatorname{div} \mathbf{v})^2 \omega^{-1} = C_T \|\operatorname{div} \mathbf{v}\|_{L^2(T \cap \Omega, \omega^{-1})}^2, \end{aligned}$$

and thus,

$$\|\operatorname{div}_h \Pi_h \mathbf{v}\|_{L^2(\Omega, \omega^{-1})}^2 \leq C_h \|\operatorname{div} \mathbf{v}\|_{L^2(\Omega, \omega^{-1})}^2. \tag{9}$$

Finally, taking into account (8) and (9) we conclude the proof.

□

Theorem 1 *The bilinear form b_h satisfies the discrete inf-sup condition* (6)*, with $\beta_h := \beta / \max(1, \sqrt{C_h})$.*

Proof. Let $q_h \in \mathbb{Q}_h$. Since b satisfies the (continuous) inf-sup condition (3) and $q_h \in \mathbb{Q}$, it follows that there exists $\mathbf{v} \in \mathbb{V}$ such that $b(\mathbf{v}, q_h) \geq \beta \|\mathbf{v}\|_{\mathbb{V}} \|q_h\|_{\mathbb{Q}}$. If $\mathbf{v}_h := \Pi_h \mathbf{v} \in \mathbb{V}_h$, from Lemma 2 it follows that

$$b_h(\mathbf{v}_h, q_h) = b(\mathbf{v}, q_h) \geq \beta \|\mathbf{v}\|_{\mathbb{V}} \|q_h\|_{\mathbb{Q}} \geq \frac{\beta}{\max(1, \sqrt{C_h})} \|\mathbf{v}_h\|_{\mathbb{V}_h} \|q_h\|_{\mathbb{Q}_h}.$$

□

It is important to know if there exists $\beta_0 > 0$ such that $\beta_0 \leq \beta_h$ for all h. This assertion holds if and only if there exists $C_0 > 0$ such that $C_h \leq C_0$ for all h. It is not a very strong restriction on the meshes. In fact, it is easy to construct meshes $\mathcal{T}_h$ in a way that C_h is uniformly bounded above. On the one hand, note that if T_{x_1} is the triangle with vertices $(0,0)$, $(x_1, 0)$ and (x_1, x_1^γ), then, $C_{T_{x_1}} = \fint_{T_{x_1} \cap \Omega} \omega \fint_{T_{x_1} \cap \Omega} \omega^{-1}$ is bounded above by a constant independent of x_1. On the other hand, for triangles T whose vertices are different from $(0,0)$, it is possible to generate a grading in the mesh to obtain that $\frac{\max_{x \in T} \omega(x)}{\min_{x \in T} \omega(x)}$ is also bounded above by a constant independent of T and h. As an example, we can define $\mathcal{T}_h$, with $h = 1/n$ using the method given in [5]: Consider the partition of the interval $[0,1]$ given by $x_j = \left(\frac{j}{n}\right)^{\frac{2}{3-\gamma}}$, for $0 \leq j \leq n$, and divide the domain using the vertical lines $x = x_j$, for $j \geq 1$.

References

[1] G. Acosta, R.G. Durán and M.A. Muschietti, *Solutions of the divergence operator on John domains*, Adv. Math., 206 (2) (2006), pp.373–401.

[2] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129-151.

[3] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[4] R.G. Durán and F. López García, *Solutions of the divergence and Korn inequalities on domains with an external cusp*, To appear in Ann. Acad. Sci. Fenn. Math.

[5] P. Grisvard, *Elliptic problems in nonsmooth domains*, Pitman, Boston, 1985.

APORTE DEL ANÁLISIS MULTIRRESOLUCIÓN EN UN CONTEXTO WAVELET-GALERKIN

Victoria Vampa[♭], María T. Martín[♭] y Eduardo Serrano[†]

[♭] *Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina, victoriavampa@gmail.com, mtmartin@fisica.unlp.edu.ar*

[†] *Universidad Nacional de San Martín, Buenos Aires, Argentina, eserrano@unsam.edu.ar*

Resumen: En los últimos años se ha extendido el desarrollo de métodos basados en wavelets para la resolución numérica de ecuaciones diferenciales. Utilizando una aproximación de Galerkin los autores presentaron recientemente un método en el que la solución en un nivel j determinado es expandida en funciones de escala. En el presente trabajo se propone un esquema numérico que aprovecha plenamente las ventajas del análisis multirresolución (AMR) usando wavelets sobre intervalos, permitiendo pasar de la aproximación en una escala a la siguiente más fina, con el menor esfuerzo computacional. Se exponen resultados numéricos obtenidos en distintas aplicaciones y se realizan comparaciones con otros métodos.

Palabras clave: *wavelets, MRA, B-splines*
2000 AMS Subject Classification: 21A54 - 55P54

1. P.V.C. DE SEGUNDO ORDEN

Consideremos el siguiente problema unidimensional de valores de contorno en el intervalo $I = [0, 1]$:

$$\begin{aligned} Lu = u''(x) + p(x)u'(x) + q(x)u(x) = f(x) \\ u(0) = u(1) = 0 \end{aligned} \tag{1}$$

donde $p(x)$, $q(x)$ and $f(x)$ son funciones continuas en I. La Ec.(1) puede asociarse con un problema variacional de la forma $a(u, v) = l(v)$. La forma bilineal es,

$$a(u, v) = \int_0^1 -u'(x)v'(x) + p(x)u'(x)v(x) + q(x)u(x)v(x)dx \tag{2}$$

y la forma lineal, $l(v) = \int_0^1 v(x)f(x)\,dx$, para $u,\ v \in H_0^1(0, 1)$.

En una primera etapa se determinará una aproximación $\tilde{u}$ de u en un subespacio V^0 de dimensión finita utilizando el procedimiento *Wavelet-Galerkin* desarrollado por los autores en [4]. La estructura multirresolución de las funciones B*-spline* [3] permitirá, en una segunda etapa, mejorar la aproximación $\tilde{u}$ con bajo costo computacional usando wavelets sobre intervalos.

Cabe recordar que las funciones B*-spline de orden* $m + 1$ son funciones definidas en $\mathbb{R}$, polinomiales a trozos y de grado m con $(m - 1)$ derivadas continuas. Pueden generarse recursivamente mediante convoluciones de la forma siguiente [1]:

$$\begin{aligned} \varphi_1(x) &= \chi_{[0,1]}(x) \\ \varphi_{m+1}(x) &= \varphi_m * \varphi_1(x) \end{aligned} \tag{3}$$

Teniendo en cuenta que las B-splines constituyen las funciones de escala de una estructura AMR sobre toda la recta real [3], para el problema de segundo orden Ec.(1) es necesario adaptar el AMR generado por las B-splines, al intervalo I. En este trabajo, se utilizan B-splines cúbicas, cuyo soporte está incluido en $[0, 4]$.

1.1. APROXIMACIÓN EN ESCALA j

Denotaremos $\varphi_{4,j,k} = 2^{j/2}\varphi_4(2^j x - k)$ a las splines cúbicas en $\mathbb{Z}/2^j$ trasladadas, cuyo soporte es $[2^{-j}k, 2^{-j}(k + 4)]$ y consideraremos los siguientes espacios según los lineamientos descriptos en [4]:

$$V_j^I = span\{\varphi_{4,j,k}, 0 \le k \le 2^j - 4\} \quad (4)$$

$$\widehat{V}_j^I = span\{\varphi_{4,j,k}, -3 \le k \le 2^j - 1\} \quad (5)$$

de dimensiones $(2^j - 3)$ y $(2^j + 3)$, respectivamente.

Los espacios definidos en la Ec.(4) satisfacen $V_j^I \subset V_{j+1}^I$ y además definen un *análisis multirresolución* en $L^2(I)$, [2, 3, 5]. Sin embargo, como ya fue estudiado por los autores [4], la estructura de estos espacios es insuficiente para obtener una buena aproximación, por lo cual se trabajará en el espacio ampliado $\widehat{V}_j^I$, definido en la Ec.(5), que incluye las splines interiores y las de borde. Los pasos para obtener la $\widehat{u}_j \in \widehat{V}_j^I$, son los siguientes (ver [4]):

I) *Sistema algebraico de la aproximación en la escala* j:

a) *Ecuaciones variacionales*: son las obtenidas a partir de la formulación variacional Ec.(2), considerando que la incógnita u está en $\widehat{V}_j^I$ mientras que la función de prueba v está en V_j^I. Esto conduce a un sistema rectangular de dimensión $(2^j - 3) \times (2^j + 3)$:

$$\widehat{A}_{4,j}\widehat{\alpha}_j = \widehat{b}_{4,j} \quad (6)$$

donde los elementos de la matriz del sistema, teniendo en cuenta propiedades que satisfacen las B-splines (ver [4]), tienen la siguiente expresión simplificada:

$$A_{4,j}(n,k) = -2^{2j}\varphi_8''(4+n-k) + 2^j p_j(n,k)\varphi_8'(4+n-k) + q_j(n,k)\varphi_8(4+n-k) \quad (7)$$

donde, φ_8 es la B-spline de orden 7.

b) *Ecuaciones de colocación* : obtenidas a partir del requerimiento que el residuo se anule en los extremos del intervalo y en los puntos de colocación, 2^{-j} y $1 - 2^{-j}$.

c) *Condiciones de borde*: obtenidas a partir de las condiciones que debe cumplir la solución en los extremos del intervalo.

II) *Aproximación en* $\widehat{V}_j^I$: Resolviendo el sistema se hallan los $(2^j + 3)$ coeficientes $\widehat{\alpha}_{jk}$ correspondientes a la expansión de $\widehat{u}_j$.

1.2. Utilización de las wavelets para pasar a la escala $j + 1$

Si se quiere avanzar a la escala $j+1$, una posibilidad es repetir el proceso en la nueva escala, considerando las $2^{j+1} - 3$ ecuaciones variacionales

$$\langle L\widehat{u}_{j+1}, \varphi_{j+1,n}\rangle = \langle f, \varphi_{j+1,n}\rangle \qquad 0 \le n \le 2^{j+1} - 4 \quad (8)$$

y las 6 ecuaciones de colocación y de borde similares a las utilizadas en la escala j, obteniendo la siguiente expansión,

$$\widehat{u}_{j+1} = \sum_{k=-3}^{2^{j+1}-1} \widehat{\alpha}_{j+1,k}\ \varphi_{4,j+1,k} \quad (9)$$

Otra alternativa es considerar otra base de $\widehat{V}_{j+1}$ de acuerdo a lo siguiente:

$$\widehat{u}_{j+1} = \sum_{k=0}^{2^j-4} b_{j,k}\varphi_{4,j,k} + \sum_{k=1}^{2^j} c_{j,k}\psi_{4,j,k} + \sum_{k=-3}^{-1} \widehat{\alpha}_{j+1,k}\ \varphi_{4,j+1,k} + \sum_{k=2^{j+1}-3}^{2^{j+1}-1} \widehat{\alpha}_{j+1,k}\ \varphi_{4,j+1,k} \quad (10)$$

en donde se ha utilizado la importante relación:

$$V_{j+1} = V_j \oplus W_j \quad (11)$$

proveniente del AMR considerado ([2] y [3]).

En esta nueva base las $2^{j+1}-3$ ecuaciones variacionales pueden ser reemplazadas por las siguientes:

$$\langle L\widehat{u}_{j+1}, \varphi_{j,n}\rangle = \langle f, \varphi_{j,n}\rangle \qquad 0 \le n \le 2^j - 4 \tag{12}$$

$$\langle L\widehat{u}_{j+1}, \psi_{j,n}\rangle = \langle f, \psi_{j,n}\rangle \qquad 1 \le n \le 2^j \tag{13}$$

1.3. Reducción del número de incógnitas

Escribamos la aproximación en la escala $j+1$ en el espacio $\widehat{V}_{j+1}$, de la siguiente forma:

$$\widehat{u}_{j+1} = \widehat{u}_j + [\widehat{u}_{j+1} - \widehat{u}_j] = \widehat{u}_j + \widehat{v}_j \tag{14}$$

Ya que $\widehat{v}_j \in \widehat{V}_{j+1}$, tendrá una expansión de la forma:

$$\widehat{v}_j = \sum_{k=-3}^{2^{j+1}-1} \gamma_{j+1,k}\varphi_{4,j+1,k} \tag{15}$$

Nota: Observar que $\langle \widehat{v}_j, \widehat{u}_j\rangle \neq 0$, , es decir, $\widehat{v}_j$ y $\widehat{u}_j$, no son ortogonales.

Reemplazando Ec.(14) en la Ec.(12), se tiene,

$$\langle L\widehat{u}_j, \varphi_{j,n}\rangle + \langle L\widehat{v}_j, \varphi_{j,n}\rangle = \langle f, \varphi_{j,n}\rangle \tag{16}$$

Teniendo en cuenta la ecuación variacional de la escala j, $\langle L\widehat{u}_j, \varphi_{j,n}\rangle = \langle f, \varphi_{j,n}\rangle$, la Ec.(16) consiste en un sistema lineal homogéneo de $2^j - 3$ ecuaciones y $2^{j+1}+3$ incógnitas,

$$\langle L\widehat{v}_j, \varphi_{j,n}\rangle = 0 \qquad 0 \le n \le 2^j - 4 \tag{17}$$

Reemplazando Ec.(15) en Ec.(17) se tiene la siguiente ecuación matricial:

$$D_j\gamma_{j+1} = \vec{0} \tag{18}$$

donde D_j es una matriz rectangular de productos que depende del operador L y $\gamma \in Null[D_j]$ de dimensión $2^{j+1}+3-(2^j-3) = 2^j+6$. Por lo tanto es posible construir una matriz N_j de dimensión $(2^{j+1}+3)\times(2^j+6)$, de estructura simple y recursiva tal que

$$\gamma_{j+1} = N_j\alpha_{j+1} \tag{19}$$

Luego se reduce la cantidad de incógnitas de $(2^{j+1}+3)$ a (2^j+6), en un factor 2.

Usando la descomposición (14) en las ecuaciones que restan (13), quedan 2^j ecuaciones variacionales,

$$\langle L\widehat{u}_j, \psi_{j,n}\rangle + \langle L\widehat{v}_j, \psi_{j,n}\rangle = \langle f, \psi_{j,n}\rangle \tag{20}$$

donde el primer término es conocido de la escala j. Agregando las 6 condiciones de borde, escritas en las nuevas incógnitas γ, se obtiene el sistema algebraico a resolver para obtener los coeficientes de la aproximación en la escala $(j+1)$. En resumen, la estructura del MRA permite obtener $\widehat{u}_{j+1}$ en una forma eficiente resolviendo un sistema lineal de dimensión 2^j+6 en la escala $j+1$ en el espacio V^I_{j+1} (ver Ec.(16)).

2. Ejemplo numérico: perturbación de la ecuación de reacción-difusión

Se considera el problema

$$Lu = \varepsilon\, u''(x) - u(x) = \cos^2(\pi x) + 2\varepsilon\pi^2\cos(2\pi x) \tag{21}$$

que tiene solución exacta ,

$$u(x) = \frac{exp((x-1)/\sqrt{\varepsilon}) + exp(-x/\sqrt{\varepsilon})}{1+exp(-1/\varepsilon)} - \cos^2(\pi x) \tag{22}$$

Utilizando las funciones spline, la matriz del sistema algebraico es banda, casi Toeplitz lo cual implica que los cálculos pueden realizarse mediante métodos recursivos en forma eficiente. En la Tabla 1 se muestra la convergencia para distintos valores del parámetro ε. Los errores obtenidos para las distintas escalas j evidencian una muy buena performance del método de refinamiento mediante el uso de wavelets, superando los resultados presentados por Kumar en su trabajo (ver [6]), obtenidos con un método de colocación adaptativo . La figura 1 muestra la convergencia para $\varepsilon = 2^{-15}$.

j	$\|u-\tilde{u}_j\|_{j,\infty}$, $\varepsilon = 2^{-5}$	$\|u-\tilde{u}_j\|_{j,\infty}$, $\varepsilon = 2^{-10}$	$\|u-\tilde{u}_j\|_{j,\infty}$, $\varepsilon = 2^{-15}$
6	$3{,}3 \times 10^{-5}$	$5{,}9 \times 10^{-2}$	$3{,}6 \times 10^{-1}$
7	$5{,}8 \times 10^{-6}$	$1{,}2 \times 10^{-2}$	$2{,}0 \times 10^{-1}$
8	$5{,}5 \times 10^{-7}$	$1{,}6 \times 10^{-3}$	$1{,}1 \times 10^{-1}$
9	$4{,}6 \times 10^{-8}$	$1{,}4 \times 10^{-4}$	$2{,}9 \times 10^{-2}$
10	$2{,}9 \times 10^{-9}$	$1{,}1 \times 10^{-5}$	$4{,}6 \times 10^{-3}$

Tabla 1: Errores relativos para distintas escalas

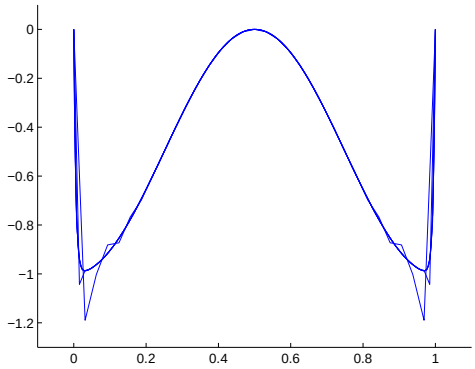


Figura 1: Solución aproximada en las escalas $j = 6, 7, 8$

3. CONCLUSIONES

Los resultados obtenidos son muy alentadores, en el sentido que abren la posibilidad de diseñar con el uso de wavelets un esquema adaptativo y acoplarlo al método de refinamiento propuesto. Los coeficientes wavelet pueden tratarse como estimadores locales del error para evitar un incremento significativo en la cantidad de wavelets, refinando la solución únicamente en las zonas de interés.

Un problema aún abierto en nuestra línea de desarrollo y en el que estamos actualmente trabajando, es la estimación eficiente de los productos escalares cuando los coeficientes de la ecuación de segundo orden, $p(x)$ y $q(x)$, varían con x.

REFERENCIAS

[1] C. K. CHUI, *An introduction to wavelets*, Academic Press, New York, 1992.

[2] S. G. MALLAT, *A Wavelet Tour of Signal Processing - The Sparse Way*, ACADEMIC PRESS - ELSEVIER - MA EEUU, 2009.

[3] D. WALNUT, *An Introduction to Wavelet Analysis*, Birkhauser, 2001.

[4] V. VAMPA, M. T. MARTÍN, AND E. SERRANO, *A hybrid method using wavelets for the numerical solution of boundary value problems on the interval*, Appl. Math. Comput., Vol 217, 7, (2010), pp.3355-3367.

[5] A. CAMMILLERI AND E. P. SERRANO, *Spline multiresolution analysis on the interval*, Latin American Applied Research, 31 (2001), pp.65-71.

[6] V. KUMAR, M. MEHRA, *Cubic spline adaptive wavelet scheme to solve singularly per- turbed reaction diffusion problems*, International Journal of wavelets, multiresolution and information processing Vol. 5 (2007), pp.317-331.

THE CONTRACTION PROPERTY OF TOTAL ERROR IN INEXACT AFEM FOR QUASI-LINEAR PROBLEMS

Carlos Zuppa[♭]

[♭]*Departamento de Matemáticas, Universidad Nacional de San Luis, Chacabuco 917, 5700 San Luis, Argentina, carlos.zuppa@gmail.com*

Abstract: In this article we prove the Contraction Property of the total error of adaptive finite element methods for nonlinear elliptic equations of monotone type with inexact finite solutions. The adaptive algorithm is based on type residual a posteriori error estimators and the Dörfler's strategy is assumed for marking.

Keywords: *nonlinear elliptic equations, adaptive finite element methods inexact solutions.*
2000 AMS Subject Classification: 65N30 - 65N12

1 INTRODUCTION

Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with a Lipschitz boundary. In particular, we suppose that Ω is a polygonal domain if $d = 2$ and a polyhedral domain if $d = 3$.

Roughly speaking, we shall consider a class of Dirichlet problems of the form

$$\begin{aligned} -\text{div } (\mathbf{A}(x, u, \nabla u)\, \nabla u) &= f(x), \qquad \text{in } \Omega, \\ u \,|\, \partial\Omega &= 0. \end{aligned} \tag{1}$$

Main example: nonlinear stationary conservation laws

$$-\text{div } \big(\alpha(., |\nabla u|^2)\nabla u\big) = f,$$

where α is a suitable function. This kind of problems arise in many practical situations, for example, in shock-free airfoil design, coarse grained porous media, and in some glaciological problems [3].

We consider the adaptive cycle (AFEM)

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE}$$

in order to approximate the exact solution in a controled way. In [1] we have analysed the optimal complexity of the standard adaptive loop of this form, based on classical residual-type a posteriori error estimators, where the Galerkin discretization for problem (1) is considered. At this point, it is important to remark that the discrete problem is also *nonlinear*, and for our analysis we have assumed that it can be solved exactly in every mesh $\mathcal{T} \in \mathbb{T}$. However, this assumption is usual even though in practice, even for discrete *linear* problems, we *compute* only approximations to the solution of discrete problems. In this work we are concern with inexact solutions in the AFEM cycle and we prove a contraction property of the total error which garanties the convergence and it is a useful tool in proving the quasi-optimal complexity of AFEM.

We state now precisely the continuous problem that we study and mention some of its properties (see [1] for details).

$||\cdot||_1 := ||\nabla \cdot||_2$.

Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be defined by $a(u; v, w) := \int\limits_\Omega \mathbf{A}(x, u, \nabla u)\, \nabla v \cdot \nabla w \, dx$, and the operator $A : H_0^1(\Omega) \to \big(H_0^1(\Omega)\big)'$ defined by $< A(u), v >:= a(u; u, v), \forall v \in H_0^1(\Omega)$. We assume that A is strongly monotone with constant $\delta > 0$, and locally Lipschitz.

The function $\mathcal{J} : H_0^1(\Omega) \to \mathbb{R}$ defined by $\mathcal{J}(u) := \int\limits_0^1 < A(su), u > ds, \forall u \in H_0^1(\Omega)$, is Gateaux differentiable and $\mathcal{J}'(u) = A(u)$, for all $u \in H_0^1(\Omega)$.

We consider now the functional $\mathcal{J}_f : H_0^1(\Omega) \to \mathbb{R}$ defined by

$$\mathbf{R}(u) := \mathcal{J}_f(u) := \mathcal{J}(u) - L_f(v), \forall v \in H_0^1(\Omega),$$

where

$$L_f(v) := \int_\Omega fv.$$

It follows easily that the minimum problem

$$\mathcal{J}_f(u) = \min!, \forall u \in H_0^1(\Omega),$$

and the operator equation

$$\mathbf{R}(u) = A(u) - L_f = 0,$$

are equivalent. Both problems have a unique solution $u \in H_0^1(\Omega)$.

2 THE DISCRETE PROBLEM

In order to define the discrete approximations we will consider triangulations $\mathcal{T}$ of the domain Ω, that is, partitions of Ω into d-simplices. Let $\mathbb{V}_\mathcal{T}$ be the finite element space consisting of continuous functions vanishing on $\partial\Omega$ which are polynomials of degree ≤ 1 in each element of $\mathcal{T}$, i.e,

$$\mathbb{V}_\mathcal{T} := \{v \in H_0^1(\Omega) : \quad v|_T \in \mathcal{P}_1(T), \quad \forall\, T \in \mathcal{T}\}. \tag{2}$$

Obviously, $\mathbb{V}_\mathcal{T} \subset H_0^1(\Omega)$ and if $\mathcal{T}_*$ is a refinement of $\mathcal{T}$, then $\mathbb{V}_\mathcal{T} \subset \mathbb{V}_{\mathcal{T}_*}$.

The discrete problem is: find $U_\mathcal{T} \in \mathbb{V}_\mathcal{T}$ such that

$$a(U_\mathcal{T}; U_\mathcal{T}, V) = L_f(V), \qquad \forall V \in \mathbb{V}_\mathcal{T}. \tag{3}$$

From now on, whenever we write $A \lesssim B$ we mean that $A \leq C\,B$ with a constant C that may depend on $\mathbf{A}$, the domain Ω and the regularity of the triangulation $\mathcal{T}$, but not on other properties of $\mathcal{T}$ such as element size or uniformity.

An a posteriori estimation of the energy error $\eta_\mathcal{T}(U_\mathcal{T})$ is defined in [1].

2.1 AN A POSTERIORI ERROR ESTIMATOR

Given $U \in \mathbb{V}_\mathcal{T}$, we set $e(U) := U - u$. Then, by strong monotonicity

$$\delta\, ||\, e(U)||_1^2 \lesssim \langle \mathbf{R}(U), e(U)\rangle\,,$$

and we can write now

$$\langle \mathbf{R}(U), e(U)\rangle = \langle \mathbf{R}(U), e(U) - \mathcal{P}_\mathcal{T}(e(U))\rangle + \langle \mathbf{R}(U), \mathcal{P}_\mathcal{T}(e(U))\rangle\,,$$

where $\mathcal{P}_\mathcal{T} : H_0^1(\Omega) \to \mathbb{V}_\mathcal{T}$ is the Scott-Zang operator.

From the first term we get the standard Residual Estimator

$$\eta_\mathcal{T}^2(U) =_{T\in\mathcal{T}} \eta_\mathcal{T}^2(U, T).$$

and we arrive easily at the estimation

$$||\, e(U)||_1 \lesssim \eta_\mathcal{T}(U) + ||\mathbf{R}(U)||_{\mathbb{V}'_\mathcal{T}}\,. \tag{4}$$

3 THE INEXACT ADAPTIVE PROCESS IAFEM

Let u denotes the exact weak solution of problem 1, and U_k, $\{\eta_k(T)\}_{T\in\mathcal{T}_k}$, $\mathcal{M}_k$, $\mathcal{T}_k$ will denote the outputs of the corresponding modules SOLVE, ESTIMATE, MARK and REFINE of the Adaptive Algorithm when iterated after starting with a given initial mesh $\mathcal{T}_0$. Modules ESTIMATE, MARK and REFINE are the same as in [1].

We recall that (3) is a nonlinear problem that it is solved by an iterative method which we call QK. We require from the method the key condition

$$||\mathbf{R}(U_{k,n})||_{(H_0^1(\Omega))'} \lesssim ||U_{k,n} - U_{k,n-1}||_1. \tag{5}$$

Several efficient iterative methods satisfy this requirement (see [2]).

The module SOLVE. Given $(\mathcal{T}_{k-1}, U_{k-1})$ and $(\mathbb{V}_k, \mathcal{T}_k)$ by refining, we set $U_{k,0} = U_{k-1}$ and run several steps of the iterative method $U_{k,n} = QK(U_{k,n-1})$ until the two conditions stated bellow are satisfied.

3.0.1 Condition I

Definition 1 *$U_{k,n}$ satisfies the $\eta-$Condition iff*

$$||U_{k,n} - U_{k,n-1}||_1 \leq \eta_{\mathcal{T}_k}(U_{k,n}).$$

If $U_{k,n}$ satisfies this condition, we get from (4) and (5) the estimate

$$||\, e(U_{k,n})||_1 \lesssim \eta_{\mathcal{T}}(U_{k,n}). \tag{6}$$

In order to state the second condition we need first another fundamental estimate related to the hessian of $\mathcal{J}_f$.

3.1 THE HESSIAN OF $\mathcal{J}_f$.

$v, w \in W \subseteq H_0^1(\Omega)$. Define $\phi(t) := (1-t)w + tv$. Then, by Taylor

$$\mathcal{J}_f(v) - \mathcal{J}_f(w) = \left\langle \mathcal{J}_f'(w), v-w \right\rangle + \int_0^1 \int_\Omega D_2^2\gamma(\cdot, \nabla\phi(t))\, \nabla(v-w)\cdot\nabla(v-w)(1-t)\, dxdt. \tag{7}$$

Since $D_2^2\gamma$ is uniformly elliptic, there exist constant C_A, c_A such that

$$\frac{c_A}{2}||\nabla(v-w)||_2^2 \leq \int_0^1 \int_\Omega D_2^2\gamma(\cdot, \nabla\phi(t))\, \nabla(v-w)\cdot\nabla(v-w)(1-t)\, dxdt \leq \frac{C_A}{2}||\nabla(v-w)||_2^2. \tag{8}$$

the following result will be a basic tool in the quasi-optimality of a inexact adaptive FEM.

Corollary 1 *If* $\left|\left|\mathcal{J}_f'(w)\right|\right|_{W'} \leq \sigma\, ||\nabla(v-w)||_2$, $\sigma < \frac{c_A}{2}$, *then*

$$\left|\left\langle \mathcal{J}_f'(w), v-w \right\rangle\right| \leq \sigma\, ||\nabla(v-w)||_2^2,$$

and

$$\widetilde{c}_A\, ||\nabla(v-w)||_2^2 \leq \mathcal{J}_f(v) - \mathcal{J}_f(w) \leq \widetilde{C}_A\, ||\nabla(v-w)||_2^2,$$

with $0 < \widetilde{c}_A = \frac{c_A}{2} - \sigma < \widetilde{C}_A = \frac{C_A}{2} + \sigma$.

3.1.1 Condition II

We need to estimate $\left|\left|\mathcal{J}_f'(U_{k,n})\right|\right|_{\mathbb{V}_k'}$. We use

$$\left|\left|\mathcal{J}_f'(U_{k,n})\right|\right|_{\mathbb{V}_k'} \leq \left|\left|\mathcal{J}_f'(U_{k,n}) - \mathcal{J}_f'(U_{k,n-1})\right|\right|_{\mathbb{V}_k'} + C_A \ ||d_n||_1 \,,$$

and we note that

$$\left|\left|\mathcal{J}_f'(U_{k,n}) - \mathcal{J}_f'(U_{k,n-1})\right|\right|_{\mathbb{V}_k'} \leq \left|\left|\alpha(\cdot, |\nabla U_{k,n}|^2)\nabla U_{k,n} - \alpha(\cdot, |\nabla U_{k,n-1}|^2)\nabla U_{k,n-1}\right|\right|_2 .$$

Note that the expression

$$\sigma(k,n) := \left|\left|\alpha(\cdot, |\nabla U_{k,n}|^2)\nabla U_{k,n} - \alpha(\cdot, |\nabla U_{k,n-1}|^2)\nabla U_{k,n-1}\right|\right|_2 + C_A \ ||\, d_n||_1 \,,$$

is easily computable.

Definition 2 *Given a fixed $\sigma < \frac{c_A}{2}$, we said that $U_{k,n}$ satisfies the $\sigma-$Condition iff*

$$\sigma(k,n) \leq \sigma \ ||U_{k,n} - U_{k-1}||_1 \,.$$

In this case we have

$$\widetilde{c}_A \, ||\nabla(U_k - U_{k-1})||_2^2 \leq \mathcal{J}_f(U_{k-1}) - \mathcal{J}_f(U_k) \leq \widetilde{C}_A \, ||\nabla(U_k - U_{k-1})||_2^2. \tag{9}$$

Now, we set $U_k = U_{k,n}$ if $U_{k,n}$ satisfies both $\eta-$Condition and $\sigma-$Condition. Clearly, these two conditions are reached at a finite number of iterations.

We denote $\varepsilon_k := \mathcal{J}_f(U_k) - \mathcal{J}_f(u)$.

In the same way as in [1], we can prove now our fundamental result

Theorem 1 *Assume that at each stage k of AFEM, U_k is an approximate solution of the finite problem (3) which satisfies conditions I and II. Then, there exist $0 < \alpha < 1, \gamma > 0$ such that*

$$\varepsilon_k + \gamma \, \eta_k^2 \leq \alpha \left(\varepsilon_{k-1} + \gamma \, \eta_{k-1}^2\right).$$

REFERENCES

[1] E. M. GARAU, P. MORIN, AND C. ZUPPA, *Optimal complexity of AFEM for quasi-linear problems*, submited to NMTMA (2010).

[2] J. SPEDALETTI AND C. ZUPPA, *A quasi-Kačanov iterative method for quasi-linear problems*, MACI2011, Bahía Blanca, Argentina (2011).

[3] E. ZEIDLER, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, (1990), Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.

A QUASI-KAČANOV ITERATIVE METHOD FOR QUASI-LINEAR PROBLEMS

Juan Spedaletti[♭] and Carlos Zuppa[♭]

[♭]*Departamento de Matemáticas, Universidad Nacional de San Luis, Chacabuco 917, 5700 San Luis, Argentina, carlos.zuppa@gmail.com*

Abstract: In solving problems like nonlinear stationary conservation laws the Kačanov iterative method have been revealed very efficient. This method requires, however, a key condition on the coefficient function. In this paper we modify the Kačanov method with a line search procedure in order to deal with the general case.

Keywords: *nonlinear elliptic equations, Kačanov iterative method.*
2000 AMS Subject Classification: 65N30 - 65N12

1 INTRODUCTION

Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with a Lipschitz boundary. Roughly speaking, we shall consider a class of Dirichlet problems of the form

$$\begin{aligned} -\text{div}\,\left(\alpha(.,|\nabla u|^2)\nabla u\right) &= f(x), \qquad \text{in } \Omega, \\ u\,|\,\partial\Omega &= 0. \end{aligned} \tag{1}$$

where α is a suitable function. If $\alpha' \leq 0$, the secant method of Kačanov produces a convergent sequence of approximation of the solution.

The aim of this paper is to slightly modify the secant method by a line search procedure in order to deal with the general case.

2 SETTINGS

We state now precisely the continuous problem that we study and mention some of its properties. We consider the problem

$$\begin{aligned} -\text{div}\,\left(\alpha(.,|\nabla u|^2)\nabla u\right) &= f, \\ u|\partial\Omega &= 0. \end{aligned}$$

with a suitable function α (see [1], [2]). This equation comprehends many completely different physical problems in hydridynamics and gas dynamics, electrostatic, heat conductions, elasticity and plasticity, etc.

Given $W \subseteq H_0^1(\Omega)$, we consider the problem: find $u \in W$ such that

$$a(u;u,v) = L_f(v), \qquad \forall v \in W, \tag{2}$$

where

$$L_f(v) := \int_\Omega f v\, dx.$$

An iterative procedure which produces a convergent sequence $u_n \to u$ must be used to deal with this nonlinear problem.

If $\partial\alpha/\partial t \leq 0$, we have the key condition

$$\mathcal{J}_f(v) - \mathcal{J}_f(w) \leq \frac{1}{2}(a(w;v,v) - a(w;w,w)), \qquad \forall w, v \in W \tag{3}$$

An iterative method for solving (6) is the following. given an initial guess $u_0 \in W$, we define a sequence of iterates $\{u_n\} \subset W$ by

$$a(u_{n-1}; u_n, v) = L_f(v), \qquad \forall v \in W, \tag{4}$$

for $n = 1, 2, ...$ Notice that in (8) we solve a sequence of linear problems instead of the nonlinear problem (6).

The key condition can be used to prove that this secant method of Kačanov [2] produces a $\mathcal{J}_f-$strictly decreasing sequence which converge to u. Furthermore,

$$||u - u_n||_1 \lesssim ||u_n - u_{n-1}||_1 .$$

3 AN ITERATIVE METHOD AND ITS CONVERGENCE

By (8) we have

$$a(u_{n-1}; u_n, v) - a(u_{n-1}; u_{n-1}, v) = -(a(u_{n-1}; u_{n-1}, v) - L_f(v)), \qquad \forall v \in W.$$

Setting $d_n := u_n - u_{n-1}$, the last equation is equivalent to

$$a(u_{n-1}; d_n, v) = -\left\langle \mathcal{J}_f'(u_{n-1}), v \right\rangle, \qquad \forall v \in W.$$

That is, the Kačanov method is a Newton's like method where $u_n = u_{n-1} + d_n$.

If the key condition $\partial\alpha/\partial t \leq 0$ is not satisfied, we can proceed almost in the same way with the addition of a line search procedure. Roughly speaking, the algorithm runs as follows:

First solve

$$a(u_{n-1}; \overline{u}_n, v) = L_f(v), \qquad \forall v \in W,$$

and define $d_n := \overline{u}_n - u_{n-1}, t_n := 1$.

We note that

$$||d_n||_1^2 \lesssim a(u_{n-1}; d_n, d_n) = -\left\langle \mathcal{J}_f'(u_{n-1}), d_n \right\rangle .$$

Then, $\left\langle \mathcal{J}_f'(u_{n-1}), d_n \right\rangle$ is a strictly negative number.

Now, we decrease t_n until

$$\left\langle \mathcal{J}_f'(u_{n-1} + t_n d_n), d_n \right\rangle < 0.$$

The rest of the section is dedicated to formalize this algorithm and show that it produces a $\mathcal{J}_f-$strictly decreasing sequence which converge to u in a controled way.

We call this procedure Quasi-Kačanov method, and we write $u_n = QK(u_{n-1})$ the process of obtain u_n from u_{n-1}.

We state the procedure in a more general form which applies in other situation. We make the following assumptions:

(i) The functional $\mathcal{J} : X \to \mathbb{R}$ is differentiable on the real Hilbert space X.

(ii) $\mathcal{J}$ is coercive, i.e.

$$\lim_{||u||_X \to \infty} \frac{\mathcal{J}(u)}{||\, u \,||_X} = \infty.$$

(iii) $\mathbf{R} := D\mathcal{J} : X \to X'$ is strongly monotone. That is,

$$\langle \mathbf{R}(u) - \mathbf{R}(v), u - v \rangle \geq \delta \,||\, u - v \,||_X^2, \qquad \forall u, v \in X,$$

where $\langle \cdot, \cdot \rangle$ is the dual pairing between X and its dual space X'.

Then,

(a) The minimum problem

$$\mathcal{J}(u) = \min!, \qquad u \in X, \tag{5}$$

and the operator equation

$$\mathbf{R}(u) = 0, \qquad u \in X, \tag{6}$$

are equivalent.

(b) Both problems have a unique solution $u \in X$ ([2]).

Assumption:

(iv) $\mathbf{R}$ is locally Lipschitz. That is, for every real number $R > 0$, there exists a positive constant $L(R)$ such that

$$|| \mathbf{R}(u) - \mathbf{R}(v) ||_{X'} \leq L(R) || u - v ||_X, \qquad \forall u, v \in X : || u ||_X, || v ||_X \leq R.$$

Given an initial $u_0 \in X$, we shall produce a sequence $u_n, n = 1, 2, ...$, such that $u_n \to u$.

Taking into account the coercive condition (ii) and the fact that the iterative method is $\mathcal{J}-$decreasing, it will be evident that we are dealing always with a fixed Lipschitz constant $L(R)$. Then, we omit any reference to R and we proceed as if $\mathbf{R}$ being globally Lipschitz.

We now describe the algorithm in detail. given a current iterate u_n, we first solve the linear problem

$$G_n d_n = -\mathbf{R}(u_n),$$

where $G_n : X \to X'$ is a linear coercive operator.

We shall assume that the sequence of operators $\{G_n\}, n = 1, 2, ...$, are uniformly bounded and uniformly coercive. That is,

(v) There exist constant $A, B > 0$ such that

$$\begin{aligned} || G_n || &\leq A, \qquad \forall n = 1, 2..., \\ \langle G_n v, v \rangle &\geq B || v ||_X^2, \qquad \forall n \text{ and } \forall v \in X. \end{aligned}$$

Then, we have

$$\langle G_n d_n, d_n \rangle = - \langle \mathbf{R}(u_n), d_n \rangle \geq B || d_n ||_X^2,$$

and $\langle \mathbf{R}(u_n), d_n \rangle$ is strictly negative.

Now, because a full step $u_{n+1} := u_n + d_n$ in the obtained Newton direction may not satisfy the decreasing condition $\mathcal{J}(u_{n+1}) < \mathcal{J}(u_n)$, the second step is a line search procedure. We consider

$$\gamma(t) := \mathcal{J}(u_n + t d_n), \qquad t \in \mathbb{R}.$$

Then

$$\gamma'(t) = \langle \mathbf{R}(\gamma(t)), d_n \rangle .$$

Since $\mathbf{R}$ is strictly monotone, the function γ' is monotone increasing and we know that $\gamma'(0) < 0$. Our line search procedure runs as follows

---

$t_n \leftarrow 1$;
while $(\gamma'(t_n) > 0)$ **do**
  $t_n \leftarrow t_n/2$;
end while

We have to prove that the algorithm always terminates.

There exists a number $\alpha > 0$ such that

$$t_n \geq \alpha, \qquad \forall n = 1, 2, ...$$

For any $t > 0$ we have

$$\begin{aligned} \langle \mathbf{R}(\gamma(t)), d_n \rangle &= \langle \mathbf{R}(\gamma(t)) - \mathbf{R}(\gamma(0)) + \mathbf{R}(\gamma(0)), d_n \rangle \\ &= - \langle G_n d_n, d_n \rangle + \langle \mathbf{R}(\gamma(t)) - \mathbf{R}(\gamma(0)), d_n \rangle \\ &\leq || d_n ||_X^2 \, (-B + Lt) . \end{aligned}$$

We see that $\alpha = 2^{-k}$, where k is the first nonnegative integer such that $2^{-k} < B/L$, satisfies the requirement.

We define now

$$u_{n+1} = u_n + t_n d_n.$$

Moreover, writing

$$\gamma(0) - \gamma(t_n) = -\int_0^{t_n} \langle \mathbf{R}(\gamma(s)), d_n \rangle \, ds,$$

and proceeding exactly as above we arrive at

$$\gamma(0) - \gamma(t_n) \geq t_n \, || \, d_n \, ||_X^2 \, \left(B - \frac{L}{2} t_n \right).$$

It follows easily that there exists $\beta > 0$ such that

$$\mathcal{J}(u_n) - \mathcal{J}(u_{n+1}) \geq \beta \, || \, d_n \, ||_X^2 \, .$$

In particular, $\{\mathcal{J}(u_n)\}$ is a bounded bellow sequence which must have a limit and this also implies

$$\lim_{n \to \infty} || \, d_n \, ||_X = 0,$$

and

$$\lim_{n \to \infty} || \, u_{n+1} - u_n \, ||_X = 0.$$

We proceed now to estimate $|| \, \mathbf{R}(u_n) \, ||_{X'}$. We can write

$$\mathbf{R}(u_n) = \mathbf{R}(u_n) - \mathbf{R}(u_{n-1}) + G_{n-1} d_{n-1},$$

which implies

$$|| \, \mathbf{R}(u_n) \, ||_{X'} \leq || \, \mathbf{R}(u_n) - \mathbf{R}(u_{n-1}) \, ||_{X'} + A \, || \, d_{n-1} \, ||_X. \tag{7}$$

This inequality is util when we want to estimate $|| \, \mathbf{R}(u_n) \, ||_{X'}$ without the recourse to Lipschitz constant L.

Using the estimate above we get easily

$$|| \, \mathbf{R}(u_n) \, ||_{X'} \lesssim || \, u_n - u_{n-1} \, ||_X.$$

It remains to show that effectively $u_n \to u$. But this follows immediately by the strongly monotonicity.

$$\delta \, || \, u_n - u \, ||_X^2 \leq \langle \mathbf{R}(u_n), u_n - u \rangle \leq || \, \mathbf{R}(u_n) \, ||_{X'} \, || \, u_n - u \, ||_X,$$

and

$$\delta \, || \, u_n - u \, ||_X \leq || \, \mathbf{R}(u_n) \, ||_{X'} \lesssim || \, u_n - u_{n-1} \, ||_X.$$

REFERENCES

[1] E. M. GARAU, P. MORIN, AND C. ZUPPA, *Optimal complexity of AFEM for quasi-linear problems*, in preparation (2010).

[2] E. ZEIDLER, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, (1990), Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.

ERROR ESTIMATES FOR THE FINITE ELEMENT APPROXIMATION OF A CLASS OF BOUNDARY OPTIMAL CONTROL SYSTEMS

Pablo Gamallo[♭], Erwin Hernández[†] and Andres Peters[†,‡]

[♭]*Tecnologías Avanzadas Inspiralia (ITAV), PERA group, Estrada 10 5-B, 28034, Madrid, España Pablo.Gamallo@itav.es*
[†]*AM2V, Departamento de Matemática, Universidad Técnica Federico Santa María, Avda España 1680, Valparaíso, Chile erwin.hernandez@usm.cl, www.usm.cl*
[‡]*The Hamilton Institute, The National University of Ireland, Maynooth, Co Kildare, Ireland Andres.Peters@nuim.ie*

Abstract: In this paper we consider an application of the abstract error estimate for a class of optimal control systems governed by a linear partial differential equation. The control is applied in the boundary and we consider both, Dirichlet and Neumann optimal control problems. A finite element method is proposed in order to approximate the solution of the control problem, considering a discretization of the variational inequality resulting from the optimality conditions; this approach is known as the classical one. We obtain optimal order error estimates for the control variable and numerical examples are included to illustrate the results.

Keywords: *Optimal Control, Finite Element Approximation*
2000 AMS Subject Classification: 65N30

1 INTRODUCTION

According to [1], let us consider the following spaces

- Space of states V: a Hilbert space with inner product $(\cdot,\cdot)_V$.
- Space of controls $\mathcal{U}$: a Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{U}}$.
- Space of observations $\mathcal{H}$: a Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{H}}$.

The set of admissible controls will be a non-empty convex set $\mathcal{U}_{\rm ad} \subset \mathcal{U}$ and V' will stand for the dual space of V. For $S = V, \mathcal{U}$, or $\mathcal{H}$, we will denote by $\|\cdot\|_S$ the induced norm.

We consider the state y as the solution of the elliptic linear partial differential equation, with Neumann or Dirichlet boundary condition, posed in a functional framework as follow: Let $\mathbf{A} \in \mathcal{L}(V, V')$ be the linear continuous operator defined as

$$\mathbf{A}y := a(y,\cdot) : V \to \mathbb{R},$$

where $a : V \times V \to \mathbb{R}$ is a continuous bilinear form. Given $f \in V'$ and a control $u \in \mathcal{U}$, the state of the problem will be given by the solution y of

$$\mathbf{A}y = f + \mathbf{B}u.$$

Here $\mathbf{B} : \mathcal{U} \to V'$ is a continuous linear operator, defined by taking into account the boundary conditions.

The optimal control problem we are interested in is

Find $u_{\rm op} \in \mathcal{U}_{\rm ad}$ *such that*

$$J(u_{\rm op}) = \inf_{u \in \mathcal{U}_{\rm ad}} J(u).$$

The cost function will be given by the quadratic functional

$$J(u) = \frac{1}{2}\|\mathbf{C}y(u) - y_d\|^2_{\mathcal{H}} + \frac{\nu}{2}\|u\|^2_{\mathcal{U}},$$

where y_d denotes the desired state, $\nu > 0$ is the cost, and the continuous linear operator $\mathbf{C} : V \to \mathcal{H}$ is known as the observation operator.

Now, we consider a consistent and regular family of discrete spaces $\{V_h\}_h$ of V and introduce the operator $\mathbf{A}_h : V_h \to V_h'$

$$\mathbf{A}_h y_h := a(y_h, \cdot) : V_h \to \mathbb{R},$$

which is the discrete counterpart of $\mathbf{A}$. Let $\{\mathcal{U}_{h'}\}_{h'}$ be a consistent and regular family of discrete spaces, we set $\mathcal{U}_{\mathrm{ad}h'} = \mathcal{U}_{h'} \cap \mathcal{U}_{\mathrm{ad}}$. Additionally we assume that $\mathcal{U}_{\mathrm{ad}h'}$ is not empty for all h'. Then, in the same way of the continuous case, the discrete state of the system $y_h(u_{h'}) \in V_h$, in terms of the control $u_{h'} \in \mathcal{U}_{\mathrm{ad}h'}$, will be the solution of

$$\mathbf{A}_h y_h(u_{h'}) = \mathbf{R}_h f + \mathbf{R}_h \mathbf{B} u_{h'}, \tag{1}$$

and thus, for a given control $u_{h'} \in \mathcal{U}_{\mathrm{ad}h'}$, we will write

$$y_h(u_{h'}) = y_h(0) + \mathbf{A}_h^{-1} \mathbf{R}_h \mathbf{B} u_{h'},$$

being $y_h(0) = \mathbf{A}_h^{-1} \mathbf{R}_h f$, where $\mathbf{R}_h$ is a restriction operator in the dual space from V' onto V_h'.

The optimal control problem we are interested in is

Find $u_{\mathrm{op}_{h,h'}} \in \mathcal{U}_{\mathrm{ad}h'}$ such that

$$J_h(u_{\mathrm{op}_{h,h'}}) = \inf_{u_{h'} \in \mathcal{U}_{\mathrm{ad}h'}} J_h(u_{h'}).$$

where

$$J_h(u_{h'}) := \frac{1}{2} \|\mathbf{C} y_h(u_{h'})\|_{\mathcal{H}}^2 + \frac{\nu}{2} \|u_{h'}\|_{\mathcal{U}}^2.$$

According to section 2 of [1], the main abstract result on the approximation error is:

Theorem 1 *There exist a constant C independent of the discrete space $\mathcal{U}_{h'}$ such that*

$$\begin{aligned} \|u_{\mathrm{op}} - u_{\mathrm{op}_{h,h'}}\|_{\mathcal{U}} \quad \leq \quad & C \left(\inf_{v_{h'} \in \mathcal{U}_{h'}} \|u_{\mathrm{op}} - v_{h'}\|_{\mathcal{U}} \right. \\ & + \sup_{w_{h'} \in \mathcal{U}_{h'}} \frac{|\pi(u_{\mathrm{op}}, w_{h'}) - \pi_h(u_{\mathrm{op}}, w_{h'})|}{\|w_{h'}\|_{\mathcal{U}}} \\ & \left. + \sup_{w_{h'} \in \mathcal{U}_{h'}} \frac{|L(w_{h'}) - L_h(w_{h'})|}{\|w_{h'}\|_{\mathcal{U}}} \right). \end{aligned}$$

Then, we use this abstract result in the case of elliptic linear partial differential equation, with Neumann or Dirichlet boundary condition [2, 3]

ACKNOWLEDGMENTS

This work has been partially supported by Programa Basal CMM. U. de Chile and by Conicyt-Chile thought FONDECYT No. 1100490.

REFERENCES

[1] P. Gamallo and E. Hernández. Error estimates for the approximation of a class of optimal control systems governed by linear PDEs. *Numer. Funct. Anal. Optim.* **30** (2009), 523–547.

[2] S. May and R. Rannacher and B. Vexler. Error Analysis for a Finite Element Approximation of Elliptic Dirichlet Boundary Control Problems. Submitted.

[3] E. Casas and M. Mateos. Error estimates for the numerical approximation of Neumann control problems. *Comput. Optim. Appl.* **39** (2008), 265–295.

ERROR ANALYSIS OF A MESHFREE METHOD WITH DIFFUSE DERIVATIVES

Mauricio Osorio[♭] and Donald French[†]

[♭]*Escuela de Matematicas, Universidad Nacional de Colombia, Apartado Aereo 3840, Medellín, Colombia, maosorio@unal.edu.co*
[†]*Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA, french@math.uc.edu*

Abstract: A meshfree method with diffuse derivatives and a penalty stabilization is developed. An error analysis for the approximation of the solution of an elliptic differential equation with Neumann boundary conditions in several dimensions is provided. Theoretical and numerical results show that the approximation error and the convergence rate are better than the diffuse element method.

Keywords: *Meshfree methods, Diffuse derivatives, Error estimates.*
2000 AMS Subject Classification: 65N15

1 INTRODUCTION

Numerical methods based on moving least square (MLS) approximations and Galerkin formulations form a popular class of meshfree schemes. However, the high computational expense in the evaluation of the shape functions and their derivatives are drawbacks to the Galerkin approach. An alternative for the computation of derivatives, the diffuse derivative, was used by Nayroles in [9] in the diffuse element method (DEM). In the diffuse derivative approximation, only the derivatives of the polynomial basis need to be included in computing the gradients of the local field variables. Belytshko *et al.* [3, 7] argued that diffuse derivatives are not attractive in Galerkin methods because they degrade the accuracy due to their lack of integrability. However, recently, the diffuse derivative has been used in a class of novel meshfree methods (Huerta *et al* [5]) for Stokes problems. Because of their simplicity, diffuse derivatives, unlike the full derivatives, retain the same subspace structure as their defining functions. This special feature allowed Huerta *et al* [5] to circumvent the complicated incompressibility constraint and define a class of divergence free meshfree approximation functions. Beyond fluid mechanics, we think this new approach could be used to enhance the common mixed method approach.

In this work we present a complete mathematical analysis of a Galerkin meshfree method with diffuse derivatives. We introduce a penalty term that stabilizes the traditional DEM and improves its accuracy. We call this approach penalized-DEM (P-DEM), and apply it to an elliptic differential equation with Newmann boundary conditions. We use approximation results from Kim and Kim [6] on L^p-norm convergence of diffuse derivatives, and Huerta *et al.* [5] on the pointwise accuracy of diffuse derivatives. The analysis in Armentano [1, 2] was crucial in the development of our penalty term. We will show that our penalized diffuse derivative Galerkin approach (P-DEM) is accurate and its rate of convergence increases as the approximating polynomial degree increases and the width of the support domain (R) decreases. A drawback for our approach, which renders our scheme suboptimal, is that the penalty term, though small, is not zero when applied to a true BVP solution.

2 MOVING LEAST SQUARES APPROXIMATIONS (MLS)

Let $\Lambda = \{x_1, x_2, ..., x_N\}$ be a set of N distinct points inside and in the boundary of $\Omega \subset \mathbb{R}^n$ which is an open and bounded set with Lipschitz boundary $\partial\Omega$ and $u_1, u_2, ...u_N$ be the values of an unknown scalar function $u(x)$ at the points in Λ (i.e. $u_i = u(x_i), 1 \leq i \leq N$). Also let $R > 0$ and consider a positive even weight function $W(x)$ with compact support in $\overline{B_1(0)}$ and $\int_{\mathbb{R}^n} W\, dx \cong 1$. Define $W_R(x) := W(x/R)$ and note W_R has compact support in $\overline{B_R(0)}$.

Let $m << N$ and $\mathbf{p}(z) = \{p_0(z), p_1(z), p_2(z), ..., p_s(z)\}^T$ be a basis of the subspace of polynomials of degree less or equal than m (denoted $\mathcal{P}_m$) in $\mathbb{R}^n$, placed in multi-index notation. Note that $s + 1 = (n + m)!/(n!m!)$. For each $x \in \Omega$ consider

The local moving least square approximation of a function u is defined as (see [1])

$$P_u(x,y) = \sum_{i=1}^{N} \mathbf{p}^T \left(\frac{y-x}{R}\right) M^{-1}(x) \mathbf{p} \left(\frac{x_i - x}{R}\right) W \left(\frac{x_i - x}{R}\right) u_i. \tag{1}$$

and letting $y \mapsto x$

$$u(x) \approx u_R(x) = P_u(x,x) = \sum_{i=1}^{N} \mathbf{p}^T(0) M^{-1}(x) \mathbf{p} \left(\frac{x_i - x}{R}\right) W \left(\frac{x_i - x}{R}\right) u_i = \sum_{i=1}^{N} \varphi_i(x) u_i, \tag{2}$$

where $\varphi_i(x), 1 \leq i \leq N$, are called the shape functions, we obtain, $u_R(x)$, the (global) moving least squares approximation of the function u at the point $x \in \Omega$.

2.1 Some error estimates for MLS approximations

In [1], [2], [4], [8], under certain assumptions we can find the following result:

Theorem 1 *Let $m+1 > \frac{n}{2}$. If $u \in H^{m+1}(\Omega)$, then there exists constants C_1 and C_2 independent of R such that*

$$||u - u_R||_{L^2(\Omega)} \leq C_1 R^{m+1} |u|_{H^{m+1}(\Omega)}$$

and

$$||\nabla u - \nabla u_R||_{L^2(\Omega)} \leq C_2 R^m |u|_{H^{m+1}(\Omega)}.$$

2.2 The diffuse derivative

The approximation of the derivative of u arises from the corresponding derivative of $P_u(x,y)$. That is, in the one dimensional case

$$\frac{du(x)}{dx} \approx \lim_{y \to x} \frac{\partial P_u(x,y)}{\partial x} = \lim_{y \to x} \left[\frac{\partial \mathbf{p}^T((y-x)/R)}{\partial x} \mathbf{a}(x) + \mathbf{p}^T \left(\frac{y-x}{R}\right) \frac{d\mathbf{a}(x)}{dx} \right]. \tag{3}$$

The second term on the right-hand-side (rhs) is not trivial since derivatives of vector **a** require the solution of a linear system of equations with matrix M. The concept of diffuse derivative, proposed in [9], involves only the first term on the rhs of (3), i.e.

$$\begin{aligned} \frac{du(x)}{dx} \approx \delta u(x) := \delta u_R(x) &= \lim_{y \to x} \frac{\partial P_u(x,y)}{\partial y} \\ &= \lim_{y \to x} \frac{\partial \mathbf{p}^T((y-x)/R)}{\partial y} \mathbf{a}(x) := \sum_{i=1}^{N} \delta\varphi_i(x) u_i, \end{aligned} \tag{4}$$

A good description of this derivative can be found in (see [6]):

Now consider a function $V \in \mathcal{V}_R = span\{\varphi_1, \varphi_2, ..., \varphi_N\}$ defined as

$$V(x) = \sum_{i=1}^{N} \varphi_i(x) \overline{V}_i = \lim_{y \to x} P_{\overline{V}}(x,y) = \lim_{y \to x} \mathbf{p}^T \left(\frac{y-x}{R}\right) \mathbf{a}(x) = \mathbf{p}^T(0) \mathbf{a}(x), \tag{5}$$

where $\overline{V}_i, 1 \leq i \leq N$, are constants and $\varphi_i(x)$ are the MLS shape functions.

Following closely the ideas on [1], and the same assumptions there, we can prove:

Lemma 1 *If $|\alpha| = 1$ then there exists a constant C, independent of R, such that*

$$||D^\alpha V - \delta^\alpha V||^2_{L^2(\Omega)} \leq \frac{C}{R^2} \int_\Omega \sum_{x_k \in \Lambda(x)} |P_{\overline{V}}(x, x_k) - \overline{V}_k|^2 dx$$

where, $\Lambda(x) = \{x_j \in \Lambda \ : \ x \in supp(W(x_j - \cdot)) \cap \overline{\Omega}\} = \{x_j \in \Lambda \ : \ x \in \overline{B_R(x_j)} \cap \overline{\Omega}\}$.

3 A PENALIZED DEM

For a bounded domain Ω with C^1-boundary $\partial\Omega$ we consider problems of the form:

$$\begin{cases} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)(x) + c(x)u = f(x) & x \in \Omega \\ \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial u}{\partial x_j}\nu_i(x) = g(x) & x \in \partial\Omega \end{cases} \tag{6}$$

where $a_{ij}, c \in L^\infty(\Omega), f \in L^2(\Omega), g \in L^2(\partial\Omega), a_{ij} \in L^\infty(\partial\Omega)$ and ν is a unit normal vector to $\partial\Omega$.

Here $c(x) \geq c_1 > 0$ for all $x \in \Omega$ and $A(x) = (a_{ij}(x))_{n\times n}$ is assumed to be uniformly elliptic in Ω. Assume also that $u \in C^{m+1}(\Omega)$ is the true solution.

For $g = 0$, define the bilinear form

$$B(u,v) = \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}(x)\frac{\delta u}{\delta x_j}\frac{\delta v}{\delta x_i} + c(x)uv\right)dx := (A\delta u, \delta v) + (cu, v) \tag{7}$$

where

$$\delta u = \left[\frac{\delta u}{\delta x_1}, \frac{\delta u}{\delta x_2}, \cdots, \frac{\delta u}{\delta x_n}\right]^T$$

and A is a matrix whose ij entry is a_{ij}

This variational problem has a unique solution by the Lax-Milgram theorem.

Set $\mathcal{V}_R = span\{\varphi_1, \varphi_2, ..., \varphi_N\}$ and for $U, V \in \mathcal{V}_R$ defined as in (5), define

$$\mathcal{M}(V)(x, x_l) = P_{\overline{V}}(x, x_l) - \overline{V}_l,$$

and the penalty function

$$\mathcal{P}(\mathcal{M}(U), \mathcal{M}(V)) = \int_\Omega \left[\sum_{x_l \in \Lambda(x)} \mathcal{M}(U)(x, x_l)\mathcal{M}(V)(x, x_l)\right]dx. \tag{8}$$

So, $\|D^\alpha V - \delta^\alpha V\|_{L^2(\Omega)} \leq \frac{C}{R^2}\mathcal{P}(\mathcal{M}(V), \mathcal{M}(V))$.

Consider the numerical scheme:

$$\textbf{P-DEM:} \begin{cases} \text{Find } U \in \mathcal{V}_R \text{ so that} \\ B(U,\beta) + R^{-2\gamma}\mathcal{P}(\mathcal{M}(U), \mathcal{M}(\beta)) = (f,\beta) \quad \forall\, \beta \in \mathcal{V}_R. \end{cases}$$

We can prove (see [10]):

Theorem 1 *: Let $u \in C^{m+1}(\Omega)$ be the exact solution of (6), u_R be its MLS approximation and U be the solution given by the numerical scheme **P-DEM**, then if $\gamma = m/2 + 1$ and $e = u_R - U$, there exists a constant C independent of R such that,*

$$\|\delta e\|^2_{L^2(\Omega)} + \|e\|^2_{L^2(\Omega)} \leq CR^m, \textit{and therefore:} \quad \|u - U\|_{L^2(\Omega)} \leq CR^{m/2} \quad \|\delta u - \delta U\|_{L^2(\Omega)} \leq CR^{m/2} \tag{9}$$

4 NUMERICAL RESULTS

For the two dimensional case we consider the problem

$$\begin{cases} -\Delta u + \pi^2 u = 3\pi^2 cos(\pi x)cos(\pi y) & (x,y) \in \Omega, \\ \frac{\partial u}{\partial n} = 0 & (x,y) \in \partial\Omega, \end{cases} \tag{10}$$

where $\Omega = [0,1] \times [0,1]$ and n is the unit normal vector to the boundary of Ω. The exact solution of (10) is $u(x,y) = cos(\pi x)cos(\pi x)$.

For this example, we divide Ω into $11 \times 11, 21 \times 21$ and 31×31 uniformly distributed points, and the dilation parameter (R/h) is again kept constant for each m, so that the hypotheses from section 2 are satisfied.

Tables 1 and 2 show numerical results comparing EFG, DEM and P-DEM for different dimensions (m) of the polynomial basis used in the MLS approximation. We report errors for both the numerical approximation of the solution of (10) and the approximation of its first derivative with respect to x using diffuse derivatives, in the L^2 norm. It is important to notice that similar results can be obtained for the full and diffuse derivatives with respect to y, and therefore for the gradient and diffuse gradient of u. The convergence rates are summarized in the tables. It can be seen that the results obtained using the EFG present the lowest error and the fastest convergence rates, also as before the convergence rate of the DEM is constant and independent of the dimension of the polynomial basis used, although slightly better than in the one dimensional problem studied above. The DEM gives acceptable results for $m = 1$, and it is comparable or sometimes better than our P-DEM, however the P-DEM performs better for $m > 1$ and the convergence rate increases as the dimension of the polynomial basis increases.

Table 1: Convergence Rates for Solution Approximation in L^2

Method	$m = 1$	$m = 2$	$m = 3$
DEM	1.96	2.06	1.99
P-DEM	1.63	1.92	2.49
EFG	2.49	3.15	4.10

Table 2: Convergence Rates for Diffuse Derivative Approximation (with respect to x) in L^2

Method	$m = 1$	$m = 2$	$m = 3$
DEM	1.491	1.48	1.48
P-DEM	0.95	1.54	1.88

Remark: A proof of the enhanced L^2 convergence observed remains an open question. The fact that our penalty term is not exactly zero at the true solution and integration by parts with diffuse derivatives is not possible, seem to make the standard duality argument impossible. impossible.

REFERENCES

[1] M. ARMENTANO, *Error estimates in Sobolev spaces for moving least square approximations*, SIAM J. Numer. Anal., 39 (2001), pp. 38–51.

[2] M. ARMENTANO AND R. DURAN, *Error estimates for moving least square approximations*, Applied Numerical Mathematics, 37 (2001), pp. 397–416.

[3] T. BELYTSCHKO, Y. Y. LU, AND L. GU, *Element-free Galerkin methods*, Int. J. Numer. Methods Engrg., 37 (1994), pp. 229–256.

[4] W. HAN AND X. MENG, *Error analysis of the reproducing kernel particle method*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6157–6181.

[5] A. HUERTA, Y. VIDAL, AND P. VILLON, *Pseudo-divergence-free element free galerkin method for incompressible fluid flow*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1119–1136.

[6] D. KIM AND Y. KIM, *Point collocation methods using the fast moving least square reproducing kernel approximation*, Int. J. Numer. Meth. Engrg., 56 (2003), pp. 1445–1464.

[7] Y. KRONGAUZ AND T. BELYTSCHKO, *A Petrov-Galerkin diffuse element method (PG DEM) and its comparison to EFG*, Computational Mechanics, 19 (1997), pp. 327–333.

[8] W. K. LIU AND T. BELYTSCHKO, *Moving least-square reproducing kernel methods. Part I: Methodology and convergence*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 113–154.

[9] B. NAYROLES, G. TOUZOT, AND O. VILLON, *Generating the finite element method: diffuse approximation and diffuse elements*, Comput. Mech., 10 (5) (1992), pp. 307–318.

[10] M. OSORIO, *Error analysis of Meshfree methods with diffuse derivatives*, PhD thesis, (in preparation), University of Cincinnati, 2010.

INTERPOLATION ERROR ESTIMATES FOR B-SPLINES APPROXIMATION ON ANISOTROPIC RECTANGULAR MESHES

Ariel L. Lombardi[♭]

[♭]*Instituto de Ciencias, Universidad Nacional de General Sarmiento, aldoc7@dm.uba.ar*

Abstract: In this short note we consider the approximation by B-splines of general elliptic problems, on rectangular meshes that may contain narrow or flat elements. We show that the error estimates do not deteriorate when the aspect ratio of the elements degenerates. The same conclusion can then be applied for the NURBS-based isogeometric approach for elliptic problems on more general physical domains.

Keywords: *B-splines, finite elements, h-refinement, anisotropic meshes*
2000 AMS Subject Classification: 65N30

1 INTRODUCTION

In this short note we present some results showing that the use of rectangular meshes with highly narrow elements does not deteriorates the approximation by B-splines [2, 3, 5] of elliptic variational problems. Our interest in this kind of approximation, resides in the fact that B-splines are the bases for the NURBS-based Isogeometric Analysis [1, 4]. In fact, it follows from what was just said, that the isogeometric approximations of elliptic problems does not deteriorates if the mesh in the parametric domain contains arbitrarily flat elements.

Isogeometric analysis based on NURBS (non-uniform rational B-splines) was introduced in [4], with the aim of improving the connection between numerical simulation of physical phenomena and the Computer Aided Design (CAD) system. Indeed, one of its most important features, is to eliminate (or at least reduce) the approximation of the computational domain and the re-meshing by the use of the "exact" geometry directly on the coarsest level of the discretization. This is achieved by using B-splines or NURBS for the geometry description as well as for the representation of the unknown fields, combined with the classical isoparametric concepts of finite element method.

This work is devoted to the basic situation of B-spline approximation, that corresponds to the isogeometric approach when the parametric and physical domains coincide. We see this as a starting point to consider the general situation.

Situations in which the use of parametric meshes with narrow element is needed can appear for example when the solution to be approximated contain sharp gradients, like boundary or internal layers. This is the case of singularly perturbed reaction-convection-diffusion equations. Other situations can appear related with the map that applies the parametric domain onto the physical one: in order to obtain a quasiuniform subdivision of the physical domain, may happen that a mesh with some refined regions is necessary in the parametric domain.

In the next Section we introduce the B-spline approximation of a general elliptic problem, and we define a quasi-interpolation that helps to analyze the approximation error. Then, in Section 3, the interpolation error estimates are presented.

2 PRELIMINARIES

Let $\Omega = [0,1]^2$. Consider a variational problem

$$u \in H_0^1(\Omega): \qquad a(u,v) = F(v) \quad \forall v \in H_0^1(\Omega), \tag{1}$$

where a is a continuous and coercive bilinear form in $H_0^1(\Omega)^2$, and F is a continuous linear form in $H_0^1(\Omega)$.

Now we consider a discrete approximating spaces.

Let $m > 0$ be an integer. For $j = 1, 2$, given n_j, let the open knot vector

$$\Xi_j = \{0 = y_{j,1}, y_{j,2}, \ldots, y_{j,n_j+m} = 1\}$$

with $y_{j,1} = \ldots = y_{j,m} = 0$ and $y_{j,n_j+1} = \ldots = y_{j,n_j+m} = 1$. The allowed maximum multiplicity of the knots is m. We will denote

$$Q_{i_1 i_2} = (y_{1,i_1}, y_{1,i_1+1}) \times (y_{2,i_2}, y_{2,i_2+1}),$$

for $i_j = 1, \ldots, n_j+m-1$. We note that may occur $Q_{i_1 i_2} = \emptyset$. We also use the notation $h_{j,i_j} = y_{j,i_j} - y_{j,i_j-1}$.

Let $\{N^j_{i_j,m}\}, i_j = 1, \ldots, n_j$ be the B-spline basis of order m (and degree $m-1$) associated with the knots Ξ_j. These basis can be defined by an iterative procedure presented for example in [4], Section 2.2. Then define the bidimensional basis $\{N_{i_1 i_2}\}$ by

$$N_{i_1 i_2} = N^1_{i_1,m} N^2_{i_2,m}.$$

The discrete space V_h is then defined as the space generated by

$$\{N_{i_1 i_2} : i_1 = 2, \ldots, n_1 + m - 1, i_2 = 2, \ldots, n_2 + m - 1\}.$$

The index h, representing the global mesh-size, is $h = \max_{i_1,i_2}\{\operatorname{diam} Q_{i_1 i_2}\}$. We remark that

$$V_h \subset H^1_0(\Omega),$$

that is a consequence of the fact that the basis function defining V_h are contained in $H^1_0(\Omega)$ [1]. Then we define the approximating problem as:

$$u_h \in V_h : \qquad a(u_h, v) = F(v) \quad \forall v \in V_h, \tag{2}$$

Existence of solutions u and u_h follows from Lax-Milgram Theorem, while from Céa Lemma we have the error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v \in V_h} \|u - v\|_{H^1(\Omega)}. \tag{3}$$

So, we can estimate the error by taking above $v = \Pi_h u$, where $\Pi_h : H^1_0(\Omega) \to V_h$ is an appropriated (quasi) interpolation operator. In order to define this operator we introduce the following. For $j = 1, 2$ and $1 \leq i_j \leq n_j$, define the functionals

$$\lambda^j_{i_j} f = \int_{y_{j,i_j}}^{y_{j,i_j+m}} f(t) D^m \psi_{j,i_j}(t)\, dt.$$

where

$$\psi_{j,i_j}(t) = G_{j,i_j}(t) \varphi_{j,i_j}(t)$$

and

$$G_{j,i_j}(t) = g\left(\frac{2t - y_{i_j} - y_{i_j+m}}{y_{i_j+m} - y_{i_j}}\right)$$

with g the transition function defined in [5], Theorem 4.37. Finally

$$\varphi_{j,i_j}(t) = \frac{1}{(m-1)!}(t - y_{j,i_j+1}) \cdots (t - y_{j,i_j+m-1}).$$

Then, the functional $\lambda_{i_1 i_2}$ for $i_1 = 1, \ldots, n_1+m$, $i_2 = 1, \ldots, n_2+m$, is the tensor product of the functionals $\lambda^1_{i_1}$ and $\lambda^2_{i_2}$:

$$\lambda_{i_1 i_2} = \lambda^1_{i_1} \cdot \lambda^2_{i_2}$$

Following [1], Section 3.4, for a function $v \in H^1_0(\Omega)$ we introduce the quasi-interpolation $\Pi_h v$ defined by

$$\Pi_h v = \sum_{i_1=2}^{n_1+m-1} \sum_{i_2=2}^{n_2+m-1} (\lambda_{i_1 i_2} v) N_{i_1 i_2}$$

We note that $\Pi_h v|_{\partial\Omega} = 0$.

In what follows we will need the notation

$$\widetilde{Q}_{s_1 s_2} = (y_{1,s_1-m}, y_{1,s_1+m+1}) \times (y_{2,s_2-m}, y_{2,s_2+m+1}).$$

GEOMETRICAL ASSUMPTION ON Ξ_j

We assume that the partitions defined by Ξ_j, $j = 1, 2$, are locally quasi-uniform. That is, if $\{\xi_{j,s}\}_{s=1}^{r_j}$ are the knots without repetitions, then we assume that there exists a constant $\mu > 0$ such that

$$\frac{1}{\mu} \leq \frac{\xi_{j,s} - \xi_{j,s-1}}{\xi_{j,s+1} - \xi_{j,s}} \leq \mu, \qquad 1 < s < r_j, j = 1, 2.$$

Observe that this assumption allows for arbitrarily narrow elements, since that no restriction is imposed on the relationship between the lengths of elements along different directions.

It follows from this assumption that if $h_{j,i_j} \neq 0$, $j = 1, 2$, then the length of $\widetilde{Q}_{i_1 i_2}$ along the j-direction is comparable with h_{j,i_j}, with constants in this equivalence depending only on μ and m.

3 ERROR ESTIMATES

We briefly state the some results that we have obtained.

Fix s_1 and s_2 such that $Q_{s_1 s_2} \neq \emptyset$. The (anisotropic) estimate for the L^2-norm of the error follows as in [1], Sections 3.1 and 3.4.

Proposition 1 *If $u \in H^k(\widetilde{Q}_{s_1 s_2})$, $2 \leq k \leq m$, such that $u = 0$ on $\partial\widetilde{Q}_{s_1 s_2} \cap \partial\Omega$. Then*

$$\|u - \Pi_h u\|_{0,Q_{s_1 s_2}} \leq C(h_{1,s_1}^k \|\partial_{x_1}^k u\|_{0,\widetilde{Q}_{s_1 s_2}} + h_{2,s_2}^k \|\partial_{x_2}^k u\|_{0,\widetilde{Q}_{s_1 s_2}})$$

with the constant C depending only on μ and m.

The anisotropic estimate for the H^1-seminorm of the error $u - \Pi_h u$ needs some additional effort. Our main result is the following.

Proposition 2 *Let $u \in H_0^1(\Omega)$. Let $p_{s_1 s_2} \in \mathcal{Q}_{m-1}(\widetilde{Q}_{s_1 s_2})$ be a polynomial vanishing on $\partial\Omega \cap \partial\widetilde{Q}_{s_1 s_2}$. Then*

$$\begin{aligned} \|\partial_{x_1}(u - \Pi_h u)\|_{0,Q_{s_1 s_2}} &\leq C \|\partial_{x_1}(u - p_{s_1 s_2})\|_{0,\widetilde{Q}_{s_1 s_2}} \\ \|\partial_{x_2}(u - \Pi_h u)\|_{0,Q_{s_1 s_2}} &\leq C \|\partial_{x_2}(u - p_{s_1 s_2})\|_{0,\widetilde{Q}_{s_1 s_2}} \end{aligned}$$

The previous Proposition allows for obtaining, for B-spline approximation, practically the same anisotropic estimates that were proved for standard finite elements, by simply choosing the adequate interpolation (or quasi-interpolation) polynomial $p_{s_1 s_2}$. For example, if it is taken as the Lagrange interpolation on $\widetilde{Q}_{s_1 s_2}$, we have the next result.

Proposition 3 *Let s_1, s_2 be fixed. Suppose that for $2 \leq k \leq m$ we have $u \in H^k(\widetilde{Q}_{s_1 s_2})$, such that $u = 0$ on $\partial\widetilde{Q}_{s_1 s_2} \cap \partial\Omega$. Then*

$$\begin{aligned} \|\partial_{x_1}(u - \Pi_h u)\|_{0,Q_{s_1 s_2}} &\leq C \left(h_{1,s_1}^{k-1} \|\partial_{x_1}^k u\|_{0,\widetilde{Q}_{s_1 s_2}} + h_{2,s_2}^{k-1} \|\partial_{x_2}^{k-1} \partial_{x_1} u\|_{0,\widetilde{Q}_{s_1 s_2}} \right) \\ \|\partial_{x_2}(u - \Pi_h u)\|_{0,Q_{s_1 s_2}} &\leq C \left(h_{1,s_1}^{k-1} \|\partial_{x_1}^{k-1} \partial_{x_2} u\|_{0,\widetilde{Q}_{s_1 s_2}} + h_{2,s_2}^{k-1} \|\partial_{x_2}^k u\|_{0,\widetilde{Q}_{s_1 s_2}} \right) \end{aligned}$$

where the constant C depends on m, k and the constant μ in the geometrical assumption.

This result together with (3) gives optimal error estimates for the B-spline approximation (2) of problem(1).

REFERENCES

[1] Y. Bazilevs, L. Beirão da Veiga, J. A. Cottrell, T. J. R. Hughes, G. Sangalli, *Isogeometric Analysis: approximation, stability and error estimates for h-refined meshes*, Math. Models Meth. Appl. Sci., 16(7), 1031–1090, 2006.

[2] C. de Boor, A practical guide to splines, "Applied Mathematical Sciences", Vol. 27, Springer, New York, 2001.

[3] C. de Boor, G. J. Fix, *Spline Approximation by Quasiinterpolants*, J. Approx. Theory, 8, 19–45, 1973.

[4] T. J. R. Hughes, J. A. Cottrell, Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry, and mesh refinement*, Comp. Meth. Appl. Mech. Engrg., 194, 4135–4195, 2005.

[5] L.L. Schumaker, Spline Functions: Basic Theory, Cambridge University Press, Cambridge, 2007.

ERROR ESTIMATES FOR NONLINEAR ADAPTIVE PARABOLIC FINITE ELEMENT METHOD IN CATALYTIC REACTOR MODELING

Marta Beatriz Bergallo and Carlos-Enrique Neuman Meira

Mathematics Department (FIQ), Universidad Nacional del Litoral, Santiago del Estero 2829, 3000 Santa Fe, Argentina, bergallo,ceneuman@fiq.unl.edu.ar

We dedicate this work to Professor Carlos E. Kenig

Abstract: Several numerical properties associated to the modeling of chemical reactors as sets of parabolic problems with nonlinear boundary conditions of Robin type are studied. The main issue in the modelisation and solution procedure is the nonlinearity and presence of boundary singularities that are essential to these systems and their relationship with the error estimations for adaptivity. The focus of this work is, in consequence, to assess different types of error estimators associated to linearization based in a Picard-like scheme. The models are stated and solved in the adaptive finite element software environment ALBERTA. The numerical examples are associated to problems drawn from Chemical Engineering (Simplified Catalytic Reactors) with the aim of providing approximate solutions and *a posteriori* error estimators oriented to the adaptation of meshes and timesteps. The schematic and simplified modeling and simulation of the reactor is an issue in this article, due to the characteristics of the original problem.

Keywords: *Parabolic problems, Border singularities,* A posteriori *error estimation, Robin boundary conditions, Catalytic Reactor Modeling, ALBERTA*
2000 AMS Subject Classification: 65N30 - 80A20

1 INTRODUCTION

In this work several properties associated to the modeling of catalytic reactors as sets of parabolic problems with mixed boundary conditions are studied. The focus of this work is to study parabolic problems where the domains present border nonlinearities and singularities in order to assess different types of error estimators. Catalytic reactors are modeled as sets of nonlinear parabolic problems. The examples are, in consequence, associated to problems drawn from Chemical Engineering (Simplified Catalytic Reactors) with the aim of providing reasonable solutions and acceptable *a posteriori* error estimates.

1.1 MONOLITH CONVERTERS

Monolith converters (i.e. with impermeable walls) have been thoroughly used in the last 35 years as afterburners of combustion engine exhausts and, e.g, for catalytic reactions in Petroleum Chemistry (see, for example the review of Cybulski and Moulijn [3], and the references therein). The catalytic oxidation of hydrocarbons, nitrogen oxides, carbon monoxide and other contaminants of exhaust gases of automobiles is performed on the walls of thousands of (near parallel) passages of the monolith converter. The main parts of this article related to monolith converters, begin with the modeling of the problem introducing several simplifications. The next step is to develop a finite element model of the variational (weak) equations. We set and test several error indicators in order to estimate the errors of approximation [1].

2 BASIC MODEL

We model a single passage of the monolith. Being considered the hydrocarbon oxidation, the carbon monoxide combustion, and the nitrogen oxides elimination. The advection-diffusion molar balances for the chemical species i, $i = 1{:}s$, in the r reactions are omitted for reasons of space in this resumed work since we concentrate on the heat balance. The heat balance reads[2, 4, 5]

$$\frac{1}{\rho C_p}\nabla(K\nabla T) - V\nabla T - \sum_{i=1}^{s}\frac{M_i\Delta H_i}{\rho C_p}\sum_{j=1}^{r}\nu_{ij}R_{jH} = \frac{\partial T}{\partial t} \tag{1}$$

where the thermal diffusivity $K = K(T)$ is temperature dependent, M_i are the molecular weight, and ΔH_i the enthalpy of species i. The specific heat of the gas is C_p. The cylindrical symmetry allows us to set boundary conditions in the four sides of a rectangular half section of the tube. In the inner side and the outlet side we pose natural conditions. On the wall side, where the catalytic reaction actually performs, is necessary to consider the heterogeneous reaction rates (faster than the homogeneous ones at operation temperatures) and the heat production (chemical reactions) and heat dissipation on the monolith. In the inlet side the conditions are essential and it is necessary to know the molar fractions and temperature of the gas.

3 EQUATION

The simplified heat equation we use in this model is

$$u_t - \operatorname{div}(A\nabla u) + b\nabla u + (cw/u)\exp(-H_h/u) = f, \qquad b = v(1-y^2,0) \tag{2}$$

where $\Omega = [0,1]\times[0,1]$; $u = u_0,\ t = 0$; $u = g,\ x = 0\ (0 \le y \le 1)$; $\nu A\nabla u = h_n,\ \Gamma_N:\ x = 1\ (0 \le y \le 1),\ y = 0\ (0 \le x \le 1)$; $\nu A\nabla u + \left((qw/u^2)\exp(-H_w/u) - \sigma u^3\right)u = h_r,\ \Gamma_R:\ y = 1\ (0 \le x \le 1)$ and w is the solution of the other equations (mass balances). From now on we use the notation $E_a(b) = \exp(-H_a/b)/b^2$.

4 WEAK LINEAR FORM

The weak (Picard-)linearized form is

$$\int_\Omega \left(\frac{u_h^n - u_h^{n-1}}{\Delta t}\right)\varphi + \int_\Omega A\nabla u_h^n\nabla\varphi + \int_\Omega b\nabla u_h^n\varphi + \int_\Omega cw_h^{n-1}E_h(u_h^{n-1})u_h^n\varphi$$
$$+ \int_{\Gamma_R}\left(qw_h^{n-1}E_w(u_h^{n-1}) - \sigma(u_h^{n-1})^3\right)u_h^n\varphi = \int_\Omega f\varphi + \int_{\Gamma_N} h_n\varphi + \int_{\Gamma_R} h_r\varphi \tag{3}$$

We solve this problem by AFEM with the information of the other equations of the system included as "w".

5 LOCAL ERROR ESTIMATES

The error estimates are proposed and tested in this work. They are classified as linear, nonlinear, total, and global. *A posteriori* error estimates for nonlinear parabolic problems consist of five different types of terms, *i.e.* terms estimating the (1) initial error; (2) error from discretization in space; (3) error from mesh coarsening between time steps; (4) error from linearization of the weak form; and (5) error from discretization in time.

5.1 LINEAR

We introduce the estimators (we omit several error estimates):

$$\eta_S^2 = C_0^2 h_S^4 \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t} - \operatorname{div}(A\nabla u_h^n) + b\nabla u_h^n + cw_h^{n-1}E_h(u_h^{n-1})u_h^n - f \right\|^2 \tag{4}$$

$$\eta_R^2 = C_5^2 h_R^2 \left\| \nu A\nabla u_h^n + \left(qw_h^{n-1}E_w(u_h^{n-1}) - \sigma(u_h^{n-1})^3\right)u_h^n - h_r \right\|^2 \tag{5}$$

So the linear part of the estimator is $\eta_{h,lin}^2 = \eta_S^2 + \eta_{Sc}^2 + \eta_{SJ}^2 + \eta_N^2 + \eta_R^2$.

5.2 NONLINEAR

We propose, for (2), the nonlinear estimates:

$$\eta_{S,NL}^2 = C_4^2 h_S^4 \left\| E_h(u_h^n) - E_h(u_h^{n-1}) \right\|^2 \tag{6}$$

$$\eta_{R,NL}^2 = C_6^2 h_R^2 \left(\left\| w_h^{n-1}\left(E_w(u_h^n) - E_w(u_h^{n-1})\right) \right\|^2 + \sigma^2 \left\| (u_h^n)^3 - (u_h^{n-1})^3 \right\|^2 \right) \tag{7}$$

so the nonlinear part of the estimate is $\eta_{h,NL}^2 = \eta_{S,NL}^2 + \eta_{R,NL}^2$.

5.3 TOTAL

The total estimate is, in consequence, $\eta_h^2 = \eta_{h,lin}^2 + \eta_{h,NL}^2$

6 GLOBAL ERROR ESTIMATES

The global error estimates are needed in the adaptive algorithms decisions, they are $\eta_0 = C_8 \left\| u_0 - u_h^0 \right\|$, $\eta_\tau = C_3 \left\| u_h^n - u_h^{n-1} \right\|$, $\eta = \sqrt{\sum \eta_h^2}$

7 ADAPTIVE ALGORITHMS

The basic iteration of an adaptive finite element code for a stationary problem is (1) assemble and solve the discrete system; (2) calculate the error estimate; and (3) adapt the mesh, when needed. For time dependent problems, such an iteration is used in each time step, and the step size of a time discretization may be subject to adaptivity, too.

We summarize the time and space adaptive algorithms

Algorithm `adapt`

```
adapt_timestep
while sp_est > tol or sp_est << tol
  mark elem. for refinem. or coarsen.
  if elements are marked
        adapt mesh
        assemble(linearized problem)
        solve => u(n+1) [using u(n)]
        compute estimates
  end if
  adapt_timestep
end while
```

Subalgorithm `adapt_timestep`

```
while t_est > tol or t_est << tol
    modify timestep
    increment time
    assemble(linearized problem)
    solve -> u(n+1)  [using u(n)]
    compute estimates
end
```

8 REFINEMENT STRATEGY

The number of mesh elements in the mesh S_k is N_k. We suppose equidistributed error indicators, *i.e.* $\eta_S \approx \eta_{S'}, \quad \forall S, S' \in S_k$, then

$$\eta = (\sum_{S \in S_k} \eta_S^p)^{1/p} \approx N_k^{1/p} \eta_S \simeq \text{tol} \tag{8}$$

in consequence we mark for refinement and coarsening all elements where

$$\eta_S > \theta \quad \text{tol}/N_k^{1/p} \qquad \text{and} \qquad \eta_S^p + \eta_{c,S}^p < \theta_c^p \quad \text{tol}^p/N_k \tag{9}$$

respectively, where $\theta = 0.9$ is a robustness parameter, and $\theta_c = 0.165$ is a parameter for fine numerical adjustment of coarsening.

9 ERROR ESTIMATION AND ERROR

With the objective of analyzing the numerical behavior of the proposed estimates of discretization errors we consider a similar pseudoproblem with exact polinomial solution in the region of space corresponding to the actual solution. In Fig. 1 we observe that the actual errors and the estimates are similar. Also we observe the convergence of the whole process and the numerical stability of the solution of the parabolic problem. We use coarsening and refinement in our adaptive algorithm in order to achieve the quasi uniformity of errors of approximation. In Fig. 2 we plot the errors, with the aim to show the quality of our estimates. In this article we compare with results obtained in our previous work ([6, 1]). We also show a sketch of the approximate solution (Fig. **??**).

10 CONVERGENCE OF AFEM

In the studied examples (with known solution of the approximate pseudoproblem), we have observed the following convergence result for (2): The AFEM with our estimators achieves

$$\lim_{k \to \infty} \| u_{h_k}^{(k)} - u(\cdot, t_k) \| = 0 \tag{10}$$

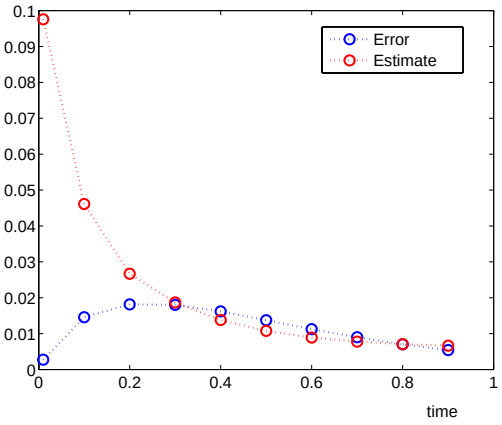


Figure 1: Errors and estimators

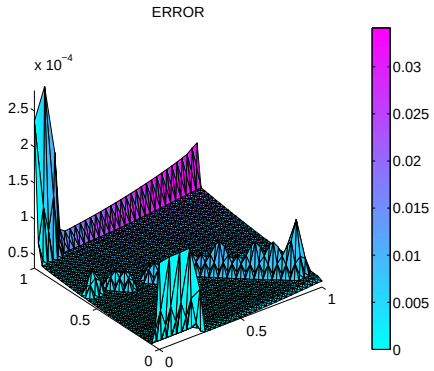


Figure 2: Errors.

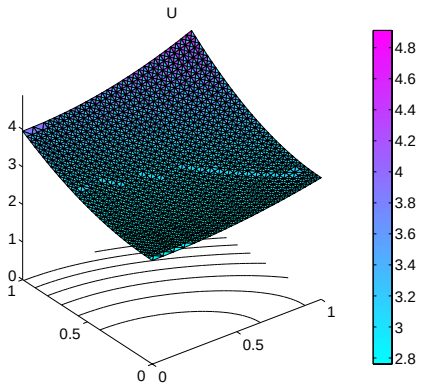


Figure 3: Approximate solution.

where the limit is understood in a time-dynamic way. As time passes, the meshes are refined in the areas of greatest error in order to achieve equidistributed estimated error.

11 Conclusions

In this work adaptive FEM with *a posteriori* error detection in boundary singular systems was addressed. And modeling of Monolith Catalytic Reactors in a simplified form was performed. Convergent non-linear (Picard linearized) systems associated to the converters were devised and good responses of adaptivity, error estimation, convergence and stability were observed

References

[1] Bergallo, M.B., Neuman, C.E. and Sonzogni, V.E.: "Errores *A Posteriori* y Mejoramiento de la Valuación de Opciones Financieras Dependientes del Camino", *Mecánica Computacional*, **25**, 1051–1069, (2006).

[2] Bird, R.B., Stewart, W.E., and Lightfoot, E.N.: *Fenómenos de Transporte.*, Reverté, Barcelona, (1964).

[3] Cybulski, A., and Moulijn, J.A.: "Monoliths in heterogeneous catalysis", *Catalysis Reviews, Science and Engineering*, **36**(2), 179–270, (1994).

[4] Froment, G.F., and Bischoff, K.B.: *Chemical Reactor Analysis and Design*, J. Wiley & Sons, New York, (1990).

[5] Hayes, R.E., Kolaczkowski, S.T., and Thomas, W.J.: "Finite–Element Model for a Catalytic Monolith Reactor", *Computers & Chemical Engineering*, **16**(7), 645–657, (1992).

[6] Neuman, C.E.: "Modeling Catalytic Reactors as Elliptic Problems with Essential Border Singularities", *Mecánica Computacional*, **23**, 2845–2862, (2004).

SUPERCONVERGENCE FOR FINITE ELEMENT APPROXIMATIONS OF A SINGULARLY PERTURBED PROBLEM USING GRADED MESHES

Ricardo G. Durán, Ariel L. Lombardi and Mariana I. Prieto

Departamento de Matemática, FCEyN, Universidad de Buenos Aires, Argentina.

Abstract: We consider the approximation of a model singularly perturbed reaction-diffusion equation by standard bilinear finite elements using graded meshes. We prove superconvergence in the energy norm which is valid almost uniformly in the singular perturbation parameter ε, i. e., the constant depends only on the logarithm of ε.

Keywords: *finite elements, reaction-diffusion, superconvergence, graded meshes.*
2000 AMS Subject Classification: 65N30

1 INTRODUCTION

As is well known, the approximation of problems with boundary layers require some special treatment in order to avoid oscillations. A lot of work have been done using special a priori adapted meshes both for finite difference and finite element methods. The most well known meshes are the Shishkin type ones (see for example the book [6] and its references). A different approach is to use graded meshes as those considered in [3, 4]. The advantage of this kind of meshes over the Shishkin ones for reaction-diffusion problems is that almost optimal order estimates in the energy norm can be obtained with meshes independent of the singular perturbation parameter.

In this work we obtain superconvergence error estimates for standard bilinear finite element approximations of a singularly perturbed reaction diffusion problem using graded meshes. Namely, we prove that the energy norm of the difference between the numerical solution u_h and the Lagrange interpolation u_I of the exact solution is of higher order than the energy norm of the error $u - u_h$.

As a consequence of our results we obtain almost optimal order error estimates in the L^2-norm.

We consider the model problem,

$$\begin{aligned} -\varepsilon^2 \Delta u + u &= f \qquad \text{in } (0,2)^2 \\ u &= 0 \qquad \text{on } \partial(0,2)^2 \end{aligned} \tag{1}$$

where ε is a small positive parameter. We assume that $f \in \mathcal{C}^2([0,2]^2)$ and satisfies the following compatibility conditions,

$$f(0,0) = f(2,0) = f(0,2) = f(2,2) = 0$$

which ensure that the solution u of problem (1) belong to $\mathcal{C}^4((0,2)^2) \cap \mathcal{C}^2([0,2]^2)$. Such compatibility conditions are sufficient for the a priori estimates of the solution that we will use in our analysis.

For a domain D we use the standard notation for Sobolev spaces, norms and seminorms, namely,

$$\|u\|_{m,D} := \left\{ \sum_{\alpha \le m} \|\mathcal{D}^\alpha u\|^2_{L^2(\Omega)} \right\}^{1/2}, \qquad |u|_{m,D} := \left\{ \sum_{\alpha = m} \|\mathcal{D}^\alpha u\|^2_{L^2(\Omega)} \right\}^{1/2}.$$

In particular $\|u\|_{0,D}$ denotes the L^2-norm of u. When $D = [0,1]^2$, and no confusion can arise, we will write $\|u\|_0$ instead of $\|u\|_{0,D}$.

$\mathcal{P}_k$ and $\mathcal{Q}_k$ denote the spaces of polynomials of total degree less than or equal to k and of degree less than or equal to k in each variable respectively.

The standard weak formulation of Problem (1) is given by

$$\mathcal{B}(u,v) = \int_{(0,2)^2} f\, v\, dx \qquad \forall v \in H^1_0((0,2)^2), \quad \text{where} \quad \mathcal{B}(u,v) = \int_{(0,2)^2} (\varepsilon \nabla u \cdot \nabla v + uv)\, dx. \tag{2}$$

We denote $\|.\|_\varepsilon$ the norm associated with the bilinear form $\mathcal{B}$, i.e., $\|v\|_\varepsilon^2 := \mathcal{B}(v,v)$.

In [4] an analysis for the approximation of Problem (1) by standard bilinear finite elements, using appropriate graded meshes, was developed. Almost optimal order of convergence independent of ε was proved in that paper. Here we prove superconvergence for the same approximation considered in [4], i.e., that the difference between the finite element solution and the Lagrange interpolation of the exact solution, in the ε-weighted H^1-norm, is of higher order than the error itself. The constant in our estimate depends only weakly on the singular perturbation parameter.

By symmetry it is enough to prove the estimates in $\Omega = (0,1)^2$.

Our results are based on some a priori weighted estimates which are given in the following lemma. These estimates can be proved by using known point-wise estimates (see for example [5]).

Lemma 1 *There exists a constant C such that*

$$\varepsilon^\alpha \left\| y^\beta \frac{\partial^2 u}{\partial y^2} \right\|_{0,\Omega}, \varepsilon^\alpha \left\| x^\beta \frac{\partial^2 u}{\partial x^2} \right\|_{0,\Omega} \le C \quad \textit{for} \quad \alpha+\beta \ge 3/2,\ \alpha \ge 0, \beta > -1/2.$$

$$\varepsilon^\alpha \left\| y^\beta \frac{\partial^2 u}{\partial x \partial y} \right\|_{0,\Omega}, \varepsilon^\alpha \left\| x^\beta \frac{\partial^2 u}{\partial x \partial y} \right\|_{0,\Omega} \le C \quad \textit{for} \quad \alpha+\beta \ge 1,\ \alpha \ge 1/2, \beta > -1/2.$$

$$\varepsilon^\alpha \left\| y^\beta \frac{\partial^3 u}{\partial y^3} \right\|_{0,\Omega}, \varepsilon^\alpha \left\| x^\beta \frac{\partial^3 u}{\partial x^3} \right\|_{0,\Omega} \le C \quad \textit{for} \quad \alpha+\beta \ge 5/2,\ \alpha \ge 0, \beta > -1/2.$$

$$\varepsilon^\alpha \left\| x^\beta y^\gamma \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{0,\Omega} \le C \quad \textit{for} \quad \alpha+\beta \ge 3/2, \alpha+\gamma \ge 1/2, \alpha+\beta+\gamma \ge 2, \alpha \ge 0, \beta \ge -1/2, \gamma > -1/2.$$

$$\varepsilon^\alpha \left\| x^\beta y^\gamma \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{0,\Omega} \le C \quad \textit{for} \quad \alpha+\beta \ge 1/2, \alpha+\gamma \ge 3/2, \alpha+\beta+\gamma \ge 2, \alpha \ge 0, \beta \ge -1/2, \gamma > -1/2.$$

Another key point in our analysis are the following weighted trace theorems.

Lemma 2 *Let $w \in H^1(R)$ and l be one of the horizontal edges of a reference rectangle R. Then, for $0 \le \alpha < 1/2$,*

$$\|w\|_{L^1(l)} \le C \left\{ \|w\|_{L^2(R)} + \frac{1}{1-2\alpha} \left\| x^\alpha \frac{\partial w}{\partial y} \right\|_{L^2(R)} \right\}$$

$$\|w\|_{L^1(l)} \le C \left\{ \|w\|_{L^2(R)} + \frac{1}{1-2\alpha} \left\| y^\alpha \frac{\partial w}{\partial y} \right\|_{L^2(R)} \right\}$$

Let us recall the definition of the graded meshes introduced in [4]. Given a parameter h, $0 < h < 1$, and $\gamma \in (0,1)$ we consider the partition $\{\xi_i\}_{i=0}^M$ of the interval $[0,1]$ given by $\xi_0 = 0, \xi_1 = h^{\frac{1}{1-\gamma}}, \xi_{i+1} = \xi_i + h\xi_i^\gamma$, for $1 \le i \le M-2$ where M is such that $\xi_{M-1} < 1$ and $\xi_{M-1} + h\xi_{M-1}^\gamma \ge 1$, and $\xi_M = 1$. If $1 - \xi_{M-1} < \xi_{M-1} - \xi_{M-2}$ we modify the definition of ξ_{M-1} taking $\xi_{M-1} = (1+\xi_{M-2})/2$.

We define $R_{ij} = [\xi_{i-1}, \xi_i] \times [\xi_{j-1}, \xi_j]$, and the graded mesh $\mathcal{T}_h = \{R_{ij}\}_{i,j=1}^M$ in Ω. Also we define $h_i = \xi_i - \xi_{i-1}$.

The finite element approximation $u_h \in V_h$ is given by

$$\mathcal{B}(u_h, v) = \int_\Omega f v \, dx \quad \forall v \in V_h \quad \text{with} \quad V_h = \left\{ v \in \mathcal{C}(\Omega) : v\,|_{R_{ij}} \in \mathcal{Q}_1(R_{ij}), 1 \le i,j \le M \right\}$$

2 ERROR ESTIMATES

Given a continuous function u, we introduce $u_I \in V_h$ its Lagrange interpolation. We have dropped the dependence on h in the notation u_I to simplify notation.

We have the following interpolation error estimates.

Lemma 3 *For $0 \leq \alpha < 1/2$ we have*

$$\|u - u_I\|^2_{L^2(R_{ij})} \leq \frac{h_i^{2-2\alpha}}{1-2\alpha} \left\{ \left\| (x - x_{i-1})^\alpha \frac{\partial u}{\partial x} \right\|^2_{L^2(R_{ij})} + \left\| (x - x_{i-1})^\alpha \frac{\partial u_I}{\partial x} \right\|_{L^2(R_{ij})} \right\}$$

$$\|u - u_I\|^2_{L^2(R_{ij})} \leq \frac{h_j^{2-2\alpha}}{1-2\alpha} \left\{ \left\| (y - y_{j-1})^\alpha \frac{\partial u}{\partial y} \right\|^2_{L^2(R_{ij})} + \left\| (y - y_{j-1})^\alpha \frac{\partial u_I}{\partial y} \right\|_{L^2(R_{ij})} \right\}$$

$$\left\| \frac{\partial}{\partial x}(u - u_I) \right\|^2_{L^2(R_{ij})} \leq \frac{C}{(1-2\alpha)^2} \left\{ h_i^{2-2\alpha} \left\| (x - x_{i-1})^\alpha \frac{\partial^2 u}{\partial x^2} \right\|^2_{L^2(R_{ij})} + h_i^{-2\alpha} h_j^2 \left\| (x - x_{i-1})^\alpha \frac{\partial^2 u}{\partial x \partial y} \right\|^2_{L^2(R_{ij})} \right\}$$

$$\left\| \frac{\partial}{\partial x}(u - u_I) \right\|^2_{L^2(R_{ij})} \leq \frac{C}{(1-2\alpha)^2} \left\{ h_i^2 h_j^{-2\alpha} \left\| (y - y_{j-1})^\alpha \frac{\partial^2 u}{\partial x^2} \right\|^2_{L^2(R_{ij})} + h_j^{2-2\alpha} \left\| (y - y_{j-1})^\alpha \frac{\partial^2 u}{\partial x \partial y} \right\|^2_{L^2(R_{ij})} \right\}$$

Finally, another ingredient of our proofs is the following polynomial approximation result.

Lemma 4 *Given $u \in H^3(R_{ij})$, there exists $p \in \mathcal{P}_2(R_{ij})$ such that*

$$\left\| (x - x_{i-1})^\alpha \frac{\partial^2 (u-p)}{\partial x^2} \right\|_{0,R_{ij}} \leq C \left\{ \left\| (x - x_{i-1})^{\alpha+1} \frac{\partial^3 u}{\partial x^3} \right\|_{0,R_{ij}} + h_i^{-2} h_j^2 \left\| (x - x_{i-1})^{\alpha+1} \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{0,R_{ij}} \right\},$$

$$\left\| (x - x_{i-1})^\alpha \frac{\partial^2 (u-p)}{\partial x \partial y} \right\|_{0,R_{ij}} \leq C \left\{ \left\| (x - x_{i-1})^{\alpha+1} \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{0,R_{ij}} + h_i^{-2} h_j^2 \left\| (x - x_{i-1})^{\alpha+1} \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{0,R_{ij}} \right\},$$

Indeed, this theorem can be proved taking p as an averaged Taylor polynomial of u (see for example [1]) and using the weighted Poincaré inequality on the reference square $R = [0,1]^2$

$$\left\| x^\alpha (f - \bar{f}) \right\|_{L^2(R)} \leq C \left\| x^{\alpha+1} \nabla f \right\|_{L^2(R)}.$$

This inequality can be proved using the arguments given in [2].

Combining all the lemas we can prove our main results which are stated in the following two theorems.

Theorem 1 *We have the following interpolation error estimates,*

$$\|u - u_I\|_{L^2(\Omega)} \leq C \log(1/\varepsilon) h^2 \quad \textit{and} \quad \varepsilon \left\| \frac{\partial}{\partial x}(u - u_I) \right\|_{L^2(\Omega)} \leq C \log(1/\varepsilon) h$$

Theorem 2 *The following superconvergence result holds,*

$$\|u_h - u_I\|_\varepsilon \leq C h^2 \log(1/\varepsilon)$$

An immediate consequence of these two theorems is the optimal order convergence in the L^2-norm.

Finally, it is important to see the dependence of the error in terms of the number of nodes N. It can be seen (see [3]) that

$$h \leq \frac{C}{1-\gamma} \frac{\log N}{\sqrt{N}}$$

and therefore,

$$\|u_h - u_I\|_\varepsilon \leq C \log(1/\varepsilon) \frac{(\log N)^2}{N}$$

where the constant depends on γ.

REFERENCES

[1] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics 15, Springer, Berlin, 1994.

[2] I. DRELICHMAN AND R. G.DURÁN, *Improved Poincar inequalities with weights*, J. Math. Anal. Appl., 347(2008), pp. 286293.

[3] R. G. DURÁN AND A. L. LOMBARDI, *Error estimates on anisotropic Q_1 elements for functions in weighted Sobolev spaces*, Math. Comp., 74(2005), pp.1679-1706.

[4] R. G. DURÁN AND A. L. LOMBARDI, *Finite element approximation of convection-diffusion problems using graded meshes,* Appl. Numer. Math., 56(2006), pp.1314-1325.

[5] J. LI AND M.WHEELER, *Uniform convergence and superconvergence of mixed finite element methods on anisotropically refined grids*, SIAM J. Numer.Anal., 38, vol 3 (2000) pp.770-798.

[6] H. G. ROOS, M. STYNES, AND L. TOBISKA, *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems*. Springer Series in Computational Mathematics, 24(2008) Springer-Verlag, Berlin.

LONG-TIME INTEGRATION OF STOCHASTIC DIFFERENTIAL EQUATIONS BY EXPONENTIAL LL-BASED METHODS

H. de la Cruz[♭] and J. P. Zubelli[♭]

[♭]*IMPA. Estrada Dona Castorina 110, Rio de Janeiro, Brazil., hugo@impa.br, zubelli@impa.br*

Abstract: We propose and analyze A-stable numerical methods for the long-time integration of stochastic differential equations with additive noise. These methods are able to reproduce key features related to the asymptotic behaviour of a linear stochastic oscillator: they mimic the linear growth of the second moment of the solution, an infinitely-oscillation property and the symplectic structure of this Hamiltonian system. Computer experiments illustrate the theoretical findings and the advantages over long periods of time of the proposed method in comparison with some conventional integrators.

Keywords: *Stochastic differential equations, additive noise, numerical methods, long-time integration.*
2000 AMS Subject Classification: 65C30 - 60H35

1 INTRODUCTION

In the last years the stochastic modeling of physical systems has gained importance in different fields of science and engineering. In particular stochastic differential equations (SDEs) with additive noise arise as natural mathematical models for describing random processes in a variety of application areas. For example, models for celestial mechanics, the blood clotting systems, electrical activity of neural masses, electrical circuit engineering, finance and noisy oscillators in a diversity of physical systems. Since in general analytic solutions of these equations are not known or are computationally unfeasible to simulate, efficient numerical schemes capable of preserving as much as possible the key features of the original equation are then required.

Numerical integrators commonly used to solve SDEs are traditionally designed over a relative short time interval (cf. [4], [6], [8]). Consequently these integrators may perform poorly for long-time computations and not explain the complete range of behaviour that can be observed in the SDE as the integration time goes to infinity.

In order to investigate the ability of numerical methods to mimic the asymptotic behaviour of SDEs under discretization, one useful idea has been to study the numerical method on a simple test problem. In this spirit two main type of equations have been considered in the literature for the additive noise scenario (see e.g., [3], [4], [5], [8], [9], [10]):

i) the linear equation

$$dX\left(t\right)=\mathbf{A}X\left(t\right)dt+dW_{t},\qquad t\in\mathbb{R}, \tag{1}$$

with the matrix $\mathbf{A}$ having eigenvalues with negative real part and W is a (two-sided) standard Wiener process; and more recently

ii) the linear stochastic oscillator with additive noise

$$d\mathbf{X}\left(t\right)=\mathbf{A}\mathbf{X}\left(t\right)dt+\mathbf{b}dW_{t} \tag{2}$$

with

$$\mathbf{A}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix},\qquad \mathbf{b}=\begin{pmatrix}0\\ \sigma\end{pmatrix},\qquad \sigma>0,\ \ W \text{ an standard Wiener process}$$

where W is a standard Wiener process.

The equation (1) has an asymptotically stable solution and from a random dynamical viewpoint it has a random attractor of the flow of solutions, which is determined by its unique stationary solution [1]. The numerical preservation of these features has been studied by mean of the absolute stability ($A-$stability) property of numerical integrators. See [2] for additional details.

On the other hand, the stochastic oscillator equation (2) has distinctive long-time properties. Namely, it satisfies a linear growth property for the second moment [9]: $\mathbf{P1})\quad E\left(\|\mathbf{X}(t)\|^2\right)=\mathbf{x}_0+\mathbf{y}_0+\sigma^2(t-t_0)$ and the oscillatory property [9]: $\mathbf{P2})\quad \mathbf{X}^1(t)$ has infinitely many zeros, on $[t_0,\infty)$. On the other hand, since the oscillator (2) is a Hamiltonian system with additive noise, it preserves symplectic structure (cf. [6]). That is, it possess the property [10]: $\mathbf{P3})\quad d\mathbf{X}^1(t)\wedge d\mathbf{X}^2(t)=d\mathbf{x}_0\wedge d\mathbf{y}_0,\ \forall t\geq 0$.

In the present work we propose computationally viable integrators that in addition to replicate the qualitative behaviour of equation (1), they result in symplectic methods that reproduce the properties $\mathbf{P1}-\mathbf{P3}$ when applied to (2) and are optimal in the sense that use only increments of the Brownian path in its formulation. The schemes presented here are useful for many physical applications mainly when a stable long-time integration of the system is required.

2 THE PROPOSED METHODS

Let $(\Omega,\mathcal{F},P)$ be a complete probability space, and $(\mathcal{F}_t)_{t\geq 0}$ be an increasing right continuous family of complete sub σ-algebras of $\mathcal{F}$. Consider the d-dimensional SDE with additive noise

$$d\mathbf{X}(t)=\mathbf{f}\left(\mathbf{X}(t)\right)dt+\sum_{j=1}^{m}\mathbf{g}_j\left(t\right)dW_t^j,\qquad t\in[t_0,T],\qquad \mathbf{X}(t_0)=\mathbf{X}_{t_0},\tag{3}$$

where W_t^j, $j=1,\ldots,m$ are $\mathcal{F}_t$-adapted, uncorrelated standard Wiener processes. It is assumed that the $\mathbb{R}^d$-valued measurable functions $\mathbf{f},\mathbf{g}_j$, satisfy the standard conditions to ensure existence and uniqueness of a strong solution to the problem (3), [4]. Furthermore, let us suppose that $\mathbf{f}\in\mathcal{C}^1\left(\mathbb{R}^d,\mathbb{R}^d\right)$ and denote by $\mathbf{f}_\mathbf{x}$ the Jacobian of the function $\mathbf{f}$. Let $(t)_h=\{t_n:n=0,1,\ldots,N\}$ be a partition of the time interval $[t_0,T]$, with maximum stepsize $h<1$. Starting from the initial value $\mathbf{X}_0=\mathbf{X}_{t_0}$, approximations $\{\mathbf{X}_i\}$ to $\{\mathbf{X}(t_i)\}$, $(i=1,2,\ldots,N)$ can be obtained recursively as follows:

$$\mathbf{X}_{n+1}=\mathbf{X}_n+\mathbf{\Phi}(t_{n+1};t_n,\mathbf{X}_n)+\mathbf{Z}_n\left(t_{n+1}\right),\tag{4}$$

where

$$\mathbf{\Phi}(t;t_n,\mathbf{X}_n)=[\mathbf{I}_{d-1\times d-1}\ \ \mathbf{0}_{d-1\times 1}]\exp\left(\left[\begin{array}{cc}\mathbf{f}_\mathbf{x}(\mathbf{X}_n) & \mathbf{f}(\mathbf{X}_n)\\ 0 & 0\end{array}\right](t-t_n)\right)[\mathbf{0}_{1\times d-1}\ \ \mathbf{1}]^T$$

and $\mathbf{Z}_n\left(t_{n+1}\right)$ is an approximation in $t=t_{n+1}$ to the solution $\mathbf{R}(t)$ of the equation

$$d\mathbf{R}(t)=\mathbf{q}^{t_n,\mathbf{X}_n}\left(t,\mathbf{R}(t)\right)dt+\sum_{j=1}^{m}\mathbf{g}_j\left(t\right)dW_t^j,\qquad t\in[t_n,t_{n+1}],\qquad \mathbf{R}(t_n)=\mathbf{0},\tag{5}$$

with

$$\mathbf{q}^{t_n,\mathbf{X}_n}(t,\mathbf{R})=\mathbf{f}(\mathbf{X}_n+\mathbf{\Phi}(t;t_n,\mathbf{X}_n)+\mathbf{R})-\mathbf{f}_\mathbf{x}\left(\mathbf{X}_n\right)\mathbf{\Phi}(t;t_n,\mathbf{X}_n)-\mathbf{f}\left(\mathbf{X}_n\right).$$

A number of algorithms may be applied to compute the involved exponentials. In particular, those algorithms based on rational (p,q)-Padé approximation ($p\leq q\leq p+2$) combined with the "scaling and squaring" strategy, provide stable approximations to the matrix exponential. See the review in [7].

Clearly, from the mechanism described above, a variety of numerical methods can be constructed for solving equation (3) by applying a one-step numerical integrator to the auxiliary problem (5).

In this direction, when the Forward Euler-Maruyama (EM) method is used to integrate the auxiliary equation (5), the corresponding resultant method is (**LL_Euler method**):

$$\mathbf{X}_{n+1}=\mathbf{X}_n+\mathbf{\Phi}(t_{n+1};t_n,\mathbf{X}_n)+\sum_{j=1}^{m}\mathbf{g}_j\left(t_n\right)\Delta W_n^j,\tag{6}$$

with $\Delta W_n^j=W_{t_{n+1}}^j-W_{t_n}^j$.

Other explicit methods can be obtained by using predictor-corrector methods (P(EC)^1) to integrate (5). In this way, by taking the (explicit) Euler method as predictor and any $\theta-$method as corrector, new integrators are obtained. Hence for $\theta = 1$, the corresponding integrator in this case is (**LL_P(EC)^1**, with **predictor Euler, corrector Backward Euler (BE)**):

$$\mathbf{X}_{n+1} = \mathbf{X}_n + \mathbf{\Phi}(t_{n+1}; t_n, \mathbf{X}_n) + \left(\mathbf{q}^{t_n, \mathbf{X}_n} \left(t_{n+1}, \sum_{j=1}^{m} \mathbf{g}_j (t_n) \Delta W_n^j \right) \right) + \sum_{j=1}^{m} \mathbf{g}_j (t_n) \Delta W_n^j. \quad (7)$$

By following this line some other methods can be also devised. The flexibility of the proposed formulation allows us to select the appropriate method to combine with in concrete situations.

For the proposed methods we have the important result:

Theorem 1 *The numerical schemes obtained from (4) are A-stable (i.e, appropriate for integrating equation (1)) and when applied to the stochastic oscillator (2) they reproduce the long-time properties* $\mathbf{P1}$, $\mathbf{P2}$, $\mathbf{P3}$ *of this system.*

3 NUMERICAL EXPERIMENTS

In this section different numerical experiments are reported. For the sake of comparison, in each figure below, we use the same generated sample path of the Wiener processes in all the involved computations. The first experiment presents the evolution in the phase plane of system (2) with $\sigma = 1$ and for different initial conditions which are taken over the unit circle with centre at the origin. In Figure 1 the exact solution (which is simulated as indicated in [6]), the solution obtained by the P(EC)^1 method (with EM as predictor, BE as corrector), and the solution obtained by the corresponding (symplectic) LL_P(EC)^1 method are plotted at three time moments $T_1 = 30$, $T_2 = 70$, and $T_3 = 100$. We have used the stepsize $h = 0.05$. It is known (and can be observed in the figure) that for the linear stochastic oscillator, the exact images of the unit circle are circles with the unit radius shifted from the origin due to the influence of noise. For the numerical method we are testing, the images also keep this shape, but for the standard P(EC)^1 method the approximation worsens as the time of integration grows. Remarkably, the LL_P(EC)^1 method, which is a symplectic method, reproduces the exact image much better for every the moment of time.

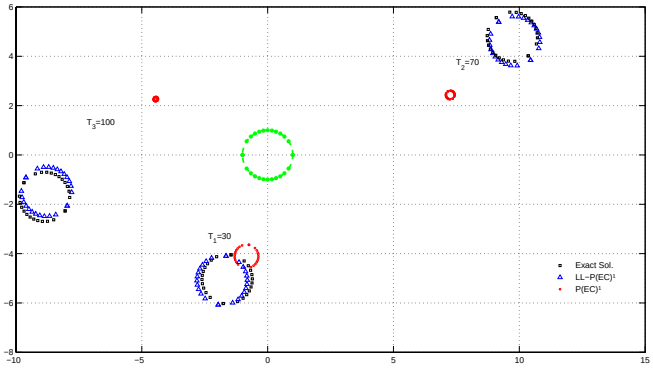


Figure 1: Trajectories in the phase plane of the exact solution and of the numerical approximation to the system (2) obtained by the P(EC)^1 method (with EM as predictor, BE as corrector) and the symplectic LL_P(EC)^1 method. The trajectories are obtained from different initial conditions, which are taken over the unit circle centred at the origin, at three moments of time, $T_1 = 30$, $T_2 = 70$ and $T_3 = 100$. The stepsize $h = 0.05$ is used for both integrators.

The next example illustrates the behaviour of the proposed methods in the integration of a Lotka-Volterra model with additive noise. Namely

$$\begin{aligned} d\mathbf{X}^1(t) &= \mathbf{X}^1(t) \left(\mathbf{X}^2(t) - 2 \right) dt \\ d\mathbf{X}^2(t) &= \mathbf{X}^2(t) \left(1 - \mathbf{X}^1(t) \right) dt + \sigma dW_t. \end{aligned} \quad (8)$$

The effect of the length of the integration interval on the performance of the standard and the proposed integrators is shown in Figure 2. Here, the stepsize is fixed at $h = 0.01$ for the Euler method, and a bigger stepsize $h = 0.2$ is considered for the LL_Euler method. The simulations are carried out at three moments of time, $T_1 = 10$, $T_2 = 35$, and $T_3 = 50$ and the initial condition was chosen as $(4; 2)$. As expected, the quality of the numerical approximations get worse as the time of integration increases. For the highest times T_2 and T_3 the standard Euler method results in explosive trajectories (for visualization purposes, in plotting the figure, coordinate axes were conveniently bounded in such a way that explosive behaviour is reflected by trajectories reaching the bounding frames). On the contrary it is observed that the proposed methods (exemplified by the LL_Euler) work perfectly, preserving the qualitative behaviour of the exact flow (the top row of the figure can be thought of as the actual solution), even though a higher stepsize $h = 0.2$ was used for the simulations in this case. We have corroborated the effectiveness of proposed methods and their superiority over the conventional ones.

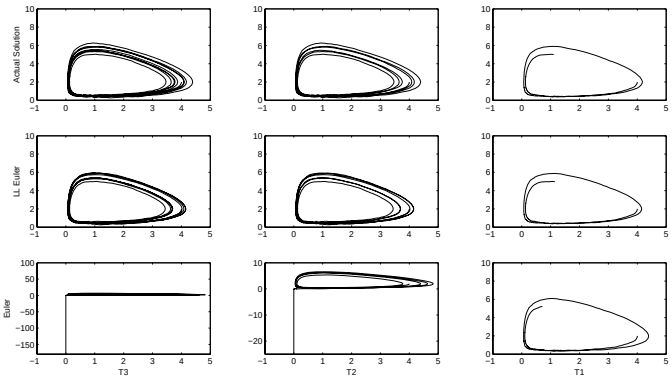


Figure 2: Trajectories in the phase plane obtained from numerical integration of equation (8) at three moments of time, $T_1 = 10$, $T_2 = 35$ and $T_3 = 50$. The Euler method and the corresponding LLEuler method were used in this comparison. The stepsizes $h = 0.01$ and $h = 0.2$ were used by the Euler and the LL_Euler integrator respectively.

ACKNOWLEDGMENTS

We thank the research support provided by CNPq under grant no. 500298/2009-2.

REFERENCES

[1] L. ARNOLD, *Random Dynamical Systems*, Springer-Verlag, Heidelberg, 1998.

[2] H. DE LA CRUZ, R.J. BISCAY, J.C. JIMENEZ, F. CARBONELL AND T. OZAKI, *High Order Local Linearization methods: an approach for constructing A-stable high order explicit schemes for stochastic differential equations with additive noise*, BIT, 50 (2010), pp. 509–539.

[3] D.B. HERNANDEZ AND R. SPIGLER, *A-stability of Runge-Kutta methods for systems with additive noise*, BIT Numerical Mathematics, 32 (1992), pp. 620-633.

[4] P.E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, Third Edition, 1999.

[5] G.N. MILSTEIN, *Numerical Integration of Stochastic Differential Equations*, 1995.

[6] G.N. MILSTEIN AND M.V. TRETYAKOV, *Stochastic Numerics for Mathematical Physics*, Springer, 2004.

[7] C. MOLER AND C.F. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Review, 45 (2003), pp. 3-49.

[8] H. SHURZ, *Numerical Analysis of Stochastic Differential Equations Without Tears*. In: Handbook of Stochastic Analysis and its Applications. Marcel Dekker, Inc: New York, 2002.

[9] A. H. STRØMMEN AND D. J. HIGHAM, *Numerical simulation of a linear oscillator with additive noise*, Appl. Numer. Math, 51 (2004), pp. 89–99.

[10] A. TOCINO, *A comparison among several numerical integrators to solve a linear stochastic oscillator*, AIP Conference Proceedings, 1048 (2008), pp. 994-998.

APLICACIÓN DE MÉTODOS DE CONTINUACIÓN NUMÉRICA AL CÁLCULO DE HIPER-LÍNEAS DE INTERÉS EN EL CAMPO DE LA TERMODINÁMICA DEL EQUILIBRIO ENTRE FASES

S. Belén Rodriguez-Reartes†, Juan I. Ramello†‡*, Gerardo Pisoni†‡*, Martín Cismondi‡* y Marcelo S. Zabaloy†

† Planta Piloto de Ingeniería Química - Universidad Nacional del Sur – CONICET
CC 717 - 8000 Bahía Blanca, Prov. Buenos Aires. ARGENTINA, mzabaloy@plapiqui.edu.ar
** IDTQ – Grupo Vinculado PLAPIQUI- CONICET*
‡ Facultad de Ciencias Exactas Físicas y Naturales. Universidad Nacional de Córdoba – Av. Vélez Sarsfield 1611, Ciudad Universitaria. X5016GCA – Córdoba – ARGENTINA.

En el campo de la termodinámica del equilibrio entre fases se requiere con frecuencia computar curvas que existen en un espacio multidimensional (hiper-curvas). Cada punto de una dada hiper-curva (hiper-punto) está definido por *n* componentes. De las *n* componentes de un dado hiper-punto, los valores de (*n-1*) de ellas surgen de resolver un sistema no lineal de ecuaciones algebraicas de dimensión (*n-1*) × (*n-1*). La componente restante constituye el único grado de libertad del sistema no lineal de ecuaciones asociado al hiper-punto considerado. Las hiper-líneas de interés, las cuales suelen ser altamente no lineales, pueden computarse en forma robusta si se utilizan los llamados métodos de continuación numérica (MCN). En este trabajo se aplican las formas más simples de los MCN al cálculo de hiper-lineas de distinto tipo: críticas, de equilibrio sólido-fluido, etc. Los resultados obtenidos permiten alcanzar una comprensión más profunda de los fenómenos físicos a los que corresponden las hiper-líneas computadas.

Palabras claves: *métodos de continuación numérica, termodinámica del equilibrio entre fases*

1. INTRODUCCIÓN

Cuando una mezcla de compuestos químicos, los cuales se encuentran en proporciones definidas (composición definida) en la mezcla, se somete a una temperatura y a una presión constante, durante suficiente tiempo, el sistema alcanza un estado de equilibrio. Bajo tal condición el sistema puede estar constituido por una única fase (fluida o sólida), o por dos o más fases, las cuales pueden ser sólidas o fluidas. Una fase en equilibrio es una porción del sistema en equilibrio en la que las propiedades macroscópicas son constantes (independientes del tiempo y de la posición espacial). El estado de equilibrio se caracteriza por la ausencia de fuerzas impulsoras que produzcan cambios en las propiedades macroscópicas del material. Una vez que se alcanza el estado de equilibrio, ya no se observarán cambios en el sistema con el paso del tiempo. En general, las fases que se encuentran en equilibrio tienen composiciones (y otras propiedades) distintas. Esto se aprovecha en diversos procesos industriales para separar las mezclas en fracciones que difieren en su composición. La termodinámica del equilibrio entre fases provee las ecuaciones que permiten describir matemáticamente las diversas situaciones de equilibrio de un dado material. Todo punto de equilibrio está representado por un sistema no lineal de ecuaciones algebraicas. El número de grados de libertad de tal sistema puede conocerse rápidamente aplicando la llamada Regla de las Fases de Gibbs. Con frecuencia se imponen valores fijos para todos los grados de libertad excepto uno. Ello define una hiper-línea en un espacio multidimensional. Ejemplos de tales hiper-lineas son, para el caso de mezclas de dos compuestos químicos (sistemas binarios), los siguientes: isotermas de equilibrio líquido-vapor, isobaras de equilibrio líquido-líquido, líneas de equilibrio líquido-líquido-vapor, líneas críticas, líneas de equilibrio sólido-líquido-vapor, etc. Las hiper-lineas de interés pueden tener un comportamiento altamente no lineal. Las mismas con frecuencia se calculan en forma incompleta en la literatura. Ello se debe al uso de métodos numéricos enfocados en el cálculo de puntos individuales (ej. Método de Newton-Raphson) en los que no se aprovecha en forma inteligente la información de un dado punto (convergido) de la hipercurva para facilitar el cómputo del siguiente punto. Tal aprovechamiento es característico de los llamados métodos de continuación numérica (MCN), cuyo uso no está generalizado en la comunidad de investigadores en el campo de la termodinámica del equilibrio entre fases. El propósito de este trabajo consiste en aplicar un método básico de continuación numérica al cómputo, en forma completa, de distintas hiper-líneas termodinámicas, de grado de no linealidad variable,

para sistemas materiales de dos compuestos químicos (sistemas binarios) y de tres compuestos químicos (sistemas ternarios).

2. MÉTODO BÁSICO DE CONTINUACIÓN NUMÉRICA

Consideremos un conjunto de *(n-1)* funciones reales no lineales, F_1 a F_{n-1}, de las variables reales X_1 a X_n. El número (*n*) es un número natural. El sistema no lineal de ecuaciones $\{F_i = 0\,, [i = 1, (n-1)]\}$ tiene un grado de libertad, pues el número de variables es igual a (*n*) mientras que el número de ecuaciones es igual a *(n-1)*. Identificaremos a tal sistema como el sistema P_{n-1}. Por poseer un único grado de libertad, el sistema P_{n-1} define una hiper-línea en el espacio (*n*)-dimensional de las variables X_1 a X_n. El grado de libertad disponible lo imponemos agregando al sistema la ecuación $\{F_n = 0\}$ donde $\{F_n = X_p - S\}$, siendo S un parámetro real y (*p)* un número natural tal que $1 \le p \le n$. Es claro que X_p puede ser cualquiera de las variables del vector $\vec{X}$ (cuyas componentes son X_1 a X_n). El vector $\vec{F}$ es aquel cuyas componentes son F_1 a F_n. El sistema $\{\vec{F} = 0\}$ es un sistema no lineal de (*n*) × (*n*) dependiente del parámetro S. Debido que el sistema $\{\vec{F} = 0\}$ depende del parámetro S, diremos que $\{\vec{F} = 0\}$ es un sistema "paramétrico" de ecuaciones no lineales, que identificaremos como el sistema P_n. Las variables del sistema P_n son las componentes del vector $\vec{X}$. Una vez impuesto un valor para el parámetro S el sistema P_n puede ser resuelto utilzando por ejemplo el método de Newton-Raphson. Supongamos que el primer punto de la hiper-línea que se quiere construir se obtiene para $S = S_I$. La información contenida en el Jacobiano del sistema P_n puede utilizarse para computar el llamado vector de sensitividades $\left(d\vec{X}/dS\right)\Big|_{S=S_I}$, el cual establece cómo cambia la solución del sistema P_n ante un cambio diferencial en el valor del parámetro S [1]. La información contenida en el vector de sensitividades se utiliza, vía extrapolación lineal, para predecir el vector solución $\vec{X}$ correspondiente a un nuevo valor para el parámetro S, como por ejemplo $S = S_{II}$. El vector de sensitividades también se utiliza para identificar la variable óptima a ser especificada para computar el siguiente punto de la hiper-línea: es aquella correspondiente a la componente del vector de sensitividades con máximo valor absoluto. Si la variable especificada difiere para dos puntos consecutivos de la hiper-línea, también el índice p, y consecuentemente la función F_n, difieren para tales puntos. El uso descripto del vector de sensitividades permite construir hiper-curvas altamente no lineales, las cuales pueden incluso autointersectarse. El hecho de especificar la variable óptima para el cálculo de cada punto de la hiper-línea, hace posible mantenerse, al definir el único grado de libertad disponible, dentro del rango de existencia de la hiper-línea. Consecuentemente se evitan problemas de convergencia originados por una especificación del grado de libertad correspondiente a la inexistencia de solución para el sistema P_n. Es posible definir métodos de continuación numérica más sofisticados que el descripto.

3. APLICACIÓN A PROBLEMAS DE TERMODINÁMICA DEL EQUILIBRIO ENTRE FASES

El método de continuación numérica (MCN) brevemente descripto en la sección previa es aplicable a cálculos de diversa naturaleza, de interés en la termodinámica del equilibrio entre fases. La fig 1 ilustra los resultados del uso del MCN para dos hiper-curvas isotérmicas: una de equilibrio líquido-vapor (estabilidad global) y otra correspondiente al límite de estabilidad intrínseca (línea de guiones). El sistema es un mezcla de dos compuestos químicos, esto es, dióxido de carbono (CO_2) e Isobutano. El modelo termodinámico

utilizado es el conocido modelo de Peng-Robinson, el cual depende de cierto parámetro k_{12} fijado en un valor de 0.134 durante la obtención de los resultados de la fig 1. La línea de guiones es una hiper-línea para la que cada hiper-punto corresponde a un sistema no lineal "paramétrico" de 3 × 3, siendo las variables del mismo la presión (P), la composición (xCO2) y la densidad ρ, mostrándose sólo la relación entre las dos primeras en la fig 1 (línea de guiones). La línea de guiones es altamente no lineal. La misma se intersecta a sí misma para un valor de P de aproximadamente 33 bar. Tal intersección no es, en general, estrictamente tal en el espacio multidimensional en que existe la hiper-línea. Un alto grado de no linealidad exige un alto grado de intervención durante el cálculo por parte del usuario, si no se recurre a un MCN como el aquí utilizado, el cual permitió obtener la curva de guiones completa en una única corrida. Es claro que la hiper-linea de guiones existe en un rango limitado de presión P, esto es, entre aproximadamente 8 bar (punto D, fig 1) y 60 bar (punto C, fig 1). El MCN aquí utilizado permite evitar especificaciones inapropiadas, como por ejemplo P = 70 bar, para la cual el sistema de ecuaciones no tiene solución en el campo de los números reales (el campo de interés en termodinámica).

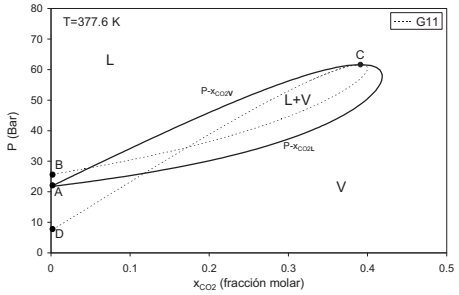


Fig. 1. Comportamiento de fases en el plano Presión-Composición para el sistema CO_2-Isobutano a 377.6 K. Modelo: PR-EOS con k_{12}=0.134. Línea contínua: hiper-curva de equilibrio Líquido-Vapor. G11: hiper-curva espinodal difusional (límite de estabilidad intrínsica).

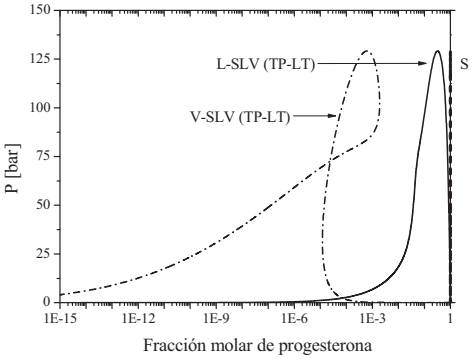


Fig. 2. Proyección presión-fracción molar de progesterona de la hiper-curva de equilibrio Sólido-Líquido-Vapor calculada para el sistema CO_2 + progesterona con el modelo de la ref [2] . Δv^{S-L} = -0.030006 m^3/Kmol.

Cada punto de la hiper-línea de trazo continuo de la fig 1 se obtiene resolviendo un sistema "paramétrico" no lineal de ecuaciones (P_n) de 5 × 5 cuyas variables son la presión "P", densidad de fase líquida ρ_L , densidad de fase vapor ρ_V , composición de fase vapor "X_{CO2V}" y composición de fase líquida "X_{CO2L}". Lo valores de las densidades no se muestran en la fig 1, mientras que "X_{CO2V}" y "X_{CO2L}" se pueden mostrar junto con la variable "P" en la fig 1 porque "X_{CO2V}" y "X_{CO2L}" son variables de naturaleza similar, a pesar de ser variables distintas. La curva "P-X_{CO2V}" muestra la relación entre la variables "P" y "X_{CO2V}" y la curva "P-X_{CO2L}" muestra la relación entre la variables "P" y "X_{CO2L}". La información que involucra a las variables P , "X_{CO2V}" y "X_{CO2L}" mostrada a través de la línea de trazo continuo de la fig 1 podría también mostrarse en un diagrama tridimensional en que cada uno de los ejes del sistema de coordenadas correspondiera a cada una de esas variables. A partir de la fig. 1 es también claro que la hiper-línea de trazo continuo es también altamente no lineal.

La fig 2 muestra una hiper-línea, calculada con el presente MCN, correspondiente al equilibrio sólido-líquido-vapor calculado para un sistema binario constituido por los compuestos químicos dióxido de carbono (CO_2) y progesterona. En este caso el sistema no lineal "paramétrico" de 7×7 correspondiente a cada punto de la hiper-linea tiene como sus variables a la presión "P", temperatura "T" , densidad de fase líquida ρ_L , densidad de fase vapor ρ_V , composición de fase vapor "xProg_vap" , composición de fase líquida "xProg_liq", y densidad del líquido hipotético constituido por progesterona pura ρ_L^{hyp} . La línea de

trazo continuo de la fig 2 muestra la relación entre "P" y "xProg_liq", mientras que la curva de guiones y puntos corresponde a la relación entre "P" y "xProg_vap". En este último caso nuevamente observamos un punto de auto-intersección, característico de un alto grado de no linealidad.

Otro caso al que nuestro grupo está actualmente aplicando el presente MCN es el cálculo de hiper-lineas críticas de fluidos ternarios. En el campo de la termodinámica del equilibrio entre fases un punto crítico de un fluido está dado por las condiciones en que el fluido es globalmente estable pero se encuentra en su límite de estabilidad intrínseca. Existen diversas formas de calcular un punto crítico. En uno de los enfoques posibles un punto crítico, para un fluido ternario particular, se obtiene resolviendo un sistema no lineal de 4 ecuaciones cuyas variables son la temperatura "T", la presión crítica "P_c", el volumen molar crítico υ_c , y las concentraciones de cada uno de los tres compuestos químicos (ω_1, ω_2 y ω_3). La figura 3 presenta la relación entre un par de las variables ("P_c" vs. υ_c) de una hiper-linea crítica isotérmica (de comportamiento altamente no lineal) del sistema ternario constituido por los compuestos químicos dióxido de carbono (CO2), agua (H2O) y Etano.

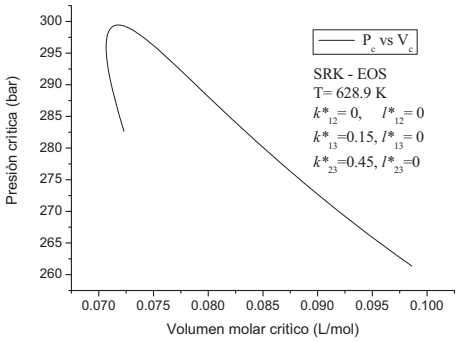


Fig. 3. Línea critica del sistema CO2(1)+H2O(2)+Etano(3), calculada utilizando el presente MCN. Modelo: SRK.

4. CONCLUSIÓN

En este trabajo se ha aplicado un método de continuación numérica (MCN) a cálculos termodinámicos. El MCN brevemente descripto en el presente trabajo resulta, más allá de su relativa simplicidad, de gran utilidad para el cómputo robusto de hiper-líneas correspondientes a problemas de interés en el campo de la termodinámica del equilibrio entre fases. Hemos mostrado ejemplos correspondientes a hiper-líneas críticas, de equilibrio líquido-vapor isotérmico, de equilibrio sólido-líquido-vapor, y de límite de estabilidad intrínseca. Existen otros tipos de hiper-líneas de interés, como es el caso de las envolventes de fases isopléticas.

El cómputo de estas últimas aplicando métodos de continuación numérica nos ha permitido recientemente comprender el comportamiento experimental de la mezcla binaria CO_2 + progesterona, la cual presenta el complejo fenómeno llamado "cristalización retrógrada".

AGRADECIMIENTOS

Se agradece el apoyo financiero del CONICET, de la Agencia Nacional de Promoción Científica y Tecnológica, de la Universidad Nacional del Sur y de la Universidad Nacional de Córdoba.

5. REFERENCIAS

[1] M. CISMONDI, M.L. MICHELSEN, M.S. ZABALOY. *Automated generation of phase diagrams for binary systems with azeotropic behavior*. Ind. Eng. Chem. Res., Vol. 47 (23), (2008), pp. 9728–9743

[2] S.B. RODRIGUEZ-REARTES, M. CISMONDI, E. FRANCESCHI, M.L. CORAZZA, J.VLADIMIR OLIVEIRA, M.S. ZABALOY. *High-pressure phase equilibria of systems carbon dioxide + n-eicosane and propane + n-eicosane*. J. Supercrit. Fluids Vol. 50 (2009), pp. 193–202.

Solución Numérica de la Potencia en un Reactor Nuclear Usando el Método de Hamming

Daniel Suescún Díaz†, Juan Felipe Flórez Ospina† y Carlos Alberto Lozano†

†Departamento de Ciencias Naturales y Matemáticas, Pontificia Universidad Javeriana Cali, A.A 26239, Cali Colombia, dsuescun@javerianacali.edu.co, www.puj.edu.co
†Departamento de Ciencias Naturales y Matemáticas, Pontificia Universidad Javeriana Cali, A.A 26239, Cali Colombia, jfflorez@javerianacali.edu.co, www.puj.edu.co
†Departamento de Ciencias Naturales y Matemáticas, Pontificia Universidad Javeriana Cali, A.A 26239, Cali Colombia, clozano@javerianacali.edu.co, www.puj.edu.co

Resumen: Se plantea un nuevo método para resolver numéricamente las ecuaciones de la cinética puntual en un reactor nuclear usado el método de Hamming, el cual converge con precisión del orden h^5 donde h es el paso en el tiempo de cálculo. El procedimiento es validado para diferentes formas de reactividad y con diferentes pasos de tiempo. Los resultados computacionales son satisfactorios, pues indican que el método es bastante aproximado y de esfuerzo computacional bajo comparado con otros métodos convencionales.

Palabras claves: ecuaciones de la cinética puntual*, Hamming, reactividad*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. Introducción

La potencia y la concentración de neutrones retrasados en un reactor nuclear se modelan por medio de ecuaciones de la cinética puntual. La solución computacional de las ecuaciones proporciona comprensión de las dinámicas de operación en un reactor nuclear, lo cual permite entender las fluctuaciones de potencia evidenciadas durante el encendido o pagado del reactor [1], particularmente cuando las barras del reactor son ajustadas.

Gran parte de los trabajos interesados en este caso de estudio se han focalizado en eliminar el confinamiento de Stiffness usando la aproximación de Padé en 1971 [2]. Chao y Attard en 1985 [3]. Luego, Sánchez implementó el método de Runge-Kutta generalizado en 1989 [4]. Más tarde Aboander y Nahla emplearon la técnica de inversión de polinomios en 2002 [5]. En este trabajo se plantea una nueva formulación para la solución numérica de las ecuaciones de la cinética puntual con diferentes reactividades mediante el método de Hamming y se realiza la comparación con los resultados obtenidos usando los métodos de Runge-Kutta y Diferencias Finitas, con el fin de validar los resultados.

2. Generalidades

Las ecuaciones de la cinética puntual se pueden obtener fácilmente a partir de la ecuación de la difusión de neutrones [6] (Duderstadt and Hamilton, 1976). Dichas ecuaciones forman un sistema de siete ecuaciones diferenciales acopladas no lineales, que describe la evolución en el tiempo de la distribución de neutrones y la concentración de los precursores de neutrones retardados en el núcleo de un reactor nuclear, donde el parámetro dependiente del tiempo en este sistema es la reactividad.

$$\frac{dP(t)}{dt} = \left[\frac{\rho(t)-\beta}{\Lambda}\right]P(t) + \sum_{i=1}^{6}\lambda_i C_i(t) \tag{1}$$

$$\frac{dC_i(t)}{dt} = \frac{\beta_i}{\Lambda}P(t) - \lambda_i C_i(t) \quad , \; i = 1,2,\ldots, 6 \tag{2}$$

Con las siguientes condiciones iniciales:

$$P(t=0) = P_0 \tag{3}$$

$$C_i(t=0) = \frac{\beta_i}{\Lambda\lambda_i} P_0 \tag{4}$$

Donde,

P(t) = potencia nuclear
$C_i(t)$ = concentración del i-ésimo grupo de precursores de neutrones retardados
$\rho(t)$ = reactividad
Λ = tiempo de generación de los neutrones instantáneos
β_i = fracción efectiva del i-ésimo grupo de neutrones retardados
β = fracción efectiva total de neutrones retardados $\left(\beta = \sum_i \beta_i\right)$
λ_i = constante de decaimiento de él i-ésimo grupo de los precursores de neutrones retardados.

3. Esquema predictor – corrector de Hamming

Sea la ecuación diferencial de la forma $y' = f(t_k, y_k)$ y dado que los esquemas del corrector predictor son de pasos múltiples, se requiere información de varios puntos previos a y_k para encontrar y_{k+1}, generalmente dichos puntos previos se calculan con el método de Runge-Kutta de cuatro orden.

El esquema predictor-corrector de Hamming utiliza el predictor de Milne, para aproximar la solución de la EDO haciendo uso del teorema fundamental del cálculo y el polinomio interpolador de Newton-Cotes de grado n igual a tres, para n valores equi-espaciados, la ecuación está dada por:

$$P_{k+1} = y_{k-3} + \frac{4h}{3}\left(2f_{k-2} - f_{k-1} + 2f_k\right) \tag{5}$$

Mediante formas de integración implícita y el método de los coeficientes indeterminados se da solución a la función f(t,y) teniendo en cuenta un orden P y k puntos anteriores. Al partir de la ecuación en diferencias con coeficientes indeterminados para un corrector generalizado de orden cuatro y k igual a tres se obtiene que:

$$y_{k+1} = \alpha_1 y_k + \alpha_2 y_{k-1} + \alpha_3 y_{k-2} + h\left[\beta_0 y_{k+1} + \beta_1 y_k + \beta_2 y_{k-1}\right] \tag{6}$$

En la ecuación (6) se pueden determinar los coeficientes en términos de α_2 (ver Tabla 1). Para α_2=1 se obtiene el corrector de Milne. Por otro lado, el análisis de estabilidad para la ecuación en diferencias del corrector generalizado realizado por Hamming, presenta que el valor de α_2 debe estar en el intervalo -0.6< α_2<1. Para α_2 igual a 1, 9/17, 1/9, 0, -1, -1/7 se obtiene buenos resultados, pero el mejor es obtenido con α_2=0, pues implica mayor zona de estabilidad.

$\alpha_1 = (1/8)(9 - 9\alpha_2)$	0	9/17	1	9/8	9/7	45/31	9/5
$\alpha_2 = \alpha_2$	1	9/17	1/9	0	-1/7	-9/31	-3/5
$\alpha_3 = -(1/8)(1 - \alpha_2)$	0	-1/17	-1/9	-1/8	-1/7	-5/31	-1/5
$\beta_0 = -(1/24)(9 - \alpha_2)$	1/3	6/17	10/27	3/8	8/21	12/31	2/5
$\beta_1 = (1/12)(9 + 7\alpha_2)$	4/3	18/17	22/27	3/4	2/3	18/31	2/5
$\beta_2 = (1/24)(-9 + 17\alpha_2)$	1/3	0	-8/27	-3/8	-10/21	18/31	-4/5

Tabla 1. Coeficientes en función de α_2 para optimizar el método de Hamming

Por lo anterior el corrector de Hamming queda expresado de la siguiente forma:

$$C_{k+1} = \frac{-y_{k-2} + 9y_k}{8} + \frac{3h}{8}\left(-f_{k-1} + 2f_k + f_{k+1}\right) \tag{7}$$

Se introduce un modificador para mejorar la precisión dada por las ecuaciones (5) y (7), este valor esta dado por:

$$m_{k+1} = P_{k+1} - \frac{112}{121}\left(P_k - C_k\right) \tag{8}$$

Sustituyendo el valor modificando en la ecuación (7), se obtiene:

$$C_{k+1} = \frac{-y_{k-2} + 9y_k}{8} + \frac{3h}{8}\left(-f_{k-1} + 2f_k + f(t_{k+1,} m_{k+1})\right) \tag{9}$$

Finalmente el valor de salida es dado por:

$$y_{k+1} = C_{k+1} + \frac{9}{121}\left(P_{k+1} - C_{k+1}\right) \tag{10}$$

Para la estabilidad del método se tiene la condición:

$$h < \frac{0.69}{\left|f_y\left(t - y\right)\right|} \tag{11}$$

4. Resultados

A continuación, se presentan los resultados del método implementado y su comparación con los métodos de Runge-Kutta y Diferencias Finitas, en los cuadros comparativos se considera el tamaño de paso h y el error máximo absoluto e_{max} en los puntos de referencia abordados en el presente estudio. Se observa que el método de Hamming tiene más precisión para la reactividades 800 pcm (partes por cien mil).

Método (300 pcm)	t=1s	t=10s	t=20s	h	e_{max}
Valor Exacto	2.2098	8.0192	28.2970	-	0
Hamming	2.209840	8.019199	28.297399	10^{-3}	0.000399
Runge Kutta	2.209840	8.019199	28.297399	10^{-2}	0.000399
Diferencias Finitas	2.2098	8.0192	28.2974	10^{-6}	0.000400

Tabla 2. Cuadro comparativo de las soluciones para cada método dada una reactividad de 300 pcm

Método (550 pcm)	t=0.1s	t=2s	t=10s	h	e_{max}
Valor Exacto	5.2100	43.025	$1.3886x10^5$	-	0
Hamming	5.215306	43.025143	$1.38860x10^5$	10^{-2}	0.005306
Runge Kutta	5.209959	43.025143	$1.38860x10^5$	10^{-2}	0.000041
Diferencias Finitas	5.2100	43.0251	$1.38860x10^5$	10^{-6}	0.000000

Tabla 3. Cuadro comparativo de las soluciones para cada método dada una reactividad de 550 pcm

Método (700 pcm)	t=0.01s	t=0.5s	t=2s	h	e_{max}
Valor Exacto	4.5088	$5.3459x10^3$	$2.0591x10^{11}$	-	0
Hamming	4.508851	$5.34588\ x10^3$	$2.059159\ x10^{11}$	10^{-2}	0.000059
Runge Kutta	4.508851	$5.34584\ x10^3$	$2.059089\ x10^{11}$	10^{-2}	0.000051
Diferencias Finitas	4.508855	$5.34570\ x10^3$	$2.05887\ x10^{11}$	10^{-6}	0.00023

Tabla 4. Cuadro comparativo de las soluciones para cada método dada una reactividad de 700 pcm

Método (800 pcm)	t=0.01s	t=0.1s	t=1s	h	e_{max}
Valor Exacto	6.2029	$1.4104x10^3$	$6.1634x10^{23}$	-	0
Hamming	6.20285	$1.41042\ x10^3$	$6.16335x10^{23}$	10^{-3}	0.00005
Runge Kutta	6.20285	$1.41042\ x10^3$	$6.16331\ x10^{23}$	10^{-3}	0.00009
Diferencias Finitas	6.202683	$1.410224x10^3$	$6.154747x10^{23}$	10^{-6}	0.008653

Tabla 5. Cuadro comparativo de las soluciones para cada método dada una reactividad de 800 pcm

Es bueno también observar la condición (11), ya que ella establece que $h < \frac{0.69 * \Lambda}{|\rho - \beta|}$, siendo ρ=β=700 pcm. Esto muestra que el paso de cálculo depende del tiempo de generación de los neutrones instantáneos Λ y

no podemos aumentar el tamaño de cálculo pues causaría oscilación en el valor de la potencia nuclear.

La Figura 1. muestra la potencia obtenida cuando la reactividad tiene la forma $\rho(t) = 0.005333 \sin\left(\frac{\pi}{T} t\right)$ con periodo T=100 s, idéntico resultado al obtenido por el método de Kinard y Allen en el 2004 [7].

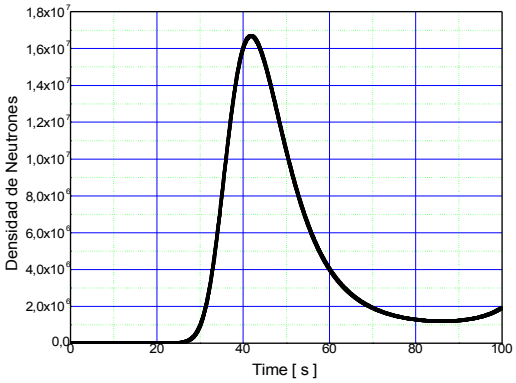


Fig 1. Densidad de la potencia nuclear en función de el tiempo

5. Conclusión

Se implementó el método de Hamming para resolver numéricamente las ecuaciones de la cinética puntual con un paso de $h=10^{-3}$ con una diferencia máxima de 0.000399 para reactividad con valor de 300 pcm. Para reactividades de 700 pcm el paso es aumentado 10 veces, es decir $h=10^{-3}$. Para reactividades de 800 pcm para obtener una buena aproximación el paso debe disminuir en 10 veces. Adicionalmente usando el método de Hamming para toda reactividad se disminuye el tiempo de computo en 276 s tomando como referente el método de diferencias finitas que arroja un tiempo de computo de 277.9 s para un intervalo de tiempo [0,20] en segundos.

También, cabe mencionar que dada la naturaleza de los algoritmos de Runge-Kutta a reactividades bajas o igual a 300 pcm su zona de estabilidad es más amplia, lo cual se evidencia en el tamaño del paso h, sin embargo la ventaja de Hamming con respecto a este método está dada por el número de funciones que deben evaluarse para obtener la solución, al comparar estos métodos el tiempo de cálculo del método Hamming es 0.6859 veces el tiempo de Runge-Kutta lo que garantiza mayor velocidad en la solución de problemas numéricos. Los resultados fueron obtenidos usando como lenguaje de programación Matlab y el computador con procesador Intel ® Core (TM) 2 Duo CPU T6500 @ 2.1 GHz, RAM de 4.00 GB con sistema operativo de 64 bits.

6. Referencias

[1] D.D. Suescun and A.M. Senra, *A Finite difference with exponential filtering in the calculation of reactivity.* Kerntechnik., 75 (2010), pp. 210-213.

[2] J.A.W. Da Norbrega, *A new solution of the point kinetics equations.* Nuclear Science and Enginneering., 46 (1971), pp. 366-375.

[3] Y. chao, and A. Attard, *A resolution to the stiffness problem of reactor kinetics.* Nuclear Science and Enginneering., 90 (1985), pp. 40-46.

[4] J. sanchez, *On the numerical solution of the point kinetics equations by generalized Runge-Kutta methods.* Nuclear Science and Enginneering., 103 (1989), pp. 94-99.

[5] A.E. aboander, and A.A. nahla, *Generalization of the analytical inversion method for the solution of the point kinetics equations.* Journal of Physics A: Mathematical and General., 35 (2002), pp. 3245-3263.

[6] J. duderstadt, and L. Hamilton, *Nuclear reactor analysis*, Wiley and Sons, 1976.

[7] M. Kinard, and E.J. Allen, *Efficient numerical solution of the point kinetics equations in nuclear reactor dynamics.* Annals of Nuclear Energy., 31 (2004), pp. 1039-1051.

CONSISTENT SPATIAL DISCRETIZATION OF THE KPZ EQUATION

Horacio S. Wio[♭], Jorge A Revelli[†] and Roberto R. Deza[‡]

[♭]*IFCA (UC and CSIC), Avda. de los Castros, s/n, E-39005 Santander, Spain, wio@ifca.unican.es, www.ifca.unican.es/users/wio*
[†]*IFEG (UNC and CONICET), Av. Luis Medina Allende, Ciudad Universitaria, 5000 Córdoba, Argentina, revelli@famaf.unc.edu.ar*
[‡]*IFIMAR (UNMdP and CONICET), Deán Funes 3350, B7602AYL Mar del Plata, Argentina, deza@mdp.edu.ar, fisica.mdp.edu.ar/CV/rdeza/personal*

Abstract: The feature which has made the stochastic nonlinear PDE known as *the Kardar–Parisi–Zhang* (KPZ) *equation* so successful as a mesoscopic model of surface and interface growth processes, is the phenomenologically grounded term $\frac{\lambda}{2}\left(\partial_x h\right)^2$ (a.k.a. "the KPZ term"). The usual discretization choices in finite-difference scheme, have been centered Laplacian $a^{-2}(h_{j+1}-2h_j+h_{j-1})$ and gradient $(2a)^{-1}(h_{j+1}-h_{j-1})$. This choice respects a discrete form of Galilean invariance (related to tilting invariance of the particle flow in microscopic growth models and thought to be at the heart of exact scaling relations in one spatial dimension). Other choices respect instead a fluctuation–dissipation theorem (also peculiar of 1D). We show that all those choices are inconsistent and moreover, that neither Galilean invariance nor the fluctuation–dissipation theorem are really an issue. We even propose a highly accurate spatial discretization scheme, which accelerates the setting of the asymptotic scaling regime.

Keywords: *KPZ, Variational Formulation, Galilean Invariance*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

Besides their crucial role in biology and environmental chemistry, surface and interface growth processes lie at the heart of multibillion-dollar industries, like semiconductor and pharmaceutical ones. That is why the appearance a quarter of century ago of a highly successful phenomenological mesoscopic model, the KPZ equation [1]

$$\partial_t h = \nu\,\partial_x^2 h + \frac{\lambda}{2}\left(\partial_x h\right)^2 + F + \varepsilon\,\xi(x,t), \tag{1}$$

was so celebrated. Its suitability for analytical work, its explicit symmetries, and its prediction of an exact dynamic scaling relation for a one-dimensional (1D) substratum, led people to adopt it as the "standard" model to describe the growth of rough interfaces, and it readily became a paradigm of nonequilibrium growth processes [2, 3]. Some of its interesting properties from a theoretical point of view are its exact mappings to the Burgers equation [4] and to a diffusion equation with *multiplicative* noise, whose field $\phi(x,t)$ can be interpreted as the restricted partition function of the directed polymer problem.

The KPZ equation in 1D has two main symmetries: Galilean invariance and the fluctuation–dissipation relation. Galilean invariance has been traditionally linked to the exactness of the relation $\alpha + z = 2$ among the critical exponents, in any spatial dimensionality [the roughness exponent α, characterizing the surface morphology in the stationary regime, and the dynamic exponent z, indicating the correlation length scaling as $\xi(t) \sim t^{1/z}$]. However, this interpretation has been criticized in this and other nonequilibrium models. The second symmetry essentially tells us that in 1D, the nonlinear KPZ term is not operative at long times.

Notwithstanding its theoretical interest, it is clear that in order to describe experiments and be able to extract results of interest to the industry, one must resort to numerical simulations and thus prescribe a spatial discretization scheme. We have recently shown [5, 6, 7] that the known fact that the KPZ equation stems (through a Hopf–Cole transformation) from a diffusion equation with multiplicative noise, strongly restricts the arbitrariness in the choice of spatial discretization schemes. On one hand, the discretization prescriptions for the Laplacian and the nonlinear (KPZ) term cannot be independently chosen. On the other, since the discretization is an operation performed on *space* and the Hopf–Cole transformation is *local* both in space and time, the former should be the same regardless of the field to which it is applied.

Whereas some discretization schemes pass both consistency tests, known examples in the literature do not. This leads us to a closer scrutiny of the relevance of Galilean invariance for determining the universality class of KPZ dynamics, as well as of a fluctuation–dissipation theorem, peculiar of 1D.

2 CONSISTENT SPATIAL DISCRETIZATION SCHEME

We start from a general scalar reaction–diffusion equation with multiplicative noise

$$\partial_t \phi(x,t) = \nu\, \partial_x^2 \phi(x,t) + \gamma\, \phi(x,t) + \phi(x,t)\, \eta(x,t), \tag{2}$$

where $\eta(x,t)$ is a Gaussian white noise, with zero mean and intensity σ, and we assume the Stratonovich interpretation. Exploiting the Hopf–Cole transformation we can now define a new field $h(x,t)$, that corresponds to an interface height, $h(x,t) = \frac{2\nu}{\lambda} \ln \phi(x,t)$, whose inverse is $\phi(x,t) = \exp\left[\frac{\lambda}{2\nu} h(x,t)\right]$. The transformed equation reads

$$\partial_t h(x,t) = \nu\, \partial_x^2 h(x,t) + \frac{\lambda}{2} (\partial_x h)^2 + \frac{\lambda\gamma}{2\nu} + \xi(x,t), \tag{3}$$

which is the celebrated KPZ equation. The noise term, which had a multiplicative character in Eq. (2), becomes additive. We use the standard, nearest-neighbor discretization prescription [2, 3] as a benchmark to elucidate the constraints to be obeyed by any spatial discretization scheme, arising from the mapping between the KPZ equation and Eq. (2). The standard spatially discrete version of Eq. (2) is

$$\dot{\phi}_j = \frac{\nu}{a^2}\, (\phi_{j+1} - 2\phi_j + \phi_{j-1}) + \frac{\lambda F}{2\nu} \phi_j + \frac{\lambda\varepsilon}{2\nu} \phi_j \xi_j, \tag{4}$$

with $1 \le j \le N \equiv 0$ (because of the assumed p.b.c.) and a the lattice spacing. Then, using the discrete version of the Hopf–Cole transformation

$$\phi_j(t) = \exp\left[\frac{\lambda}{2\nu} h_j(t)\right],$$

we get

$$\dot{h}_j = \frac{2\nu^2}{\lambda a^2} \left(\mathrm{e}^{\delta_j^+ a} + \mathrm{e}^{\delta_j^- a} - 2 \right) + \varepsilon\, \xi_j, \tag{5}$$

with $\delta_j^\pm \equiv \frac{\lambda}{2\nu a}(h_{j\pm1} - h_j)$ and $\gamma = 0 = F$. By expanding the exponentials up to terms of order a^2, and collecting equal powers of a (observe that the zero-order contribution vanishes) we retrieve

$$\dot{h}_j = \frac{\nu}{a^2}\, (h_{j+1} - 2h_j + h_{j-1}) + \frac{\lambda}{4\, a^2} \left[(h_{j+1} - h_j)^2 + (h_j - h_{j-1})^2 \right] + \varepsilon\, \xi_j. \tag{6}$$

As we can see, the first and second terms on the r.h.s. of Eq. (6) are *strictly* related by virtue of the discrete Hopf-Cole transformation: the discrete form of the Laplacian in Eq. (5) constrains the discrete form of the nonlinear term in the transformed equation. Latter we show again, in another way, the tight relation between the discretization of both terms. Known proposals fail to comply with this natural requirement.

An important feature of the Hopf–Cole transformation is that it is *local*, i.e., it involves neither spatial nor temporal transformations. An effect of this feature is that the discrete form of the Laplacian is the same, regardless of whether it is applied to ϕ or h.

2.1 THE FLUCTUATION–DISSIPATION RELATION

This relation is, together with Galilean invariance, a fundamental symmetry of the one-dimensional KPZ equation. It is clear that both symmetries are recovered when the continuum limit is taken in any reasonable discretization scheme. Thus, an accurate enough partition must yield suitable results. The stationary probability distribution for the KPZ problem in 1D is known to be [2, 3]

$$\mathcal{P}_{\mathrm{stat}}[h] \propto \exp\left\{ \frac{\nu}{2\,\varepsilon} \int dx\, (\partial_x h)^2 \right\}.$$

With the discretization scheme in Eq. (6), this is

$$\mathcal{P}_{\rm stat}[h] \propto \exp\left\{\frac{\nu}{2\varepsilon}\frac{1}{2a}\sum_j\left[(h_{j+1}-h_j)^2+(h_j-h_{j-1})^2\right]\right\}. \tag{7}$$

Inserting this expression into the stationary Fokker–Planck equation, the only surviving term has the form $\frac{1}{2a^3}\sum_j\left[(h_{j+1}-h_j)^2+(h_j-h_{j-1})^2\right]\times[h_{j+1}-2h_j+h_{j-1}]$, whose continuum limit is $\int dx\,(\partial_x h)^2\,\partial_x^2 h$, that is identically zero [2, 3]. A numerical analysis of such term indicates that it is several orders of magnitude smaller than the value of the exponents' pdf [in Eq. (7)], and typically behaves as $\mathcal{O}(1/N)$, where N is the number of spatial points used in the discretization. Moreover, it shows an even faster approach to zero if expressions with higher accuracy are used for the differential operators. In addition, when the discrete form of $(\partial_x h)^2$ is used together with its consistent form for the Laplacian, the fluctuation–dissipation relation **is not** exactly fulfilled. This indicates that the problem with the fluctuation–dissipation theorem in $1+1$ can be just circumvented by using more accurate expressions.

2.2 GALILEAN INVARIANCE

This invariance means that the transformation

$$x \to x-\lambda v t \quad , \quad h \to h + v x \quad , \quad F \to F - \frac{\lambda}{2}v^2, \tag{8}$$

where v is an arbitrary constant vector field, leaves the KPZ equation invariant. The equation obtained using the classical discretization

$$\partial_x h \to \frac{1}{2\,a}(h_{j+1}-h_{j-1}), \tag{9}$$

is invariant under the discrete Galilean transformation

$$ja \to ja - \lambda v t, \quad h_j \to h_j + v j a, \quad F \to F - \frac{\lambda}{2}v^2. \tag{10}$$

However, the associated equation is known to be numerically unstable, at least when a is not small enough. No other discretization is known to be invariant under the discrete Galilean transformation and that assertion includes Eq. (6). In fact, the transformation $h \to h + vja$ yields an excess term which is compatible with the gradient discretization in Eq. (9); however this discretization does not allow to recover the quadratic term in Eq. (6), indicating that this finite-difference scheme is not Galilean-invariant.

Since Eq. (4) is invariant under the transformation indicated in Eq. (10), it is the nonlinear Hopf–Cole transformation (within the present discrete context) which is responsible for the loss of Galilean invariance. Note that these results are independent of whether we consider this discretization scheme or a more accurate one. As already argued, Galilean invariance has always been associated with the exactness of the one-dimensional KPZ exponents, and with a relation that connects the critical exponents in higher dimensions [2, 3]. If the numerical solution obtained from a finite-difference scheme as Eq. (6), which is not Galilean invariant *yields the well known critical exponents*, that strongly suggests that Galilean invariance is not a fundamental symmetry as usually considered. The numerical results shown in [5, 6] are a clear indication that this is the case.

2.3 A MORE ACCURATE DISCRETIZATION SCHEME

A Taylor expansion of ϕ_{j+l} around ϕ_j shows that the general form of the discrete Laplacian, involving up to the n–th nearest neighbors of site j, is of the form

$$L_{(n)}(\phi_j) = \frac{\sum_{l=1}^n b_l\left[\Phi_j^{j+l}+\Phi_j^{j-l}\right]}{a\sum_{l=1}^n l^2 b_l}, \tag{11}$$

where the subscript stands for the number of nearest neighbors. Repeating the steps described above, one obtains

$$L_{(n)}(h_j) = \frac{\sum_{l=1}^{n} b_l \left[H_j^{j+l} + H_j^{j-l}\right]}{a \sum_{l=1}^{n} l^2 b_l}, \quad Q_{(n)}(h_j) = \frac{\sum_{l=1}^{n} b_l \left[\left(H_j^{j+l}\right)^2 + \left(H_j^{j-l}\right)^2\right]}{2 \sum_{l=1}^{n} l^2 b_l}. \tag{12}$$

The $\mathcal{O}(a^2)$ corrections to $L_{(n)}$ [applied to h_j in Eq. (12)] are of the form $\frac{2}{4!}\frac{\sum_{l=1}^{n} l^4 b_l}{\sum_{l=1}^{n} l^2 b_l}\partial_x^4 h$. Thus, the $\mathcal{O}(a^2)$ correction to $L_{(2)}^{(\gamma)}$ is $\frac{1}{12}\frac{1+7\gamma}{1+\gamma}\partial_x^4 h$. It attains its minimum value ($\frac{1}{12}\partial_x^4 h$) precisely for $\gamma = 0$, namely for $L_{(1)}$. What is then the convenience of a more complex prescription for the Laplacian? A wise criterion for choosing b_1 and b_2 in $L_{(2)}$ is making the $\mathcal{O}(a^2)$ corrections vanish. This yields the prescription $b_1 = 16$, $b_2 = -1$, known to be accurate up to corrections of $\mathcal{O}(a^4)$ [8]. Carrying out the already sketched procedure we obtain

$$L_{(2)}(h_j) \equiv \frac{4}{3}L_{(1)}(h_j) - \frac{1}{12}\left(H_j^{j+2} + H_j^{j-2}\right), \tag{13}$$

$$Q_{(2)}(h_j) \equiv \frac{2}{3}\left[(H_j^{j+1})^2 + (H_j^{j-1})^2\right] - \frac{1}{24}\left[(H_j^{j+2})^2 + (H_j^{j-2})^2\right]. \tag{14}$$

The $\mathcal{O}(a^2)$ corrections to $Q_{(n)}$ are

$$\frac{2}{4!}\frac{\sum_{l=1}^{n} l^4 b_l}{\sum_{l=1}^{n} l^2 b_l}\left[3\left(\partial_x^2 h\right)^2 + 4\left(\partial_x^3 h\right)\left(\partial_x h\right)\right],$$

which also vanishes for $b_1 = 16$, $b_2 = -1$. Since this discretization scheme fulfills the consistency conditions, is accurate up to $\mathcal{O}(a^4)$ corrections, and its prescription is not more complex than the ones studied before, it is obvious that it will be a convenient one to be used when a higher accuracy in numerical schemes is required.

3 Conclusions:

We have shown that (i) the problem with the fluctuation–dissipation theorem in 1D is tantamount to numerical accuracy; (ii) there is a strong evidence that Galilean invariance does not play the relevant role previously assumed in defining the KPZ universality class. We have moreover suggested a highly accurate consistent scheme.

Acknowledgments

HSW acknowledges financial support fromfrom MEC (Spain)—Project CGL2007-64387/CLI—and RRD from CONICET and UNMdP (Argentina).

References

[1] M. Kardar, G. Parisi, and Y.-C. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett. 56 (1986) pp.889–892.

[2] A.-L. Barabási and H. E. Stanley, *Fractal concepts in surface growth*, Cambridge U. Press, 1995.

[3] T. Halpin-Healy and Y-Ch. Zhang, *Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics*, Phys. Rep. 254 (1995) pp.215–414.

[4] H.C. Fogedby and W. Ren, *Minimum action method for the Kardar–Parisi–Zhang equation*, Phys. Rev. E 80 (2009) 041116.

[5] H.S. Wio, J.A.Revelli, R.R.Deza, C.Escudero and M.S. de La Lama, *KPZ equation: Galilean-invariance violation, consistency, and fluctuation–dissipation issues in real-space discretization*, Europhys. Lett. 89 (2010) 40008.

[6] H.S. Wio, J.A.Revelli, R.R.Deza, C.Escudero and M.S. de La Lama, *Discretization-related issues in the Kardar–Parisi–Zhang equation: Consistency, Galilean-invariance violation, and fluctuation–dissipation relation*, Phys. Rev. E 81 (2010) 066706.

[7] H.S. Wio, J.A.Revelli, R.R.Deza, C.Escudero and M.S. de La Lama, *Recent developments on the KPZ surface-growth equation*, Phil. Trans. Roy. Soc. A 369 (2011) pp.396–411.

[8] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1965.

ARBITRARY STEP TRANSPORTATION PROBLEM

By
Ezio Marchi
Emeritus Professor IMASL

Let $i:1,...,m$ be origins and $k:1,...,p$ be destinations. The merchandise, which is assumed to be indistinguishable, goes from an origin to a destination. However, the merchandise leaving an origin goes through a deposit $j:1,...,n$ and reaches a destination. Each origin i has a capacity r_i and each destination k needs the amount t_k. We have the common condition

$$\sum_{i=1}^{m} r_i = \sum_{k=1}^{p} t_k = r$$

which is concerned with the fact that all the merchandise required is distributed. Such condition is natural in transportation problems. Thus, if $x_{ij}^1 \geq 0$ and $x_{jk}^2 \geq 0$ are the respective total amount transported from origin i to the deposit j, and from there to destination k, then the two-step transportation problem can take the following expression:

$$\begin{aligned} &\sum_{j=1}^{m} x_{ij}^1 = r_i \qquad i \in \{1,...,m\} = I \\ &\sum_{j=1}^{m} x_{jk}^2 = t_k \qquad k \in \{1,...,p\} = K \qquad (1,1) \\ &\sum_{i=1}^{m} x_{ij}^i - \sum_{k=1}^{p} x_{jk}^2 = 0 \qquad j \in \{1,...,n\} = J \end{aligned}$$

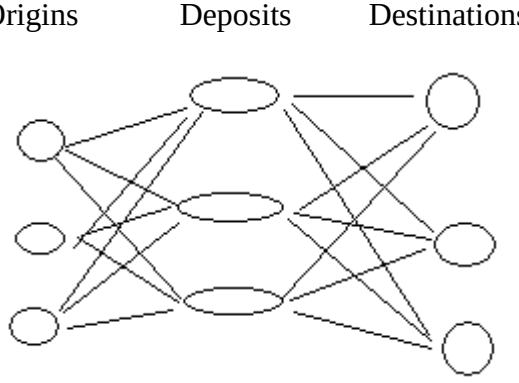
with $x = (x^1, x^2) \geq 0$

The last equation expresses the fact that at each deposits all the incoming amounts go out.

The conditions above are concerned with the total transported amounts, but the complete transportation problem is related to a cost function

$$\begin{aligned} f(x) &= c^1 x^1 + c^2 x^2 = \min! \qquad (1,2) \\ &= \sum_{ij} c_{ij}^1 x_{ij}^1 + \sum_{jk} c_{jk}^2 x_{jk}^2 = \min! \end{aligned}$$

which is linear, where $x^1 = \{x_{ij}^1\}$ and $x^2 = \{x_{ij}^2\}$. The amounts c_{ij}^1 and c_{jk}^2 are the costs to carry the unit amount from i to j and from j to k, respectively. A solution of (1,1) and (1,2) is called feasible.



It is possible to arrange the linear system (1,1) in a matrix form $Ax = b$, where A is:

1……..1	1……..1	. . . .	1……..1					↓ m ↑
				1……..1	1……..1	. . . .	1……..1	↓ p ↑
1 1	1 1	. . . . .	1 1	–1 . . . . –1	–1 . . . . –1	. . . . .	–1 . . . . –1	↓ n ↑
← mn →				← pn →				

and $b = (r_j, t_k, 0)^t$.

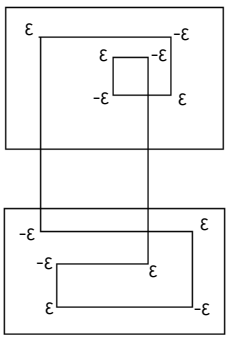
Definition: The support of $s = (y^1_{ij}, y^2_{jk})$ is the set $S(s) = \{(i,j)/\, y^1_{ij} > 0\} \cup \{(j,k)/\, y^2_{jk} > 0\}$

Definition: Let $s = (y^1_{ij}, y^2_{jk})$ a solution of the 2-step transportation problem, a cycle in the support of s is a sequence of elements of S(s), such that:

$$\sum_j y^1_{ij} = 0 \qquad \forall i$$

$$\sum_j y^2_{jk} = 0 \qquad \forall k$$

$$\sum_i y^1_{ij} = \sum_k y^2_{jk} \qquad \forall j$$



Characterization of extremals

Theorem: A feasible solution c of the 2-step transportation problem defined above is extremal if and only if it has no cycles in its support S(c).

Theorem: Each extremal of the problem has at most $m + p + n - 1$ positive components.

Algorithm: Take $(\bar{i}, \bar{j}, \bar{k})$ and consider $a_1 = \min(r_{\bar{i}}, t_{\bar{k}}) = \bar{x}^1_{\bar{i}\bar{j}} = \bar{x}^2_{\bar{j}\bar{k}}$. If $a_1 = r_{\bar{i}}$, delete the row $\bar{i}$ and take the problem in $I - \{\bar{i}\}, J, K$. The new problem has the entries r_i for $i \neq \bar{i}$, and $t_{\bar{k}} - r_{\bar{i}} \geq 0$ for $k = \bar{k}$ and t_k for $k \neq \bar{k}$. Follow as in the first step. In this way adding up all the entries in the different steps in the respective places it converges to a solution of (1,1). We call it an algorithm which is clear that converges.

Theorem: The algorithm converges and provides all the extremals.

Example: Consider a 2-step problem with 2 origins, 3 deposits and 2 destinations with matricial form:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_{11}^1 \\ x_{12}^1 \\ x_{13}^1 \\ x_{21}^1 \\ x_{22}^1 \\ x_{23}^1 \\ x_{11}^2 \\ x_{21}^2 \\ x_{31}^2 \\ x_{12}^2 \\ x_{22}^2 \\ x_{32}^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

which it has as solutions:

solution 1:

1	1			
5	2	3		
	3			3
		3		3

solution 2:

1	0.5	0.5		
5	2.5	2.5		
	1.5	1.5		3
	1.5	1.5		3

Solution 1 is extremal, solution 2 it is not because is convex combination of solutions:

1	1			
5	5			
	3			3
	3			3

1		1		
5		5		
		3		3
		3		3

The two step transportation problem cannot be formulated as a particular case of the classic transportation problem.

It is remarkable that the two step transportation problem cannot be reduced to a classic transportation problem.

Problem with constraints

The equation (1,1) plus the following one which is related with constraints on the deposits

$$\sum_{i=l}^{m} x_{ij}^1 \le s_j \qquad j = 1,...,n$$

with

$$\sum_{i}^{m} r_i = \sum_{k} t_k \le \sum_{j} s_j$$

determine the following incidence matrix A_S and it is not empty:

1……..1	*1……..1*	. . . .	*1……..1*					↓ m ↑
				1……..1	*1……..1*	. . . .	*1……..1*	↓ p ↑
1 . . . . *1*	*1* *1*	. . . . .	*1* *1*	*–1* . . . . *–1*	*–1* . . . . *–1*	. . . . .	*–1* . . . . *–1*	↓ n ↑
1 . . . . *1*	*1* *1*	. . . . .	*1* *1*					↓ n ↑
⟷ *mn*				⟷ pn				

The rank of this matrix is $A_S = (m + p + 2n - 2)$

Extremals

Extremals are obtained in a similar manner as the problem without constraints. The support of extremals is bounded by (m + p + 2n – 2). This matter is rather more difficult.

Arbit6rary step transportation problem

It is possible to extend this problem to an arbitrary number of spets in a natural way.

Bibliography:

Brualdi, R.: Introductory Combinatory. North Holland. 1978

Marchi, E. and Tarazaga, P.: About Two Step Transportation Problem. Relatorio Inter. #54. IMECC, UNICAMP, Brazil. 1978.

Tarazaga, P and Oviedo, J.: Sobre el Convexo de Transporte con Capacidades Máximas. Actas Congreso Mat. Aplicada, Rio de Janeiro, Brazil. 2002. pp 697-717

Marchi, E. A Variational formulation of the two step transportation problem. International Journal of Mathematics, Game Theory and Algebra. Vol. 19. No 3. 2010

ON VULNERABILITY OF UNITARY CAYLEY GRAPHS

Daniel A. Jaume[1], Adrián Pastine[1] and Denis E. Videla[1]

[1]*Universidad Nacional de San Luis, Ejército de los Andes 950, 5700 San Luis, Argentina, djaume@unsl.edu.ar, www.unsl.edu.ar*

Abstract: The unitary Cayley graph X_n has vertex set $Z_n = \{0, 1, \ldots, n-1\}$. Vertices a, b are adjacent, if $gcd\,(a-b, n) = 1$. In this work we give bounds for the toughness and the integrity of X_n and we found the toughness and the integrity for some subclasses of unitary Cayley graphs.

Keywords: *Unitary Cayley Graphs, toughness, integrity, vulnerability.*
2000 AMS Subject Classification: 05C25, 05C50

1 INTRODUCTION

Let Γ be a multiplicative group with identity 1. For $S \subset \Gamma$, $1 \notin S$ and $S^{-1} = \{s^{-1} : s \in S\} = S$ the *Cayley Graph* $X = \text{Cay}\,((\Gamma, S)$ is the undirected graph having vertex set $V\,(X) = \Gamma$ and edge set $E\,(X) = \{a, b : ab^{-1} \in S\}$. By right multiplication Γ may be considered as a group of automorphisms of X acting transitively on $V(X)$. The Cayley graph X is regular of degree $|S|$. Its connected components are the right cosets of the subgroup generated by S. So X is connected, if S generates Γ. More information about Cayley graph can be found in the books on algebraic graph theory by Biggs [1] and by Godsil and Royle [3].

For a positive integer $n > 1$ the *unitary Cayley graph* $X_n = \text{Cay}(Z_n, U_n)$ is defined by the additive group of the ring Z_n of integers modulo n and the multiplicative group U_n of its units. If we represent the elements of Z_n by the integers $0, 1, \ldots, n-1$ then it is well known that $U_n = \{a \in Z_n : \text{gcd}(a, n) = 1\}$. So X_n has vertex set $V(X_n) = Z_n = \{0, 1, \ldots, n-1\}$ and edge set $E(X_n) = \{a, b : a, b \in Z_n, \text{gcd}(a-b, n) = 1\}$.// The graph X_n is regular of degree $|U_n| = \varphi(n)$, where $\varphi(n)$ denotes the Euler totient function. Let p be a prime number, then $X_p = K_p$ (the complete graph on p vertices). Let α be a positive integer, the X_{p^α} is a complete p-partite graph which has the residue classes modulo p in Z_n as maximal sets of independent vertices. Unitary Cayley graphs are highly symmetric. They have some remarkable properties connecting graph theory and number theory (for example it can be proved that $\varphi(n)$ is even, for $n > 2$, via a graph theory argument, using unitary Cayley graphs).Unitary Cayley Graphs represent very reliable network, so it is important to find its toughness and integrity.

2 PRELIMINARIES

Definition 1 *Let $n \in \mathbb{N}$ we define $p_1(n)$ as the smallest p prime number such that $p|n$.*

Some properties of unitary Cayley Graphs were determined by Klotz and Sander [4]. We list here in the next theorem all the invariants from the work of Klotz and Sander that we will need

Theorem 1 *Let X_n be a unitary Cayley graph of order n then*

1. *the chromatic number $\chi(X_n) = p_1(n)$.*
2. *the clique number $\omega(X_n) = p_1(n)$.*
3. *the independence number $\beta(X_n) = \frac{n}{p_1}(n)$.*
4. *the vertex connectivity $\kappa(X_n) = \varphi(n)$.*

Definition 2 *Given a graph G we define $N(G)$ as the maximum order of a component of G. In the same way we define $k(G)$ as the number of components of G*

Definition 3 *Given a connected graph G, the **toughness** of G, a no-complete graph, $t(G)$ is defined as:*

$$t(G) = \min \frac{|S|}{k(G-S)}$$

where the minimum is taken over all the vertex-cuts sets S.

The toughness of the complete Graph K_n is defined as $t(K_n) = +\infty$

In this work we will make use the following known result about toughness:

Theorem 2 *For every noncomplete graph G,*

$$\frac{\kappa(G)}{\beta(G)} \leq t(G) \leq \frac{\kappa(G)}{2}$$

Theorem 3 *For every noncomplete graph G,*

$$t(G) \leq \frac{|V(G)| - \beta(G)}{\beta(G)}$$

Definition 4 *Given a connected graph G, the **integrity** of G, $I(G)$ is defined as:*

$$\min |S| + N(G-S)$$

where the minimum is taken over all the $S \subset V(G)$.

We will make use of the following known result about integrity:

Theorem 4 *Let G be a Graph of order $n \geq n$ with degree sequence $d_l, d_2, \ldots, d_n$, with $d_1 \geq d_2 \geq \ldots \geq d_n$. Then*

$$\min_{1 \leq t \leq n} \{\max\{t, d_t + 1\}\} \leq I(G) \leq n - \beta(G) + 1$$

3 RESULTS

To begin this section we will write the bounds from Theorem 2 and Theorem 3 in the case of Unitary Cayley Graphs:

Lemma 1 *Let X_n be a noncomplete Unitary Cayley Graph, then:*

$$\frac{\varphi(n)}{\frac{n}{p_1(n)}} \leq t(X_n) \leq \frac{\varphi(n)}{2}$$

$$t(X_n) \leq \frac{n - \frac{n}{p_1(n)}}{\frac{n}{p_1(n)}} = p_1(n) - 1$$

Its important to notice that if n is not a prime number φ properties give us the following inequality:

$$\frac{\varphi(n)}{2} \geq p_1(n) - 1$$

Lemma 2 *If $m|n$, $m \in \mathbb{Z}_+$ then X_n is m-partite, with each part having $\frac{n}{m}$ vertices.*

Proof. Given two vertices a and b, if $a \equiv_p b$ then a and b are not connected. Using this we can form a partition of X_n so that each congruence class modulo p is a part of the partition. This way each part of the partition has $\frac{n}{p}$ vertices. □

Lemma 3 *X_{pq} is p-partite with each vertex being connected to $q-1$ vertices of every other part.*

Proof. As we know from Lemma 2 we can partition X_{pq} in p parts, where each part corresponds to a congruence class modulo p. Consider the vertex 0, the vertices to which it is not connected are the ones that are not relatively primes with pq. This are the numbers of the form mp or mq, with $m \in \mathbb{Z}_+$. The numbers of the form mp are all in the same congruence class as 0. There are p numbers of the form mq, one in each congruence class. This means that 0 is connected to all but one vertex in every other part. Using now that the graph is vertex-transitive we have that this is the case for every vertex. □

Theorem 5 *Let p be a prime number and $k \in \mathbb{Z}^+$ then $t(X_{p^k}) = p - 1$*

Proof. As we know $t(X_n) \geq \frac{\kappa(X_n)}{\beta(X_n)}$. Letting $n = p^k$ we have that

$$\frac{\kappa(X_n)}{\beta(X_n)} = \frac{\varphi(p^k)}{\frac{p^k}{p}}$$

$$= \frac{p^k - p^{k-1}}{p^{k-1}}$$

$$= p - 1$$

Thus, using the upper bound from Lemma 1, we have that:

$$p - 1 \leq t(X_n) \leq p_1(n) - 1 = p - 1$$

□

Lemma 4 *Let $p < q$ be prime numbers and S vertex-cut set with $|S| < pq - q$ then $k(X_{pq} - S) = 2$.*

Proof. As we know from Lemma 2 X_{pq} is p-partite, with each part having q elements. Lets call the parts $A_0, A_1, \ldots, A_{p-1}$. We also know from Lemma 3 that any vertex is not connected to only one vertex in any other part of the partition. Suppose now that

$$|S| < pq - q$$

Then

$$|V(X_{pq} - S)| > pq - (pq - q) = q$$

This means that $X_{pq} - S$ has at least $q + 1$ vertices. As $q > p$, by the pigeonhole principle there is an i such that $|A_i \cap V(X_{pq} - S)| \geq 2$. Using the pigeonhole principle again, as $|X_{pq} - S| > q$ there is a $j \neq i$ such that $|A_i \cap V(X_{pq} - S)| \geq 1$. Without loss of generality we will assume that $i = 1$ and $j = 2$.
Take two vertices a_1 and a_2 from $A_1 \cap V(X_{pq} - S)$, clearly any vertex in any other part is connected to at least one of this vertices. Take now a vertex b in A_2, it is connected to every vertex of A_1 except one. Now we have only two components, the one that has b, and the one that has the vertex from A_1 that is not connected to b. □

Theorem 6 *Let $p < q$ be prime numbers, then $t(X_{pq}) = p - 1$*

Proof. By Lemma 4 we have that if S is a vertex-cut set with $|S| < pq - q$ then $k(X_{pq} - S) = 2$. Using that the connectivity of X_{pq} is $\kappa(X_{pq}) = \varphi(pq) = (p - 1)(q - 1)$ we have that if $S < pq - q$

$$\frac{|S|}{k(X_{pq} - S)} \geq \frac{\varphi(pq)}{2}$$

$$= \frac{(p - 1)(q - 1)}{2} \geq p - 1$$

Suppose now that $S \geq pq - q$ then we have

$$\frac{|S|}{k(X_{pq} - S)} \geq \frac{pq - q}{\beta(X_{pq})} = p - 1$$

And so we have that the toughness is $p - 1$. □

We make a stop here to point out that in both cases we found that the upper bound was actually the toughness, $p_1(n) - 1$, we conjecture that this is true for every noncomplete Unitary Cayley Graph.

Before going on with the results on integrity we will write the bounds from Lemma 4 in the language of Unitary Cayley Graphs:

$$\varphi(n) + 1 \leq I(X_n) \leq n - \frac{n}{p_1(n)} + 1$$

Theorem 7 *Let p be a prime number and $k \in \mathbb{Z}_+$, then $I(X_{p^k}) = p^k - p^{k-1} + 1$*

Proof. We know that

$$I(X_n) \geq \kappa(X_n) + 1$$

Letting $n = p^k$

$$I(X_{p^k}) \geq \kappa(X_{p^k}) + 1 = p^k - p^{k-1} + 1$$

But we also have that:

$$I(X_{p^k}) \leq |V(X_{p^k})| - \beta(X_{p^k}) + 1 = p^k - p^{k-1} + 1$$

Thus $I(X_{p^k}) = p^k - p^{k-1} + 1$ as we wanted to prove. □

Theorem 8 *Let $p < q$ be prime numbers, then $I(X_{pq}) = pq - q + 1$*

Proof. Let S be a vertex-cut set such that $|S| < pq - q$. By Lemma 4 we have that $k(X_{pq}) - S = 2$. Using the pigeonhole principle we have that:

$$N(X_{pq} - S) \geq \frac{pq - |S|}{2}$$

And so we have, using that $p < q$ and that, as S is a vertex-cut, $S \geq \kappa(X_{pq}) = (p-1)(q-1)$:

$$|S| + N(X_{pq} - S) \geq |S| + \frac{pq - |S|}{2} \geq \frac{S + pq}{2} \geq pq - \frac{p+q}{2} + \frac{1}{2} > pq - q$$

Assume now that $|S| \geq pq - q$, then:

$$|S| + N(X_{pq} - S) \geq pq - q + 1$$

And so $I(X_{pq}) = pq - q + 1$. □

Notice that in both cases the integrity is equal to the upper bound $n - \frac{n}{p_1(n)} + 1$, we conjecture that this is the case for every Unitary Cayley Graph.

REFERENCES

[1] N. BIGGS, *Algebraic graph theory, Second Edition*, Chematical library. Cambridge University Press, 1993.

[2] G. CHARTRAND AND L. LESNIAK, *Graphs and Digraphs, Third Edition*,Chapman and Hall, 1996.

[3] C. GODSIL AND R. ROYLE, *Algebraic graph theory*, Graduate Text in Mathematics. Springer, 2001.

[4] W. KLOTZ AND T. SANDER, *Some properties of unitary Cayley graphs*, The Electronic Journal of Combiantorics 14 (2007), R45, pp. 1-12.

PERFIL DE POTENCIAL ELECTROQUÍMICO EN TUBOS DE CONDENSADOR DE CENTRAL DE GENERACIÓN DE ENERGÍA

Mariana Corengia, Víctor Martínez – Luaces y Mauricio Ohanian

IIQ, Facultad de Ingeniería, UDELAR, Julio Herrera y Reissig 565, Montevideo, Uruguay, corengia@fing.edu.uy

Resumen: La disímil característica de las aleaciones utilizadas en la construcción de condensadores de plantas de generación de potencia puede inducir una severa corrosión galvánica. Las patologías presentes incluyen picado en los tubos de intercambio, los cuales ocasionan paradas no previstas de planta, con el consiguiente perjuicio económico.
En el presente trabajo se analiza la distribución de potencial en tubos del condensador cuando se aplica un potencial dado a sus extremos. El procedimiento empleado incluye un la linealización de la condición de borde determinada en un ensayo electroquímico independiente. La EDP del potencial escalar en geometría cilíndrica se simplifica a una dimensión y es resuelto analíticamente empleando una discretización basada en cajas móviles. El desarrollo permite la toma de decisiones en el diseño (potencial de los extremos) de sistemas en los que cuales se emplea protección catódica tanto por corriente impresa o por ánodos de sacrificio.

Palabras claves: *corrosión, protección catódica, geometría cilíndrica*
2000 AMS Subjects Classification: 34B15, 34B60, 65L10

1. INTRODUCCIÓN

Debido a los distintos requerimientos de los materiales utilizados en la construcción de condensadores de plantas de generación de potencia (mecánicos, de transferencia de calor, de confiabilidad o económicos), es común que las distintas partes del condensador estén construidas en distintos materiales. Lo usual es encontrar condensadores fabricados con su caja de acero al carbono, tubos de titanio, acero inoxidable o aleaciones de cobre y placas de tubos de una aleación de cobre distinta. Todas estas aleaciones mecánicas presentan diferentes características respecto a la corrosión electroquímica, y debido a que las mismas se encuentran en contacto eléctrico se puede inducir corrosión galvánica. La llamada serie galvánica provee una indicación de cuál combinación de metales son susceptibles de generar un par [1]. La velocidad de corrosión no puede ser predicha por los potenciales de corrosión de la serie galvánica, sino que depende de factores geométricos, las resistencias a la polarización de la superficie y la relación de áreas de los electrodos. Debido a la complejidad del caso considerado –condensador de central de generación eléctrica- la distribución de corriente no es uniforme y por tanto utilizar los métodos tradicionales de predicción no tiene validez. En el caso del condensador es de esperar altas densidades de corriente galvánica en el contacto caja de condensador/placa de tubos, y en la unión placa de tubos/tubos, cayendo la corriente a valores insignificantes en el tubo a medida que nos alejamos de la placa. En el caso de que haya protección catódica aplicada –en la caja-, se polarizará la placa catódicamente, en tanto su efecto puede no ser visto por la totalidad del tubo, actuando el mismo – o parte de él – como ánodo. Debido a lo anterior, el potencial a lo largo del tubo no es constante, y es de interés determinar el perfil de potencial con respecto a su largo, teniendo como parámetros los potenciales aplicados a los bordes.

En el presente trabajo se presenta la determinación del perfil de potencial del interior de un tubo, simplificando el Laplaciano de potencial a la dimensión longitudinal del mismo. Se presenta la determinación de la curva corriente vs. potencial, en un ensayo electroquímico de laboratorio, la cual es una de las condiciones de borde del campo mencionado.

2. PARTE EXPERIMENTAL

Los ensayos electroquímicos se realizan en una celda de tres electrodos, con un potenciostato Voltalab PGZ301. Se utilizó *Admiralty brass* como electrodo de trabajo, platino platinado como contraelectrodo y un electrodo saturado de calomel (SCE, *saturated calomel electrode*) como referencia. El electrolito es sulfato de sodio 0.1M. La fluidodinámica de la determinación es de régimen completamente turbulento. Temperatura de trabajo 20±2 ºC. Aeración natural a través de la superficie de electrolito. Se realizó un barrido de potencial de –0.45 V a –0.05 V vs. SCE, a una velocidad de 10 mV/s.

3. DESARROLLO DEL MODELO

Atsley [2] presenta el modelado del problema de distribución de corriente y potencial en geometría cilíndrica. El autor reseña condiciones del sistema (geométricas y físicas) para las cuales es válido realizar una aproximación unidireccional en el sentido de flujo, la cual expresa como:

$$R < 2\left(\rho\left(di/dE\right)\right)^{-1} \tag{1}$$

Donde R [m] y ρ [Ω m] son el radio del tubo y la resistividad del electrolito respectivamente y *(di/dE)* la pendiente de la curva densidad de corriente vs. potencial.

De esta manera se elimina la consideración del término radial. Por otra parte, por razones de simetría, se elimina el término angular del Laplaciano correspondiente al balance de potencial eléctrico. En las condiciones reseñadas Atsley reporta que:

$$\frac{\partial^2 E}{\partial z^2} = \pm\frac{2\rho i}{R} \tag{2}$$

Donde E [V] es el potencial electroquímico dependiente de la posición longitudinal z e i es la densidad de corriente [A m^{-2}] circulante normal a las paredes del tubo.

Para la pared del tubo las condiciones se obtienen a partir de experiencias electroquímicas de laboratorio; el resultado es el vínculo entre la corriente circulante y potencial aplicado (llamadas curvas de polarización). En la Figura 1 se representa la condición mencionada para *Admiralty brass* en sulfato de sodio 0.1M.

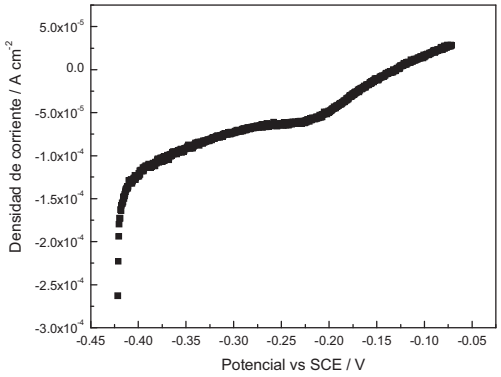


Figura 1: Curva de polarización de *Admiralty brass* en sulfato de sodio 0.1M.

Dichas curvas experimentales no son representables con expresiones funcionales simples y/o de utilidad práctica. A menudo se representan, para condiciones electroquímicas dadas, la dependencia en ciertos rangos de potencial como aproximaciones lineales o lineales semilogarítmicas del potencial contra la corriente (aproximación lineal o de Tafel respectivamente), las cuales no son generalizables a todas las condiciones presentadas en nuestro problema.

En resumen se tiene la EDO que representa el campo de potencial en el sistema (ecuación 2), la condición de borde en las paredes del tubo que surge de la Figura 1 y se asume potencial constante en los extremos del tubo:

$$E\big|_{z=0} = E_1 \quad E\big|_{z=L} = E_2 \tag{3}$$

3.1. RESOLUCIÓN DEL PROBLEMA

En las condiciones consideradas en el presente trabajo, se verifica la desigualdad de (1).

De modo de resolver analíticamente la EDO, la condición de borde dada por la curva experimental representada en la Figura 1 se linealiza. Para cada par experimental de potencial y densidad de corriente, se

trabaja con una caja que contiene los diez pares de puntos anteriores y los diez puntos posteriores, además del propio punto considerado. A esta población (pares (*E*, *i*)) se le ajusta un modelo lineal simple por mínimos cuadrados, y de dicho modelo se asigna –punto a punto del registro (*E*, *i*)– la ordenada en el origen y la pendiente de la tangente de la curva. Por tanto tenemos en cada punto una dependencia de tipo:

$$\frac{\partial E}{\partial r} = i = a + bE \tag{4}$$

La solución para el campo escalar de potencial y la condición de borde local de (4) es para el caso considerado (en que la curva experimental es creciente):

$$E = c_1 \exp(\psi z) + c_2 \exp(-\psi z) - \frac{a}{b} \tag{5}$$

Siendo

$$\psi = \sqrt{2\rho b / r} \tag{6}$$

Imponiendo la condición de borde (4) se obtienen los coeficientes c_1 y c_2

$$c_1 = -\frac{(E_1 + a/b)\exp(-\psi L) - E_2 - a/b}{2\sinh(\psi L)}$$

$$c_2 = -\frac{E_2 + a/b - (E_1 + a/b)\exp(\psi L)}{2\sinh(\psi L)} \tag{7}$$

3.2. AJUSTE DE LOS DATOS EXPERIMENTALES AL MODELO

Se realizan los cálculos con las siguientes condiciones: largo del tubo L=9 m, radio del tubo r=0,0254 m, resistividad de la solución ρ=0,40 mho, paso h=0,01m.

3.2.1 Cálculo preliminar

Se calcula el perfil de potencial vs *z* (distancia al extremo) en dos sentidos diferentes *Creciente* y *Decreciente* de la siguiente manera:

a- Método *Creciente* en *z*

$$E_1\big|_{z=ih} = E\big|_{z=(i-1)h}$$

$$E_1\big|_{z=1\times h} = -0{,}42$$

$$E_2 = -0{,}38 \tag{8}$$

$$L\big|_{ih} = 9 - ih$$

Con (8) se calculan las constantes c_1 y c_2 mediante (7) y posteriormente se calcula el potencial de la caja con (5). El potencial calculado sirve como E_1 en la primer ecuación de (8) en el cálculo de la caja siguiente en *z*.

a- Método *Creciente* en *z*

$$E_1\big|_{z=ih} = E\big|_{z=(i+1)h}$$

$$E_1\big|_{z=1\times h} = -0{,}38$$

$$E_2 = -0{,}42 \tag{9}$$

$$L\big|_{ih} = 9 - ih$$

Mediante similar procedimiento al mostrado en Fordward se realiza el cálculo en el sentido opuesto.

Posteriormente, como semilla del proceso iterativo del cálculo de perfil de potencial se toma un promedio ponderado, en base a la distancia, del cálculo en ambos sentidos:

$$E_i = \frac{L - z}{L} * E_{i,Creciente} + \frac{z}{L} * E_{i,Decreciente} \tag{10}$$

3.2.2 Proceso iterativo

Se realiza el ajuste mediante un proceso iterativo, teniendo como criterio de cierre que el cálculo de la diferencia media cuadrática (RMSD, ecuación (11)) entre los valores de dos iteraciones sucesivas sea menor a 1‰.

$$RMSD = \sqrt{\sum_{z=0}^{L} \left(E_z^j - E_z^{j-1}\right)^2} \tag{11}$$

Parámetros de iteración

$$E_1\Big|_{z=ih}^{j} = E\Big|_{z=(i-1)h}^{j-1}$$

$$E_2\Big|_{z=ih}^{j} = E\Big|_{z=(i+1)h}^{j-1}$$

$$E\Big|_{z=0} = -0{,}42$$

$$E\Big|_{z=L} = -0{,}38 \tag{12}$$

$$L\Big|_{ih} = 0{,}01$$

En la Figura 2 se representan los resultados de los ajustes, representándose los valores iniciales Backward y Fordward y los valores de iteraciones primera y segunda. El RMSD $_{j=2}$=0,5‰.

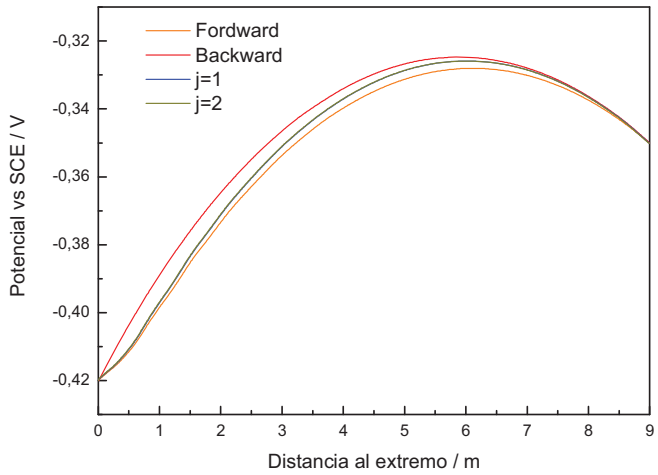


Figura 2: Resultados de perfil de potencial para procedimiento fordward, backward y primeras iteraciones.

4 CONSIDERACIONES FINALES

Mediante el procedimiento descripto se calcula el prefil de potencial a lo largo de un tubo, pudiéndose obtener valores críticos de protección frente a la corrosión (i.e. máximos de potencial) y/o fijar diferentes potenciales en extremos.

REFERENCIAS

[1] BAECKMAN, W., SCHWENK, W., PRINZ, W. "Handbook of Cathodic Corrosion Protection". Gulf Professional Publishing, 1997, pp 27-78.

[2] ASTLEY, D.J., "Use of the microcomputer for calculation of the distribution of galvanic corrosion and cathodic protection in seawater systems". *Galvanic corrosion*, ASTM STP 978, H. P. Hack, Ed., American Society for Testing and Materials, Philadelphia, 1988, pp. 53-78

DIAGNÓSTICO DE FALLAS EN MATERIAL COMPUESTO DE FIBRA DE CARBONO (CFRP) USANDO REDES NEURONALES

Adriana Zapico†+, Leonardo Molisani†, Ronald O'Brien†, Juan C. del Real‡, Yolanda Ballesteros‡ y Nicolás Ponso†

†Universidad Nacional de Río Cuarto, Enlace Ruta 36 Km. 601, X5800BYA Río Cuarto, Argentina Teléfono:+54-9358-4031155, e-mail:adrianazapico@gmail.com

+Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Av. Rivadavia 1917 CP C1033AAJ Ciudad de Buenos Aires, Argentina

‡ Universidad Pontificia Comillas de Madrid, Alberto Aguilera, 23, 28015 Madrid, España

Resumen: El uso del sonido para diagnosticar fallas data de la antigüedad. En este trabajo se usa el nivel de presión sonora para diagnosticar fallas como método evaluador global no destructivo en vigas de material compuesto de epoxi reforzado con fibras de carbono. Disminuyendo drásticamente el costo de los instrumentos de diagnóstico y evitando el indeseado efecto de masa agregada que ocurre cuando se usa un acelerómetro en pequeñas probetas del material. Es habitual que en el diagnostico de fallas se requieran varias medidas para evaluar el daño, en este trabajo nos concentraremos en un análisis global que solo requiere una medición por viga. Los niveles de presión sonora varían según las vigas estén sanas o dañadas y teniendo en cuenta la profundidad del daño. Se usa una red neuronal supervisada con conexiones hacia delante con algoritmo de entrenamiento backpropagation á la Levenberg-Marquardt para clasificar las vigas según el daño, teniendo en cuenta el nivel de presión sonora emitido por la viga. Este es un trabajo preliminar, en un próximo paso se testeará con mayor cantidad de vigas y con vigas ligeramente mayores la función de la red para clasificar. La idea principal de este trabajo es que una vez fijada la estructura se desarrolle una red neuronal para cada caso particular y se instrumente electrónicamente sobre la estructura para detectar las fallas propias de la estructura en condiciones operacionales

Palabras claves: *Diagnostico de Fallas, Redes Neuronales, Nivel de Presión Sonora.*

1. INTRODUCCIÓN

El uso de sonido en el diagnostico de fallas data de la antigüedad ya que golpeando un objeto se puede distinguir cuando el objeto esta fallado. En general, la evaluación no destructiva es utilizada para detectar y localizar defectos usando señales con una longitud de onda menor o igual que el defecto a ser detectado. Actualmente, las técnicas de ultrasonido son comúnmente usadas en ingeniería para la determinación de materiales, medición de espesores, adherencia de capas pegadas, y en metalurgia para la determinación de la calidad de las soldaduras en piezas metálicas. Pero esta técnica requiere que el objeto sea analizado en muchas secciones pequeñas. Las técnicas convencionales requieren la observación en largos intervalos o el uso de poderosos mecanismos para imponer movimientos en la estructura. Las fallas en las estructuras metálicas causan pequeños cambios en las frecuencias naturales de vibrar. Por otro lado, experiencias previas muestran que pequeños cambios en la resonancia producen variaciones en las propiedades dinámicas de la estructura.

En otros trabajos de la literatura pública se han utilizado funciones de respuesta en frecuencia (FRF) [1-6], para diagnosticar fallas en vigas de acero, en vigas de aluminio y en vigas de material compuesto en ensayos globales. Hay dos razones para cambiar la técnica ya propuesta [2] para clasificar las vigas de material compuesto de epoxi reforzado con Fibra de Carbono (CFRP). La primera razón es de origen técnico y es que cuando se usan las funciones de respuesta en frecuencia se usa un acelerómetro, pero el peso de un acelerómetro normal de bajo costo supera el 10 % del peso de las vigas de material compuesto con lo cual se crea el denominado efecto de masa agregada, esto es que se modifican las propiedades de comportamiento dinámico de la viga y por ende la caracterización de las propiedades del material. La segunda razón es los costos del equipamiento de mediciones que limita su uso a proyectos con un alto presupuesto. Por estas razones se utiliza un micrófono para medir los niveles de presión sonora emitido por las vigas sin afectar el peso y por lo tanto la dinámica de comportamiento de la viga. Para ello se utilizan

como entrada a la red neuronal [7] los niveles de presión sonora *Lp* emitido por las vigas bajo una carga impulsiva. La red clasifica las vigas según su daño en sanas o con ranuras (en el medio de la viga) de 1 mm, 1.7 mm o 2.3 mm.

En la próxima Sección introduciremos el concepto de nivel de presión sonora. Luego en la Sección 3 incluiremos la estructura de red necesaria para resolver este problema. Finalmente, en la última Sección incluiremos algunas conclusiones y trabajos futuros.

2. NIVELES DE PRESIÓN SONORA

La señal acústica proveniente de la estructura radiante es captada por un micrófono. El nivel de presión sonora es definido por

$$L_p(\omega) = 10\, Log_{10}\left(\frac{P_{rms}^2}{P_{ref}^2}\right) \tag{1}$$

Siendo P_{rms} la raíz cuadrada media de la presión acústica y P_{ref} la presión de referencia de *20 μPa* que es la mínima variación de presión acústica que distingue el odio humano.

Debido a que no tenemos la posibilidad de medir la intensidad del golpe impulsivo para obtener unicidad. Los datos se normalizaron con respecto al valor máximo obtenido en medición sobre la viga sana.

$$\bar{\bar{L}}_p^{(i)}(\omega) = \frac{\bar{L}_p^{(i)}(\omega)}{\max_{\omega}\left[\bar{L}_p^{(1)}(\omega)\right]} \tag{2}$$

Donde

$$\bar{L}_p^{(i)}(\omega) = \frac{1}{2}\left[L_p^{(i)}(\omega) - 20\, Log_{10}\left(\frac{s}{\sqrt{2}P_{ref}}\right)\right] \tag{3}$$

Siendo *s* la sensibilidad del micrófono cuyo valor para este caso particular es de 125 Pa/mV.

En la figura 1 se observa como los modos de vibrar primero, tercero y quinto se encuentran modificados teniendo en cuenta la presencia de daño y tamaño de la falla. Aumentando la profundidad de la falla se produce mayor desplazamiento hacia la izquierda. Así la línea azul representa la viga sana y la línea roja representa la más dañada. Siendo las curvas verdes y cian tamaños de falla intermedios.

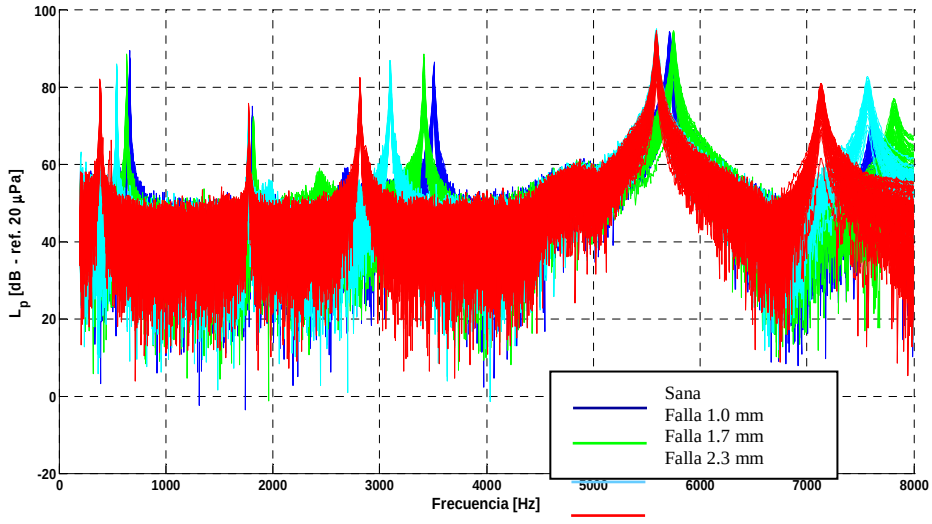


Figura 1 : Nivel de Presión Sonora L_p.

3. UNA RED NEURONAL PARA CLASIFICAR DAÑOS EN BASE A LOS NIVELES DE PRESIÓN SONORA

En este trabajo de investigación se utilizaron vigas de epoxi reforzado con fibras de carbono, son vigas de 175 mm de longitud, 25 mm. de ancho y 3 mm. de espesor, con ranuras de 1 mm, 1.7mm y 2.3 mm. de profundidad para simular el daño en el material. Se cuenta con 4 vigas, una con cada tipo de falla. Para el entrenamiento y validación de la red se tomaron 60 mediciones individuales de cada tipo de daño, utilizándose un 60% de los datos para entrenar, un 20% para testear y un 20% para validar. La condición de apoyo de las probetas es libre-libre.

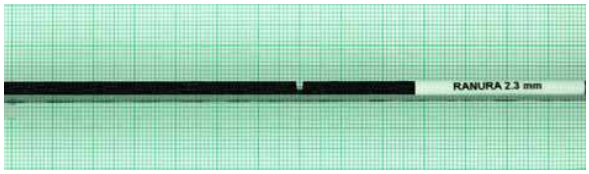


Figura 2: Viga de epoxi reforzado con fibra de carbono con una ranura de 2.3mm

Se utiliza un banco de ensayos que es un soporte universal de cuerdas y un micrófono para obtener los niveles de presión sonora que son procesados por un software especialmente diseñado. Estos niveles de presión sonora son las entradas de la red neuronal. Los soportes de cuerdas se ubican en lugares que evitan el bloqueo de los modos de vibrar de la probeta, y son indicados por el software desarrollado.

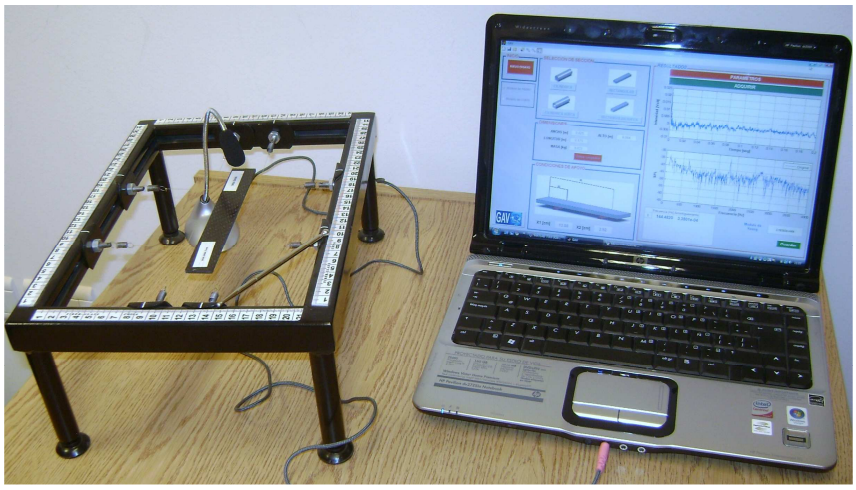


Figura 3: Equipamiento usado para medir los niveles de presión sonora.

Se utilizó una red neuronal artificial totalmente interconectada con conexiones hacia delante (feedforward) con aprendizaje supervisado con algoritmo de entrenamiento backpropagation á la Levenberg-Marquardt

La estructura de la red tiene una capa oculta. La primera función de transferencia en la red es una función sigmoidea y la segunda es una función lineal. La estructura de la red es de 230-8-2 neuronas. La entrada de las neuronas es una selección(con un salto de 25Hz) de las líneas espectrales de las curvas de nivel de presión sonora en el intervalo 2000-7000 Hz, donde como se puede apreciar en la Figura 1 se presenta un notable corrimiento de las curvas. Las dos neuronas de la capa de salida sirven una para clasificar según estén dañadas o sanas y la otra para determinar la profundidad del daño. La red converge (en un promedio de 10 épocas) con un error medio promedio (MSE) de entre 1.20%-1.60%. De esa manera una vez obtenida la convergencia puede utilizarse la técnica para detectar daños en las vigas de material compuesto CFRP.

4. CONCLUSIONES Y TRABAJO FUTURO

En este esfuerzo se ha logrado medir y detectar el efecto de fallas sin el efecto de masa agregado de sensores como por ejemplo acelerómetros. Se ha obtenido una disminución del costo del equipamiento de toma de mediciones para el diagnostico de fallas en vigas de epoxi reforzado con fibras de carbono. Se obtuvo una red neuronal que clasifica según los cuatros estados simulados de daños de la viga. Este es un trabajo preliminar se testeará con mayor cantidad de vigas y con vigas ligeramente mayores la función de la red para clasificar. Es un paso preliminar necesario para una investigación más amplia y con mayor cantidad de datos. Esto es razonable dadas las dificultades en encontrar un conjunto representativo de vigas falladas con fallas no artificiales, que realmente constituyan un conjunto de entrenamiento realista (sin necesidad de producirlo artificialmente, lo que requiere de inversiones en dinero). La idea principal de este trabajo es que una vez fijada la estructura se desarrolle una red neuronal para cada caso particular y se instrumente electrónicamente sobre la estructura para detectar las fallas propias de la estructura en condiciones operacionales. El bajo número de épocas y el reducido error obtenido se deben a que se trabaja con pocas vigas en realidad, simulándose la variación mediante diferentes mediciones independientes.

Figure 4. Colapso de una viga de material compuesto del brazo de un elevador mecánico

En una investigación previa [9], se trabajó con vigas de epoxi reforzadas con fibra de vidrio. Allí se utilizó también los niveles de presión sonara como entrada de una red neuronal para clasificar el daño de las vigas. Claramente la detección de fallas con mediciones globales usando redes neuronales es una herramienta de gran interés y bajo costo que es posible de ser utilizada en dispositivos en tiempo real.

Se planea usar esta técnica en elevadores mecánicos de mantenimiento de líneas eléctricas de alto voltaje donde se montan operarios que trabajan con la línea con alta tensión. Los brazos de estos tipos de elevadores mecánicos están constituidos por material compuesto y acero. El desafío en estos elevadores es incorporar mediciones en tiempo real y que se utiliza la combinación de varios materiales, en general material compuesto y acero. La idea es evitar los colapsos de los brazos de los elevadores mecánicos que pueden ocasionarse por fatiga del material o defectos en el material involucrado.

REFERENCIAS

1. Zapico, A. and L. Molisani. *USO DE REDES NEURONALES PARA LA DETECCION DE FALLAS EN VIGAS DE ACERO*. in *MACI 2009*. 2009. Rosario.
2. Zapico, A., et al. *Determinación No Destructiva de Fallas en Materiales Compuestos Utilizando Redes Neuronales*. in *XI Congreso de Adhesión y Adhesivos*. 2010. Madrid.
3. Zapico, A., et al., *Global Fault Detection Using Artificial Intelligence.* Journal of Adhesion Science and Technology, 2010. **Submitted**.
4. N Mohamad, et al. *Artificial Neural Network for the Classification of Steel Hollow pipe*. in *International Conference on Applications and Design in Mechanical Engineering*. 2009. Penang,Malasya.
5. Zang, C. and M. Imregun, *Structural damage detection using artificial neural networks and measured frf data reduced by principal component projection.* Journal of Sound and Vibration 2001. **24**(5): p. 813–827.
6. Zapico, A. and L. Molisani, *Fault Diagnosis on steel structures using artificial neural networks.* Mecánica Computacional 2009. **Vol XXVIII**: p. 181-188.
7. Bishop, C.M., *Neural Networks for Pattern Recognition* 1995: Oxford University Press.
8. Hagan, M. and M. Menhaj, *Training feed-forward networks with the Marquardt algorithm.* IEEE Transactions on Neural Networks 1994. **5**(6): p. 989-993.
9. Zapico, A., et al. *Fault Analysis in Composed Material: A Neural Net Application Using Acoustical Signal*. in *MECOM 2010-CILAMCE 2010*. 2010. Buenos Aires.

DETECCIÓN DE FUENTES SONORAS MEDIANTE EL USO DE IMÁGENES ACÚSTICAS

Ronald J. O'Brien†, Leonardo Molisani†, Ricardo Burdisso‡

†*Grupo de Acústica y Vibraciones (GAV), Universidad Nacional de Río Cuarto, Ruta Nac. Nº 36 Km 601, 5800 Río Cuarto, Argentina, e-mail: robrien@ing.unrc.edu.ar, lmolisani@ing.unrc.edu.ar.*
‡*Vibration and Acoustics Laboratories, 153 Durhan Hall, Virginia Tech, Blacksburg, VA, 24061-0238, USA, rburdiss@vt.edu, http://www.val.me.vt.edu*

Resumen: Actualmente para medir los niveles máximos de emisión sonora las regulaciones se establecen a través de mediciones globales. Con respecto al límite de los mismos las regulaciones son cada vez más exigentes debido al efecto nocivo que produce el ruido en la población y medioambiente. Por lo tanto es necesario determinar que partes de la fuente emisora contribuyen principalmente al ruido global para así establecer estrategias de control. En este trabajo se utilizó la técnica de construcción de imágenes acústicas mediante beamforming a través del uso de una antena ubicada a cierta distancia de la fuente emisora. La antena acústica está constituida por un arreglo espacial de sensores de presión. La distribución de sensores de presión fue optimizada y constatada experimentalmente. La distribución espacial de los sensores de presión afecta al rango dinámico de la antena y se manifiesta a través del Máximo Lóbulo Lateral (MLL). Mediante algoritmos genéticos se optimizó el parámetro MLL de la antena acústica. La técnica de imágenes acústicas es utilizada en este esfuerzo comparando distintas distribuciones de sensores de presión, obteniéndose resultados satisfactorios entre simulaciones y experimentos. Esta tecnología permite localizar espacialmente, en tiempo y frecuencia que partes de las fuentes sonoras contribuyen al ruido global. Conocer los puntos de emisión de las distintas partes de la fuente de ruido es necesario para el control de las mismas, permitiendo cumplimentar las exigencias de las normas sobre regulaciones sonoras emitidas hacia el ambiente.

Palabras claves: *imágenes acústicas, beamforming, Máximo Lóbulo Lateral, algoritmos genéticos.*

1. INTRODUCCIÓN

La capacidad de detectar fuentes de ruido [1] producidas por partes mecánicas móviles y ruido aerodinámico permitirá tomar los recaudos necesarios para cumplir con regulaciones, normas de emisión de ruido y control de ruido [2]. La primera aplicación de un phased array bidimensional (2D) para realizar mediciones en aeronáutica fue realizada por Brooks esté usó por primera vez un arreglo de sensores de presión 2D, para la cual utilizó la técnica delay and sum en el dominio de la frecuencia. Con el paso de los años se fueron mejorando los algoritmos para la detección de fuentes sonoras, en 1974 Högbom desarrollo CLEAN [3] usando la deconvolución de las imágenes emitidas por los radioastronomos, desde allí se adaptó para el uso en imágenes acústicas. Dougherty and Underbrink usaron un arreglo 2D en el año 86, implementando la técnica convencional de beamforming. En el año 1987 Cox desarrollo el algoritmo "Robust Adaptative Beamforming" [4], esta técnica permite reducir el ruido blanco no correlacionado. Los algoritmos aplicados a la técnica beamforming más recientes son DAMAS [5], LORE [6] y CLEAN-SC [7], los cuales logran detectar y separar fuentes con mejor resolución que el Beamforming Convencional. En este trabajo a partir de algoritmos genéticos se optimiza la distribución de sensores de presión a modo de evitar obtener mediciones erróneas, aumentar la resolución de la antena y su rango dinámico. De esta forma disminuye la carga computacional en el post-proceso de datos experimentales.

2. TECNOLOGÍA DEL ARREGLO DE MICRÓFONOS

El propósito de esta sección es presentar los conceptos físicos y vocabulario sobre el beamforming convencional y mostrar una aplicación concreta de una antena de micrófonos para la detección de fuentes sonoras. La técnica Delay and Sum y la técnica de beamforming, la cual es utilizada para realizar el post-proceso de datos experimentales, son descriptas a continuación.

2.1. TÉCNICA DE DELAY AND SUM BEAMFOMER

En el dominio de la frecuencia la salida del beamforming se expresa de la siguiente manera [8]

$$b(\mathbf{k},\omega) = \sum_{m=1}^{M} P_m(\omega)\, e^{i\mathbf{K}.r_m} \qquad (1)$$

En la ecuación (1) ω es la frecuencia angular, $\mathbf{K} = -k\mathbf{k}$ es el número de onda de la onda incidiendo desde la dirección **k** en la cual el arreglo de micrófonos está enfocándose y $k = \omega/c$ es el número de onda libre. Implicitamente en la ecuación (1) existe un factor igual a $e^{i\omega t}$. El arreglo de micrófonos es enfocado en la dirección **k.** Teniendo en cuenta que generalmente hay ondas que llegan de otras direcciones, el

número de onda será distinto y denominado $\mathbf{K_o}$. De esta forma se investiga cómo influyen las ondas que llegan desde otras direcciones al arreglo de micrófonos. De acuerdo con la ecuación (1) la salida del beamforming será,

$$b(\mathbf{k},\omega) = P_o \sum_{m=1}^{M} e^{i(k-k_o).r_m} = P_o W(\mathbf{K}-\mathbf{K_o}) \tag{2}$$

La función W se denomina Array Patern. El Array Patern está definido enteramente por la distribución de micrófonos. La distribución de micrófonos se realiza en el plano x – y, por lo tanto la coordenada $z = 0$, esto implica que el Array Patern es independiente de $\mathbf{K_z}$. En la ecuación (2) se observa mayor sensibilidad del arreglo de micrófonos en la dirección $\mathbf{K}$ en el cual aparece el llamado Lóbulo Principal, en otras direcciones se presentan los llamados Lóbulos Laterales. La aparición de los Lóbulos Laterales influencia la medida del Lóbulo Principal produciendo falsas medidas en el mapa acústico. Una buena distribución de micrófonos se caracteriza por tener un bajo Máximo Lóbulo Lateral (MLL) relativo al Lóbulo Principal.

$$MLL(\mathbf{K}) = 10Log_{10}\left(\frac{\max\limits_{K_{min}^0<|\mathbf{K}|\leq K} |W(\mathbf{K})|^2}{M^2}\right) \tag{3}$$

2.2. TÉCNICA DE BEAMFOMING

La técnica de beamforming presupone un monopolo como fuente sonora emitiendo en cada punto de una grilla de escaneo espacial. Luego, asumiendo ondas esféricas, la presión acústica en un punto n de la grilla se puede calcular conociendo la ubicación de los micrófonos y la del punto a considerar, como se observa en la Figura 2. El retardo de fase en cada micrófono permite escribir la contribución de la presión acústica total del punto mediante la siguiente ecuación,

$$P_n = \frac{\vec{w}^{\dagger}\vec{p}}{M} \tag{4}$$

Donde $\vec{w}^{\dagger}$ representa el vector Hermitiano de pesos, este vector se utiliza para direccionar el arreglo de micrófonos hacia un punto de la grilla otorgando a cada señal un retraso de fase correcto, p_m es la presión en el dominio de la frecuencia del micrófono M. Asumiendo que para una fuente monopolo en campo libre las componentes del vector de propagación del arreglo están dadas por una función de Green definida como

$$C_M(x_n) = \frac{e^{-ikr_M}}{4\pi r_M} \tag{5}$$

En la cual r_M es la distancia euclidea desde el microfono M hasta el punto de la grilla n. Para maximizar la salida del beamforming, el vector de propagación es elegido paralelo al vector dirección.

Luego, la salida beamforming del arreglo de micrófonos es definido por

$$b_n(r,f) = \frac{w_n{}^{\dagger} CSM w_n}{M^2} \tag{6}$$

Donde CSM representa la matriz espectral cruzada (“Cross Spectral Matrix”). La matriz espectral cruzada contiene los aportes de las señales de cada uno de los micrófonos en el dominio de la frecuencia. La matriz se define de la siguiente manera

$$CSM_{ij} = \frac{p_i(f)p_j{}^*(f)}{2} \tag{7}$$

Notar que la matriz dada por la ecuación (7) contiene el espectro cruzado fuera de la diagonal principal. De acuerdo con la referencia [9] es beneficioso quitar el auto espectro de la matriz, por lo tanto los elementos de la diagonal principal son nulos. Luego, cada punto de la grilla es evaluado con la ecuación (6), si en el punto se encuentra una fuente emisora, las señales de los micrófonos funcionan aditivamente, en cambio si no se encuentra la fuente en ese punto las señales se suman destructivamente.

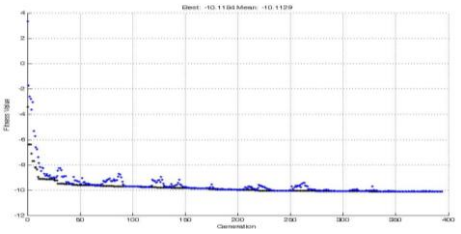
3. SIMULACIÓN DE CONFIGURACIONES DE MICRÓFONOS.

Cuando la distribución de micrófonos se vuelve regular, en el patrón de directividad de la antena comienzan a aparecer lóbulos laterales grandes, para poder reducirlos se debe optimizar la distribución de micrófonos de manera que el Máximo Lóbulo Lateral (MLL) sea mínimo.

Para realizar el proceso de optimización se decidió el uso de Algoritmos Genéticos (AG) ya que estos se adaptan muy bien a la resolución de problemas discontinuos, no diferenciables, estocásticos y altamente

no lineales. También al problema se le pueden agregar otros parámetros a optimizar como ser el costo de los micrófonos, la cantidad de micrófonos, etc. Los AG permiten resolver problemas con y sin restricciones a través de la selección natural. En nuestro caso la distribución de micrófonos se restringe a un área determinada y se castiga la superposición de los mismos sensores de presión. Los AG se diferencian de los algoritmos de optimización clásicos en dos puntos. Los AG generan una población de puntos en cada iteración, donde el mejor punto es el que aproxima a la solución óptima, mientras que los algoritmos clásicos generan un solo punto en cada iteración, en el cual la secuencia de puntos se aproxima a la solución óptima. Y como segunda diferencia los AG seleccionan la siguiente población de forma estocástica mientras que los algoritmos clásicos para seleccionar el siguiente punto en la secuencia lo hace de forma determinística.

(a)



(b)

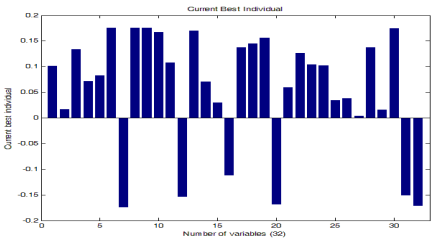


Figura 1: a) Se observa la cantidad de generaciones y b) los valores para cada variable.

La Figura 1 muestra la cantidad de generaciones utilizadas en la optimización para la función de fitness propuesta, la cual tiene como propósito la minimización de la ecuación (3) a partir de distintas distribuciones de los sensores de presión (variables). Los valores tomados por las variables que tiene el problema, son las coordenadas x – y de cada micrófono, en nuestro caso son 16 por lo tanto tenemos 32 coordenadas. Se utilizó una población inicial de 20 individuos y la tolerancia para el criterio de parada fue 1e-10.

Para corroborar los resultados de la optimización se procedió a simular un campo de presiones de acuerdo a la Figura 2 para luego comparar experimentalmente los resultados. El campo de presiones simulado se introdujo al programa que realiza beamforming. La fuente simulada consistió en un monopolo emitiendo a una frecuencia de 3 kHz en las coordenadas (0, 0, 1) y se agregó otro monopolo a la misma frecuencia pero desfasado $\pi/4$ (Figura 2), por lo tanto la señal será incoherente y no correlacionada.

Como se observa en la Figura 2 al romper la regularidad de la distribución de los micrófonos se logra disminuir los Lóbulos Laterales en gran medida y de esa forma evitamos obtener mediciones falsas. Las mediciones falsas consisten en fuentes que aparecen por la influencia de los Lóbulos Laterales pero que no son reales (fuentes fantasmas). La Figura 3 muestra las simulaciones que se realizaron para obtener el campo acústico anteriormente mencionado. Comparando experimentalmente los resultados obtenidos con la simulación se observa que la distribución de micrófonos obtenida mediante la optimización utilizando Algoritmos Genéticos es la que permite disminuir la influencia de los Lóbulos Laterales.

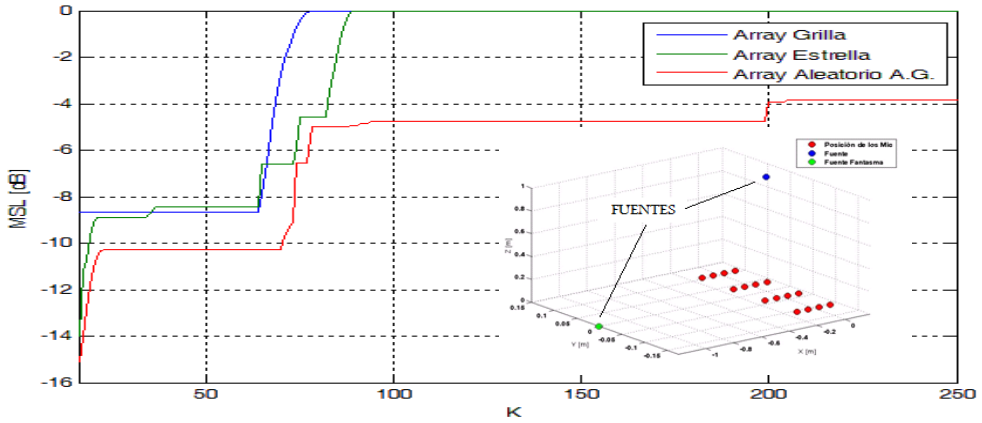


Figura 2: MLL en función del K (número de onda) para 3 tipos de distribución de micrófonos.

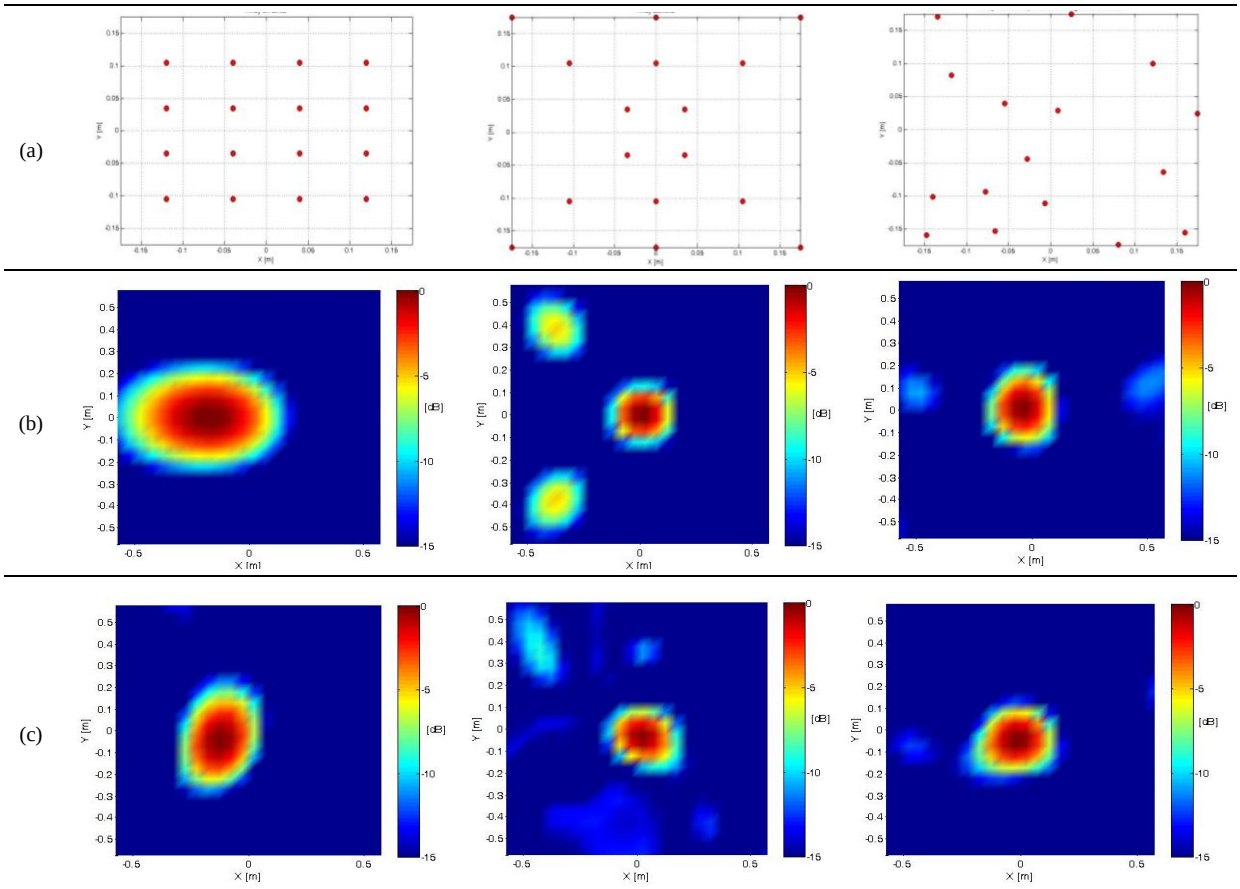


Figura 3: (a) Distintos arreglos de sensores y las comparaciones de las (b) simulaciones con los (c) resultados experimentales.

AGRADECIMIENTOS

Los autores agradecen la financiación aportada por la Agencia Nacional de Promoción Científica y Tecnológica a través del Fondo para la Investigación Científica y Tecnológica (FONCyT).

REFERENCIAS

[1] THOMAS J., MUELLER, CHRISTOPHER S., ALLEN, WILLIAM K., BLAKE, *Aeroacoustic Measurements*, SPRINGER VERLAG, ISBN-13: 978-3540417576, 2007.

[2] J.J. CHRISTENSEN AND J. HALD, *Beamforming, Brüel&Kjrer Sound&Vibration Measurement NS*, 2004.

[3] HÖGBOM, J. A., *A&A suppl.*, 15, 417, 1974.

[4] HENRY COX, ROBERTM ZESKIND AND MARKM OWEN, "*Robust adaptive beamforming"* IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. ASSP-35, pp. 1365-1376. Oct. 1987.

[5] THOMAS F. BROOKS, WILLIAM M. HUMPHREYS, JR. *"A Deconvolution Approach for the Mapping of Acoustic Sources (DAMAS) Determined from Phased Microphone Arrays",* 10th AIAA/CEAS Aeroacoustics Conference, Manchester, UK, May 10-12, 2004.

[6] PATRICIO A. RAVETTA, RICARDO A. BURDISSO AND WING F. NG, *"Noise Source Localization and Optimization of Phased Array Results (LORE)",* 12th AIAA/CEAS Aeroacoustics, Conference (27th AIAA Aeroacoustics Conference), Cambridge, Massachusetts, 8 - 10 May 2006.

[7] PIETER SIJTSMA, *"CLEAN Based on Spatial Source Coherence",* 13th AIAA/CEAS Aeroacoustics Conference (28th AIAA Aeroacoustics Conference), AIAA 2007-3436.

[8] J. HALD, J. J. CHRISTENSEN, *"A class of optimal broadband phased array geometries designed for easy construction",* The 2002 International Congress and Exposition on Noise Control Engineering Dearborn, MI, USA. August 19-21, 2002.

[9] G. ELIAS, *"Source Localization with a Two-dimensional Focused Array: Optimal Signal Processing for a Cross-shaped Array",* Proceedings of Inter-Noise 95, 1175-1178, 1995.

UN ALGORITMO PARA EL PROBLEMA *Cutting stock* EN DOS DIMENSIONES

Ignacio Ojea†

†*Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina, iojea@dm.uba.ar*

Resumen: El problema de *cutting-stock* en dos dimensiones, se presenta en diversas ramas de la industria. Se trata, esencialmente, de cortar una plancha de un determinado material (vidrio, madera, papel, acero, etc.) para obtener una cantidad de piezas rectangulares de menor tamaño que constituyen el pedido del cliente. La maquinaria con la que se realizan los cortes impone algunas restricciones. Presentamos aquí un algoritmo que fue probado para el caso de la madera en aglomerado, y que está basado en otro, propuesto por Andres Fritsch y Oliver Vornberger. Analizamos los resultados obtenidos y anticipamos posibles alternativas que están siendo investigadas.

Palabras clave: *Corte de Stock, matching, metarectángulo.*

1. INTRODUCCIÓN: EL PROBLEMA

El problema de Corte de Stock en dos dimensiones se presenta en distintas áreas de la industria, (maderera, metalúrgica, del vidrio, del papel, etc.). Puede describirse del siguiente modo: dada una demanda de piezas rectangulares $\{r_i\}$, $i = 1, \ldots, n$, que constituyen el pedido del cliente, y una serie idealmente infinita de placas también rectangulares R (todas iguales), se quiere cortar éstas últimas de manera de obtener como resultado las piezas $\{r_i\}$ más un cierto sobrante que se considera desperdicio, utilizando el menor número posible de placas. Cabe señalar que, dado que las piezas del pedido son un dato, una vez establecida la cantidad de placas que es necesario usar, el desperdicio queda determinado.

Puesto que tanto las piezas como las placas son rectangulares, pueden ser definidas por pares ordenados de la forma (ancho,largo): $r_i = (a_i, \ell_i)$, $R = (A, L)$. Eventualmente, las piezas pueden no tener una orientación fija, de modo que deberán ser consideradas las dos alternativas: $r_i = \{(a_i, \ell_i), (\ell_i, a_i)\}$. La principal restricción del problema es que todos los cortes deben ser rectos, paralelos a los bordes de la placa y de tipo *guillotina*, cortando la placa de lado a lado (Figura 1).

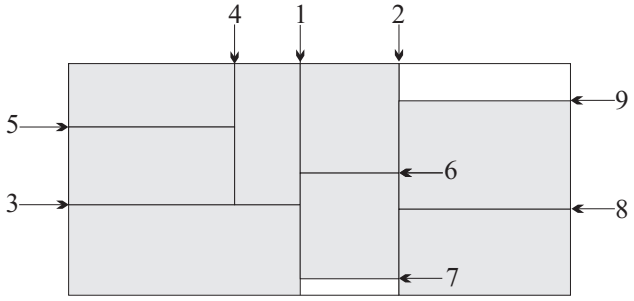


Figura 1: Los cortes deben ser de tipo *guillotina*

Este problema fue estudiado por Gilmore y Gomory, que lo abordaron usando programación lineal (1961, 1965). Hertz, en [2], mejoró las técnicas desarrolladas por ellos. Trabajos posteriores, de Whitlock y Christofides (1977), Cani (1979) y Coffman, Garey, Johnson y Trajan (1980), investigaron otros enfoques, proporcionando nuevas técnicas para su resolución. La mayor limitación de estos enfoques es que imponen fuertes restricciones al número de piezas que deben cortarse. Por otro lado, necesitan que la orientación de las piezas sea única, no admitiendo rotaciones. En el caso del algoritmo de Hertz se asume, además, que la cantidad de apariciones de una determinada pieza en el patrón de corte final no está acotada.

En este trabajo estudiaremos un algoritmo propuesto por Andreas Fritsch y Oliver Vornberger ([1]) basado en la aplicación de matching iterado; y presentaremos algunas modificaciones para su adaptación al caso particular del corte de placas de aglomerado. La principal ventaja de este algoritmo radica en el hecho de que no requiere que la orientación de las piezas sea fijada de antemano. Por otro lado, soporta pedidos de más de 100 piezas.

2. EL CASO PARTICULAR

Fritsch y Vornberger desarrolaron su método para la industria del vidrio. Nosotros estudiamos su utilización para el caso de la madera y, más particularmente, del aglomerado enchapado en melamina. Nuestro objetivo es desarrollar una aplicación rápida que pueda ser utilizada en pequeñas fábricas de muebles o carpinterías minoristas, para agilizar el diseño de los diagramas de corte. Si bien el problema, desde el punto de vista matemático, es esencialmente el mismo, el cambio de material implica ciertas variaciones que limitan o potencian la eficacia de uno u otro procedimiento. El algoritmo de Fritsch y Vornberger, tal cual es presentado en su trabajo, no dio resultados satisfactorios en el aglomerado. Fue necesario, por lo tanto, implementar modificaciones que lo adaptaran a las dificultades propias del caso.

En el caso del vidrio, por tratarse de un material isótropo, la orientación de las piezas es indistinta. Existe, además, una restricción importante: dado que las placas de vidrio son muy largas (típicamente 6 mts.) para evitar roturas los primeros cortes deben realizarse a lo ancho.

El aglomerado se comercializa en placas de $1,83 \times 2,6$ mts. enchapadas en melamina. Este enchapado puede ser de color liso o imitar la veta de la madera real. En el primer caso, la orientación de las piezas será indistinta. En el segundo, suele ser necesario fijar una orientación. Existen también problemas mixtos, en los que, para un mismo pedido, algunas piezas tienen una orientación fija y otras no. La mayor diferencia entre el vidrio y la madera se encuentra en el tamaño de las placas y, más precisamente, en la relación entre placas y piezas: la placa de vidrio es mucho mayor que la de madera, mientras que las piezas, si bien de tamaños variados, son en general similares en ambos casos. Esta pequeña diferencia será, sin embargo, crucial.

3. EL MÉTODO ORIGINAL

La idea básica del algoritmo consiste en agrupar las piezas por matching iterado. Para ello, la unión de dos piezas es considerada como un nuevo rectángulo, con cierto desperdicio, (Figura 2).

Figura 2: Unión de dos piezas

Sin embargo, dos piezas pueden unirse de diversos modos. En la Figura 3 se muestran todas las posibles disposiciones, admitiendo rotaciones. Llamaremos *instrucción de corte* a cada una de estas alternativas, y *metarectángulo* al conjunto de todas ellas. El algoritmo manipulará metarectángulos, de modo que en todo momento tendrá en cuenta *todas* las posibles disposiciones de las piezas agrupadas. Observemos que si se admiten rotaciones, una misma pieza es un metarectángulo con dos instrucciones de corte. Llamaremos *transversal* a una instrucción de corte cuyo ancho sea aproximadamente el de la placa y cuyo largo no exceda el tercio del de ésta (para garantizar que los primeros cortes sean a lo ancho). El matching iterado se utiliza para agrupar piezas hasta formar transversales.

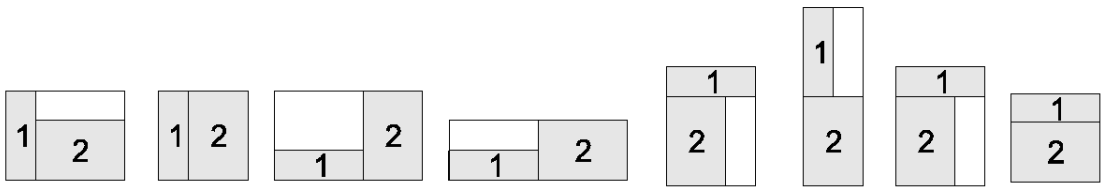


Figura 3: Metarectángulo: todas las posibles formas de unir dos piezas.

En cada iteración se construye un grafo en el que cada nodo representa a uno de los metarectángulos disponibles. Cada rama representa la posibilidad de unir los nodos adyacentes, y tiene asignado un peso cuyo valor es el desperdicio promedio de todas las instrucciones de corte del metarectángulo unión. Sobre este grafo se encuentra un matching de mínimo peso mediante el algoritmo de Edmonds (ver [3]). De este

modo se realizaran las uniones que impliquen un menor desperdicio promedio en la próxima iteración. Los metarectángulos que contengan instrucciones de corte transversales, serán almacenados pero no participarán de los siguientes apareamientos. El algoritmo de Edmonds tiene una complejidad del orden de n^3, siendo n el número de nodos del grafo. El crecimiento exponencial de los metarectángulos garantiza que el número de iteraciones sea bajo.

Los conglomerados de piezas así formados pueden ser interpretados como árboles binarios con raíz, en los que las hojas son las piezas. Esta situación se ilustra en la Figura 4.

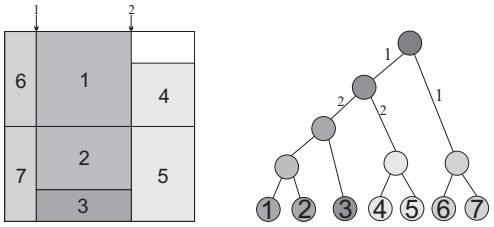


Figura 4: Árbol de corte.

Dado que el matching iterado es una estrategia ambiciosa, que no garantiza la minización del desperdicio al final del proceso, sino sólo dentro de cada iteración, los transversales deben ser sometidos a una etapa de refinamiento, a través de la cual se procura disminuir el desperdicio. Este refinamiento consiste en tomar dos transversales y estudiar permutaciones de sus sub-árboles que impliquen una reducción del desperdicio. Es importante señalar que no todos los bloques son transversales. Sin embargo, al llegar a este punto el algoritmo los trata como tales, asumiendo que el resto es desperdicio.

Una vez refinados, los bloques resultantes son depurados: se eliminan de los metarectángulos todas las instrucciones de corte que no son efectivamente transversales. Finalmente, los transversales son distribuidos sobre las placas. Dado que el ancho de aquellos se asume idéntico al de éstas, sólo debe tomarse en cuenta el largo. Este último paso es realizado por el método heurístico *First Fit Decreasing*.

El algoritmo original presentado por Fritsch y Vornberger responde, por lo tanto, al siguiente esquema:

1. Construcción de transversales por matching iterado.
2. Refinamiento de los transversales por permutación de sub-árboles.
3. Distribución de los transversales en las placas por *First Fit Decreasing*.

4. MODIFICACIONES

Puesto que el algoritmo tal como está presentado en [1] no dio buenos resultados para el caso del aglomerado, exponemos aquí las modificaciones que implementamos para mejorar su rendimiento. Exploramos dos posibles alternativas:

La primera de ellas consistió en eliminar el concepto de transversal, que no es necesario en el caso de la madera. En esta variante, que llamamos *Directa*, el procedimiento de matching se reitera hasta que las agrupaciones de piezas completan el tamaño de una placa, en ancho y largo. Los resultados obtenidos por este camino fueron buenos en pedidos de pocas piezas, que ocupan una única placa, pero no fueron satisfactorios para pedidos de más de 25 o 30 piezas.

La segunda variante resultó más eficiente y procuró atacar directamente los inconvenientes particulares de la madera: el principal problema en el caso del aglomerado es que el matching iterado no basta, por sí solo, para construir transversales. Por lo general, para completar el ancho de la placa son necesarias 5 o 6 piezas. El matching iterado agrupa de a 2, de a 4 y luego de a 8. Cuando una unión sobrepasa el ancho posible es descartada, de modo que al cabo de la primera etapa se obtiene un gran número de bloques con pocas piezas cada uno y que no alcanzan a cubrir el ancho de la placa, dejando un gran espacio de desperdicio. Implementamos, por lo tanto, una serie de rutinas de permutación y reagrupamiento de sub-árboles con el objeto, no de disminuir el desperdicio, sino simplemente de formar transversales. Al cabo de este proceso se obtienen algunos transversales, pero también algunos pequeños bloques más pequeños, para los cuales es necesario aplicar una nueva etapa de matching iterado.

Una vez obtenidos los transversales se eliminan las instrucciones de corte superfluas y se los distribuye en la placas. Aquí fue necesario implementar una nueva modificación. En el aglomerado no puede imponerse como restricción que los transversales tengan a lo sumo un tercio del largo de la placa, puesto que a veces las mismas piezas son más largas que eso. Definimos, pues, los transversales, como agrupaciones que ocupan todo el ancho de la placa y menos del $60\,\%$ de su largo. El método *First Fit Decreasing* arrojó muy malos resultados al distribuir este tipo de transversales. Implementamos, por lo tanto, un conjunto de rutinas heurísticas, variantes del *First Fit Decreasing*. El algoritmo las evalúa todas y toma el mejor resultado.

El esquema de nuestro algoritmo es el siguiente:

1. Construcción de bloques por matching iterado.
2. Refinamiento de los bloques para obtener transversales.
3. Reagrupamiento de los bloques pequeños, por marching iterado.
4. Refinamiento de los transversales por permutación de sub-árboles.
5. Distribución de los transversales en las placas por diversas heurísticas.

5. Resultados obtenidos y trabajo futuro

Los resultados obtenidos con el método por Transversales modificado fueron muy buenos. Se hicieron pruebas con varios ejemplos reales, obteniéndose en todos los casos el número óptimo de placas. Se evaluaron instancias de hasta 250 piezas, sin problemas de memoria, y en pocos segundos. La Tabla 5 muestra los resultados de distintos ejemplos. Constan en ella el número de piezas, el desperdicio relativo (desperdicio en el interior de los bloques) y el desperdicio neto (el área de las placas menos el área de las piezas).

Cant.Piezas	Placas	Desp. Rel.	Desp. Neto
19	1	$9,1\,\%$	$21,25\,\%$
35	4	$4,24\,\%$	$21,66\,\%$
56	3	$5,19\,\%$	$9,61\,\%$
100	4	$4,70\,\%$	$15,13\,\%$
250	10	$2,17\,\%$	$8,98\,\%$

El método Directo arrojó menores desperdicios relativos, pero requirió, en general, más placas, aumentando considerablemente el desperdicio neto. El objetivo principal del algoritmo es minimizar el número de placas. Sin embargo, un objetivo secundario es el de obtener desperdicios compactos (es decir: dismiuir el desperdicio relativo), que puedan ser reutilizados, o comercializados como piezas sueltas. Una posible continuación de este trabajo, por lo tanto, es intentar conjugar las virtudes de ambos métodos para minimizar el número de placas y también el desperdicio relativo.

Por otra parte, nuestra implementación, provisoria, admite múltiples mejoras: hay algunas estructuras y rutinas que, al ser optimizadas, permitirían aumentar y mejorar los procedimientos de refinamiento. Otro punto para mejorar es el del criterio con el cual se refina. Nuestra implementación se limita a probar permutaciones, realizando inmediatamente aquellas que disminuyen el desperdicio, sin importar si hay alguna otra cuyo resultado sea mejor. Ningún criterio de decisión garantizará, en este caso, un resultado óptimo. Estamos estudiando, sin embargo, la posibilidad de establecer algún mecanismo heurístico o incluso aleatorio inspirado en las técnicas de recocido simulado.

Referencias

[1] Fritsch, Andreas y Vornberger, Oliver; **Cutting stock problem by iterated matching**; `http://www.inf.uos.de/papers_html/or_94.`

[2] Hertz, J.C.; **Recursive Computational Procedure for Two-dimensional Stock Cutting**; IBM J. Res. Develop.; Sept. 1972.

[3] Papadimitrou, C.H. y Steiglitz, K.; **Combinatorial Optimization - Algorithms and Complexity**; Dover; 1998.

MODELAMIENTO NO PARAMÉTRICO DE DATOS GNSS PARA IMPLEMENTAR UN SIG EN 4D DESTINADO A LA ADMINISTRACIÓN VIAL DE COLOMBIA.

Saúl Becerra Ospina, Hernán Estrada B y Jorge M. Ruíz V

Departamento de Matemáticas, Universidad Nacional de Colombia Bogotá D.C. Colombia, www.unal.edu.co

Resumen: El Instituto Nacional de Vías, INVIAS y el Instituto Geográfico o Agustín Codazzi, IGAC desarrollaron un Sistema de Información Geográfico destinado a la administración vial de Colombia. La información cartográfica de las vías, se levanto con una campaña GNSS. En este trabajo se presenta la metodología para el modelamiento de datos geodésicos y ajuste de trayectorias observadas de las vías de primer orden de Colombia realizado para implementar adecuadamente un Sistema Lineal de Referencia, SLR.

Palabras clave: *GNSS, SLR.*
2000 AMS Subject Classification: 62G08

1. INTRODUCCIÓN

En la administración de infraestructura de transporte, es útil contar con información cartográfica tanto de las vías como de los elementos viales dispuestos para facilitar el flujo vehicular [1]. En el marco de un Sistema de Información Geográfica (SIG), es posible manejar información espacial georreferenciada en tres dimensiones. Las vías y los elementos lineales son abstraídas como un conjunto finito de puntos que definen una curva en $\mathbb{R}^3$. Otros objetos como señalizaciones son considerados puntos.

Un gran reto es levantar la información cartográfica en tres dimensiones de todas las vías y elementos viales, teniendo en cuenta que, a través de técnicas convencionales como fotogrametría o interpretación de imágenes, solo es posible obtener información horizontal. Una mejor alternativa es utilizar Sistemas Globales de Navegación Satelital (GNSS), sin embargo para los elementos viales la cantidad de sitios a ocupar pueden superar facilmente los cien mil y considerando los tiempos de rastreo por elemento vial, el traslado de equipos y su respectiva disposición para una correcta observación, resulta muy costoso y por lo tanto no es factible.

En Colombia, esta implementado un Sistema de Referencia Lineal (SRL) que consta de postes de referencia (PR) dispuestos a lo largo de los tramos viales, a partir de los cuales es posible localizar puntos especificando el código de un PR y una distancia medida sobre la vía. Adoptar un SRL e implementarlo dentro de un SIG es ventajoso, ya que solo exige georreferenciar las vías y los PR's, los demás elementos pueden ser referidos a los PR's. Con esto se evita levantar con GNSS cada uno de los objetos que conforman la infraestructura vial.

Para georreferenciar las vías y los postes de referencia el Instituto Nacional de Vías (INVIAS) en convenio con el Instituto Geográfico Agustín Codazzi (IGAC), llevaron a cabo una campaña con (GNSS), en la cual se observaron las vías de primer orden en un levantamiento en modo cinemático y los PR's en un levantamiento en modo stop and go. Este tipo de levantamientos son de tercer orden y particularmente la obtencion de alturas tanto la exactitud como la presión están sensiblemente comprometidas.

Para localizar los objetos con una aceptable precisión, es necesario que los perfiles viales describan trayectorias suaves. No obstante, los datos obtenidos en campo no se comportan de esta manera, por lo tanto se requiere ajustar las series de datos que definen un tramo vial. Dadas las condiciones topográficas de Colombia, un modelo de ajuste paramétrico no es posible ya que no se puede asociar funciones conocidas con los perfiles viales.

En este trabajo se presenta la metodología adoptada para el modelamiento no paramétrico de las series de datos observadas por cada tramo vial. Dicha metodología se aplicó en la implementación de un SIG para el INVIAS.

[1] Por ejemplo, puentes, señales verticales y horizontales, bermas, bordillos.

2. Sistemas Lineales de Referencia

Los Sistemas Lineales de Referencia SLR permiten localizar objetos sobre las vías. Para implementar un SLR se materializan postes de referencia (PR) a partir de los cuales se toman distancias hasta los objetos a ubicar, tal como se ilustra en la Figura 1 (a). Para incluir un SLR en un SIG, es necesario alojar datos espaciales de las vías y los PR, esto quiere decir, que debemos disponer de las coordenadas de puntos que pertenecen a una determinada vía.

En la práctica, mediante un levantamiento GNSS en modo cinemático, se registran coordenadas tridimensionales cada segundo de una antena desplazándose sobre las vías a una velocidad entre los 30 y 50 kilometros por hora. Producto del levantamiento, se obtienen coordenadas geográficas y la altura sobre el nivel del mar de puntos sobre la vía

$$P_i = (\mathbf{x}_i, h_i), \quad \text{con} \quad \mathbf{x} = (\varphi_i, \lambda_i), \text{ para } i = 1, 2, \ldots, n. \tag{1}$$

Si aproximamos la vía con un conjunto de puntos, las distancias sobre la vía entre dos puntos a y b, se estiman, aproximando la longitud de arco de la siguiente manera

$$d \approx \sum_{k=1}^{m} |P_k - P_{k-1}|, \text{ para } k = 1, 2, \ldots, m, \tag{2}$$

donde P_k son puntos sobre la vía, $P_0 = a$ y $P_m = b$. En este caso de estudio a, corresponde al PR y b es el respectivo objeto a referenciar, ver Figura 1 (a).

El problema surge porque un levantamiento GNSS cinemático es un proceso de medición, que por supuesto tiene incertidumbre, particularmente crítica en la determinación de alturas sobre el nivel del mar, donde el error con buenas condiciones de rastreo[2] puede ser ± 2 metros [3].

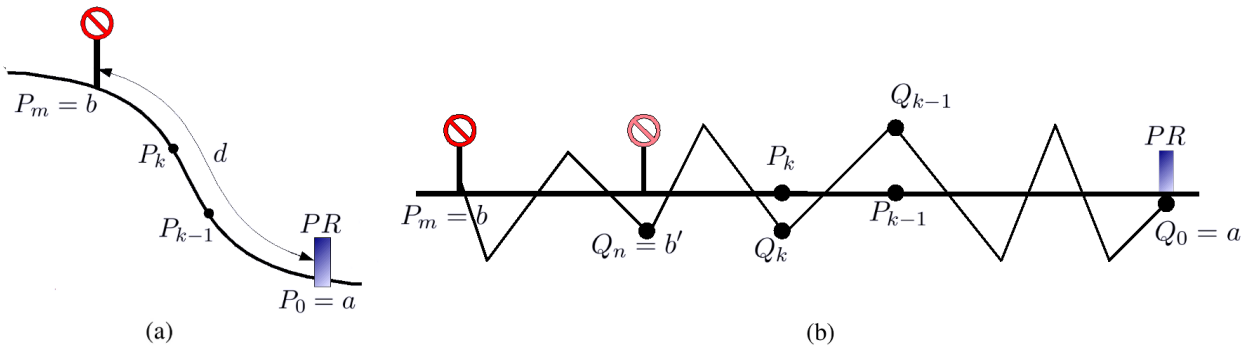


Figura 1: Esquema de un SLR.

De acuerdo con la discusión anterior, cuando se calcula la distancia entre cada par de puntos observados se introduce un error debido a la imprecisión para determinar alturas sobre el nivel del mar. Las distancias horizontales entre dichos puntos están entre los 8 y 13 metros, por lo tanto la máxima variación de altura en vías con pendientes considerables es menor a un metro. En términos generales, la variación de altura, denotada por Δh_i debe cumplir que $\Delta h_i \leq 100^{-1} I |P_i - P_{i-1}|$, siendo I la pendiente máxima de una vía en porcentaje. Por ejemplo, para una pendiente del $9\,\%$ el cambio de altitud esta alrededor de $0{,}72$ metros. Sin embargo, en un levantamiento GNSS, podríamos encontrar diferencias de hasta 4 metros y si no corregimos este error, la ubicación de los elementos viales no es correcta. Para explicar este hecho, en la Figura 1 (b).

[2] Más de cuatro satélites disponibles, sin obstáculos y con buenas condiciones meteorológicas.

[3] En posicionamiento GNSS, la precisión de coordenadas horizontales (φ, λ) es submétrica, $\pm 0{,}7$ metros y 2 o 3 veces más precisas que la altura h, por esto el error de h puede estar alrededor de ± 2

se esquematiza que

$$d = \sum_{k=1}^{m} |P_k - P_{k-1}| = \sum_{k=1}^{n} |Q_k - Q_{k-1}|, \tag{3}$$

donde Q_i son los puntos observados en campo con su respectivo error en altura y P_i representa puntos con las mismas coordenadas horizontales sin error de altura, es decir puntos sobre la vía. Adicionalmente, es sencillo ver que $|P_k - P_{k-1}| \leq |Q_k - Q_{k-1}|$, lo cual implica que $n \leq m$. El resultado es que el elemento se ubica en b' y no en b donde realmente se encuentra.

3. MODELAMIENTO DE PERFILES VIALES

Se requiere ajustar los datos de altura observados para obtener una curva suave que se aproxime a cada perfil vial. Los métodos paramétricos son muy limitados y además difíciles de aplicar para analizar las alturas observadas en la vías de Colombia, ya que sus condiciones topográficas son complejas y por lo tanto los perfiles no describen curvas de funciones conocidas, de modo que encontrar una función para cada vía y estimar sus parámetros resulta engorroso.

La altura topográfica observada con GNSS, se puede ver como una función de las coordenadas horizontales y se puede escribir como la suma de la altura topográfica más el error de observación [1]

$$h_{obs}(\mathbf{x}_i) = h(\mathbf{x}_i) + u_i \quad \text{con} \quad \mathbf{x}_i = (\varphi_i, \lambda_i). \tag{4}$$

La altura esperada es

$$h(\mathbf{x}) = E(h(\mathbf{x}_i)|\mathbf{x}_i = \mathbf{x}). \tag{5}$$

Los puntos observados describen cambios fuertes de altura que no reflejan el verdadero comportamiento de las vías. Para estimar una curva que se ajuste adecuadamente al perfil vial, se requiere un método que permita suavizar la trayectoria observada. *Smooth spline*[3] introduce un factor de penalización para la rugosidad [2], por lo tanto resulta ser una adecuada función de regresión. La rugosidad de una funcion f definida en un intervalo $[a, b]$ es

$$\int_a^b \{f^{(k)}(u)\}^2 \, du, \quad \text{para un} \quad k \geq 1, \tag{6}$$

de modo que un estimador de (5) es el siguiente criterio de mínimos cuadrados penalizado [2]

$$\sum_{i=0}^{n} u_i^2 + \lambda \int_a^b \{f^{(k)}(u)\}^2 \, du \tag{7}$$

donde $\lambda > 0$ es el parámetro de suavizado.

Finalmente, presentamos el ajuste realizado para una vía del Valle del Cauca. En la Figura 2 se gráfica un segmento desde el kilómetro 8 al 15. Con un rectángulo rojo, se indica un recuadro ampliado en la Figura 3. donde se aprecia una variación de altura Δh aproximadamente de 4 metros, encontrándose dentro de la precisión aceptada.

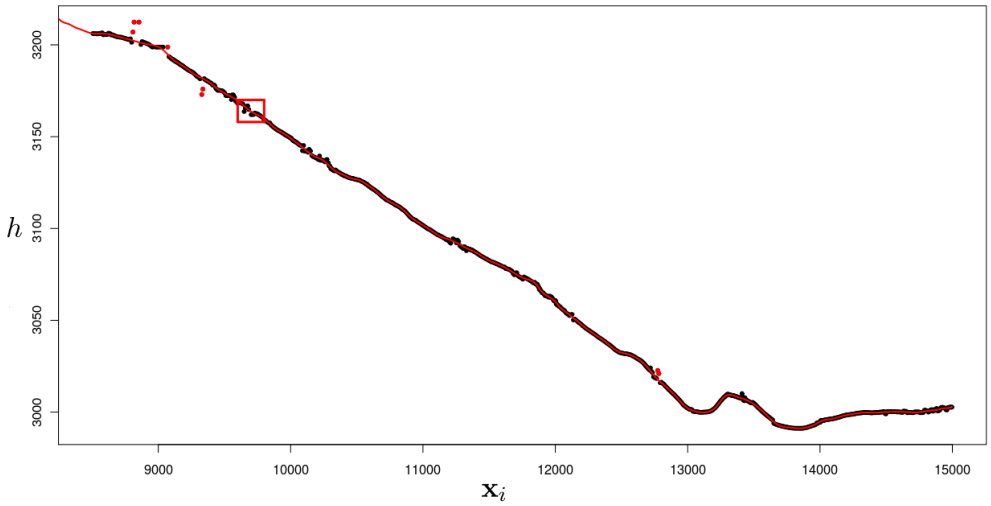


Figura 2: Ajuste de un segmento de un tramo vial del Valle del Cauca.

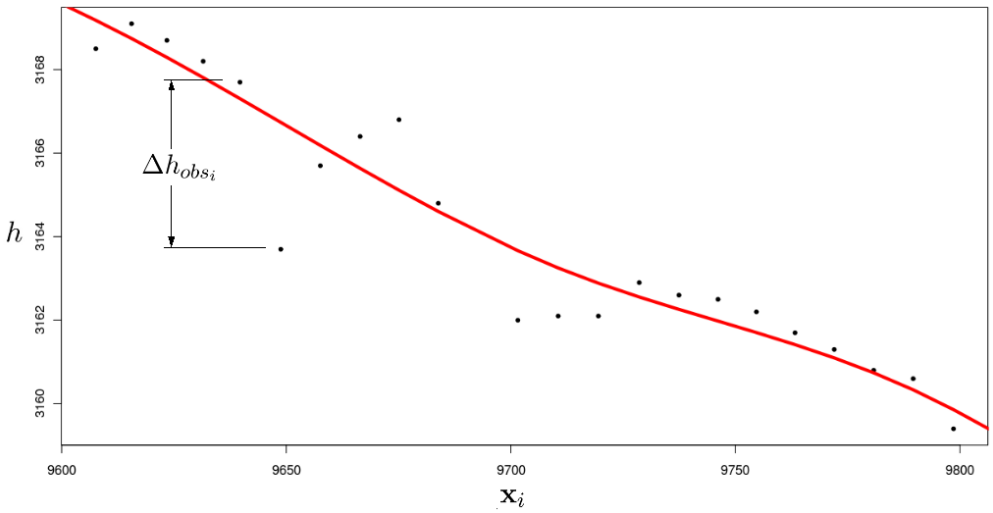


Figura 3: Perfil vial ajustado.

REFERENCIAS

[1] BORSA A., MINSTER J., BILLS, B. Y FRICKER, H., *Modeling long-period noise in kinematic GPS applications*, Springer-Verlag, J Geodesy, vol 81, pp 157–170, 2007

[2] HULIN WU, JIN-TING ZHANG, *Nonparametric Regression Methods for Longitudinal Data Analysis*, John Wiley & Sons, Inc., 2006.

[3] TAKEZAWA K., *Introduction to nonparametric regression.*, John Wiley & Sons, Inc., 2006.

UNA APLICACIÓN DE REDES NEURONALES ARTIFICIALES EN LA ESTIMACIÓN DE LA RESISTENCIA A LA PENETRACIÓN EN SUELOS

Nidia J. Valdés-Holguín†, Luis O. González-Salcedo†

†Grupo de Investigación en Materiales y Medio Ambiente, Universidad Nacional de Colombia Sede Palmira, Carrera 32 vía Candelaria, 1 Palmira, Colombia, logonzalezsa@unal.edu.co, www.palmira.unal.edu.co

Resumen: Las Redes Neuronales Artificiales, simuladoras del proceso de aprendizaje de las neuronas biológicas, han sido utilizadas con éxito en la estimación de parámetros, en diversos problemas de ingeniería donde sus variables involucradas tienen una alta relación entre sí no lineal y donde otras técnicas de modelación no son posibles de representar el problema mediante una ecuación. En el presente reporte se muestra la aplicación de una red neuronal para estimar la resistencia a la penetración a diferentes profundidades de un suelo. Los resultados muestran una mejor estimación para profundidades de 20 a 30 centímetros, considerando como variables de entrada la humedad y densidad del suelo, carga estática y presión de inflado.

Palabras claves: *Inteligencia Artificial, Redes Neuronales Artificiales, Suelos*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

La compresión del suelo consiste en la disminución de su volumen por aplicación de carga alta, y se denomina consolidación cuando el proceso ocurre en suelos saturados con exclusión del agua, y compactación cuando ocurre en suelos nos saturados con exclusión del aire [Bradford & Gupta 1986; Pinzón & Amézquita 1989]; la compactación causa cambios en el contenido de humedad y en el intercambio de gases entre el suelo y la atmósfera, e impide el desarrollo de las raíces [Foloni et al. 2003], y en el horizonte superior del suelo ocasiona cambios físicos y pedogenéticos que son de interés determinar [Pinzón & Amézquita 1991], como se muestra en la figura 1. La compactación en los suelos es un problema con efectos negativos desde el punto de vista económico y ecológico, aumenta la resistencia a la penetración de las raíces disminuyendo la capacidad de absorción radicular, se reduce el número y tamaño de los macroporos con lo cual se dificulta la aireación del suelo y la infiltración y provocando fenómenos de erosión por escorrentía [Colmer 2003]. Estudios relacionan la compactación de los suelos con problemas ambientales globales, y se afirma que contribuye al calentamiento global al disminuir el efecto sumidero del suelo y reducir la concentración de gases como el CO_2, CH_4, y N_2H [Horn et al. 1995].

Figura 1: Efectos negativos de la compactación en los suelos en la limitación de la zona de desarrollo de las raíces de las plantas [Alliaume & Hill 2008].

A la compactación por tránsito se le ha asociado un serio deterioro de la estructura de los suelos siendo uno de los problemas presentados la pérdida de su porosidad conllevando a la impedancia mecánica generada por dicho tránsito [Gerster & Bacigaluppo 2004]; como alternativas para reducir la impedancia se ha sugerido la descompactación mecánica, sin embargo los resultados de estas experiencias han sido contradictorios [Balbuena 2009]. En estudios realizados se ha sugerido la resistencia mecánica a la penetración como un indicador para determinar el grado de impedancias físicas en el suelo, en razón a que muestran a partir de un valor determinado, una disminución en los rendimientos de cultivo [Díaz-Zorita

2004]. En consecuencia, la resistencia mecánica a la penetración puede ser un indicador sensible para estudiar los efectos de la descompactación mecánica y la secuencia de cultivos sobre el rendimiento del cultivo [Gerster et al. 2010].

La amplia utilización de la resistencia mecánica a la penetración en suelos como un identificador y caracterizador de capas densificadas por efectos del laboreo, ha conllevado a que sus resultados se correlacionen con el crecimiento de las raíces y la productividad de los cultivos [Ehlers et al. 1983], el contenido hídrico por horizontes [Cerana et al. 2005], la humedad del suelo y la densidad aparente como indicador de calidad del suelo [Díaz et al. 2010]. Sin embargo, aun no han sido exploradas otras variables que se involucran en la caracterización del suelo cuando es medida su resistencia a la penetración, y no todas las variables son relacionadas en un solo modelo matemático que permita mostrar una relación compleja entre ellas; una causa de este vacío es la alta relación y dinámica de las variables en los problemas relacionados de estimación en la Ciencia del Suelo [Valdés 2010].

En el campo de la modelación, las redes neuronales artificiales son modelos de caja negra, desarrollados para resolver problemas en los que las relaciones de los diferentes componentes son muy complejas, las variables o reglas no son fáciles de obtener, donde hay escaso conocimiento, pero existe la experiencia de una serie de datos [López & Caicedo 2006]. Las redes neuronales artificiales, son una similitud del funcionamiento de una neurona biológica; y desde este punto de vista funcional, constituyen procesadores de información, con un canal de entrada de información y un canal de salida, con alta capacidad de comunicarse y unirse entre sí, cuya unión es denominada también sinapsis [Isasi 2007]. Las redes neuronales artificiales han sido utilizadas en la estimación de parámetros de diversos problemas de la Ciencia del Suelo [Buendía et al. 2002; Maneta & Schnabel 2003; Mena & Montecinos 2006; Bocco et al. 2007], siendo de interés estimar la resistencia a la penetración a partir de una red neuronal artificial. En el presente documento se muestra la elaboración de redes neuronales artificiales para estimar la resistencia mecánica a la penetración de un suelo.

2. MATERIALES Y MÉTODOS

2.1. BASE DE DATOS

Para el entrenamiento y validación de la red neuronal artificial, se elaboró una base de datos conformada por un conjunto de 192 vectores de información, disponibles en reportes de ensayos de resistencia mecánica a la penetración, y cuya metodología está basada en el ensayo estandarizado del índice de cono, usando el penetrómetro de impacto, medido a profundidades de 0-10 cm, 10-20 cm, 20-30 cm, y 30-40 cm, en el cual se mide la profundidad por golpe y número de golpes, que posteriormente son convertidos en unidades de presión, KPa [Valdés 2010]; los reportes disponen de las siguientes variables: contenido de humedad (H, %), masa de suelo en estado húmedo (MSH, g), masa de suelo en estado seco (MSS, g), densidad aparente (Da, g/cm^3), carga estática aplicada al suelo (Ce, KN), presión de inflado de la llanta (Prin, psi), porosidad (Poros, %), Relación de vacíos (Rel_Vacios, %) y los índices de cono para las resistencias a la penetración mencionadas (IC-H_0, IC-H_{10}, IC-H_{20}, y IC-H_{30}, medidos en KPa).

2.2. REDES NEURONALES ARTIFICIALES

Se elaboraron seis redes neuronales artificiales para estimar las resistencias a la penetración IC-H_{10}, IC-H_{20}, y IC-H_{30}. La tipología de las redes neuronales artificiales corresponden a una red multicapa, conformada por una capa de entrada, una capa oculta, y una capa de salida, y una sola neurona de salida conformó la capa de salida, correspondiente al valor de la resistencia a la penetración en una profundidad de referencia (IC-H_{10}, IC-H_{20}, y IC-H_{30}). Todas las redes usaron una metodología de aprendizaje de propagación hacia atrás, *backpropagation*. En cuanto a la división de los datos en Entrenamiento y Validación, se utilizó la técnica de *K-Fold Cross Validation*, con k=3. En este caso se divide el conjunto en tres partes iguales al azar, y se usan alternativamente dos como Conjunto de Entrenamiento y el tercero como Validación. Tiene la ventaja de que elimina cualquier sesgo de esta elección, y da una mejor idea de los errores de Validación y de la existencia de *Outliers*. Como lenguaje de programación, para la elaboración del algoritmo de la red neuronal artificial, se usó MATLAB®.

3. RESULTADOS Y ANÁLISIS

Los resultados de estimación muestran que el mejor resultado se presenta cuando se estima la resistencia a la penetración a una profundidad de 20-30 cm, IC-H_{20}, cuyas características específicas de la red son las siguientes:

- Tipo: *MultiLayer Perceptrón*.
- Arquitectura: 1 capa oculta, 40 neuronas.
- Algoritmo de Entrenamiento: *Bayesian Regulation Backpropagation*.
- Variables de Entrada: H, MSH, MSS, Da, Ce, Prin, IC-H_0, IC-H_{10}
- Variable de Salida: IC-H_{20}

La figura 2 muestra la validación de la red con la mejor estimación. El error relativo de validación obtenido para cada una de las 3 redes neuronales utilizadas es e1=10.7389%, e2=13.4008%, y e3=11.9886% en la estimación de la resistencia a la penetración del suelo.

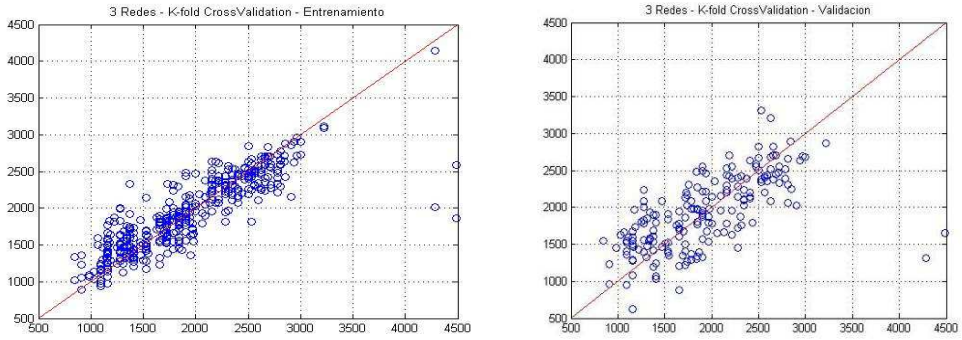


Figura 2: Validación de la red con la mejor estimación correspondiente a la resistencia a la penetración a una profundidad de 20-30 cm, IC-H_{20}.

4. CONCLUSIONES Y RECOMENDACIONES

Se observa en la red encontrada la existencia de un pequeño número de datos *outliers*, y un error de entrenamiento y validación relativamente alto, pero razonable dado los altos errores a considerar en las variables de entrada y la poca cantidad de datos disponibles. Es importante notar que esto es efecto de la técnica de CrossValidation utilizada que da una idea cabal de los errores involucrados, ya que normalmente queda enmascarado por la división de datos y otros efectos (Varmuza & Filmzoser 2009). Este error abre además una agenda futura en este campo, proponiéndose la elaboración de nuevas redes mediante una clasificación de los vectores de información componentes de la base de datos, y la confección una base de datos mas amplia, involucrando más variables y más datos por cada variable, lo cual va a redundar claramente en un mejor error de predicción.

A pesar del error obtenido en la estimación, es importante notar que la mejor estimación se presenta para una de las mayores profundidades (20-30 cm), donde se ha tomado en adición como variables de entrada la resistencia a la penetración a profundidades menores (0-10 cm y 10-20 cm), puesto que en el estudio de la resistencia a la penetración, sólo será necesaria la realización del ensayo a menores profundidades, permitiendo un ahorro significativo del recurso tiempo y costo.

La realización del presente trabajo muestra que mediante nuevas redes neuronales artificiales con mejores estimaciones, se aporta una contribución importante en la aplicación investigativa y profesional en la Ciencia del Suelo, en consideración a lo siguiente:

- La resistencia mecánica a la penetración, obtenida mediante el ensayo estandarizado del índice de cono es usado como un indicador confiable para la correlación de productividad o rendimiento del cultivo y el suelo; sin embargo, debe realizarse con una amplia cobertura en las zonas de laboreo y a cuatro diferentes profundidades. La disponibilidad de una herramienta de estimación, permite un

ahorro importante en la utilización del recurso tiempo y en el costo de la realización de los ensayos.

- La profundidad de arraigamiento es definida como el espesor de la zona más apta para el desarrollo de raíces [Alliaume & Hill 2008], clasificándose en tres grupos: superficial (0-15 cm), media (15-30 cm), y profunda (>30 cm). Es significativo entonces, que la herramienta pueda reportar estimaciones confiables de la resistencia a la penetración mecánica a mayores profundidades, puesto que permite conocer las condiciones del suelo para un desarrollo adecuado en aquellas especies vegetales de importancia económica que requieran estratos por encima de los 30 cm para su crecimiento radicular.

REFERENCIAS

[1] F. ALLIAUME, AND M. HILL, *Propiedades físicas ¿en qué influyen?: Dinámica del aire, dinámica del agua, crecimiento radicular*, Presentación, Curso de Física de Suelos, Universidad Nacional de Colombia Sede Palmira, Palmira – Colombia, 2008.

[2] R. BALBUENA, *Alternativas para la descompactación mecánica de los suelos*, Actas III Taller Física de Suelos, Rio Cuarto – Argentina, 2009.

[3] M. BOCCO, G. OBANDO, S. SAYAGO, AND E. WILLINGTON, *Neural Network models for land cover classification from satellite images*, Agricultura Técnica (Chile), Vol. 67, 4 (2007), pp.414-421.

[4] J.M. BRADFORD, AND S.C. GUPTA, *Soil compressibility*, Methods of soil analysis, Madison, pp.479-492, 1986.

[5] E. BUENDÍA, E. VARGAS, A. LEYVA, AND S. TERRAZAS, *Aplicación de redes neuronales artificiales y técnicas SIG para la predicción de coberturas forestales*, Revista Chapingo Serie Ciencias Forestales y del Medio Ambiente, Año/Vol. 8, 1 (2002), pp.31-37.

[6] J. CERANA, M. WILSON, O. POZZOLO, J.J. DE BATTISTA, S. RIVAROLA, Y E. DÍAZ, *Relaciones matemáticas entre la resistencia mecánica a la penetración y el contenido hídrico en un vertisol*, Estudios de la Zona no Saturada del Suelo, Vol. 7 (2005), pp.159-163.

[7] T.D. COLMER, *Long-distance transport of gases in plants: a perspective on internal aeration and radial oxygen loss form roots*, Plant, Cell and Environment, 26 (2003), pp.17-36.

[8] C.G. DÍAZ, R. OSINAGA, Y J. ARZENO, *Resistencia a la penetración, humedad del suelo y densidad aparente como indicadores de calidad de suelos en parcelas de largo plazo*, XXII Congreso Argentino de la Ciencia del Suelo, Rosario – Argentina, 2010.

[9] M. DÍAZ-ZORITA, *Effects of subsurface soil compaction of a Typic Hapludol on sunflower (Helianthus annus L.) production*, Ciencia del Suelo, 22 (2004), pp.40-43.

[10] W. EHLERS, U. KÖPKE, F. HESSE, AND W. BOHM, *Penetration resistance and root growth of oats in tilled and untilled loess soil*, Soil & Tillage Research, 3 (1983), pp.261-275.

[11] J. FOLONI, J. CALONEGO, E S. DE LIMA, *Efeito da compactação do solo no desenvolvimento aéreo e radicular de cultivares de milho*, Pesquisa Agropecuária Brasileira, Vol. 38, 8 (2003), pp.947-953.

[12] G. GERSTER, Y S. BACIGALUPPO, *Consecuencias de la densificación por tránsito en Arguidoles del sur de Santa Fe*, Actas XIX Congreso Argentino de la Ciencia del Suelo, Paraná – Argentina, 2004.

[13] G. GERSTER, S. BACIGALUPPO, M. BODRERO, Y F. SALVAGIOTTI, *Secuencia de cultivos, descompactación mecánica y rendimiento de soja en un suelo degradado de la región pampeana*, Para mejorar la producción, 45 (2010), pp.59-61.

[14] R. HORN, H. DOMZAL, A. SLOWINSKA-JURKIEWICKZ, AND C. VAN OUWERKERK, *Soil compaction processes and their effects on the structure of arable soils and the environment*, Soil & Tillage Research, 35 (1995), pp.23-36.

[15] P. ISASI, *Redes de neuronas*, Universidad Carlos III de Madrid, 2007.

[16] J.A. LÓPEZ Y E. CAICEDO, *Una aproximación práctica a las redes neuronales artificiales*, Universidad del Valle, 2006.

[17] M. MANETA, Y S. SCHNABEL, *Aplicación de redes neuronales artificiales para determinar la distribución espacial de la humedad del suelo en una pequeña cuenca de drenaje: Estudios Preliminares*, Estudios de la Zona no Saturada del Suelo, Vol. VI (2003), pp.295-304.

[18] C. MENA, Y R. MONTECINOS, *Comparación de redes neuronales y regresión lineal para estimar productividad de sitio en plantaciones forestales, utilizando geomática*, Bosque, Vol. 27, 1 (2006), pp.35-43.

[19] A. PINZÓN, Y E. AMÉZQUITA, *Compactación de suelos por el pisoteo de animales en pastoreo en el piedemonte amazónico de Colombia*, Pasturas Tropicales, Vol. 13, 2 (1991), pp.21-26.

[20] N.J. VALDÉS, *Exploración y elaboración de una red neuronal artificial para determinar propiedades específicas en los suelos*, Tesina (Ingeniera Agrícola), Universidad Nacional de Colombia Sede Palmira, Palmira – Colombia, 2010.

[21] K. VARMUZA & P. FILMZOSER: *·Introduction to multivariate statistics in chemometrics*, CRC Press, 2009.

ANÁLISIS DEL RENDIMIENTO DE UN CÓDIGO COMPUTACIONAL QUE IMPLEMENTA EL MÉTODO DE RED DE VÓRTICES INESTACIONARIO Y NO LINEAL

Alejandro Llanos†, Luis Ceballos‡, Sergio Preidikman‡

†Magíster en Ciencias de la Ingeniería mención Modelación Matemática. Universidad de La Frontera, Av. Francisco Salazar 01145. Temuco, Chile, a.llanos01@ufromail.cl, http://www.ufro.cl
‡Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Ruta Nacional 36 km 601, 5800 Río Cuarto, Argentina, lceballos@ing.unrc.edu.ar, www.ing.unrc.edu.ar
‡Departamento de Estructuras, Facultad de C. E. F. y N., Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, spreidik@efn.uncor.edu, http://www.efn.uncor.edu

Resumen: En este trabajo se presenta un análisis del uso de memorias cache y las mejoras en el manejo de éstas, de un código computacional que implementa un modelo de red de vórtices inestacionario y no-lineal para estudiar el comportamiento aerodinámico de vehículos aéreos no-tripulados con configuraciones de alas unidas y de gran alargamiento. El análisis consiste en identificar aquellas rutinas que posean altos índices de flujo de información, instrucciones ejecutadas y tiempo de cómputo. Una vez identificadas estas rutinas, se realizan algunas medidas correctivas que logran una mejor gestión de la memoria lo que se traduce en importantes incrementos en el speedup. Para desarrollar este trabajo se utiliza el framework Valgrind, en particular las dos herramientas: (i) Cachegrind, la cual permite identificar el comportamiento del código con las memorias cache; y (ii) Callgrind, para obtener información sobre porcentaje de tiempo que ocupa cada rutina y la interacción entre ellas en cuanto a transferencia de información.

Palabras claves: *computación de alto desempeño, aerodinámica inestacionaria y no-lineal, método de red de vórtices*

1. INTRODUCCIÓN

Este trabajo está orientado al análisis de la performance de un código computacional denominado AeroVANT [1, 2, 3]. Esta aplicación implementa el método de red de vórtices inestacionario y no-lineal (NUVLM) y permite simular el comportamiento aerodinámico de vehículos aéreos no-tripulados o "Unmanned Aerial Vehicles" (UAVs) con configuraciones de alas unidas y de gran alargamiento. Este tipo de UAVs tienen aplicaciones en actividades civiles, científicas, comerciales y militares. Estas actividades son desarrolladas a grandes altitudes, en régimen de vuelo subsónico, por un tiempo muy prolongado, y a bajo costo.

En lo que respecta a los requerimientos computacionales, numerosas implementaciones "seriales" del NUVLM [2, 4] mostraron que se insumen cuantiosos recursos de cómputo. El uso de técnicas tales como recortar las estelas o aprovechar la simetría del problema permiten que las implementaciones computacionales del NUVLM puedan ejecutarse más rápido, a costo de restringir el tipo de problemas que pueden ser atacados. Por ello, uno de los caminos a seguir para lograr mayores velocidades de ejecución consiste en implementar técnicas de performance que permitan aprovechar de una manera más eficiente los recursos computacionales que se disponen al momento de realizar las ejecuciones del código.

En un trabajo previo [3] destinado a mejorar los tiempos de ejecución, se desarrolló una estrategia de paralelización explícita, cuya implementación computacional permitió obtener un speedup del orden de 3,5 o un incremento del 72 % en el rendimiento de la aplicación, para ejecuciones realizadas con un procesador Intel® QuadCore® de 2,4 GHz ocupando sus 4 threads.

En este artículo se presenta un análisis del rendimiento o performance de la versión secuencial de este código. El análisis es realizado por medio de las componentes Cachegrind y Callgrind de la herramienta de análisis Valgrind. Estas componentes, permiten identificar las rutinas que consumen el mayor tiempo durante la ejecución del programa, conocer la interacción entre ellas y verificar problemas de acceso a las memorias cache L1 y L2.

Valgrind (http://valgrind.org/) es un sistema de herramientas de análisis de performance, de licencia GNU, que ayudan a mejorar el rendimiento de las aplicaciones [5]. Valgrind simula la ejecución de la aplicación, compilada en base a sus archivos objetos, para un análisis detallado del código. Las aplicaciones ejecutadas bajo este framework, sufren un factor de retardo de 2 a 20 veces o más, dependiendo la herramienta que sea utilizada. Algunas plataformas soportadas por Valgrind son: x86 y AMD64, entre otras.

2. NUVLM Y AEROVANT

El NUVLM permite modelar correctamente no-linealidades aerodinámicas asociadas con grandes ángulos de ataque, deformaciones estáticas, y flujos dominados por vorticidad en los que el fenómeno conocido como *vortex bursting* no ocurre. El modelo predice correctamente la emisión de vorticidad desde un cuerpo (o varios), inmerso en el seno de un fluido, hacia el campo del flujo. Esta vorticidad es transportada por el flujo de aire desde el cuerpo hacia el fluido y forma así las estelas. La distribución de la vorticidad en las estelas y la forma de las mismas son, también, parte de la solución del problema. El NUVLM es un método confiable y muy buen predictor de cargas aerodinámicas altamente inestacionarias y no-lineales [6].

La aplicación AeroVANT consta de tres partes. La principal, es un código que implementa un modelo aerodinámico basado en el NUVLM. Otra parte corresponde a la implementación de un preprocesador capaz de generar automáticamente configuraciones de UAVs de alas unidas, y preparar la geometría de esas configuraciones para que sean tratadas por el código principal de la herramienta. La tercera parte es un código que permite postprocesar los resultados para que puedan, con la ayuda de un software de visualización, representarse gráficamente. En la referencia [3] puede consultarse el algoritmo implementado computacionalmente en la parte principal del código.

Esta aplicación permite: explorar distintas configuraciones y realizar estudios paramétricos del comportamiento aerodinámico de UAVs con alas unidas; calcular saltos de presión y visualizar el cambio temporal de su distribución sobre las superficies sustentadoras; calcular coeficientes adimensionales de sustentación y trazar curvas de su variación según cambia el ángulo de ataque geométrico; y calcular la evolución espacio temporal de estelas desprendidas desde las superficies sustentadoras del vehículo [1,2].

3. ANÁLISIS DE RENDIMIENTO

En esta sección se presentan algunas mediciones de performance hechas con las herramientas computacionales de Valgrind 3.5 anteriormente mencionadas, sobre la versión secuencial del algoritmo implementado en AeroVANT, escrito en Fortran 95 y utilizando el compilador GNU de Fortran, GFortran. Las ejecuciones han sido realizadas bajo el sistema operativo Ubuntu 9.04, en un equipo de escritorio con procesador de 64 bits, Intel® Core 2 Quad® Q6600 de 2.40 GHz. También se consideran 30 pasos de simulación y se compila el código sin ninguna de las opciones de optimización del compilador GFortran.

Estas herramientas permiten visualizar la estructura del código mediante un grafo en árbol, determinar la interacción existente entre las rutinas del código, medir el número de veces que las rutinas del programa son ejecutadas o llamadas, y medir el porcentaje de tiempo que ocupan las ejecuciones de las rutinas.

Al analizar los resultados, se detecta que el tiempo de ejecución está dominado por el conjunto de rutinas encargadas de realizar la convección de partículas de fluido, con un 98,3 % del total del tiempo de ejecución. Las rutinas que más tiempo insumen son aquellas encargadas de calcular, sobre cada partícula de fluido, las influencias aerodinámicas de las sábanas vorticosas adheridas y libres.

En un análisis al código del conjunto de rutinas, se observa que estas realizan las mismas operaciones, pero con diferente número de variables y de llamados; por lo que su trabajo como rutinas independientes es una ayuda a la eficiencia del código. También se logra apreciar que las operaciones más usadas por estas rutinas son la multiplicación, suma y resta de matrices y vectores del tipo *double* y tamaño 3x3 y 1x3, respectivamente. Este tamaño pequeño de los datos descarta la búsqueda de mejores algoritmos para dichas operaciones. Otra apreciación sobre estas rutinas se refiere a que cada una tiene asociada las tres siguientes funciones: producto vectorial entre arreglos, producto escalar entre arreglos y suma de las componentes de un vector. Este conjunto de funciones representa un total del 30,6% sobre el tiempo total de ejecución del código. Además, se aprecia que la cantidad de veces que estas funciones son ejecutadas varían entre los 90 y 1.100 millones de ejecuciones según la rutina padre que las ejecute. Esto, junto al tráfico de datos, como matrices o vectores, requeridos para ejecutar cada función, implica un alto costo computacional.

Las herramientas de Valgrind permiten obtener los contadores de performance del algoritmo del NUVLM, sobre las memorias cache L1 y L2, en base a tres tipos de interacciones: la captura de instrucciones "Instruction Fetch", la lectura de datos "Data Read Access" y la escritura de datos "Data Write Access".

Los resultados aportados por el uso de estas herramientas muestran que el conjunto de rutinas utilizadas en la convección de partículas de fluido realiza cerca el 99% del total de Instruction Fetch, el 98% del total de Data Read Access y el 99% del total de Data Write Access. Una vez más se muestra que aquí se concentra casi la totalidad del costo cómputo. También, se determina que la rutina que posee la mayor cantidad de fallas en los accesos a memorias L1 y L2, es la rutina encargada de resolver el sistema

de ecuaciones algebraicas lineales en el NUVLM. Esta rutina contiene el 58% de las pérdidas de acceso a la memoria L1 y el 92% a la L2. Como este proceso se realiza solo una vez en la ejecución del programa, se descartan labores para mejorar su eficiencia. En resumen, a partir de la información obtenida con las herramientas de performance, la magnitud de cache Misses en todo el programa dista de ser un punto de preocupación en la eficiencia del algoritmo, puesto que se encuentra bajo el 0,02% de las pérdidas.

4. RESULTADOS

En esta sección se presentan las propuestas para mejorar la eficiencia del código del NUVLM y como estas propuestas influyen en la gestión de la memoria cache y en el rendimiento, en general, del programa.

De la sección 3, se concluye que: i) las rutinas encargadas de las convección de partículas de fluido, consumen el mayor porcentaje de tiempo de ejecución del programa (98,3%). ii) El análisis de estas rutinas con Cachegrind descarta problemas de acceso a la memoria cache, pero iii) se logra detectar que éstas concentran el mayor flujo de datos del programa en sus subrutinas. Por lo que se proponen las siguientes medidas correctivas sobre las rutinas utilizadas en las operaciones de convección de partículas de fluido:

A) Se quita el llamado a las funciones que realizan operaciones de productos y suma entre arreglos, y en su lugar se escriben sus códigos explícitamente como parte de las rutinas que hacían las llamadas. Este cambio, permite disminuir el flujo de datos, ahorrar el número de ciclos al fusionar rutinas similares, y por ende reducir el número de operaciones y referencias a memoria realizadas anteriormente.

B) Se comprueba el correcto acceso a memoria por parte de las operaciones matriciales realizadas; puesto que en Fortran la forma de acceso los datos difiere a la de los lenguajes C/C++. También se descarta introducir algoritmos que operen en forma de bloques sobre las operaciones matriciales, debido al pequeño tamaño de éstas y las bajas pérdidas de acceso a memoria L1 y L2 mostradas en la sección anterior.

C) Se logra un 5% de mejora al evitar divisiones explícitas de datos tipo double dentro del código.

D) Por último, solo se utiliza el nivel de optimización "-O2" del compilador.

Los resultados de las mejoras propuestas anteriormente son mostrados a continuación y se hace énfasis en el tiempo de ejecución logrado y la mejora de eficiencia en la interacción de las memorias cache. Las mediciones se realizan sobre la versión secuencial del código, y se toman los siguientes pasos de simulación: 30, 35, 40, 70, 100 y 150.

En la Figura 1, se muestra la influencia de las mejoras respecto al tiempo. Estas mediciones, no contemplan la opción compilación "-O2", puesto que la idea es ilustrar los resultados sobre las modificaciones realizadas explícitamente sobre el código del programa. Estos resultados muestran que la nueva versión tiene una mejora aproximada del 58% respecto a la anterior. Si incluimos a la opción de optimización "-O2" del compilador, esta mejora asciende a un 71%.

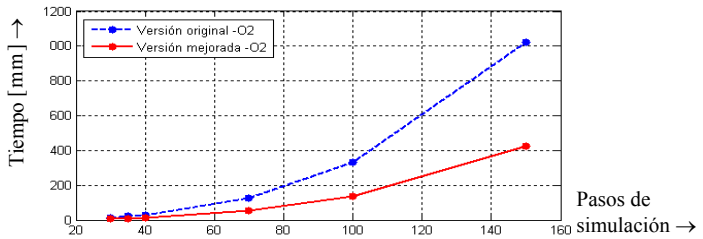


Figura 1: Comparativa de los tiempos de ejecución de ambas versiones secuenciales del código.

Otra mejora apreciada, es que el acceso a memoria cache por lectura y escritura disminuye en un 77% y 75% respectivamente. Esto, es reflejo de evitar la excesiva transferencia de datos al reescribir las funciones encargadas de los productos y suma entre arreglos.

Como se mencionó en la sección anterior, las rutinas de la convección no presentan un número importante de fallas de acceso a memoria L1 o L2, pero estos cambios realizados proporcionan un ahorro del 8,9% en el acceso a memoria L1 y de un 0,04% en el acceso a memoria L2.

La implementación de las mejoras sobre la versión paralelizada del código [3] muestra, de igual modo, un incremento en el tiempo de ejecución del código. En la Figura 2 se ilustra una comparativa del speedup obtenido, sobre el número de pasos de simulación. Estas pruebas fueron realizadas en un equipo QuadCore, utilizando sus 4 threads y para 30, 35, 40, 70, 100 y 150 pasos de simulación.

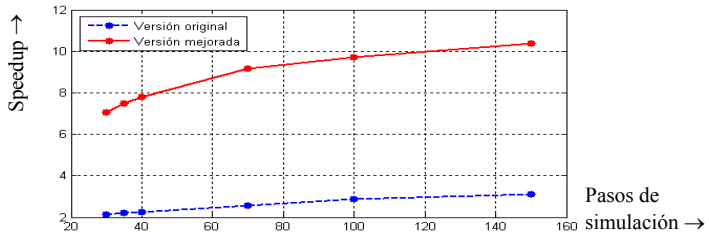


Figura 2: Comparativa del speedup, sobre ambas versiones paralelizadas del código

Los resultados muestran que el speedup alcanzado para un número de pasos igual a 150, es de 10,3 para la versión mejorada y de 3,1 para la versión sin modificaciones. Esto significa una mejora del 70%, aproximadamente, en el rendimiento de la aplicación.

5. CONCLUSIONES

En este trabajo se presentó un análisis de la performance de la herramienta computacional AeroVANT, la cual permite realizar simulaciones del comportamiento aerodinámico inestacionario y no-lineal de vehículos aéreos no-tripulados con configuraciones de alas unidas. En trabajos anteriores, investigadores de las Universidades Nacionales de Río Cuarto y Córdoba (Argentina), desarrollaron versiones secuenciales y paralelizadas de un código computacional robusto, cuyo núcleo principal implementa un modelo aerodinámico basado en el método de red de vórtices inestacionario y no-lineal.

El análisis de la performance del código AeroVANT se realizó con la asistencia del software Valgrind; un framework que posee herramientas de análisis dinámico y estático. El uso de estas herramientas permitió obtener una mejor comprensión del comportamiento del código durante su ejecución. Este análisis logró: i) identificar aquellas rutinas que consumen las mayores cantidades de tiempo de ejecución, ii) conocer la interacción de estas rutinas con otras, e identificar el número de llamados, y iii) observar una reducida cantidad de problemas en el acceso a las memorias cache L1 y L2.

El análisis del comportamiento del código, permitió delinear algunas acciones correctivas para modificar la versión original y lograr, así, un incremento en su rendimiento. Las acciones correctivas consistieron en: fusionar rutinas, para evitar la excesiva referencia a variables y aprovechar esta instancia para reorganizar bucles internos, con el afán de reducir comparaciones; evitar divisiones directas con la introducción de nuevas variables; y aprovechar la opción "-O2" de optimización del compilador.

La implementación de estas acciones correctivas, permitió obtener mejoras en la versión secuencial del código y en su versión paralelizada, alcanzando incrementos del 71% y del 70% respectivamente. Por último, la mejora en la versión secuencial del código pudo ser apreciada, no solo en el ahorro de tiempo, sino también en el comportamiento del código con los niveles de memoria L1 y L2.

REFERENCIAS

[1] L. CEBALLOS, S. PREIDIKMAN, J. MASSA, *Generador paramétrico de geometrías de UAVs de alas unidas orientado al método no-lineal e inestacionario de red de vórtices*, Mecánica Computacional, Vol. 27 (2008), pp. 2983-3007.

[2] L. CEBALLOS, S. PREIDIKMAN, J. MASSA, *Herramienta computacional para simular el comportamiento aerodinámico de vehículos aéreos no tripulados con una configuración de alas unidas*, Mecánica Computacional, Vol. 27 (2008), pp. 3169-3188.

[3] L. CEBALLOS, A. BARONE, A. FLORES, S. PREIDKMAN, *Desarrollo de una estrategia de paralelización explícita para el método de red de vórtices inestacionario y no lineal* en Actas del II Congreso Argentino de Ingeniería Mecánica, Universidad Nacional de San Juan, San Juan, Argentina, (2010).

[4] B. ROCCIA, S. PREIDIKMAN, J. MASSA, *Implementación de un modelo no-lineal e inestacionario para estudiar la aerodinámica de un micro-vehículo aéreo en "hover"* en Actas del II Congreso Argentino de Ingeniería Mecánica, Universidad Nacional de San Juan, San Juan, Argentina, (2010).

[5] N. NETHERCOTE Y J. SEWARD, *Valgrind: A program supervision framework*, Electronic Notes in Theoretical Computer Science 89 No. 2, (2003).

[6] S. PREIDIKMAN, *Numerical simulations of interactions among aerodynamics, structural dynamics, and control systems*, Ph.D. Dissertation, Department of Engineering Science and Mechanics. Virginia Polytechnic Institute and State University, Blacksburg, VA, (1998).

AVIONES NO-TRIPULADOS CON ALAS QUE MUTAN: ASPECTOS ESTRUCTURALES

Marcos L. Verstraete[1,2], Sergio Preidikman[1,2,3] y Julio C. Massa[1,3]

[1]*Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Campus Universitario, Ruta Nacional 36 Km. 601, 5800 Río Cuarto, Argentina. Tel/Fax: 0358-4676246, mverstraete@ing.unrc.edu.ar, http://www.ing.unrc.edu.ar*

[2]*CONICET – Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917, Buenos Aires, Argentina, spreidikman@efn.uncor.edu, www.conicet.gov.ar*

[3]*Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, jmassa@efn.uncor.edu, http://www.efn.uncor.edu*

Resumen: En este trabajo se presenta el desarrollo de un modelo estructural de alas flexibles para vehículos aéreos no tripulados (UAVs), inspirados en el vuelo de las aves, con alas que cambian de forma (*morphing-wings*). El modelo incorpora actuadores piezoeléctricos inmersos y distribuidos espacialmente en la estructura del ala del UAV. Estos actuadores inducen deformaciones mecánicas en la estructura del ala con el objetivo de cambiar su configuración, y así modificar las características aerodinámicas de acuerdo a los requerimientos de las diferentes fases del vuelo. Las ecuaciones de movimiento que describen la dinámica del modelo estructural de alas flexibles son discretizadas mediante el método de elementos finitos.

Palabras claves: UAVs, Morphing-wings, Alas flexibles, Actuadores piezoeléctricos,

1. INTRODUCCIÓN

Un tópico de investigación que ha acaparado la atención de muchos investigadores es el desarrollo de vehículos aéreos no tripulados (UAVs) con alas que cambian de forma (*morphing-wings*), inspirados en el vuelo de las aves. La adaptación de su sistema muscular y de sus huesos, permite a un ave deformar sus alas. Para realizar maniobras de vuelo, el sistema muscular cambia la posición relativa de los huesos que conforman el esqueleto [1] [2]. En las alas de un UAV, es posible reproducir este efecto mediante la implementación de un sistema de control, compuesto de sensores y actuadores inmersos de manera conveniente en la estructura elástica del ala. La elección del tipo de actuador para inducir deformaciones del ala flexible del UAV es, quizás, el aspecto más importante para hacer viable esta idea. Se requiere de actuadores con poco consumo de energía y poco peso, compactos, de alta eficiencia y larga vida útil. Recientemente, los actuadores piezoeléctricos han sido utilizados para inducir deformaciones controladas en superficies flexibles [3] debido a que cumplen con la mayoría de los requisitos mencionados anteriormente.

En esta instancia preliminar se presenta el desarrollo de un modelo estructural de alas flexibles con actuadores piezo-cerámicos del tipo PZT, inmersos y distribuidos espacialmente sobre la estructura del ala del UAV (Figura 2a) donde no se consideran ni el algoritmo de control, ni los sensores como parte del modelo. La herramienta en desarrollo ayudará a comprender el comportamiento de alas flexibles de UAVs sometidas a deformaciones mecánicas controladas.

Este trabajo es el inicio de un proyecto de mayor envergadura cuyo objetivo final es desarrollar herramientas numéricas que permitan investigar el comportamiento aeroservolástico de UAVs con alas que cambian dinámicamente de forma durante las diferentes fases del vuelo.

2. ACTUADORES PIEZOELÉCTRICOS

Los denominados *materiales piezoeléctricos* exhiben un comportamiento muy interesante. Cuando se someten a una deformación mecánica generan un campo eléctrico (*efecto directo*), y de manera inversa, al someterse a un campo eléctrico experimentan una deformación mecánica (*efecto inverso*). En estos materiales existen relaciones entre las variables eléctricas y las mecánicas que pueden ser descriptas a través de ecuaciones constitutivas lineales que pueden ser derivadas a partir de consideraciones energéticas [4]:

$$S = s^E T + d E \qquad D = d T + \varepsilon^T E \tag{1}$$

donde S es el tensor de deformación mecánica, s^E es el tensor de flexibilidad a campo eléctrico constante, T es

el tensor de tensión mecánica, *d* es el tensor de acoplamiento electromecánico que contiene los coeficientes de deformación piezoeléctrica, *E* es el campo eléctrico, *D* es el desplazamiento eléctrico y ε^T es el tensor de permitividad eléctrica a tensión mecánica constante.

En la Figura 1a se muestra una lámina rectangular de PZT (actuador) referida a un sistema de referencia de ejes cartesianos y ortogonales 1, 2 y 3. De acuerdo a la primera de las ecuaciones constitutivas, si se aplica un campo eléctrico, E_3, en la dirección 3, la lámina de PZT experimentará una deformación mecánica en la dirección 1. Es oportuno destacar que el campo eléctrico es el producto de la aplicación de un voltaje, *V*, sobre electrodos que están adheridos a las caras de la lámina. Si se adhieren perfectamente un par de láminas idénticas de PZT en las superficies superior e inferior de una viga (Figura 1b), y luego a cada una de estas láminas se les aplica un campo eléctrico de la misma intensidad pero de sentido opuesto, se causará un acortamiento en las fibras superiores de la viga y un alargamiento en las fibras inferiores de la misma, lo que resultará en una flexión pura del sistema. Para más información sobre actuadores piezoeléctricos, y la influencia que ellos ejercen sobre estructuras flexibles puede consultarse la referencia [5].

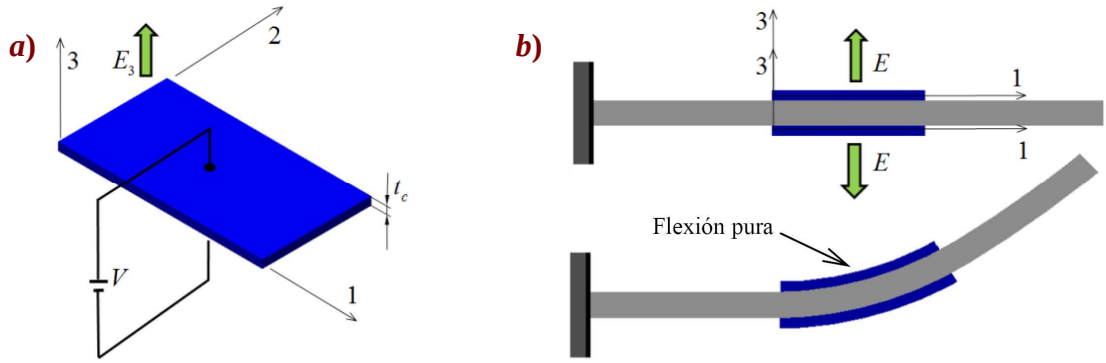


Figura 1: ***a)*** Esquema de un actuador PZT. ***b)*** Efecto de aplicar un par de actuadores sobre una viga.

3. MODELO DE ALA FLEXIBLE

Como se muestra en la Figura 2a, la estructura propuesta para el ala del UAV está compuesta por una viga principal que cubre, casi completamente, la envergadura del ala y por varias vigas secundarias en voladizo, orientadas a lo largo de la cuerda, y conectadas en uno de sus extremos a la viga principal. Notar que se considera únicamente una semiala del UAV empotrada en su raíz. Sobre la estructura del ala se distribuyen pares de actuadores piezo-cerámicos del tipo PZT con la finalidad de inducir deformación sobre dicha estructura, y así cambiar su configuración. El modelo de ala flexible con actuadores piezoeléctricos inmersos permite variar del ángulo de diedro del ala mediante la deflexión de la viga principal, y además, cambiar la combadura o ángulo de ataque mediante la deflexión de las vigas secundarias.

4. FORMULACIÓN DE ELEMENTOS FINITOS PARA EL MODELO DE ALAS FLEXIBLES

Las vigas que componen la estructura son divididas en un número finito de elementos conectados en los nudos. Un elemento puede estar compuesto únicamente por el material elástico, o bien tener un par de láminas de PZT (actuadores) adheridas al material elástico formando una estructura "sándwich" como se ve en la Figura 2b. Las propiedades del conjunto viga/actuadores se asumen constante a lo largo de cada elemento.

Se define un sistema de referencia (*sistema global*) fijo a la raíz del ala, cuyas coordenadas cartesianas ortogonales son x_g, y_g y z_g; y un sistema de referencia (*sistema local*) fijo a cada elemento cuyas coordenadas son x_l, y_l y z_l. En el modelo se consideran tres grados de libertad por nudo; (*i*) desplazamiento u_z en la dirección z_l, (*ii*) giro ϕ_x, en la dirección x_l (torsión), y (*iii*) giro ϕ_y en la dirección y_l (flexión). En la Figura 2b se muestran los grados de libertad asociados a los nodos '*i*' y '*j*' de un elemento típico. Las ecuaciones de movimiento se obtienen para cada uno de los elementos, en el sistema local, y luego son ensambladas, en el sistema global, para modelar la estructura viga-actuadores en su conjunto.

Considerando las características geométricas y las propiedades de los materiales correspondiente a los elementos, se determina la energía cinética y la energía elástica para cada elemento. Luego las expresiones para la energía cinética y elástica son reemplazadas en las ecuaciones de Lagrange para obtener la ecuación (2)

que gobierna la dinámica de un elemento de viga con actuadores.

$$\mathbf{M}_e\,\ddot{q}_e(t) + \mathbf{K}_e\,q_e(t) = \mathbf{Q}_e \tag{2}$$

donde $\mathbf{M}_e$ y $\mathbf{K}_e$ son respectivamente la matriz de masa consistente y la matriz de rigidez del elemento con actuadores. q_e es el vector que contiene los grados de libertad de los nudos que determinan el inicio y el fin del elemento. $\mathbf{Q}_e$ es el vector de cargas generalizadas que contiene el término de la llamada fuerza bloqueada, generada por la limitación en la deformación libre del actuador al someterlo a la acción de un campo eléctrico variable en el tiempo. Las matrices de rigidez y masa, y el vector de cargas del elemento viga-actuadores deben ensamblarse para obtener las ecuaciones de movimiento del conjunto completo.

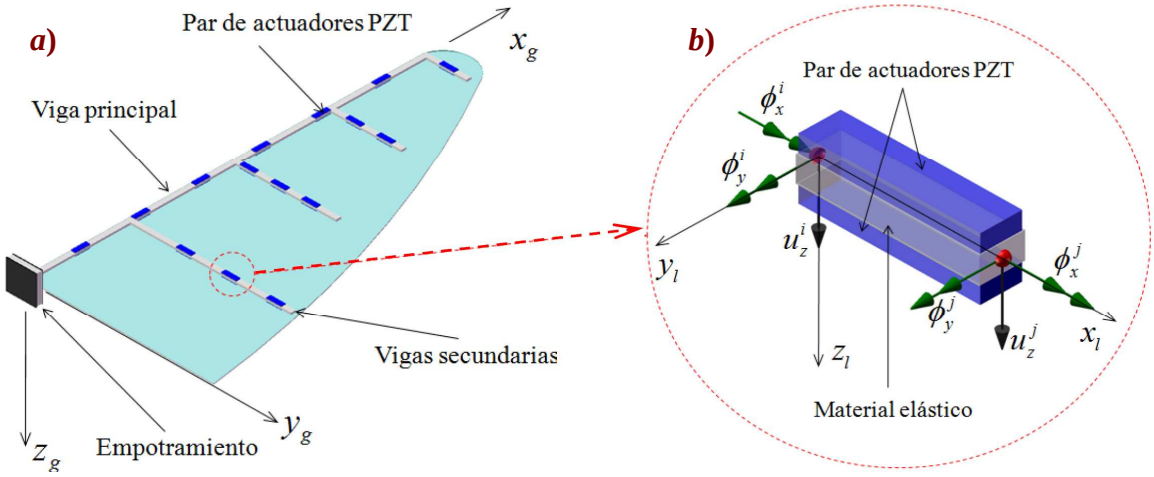


Figura 2: ***a)*** Modelo de ala flexible conteniendo actuadores PZT. ***b)*** Elemento típico producto de la discretización y grados de libertad de los nudos.

5. RESULTADOS

El modelo estructural de alas flexibles desarrollado, permite estudiar el comportamiento estático y dinámico de UAVs con alas que se someten a deformaciones controladas mediante la acción de actuadores piezo-cerámicos, inmersos y distribuidos convenientemente sobre la estructura del ala.

En esta sección se presenta un ejemplo en el que se analiza estática y dinámicamente la estructura de un ala de un UAV con actuadores PZT adheridos. La estructura del ala es de aluminio y está conformada por una viga principal y cuatro vigas secundarias. Se coloca un par de actuadores en el centro de la viga principal y otro par en el extremo libre de la misma, además se adhiere un par de actuadores en el extremo libre de cada una de las vigas secundarias (notar que este modelo es algo más simple que el mostrado en la Figura 2a). Con el fin de cambiar el estado del ala, se induce una flexión pura (Figura 2b) en las vigas aplicando un voltaje de 100 volts, excitación escalón, igual en todos los pares des actuadores PZT adheridos a la estructura.

En la Figura 3 se muestra la configuración que adquiere la estructura del ala en estado estático por la acción de los actuadores piezoeléctricos (V = cte =100 volt). La configuración puede cambiarse de acuerdo a los requerimientos de las diferentes etapas del vuelo, con el fin de mejorar el rendimiento aerodinámico. Aunque en este ejemplo las deformaciones no son significativas, las simulaciones realizadas son un punto de partida para estudiar estructuras de alas flexibles, especialmente en el caso de UAVs con alas que mutan. Para este ejemplo también se muestra la respuesta dinámica (Figura 4). El intervalo de integración es 1,23 veces el primer período natural de vibración de la estructura (T_1 = 32,5 seg). Las condiciones iniciales del conjunto viga-actuadores son: desplazamientos y velocidades igual a cero, y como se mencionó antes, el voltaje aplicado es una entrada escalón. En la Figura 4a se muestra la estructura del ala deformada en un determinado instante, y en la Figura 4b se muestra la evolución temporal de la punta del ala, desplazamiento en la dirección de z_g, en la que se puede ver que el máximo desplazamiento es 0.024 m, aproximadamente un 60 % más alto que la deformación estática. En la misma figura se presentan amplificaciones que muestran el comportamiento oscilatorio que tiene la punta del ala a frecuencias naturales altas. Como el modelo no incluye fuerzas disipativas, la excitación escalón excita todos los modos. Si se considerara el amortiguamiento la respuesta tendería a la deformación estática mostrada en la Figura 3.

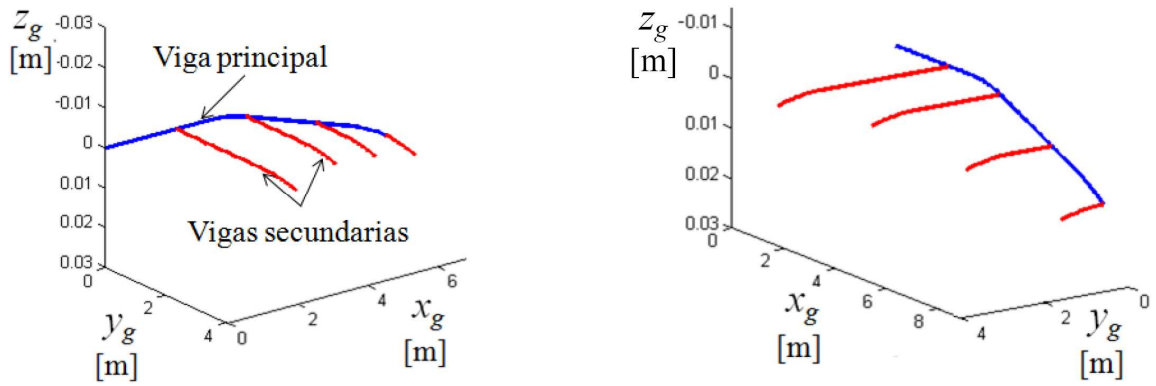


Figura 3: Configuración de la estructura del ala.

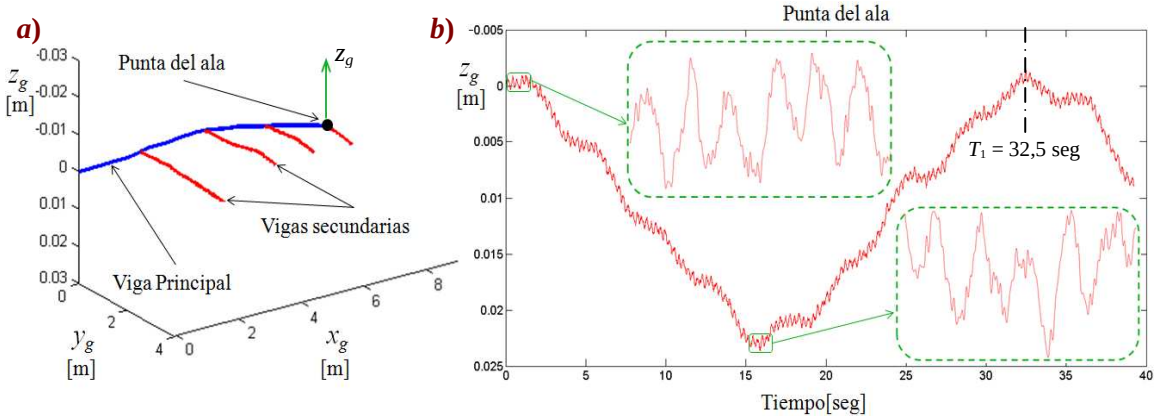


Figura 4: ***a)*** Deformación dinámica. ***b)*** Evolución temporal de la punta del ala.

6. CONCLUSIONES Y TRABAJOS FUTUROS

Se desarrolló un modelo estructural de alas flexibles para vehículos aéreos no tripulados con alas que cambian de forma por la acción de actuadores piezo-cerámicos adheridos que inducen deformaciones controladas sobre la estructura elástica del ala. Este modelo, aunque simple, permite simular el comportamiento estático y dinámico de alas de UAVs inspirados en el vuelo de las aves. El modelo permite variar el ángulo de diedro del ala mediante la deflexión de la viga principal en el plano (x_g-z_g), y además variar la combadura del ala a través de la deflexión de las vigas secundarias. En una próxima etapa, se pretende desarrollar un modelo que considere la variación del ángulo de flecha del ala mediante la deflexión de la viga principal en el plano (x_g-y_g). Como objetivo final se desea acoplar este modelo estructural de alas flexibles, con un modelo aerodinámico desarrollado por los mismos autores de este trabajo, e incorporar el algoritmo de control junto con los sensores como parte del modelo.

REFERENCIAS

[1] J. MARDEN, *Variability in the size, composition and function of insects flight muscles*. Annual Reviews, Vol. 62 (2000) pp. 157-178.

[2] W. SHYY, M. BERG AND D. LJUNGVIST, *Flapping and flexible wings for biological and micro air vehicles*. Progress in Aerospace Sciences, Vol. 35 (1999) pp. 455-505.

[3] K.B. LIM, R.C. LAKE. AND J. HEEG, *Effective selection of piezoceramic actuators for an experimental flexible wing*, Journal of Guidance, Control and Dynamics, Vol. 21 (1998) pp. 704-709.

[4] C. CHEE, *Static shape control of laminated composite plate smart structure using piezoelectric actuators*, Ph.D. Dissertation, Department of Aeronautical Engineering, Sydney 2000.

[5] B. BANDYOPADHYAY, T.C. MANJUNATH AND M. UMAPATHY, *Modeling, control and implementation of smart structures*, Springer-Verlag Berlin Heidelberg 2007.

AVIONES NO-TRIPULADOS CON ALAS QUE MUTAN: ASPECTOS AERODINÁMICOS

Marcos Verstraete[1,2], Mauro Maza[2,3] Sergio Preidikman[1,2,3] y Julio Massa[1,3]

[1]*Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Campus Universitario, Ruta Nacional 36 Km. 601, 5800 Río Cuarto, Argentina. Tel/Fax: 0358-4676246, mverstraete@ing.unrc.edu.ar, http://www.ing.unrc.edu.ar*

[2]*CONICET – Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917, Buenos Aires, Argentina, mauro-maza@hotmail.com, www.conicet.gov.ar*

[3]*Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, spreidikman@efn.uncor.edu, http://www.efn.uncor.edu*

Resumen: Se presenta el desarrollo de una herramienta numérica que simula el comportamiento aerodinámico no-lineal e inestacionario de vehículos aéreos no tripulados (UAVs) inspirados en el vuelo natural con alas que cambian de forma. Para simular numéricamente estos UAVs se utiliza: *i*) un modelo aerodinámico que predice el campo de movimiento del fluido alrededor de la estructura del ala que muta dinámicamente, y las cargas aerodinámicas actuantes sobre esta estructura; *ii*) un modelo cinemático para alas de vehículos aéreos no-tripulados inspirados en el vuelo natural con alas que cambian de forma dinámicamente; y *iii*) un método que permite combinar estos dos modelos. La herramienta desarrollada permite, en un mismo entorno, generar la geometría de la planta alar y una malla adecuada para implementar el modelo aerodinámico, describir la cinemática del ala, realizar simulaciones numéricas del comportamiento aerodinámico de alas que mutan dinámicamente, y visualizar los resultados provenientes de estas simulaciones.

Palabras claves: UAVs, Morphing-wings, Aerodinámica No-lineal e Inestacionaria.

1. INTRODUCCIÓN

La idea de investigar UAVs con alas que cambian de forma, se debe a ellos podrían seguir trayectorias complejas, generar maniobras extremadamente complicadas, y acomodarse de manera ágil para optimizar el rendimiento aerodinámico en múltiples regímenes de vuelo [1], características que no poseen los aviones convencionales en la actualidad. Los métodos tradicionales usados para estudiar el comportamiento aerodinámico de aeronaves [2, 3, 4] resultan inadecuados para estos avanzados UAVs. Por ello, es necesario desarrollar nuevas herramientas numéricas para la predicción de fenómenos aeroelásticos complejos.

En este artículo se presenta una herramienta numérica que implementa un modelo aerodinámico, un modelo cinemático y un método para combinar ambos modelos. La idea fundamental consiste en tratar el flujo de aire y la estructura del ala del UAV como elementos de un único sistema dinámico; e integrar numéricamente, en forma simultánea e iterativa en el dominio del tiempo todas las ecuaciones gobernantes. Para obtener las cargas aerodinámicas se utiliza una técnica conocida como método de red de vórtices no-lineal e inestacionario (NULMV), el cual permite tener en cuenta todas las posibles interferencias aerodinámicas.

2. DESCRIPCIÓN GEOMÉTRICA DEL ALA

En el presente trabajo se consideran solamente aspectos relacionados con la reconfiguración del ala, mientras el resto de la aeronave (fuselaje, empenajes, etc.) no cambia de forma. El ala posee un perfil simétrico y es modelada como una superficie plana sin espesor. La mitad del ala, o semiala, está constituida por tres regiones rígidas diferentes: A, B y C, que se encuentran en la parte interna, central y exterior de la semienvergadura, respectivamente. Cada región es generada en el plano mediante las coordenadas de sus cuatro vértices, esto permite obtener diferentes configuraciones de la planta alar.

El modelo aerodinámico utilizado en este trabajo requiere que el ala sea representada por una sábana vorticosa, discretizada mediante cuadriláteros cuyos vértices son denominados nodos aerodinámicos. La generación de la geometría del ala es realizada en forma automática mediante un código computacional. El código requiere ciertos parámetros geométricos y datos específicos de la misma malla (ver Figura 1). Los parámetros y datos necesarios son: Coordenadas de los vértices (x, y) que definen las regiones explícitamente; Número de nodos a lo largo de la envergadura: N_A en la Región A, N_B en la Región B, N_C en la Región C y el número de nodos a lo largo de la cuerda M (igual para las tres regiones).

En la Figura 1 se muestran con puntos azules y enumerados con números romanos los vértices que definen las

regiones A, B, y C y se indican los parámetros N_A, N_B, N_C y M. Para armar la malla aerodinámica deben generarse en forma ordenada las coordenadas de los nodos aerodinámicos y de los puntos de control (en el centroide del cuadrilátero), las conectividades entre estos últimos y los nodos y deben especificarse cuales serán los nodos de los paneles involucrados en la convección de la vorticidad a el seno del fluido.

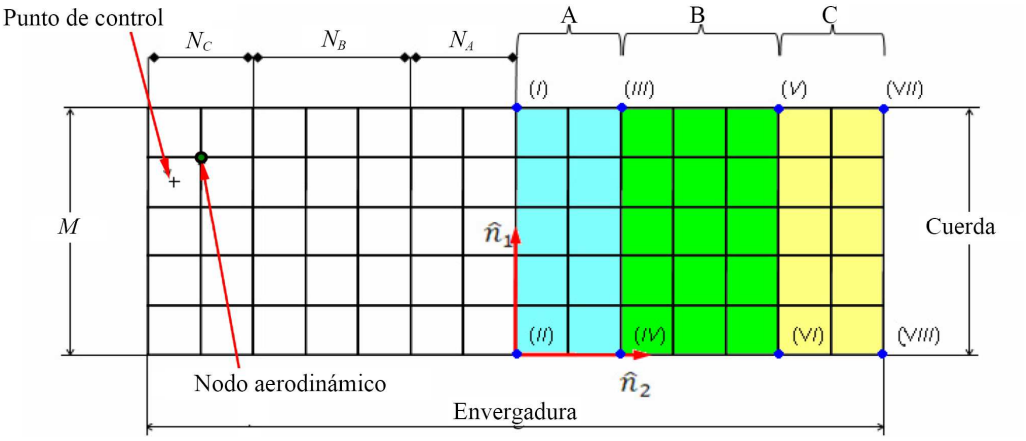


Figura 1: Datos geométricos que definen el ala.

3. MODELO CINEMÁTICO

3.1. SISTEMAS DE REFERENCIAS

Se utilizaron cuatro sistemas de referencia, uno Newtoniano o inercial (N) formado por la terna ($\hat{n}_1, \hat{n}_2, \hat{n}_3$), un sistema fijo a la región A formado por la terna ($\hat{a}_1, \hat{a}_2, \hat{a}_3$), otro fijo a la región B formado por la terna ($\hat{b}_1, \hat{b}_2, \hat{b}_3$) y el último fijo a la región C formado por la terna ($\hat{c}_1, \hat{c}_2, \hat{c}_3$). Estos sistemas de referencia ortogonales son mostrados en la Figura 2a. También se identifican las regiones A, B, y C que componen la planta alar.

3.2. GRADOS DE LIBERTAD

El modelo adoptado posee tres grados de libertad θ_1, θ_2, y θ_3 asociados a los ángulos diedros de cada región como se indica en la Figura 2b. Considerar el ala como simétrica simplifica el problema y permite concentrar la atención sólo en la semiala derecha. Para poder incorporar esta cinemática al modelo aerodinámico es necesario conocer la posición y velocidad de los puntos que componen cada región de la semiala respecto del sistema N. Para ello es necesario hacer una transformación de coordenadas. Para obtener las matrices de transformación se siguió el desarrollo que utilizó [5]. En las ecuaciones (1) a (3) se muestran las operaciones necesarias para realizar esta transformación de coordenadas.

$$[A]=[T_{AN}][N] \qquad [B]=[T_{BA}][A] \qquad [C]=[T_{CB}][B] \tag{1}$$

$$[N]=[T_{AN}]^T[A] \qquad [N]=[T_{AN}]^T[T_{BA}]^T[B] \qquad [N]=[T_{AN}]^T[T_{BA}]^T[T_{CB}]^T[C] \tag{2}$$

$$[N]=[T_{NA}][A] \qquad [N]=[T_{NB}][B] \qquad [N]=[T_{NC}][C] \tag{3}$$

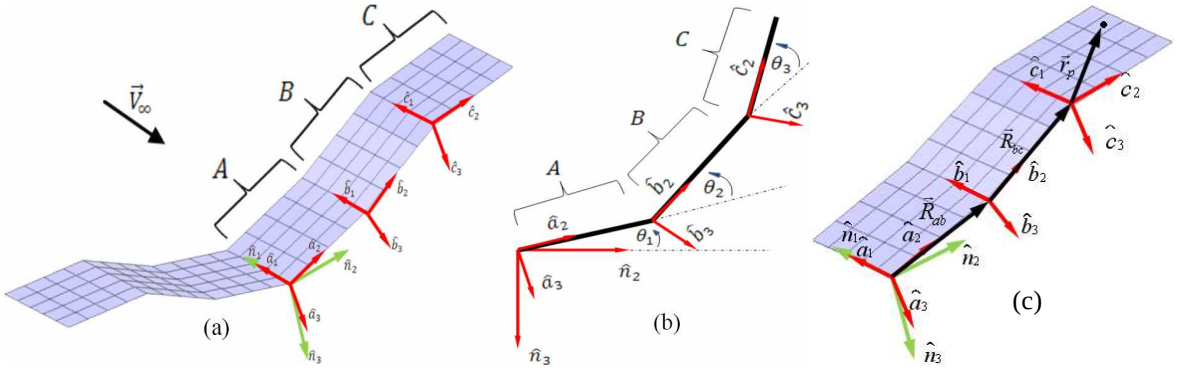


Figura 2: (a) Sistemas de referencia; (b) Grados de libertad; (c) Posición de un partícula *p* en la región C.

3.3. POSICIÓN Y VELOCIDAD

El vector posición, $\vec{R}_p$, de una partícula *p* ubicada en la región C respecto del sistema N, (ver Figura 2c), y su vector velocidad, ${}^N\vec{V}_p$, están dados por (4). Los detalles de la deducción pueden consultarse en [7]

$$\vec{R}_p = \vec{R}_{ab} + \vec{R}_{bc} + \vec{r}_p \qquad \rightarrow \qquad {}^N\vec{V}_p = {}^N\omega^A \times \vec{R}_{ab} + {}^N\omega^B \times \vec{R}_{bc} + {}^N\omega^C \times \vec{r}_p \tag{4}$$

$\vec{R}_{ab}$ es el vector que indica la posición del origen del sistema de referencia B respecto del sistema A, $\vec{R}_{bc}$ es el vector posición del origen del sistema de referencia C respecto del sistema B, y $\vec{r}_p$ es la posición de la partícula *p* respecto del sistema C. Las velocidades angulares de los sistemas fijos a las regiones A, B, y C respecto del sistema N se obtienen mediante el teorema de adición [6].

$${}^N\omega^A = -\dot{\theta}_1 \hat{a}_1 \qquad {}^N\omega^B = {}^N\omega^A + {}^A\omega^B = -\dot{\theta}_1 \hat{a}_1 - \dot{\theta}_2 \hat{b}_1 \qquad {}^N\omega^C = {}^N\omega^A + {}^A\omega^B + {}^B\omega^C = -\dot{\theta}_1 \hat{a}_1 - \dot{\theta}_2 \hat{b}_1 - \dot{\theta}_3 \hat{c}_1 \tag{5}$$

La velocidad expresada en el sistema de referencia N es:

$$\left\{ {}^N\vec{V}_p \right\}_N = [T_{NA}] \left\{ {}^N\omega^A \right\}_A \times \left\{ \vec{R}_{ab} \right\}_A + [T_{NB}] \left\{ {}^N\omega^B \right\}_A \times \left\{ \vec{R}_{bc} \right\}_B + [T_{NC}] \left\{ {}^N\omega^C \right\}_C \times \left\{ \vec{r}_p \right\}_C \tag{6}$$

Los vectores posición de los puntos de las regiones A y B y las expresiones para sus velocidades se obtienen fácilmente observando la Ec. (4) y (6). Los detalles pueden consultarse en [7].

4. EL MODELO AERODINÁMICO

El modelo aerodinámico implementado en este esfuerzo es el método general red de vórtices no-lineal e inestacionario (*non-linear unsteady vortex lattice method* o NUVLM) que permite modelar correctamente no-linealidades aerodinámicas asociadas con grandes ángulos de ataque, deformaciones estáticas, y flujos dominados por vorticidad en los que el fenómeno conocido como *vortex bursting* no ocurre. El modelo predice adecuadamente la emisión de vorticidad desde los bordes filosos de las superficies sustentadoras hacia el seno del fluido que es transportada por el flujo de aire desde las alas hacia el fluido y forma así las estelas. La distribución de la vorticidad en las estelas y la forma de las mismas son, también, parte de la solución del problema. Se escogió un método basado en el NUVLM porque existen numerosas aplicaciones previas de dicho método en las cuales se ha probado que es confiable y un muy buen predictor de las cargas aerodinámicas inestacionarias y no-lineales [8].

En flujos sobre superficies sólidas donde el número de Reynolds es alto, se genera vorticidad por efectos viscosos en capas muy delgadas, llamadas capas límites, que están pegadas a las superficie del sólido inmerso en el seno fluido. Los efectos viscosos son responsables de la existencia de las capas límites. Parte de esta vorticidad es emitida desde el borde de fuga y la punta de las alas, es transportada por el fluido, y forma las estelas. El campo de velocidades asociado con toda esta vorticidad interactúa con la llamada corriente libre: mientras las condiciones de borde de no-penetración y no-deslizamiento son satisfechas sobre las superficies sólidas generadoras de vorticidad, la vorticidad en las estelas se mueve libremente en el fluido de forma tal que no se produzcan saltos de presión a través de las estelas. Más detalles acerca de los fundamentos matemáticos y de la implementación numérica del NUVLM pueden consultarse en la referencia [8].

5. COMBINACIÓN ENTRE EL MODELO AERODINÁMICO Y EL CINEMÁTICO

Para calcular las incógnitas del problema, en cada paso de tiempo, se debe tener en cuenta la variación temporal de la velocidad de la superficie sustentadora. Las deformaciones dinámicas del sólido perturban la estela cercana a la superficie y esto influye de manera significativa en el cálculo de las cargas aerodinámicas.

La combinación de ambos modelos se logra mediante la condición de contorno de *no penetración*, que implica que la velocidad normal de una partícula de fluido relativa a la superficie sustentadora es cero, esto es para toda partícula que se encuentre en la región de la sabana adherida S. Matemáticamente se puede expresar como:

$$\left(\vec{V} - \vec{V}_s\right) \cdot \vec{n} = 0 \qquad \text{en S} \tag{7}$$

donde $\vec{V}$ es la velocidad absoluta de las partículas de fluido, $\vec{V}_S$ es la velocidad de la superficie y $\vec{n}$ es el vector normal unitario a la superficie. La velocidad de la superficie es obtenida del modelo cinemático y está asociada a la variación de los ángulos diedros de cada región.

6. RESULTADOS

La herramienta numérica desarrollada en este trabajo permite determinar los coeficientes aerodinámicos adimensionales en función del tiempo (CN, C_l, C_d, C_y), la distribución de presión sobre la superficie sustentadora (C_p) y la visualización de los datos provenientes de las simulaciones numéricas.

A modo de ejemplo se analiza una planta alar con un ángulo de flecha de 27°y ángulo de ataque igual a 10º, que se ha discretizada en 24 paneles a lo largo de la envergadura y 5 paneles a lo largo de la cuerda. Su velocidad de traslación es nula y está inmersa en una corriente uniforme de velocidad unitaria. Sus ángulos diedros están sometidos a una variación temporal que alteran la distribución de presiones sobre la superficie. La posición y velocidad de las regiones asociadas a los ángulos diedros se presentan en la Figura 3.

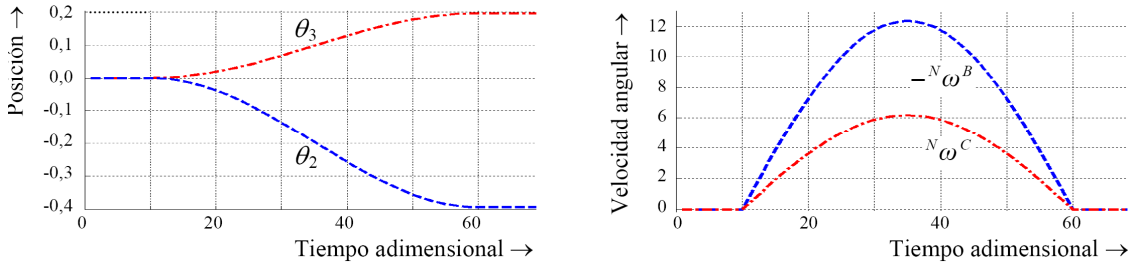


Figura 3: Posición y velocidad angular de los ángulos diedros

En la Figura 4 se muestra el comportamiento de los coeficientes de sustentación (C_l) y resistencia (C_d) en estado transitorio para las condiciones mencionadas anteriormente. Se observa un máximo de sustentación y de resistencia en la etapa donde la planta alar está en movimiento. El valor máximo de C_l es aproximadamente un 18 % mas elevado que el la magnitud que toma en estado estacionario. Para todos los puntos calculados el valor del coeficiente de resistencia toma un valor que igual al 17.6 % del valor del coeficiente de sustentación.

En la Figura 5 se presenta la evolución de la estela luego de un arranque impulsivo. Post-procesando los datos provenientes de las simulaciones numéricas se obtuvo una vista en perspectiva de los ángulos de diedro utilizados y de las estelas.

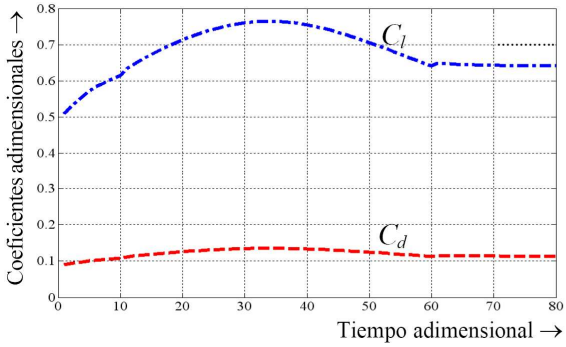


Figura 4: Variación temporal de C_l y C_d.

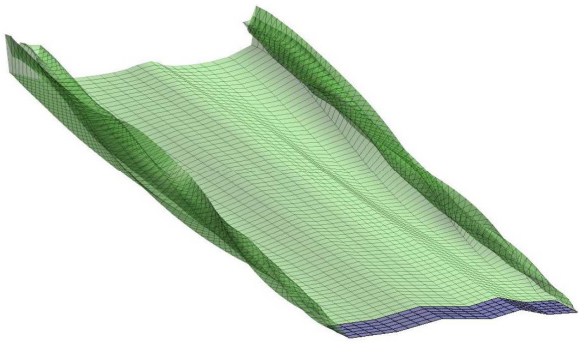


Figura 5: Evolución de la estela (vista en perspectiva)

7. CONCLUSIONES

La herramienta desarrollada genera una malla adecuada para el modelo aerodinámico implementado, aplica un modelo cinemático para el ala, acopla el modelo cinemático y el aerodinámico con el fin de obtener la distribución de presiones sobre la superficie sustentadora y en la parte final post-procesa los datos provenientes de las simulaciones numéricas para visualizarlos. La utilización del NUVLM permite predecir las cargas aerodinámicas actuantes como así también las interacciones aerodinámicas entre la estela y las alas del vehículo.

REFERENCIAS

[1] J. BOWMAN, B. SANDERS, AND T. WEISSHAAR, *Evaluating the impact of morphing technologies on aircraft performance*, AIAA Paper 2002-1631, April 2002.

[2] J.D. ANDERSON, *Fundamentals of Aerodynamics*, McGraw-Hill Science/Engineering/Math, January 2001.

[3] M. LANDAHL AND H. ASHLEY, *Aerodynamics of Wings and Bodies*, Dover Publications, 1985.

[4] L.M. MILNE-THOMSON, *Theoretical aerodynamics*, Dover Publications, June, 1973.

[5] B. ROCCIA, S. PREIDIKMAN Y J. MASSA, *Desarrollo de un código computacional para simular y analizar la cinemática de alas batientes*, Mecánica Computacional, Vol. 26 (2007), pp. 3224-3245.

[6] T.R. Kane, *Dynamics*. Holt, Rinehart and Winston, Inc., 1968.

[7] M.. VERSTRAETE, L. CEBALLOS, Y S. PREIDIKMAN, *Aviones no-tripulados inspirados en el vuelo natural con alas que mutan: Aspectos aerodinámicos*, Mecánica Computacional, Vol. 28, (2009), pp. 2975-2993.

[8] P. KONSTANDINOPOULOS, D. MOOK AND A.. NAYFEH, *A numerical method for general, unsteady aerodynamics*, AIAA-81-1877, 1981.

ASPECTOS DE DISEÑO DE UN SIMULADOR DE VUELO

Alejandro C. Limache[a], Marina H. Murillo[a], Pablo S. Rojas Fredini[a] y Leonardo Giovanini[b]

[a] *International Center of Computational Methods in Engineering (CIMEC), INTEC-CONICET. Santa Fe, Argentina., alejandrolimache@gmail.com, http://www.cimec.gov.ar/*
[b] *Laboratorio de Investigación en Señales e Inteligencia SINC(i). UNL . Santa Fe, Argentina.*

Resumen: Este trabajo presenta algunos tópicos del Simulador de Vuelo que se está desarrollando en el Centro Internacional de Métodos Computacionales en Ingeniería (CIMEC). El simulador permite recrear por computadora el comportamiento de un avión real y está capacitado para responder mediante hardware (joysticks y pedales) al comando de un piloto. Un sistema de visualización virtual permite ver las propiedades del terreno y visualizar en tiempo real el movimiento del avión con respecto al mismo. También se puede ver el movimiento del avión desde tierra, desde el aire o inclusive desde el punto de vista del piloto. Además de las aplicaciones educativas o de entrenamiento que tiene un simulador, existen aplicaciones científicas diversas. Una de ellas es que mediante el simulador se pueden diseñar y testear en tiempo real sistemas de control automático. Los sistemas de control automático y la lectura de señales de los Sistemas de Navegación Inercial forman la piedra angular para el desarrollo de Vehículos Aéreos No-Tripulados (UAVs). En el futuro, el simulador será utilizado como plataforma de desarrollo virtual de UAVs. Aquí se describen algunos conceptos modernos de diseño y se muestra como se simulan las distintos sistemas que forman el avión y como éstos se pueden programar independientemente como simples plugins. Se muestran las técnicas de visualización que corren en paralelo en hilos de proceso independiente.

Palabras clave: *flight simulator, real-time, terrain visualization, subsystems, plugins.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Un simulador es una réplica fiel, precisa y tan semejante a la realidad como sea posible de una determinada aeronave. El simulador se construye mediante la utilización de computadoras, hardware y sistemas de visualización. Para muchos, un simulador de vuelo es la experiencia más próxima a poder «volar» un avión.

Los usos que tienen estos simuladores de vuelo son de diversa índole, por ejemplo tienen mucha utilidad en la simulación de desperfectos técnicos como pueden ser fallas en el motor, turbinas e instrumentos de medición. También se utilizan para el desarrollo de aeronaves puesto que el mismo permite el testeo de hardware y software en tiempo de vuelo. Quizás la ventaja más importante de utilizar un simulador de vuelo en tiempo real, es que a través del mismo es posible «entrenar» a pilotos (y futuros pilotos) a un menor costo económico, pudiendo experimentar situaciones durante el vuelo que no podrían tener lugar en la vida real puesto que implicaría un riesgo a la vida humana. Frente a estas situaciones extremas el piloto adquiere los conocimientos necesarios para saber que acciones llevar a cabo en caso de encontrarse en una circunstancia similar en la vida real. Es una forma indirecta de adquirir experiencia y habilidad que pueden trasladarse directamente a una realidad de vuelo dada.

Se debe destacar que además de las aplicaciones estandard mencionadas arriba, el simulador nos permitirá diseñar diversos sistemas de control usandolo como banco de prueba. Los sistemas de control podrán luego aplicacarse en aviones reales o en aviones autónomos no-tripulados.

2. DISEÑO DEL SIMULADOR DE VUELO

En esta sección desarrollaremos brevemente la estructura general que presenta nuestro simualdor de vuelo. La misma puede verse en la Fig. 1. El diagrama de la Fig. 1 describe de manera sintética los bloques que forman parte de nuestro simulador de vuelo. Este esquema es el que se utliza mayormente para el desarrollo de los *Full Flight Level D Simulators* [3].

Del diagrama de Fig. 1, podemos ver que el diseño del simulador de vuelo puede dividirse en seis grandes módulos:

1. *Cockpit Flight Controls* está formado por el hardware típico que puede encontrarse en la cabina del piloto de un avión real. Es allí donde se encuentran los comandos principales del avión: columna,

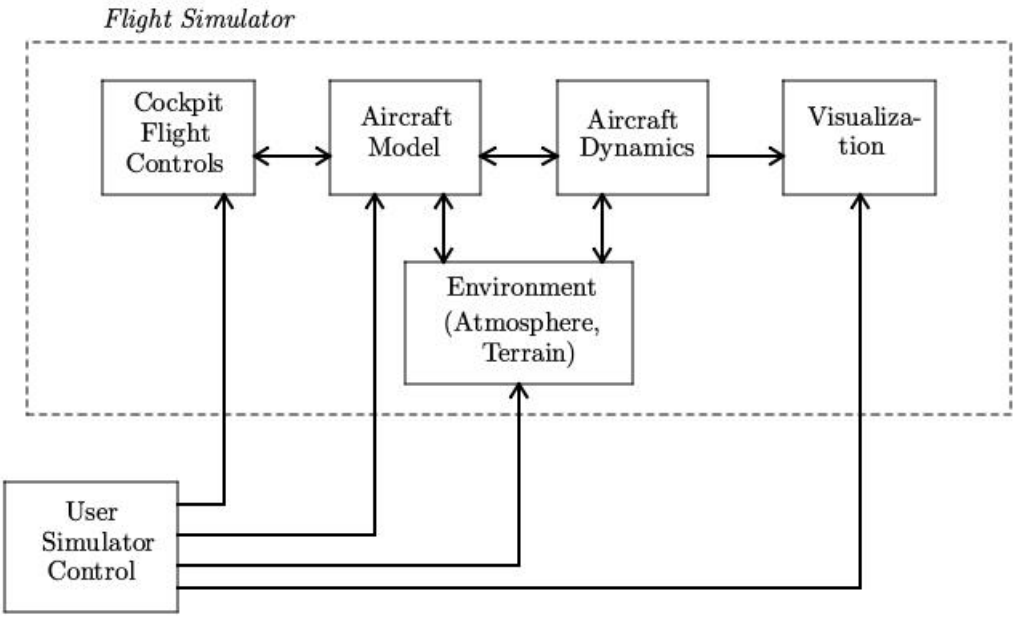
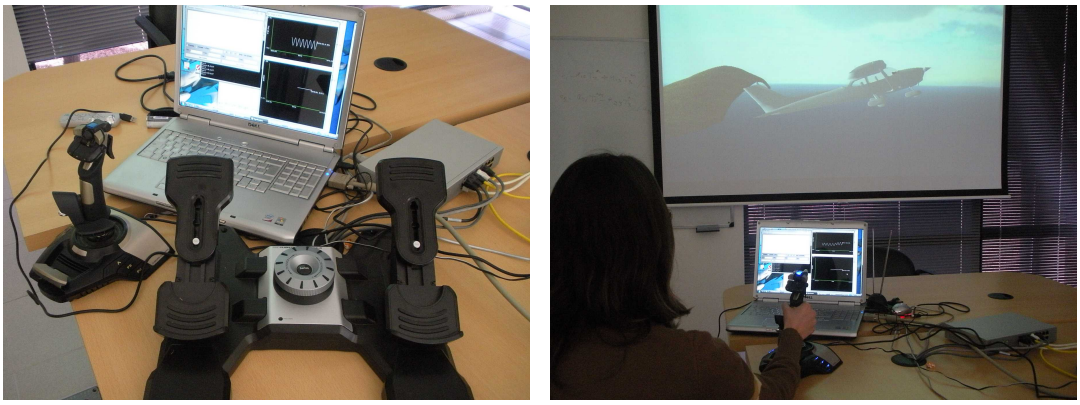


Figura 1: Diagrama en Bloques del Simulador de Vuelo

timón, pedales y propulsión. En nuestro simulador, las entradas de estos comandos se realizan mediante un Joystick similar al que se utiliza en aviones de combate. Con el mismo podemos controlar la posición de la columna, del timón y la propulsión. Además contamos con pedales para simular el mismo efecto que se produce en un avión real al pisar los mismos, es decir, producir el movimiento de guiñada. En las Fig. 2(a) y Fig. 2(b) podemos observar el hardware con el cual estamos desarrollando nuestro simulador de vuelo.



(a) Hardware Típico de un Avión Real (b) Ejemplo de Simulación

Figura 2: Simulación de Vuelo mediante el Hardware adquirido

Eventualmente es posible comandar nuestro simulador de vuelo mediante un teclado estándar. Para administrar los dispositivos de entrada de hardware utilizamos la popular librería multiplataforma [2]. La misma permite leer el estado de los dispositivos y enviar efectos de force-feedback a aquellos que lo soporten.

2. En *Aircraft Model* se describe el tipo de avión que se utilizará durante la simulación de vuelo. En base al tipo de avión seteado es posible obtener las diferentes fuerzas aerodinámicas y de propulsión que permiten que el avión «vuele».

3. *Aircraft Dynamics* es el bloque en el cual se procesan las ecuaciones de movimiento donde se realiza

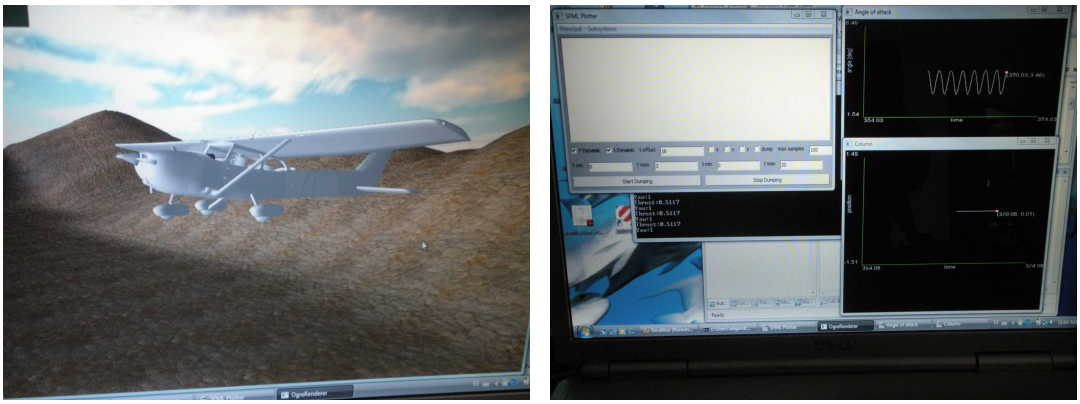
la integración temporal de las mismas. Este bloque recibe como entradas las fuerzas aerodinámicas y de propulsión obtenidas en el módulo *Aircraft Model*.

4. Con *Environment* nos referimos al bloque que se encarga de simular diversas condiciones ambientales, como pueden ser el tipo de atmósfera en que se está volando, luminosidad, lluvias, pistas de aterrizaje, etc.

   Es este módulo el que nos permite también determinar el terreno sobre el cual nuestro avión está volando, es decir, nos permite observar las características del suelo como ser montañas, océanos, ciudades, etc.

5. El módulo *Visualization* es el que hace las veces de interfaz entre el usuario y la computadora. Con él es posible observar el vuelo del avión desde distintos puntos de vista, como por ejemplo desde la cabina del piloto o bien desde tierra. En la Fig. 3(a) puede observarse el modelo del avión utilizado durante nuestra simulación.

   Además de la función mencionada anteriormente, este módulo nos permite observar el comportamiento de una diversidad de variables. Esto nos permite verificar las condiciones en las cuales el avión se encuentra volando y resultan de mucha utilidad en la etapa de depuración. Esto puede verse en la Fig. 3(b).



(a) Modelo del Avión (b) Ploteo de Variables

Figura 3: Ejemplos de Simulaciones

Para la visualización del entorno simulado utilizamos la librería multiplataforma [1]. La misma permite la ejecución del simulador en distintos sistemas operativos y hardware gráfico de manera transparente. Para el ploteo de variables se desarrolló una herramienta propia utilizando la librería [4] para gráficos en 2D.

6. El módulo *User Simulator Control* es quien posibilita la interacción entre el usuario y algunos de los módulos accesibles por él:

   – En *Cockpit Flight Controls* el usuario puede, por ejemplo, setear que las funciones que deben cumplir el joystick y los pedales sean distintas a las definidas por default, o también cambiar las funciones asignadas a las distintas entradas por teclado.

   – En *Aircraft Model* es posible modificar en tiempo de ejecución el tipo de avión que se está volando durante la simulación.

   – En *Environment* el usuario puede ir modificando la apariencia visual de la simulación intercambiando por ejemplo el día con la noche, simulando tormentas o lluvias, aterrizajes en diversas condiciones climáticas, también es posible modificar los aeropuertos de partida o de llegada, las cuidades sobre las que se vuela, etc.

– En *Visualization* el usuario puede ir ajustando o modificando la visión que se tiene desde la cabina del piloto a través de una cámara.

3. CONCLUSIONES

En este trabajo se describen algunos aspectos de diseño del Simulador de Vuelo que se está desarrollando en el CIMEC. El Simulador de Vuelo permite recrear en tiempo real el vuelo de un avión y está capacitado para responder mediante hardware (joystick y pedales) al comando de un piloto. Un sistema de visualización virtual permite ver las propiedades del terreno y observar en tiempo real el movimiento del avión. La aeronave se puede observar desde distintos puntos de vista: desde tierra, desde el aire o inclusive desde el punto de vista del piloto. De este modo es posible observar de distintas maneras su desplazamiento y su respuesta frente a los comandos del piloto. Además un sistema de visualización en paralelo que también se ha desarrollado permite acceder en tiempo real a la evolución de todas las variables físicas que describen el estado del avión así como a información de los comandos dados en los controles de cabina. Al día de hoy, se sigue trabajando en el simulador de vuelo.

REFERENCIAS

[1] OGRE3D, *http://www.ogre3d.org/*, 2010.

[2] OIS, *http://www.wreckedgames.com/*, 2010.

[3] ROCKWELL-COLLINS, *http://www.rockwellcollins.com/*, 2010.

[4] SFML, *http://www.sfml-dev.org/*, 2010.

COUPLING STRATEGY BETWEEN 0D/1D AND MULTI-D CODES FOR THE SIMULATION OF COMPRESSIBLE FLOW PROBLEMS

Ezequiel J. López[♭] and Norberto M. Nigro[†]

[♭]*Dpto. de Mecánica Aplicada, Facultad de Ingeniería, Universidad Nacional del Comahue, Buenos Aires 1400, 8300 Neuquén, Argentina, ezequiel.jose.lopez@gmail.com*

[†]*Centro Internacional de Métodos Computacionales en Ingeniería (CIMEC), INTEC-CONICET, Universidad Nacional del Litoral, Güemes 3450, 3000 Santa Fe, Argentina, nnigro@santafe-conicet.gov.ar, www.cimec.org.ar*

Abstract: This article presents a strategy to link dimensionally heterogeneous models for compressible flows and an algorithm to solve the coupled problem. The algorithm is based on a 'loose' coupling between the codes with a stage loop in order to reach a strong coupling when such loop converges. Some preliminary results of the application of the proposed algorithm to the realization of in-cylinder flow computations in internal combustion engines are presented.

Keywords: *Code coupling, Geometric multiscale model, In-cylinder flows, Internal combustion engine simulation*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

Multidimensional CFD (Computational Fluid Dynamics) codes allow a detailed computation of the flow in problems of interest in science and engineering, being able to assess the impact of geometry and operating conditions therein. In some problems, however, its high computational cost becomes unviable to use to simulate simultaneously all components of the system under study (for instance, in the simulation of internal combustion engines). On the other hand, if only a part of a given system is simulated with a multidimensional (multi-D) model, the imposition of boundary conditions to this model could not be a simple task due to the influence of the rest of the system on the part which is solved in a detailed way. Currently, a typical approach is to simulate a specific part of the system of interest with a CFD multi-D code and the rest with simplified models (0D/1D). Thus, the 0D/1D code provides appropriate boundary conditions for the multidimensional computation. This approach is known as Geometric Multiscale method and allows a substantial reduction of the numerical complexity. The use of this method leads to the need to couple properly the dimensionally heterogeneous models, both in the mathematical formulation as in the computational implementation. This work presents some preliminary results of the coupling of CFD 0D/1D and multi-D codes for the realization of in-cylinder flow computations in internal combustion engines.

2 COUPLING STRATEGY

We consider two dimensionally heterogeneous models for compressible flows: a multi-D model represented by the Navier-Stokes for compressible flows in deformable domains [2], and a 1D model consisting in the Euler equations for quasi-one-dimensional flows [3]. The multi-D model and the numerical technique used can be found in [6, 4]. A complete description of the 1D model is presented in [5].

Now, we assume that a given domain is splitted in two or more sub-domains, where in some sub-domains the multi-D model is applied, and in the remaining sub-domains the 1D model is used. For the sake of simplicity in the analysis that follows, we consider only two sub-domains. The sub-domains interchange mass, momentum and energy through the *coupling interface* Γ_c. Thus, from the point of view of each sub-problem, the coupling interface is an inlet/outlet boundary and could be solved by using absorbing boundary conditions [9]. In this case, the reference state for computing the absorbing boundary condition is provided by the corresponding state at Γ_c.

Let $\Omega_{1D} = (x_0,\, x_c)$ the spatial 1D domain, on which the 1D model governs the fluid flow. The problem is completely defined once the initial and boundary conditions are provided. At both end points, we apply absorbing boundary conditions. The reference state at point x_0 is defined by either atmospheric conditions or by the resulting state of the left sub-domain (0D or 1D). The condition at x_c (the coupling 'interface' for

Ω_{1D}) could be imposed as

$$\mathbf{\Pi}^{-}_{U_p}[\mathbf{U}_{p1D}(x_c,\, t) - \overline{\mathbf{U}}_{pMD\rightarrow 1D}(t)] = \mathbf{0}, \quad t \geq 0 \tag{1}$$

where $\mathbf{U}_{p1D} = [\rho_{1D}, u_{1D}, p_{1D}]^T$ is the primitive variables vector [3], ρ_{1D}, u_{1D} and p_{1D} being the density, velocity, and pressure, respectively; $\mathbf{\Pi}^{-}_{U_p}$ is the projection matrix onto left-going characteristics [9]; and $\overline{\mathbf{U}}_{pMD\rightarrow 1D}(t) = [\overline{\rho}_{MD}(t),\, \overline{u}_{MD}(t),\, \overline{p}_{MD}(t)]^T$ is a reference state arising from the *condensation* of the variables at the coupling interface of the multi-D domain. For scalar components of the state vector ($\overline{\rho}_{MD}$ and $\overline{p}_{MD}$) this condensation is simply the mean value of the variable on the coupling surface

$$\overline{\rho}_{MD}(t) = \frac{1}{\mathrm{meas}(\Gamma_c)} \int_{\Gamma_c} \rho_{MD}(\mathbf{x},\, t) d\sigma, \quad \overline{p}_{MD}(t) = \frac{1}{\mathrm{meas}(\Gamma_c)} \int_{\Gamma_c} p_{MD}(\mathbf{x},\, t) d\sigma \tag{2}$$

where $\mathrm{meas}(\Gamma_c) = \int_{\Gamma_c} d\sigma$. For vector components (the fluid velocity),

$$\overline{u}_{MD}(t) = \frac{-n_{1D}}{\mathrm{meas}(\Gamma_c)} \int_{\Gamma_c} \mathbf{u}_{MD}(\mathbf{x},\, t) \cdot \mathbf{n} d\sigma \tag{3}$$

$\mathbf{n}$ being the unit outward normal vector to Γ_c, and n_{1D} is the outward normal to Ω_{1D} ($n_{1D} = -1$ for the left end and $n_{1D} = 1$ for the right end).

For the multi-D problem the condition applied on Γ_c must ensures the well-posedness of the problem. Then, we would need for all $\mathbf{x}$ on Γ_c n_- boundary conditions, n_- being the rank of $\mathbf{\Pi}^{-}_{U_pn}$ and where $\mathbf{\Pi}^{-}_{U_pn}$ is the projection matrix onto the left-going characteristics of the advective jacobian projected onto the normal direction to the coupling interface [9]. In this study we propose to impose at every $\mathbf{x} \in \Gamma_c$ the same reference state $\overline{\mathbf{U}}_{p1D\rightarrow MD}(t)$, *i.e.*

$$\mathbf{\Pi}^{-}_{U_pn}\left[\mathbf{U}_{pMD}(\mathbf{x},\, t) - \overline{\mathbf{U}}_{p1D\rightarrow MD}(t)\right] = \mathbf{0}, \quad \forall \mathbf{x} \in \Gamma_c \tag{4}$$

In the last equation $\mathbf{U}_{pMD} = [\rho_{MD}, \mathbf{u}_{MD}, p_{MD}]^T$, and $\overline{\mathbf{U}}_{p1D\rightarrow MD} = [\rho_{1D}(x_c), -u_{1D}(x_c) n_{1D}\mathbf{n}, p_{1D}(x_c)]^T$.

2.1 COUPLING ALGORITHM

We use the `PETSc-FEM` [8] code for the multi-D simulation and the 0D/1D models are integrated into the internal combustion engine simulator `ICESym` [7]. In terms of computational cost, the multi-D model determines practically the cost of the coupled problem. Therefore, we designed an algorithm taking into account this issue. The time step of the coupled simulation is set by the flow problem in the multi-D domain Δt_{MD}. Due to the 0D/1D code utilizes an explicit scheme for time integration, the maximum time step allowed for this code (Δt_{1D}) will be limited by the Courant-Friedrichs-Levy (CFL) condition [3]. Generally Δt_{1D} is smaller than Δt_{MD} thus, a sub-cycling strategy is needed in order to maintain the run synchronization. Since both codes interact through the reference state at the coupling interface, the multi-D problem only can provide reference states at times t and $t + \Delta t_{MD}$ for the 0D/1D code (and not at every time in the sub-cycling time steps). Then, we assume a linear interpolation between $\overline{\mathbf{U}}_{\mathrm{ref},MD\rightarrow 1D}(t)$ and $\overline{\mathbf{U}}_{\mathrm{ref},MD\rightarrow 1D}(t+\Delta t_{MD})$. The proposed algorithm has a loose coupling between the multi-D and the 0D/1D code, thus, a 'stage loop' was added in order to reach a strong coupling when such loop converges. The basic algorithm could be stated as

1: initialize variables
2: **for** $n = 0$ to n_{step} **do** {main time loop}
3: $\quad t^n = n\Delta t_{MD}$
4: $\quad$ **for** $i = 0$ to n_{stage} **do** {stage loop}
5: $\quad\quad$ **for** $k = 0$ to n_{nwt} **do** {newton loop of MD code}
6: $\quad\quad\quad \mathbf{U}^{n+1,i+1}_{MD} = \text{CFD-MD}(\mathbf{U}^{n}_{MD}, \overline{\mathbf{U}}^{n+1,i}_{\mathrm{ref},1D\rightarrow MD})$
7: $\quad\quad$ **end for**
8: $\quad\quad$ compute $\overline{\mathbf{U}}^{n+1,i+1}_{\mathrm{ref},MD\rightarrow 1D}$ from $\mathbf{U}^{n+1,i+1}_{MD}$

9: send $\overline{\mathbf{U}}^{n+1,i+1}_{\mathrm{ref},MD\to 1D}$ to the 0D/1D code
10: $m = 0$ {sub-cycling counter of the 0D/1D code}
11: $t^m_{1D} = t^n$ {0D/1D time initialization}
12: $\tilde{\mathbf{U}}^0_{1D} = \mathbf{U}^n_{1D}$
13: **while** $t^m_{1D} \leq t^n + \Delta t_{MD}$ **do** {0D/1D time loop}
14: compute Δt^m_{1D} {time step computation satisfying the CFL condition}
15: **if** $t^m_{1D} + \Delta t^m_{1D} > t^n + \Delta t_{MD}$ **then**
16: $\Delta t^m_{1D} = t^n + \Delta t_{MD} - t^m_{1D}$
17: **end if**
18: $t^{m+1}_{1D} = t^m_{1D} + \Delta t^m_{1D}$
19: compute $\tilde{\mathbf{U}}^{m+1}_{\mathrm{ref},MD\to 1D} = \overline{\mathbf{U}}^n_{\mathrm{ref},MD\to 1D} + \frac{t^{m+1}_{1D} - t^n}{\Delta t_{MD}}(\overline{\mathbf{U}}^{n+1,i+1}_{\mathrm{ref},MD\to 1D} - \overline{\mathbf{U}}^n_{\mathrm{ref},MD\to 1D})$ {linear interpolation of the reference state during the sub-cycling iteration}
20: $\tilde{\mathbf{U}}^{m+1}_{1D} = \text{CFD-0D/1D}(\tilde{\mathbf{U}}^m_{1D}, \tilde{\mathbf{U}}^{m+1}_{\mathrm{ref},MD\to 1D})$
21: $m = m + 1$
22: **end while**
23: $\mathbf{U}^{n+1,i+1}_{1D} = \tilde{\mathbf{U}}^m_{1D}$
24: compute $\overline{\mathbf{U}}^{n+1,i+1}_{\mathrm{ref},1D\to MD}$ from $\mathbf{U}^{n+1,i+1}_{1D}$
25: send $\overline{\mathbf{U}}^{n+1,i+1}_{\mathrm{ref},1D\to MD}$ to the MD code
26: **end for**
27: **end for**

In the algorithm, n_{step} is the number of time steps in the simulation, n_{nwt} is the number of Newton loops in the nonlinear problem, and n_{stage} is the number of stages in the coupling scheme. $\mathbf{U}^{n+1,i+1}_{MD} = \text{CFD-MD}(\mathbf{U}^n_{MD}, \overline{\mathbf{U}}^{n+1,i}_{\mathrm{ref},1D\to MD})$ is the operator inside the CFD multi-D code that advances the multi-D fluid state using the reference state $\overline{\mathbf{U}}^{n+1,i}_{\mathrm{ref},1D\to MD}$; whereas $\tilde{\mathbf{U}}^{m+1}_{1D} = \text{CFD-0D/1D}(\tilde{\mathbf{U}}^m_{1D}, \tilde{\mathbf{U}}^{m+1}_{\mathrm{ref},MD\to 1D})$ is the operator inside the 0D/1D code which use the reference state $\tilde{\mathbf{U}}^{m+1}_{\mathrm{ref},MD\to 1D}$ in the coupling interface at time t^m_{1D} in the sub-cycled time. Lines 10 to 23 in the algorithm could be encapsulated into an operator of the form $\mathbf{U}^{n+1,i+1}_{1D} = \text{CFD-0D/1D\_2}(\mathbf{U}^n_{1D}, \mathbf{U}_{\mathrm{ref},MD\to 1D}, t^n, t^n + \Delta t_{MD})$, where t^n is the initial time, $t^n + \Delta t_{MD}$ is the final time, and $\mathbf{U}_{\mathrm{ref},MD\to 1D}$ is a function of time (in this case, the prescribed linear interpolation computed at line 19 in the algorithm).

If the reference state at the coupling interface for the multi-D problem has a small (relative) variation between two consecutive times, the basic algorithm could converge slowly and, thus, incrementing the computational cost. For these cases, we propose to include the resolution of the 0D/1D problem *inside* the Newton loop of the multi-D code; *i.e.*, to solve the 0D/1D problem (from t^n to t^{n+1}) for each non-linear iteration. In the structure of the basic algorithm, this modification could be performed taking $n_{\mathrm{nwt}} = 1$ and checking the convergence of the non-linear iteration in the stage loop. Of course, this strategy demands to have access to the source code in order to control the non-linear loop. We found a good convergence rate when the cited strategy is applied in problems where the coupling surface remains unchanged, with only a few number of non-linear iterations respecting to the basic algorithm. However, the strategy could fail if the area of Γ_c changes in time with a high area ratio.

3 Numerical results. Motored opposed-piston engine

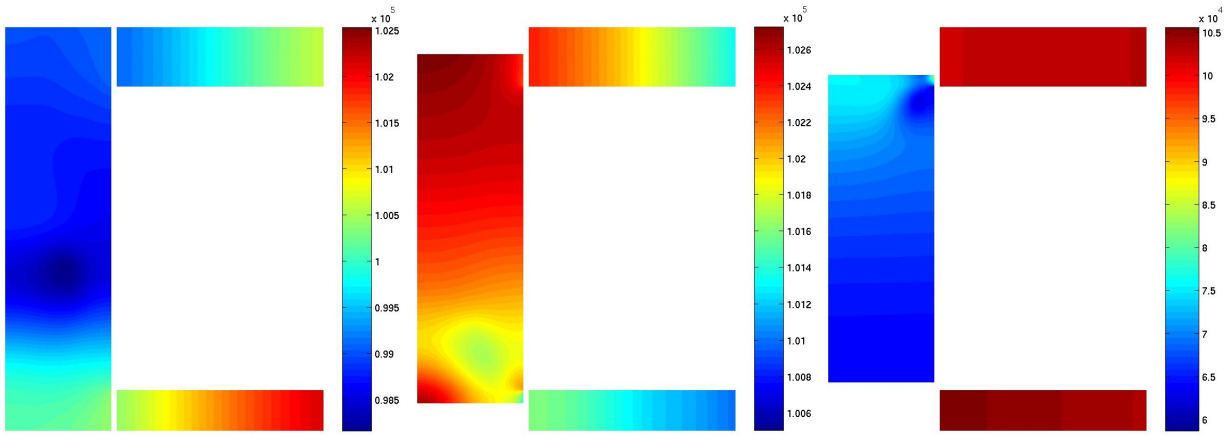
This case consists in the resolution of the fluid flow inside the cylinder of an opposed-piston engine under cold conditions, *i.e.* without firing it. The engine geometry was taken from the KIVA-3 tutorial [1]. The cylinder bore is 100 mm, the stroke of each piston is 85 mm, and the geometric compression ratio is 9.5:1. The cylinder has 8 exhaust ports evenly distributed in the circumferential direction and 12 intake ports uniformly separated also. Assuming the reference angle as the EDC (External Dead Center), the timing of the ports are: Intake Port Opening (IPO) = 295.13°; Intake Port Closing (IPC) = 64.87°; Exhaust Port Opening (EPO) = 280.2°; Exhaust Port Closing (EPC) = 79.8°.

In order to simplify the problem, the flow domain is assumed to have axial symmetry around the cylin-

der axis. The actual domain is not axisymmetric since the intake and exhaust ports are not continuously distributed as a 'ring' around the cylinder. However, the proposed simplification is perfectly valid for the purpose of this study.

The mesh was generated with the pistons at EDC (ports totally opened) and has 19K hexahedra and 38.6K nodes. The mean element size is $h = 1$ mm. Due to the simplicity of the geometry and the boundary movement, mesh dynamics is solved using an algebraic law following a linear distribution with respect to the position of pistons at IDC (Internal Dead Center). No-slip condition is imposed at solid walls. In addition, these walls are assumed to be insulated. Mixed absorbing/wall boundary conditions are used to model the ports [4]. Turbulence is modeled applying the simplest LES (Large Eddy Simulation) Smagorinsky model [10], which takes the Smagorinsky coefficient as constant. The engine speed is 3000 rpm. The time step used in the simulation was $\Delta t = 1 \times 10^{-5}$ s. Three cycles were simulated, until the convergence to a periodic state.

Figure 1 shows the pressure field at some instants in the last simulated cycle. These instants are expressed in crank angle degrees (CAD) with respect to EDC. Figure represent the pressure into the cylinder (multi-D) and in the intake and exhaust pipes (1D).



(a) 0 CAD (b) 54 CAD (c) 288 CAD

Figure 1: Pressure field ([Pa]) into the cylinder (left), intake pipe (bottom) and exhaust pipe (top).

REFERENCES

[1] A. AMSDEN, *KIVA-3: A KIVA program with block-structured mesh for complex geometries*, Los Alamos, New Nexico, 1993.

[2] J. DONEA, S. GIULIANI, AND J. HALLEUX, *An arbitrary, Lagrangian-Eulerian finite element method for transient dynamic fluid-structure interactions*, SIAM Journal on Scientific Computing, Vol. 33 (1982), pp. 689-700.

[3] C. HIRSCH, *Numerical Computation of Internal and External Flows. Volume 2: Computational Methods for Inviscid and Viscous Flows*, John Wiley & Sons, 1990.

[4] E. LÓPEZ, *Methodologies for the numerical simulation of fluid flow in internal combustion engines*, Tesis doctoral, CIMEC-INTEC, Universidad Nacional del Litoral, 2009.

[5] E. LÓPEZ, AND N. NIGRO, *Validation of a 0D/1D computational code for the design of several kind of internal combustion engines*, Latin-American Applied Research, Vol. 40 (2010), pp. 175184.

[6] E. LÓPEZ, N. NIGRO, S. SARRAF, AND S. MÁRQUEZ DAMIÁN, *Stabilized finite element method based on local preconditioning for unsteady compressible flows in deformable domains with emphasis on the low mach number limit application*, International Journal for Numerical Methods in Fluids (2010), submitted.

[7] N. NIGRO, E. LÓPEZ, AND J. GIMENEZ, *ICESym. An Internal Combustion Engine Simulator*, Copyright ©2010, `http://code.google.com/p/icesym/`.

[8] M. STORTI, N. NIGRO, R. PAZ, L. DALCÍN, AND E. LÓPEZ, *PETSc-FEM. A general purpose, parallel, multi-physics FEM program for CFD applications based on MPI/PETSc*, Copyright ©1999-2010, `http://www.cimec.org.ar/petscfem`.

[9] M. STORTI, N. NIGRO, R. PAZ, AND L. DALCÍN, *Dynamic boundary conditions in computational fluid dynamics*, Computer Methods in Applied Mechanics and Engineering, Vol. 197 (2008), pp. 1219-1232.

[10] D. WILCOX, *Turbulence Modeling for CFD*, D C W Industries, 2 Ed., 2002.

SIMULACIÓN DE FLUIDOS INTERACTIVOS EN TIEMPO REAL

P.S. Rojas Fredini† y A.C. Limache†

†*International Center of Computational Methods in Engineering (CIMEC) INTEC-CONICET. Santa Fe, Argentina , http://www.cimec.gov.ar/*

Resumen:

Este trabajo presenta las principales características de una formulación computacional desarrollada por los autores y basada en el método llamado Smoothed Particle Hydrodynamics. La formulación resuelve numéricamente las ecuaciones de Navier-Stokes permitiendo la simulación de dinámica de fluidos, tanto compresibles como casi-incompresibles. El método es simple, explícito, computacionalmente rápido y apto para la computación en paralelo. Estas características, junto con el empleo de técnicas avanzadas de computación y visualización han sido utilizadas para el desarrollo de una plataforma de simulación virtual de dinámica de fluidos con la que se puede cambiar interactivamente propiedades físicas del fluido.

Palabras clave: *smoothed particle hydrodynamics, ghost-particles, interactive simulation, multi-threading*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Smoothed particle hydrodynamics (SPH) es un método de partículas que no utiliza mallas (meshless) basado en la convolución de propiedades puntuales de un campo y un kernel de interpolación[1, 2]. En este trabajo se describe una formulación basada en SPH orientada a resolver las ecuaciones de Navier-Stokes para fluidos compresibles y pseudo-incompresibles que han desarrollado los autores. Durante el proceso de desarrollo del método se construyó una plataforma de visualización que ha evolucionado hasta convertirse en una plataforma interactiva de gran utilidad. La posibilidad de realizar simulaciones interactivas en tiempo real tiene infinidad de aplicaciones, desde simuladores para entrenamiento de profesionales (en ingeniería, medicina, etc.), videojuegos, control de procesos, diseño de prototipos, etc.

2. FORMULACIÓN DE SPH

SPH se basa en aplicar dos conceptos fundamentales a los términos de una ecuación diferencial: representación integral y aproximación de partículas [2, 1, 3]. El primero se refiere a la representación de una función mediante una convolución con una función de suavizado. El segundo concepto consiste en la aproximación que se lleva a cabo para llevar al mundo discreto la representación integral.

2.1. REPRESENTACIÓN INTEGRAL

Una función $f(x)$ puede ser aproximada en forma integral como una convolución con una función de suavizado denominada *kernel*[4]

$$< f(x) >= \int_V f(x')W(x - x', h)dV' \tag{1}$$

siendo h la distancia entre partículas. De la misma forma se puede aproximar la derivada espacial de una función [2]:

$$< \nabla f(x) >= -\int_V f(x') \left[\nabla' W(x - x', h)\right] dV' \tag{2}$$

2.2. APROXIMACIÓN DE PARTÍCULAS

En SPH los fluidos se representan como un conjunto finito de partículas donde cada partícula P_i posee una masa m_i y ocupa un volumen dV_i. Por conservación de masa dicho volumen puede ser representado en función de la densidad alrededor de P_i de la siguiente manera:

$$dV_i = \frac{m_i}{\rho_i} \tag{3}$$

La aproximación de partículas consiste en calcular las representaciones integrales de la sección 2.1 como una sumatoria de las contribuciones de los volúmenes discretos alrededor de cada partícula. Entonces las versiones discretas de las Ecs. (1) y (2) evaluadas sobre P_i se calculan como:

$$< f_i >=< f(x_i) >= \sum_{j=1}^{N} m_j \frac{f_j}{\rho_j} W_{ij} \tag{4}$$

$$< \nabla_i f >= -\sum_{j=1}^{N} m_j \frac{f_{ij}}{\rho_j} \nabla_i W_{ij} \tag{5}$$

o alternativamente:

$$< \nabla_i f >= \sum_{j=1}^{N} m_j \left(\frac{f_i + f_j}{\rho_j} \right) \nabla_i W_{ij} \tag{6}$$

con $f_j = f(x_j),\ \ W_{ij} = W(x_i - x_j, h),\ \ \nabla_i W_{ij} = \frac{\partial}{\partial x_i} [W(x_i - x_j, h)]$. Una cantidad f con dos índices de partícula indica diferencia.

2.3. FUNCIONES DE KERNEL

Existen varias alternativas para la elección del kernel. Cuanto más parecido sea a la función Delta más correcta será nuestra aproximación ya que como se mostró en [1] el error crece cuadráticamente con h.Para el presente trabajo se utilizó el kernel conocido como *spiky*. El mismo se define como:

$$W(r,h) = \frac{K_d}{h^{n_d}} \begin{cases} (2-q)^3, & 0 \leq q \leq 2 \\ \\ 0, & 2 < q \end{cases} \tag{7}$$

donde en 2 dimensiones la constante de normalización es $K_d = \frac{5}{16\pi}$. El soporte de dicha función es $2h$.

3. APLICACION DE SPH A LAS ECUACIONES DE NAVIER-STOKES

Si se aplica la aproximación de la Ec. (4) a la densidad se obtiene:

$$\rho_i = \sum_{j=1}^{N} m_j W_{ij} \tag{8}$$

Usando una descripción lagrangiana, la ecuación de momento puede ser escrita en forma de divergencia de la siguiente manera:

$$\rho \frac{Dv}{Dt} = -\nabla p + \nabla \cdot \boldsymbol{\tau} \tag{9}$$

siendo τ el tensor de tensiones viscosas para fluidos newtonianos. Usando (6) en (9) para el cálculo de los términos de presión y viscosidad obtenemos:

$$\frac{Dv_i^\alpha}{Dt} = -\sum_{j=1}^{N} m_j \frac{p_i + p_j}{\rho_i \rho_j} \frac{\partial W_{ij}}{\partial x_i^\alpha} + \sum_{j=1}^{N} m_j \frac{\tau_i^{\alpha\beta} + \tau_j^{\alpha\beta}}{\rho_i \rho_j} \frac{\partial W_{ij}}{\partial x_i^\beta} \tag{10}$$

Al término de momento es común en SPH agregarle un término de viscosidad artificial [1].

3.1. CONDICIONES DE BORDE

Existen diferentes métodos para implementar las condiciones de borde[2, 1]. Los autores implementaron partículas ficticias reflejando las partículas reales usando conceptos similares a los utilizados en el método de las imágenes en electroestática. La densidad de las partículas ficticias o *ghost* se calcula como $\rho_{ghost} = \rho_a$ siendo ρ_a la densidad de la partícula real. Mientras que su velocidad v_{ghost} se fija de acuerdo a la siguiente extrapolación: $v_{ghost} = 2u_{wall} - v_a$ dónde u_{wall} es la velocidad local de la pared y v_a es la velocidad de la partícula real.

4. Desarrollo del simulador

4.1. Descripción general

El objetivo principal del presente trabajo fue desarrollar un simulador que permitiera ejecutar simulaciones de fluidos en tiempo real o del orden del tiempo real. Es decir simulaciones veloces durante las cuales el comportamiento del sistema pueda seguirse visualmente paso a paso. Además el simulador debía permitir modificar los parámetros del sistema de manera interactiva, sin necesidad de reiniciar toda la simulación. En las próximas secciones se hará una breve descripción de cada uno de los aspectos mas destacables a tener en cuenta.

4.2. Scripting

Para poder dotar al simulador de una gran flexibilidad de uso y adaptabilidad antes los cambios de una manera natural se recurre a los lenguajes de scripting. Estos lenguajes permiten realizar extensiones de programas compilados. Los lenguajes de scripting son interpretados, por lo tanto mucho mas lentos. Esto implica que haya que ser cuidadoso a la hora de utilizarlos. En particular para el simulador desarrollado se escogío Lua que es un lenguaje muy popular en los entornos dónde la performance del scripting es crítica: Los videojuegos.

Para implementar la extensión por scripting se incorporó una consola al simulador dónde el usuario puede enviar sus comandos de script. Lua sólo se utiliza para setear variables y llamar funciones implementadas en código nativo, esto asegura que el impacto sobre la eficiencia es mínimo.

Para conectar el simulador implementado en C++ con la consola que ejecuta una máquina virtual de Lua es necesario escribir funciones para adaptar los tipos de datos y convenciones de llamadas entre ambos lenguajes. Luabind es una librería basada en *expression templates* que permite generar dicho código de manera automática con unas sencillas llamadas a sus rutinas.

Esta consola permite al usuario acceder absolutamente a todos los parámetros configurables del simulador y cambiarlos con el simulador en ejecución.

4.3. Multi-threading

SPH es un método paralelizable por naturaleza, ya que se basa en cálculos locales de cada partícula. Además el algoritmo es explícito. Esto implica que el cálculo de las partículas puede ser desacoplado y llevado a cabo de manera independiente.

En el presente trabajo se utilizó paralelismo a nivel de CPU. El simulador utiliza $n + 1$ hilo de ejecución o *thread*, dónde n es la cantidad de hilos por hardware que soporta el microprocesador. El hilo principal se encarga de sincronizar los hilos de cálculo y llevar a cabo la visualización. Los n hilos restantes se dividen de manera equitativa el trabajo de cálculo. Los hilos de cálculo esperan una señal de sincronismo dada por el hilo principal, de manera tal que dicho hilo controla la frecuencia de actualización del simulador.

Para tener un buen balance de carga las partículas se alojan en un espacio de memoria común a todos los hilos de ejecución, y cada hilo es autosuficiente en el sentido que reconoce automáticamente el segmento que le corresponde calcular. Estos segmentos son dinámicos debido a la presencia de partículas de borde. Esta división se hace en conjuntos y permite disminuir al máximo los mecanismo de sincronicación necesarios propios del ambiente multitarea.

Para la implementación se utilizó la popular librería multiplataforma *boost::thread*. Los hilos no se crean ni se destruyen en cada paso de tiempo ya que es una tarea costosa.

4.4. Ordenamiento Espacial

Si bien el método es *meshless* para realizar a cabo el cálculo de una partícula es necesario conocer cuáles son sus vecinas espaciales que se encuentran dentro del soporte del kernel. Para realizar esto de manera eficiente se implementó un spatial hash uniforme (espacio dividio en celdas de tamaño constante) en el dominio de las partículas. Si bien es cierto que esta estructura debe ser reconstruida en cada paso de tiempo, la tarea de rescontruirla no es demasiado costosa ya que cada celda del hash se comporta como una lista

simplemente enlazada.

5. Ejemplo de uso

En esta sección veremos un ejemplo del simulador en plena ejecución. Para este caso utilizamos el caso de un fluido entre dos paredes cilíndricas concéntricas conocido como flujo de Couette-Taylor. Durante la simulación se cambió la gravedad del fluido en tiempo real. Las partículas se vieron sometidas a una fuerza de gravedad con $g = -9{,}8$ y se las dejó alcanzar el reposo inicialmente, como se muestra en la Fig. 1a. Cuando las partículas alcanzaron el reposo mediante la consola de scripting integrada se cambió la gravedad g de -9.8 a 9.8. En la Fig. 1 se puede observar cómo el fluido reacciona ante el cambio de parámetros y se reacomoda hasta alcanzar nuevamente el equilibrio bajo las nuevas condiciones impuestas. Este cambio se realizó sin detenere ni pausar la simulación permitiendo ver interactivamente las reacciones.

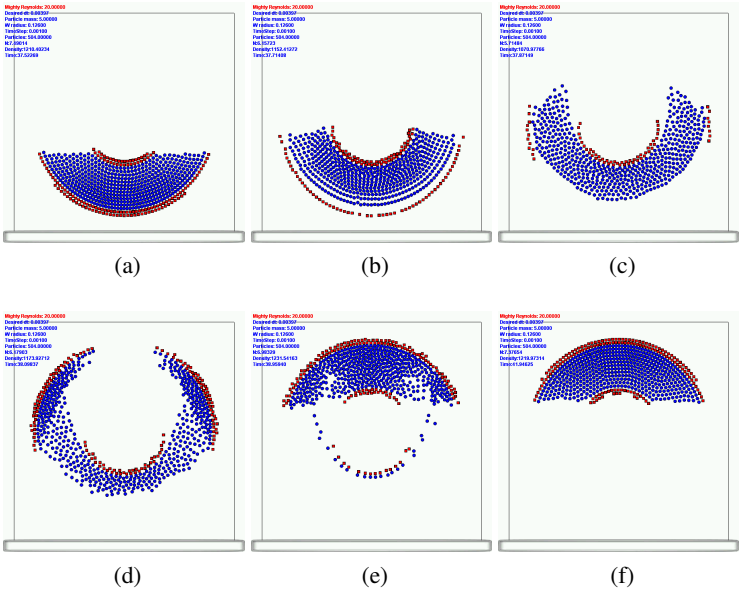


Figura 1: Secuencia cambio de gravedad

6. Conclusiones

Se presentó una formulación computacional basada en SPH para simular fluidos compresibles y pseudo-incompresibles. La misma es explícita y simple. Dichas características permiten realizar los cálculos en tiempos interactivos o en tiempo real y desarrollar herramientas de simulación que brinden resultados instantáneos.

Se introdujeron algunas ideas de cómo encarar el desarrollo de simuladores eficientes y se mostraron las ventajas de la utilización de lenguajes de scripting en el ámbito de la simulación.

Referencias

[1] J. Monaghan, "Smoothed particle hydrodynamics," *Reports on Progress in Physics*, vol. 68, pp. 1703–1759, 2005.

[2] M. Liu and G. Liu, *Smoothed Particle Hydrodynamics: A Meshfree Particle Method*. World Scientific Publishing Co. Pte. Ltd., 2003.

[3] A. Limache and P. Rojas-Fredini, "Validation of a new sph variant using analytical solutions of planar taylor-couette flows," *Computers & Fluids*, vol. submitted, 2010.

[4] J. Monaghan, "Smoothed particle hydrodynamics," *Annu. Rev. Astron. Astrophys.*, vol. 30, pp. 543–74, 1992.

MODELADO NUMÉRICO DE TORNADOS

J.P. Arroyo[a], V. Sonzogni[a] y G. Balbastro[b]

[a]*CIMEC-INTEC-UNL-CONICET, Güemes 3450, 3000 Santa Fe, Argentina, arroyo_jp@hotmail.com, sonzogni@intec.unl.edu.ar*
[b]*UTN, FR Parana y UTN, FR Santa Fe, Lavaise 610, 3000 Santa Fe, Argentina, gbalbastro@yahoo.com*

Resumen: En este trabajo se muestran algunos resultados preliminares del modelado numérico de tornados. El objetivo es el de evaluar el efecto de estos fenómenos sobre las construcciones. Para el modelado numérico se han utilizado, en esta primera etapa, códigos comerciales de elementos finitos que utilizan las ecuaciones de Navier-Stokes no estacionarias incompresibles, con un modelo de turbulencia del tipo L.E.S. (Large Eddy Simulation). El método de discretización utilizado es el de los Elementos Finitos Estabilizados mediante la formulación SUPG/PSPG. Se ha modelado un dispositivo experimental consistente en dos regiones diseñadas para reproducir el flujo de viento de un tornado, una región de convección y una región de convergencia. Los resultados obtenidos son similares a los reportados en la literatura, lo cual permite considerar que la técnica utilizada puede servir a los efectos de la evaluación de presiones sobre construcciones.

Palabras clave: *Tornado, Simulación Numérica, Large Eddy Simulation, M.E.F.*
2000 AMS Subject Classification: 76G25

1. INTRODUCCIÓN

Los tornados son fenomenos meteorológicos muy violentos que, aunque su ocurrencia es muy baja, generan pérdidas de vidas y grandes pérdidas económicas por los daños que ocasionan. En la figura 1 se observan dos pequeñas trombas, fenómenos similares a los tornados, registradas sobre el Río de la Plata el 2 de marzo de 2008. La medición en campo de presiones o velocidades producidas por un tornado son muy dificultosas, por la pequeña escala espacial, por la corta duración temporal y por lo impredecible de la ocurrencia del fenómeno. La información disponible de los mismos es muy limitada y casi inexistente en cuanto a la evaluación de fuerzas actuante de presión sobre las estructuras. Esta es la motivación de este estudio para la simulación de tornados.

Figura 1: Dos pequeñas trombas sobre el Río de la Plata el 02/02/2008.

La mecánica de fluidos computacional (CFD) ha sido utilizada exitosamente para muchos problemas ingenieriles y en particular para aerodinámica de construcciones [1]. El método de los elementos finitos es una de las técnicas numéricas utilizada. Para el tratamiento de la turbulencia, hay distintas alternativas. Una herramienta práctica es mediante simulación de grandes vórtices (LES, por Large Eddy Simulation).

Existen varios laboratorios en el mundo donde se intenta reproducir las condiciones imperantes en un tornado. Para ello se construyen dispositivos en los cuales se reproducen la geometría y los campos de

velocidades y presiones. En este trabajo se ha efectuado el modelado numérico de uno de estos dispositivos [2]. El objetivo es el de comprobar la adecuación de la técnica numérica para su uso en la estimación de presiones sobre objetos o construcciones.

2. MODELO NUMÉRICO

El modelo numérico empleado para simular el tornado, es un dispositivo que consiste en dos regiones diseñadas para reproducir el flujo de viento de un tornado, una región de convección y una región de convergencia, similar a la configuración empleada en los laboratorios experimentales [3], como se indica en la figura 2.

Esta geometría se discretizó mediante elementos finitos, utilizando el software GID/Tdyn [4]. En la figura 3 se muestra la malla de uno de los casos. La misma comporta $2{,}1 \times 10^5$ elementos tetraédricos.

Las condiciones de contorno para este problema han sido las siguientes. En la superficie cilíndrica del tubo de convección y en la tapa del tubo de convergencia hay condiciones de bordes deslizantes. En la tapa superior del tubo de convección se especificó una velocidad de salida prescripta uniforme $u_0 = 2,60\ m/s$. En la superficie cilíndrica del tubo de convergencia, la condiciones de contorno son de entrada libre. En la base del dispositivo las condiciones de bordes son no deslizantes. Además se especificó una presión de referencia p_0 en el punto central de la tapa superior.

El patrón del flujo resultante depende del número de Reynolds y de la 'relación de remolino" (*swirl ratio*). El número de Reynolds se define como $Re = Q/\nu$, donde ν es la viscosidad dinámica y $Q = u_0 A/h$ siendo u_0 la velocidad de salida; A el área del tubo de convección; y h la altura del tubo de convergencia. El denominado *swirl ratio* se define como $S = R\, tan\theta/(2h)$, siendo R el radio del tubo de convección y θ el ángulo con respecto a la normal, con que entra el flujo en la superficie cilíndrica de la zona de convergencia. Cabe mencionar que según los valores de Re y S el ángulo de entrada varía y se obtienen diferentes patrones de flujo.

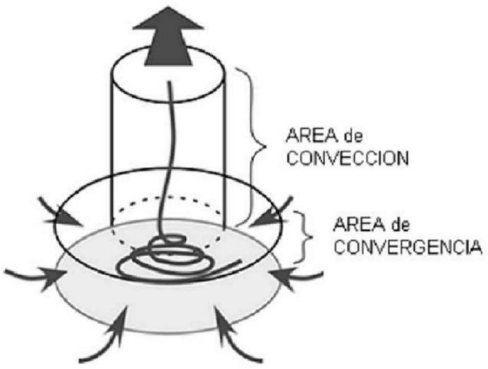


Figura 2: Esquema del Simulador de Tornados

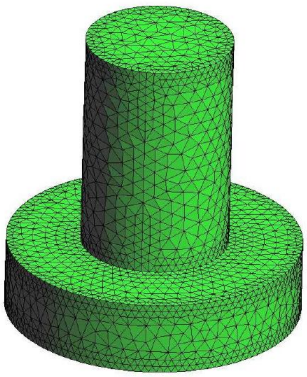


Figura 3: Mallado por Método de Elementos Finitos

3. RESULTADOS NUMÉRICOS

En la figura 4 se muestran algunos de los resultados obtenidos. Allí se pueden ver líneas de corriente donde los colores indican los módulos de velocidades. Se puede observar en esa figura el patrón del escurrimiento que asciende con corrientes helicoidales y en figura 5 el patrón publicado en [3].

En la figura 6-A se muestra un mapa con los valores de la componente de velocidad perpendicular al plano del dibujo y en la figura 6-B un mapa de los valores de presión.

Estos resultados corresponden a un número de Reynolds $Re = 450$, y tienen una "relación de remolino" $S = 0{,}15$. Estas condiciones son equivalentes a los ensayos reportados en referencia [3] y los resultados obtenidos se condicen con los indicados en ella. Para valores de número de Reynolds menores, el patrón de flujo se hace más irregular, en tanto que para valores mayores, se produce una columna ascendente sin rotaciones, prácticamente. La relación de remolino S representa la relación entre las componentes tangenciales

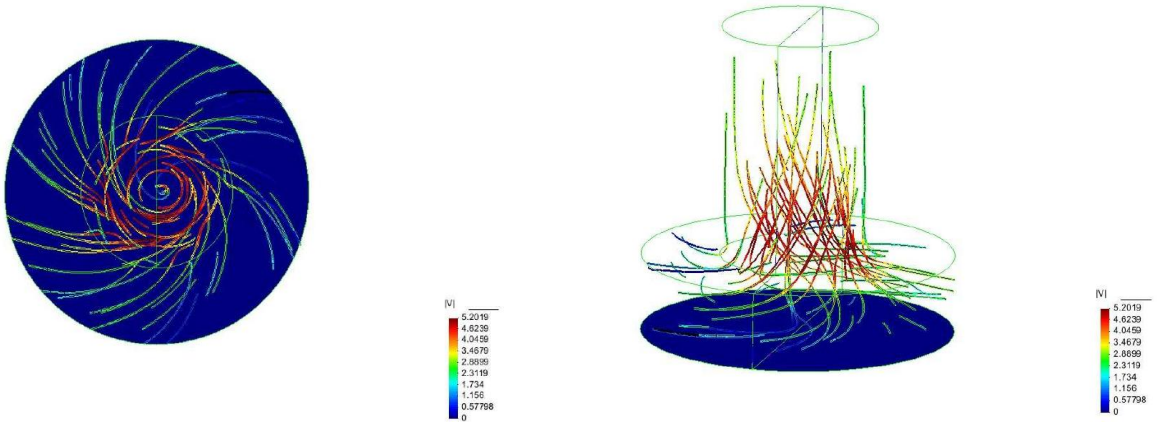


Figura 4: Resultados obtenidos: Patrón del flujo. (A)Vista Superior, (B)Vista Lateral

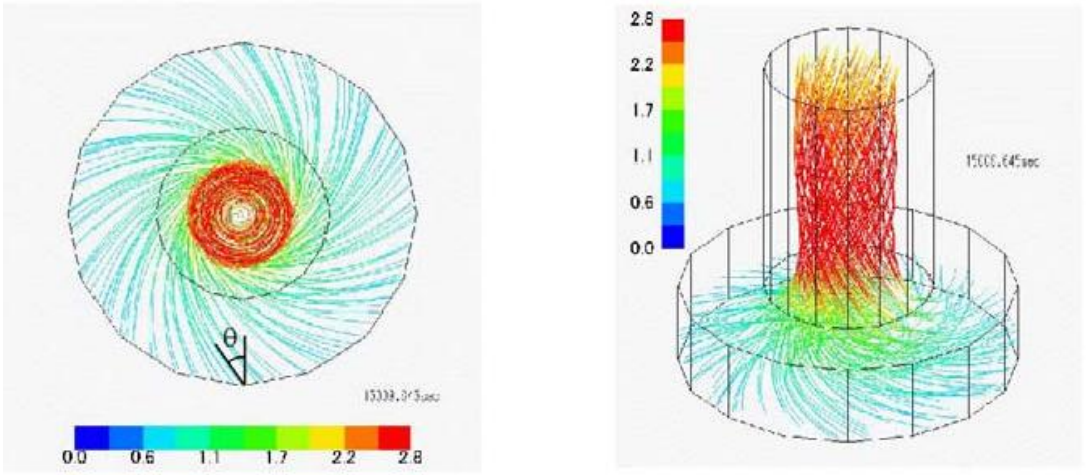


Figura 5: Patrón del flujo según ref. [3] (A)Vista Superior, (B)Vista Lateral

y axiales de la velocidad. Valores bajos de S suelen estar asociados a flujos laminares y valores entre 0.3 y 0.7, según referencia [3], a flujos turbulentos.

4. Conclusiones

En este trabajo se han mostrado algunos resultados preliminares para el modelado numérico de tornados. Debido a la carencia de datos medidos sobre fenómenos reales, se ha procedido a modelar numéricamente un dispositivo experimental utilizado para generar condiciones similares a la de los tornados. El uso de programas de CFD de elementos finitos, con técnicas de LES para representación de la turbulencia ha resultado adecuado para este estudio. Los resultados se condicen con los publicados en la bibliografía.

A partir de esto, el objetivo del proyecto es el de poder estimar numericamente las presiones sobre las estructura o construcciones, cuando estas estan sometidas a los esfuerzos producidos por el paso de un tornado [5]. Esta estimación de presiones tendrá directa aplicación para la evaluación de la seguridad de instalaciones que puedan comprometer vidas humanas, como ser instalaciones de generación de energía, plantas nucleáres, líneas de alta tensión, plantas químicas, etc.

Agradecimientos

Este trabajo se realizó con beca doctoral del CONICET. Se recibió apoyo de los proyectos PICT 2006/1506 de la ANPCYT, PIP 112-200801-2956 del CONICET, CAI+D 2009-III-4-2 de UNL y PID CCPRRA815 de UTN.

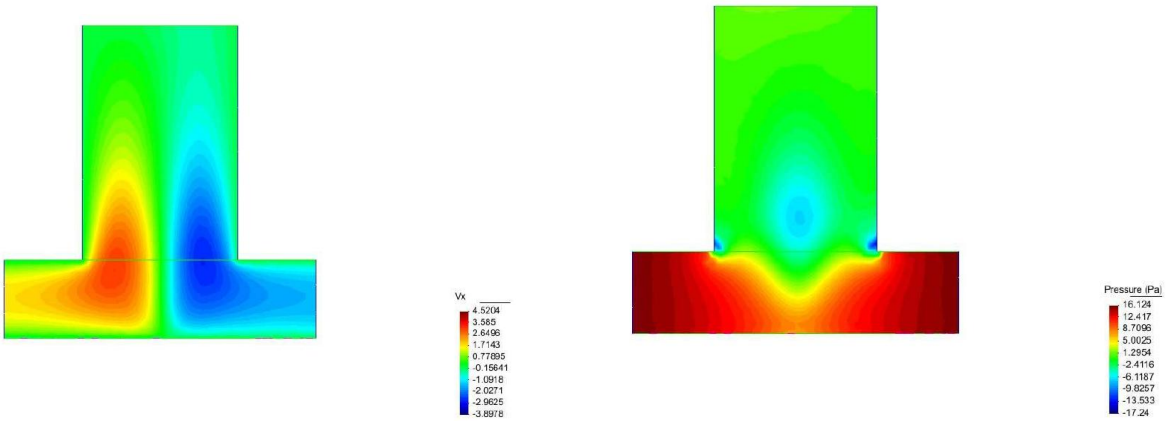


Figura 6: Resultados obtenidos. (A)Perfil de Velocidades, (B)Perfil de Presiones

REFERENCIAS

[1] G. BALBASTRO, *Coeficientes de Presión en Cubiertas Abovedadas Aisladas*, Thesis, UTN, 2009.

[2] N. MONJI, Y. MITSUTA, *A Laboratory Experiment on the multiple structure in tornado-like vortices*,Annual report of D.P.R.I. Kyoto University, 28(B-1), 427-436, 1985.

[3] T. NOMURA, S. MIYATA AND H. HASEBE, *An Attempt of Finite Element Flow Simulation of Tornado Vortices*, APCWE, 2009.

[4] GID *The personal pre and post processor program. www.gid.cimne.upc.es*.

[5] T. MARUYAMA, *A Numerically Generated Tornado-Like Vortex by Large Eddy Simulation*, APCWE, 2009.

AERODINÁMICA DE INSECTOS VOLADORES – ESTUDIO 3D DEL DESPRENDIMIENTO DE VORTICIDAD DESDE EL BORDE DE ATAQUE

Bruno Roccia[1,2,3], Sergio Preidikman[1,2,3] y Julio C. Massa[1,2]

[1] *Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Campus Universitario, Ruta Nacional 36 Km. 601, 5800 Río Cuarto, Argentina. Tel/Fax: 0358-4676246, broccia @ing.unrc.edu.ar, http://www.ing.unrc.edu.ar*

[2] *Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, jmassa@efn.uncor.edu, http://www.efn.uncor.edu*

[3] *CONICET – Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917, Buenos Aires, Argentina, spreidikman@efn.uncor.edu, www.conicet.gov.ar*

Resumen: En este trabajo se presenta el desarrollo de una herramienta de simulación numérica que permite estudiar la aerodinámica no-lineal e inestacionaria asociada al complejo movimiento de las alas de insectos y aves pequeñas. En particular se estudia la influencia del desprendimiento de vorticidad desde el borde de ataque de las alas en la generación de sustentación, para una configuración de 'vuelo suspendido' (*hover*). Los Biólogos y zoólogos conjeturan que los desprendimientos de verticidad producto del batimiento de las alas, causan el notable incremento de las cargas aerodinámicas, respecto de los valores que predicen las teorías aerodinámicas estacionarias y lineales. El movimiento de los puntos ubicados sobre las alas se describe utilizando un modelo cinemático desarrollado con anterioridad por los autores de este trabajo. La combinación entre el modelo cinemático y el modelo aerodinámico, junto con un preprocesador para generar la geometría del insecto, forman una herramienta computacional que permite considerar: *i*) diferentes cinemáticas para el movimiento de las alas, *ii*) distintas geometrías para el insecto (cabeza, tórax, abdomen y alas), predecir: *iii*) el campo de movimiento del fluido alrededor de la estructura del cuerpo y de las alas batientes, *iv*) la distribución espacio-temporal de la vorticidad adherida al cuerpo y a las alas del insecto, *v*) la distribución de vorticidad y la topología de las estelas emitidas desde los bordes filosos de las alas (incluyendo el borde de ataque), *vi*) las cargas aerodinámicas actuantes sobre éstas, y *vii*) tener en cuenta todas las posibles interferencias aerodinámicas.

Palabras claves: *Aerodinámica, Borde de ataque, Vorticidad.*

1. INTRODUCCIÓN

Cuando se estudia experimentalmente en un túnel de viento el flujo de aire alrededor de las alas de insectos en el rango de velocidades correspondientes al vuelo natural de estas criaturas, las fuerzas medidas en los experimentos son sustancialmente más grandes que aquellas predichas por las teorías aerodinámicas convencionales. Esta deficiencia, en las predicciones de la aerodinámica clásica respecto de los valores experimentales, incentivó el estudio de mecanismos aerodinámicos no-convencionales que puedan explicar el incremento en la producción de sustentación y empuje presentes en los vuelos con alas batientes.

En la década de los setenta, especialistas en diferentes áreas tales como biología e ingeniería, condujeron numerosos experimentos que permitieron identificar la presencia de un vórtice adherido al borde de ataque (*leading edge vortex*, LEV) en el vuelo de diferentes clases de insectos. Este fenómeno altamente no-estacionario explicó en gran medida los grandes valores de sustentación producidos por estas criaturas para mantenerse en vuelo. Estudios posteriores sobre el vuelo libre de insectos (sin restricciones mecánicas) y sobre modelos dinámicamente escalados (*Flapper* y *Robofly*) [1, 2], revelaron la existencia de otros mecanismos aerodinámicos no-convencionales asociados al batimiento de las alas: *i*) el retraso de la pérdida dinámica; *ii*) la sustentación rotacional; y *iii*) la captura de la estela [3].

2. DESCRIPCIÓN DEL MODELO

La geometría del modelo adoptada en este trabajo para estudiar la aerodinámica de alas batientes corresponde a una mosca de la fruta (*fruit fly*) y fue extraída de Markow y O'Grady [4]. Los principales parámetros morfológicos de la criatura que han sido preservados en este estudio son: la longitud del ala '*R*', la longitud del cuerpo '*L*', la cuerda máxima del ala y la forma del ala (ver Figura 1*a*).

Cada una de las partes del cuerpo del insecto (tórax, abdomen y cabeza) fueron modeladas, por simplicidad, como superficies de revolución. El modelo completo del insecto, incluyendo las alas, fue implementado íntegramente en MATLAB®, y se utilizó una técnica de parametrización con el fin de construir modelos de diferentes tamaños conservando las proporciones de la criatura. Se utilizaron elementos cuadriláteros simples de cuatro nodos para discretizar las superficies de revolución que componen el cuerpo del insecto, como así también las superficies planas que modelan las alas de la criatura.

2.1 CINEMÁTICA

La orientación del ala en cada instante de tiempo se define especificando la evolución temporal de tres ángulos: (*i*) el ángulo $\phi(t)$ que orienta el ala respecto del plano de batimiento, '*stroke position*'; (*ii*) el ángulo de desviación $\theta(t)$ respecto del plano de batimiento, '*stroke deviation*'; y (*iii*) el ángulo de rotación $\psi(t)$ respecto del eje longitudinal del ala.

Para llevar a cabo la transformación del sistema fijo al ala **B**, al sistema inercial **N**, primero se realiza una secuencia de rotaciones (2–3) mediante los ángulos δ y ξ, que tienen un valor constante igual a -180º y -90º respectivamente. Luego se realiza una '1' rotación dada por el ángulo $(\chi + \beta)$ para orientar el cuerpo del insecto y el plano de aleteo del mismo en el espacio 3D. Por último se utiliza una secuencia de ángulos de Euler $(1-3-2)$ mediante los ángulos $\phi(t)$, $\theta(t)$ y $\psi(t)$ definidos anteriormente. La descripción detallada de los parámetros cinemáticos, y la formulación matemática completa de la cinemática de alas batientes (posición, velocidad y aceleración) pueden consultarse en la referencia [5].

2.2 SEÑALES DE ENTRADA

Para la evolución temporal de los ángulos ϕ y ψ, en este trabajo se adoptó la cinemática utilizada por Dickinson [6], mientras que, con el objeto de simplificar el análisis, el ángulo θ se asume idénticamente igual a cero durante todo el ciclo de batimiento.

Como se puede observar en la Figura 1*b* (curva en línea continua de color rojo), la forma triangular de la función que describe la posición del ala dentro del plano de aleteo, implica que el ala se mueve con una velocidad constante en cada '*half stroke*', por otro lado, la forma trapezoidal del ángulo de rotación (curva en línea a trazos de color negro en la Figura 1*b* causa que el ala mantenga un ángulo de rotación constante durante cada '*half stroke*' y rote con una velocidad constante en cada '*reversal stroke*'.

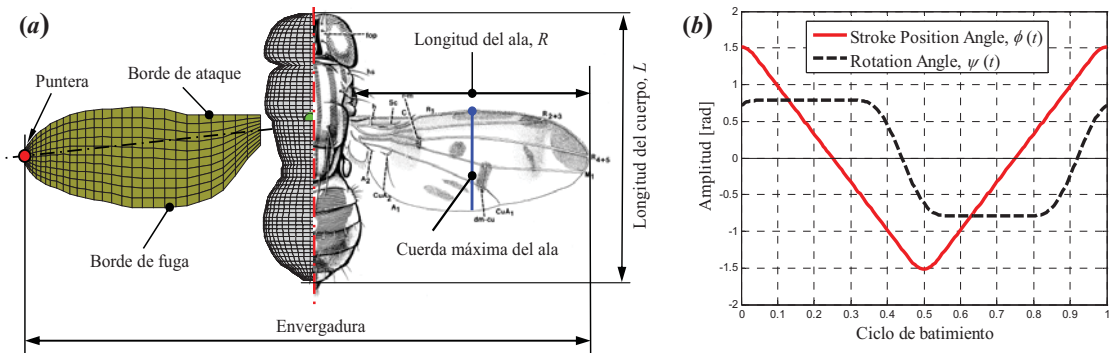


Figura 1: (*a*) Definición de los parámetros morfológicos; (*b*) Evolución temporal de los ángulos $\phi(t)$ y $\psi(t)$ durante un ciclo de batimiento

3. MODELO AERODINÁMICO

El modelo aerodinámico implementado en este esfuerzo es una versión modificada y ampliada del método general de red de vórtices no-lineal e inestacionario (*unsteady vortex lattice method* o UVLM). Este modelo permite modelar correctamente no-linealidades aerodinámicas asociadas con grandes ángulos de ataque, deformaciones estáticas y flujos dominados por vorticidad en los que el fenómeno conocido como *vortex bursting* no ocurre.

Como resultado del movimiento relativo entre el ala y el aire en reposo, se generan gradientes de velocidad que originan vorticidad concentrada en una delgada región adherida a la superficie del ala (capa límite). Esta sabana vorticosa se separa desde el borde de fuga y la puntera del ala, y es convectada hacia el seno del fluido para formar la estela.

En el modelo utilizado se restringe toda la vorticidad del flujo a las capas límites adheridas a las alas y a sus estelas; el flujo se considera irrotacional fuera de estas dos regiones. Las estelas se representan mediante sabanas vorticosas libres. Sus posiciones no son especificadas ya que pueden deformarse libremente hasta adoptar configuraciones libres de fuerzas cuando a través de las estelas no existe ningún salto de presiones. Los dos tipos de sabanas vorticosas (libre y adherida) están unidas en los bordes filosos de las alas donde es impuesta la condición de Kutta para flujos inestacionarios. A medida que el ala se mueve durante un ciclo de

batimiento, el ángulo de ataque efectivo puede alcanzar valores altos produciendo una separación adicional del flujo desde el borde de ataque, justamente ese fenómeno se incorporó al modelo en este trabajo. Para más detalles sobre este modelo el lector puede consultar las referencias [7, 8].

3.1 MODELO DE DESPRENDIMIENTO DESDE EL BORDE DE ATAQUE

El desprendimiento de vorticidad desde el borde de ataque no ocurre en forma continua durante todo el ciclo de batimiento ya que depende del ángulo que forma la dirección de la velocidad local del fluido con el plano del ala, α_e (*leading edge shedding angle*). Existen numerosos trabajos que estudian las condiciones bajo las cuales se produce este tipo de fenómeno como así también la estabilidad del mismo [1]. En este trabajo se adoptó como ángulo límite, a partir del cual ocurre el desprendimiento de vorticidad desde el borde de ataque el valor $\alpha_e = 10º$.

4. SIMULACIONES NUMÉRICAS

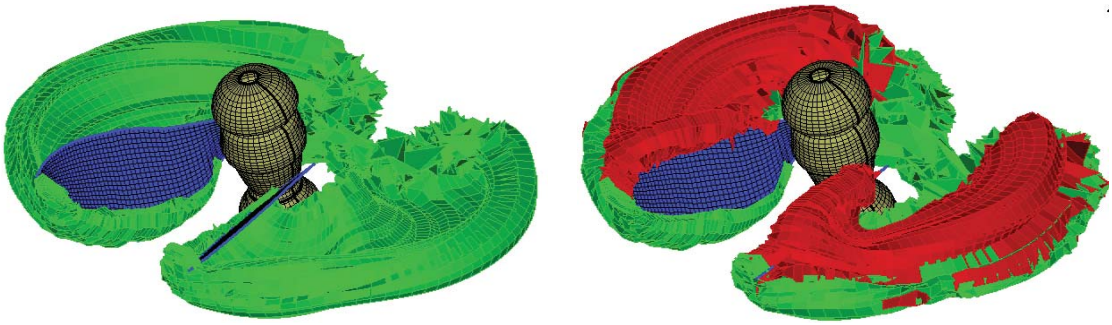
En esta sección se presenta un gráfico que muestra la influencia de considerar el desprendimiento de vorticidad desde el borde de ataque en el cálculo de las cargas aerodinámicas. Los resultados graficados fueron obtenidos con la herramienta computacional desarrollada por los autores de este trabajo. El código está escrito con Fortran 90 compilado para ser ejecutado en un sistema operativo Windows®. Para obtener mayor velocidad de ejecución se han utilizado opciones de optimización automáticas especificas para procesadores Intel® disponibles en el compilador de Fortran empleado.

Se consideraron dos casos:

Caso *a*) donde no se consideró el desprendimiento de vorticidad desde el borde de ataque.

Caso *b*) que incluye desprendimiento de vorticidad desde el borde de ataque.

Los dos casos considerados se ejecutaron para un ciclo de batimiento completo. El tiempo de ejecución en el caso ***a***) fue de 4 horas mientras que en el caso ***b***) fue de 10 horas. Ambos casos se ejecutaron en una computadora de escritorio con una memoria RAM DDR2 de 2 Gb y un procesador con una velocidad de reloj de 3 GHz, con tecnología HT, un bus frontal de 800 MHz y memoria cache de 2 Mb.



(*a*) sin desprendimiento desde el borde de ataque **(*b*)** con desprendimiento desde el borde de ataque

Figura 2: Evolución de la estela al 50 % del ciclo de batimiento;

En la Figura 2 se muestra la evolución temporal de las estelas al 50 % del ciclo de batimiento. En color verde se indica la estela desprendida desde el borde de fuga y puntera del ala y en color rojo la estela desprendida desde el borde de ataque. Se observa que el método implementado capta con gran detalle las interacciones aerodinámicas estelas-estelas y alas-estelas. Estas figuras son meramente cualitativas y son el producto de un primer análisis que tiene como objetivo futuro un estudio completo y cuantitativo de los mecanismos de vuelo asociados al complejo movimiento de las alas de un insecto inmerso en un medio fluido.

4.1 FUERZA DE SUSTENTACIÓN

En la Figura 3 se presenta un gráfico de la fuerza de sustentación, F_L, versus el tiempo adimensionalizado con respecto al período de un ciclo de batimiento. Dicho gráfico contiene dos curvas: *caso a*) sin considerar el desprendimiento de vorticidad desde el borde de ataque (línea continua de color azul) y *caso b*) considerando desprendimiento de vorticidad desde el borde de ataque (línea de trazos de color rojo).

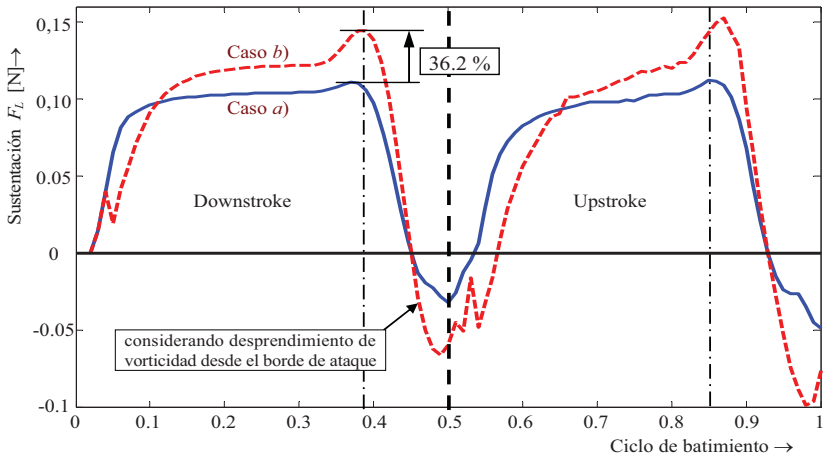


Figura 3: Fuerza de sustentación sobre la superficie sustentadora a lo largo de un ciclo de batimiento.

En la Figura 3 se puede observar que la curva correspondiente al modelo que incluye desprendimiento de vorticidad desde el borde de ataque predice un aumento en la sustentación bastante marcado respecto al modelo estándar. Este aumento se da, básicamente, durante cada fase traslacional (*downstroke-upstroke*) y hacia el final de cada una de estas fases, siendo la máxima diferencia obtenida en la fuerza de sustentación del 36,2 % y se produce exactamente al 38 % y al 86 % del ciclo de batimiento (se pueden observar dos picos bastante pronunciados en dichos puntos).

5. CONCLUSIONES

Se presentó una herramienta computacional, en desarrollo, muy versátil, basada en una ampliación y modificación del método de red de vórtices inestacionario y no-lineal en su versión tridimensional. El caso estudiado en este trabajo muestra el desprendimiento de vorticidad desde el borde de ataque tiene una incidencia significativa en el aumento de la fuerza de sustentación (máxima diferencia alrededor del 36 % hacia el final de cada fase traslacional). Este resultado es crucial porque permitiría explicar las grandes fuerzas de sustentación medidas experimentalmente sobre una gran variedad de insectos en vuelo libre.

Si bien el carácter de este trabajo es fundamentalmente cualitativo, permite demostrar que el modelo utilizado constituye un buen punto de partida para llegar a comprender de forma definitiva los mecanismos de vuelo utilizados por los insectos, como así también para combinar esta formulación con modelos de la dinámica estructural que permitan estudiar la aeroelasticidad del vuelo de insectos y aves pequeñas, y la aeroservoelasticidad de micro-vehículos aéreos de alas batientes inspirados en la biología.

REFERENCIAS

[1] C. VAN DEN BERG AND C. ELLINGTON, *The three-dimensional leading-edge vortex of a 'hovering' model hawkmoth.* Phil. Transaction Royal Society London B, 352 (1997), pp. 329-340.

[2] S.P. SANE AND M. DICKINSON, *The control of flight fore by a flapping wing: Lift and drag production*. The Journal of Experimental Biology, 204 (2001), pp. 2607-2626.

[3] M. DICKINSON, F.O. LEHMANN AND S.P. SANE, *Wing rotation and the aerodynamic basis of insect flight*. Science, 284 (1999), pp. 1954-1960.

[4] T. MARKOW AND P. O'GRADY, *Drosophila: A Guide to Species Identification and Use*, Elsevier Inc., San Diego, California, 2006.

[5] B. ROCCIA, S. PREIDIKMAN Y J. MASSA, *De la biología a los insectos robots: Desarrollo de un código computacional interactivo para estudiar la cinemática de alas batientes*. Mecánica Computacional, 27 (2008), pp. 3041-3058.

[6] S.P. SANE AND M. DICKINSON, *The control of flight fore by a flapping wing: Lift and drag production*. The Journal of Experimental Biology, 204 (2001), pp. 2607-2626.

[7] P. KONSTADINOPOULOS, D.T. MOOK AND A.H. NAYFEH, *A numerical method for general unsteady aerodynamics*. AIAA-81-1877. AIAA Atmospheric Flight Mechanics Conference, August 19–21, Albuquerque, New Mexico, 1981.

[8] S. PREIDIKMAN, *Numerical simulations of interactions among aerodynamics, structural dynamics, and control systems.* Ph.D. Dissertation, Department of Engineering Science and Mechanics, Virginia Tech, 1998.

TURBINAS EÓLICAS DE EJE HORIZONTAL Y GRAN POTENCIA: INCIDENCIA DE LA DIRECCIÓN DEL VIENTO Y LA CONICIDAD DEL ROTOR SOBRE LA POTENCIA GENERADA

Cristian Gebhardt[1], Sergio Preidikman[1,2] y Alejandro Brewer[1]

[1]*Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, cgebhardt@efn.uncor.edu, http://www.efn.uncor.edu*

[2]*Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Campus Universitario, Ruta Nacional 36 Km. 601, 5800 Río Cuarto, Argentina. Tel/Fax: 0358-4676246, spreidikman@ing.unrc.edu.ar, http://www.ing.unrc.edu.ar*

Resumen: En este trabajo se simula numéricamente el comportamiento aerodinámico no-estacionario y no-lineal de un generador eólico de eje horizontal y de gran potencia (Large Horizontal-Axis Wind Turbines o LHAWT). El rango de velocidades de operación y las grandes dimensiones del equipo dan lugar a flujos con altos números de Reynolds. Esto avala la hipótesis de considerar que los efectos viscosos están confinados únicamente a las capas límites y a las estelas. El comportamiento aerodinámico es modelado mediante la técnica de red de vórtices no-lineal y no-estacionaria, que permite determinar la magnitud y la evolución en el tiempo de las cargas aerodinámicas actuantes. Los resultados obtenidos ayudan a comprender como influyen en la eficiencia de un LHAWT, la dirección del viento y la conicidad del rotor.

Palabras claves: Turbinas Eólicas, Aerodinámica, Simulaciones, Potencia Producida.

1. INTRODUCCIÓN

La aerodinámica asociada a las LHAWT es inherentemente no-lineal y no-estacionaria debido a la presencia de condiciones ambientales complejas, vientos de amplitud y dirección cambiante, interacción aerodinámica entre el rotor y la torre portante, deformaciones estáticas, y flujos dominados por vorticidad. El rango de velocidades de operación (subsónico bajo) se conjuga con las grandes dimensiones del equipo dando lugar a flujos incompresibles cuyos números de Reynolds son altos. Este hecho avala la hipótesis de que los efectos viscosos están confinados únicamente a las capas límite y a las estelas vorticosas emitidas desde los bordes filosos de las palas y permite estimar las cargas aerodinámicas mediante una conocida técnica de la dinámica de fluidos: el método de red de vórtices no-lineal y no-estacionario, o NLUVLM (Non Linear Unsteady Vortex-Lattice Method).

2. EL MODELO AERODINÁMICO

En el modelo desarrollado se considera el flujo incompresible de un fluido no-viscoso. La velocidad absoluta de una partícula de fluido que se encuentra en la posición $\mathbf{R}$ en el instante t se denota como $\mathbf{V}(\mathbf{R};t)$. Debido a que el campo de velocidades del fluido es considerado irrotacional fuera de las capas límite y de las estelas, puede ser expresado como el gradiente del potencial total de velocidad $\Phi(\mathbf{R};t)$. La evolución espacial y temporal de este potencial esta gobernada por la ecuación de continuidad para flujos incompresibles.

$$\mathbf{V}(\mathbf{R};t) = \nabla\Phi(\mathbf{R};t) \qquad \nabla^2\Phi(\mathbf{R};t) = 0 \tag{1}$$

Para complementar la ecuación gobernante, deben considerarse un conjunto de condiciones de contorno (**CC**) [1,2]. La posición de la superficie del sólido es conocida como una función del tiempo, y la componente normal a esta superficie de la velocidad del fluido es prescripta sobre esta frontera. La primera CC requiere que la componente normal de la velocidad del fluido relativa a la superficie del sólido sea nula, lo que comúnmente se denomina condición de "no-penetración" o condición de "impermeabilidad" y se expresa como:

$$(\mathbf{V} - \mathbf{V}_S)\cdot\hat{\mathbf{n}} = (\nabla\Phi - \mathbf{V}_S)\cdot\hat{\mathbf{n}} = 0 \tag{2}$$

donde $\mathbf{V}_S$ es la velocidad de la superficie del sólido, y $\hat{\mathbf{n}}$ es el versor normal a dicha superficie. En general tanto $\mathbf{V}_S$ como $\hat{\mathbf{n}}$ varían de manera espacial y temporal. Además se debe imponer una condición de regularidad en el infinito, esta segunda CC requiere que las perturbaciones producidas en el fluido por la presencia del cuerpo (o cuerpos) disminuyan a medida que nos alejamos del mismo. Esta CC es conocida como condición de regularidad en el infinito y esta dado por:

$$\lim_{|\mathbf{R}|\to\infty}\left|\mathbf{V}(\mathbf{R};t)\right| = \lim_{|\mathbf{R}|\to\infty}\left|\nabla\Phi(\mathbf{R};t)\right| = \left|\mathbf{V}_\infty\right| \tag{3}$$

donde $\mathbf{V}_\infty$ es la velocidad de corriente libre sin perturbar. Debido a que el campo de velocidades es calculado utilizando la ley de Biot-Savart, la condición de regularidad en el infinito es satisfecha idénticamente. Para flujos potenciales incompresibles, el campo de velocidades es determinado a partir de la ecuación de continuidad, y por tal motivo la misma debe ser establecida independientemente de la presión. Una vez que el campo de velocidades es conocido, la presión es calculada mediante la versión no-estacionaria de la ecuación de Bernoulli. Además, al considerar que la velocidad del sonido es infinita, la influencia de las CC es transferida instantáneamente a todo el dominio de fluido; por lo tanto, el campo de velocidades instantáneo es obtenido a partir de las CC instantáneas. En adición a las CC, se utilizan los teoremas de Kelvin-Helmholtz y la condición no-estacionaria de Kutta para determinar la intensidad y la posición de las estelas [3].

La representación integral del campo de velocidades $\mathbf{V}(\mathbf{R};t)$ en términos del campo de vorticidad $\boldsymbol{\Omega}(\mathbf{R};t) = \nabla\times\mathbf{V}(\mathbf{R};t)$, es una extensión de la conocida ley de Biot-Savart, que tiene la siguiente forma:

$$\mathbf{V}(\mathbf{R};t) = \frac{1}{4\pi}\iint_{S(\mathbf{R}_0;t)} \frac{\boldsymbol{\Omega}(\mathbf{R};t)\times(\mathbf{R}-\mathbf{R}_0)}{\left|\mathbf{R}-\mathbf{R}_0\right|^3}\, dS(\mathbf{R}_0;t) \tag{4}$$

donde $\mathbf{R}_0$ es un vector posición en la región compacta $S(\mathbf{R}_0;t)$ del dominio del fluido. El argumento de la integral (4) es cero cuando $\boldsymbol{\Omega}(\mathbf{R};t)$ se anula, por esto la región en donde el fluido es irrotacional no produce ninguna contribución sobre $\mathbf{V}$. En cada punto, $\mathbf{V}$ puede ser computada explícitamente, e independientemente de la valuación en puntos vecinos. Como consecuencia de esta característica, que está ausente en métodos basados en diferencias finitas, la evaluación de $\mathbf{V}$ puede ser confinada a las regiones viscosas; "la distribución de vorticidad en las regiones viscosas determina el campo de fluido, tanto en la región viscosa como en la no-viscosa". Para formular la CC de "no-penetración" dada por la Ec. (2), es conveniente descomponer el potencial total de velocidades en tres partes: la primera debida a las sábanas vorticosas adheridas Φ_B, la segunda debida a las sábanas vorticosas desprendidas Φ_W, y la tercera debida a la corriente libre Φ_∞. Teniendo en cuenta esta descomposición del potencial total de velocidades, la Ec. (2) puede rescribirse como:

$$\left(\nabla\Phi_B + \nabla\Phi_W + \nabla\Phi_\infty - \mathbf{V}_S\right)\cdot\hat{\mathbf{n}} = 0 \tag{5}$$

2.1. EL MÉTODO DE RED DE VÓRTICES INESTACIONARIO

En el UVLM las sábanas vorticosas son reemplazadas por redes de segmentos vorticosos de longitud finita y circulación $\Gamma(t)$. Para conservar la circulación se utilizan anillos vorticosos cerrados de circulación $G(t)$; $\Gamma(t)$ es obtenida como suma vectorial de las circulaciones de los anillos adyacentes y $G(t)$ a partir de la condición de no-penetración.

Idealmente, sería preferible satisfacer la condición de no penetración en todos los puntos de la superficie del sólido, pero debido a que ésta se ha discretizado en un número NP de elementos, sólo es posible imponer la condición de no-penetración en un número finito de puntos, llamados puntos de control. Hay un punto de control en el centroide de los nodos de cada elemento. Para aproximar la normal a cada elemento se utiliza el producto vectorial de sus dos vectores diagonales. De la imposición de la condición de no-penetración, y su posterior manipulación algebraica, resulta un sistema de ecuaciones algebraicas lineales a coeficientes variables en el tiempo, cuyas incógnitas son las circulaciones $G_j(t)$ [3, 4].

Una vez calculadas las circulaciones, $G_j(t)$, se "convectan" las estelas. Los nodos que definen los extremos de cada segmento vorticoso en las estelas son transportados con la velocidad local del fluido, y su nueva posición se determina con la metodología desarrollada en las referencias [3] y [5].

2.2. CÁLCULO DE LAS CARGAS AERODINÁMICAS

Para calcular las cargas aerodinámicas sobre las superficies sustentadoras en cada elemento, se debe hallar el salto de presiones en el punto de control y luego multiplicarlo por el área del elemento y por el versor normal. Finalmente, se suman las fuerzas y los momentos de dichas fuerzas sobre todos los elementos. La presión en el punto de control de cada elemento se calcula mediante la ecuación de Bernoulli para flujos inestacionarios [3].

3. CONICIDAD DE LAS PALAS

La geometría del rotor de un generador eólico se define por medio de varios parámetros. Los parámetros geométricos más importantes son: la distribución de perfiles aerodinámicos, la distribución de ahusamiento y de alabeo a lo largo de la pala, y la conicidad del rotor respecto a un plano perpendicular al eje de rotación. La conicidad del rotor queda definida por medio del ángulo ξ, que es el ángulo que forma el eje longitudinal de la pala con un plano perpendicular al eje de rotación, ver Figura 1.

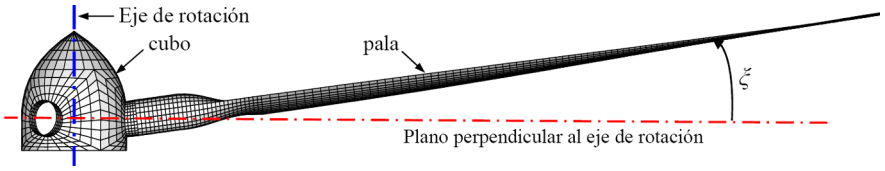


Figura 1: Ángulo de conicidad ξ de las palas de un rotor.

El ángulo de conicidad permite montar el rotor más cerca de la torre, ya que aleja las punteras de las palas evitando las colisiones que se podrían producir cuando las palas se flexionan por acción de las cargas aerodinámicas y pasan próximas a la torre. Debido a que el flujo asociado a un generador eólico es altamente complejo, es importante determinar de manera cualitativa y cuantitativa como influye sobre el comportamiento aerodinámico la variación de conicidad del rotor para diferentes condiciones de viento. Como parte del presente trabajo se estudia, ignorando la flexibilidad de las palas, como incide la conicidad de las palas del rotor en la configuración de rotor aislado.

4. INCIDENCIA DE LA DIRECCIÓN DEL VIENTO SOBRE LA POTENCIA

Los generadora eólicos operan en condiciones ambientales complejas debido a la presencia de vientos de amplitud y dirección cambiante, a los efectos de turbulencia y la existencia de la capa límite terrestre. Todo esto hace que la aerodinámica asociada a estos sistemas de generación de energía sea inherentemente no-lineal y no-estacionaria. En una primera aproximación, se puede adoptar un perfil de viento uniforme y despreciar los efectos de turbulencia y de la capa limite terrestre.

La potencia que puede generar un rotor, depende de la dirección del viento incidente. La velocidad efectiva, V_E, que capta el rotor para la producción de potencia en el eje es la proyección de la velocidad de corriente libre sobre el eje de rotación:

$$V_E = V_\infty \cos\alpha \tag{6}$$

donde α es el ángulo que forma la dirección del viento incidente respecto al eje de rotación. La potencia producida por el rotor es igual al producto escalar entre el momento aerodinámico $\mathbf{M}$ y la velocidad angular del rotor $\boldsymbol{\omega}$, por lo tanto:

$$P = \mathbf{M}\cdot\boldsymbol{\omega} = q_E \left(L_C\right)^3 C_M\, \omega \tag{7}$$

donde L_C es una longitud característica, C_M es el coeficiente de momento aerodinámico y q_E es la presión dinámica efectiva. Reemplazando la Ec. (7), en la expresión de q_E ,

$$q_E = \frac{1}{2}\rho\left(V_E\right)^2 \qquad \rightarrow \qquad q_E = \frac{1}{2}\rho\left(V_\infty\right)^2 \cos^2\alpha \tag{8}$$

se observa que la presión dinámica efectiva varía con el cuadrado del coseno de α. Reemplazando la Ec. (8), en la Ec. (7), se obtiene la siguiente expresión para el cómputo de la potencia obtenida en el eje del rotor:

$$P = \frac{1}{2}\rho\left(V_\infty\right)^2 \left(L_C\right)^3 \omega\, C_M \cos^2\alpha \tag{9}$$

Si la expresión dada en (9) se divide por P_0 (valor de potencia cuando $\alpha = 0$), se obtiene una versión adimensionalizada de la potencia obtenida en el eje del rotor, donde el cociente $C_M(\alpha)/C_M(0)$ es una función no lineal del ángulo α y que toma un valor próximo a la unidad cuando α es pequeño. Por lo tanto, en una primera aproximación, se puede estimar que la variación P/P_0 en función de α es igual a $\cos^2\alpha$.

$$\frac{P}{P_0} = \frac{C_M(\alpha)}{C_M(0)} \cos^2\alpha \tag{10}$$

5. RESULTADOS

A continuación se presentan resultados obtenidos con la herramienta computacional que está siendo desarrollada. En las simulaciones se consideró un rotor de tres palas de 70 m de diámetro rotando a 12 rpm y viento de frente con una velocidad de 20 m/s que se aplica sobre el rotor de manera impulsiva. El mismo rotor fue utilizado anteriormente por Gebhardt et al. [1, 2].

5.1. INFLUENCIA DE LA DIRECCIÓN DEL VIENTO SOBRE LA POTENCIA

Para estudiar la influencia de la dirección del viento sobre la potencia extraída se realizaron simulaciones variando la dirección del viento incidente respecto al eje del rotor. Las simulaciones fueron realizadas para una misma velocidad del viento pero variando los valores del ángulo de incidencia α desde 0 hasta 45º, con incrementos de 5º. La potencia producida por el rotor en función de α fue normalizada respecto de la potencia correspondiente a $\alpha = 0$º. Los resultado se muestran en la Figura 2a, donde además fue graficada la función $\cos^2\alpha$ con la finalidad de mostrar la incidencia del factor $C_M(\alpha)/C_M(0)$ en la potencia adimensionalizada. Se advierte que para valores de α menores a 15º, la potencia adimensionalizada varía como $\cos^2\alpha$; esto significa que el efecto de las no-linealidades sobre $C_M(\alpha)$ es pequeño. En cambio, para valores de α mayores a 15º los efectos no-lineales comienzan a ser importantes; notar que en esa zona los valores predichos por las simulaciones se separan bastante de la curva correspondiente a la función $\cos^2\alpha$.

5.2. INFLUENCIA DE LA CONICIDAD DE LAS PALAS SOBRE LA POTENCIA

Se realizaron simulaciones, variando ξ entre –15º y +15º con incrementos de 5º, para una condición de viento incidente de dirección y magnitud fijas, $V_\infty = 20$ m/s y $\alpha = 0$º. En la Figura 2b se muestra la variación de la potencia obtenida en función del ángulo de conicidad del rotor. Se observa una curva sesgada hacia la izquierda que alcanza su máximo cuando $\xi = -3{,}5$º, valor carente de importancia práctica por lo comentado anteriormente en la sección 3. En la Figura 2c se ha adimensionalizado la potencia para mostrar la importante pérdida debida a la conicidad del las palas (P_0 es la potencia cuando $\xi = 0$). Se observa que cuando la conicidad es de 7º, la potencia es un 4 % menor que la correspondiente a $\xi = 0$. Este resultado confirma que la aerodinámica de los rotores es dependiente de la configuración geométrica.

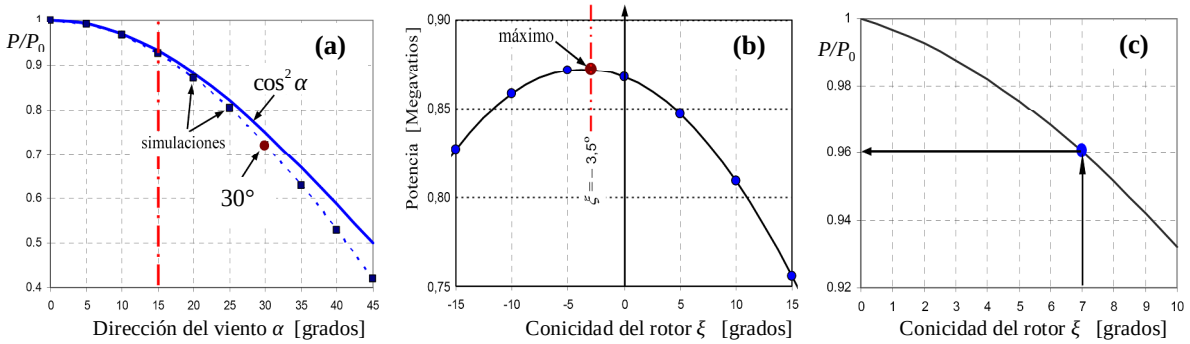


Figura 2: Potencia obtenida en función de la dirección del viento (a) y de la conicidad (b y c).

6. CONCLUSIONES

En este trabajo se presentaron y analizaron resultados obtenidos con una herramienta computacional que esta siendo desarrollada para predecir, en el dominio del tiempo, el comportamiento aerodinámico no-lineal de generadores eólicos de eje horizontal y de gran potencia. Es posible afirmar que las cargas aerodinámicas son fuertemente dependientes de la dirección de la corriente de viento, y la potencia producida se reduce cuando aumenta el ángulo de incidencia respecto del eje del rotor. Se demostró que para ángulos de hasta 15º la disminución de potencia sigue la ley del coseno cuadrado del ángulo de incidencia. También ha sido posible explicar, de forma cuantitativa, como influye la conicidad del rotor en la potencia producida.

REFERENCIAS

[1] C. GEBHARDT, S. PREIDIKMAN, J. MASSA Y G. WEBER, *Comportamiento aerodinámico y aeroelástico de rotores de generadores eólicos de eje horizontal y de gran potencia*. Mecánica Computacional 27, (2008), pp. 519-539.

[2] C. GEBHARDT, S. PREIDIKMAN Y J. MASSA, *Simulaciones numéricas del comportamiento aerodinámico de generadores eólicos de eje horizontal y de gran potencia*. 2º Congreso Iberoamericano Hidrógeno y Fuentes Sustentables de Energía. San Juan, Arg., 2009.

[3] S. PREIDIKMAN, *Numerical simulations of interactions among aerodynamics, structural dynamics, and control systems*. Ph.D. Thesis, Virginia Polytechnic Institute and State University, 1998.

[4] J. KATZ AND A. PLOTKIN, *Low-speed aerodynamics*, Cambridge University Press, 2001.

[5] O.A. KANDIL, D.T. MOOK AND A.H. NAYFEH, *Nonlinear prediction of the aerodynamic loads on lifting surfaces*, Journal of Aircraft, Vol. 13 (1976), pp. 22-28.

IMPROVED DISCRETE NON-LOCAL ABSORBING BOUNDARY CONDITION FOR HELMHOLTZ EQUATION. APPLICATIONS TO UNBOUNDED WAVE PROBLEMS

Ruperto P. Bonet[♭], Carlos Zuppa[♭] and Gloria Simonetti[♭]

[♭] *Department of Mathematics, National University of San Luis, Chacabuco y Pedernera (5700), San Luis, Argentina,rpbonet@unsl.edu.ar, www.unsl.edu.ar*

Abstract: We propose a new class of approximate non-local DNL boundary conditions to be applied on exterior boundaries when solving Helmholtz equation with variable coefficients in unbounded domain. This formulation is an extension of the construction of the classical DNL boundary condition [1],[2] to the anisotropic medium, by means of the addition of a term in the non-local relation, that computes the errors on the artificial boundary. This term can be added in two ways, as a source term or by means of an additional layer. We investigate numerically the effect of the frequency regime on the accuracy of these conditions. We also compare their performance to the classical DNL boundary conditions and to the second-order absorbing boundary condition designed by Bayliss, Gunzburger and Turkel (BGT2). Numerical experiments show clearly the superiority of the proposed methodology.

Keywords: *absorbing boundary condition, Helmholtz, anisotropic medium*
2000 AMS Subject Classification: 65M60 - 65M85

1 INTRODUCTION

Given the Helmholtz operator $\mathcal{L} = -k^2(x, y) - \Delta$ defined on a given domain $\Omega \subset \mathbb{R}^2$ we wish to solve the elliptic partial differential equation $\mathcal{L}(\phi) = 0$ in an unbounded domain Ω with appropriate boundary conditions. Analytical solutions have been reported in the past for regular geometries, however, despite their simplicity, this class of problems is not completely solved, particularly from a numerical point of view. Efforts to design a transparent boundary condition continue being a task problem in nowadays [3],[4],[5]. The present approach is to model the entire domain using finite elements on a certainly bounded region and the **DNL** method on a unbounded domain. This methodology for a finite element method on an unbounded domain is based on finding the general solution of a recurrence relationship involving the complete eigen-decomposition of the discretizated operator over an structured mesh with quadrangular elements. In this work, a non-homogeneous non-local recurrence relationship is designed, in contrast to the classical DNL methodology, and using linear and/or quadratic elements, a higher order non-local relation is also derived.

The present methodology have been applied successfully to exterior Helmholtz problems, in problems of radiation and scattering of waves. The present paper includes a review on the description of the **DNL** method in rectangular coordinates and/or circumferential coordinates, and its extensions to non-homogeneous case or higher-order case, respectively. Numerical solutions for several unbounded wave problems are obtained by means of the application of finite elements with the new approach of the **DNL** boundary condition.

2 FINITE ELEMENTS IN UNBOUNDED DOMAIN

2.0.1 Residual DNL method with one layer (RDNL1)

Using linear elements, we obtain the discrete Helmholtz equation for the layer j in the form:

$$C^j \phi^{j-1} + B^j \phi^j + A^j \phi^{j+1} = 0 \tag{1}$$

where ϕ^j is the vector containing scattered potential values for the nodes belonging to the j layer. One way solution of it can be given by the following non-homogeneous transmission relation

$$(\phi^+)^{j+1} = F^j (\phi^+)^j + R^j \tag{2}$$

where R^j vector is the source term relative to the j-layer. Substituting this relation into equation (1), are obtained the following formulas

$$F^{j-1} = -(A^j F^j + B^j)^{-1} C^j \tag{3}$$

$$R^{j-1} = -(A^j F^j + B^j)^{-1} A^j R^j \tag{4}$$

from $j = M, M-1, \cdots, 1$. The second formula is new in the framework of the condensation process, and it allows to calculate the numerical errors at each j-layer. Applying recursively these formulas from the far-field to the near-field is obtained the modified DNL with a corrector term, at the near-field. Then, the classical DNL formula adopts the new form

$$(\phi^+)^2 = F^1 (\phi^+)^1 + R^1 \tag{5}$$

at the near-field. To implement this recursive process we calculate the initial value vector R^M, at the far-field.

2.0.2 Two-layers DNL method (DNL2)

Using a piecewise quadratic basis functions system at the main direction to infinity, a higher order difference scheme is derived, and then, the second order discrete equations system (1) is replaced by the discrete equations system of fourth order

$$D^j \phi^{j-2} + C^j \phi^{j-1} + B^j \phi^j + A^j \phi^{j+1} + E^j \phi^{j+2} = 0 \tag{6}$$

corresponding to the j-layer of the semi-discrete Helmholtz equation. To achieve a higher precision of the numerical solution at the near-field a non-local homogeneous transmission relation with two layer is proposed to close the computational domain at the artificial boundary:

$$(\phi^+)^{j+1} = F^j (\phi^+)^j + G^j (\phi^+)^{j-1} \tag{7}$$

where the F^j and G^j are matrices relative to j-layer, that characterize the eigendecomposition of the discrete Helmholtz operator given in (6), and collect the outgoing modes that indicate the outer flux to the open boundary.

3 Numerical Results

We consider the exterior model problem

$$\nabla(h\nabla\phi) + k^2 h\phi = 0 \; in \Omega = (r \geq r_a, 0 \leq \theta \leq 2\pi), \tag{8}$$

$$\frac{\partial \phi}{\partial r} = 0 \; on \; (r = r_a), \tag{9}$$

$$\lim_{r\to\infty} \sqrt{r}(\frac{\partial \phi}{\partial r} - \mathrm{i}k)\phi = 0 \tag{10}$$

where $k(r \geq r_b) = k_0$ and $h(r,\theta) = max(h_a, min(\alpha * r^2, h_b))$ with k_0, h_a, h_b, α positive constants predefined. Here ∇ is the gradient operator and $k \in \mathcal{C}$ is the wave number, $Imk \geq 0$; $\mathrm{i} = \sqrt{(-1)}$ is imaginary unit. Equation (10) is the Sommerfeld radiation condition and allows only outgoing waves proportional to $exp(\mathrm{i}k_0 r)$. The radiation condition requires that energy flux at infinity be positive, thereby guaranteeing that the solution to the boundary problem (8)-(10) is unique. An appropriate representation of this condition is crucial to the reliability of any numerical formulation of this problem. On this paper we have presented two new representation of the (10) at the discrete level, which are imposed on chosen circular boundary to close the computational domain. For this numerical test we consider the geometrical relations $\frac{h_b}{h_a} \approx 10$ and $\frac{r_b}{r_a} = 3$, and a structured mesh with 20 elements by wavelength on the annular region. The computational domain is limited by the circumference of radius r_c very close to r_b, such that $\frac{r_c}{r_b} \approx 1.07$.

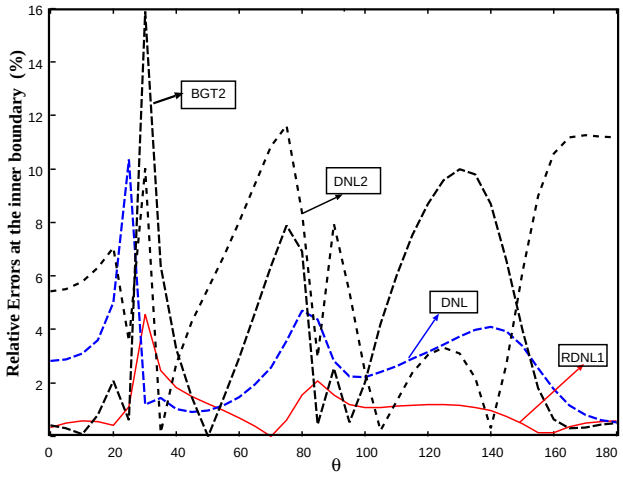


Figure 1: (%) Relative Errors at the inner boundary $r = r_a$ vs. Azimuth θ

With the objective of an evaluation of the numerical performance of the proposed methodology, we compute the relative errors between the analytical solution and the numerical solutions on the inner boundary $r = r_a$. Figure 1 shows the performance of the new class of absorbing boundary relative to known absorbing boundaries. In the figure can be noticed that $RDNL1$ reduces drastically the relative errors from the classical DNL, and of an uniform manner. Also can be observed that the $DNL2$ has a little better performance than a $BGT2$ absorbing boundary, in fact, it is according with expected behaviour, taking into account that the $DNL2$ is computed directly, without the application of the recursive process. The numerical results indicate a progress in the application of the DNL fomulation to develop absorbing boundaries conditions to Helmholtz equation in anisotropic medium.

4 Conclusions

Discrete Non-Local (DNL) boundary conditions for unbounded wave problems in two dimensions are derived and analized, defining problems that are suitable for finite element analysis and its generalizations. Two strategies to improve the classical DNL absorbing boundary conditions have been proposed. Numerical results show that the absorbing boundary derived with a recurssive process reduce drastically the errors at the open boundary. We propose the use of the Residual DNL methodology to solve unbounded wave guides problem in the framework of FEM. Higher-order DNL absorbing boundary is in progress with the design of the Two-layers DNL absorbing boundary. Also, a parallel version of these methods is in advance.

Acknowledgments

This work is supported by the Project of Faculty of Physics, Mathematics and Natural Sciences from National University of San Luis (Argentina) under the grant **PROICO** 22/F730O. We made extensive use of software distributed by GNU Fortran, the Free Software / GNU-Proyect:Linux ELF-OS, Octave, Tgif from William C.Cheng, and others.

5 References

References

[1] R.P. Bonet, N. Nigro, M.A. Storti, and S.R. Idelsohn, *A discrete non-local (DNL) outgoing boundary condition for diffraction of surface waves*,CNME, 14 (1998), pp.849-861.

[2] R.P. Bonet, N. Nigro, M.A. Storti, and S.R. Idelsohn, *Discrete Non-local absorbing discrete boundary condition for exterior problems governed by Helmholtz equation*,IJNMF, 29, (1999), pp.605-621.

[3] R.P. Bonet, *Condiciones absorbentes locales para la ecuacion de Berkhoff sobre una frontera de forma general*, RIMNE, 21,1, (2005), pp.103-133.

[4] Eliane Bécache, Dan Givoli, Thomas Hagstrom, *High-order Absorbing Boundary Conditions for anisotropic and convective wave equations*,JCP,229,(2010),pp.1099-1129.

[5] ISAAC HARARI, RABIA DJELLOULI, *Analytical study of the effect of wave number on the performance of local absorbing boundary conditions for acoustic scattering*,Appl. Num. Math., 50, (2004), pp.1547.

MULTI-CRACK IDENTIFICATION IN DAMAGED THIN-WALLED BEAMS BY MEANS OF VIBRATION ANALYSIS

Franco E. Dotti and Víctor H. Cortínez

Centro de Investigaciones de Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional FRBB, 11 de Abril 461, B8000LMI, Bahía Blanca, Argentina, vcortine@frbb.utn.edu.ar
Consejo Nacional de Investigaciones Científicas y Tecnológicas, Argentina.

Abstract: In the present article, a theoretical model for the dynamic analysis of damaged thin-walled beams, previously introduced by the authors, is employed in order to perform the identification of damage parameters, considering the presence of more than one flaw. The model incorporates bending and warping effects of shear flexibility by means of a linearized formulation based on the principle of virtual work. A seven degree-of-freedom per node finite element is employed in the discretization of governing equations. Damage is considered by modifying the sectional properties of a single finite element having an appropriate length. In order to perform identification of failure parameters, damage is treated as fatigue cracks located in a boundary of the beam cross-section. Location and depth of the cracks are identified by means of the minimization of a target function, defined in terms of differences among natural frequencies calculated with the model and experimental values.

Key words: thin-walled multi-crack damage identification
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCTION

In the present article, the dynamic behavior of cracked thin-walled beams is analyzed by means of a beam model formulation accounting for shear flexibility due to bending and warping, originally introduced by Cortínez and Rossi [3] for undamaged beams. A two-node finite element with seven degrees-of-freedom per node is used to discretize the governing equations. Damage is taken into account as a fatigue crack located at a boundary of the beam cross-section, since it represents a common failure in beam like structures. This kind of flaw is introduced by modifying the sectional properties of a single finite element, i.e. the presence of damage is modeled as a geometrical imperfection. The length of the modified element is obtained by elastic energy comparisons between the present beam model and a Fracture Mechanics model developed by Cortínez et al. [2].

Location and depth of fatigue cracks are identified by minimizing a target function defined by differences between natural frequency results of the present beam model and results of numerical experiments, with an ABAQUS' shell model. Differential Evolution (DE) [6] is employed to perform the optimization calculations.

2. THEORY

2.1. DISPLACEMENT FIELD

Figure 1 shows a sketch of a thin-walled beam with the presence of damage. The reference point C is coincident with the center of gravity (CG) while the point O corresponds to the shear center (SC) of the undamaged cross-section. B is a generic point in the middle line of the cross-sectional wall. The coordinates corresponding to points lying on this middle line are denoted as $\bar{Y}(s)$ and $\bar{Z}(s)$ (or $Y(s)$ and $Z(s)$). The present thin-walled beam theory is based on the following two assumptions: 1) The cross-section contour is rigid in its own plane, although it is free to warp out of it, and 2) The torsional warping distribution is assumed to be given by the Saint-Venant function. According to these hypotheses, the displacement field is assumed to be in the following form [3]

$$u_x = u - \bar{y}\theta_z - \bar{z}\theta_y + \omega\theta_x, \tag{1a}$$

$$u_y = v - z\phi_x, \tag{1b}$$

$$u_z = w + y\phi_x, \tag{1c}$$

where ω is the warping function [3], u, v and w are the displacements of the CG in x, y and z directions, respectively, θ_y and θ_z are bending twists, ϕ_x is the torsional twist and θ_x, the warping variable.

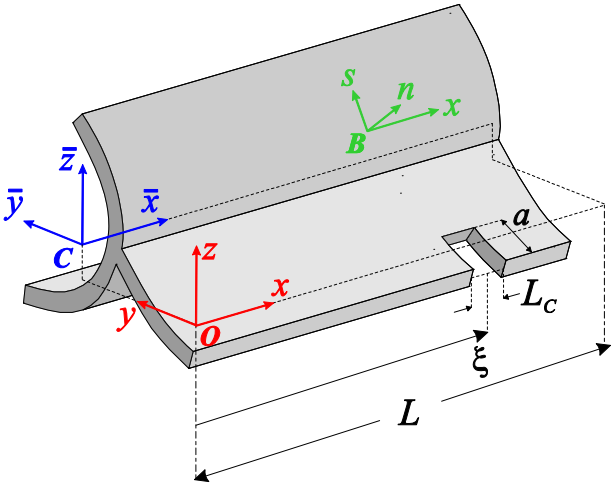


Figure 1: Generic damaged thin-walled beam and basic associated coordinate systems.

Substituting expressions (1) into the general expression of virtual work, and then integrating with respect to y and z, one can obtain the one-dimensional variational equation of motion (See reference [3]).

2.2. CONSTITUTIVE EQUATIONS

Structural damage is regarded as a geometrical imperfection. The cross-section of a segment of length L_C is modified (Fig. 1) in order to consider the presence of damage. The origins of the considered coordinate systems are not coincident with the CG and the SC of the damaged cross-section. Taking this into account, the constitutive equations for the damaged segment can be obtained in a classic way from the constitutive functional of the principle of virtual work. Therefore, they may be expressed as

$$\{Q_E\} = [J_E]\{\Delta\}, \qquad \{Q_E^{(c)}\} = [J_E^{(c)}]\{\Delta\}, \tag{2}$$

where $\{Q_E\}$ and $\{Q_E^{(c)}\}$ are the vectors of generalized beam forces corresponding to undamaged and damaged cross section, respectively, and $\{\Delta\}$ is the vector of generalized displacements. The cross-sectional properties of the beam are contained into the beam constitutive matrices, $[J_E]$ and $[J_E^{(c)}]$.

2.3. FATIGUE DAMAGE MODELING

The length L_C must be chosen in order to produce a behavior analog to the presence of a real damage. For the case of a fatigue crack, L_C must represent the behavior of the beam having a crack with depth a and location ξ and it is determined by elastic energy comparisons between the present beam model and a Fracture Mechanics model recently developed [2]. The strain energy of the beam of Fig. 1 is given by

$$U = \frac{1}{2}\int_0^{\xi-\frac{L_C}{2}} \left(\{Q_E\}^T [J_E]^{-1} \{Q_E\}\right) dx + \frac{1}{2}\int_{\xi-\frac{L_C}{2}}^{\xi+\frac{L_C}{2}} \left(\{Q_E^{(c)}\}^T [J_E^{(c)}]^{-1} \{Q_E^{(c)}\}\right) dx + \frac{1}{2}\int_{\xi+\frac{L_C}{2}}^{L} \left(\{Q_E\}^T [J_E]^{-1} \{Q_E\}\right) dx. \tag{3}$$

Solving the integrals with respect to x, Eq. (3) may be expressed as $U = U_I + U_{II} + U_{III}$, where U_I is the strain energy associated to the mode I of fracture and therefore to the axial force N, bending moments M_y, M_z and bimoment B. Griffith's criterion allows expressing the stress intensity factor of mode I, K_I, as

$$K_I(a,\xi,L_C)=\sqrt{\frac{eE}{1-(\nu)^2}\frac{\partial U_I}{\partial a}}, \tag{4}$$

being e the thickness of the beam. The factor K_I predicted by the model depends on the severity of damage, a, but also on its equivalent length L_C and its location ξ [1]. Comparing Eq. (4) with expression presented in [2], a discrete target function, F_K, can be defined as

$$\min\left[F_K=\sum_{j=1}^{n_\xi}\sum_{i=1}^{n_a}\sqrt{\left(\frac{K_I(a_i,\xi_j,L_C)-K_I^t(a_i,\xi_j)}{K_I^t(a_i,\xi_j)}\right)^2}\right], \tag{5}$$

n_a and n_ξ are the number of depths and locations employed in the optimization calculation.

3. DAMAGE IDENTIFICATION

Identification of damage parameters is performed by comparisons between experimental values of natural frequencies and those predicted by the beam model. Normalized vectors containing depths and locations of the cracks are defined as $\Lambda = \{a_1/b, a_2/b,\ldots, a_n/b\}$ and $X = \{\xi_1/L, \xi_2/L,\ldots, \xi_n/L\}$, where n is the number of existing cracks. A natural frequency k predicted by the model is denoted as $F^{(k)}(X,\Lambda)$. Components of X and Λ are identified by the minimization of a target function $T(X,\Lambda)$, that is

$$\min\left[T(X,\Lambda)=\frac{100}{q}\sum_{k=1}^{q}\sqrt{\left(\frac{F^{(k)}(X,\Lambda)-f^{(k)}}{f^{(k)}}\right)^2}\right], \tag{5}$$

where q is the number of natural frequencies employed in the calculation and $f^{(k)}$ represents an experimental measurement of the natural frequency k.

4. NUMERICAL RESULTS

4.1. ACCURACY OF THE BEAM MODEL

The potential of the beam model to reproduce the dynamic behavior of a thin-walled beam with the presence of two cracks is evaluated. Experimental results are simulated numerically with shell finite element models. Comparisons are performed for steel beams with the following properties: E = 210 GPa, G = 80.76 GPa, ν = 0.3 y ρ = 7830 kg/m^3.

Table 1: Comparisons of the first four natural frequencies of a cantilever thin-walled U beam with two cracks. Dimensions: b = 0.2 m, h = 0.2 m, e = 0.01 m, L = 2 m, L/L_C = 54.

X_1	X_2	Λ_1	Λ_2	$F^{(1)}$	$f^{(1)}$	$F^{(2)}$	$f^{(2)}$	$F^{(3)}$	$f^{(3)}$	$F^{(4)}$	$f^{(4)}$
0.25	**0.50**	**0.50**	**0.75**	21.04	20.52	30.20	30.03	70.29	69.53	105.01	104.49
0.25	**0.75**	**0.50**	**0.75**	22.08	21.61	30.09	29.95	72.52	71.78	119.45	118.68
0.12	**0.50**	**0.50**	**0.75**	20.63	19.95	30.05	29.81	69.57	68.75	101.45	98.09
0.25	**0.50**	**0.50**	**0.50**	21.97	21.50	30.34	30.26	72.21	71.58	125.04	123.37
0.25	**0.50**	**0.25**	**0.25**	22.87	22.73	30.39	30.38	74.09	73.93	129.19	128.62

Table 1 shows comparisons of the first four natural frequencies for a cantilever U beam. This beam corresponds to a sufficiently general case, involving bending-torsional couplings. Taking into account the totality of the cases in Table 1, maximum average frequency error is 2.21%. That is, average errors are

sufficiently small to allow identification of damage parameters, if a threshold of 5% is required for a good detection [4], [5].

4.2. IDENTIFICATION OF DAMAGE PARAMETERS

Minimization of the target function $T(X,\Lambda)$ is performed in order to identify damage parameters X and Λ, for two cracks. The first four natural frequencies are employed ($q = 4$) and the variables are set with the following restrictions: $0 \leq X \leq 1$, $0 \leq \Lambda \leq 1$. DE [6] is employed in the optimization: spread constant is set to 0.6, the total number of parameter vectors is set to 20 and the cross probability is set to 0.5.

Table 2: Identification of damage parameters for a cantilever thin-walled U beam with two cracks. Dimensions: $b = 0.2$ m, $h = 0.2$ m, $e = 0.01$ m, $L = 2$ m, $L/L_C = 54$.

Experimental				Estimation				Error (%)			
Location		Depth		Location		Depth		Error in Location		Error in Depth	
X_1	X_2	Λ_1	Λ_2	X_1	X_2	Λ_1	Λ_2	X_1	X_2	Λ_1	Λ_2
0.25	**0.50**	**0.50**	**0.75**	0.29	0.58	0.61	0.77	4.00	8.00	11.00	2.00
0.25	**0.75**	**0.50**	**0.75**	0.27	0.76	0.58	0.78	2.00	1.00	8.00	3.00
0.12	**0.50**	**0.50**	**0.75**	0.21	0.60	0.59	0.79	9.00	10.00	9.00	4.00
0.25	**0.50**	**0.50**	**0.50**	0.20	0.45	0.50	0.58	-5.00	-5.00	0.00	8.00
0.25	**0.50**	**0.25**	**0.25**	0.22	0.48	0.28	0.35	-3.00	-2.00	3.00	10.00

Some results are shown in Table 2. Maximum errors are in the order of 10% for location and depth of both cracks. Computation time for each iteration is in the order of 27 seconds, employing an AMD Athlon 64 5200+ processor, with 3 Gb of RAM and an ASUS M2N-MX-SE+ motherboard.

5. CONCLUSIONS

In this article, a simplified theoretical beam model that simulates the dynamic behavior of thin-walled damaged beams is tested considering the presence of more than one crack. Damage is considered by modifying the sectional properties of a single finite element having an appropriate length, which can be estimated in terms of elastic energy comparisons. In terms of natural frequencies, the present approach is consistent with numerical results obtained from higher order models.

The model is employed in the identification of damage parameters, by comparisons with experimental measures. Natural frequencies are employed as indicators, and Differential Evolution algorithm is used in optimization calculations. Errors are acceptable in estimation of location and depth of two cracks.

ACKNOWLEDGEMENTS

The authors would like to thank the support of Secretaría de Ciencia y Tecnología of Universidad Tecnológica Nacional and CONICET. The present article is part of the doctoral thesis by Franco Dotti, under the direction of Víctor Cortínez and Marcelo Piovan, at the Engineering Department of Universidad Nacional del Sur.

REFERENCES

[1] V.H. CORTÍNEZ, AND F.E.DOTTI, *Un modelo Numérico para la Dinámica de Vigas de Pared Delgada Fracturadas por Fatiga: Aplicación a la Identificación de Daño*, Mecánica Computacional, 29 (2010), pp. 431-448.

[2] V.H. CORTÍNEZ, F.E.DOTTI, AND M.T. PIOVAN, *Factor de Intensidad de Tensiones del Modo I para Vigas Abiertas de Pared Delgada*, Mecánica Computacional, 28 (2009), pp. 955-971.

[3] V.H. CORTÍNEZ, AND R.E. ROSSI, *Dynamics of Shear Deformable Thin-Walled Open Beams Subjected to Initial Stresses*, Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, 14(3) (1998), pp. 293-316.

[4] F.E. DOTTI, M.T. PIOVAN, AND V.H. CORTÍNEZ, *Vibration of Damaged Thin-Walled Beams*, Submitted.

[5] O.S. SALAWU, *Detection of Structural Damage through changes in Frequency: a review*, Engineering Structures, 19(9) (1997), pp. 718-723.

[6] R. STORN, AND K. PRICE, *Differential Evolution – A Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces*, Journal of Global Optimization, 11 (1997), pp. 341-359.

A FRACTURE CURVE GENERATED BY NEAREST FLAWS

Gonzalo Hernandez†,‡ and Robert León‡

†School of Industrial Engineering, Universidad de Valparaíso, Las Heras 06, Valparaíso, Chile, gjho@vtr.net
‡Valparaíso Center for Science and Technology, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso, Chile, roberto.leon@gmail.com

Abstract: In this work, we study a recursive algorithm for constructing a two dimensional fracture curve by visiting the nearest flaws of thee material. Our main result establishes that the algorithm is very fast and generates realistic visualizations of fracture curves.

Key words: algorithm, fracture curve, nearest flaw.
2000 AMS Subjects Classification: 74R10.

1. INTRODUCTION

Different natural and artificial high energy fragmentation processes exhibit power-law behavior for small fragment masses, see refs. [14,16,21,22]. This property has been tried to explain by several discrete and continuous fragmentation models. In the first case, the discrete characteristic of the models is defined by how the material is represented, e.g.: cellular automata, two- and three-dimensional square lattices, square lattices composed of discrete masses connected by elastic beams, see refs. [1,2,6,7,10,17,20]. In the case of continuous processes of fragmentation, numerous one- and two-dimensional models have been studied using the rate equations [18]; see for instance refs. [3,4,11,12,13,19]. Additionally, recursive and self-similar continuous fragmentation $1d$ and $2d$ process have been also studied both analytically and numerically, see refs [5,8,9,15]. For these models must be defined: how the material is modeled, how the forces are computed, which is the fracture mechanism, how the fracture curve is computed, which is the stopping condition. In ref. [9] a multiple fragmentation model was studied under these assumptions: continuous bi-dimensional material with q random point flaws; uniform and independent random distribution of the net forces; linear fracture curve, every fragment fracture stops with constant probability p. By medium-scale simulations, it was obtained an approximate power law for the fragment size distribution with an exponent that varies in the range [1.01,1.15] and depends on the stopping probability. The visualizations show patterns of fracture that simulate real fragmentation processes. Despite that these results agree with the experimental evidence, this model presents the limitation that linear cuts model the fracture curve, which diminishes the realism of the visualizations, see ref. [9].

For this reason, in this short paper we study numerically an algorithm for fast computing of fracture curves generated by visiting the nearest flaws.

2. DEFINITION OF THE ALGORITHM AND NUMERICAL RESULTS

The proposed algorithm for constructing a fracture curve is a part of a fragmentation model, whose main assumptions are:

1) Material and flaws: The initial fragment is a continuous square with its linear size equal to one. Its defects are modeled by means of q point flaws that interact with the fragmentation process.
2) Dynamic fragmentation: Each neighboring i_k fragment apply an orthogonal force on a random point of the boundary of fragment k defined by $f_{i_k,k} = \int_{\partial B_{i_k,k}} d\sigma$, where $\partial B_{i_k,k}$ is the common boundary between

fragment k and its neighbor i_k. This definition of the fracture force is an intuitive approximation of the real fracture force. The fragment k is broken as a result of the n larger forces applied.

3) The initial point of the fracture curve begin in the point of application of the larger forces and then is computed recursively by visiting the nearest flaws contained in the fragment k, according to the following algorithm:

Stage 0. Initializations:	Generate a two dimensional random polygon in $[0,1]\times[0,1]$ Generate q random point flaws. Choose an initial point IP of the fracture curve: The point of application of the larger force. Define $P = IP, C = \{P\}, 5 \le n \le 15$
Stage 1. Initial Cut	For *i=1* to *n* Compute the nearest point flaw of P : $npf(P)$ Update $C = \{P\} \bigcup npf(P)$ Update $P = npf(P)$ End
Stage 2. Main Loop:	While (condition) Compute the nearest point flaw of P : $npf(P)$ For *i=1* to *n* Compute the i-nearest point flaw of P : $npf(P)$ If $Vec(P, nfp(P))$ cross the fracture curve break End condition = distance$(P,$ any border$) \le \varepsilon$ End

Our main result establishes that the previous algorithm compute very realistic fracture curves. In figure 1 we show simulations in different fragment geometries to prove this affirmation.

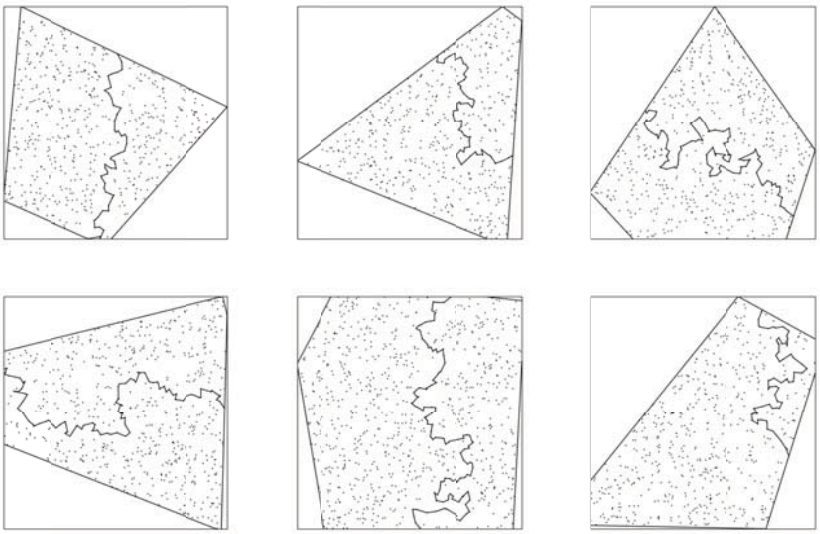


Figure 1: Examples of fractures curves in different fragment geometries.

The algorithm proposed to approximately compute the fracture curve is $O(n)$, where n is the number of flaws contained in the fragment. The visualizations of figure 1 in different fragment shapes (rhomboidal and polygonal) allows to propose the previous algorithm as generalization of linear fracture planes, see refs. [8,9].

ACKNOWLEDGEMENTS

The authors acknowledge the support of the Valparaíso Center for Science and Technology, BASAL Financing Program for Scientific and Technological Centers of Excellence. In this project Gonzalo Hernández is an Associate Researcher and Robert León a thesis student. In addition, Roberto León acknowledges CONICYT for a doctoral fellowship.

REFERENCES

[1] ASTRÖM, J. A., B. L. HOLIAN, AND J. TIMONEN, 2000, UNIVERSALITY IN FRAGMENTATION, PHYSICAL REVIEW LETTERS 84, 14, 3061-3064.
[2] ASTRÖM, J. A., R. P. LINNA, J. TIMONEN, P. F. MØLLER, AND L. ODDERSHEDE, 2004, EXPONENTIAL AND POWER-LAW MASS DISTRIBUTIONS IN BRITTLE FRAGMENTATION, PHYSICAL REVIEW E 70, 2, 026104-026110.
[3] BEN-NAI, E., P.L. KRAPIVSKY, 1997, MULTISCALING IN FRAGMENTATION, PHYSICA D, 107, 2-4, 156-160.
[4] BEN-NAI, E., P. L. KRAPIVSKY, 2000, FRAGMENTATION WITH A STEADY SOURCE, PHYSICS LETTERS A, 275, 1-2, 48-53.
[5] DOS SANTOS, 2007, F.P.M., R. DONANGELO AND S.R. SOUZA, SCHEMATIC MODELS FOR FRAGMENTATION OF BRITTLE SOLIDS IN ONE AND TWO DIMENSIONS, PHYSICA A, 374, 2, 680-690.
[6] HERNANDEZ, G. H.J. HERRMANN, 1995, DISCRETE MODELS FOR 2- AND 3-DIMENSIONAL FRAGMENTATION, PHYSICA A, 215, 420-430.
[7] HERNANDEZ, G., 2001, DISCRETE MODEL FOR FRAGMENTATION WITH RANDOM STOPPING, PHYSICA A, 300, 1-2, 13-24.
[8] HERNANDEZ, G., 2003, TWO-DIMENSIONAL MODEL FOR BINARY FRAGMENTATION PROCESS WITH RANDOM SYSTEM OF FORCES, RANDOM STOPPING AND MATERIAL RESISTANCE, PHYSICA A, 323, 1, 1-8.
[9] HERNANDEZ, G., L. SALINAS AND A. AVILA, 2006, N-ARY FRAGMENTATION MODEL WITH NEAREST POINT FLAW AND MAXIMAL NET FORCE FRACTURE, PHYSICA A, 370, 2, 565-572.
[10] KORSNES, R., S.R. SOUZA, R. DONANGELO, A. HANSEN, M. PACZUSKI, K. SNEPPEN, 2004, SCALING IN FRACTURE AND REFREEZING OF SEA ICE, PHYSICA A, 331, 291-296.
[11] KRAPIVSKY, P. L., E. BEN-NAI, 1994, SCALING AND MULTISCALING IN MODELS OF FRAGMENTATION, PHYSICAL REVIEW E, 50, 2, 3502-3507.
[12] KRAPIVSKY, P. L., E. BEN-NAI, I. GROSSE, 2004, STABLE DISTRIBUTIONS IN STOCHASTIC FRAGMENTATION, JOURNAL OF PHYSICS A, 37, 8, 2863-2880.
[13] KRAPIVSKY, P. L., I. GROSSE AND E. BEN-NAI, 2000, SCALE INVARIANCE AND LACK OF SELF-AVERAGING IN FRAGMENTATION, PHYSICAL REVIEW E, 61, 2, R993-R996.
[14] LAWN, B. R., T. R. WILSHAW, 1975, FRACTURE OF BRITTLE SOLIDS, CAMBRIDGE UNIVERSITY PRESS.
[15] MATSUSHITA, M., K. SUMIDA, 1988, HOW DO THIN GLASS RODS BREAK? (STOCHASTIC MODELS FOR ONE-DIMENSIONAL FRACTURE), CHUO UNIVERSITY, 31, 69-79.
[16] MATSUSHITA, M., T. ISHII, 1992, FRAGMENTATION OF LONG THIN GLASS RODS, DEP. OF PHYSICS, CHUO UNIVERSITY.
[17] PROSPERINI, N., D. PERUGINI, 2007, APPLICATION OF A CELLULAR AUTOMATA MODEL TO THE STUDY OF SOIL PARTICLE SIZE DISTRIBUTIONS, PHYSICA A, 383, 595 – 602.

[18] REDNER, S., 1990, IN STATISTICAL MODELS FOR THE FRACTURE OF DISORDERED MEDIA, H. HERRMANN AND S. ROUX (EDS.), RANDOM MATERIALS AND PROCESSES SERIES, ELSEVIER SCIENCE NORTH HOLLAND PUBLISHERS.
[19] RODGERS, G. J., M.K. HASSAN, 1996, STABLE DISTRIBUTIONS IN FRAGMENTATION PROCESSES, PHYSICA A, 233, 1-2, 19-30.
[20] STEACY, S., C. SAMMIS, 1991, AN AUTOMATON FOR FRACTAL PATTERNS OF FRAGMENTATION, NATURE, 353, 250-252.
[21] TURCOTTE, D.L., 1986, FRACTALS AND FRAGMENTATION, JOURNAL OF GEOPHYSICAL RESEARCH, 91, 1921-1926.
[22] TURCOTTE, D. L., 1997, FRACTALS AND CHAOS IN GEOLOGY AND GEOPHYSICS, 2ND ED., CAMBRIDGE UNIVERSITY PRESS.

ESTABILIDAD DINÁMICA: SIMULACIÓN DISCO DE FRENO APLICADO AL VEHÍCULO CITROËN C4.

José M. Ramírez† y Marcelo T. Piovan†

† Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional F.R.B.B, 11 de Abril 461, B8000LMI, Argentina, {ramirezjose,mpiovan}@frbb.utn.edu.ar, http://www.utn.frbb.edu.ar

Resumen. En el presente trabajo se efectúa el análisis de estabilidad dinámica del disco de freno del vehículo Citroën C4. En el procedimiento de frenado de un vehículo existe un problema de inestabilidad dinámica generando un ruido causado por las vibraciones de alta frecuencia. La plataforma ABAQUS/Standard es usado para calcular las frecuencias naturales de modelo de disco de freno y así poder extraer los modos complejos del sistema. Se analizan diversos casos de estudio para predecir el problema de inestabilidad dinámica. Las correspondientes modificaciones geométricas permiten eliminar el fenómeno de inestabilidad del sistema.

Palabras claves: *Vibraciones mecánicas, Estabilidad dinámica, Citroën C4.*

1. INTRODUCCIÓN

En el procedimiento de frenado de un vehículo existe un problema de inestabilidad dinámica generando un ruido causado por las vibraciones de alta frecuencia que generalmente esta en el rango de 1 y 16 kHz. Aunque se ha logrado investigar sustancialmente sobre este fenómeno, en la actualidad todavía es difícil predecir su ocurrencia debido a la complejidad de los mecanismos que causan el ruido de frenos. Se han formulado diversas teorías para explicar este fenómeno. El uso de materiales viscoelásticos puede resultar eficaz en la reducción del fenómeno, como también modificar la geometría de las pastillas o el disco para evitar el acoplamiento. En la actualidad el método que se utiliza es el de autovalor complejo, y es ampliamente usado para predecir el fenómeno. La idea principal de este método consiste en la asimetría de la matriz de rigidez y la formulación del acoplamiento de la fricción. Este es eficiente, es decir su costo computacional no es muy elevado. En el presente trabajo se realiza el estudio de estabilidad dinámica del sistema de freno del automóvil Citroën C4 mediante el uso del método de elementos finitos. ABAQUS/Standard es usado para calcular las frecuencias naturales de modelo de disco de freno y así poder extraer los modos complejos del sistema. Se realiza un análisis cuasi-estatico no lineal para calcular el efecto de contacto entre la pastilla y el disco antes de la extracción del autovalor complejo. También se observa los efectos que causan las modificaciones de los parámetros del sistema como la presión de contacto, la fricción, entre otros. Se realizan diversos estudios paramétricos para predecir el problema de ruido y las correspondientes modificaciones para eliminar este fenómeno.

2. BREVE DESCRIPCION DEL MODELO

En la Figura 1, se puede observar el sistema de freno del Citroën C4. Los componentes utilizados para realizar el análisis de estabilidad dinámica consisten en un rotor (disco de freno) y las pastillas de freno.

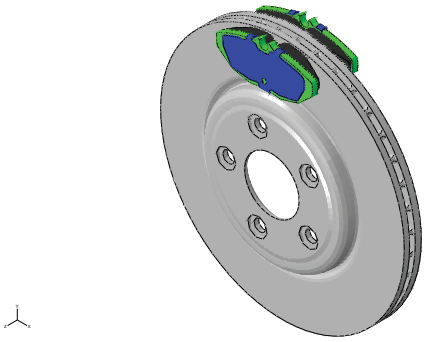


Figura 1: Modelo del disco de freno ensamblado (CAD)

3. DESARROLLO TEORICO

3.1. FORMULACIÓN GENERAL

En el presente trabajo se presenta el análisis de estabilidad dinámica de un disco de freno empleando el método de elemento finito. La ecuación diferencial para un sistema con amortiguamiento esta dada por la ecuación (1).

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F} \tag{1}$$

Siendo, $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$, $\mathbf{F}$ y $\mathbf{u}$ la matriz de masa, amortiguamiento, rigidez, el vector de fuerza y el desplazamiento generalizado respectivamente. En general, la matriz de rigidez y la matriz de masa pueden ser definidas simétricas. Para tales casos, la solución del sistema es estable. Pero existen casos donde la matriz de rigidez es asimétrica. Esta asimetría es debido a las cargas externas que interactúan con la estructura, por ejemplo la fricción en los sistemas de freno. Debido al proceso de discretización usado por el método de elemento finito, el valor crítico no es definido en términos de carga crítica, por esto es definido como tensiones de corte que esta en función de la presión de contacto entre el disco y la pastilla. Para tener en cuenta el efecto de la fricción en el sistema, se define las tensiones de corte de la siguiente forma como se puede apreciar en la ecuación (2).

$$\tau = \mu p \frac{\dot{\gamma}}{|\dot{\gamma}|} = \mu pt \tag{2}$$

Siendo el coeficiente de fricción μ dependiente de la presión p y la relación de deslizamiento $|\dot{\gamma}|$. Entonces, derivando con respecto a p, μ y t, se obtiene la expresión dada por la ecuación (3).

$$d\tau = \underbrace{\left(\mu + \frac{\partial \mu}{\partial p}\right) tdp}_{1} + \underbrace{\frac{\partial \mu}{\partial |\dot{\gamma}|} ptd|\dot{\gamma}|}_{2} + \underbrace{\mu pdt}_{3} \tag{3}$$

El término (1) contribuye a la asimetría de la matriz de rigidez, el término (2), contribuye a la matriz de amortiguamiento.

3.2. FORMULACIÓN DEL PROBLEMA DE AUTOVALORES COMPLEJO

El procedimiento de extracción del autovalor complejo utiliza un método de proyección para extraer los valores propios complejos del sistema. El problema del modelo de elementos finitos se formula de la siguiente manera, como se aprecia en la ecuación (4).

$$\left(\mu^2 \mathbf{M}^{MN} + \mu \mathbf{C}^{MN} + \mathbf{K_T}^{MN}\right)\phi^N = \mathbf{0} \tag{4}$$

Siendo μ y ϕ^N el autovalor y autovector complejo del sistema. El sistema es simetrizado omitiendo la matriz de amortiguamiento $\mathbf{C}$ y las contribuciones asimétricas de la matriz de rigidez $\mathbf{K_T}$. Ahora bien, se resuelve el problema de valores propios simétrica para encontrar la proyección. Luego las matrices originales se proyectan sobre el subespacio de vectores propios *N*.

$$\left(\mu^2 \mathbf{M}^* + \mu \mathbf{C}^* + \mathbf{K}^*\right)\phi^* = \mathbf{0} \tag{5}$$

Finalmente el problema de autovectores del sistema original puede ser obtenido por:

$$\phi = \left[\phi^1, \ldots, \phi^N\right]\phi^* \tag{6}$$

El autovalor complejo μ se puede expresar como:

$$\mu = \alpha \pm i\omega \tag{7}$$

Siendo α la parte real de μ, $\mathrm{Re}(\mu)$, indicando la estabilidad del sistema, y ω la parte imaginaria de μ, $\mathrm{Im}(\mu)$, indicando la frecuencia del modo de vibración. La relación de amortiguamiento se define como:

$$\psi = -2\alpha / |\omega| \tag{8}$$

Cuando el sistema es inestable, α es positivo. Si la relación de amortiguamiento es negativo el sistema es inestable, pero si es positivo el sistema es estable.

3.3. CONDICIONES DE BORDE

En el disco se restringe todos los desplazamientos y rotaciones en los agujeros de los tornillos y en las pastillas sólo se permite el movimiento axial (eje de rotación).

4. CASOS DE ESTUDIO

En esta sección se presenta una comparación de resultados de la simulación mediante el enfoque de elemento finito. En las Tablas 1 y 2 se aprecian las constantes elásticas que definen el material anisótropo de la pastilla de freno y las propiedades de los materiales isótropos del rotor y accesorios de la pastilla respectivamente. Se analizan diversos casos de estudio para predecir el ruido causado por vibraciones en el disco de freno y las correspondientes modificaciones geométricas (Tabla 3) para eliminar este fenómeno.

Propiedades							
Constantes elásticas	D_{1111} 5940000	D_{1122} 760000	D_{11222} 5940000	D_{1133} 980000	D_{2233} 980000	D_{3333} 2270000	D_{1212} 2590000
	D_{1313} 5940000	D_{1122} 180000	D_{2323} 180000				
Densidad (Kg/m^3)	2510						

Tabla 1: Constantes elásticas que definen el material anisótropo de la pastilla de freno

Propiedades del material	**Acero**	
Módulo de elasticidad lineal (Pa)	2.14×10^{11}	1.25×10^{11}
Módulo de Poisson	0.3	0.24
Densidad (Kg/m^3)	7820	7200

Tabla 2: Propiedades de los materiales isótropos del rotor y accesorios de la pastilla

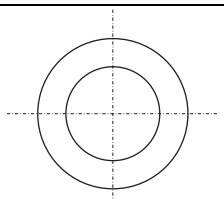
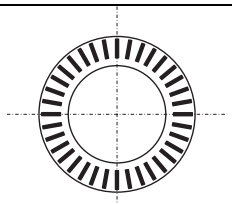
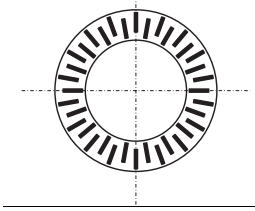
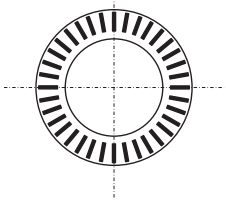
Modificación	**Descripción**	**Modificación**	**Descripción**
	M1 : Disco original delantero completamente sólido.		M2 : Disco original delantero.
	M3 : Disco delantero con las aletas ubicadas en forma alternativa.		M4 : Disco original delantero aumentando el número de aletas.

Tabla 3: Modificaciones de la geometría del disco de freno

Para todos los casos de estudio se adopta la presión ejercida sobre el disco mediante la pastilla de freno en 50 MPa, siendo el coeficiente de fricción de 0.3, en el momento del frenado la velocidad es 5 rad/seg, y el espesor del disco 20 mm. En el primer caso de estudio (caso M1) se trata de un disco completamente sólido, y como se observa en la Figura 2 no se produce inestabilidad del sistema debido a que la relación de amortiguamiento es cero para todo el rango de frecuencias. En el segundo caso de estudio (caso M2) se trata del disco que viene por defecto en el vehículo Citroën C4. Un disco que posee ventilación mediante unas aletas en su interior. En la Figura 3 se aprecia que existe inestabilidad en el modo 7 con su

correspondiente valor de frecuencia 1952 Hz. En el tercer caso de estudio (caso M3), las aletas de ventilación están dispuestas en forma alternativa. En la Figura 4 se aprecia que existe inestabilidad del sistema en el modo 7 con su correspondiente autovalor 1769 Hz. Se observa que las modificaciones solo reduce parcialmente el fenómeno. En el último caso de estudio (caso M4) posee un mayor número de aletas de ventilación. Con esta disposición geométrica se logra rigidizar el disco sin perder la capacidad de refrigeración ni comprometer el funcionamiento del disco. Nótese en la Figura 4 que no se produce inestabilidad del sistema en el rango de frecuencias.

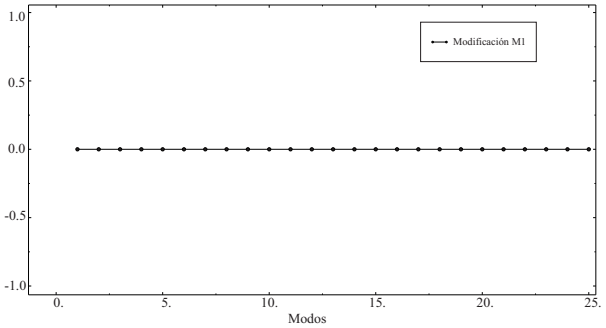


Figura 2: Relación de amortiguamiento (caso M1)

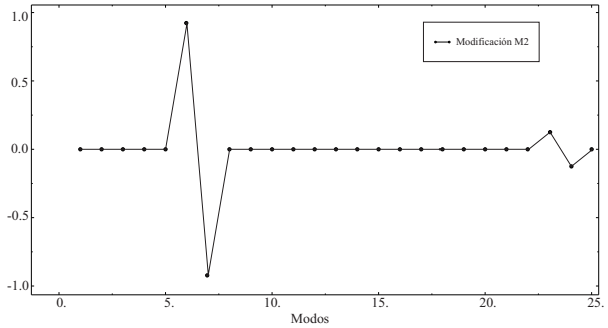


Figura 3: Relación de amortiguamiento (caso M2)

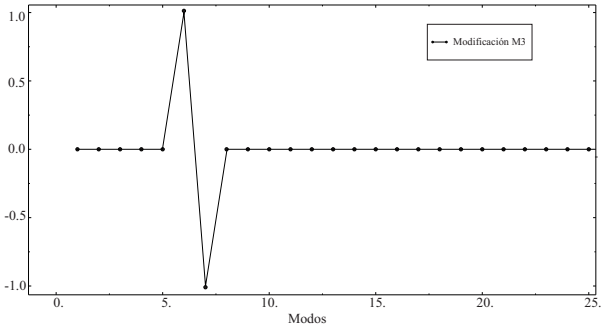


Figura 4: Relación de amortiguamiento (caso M3)

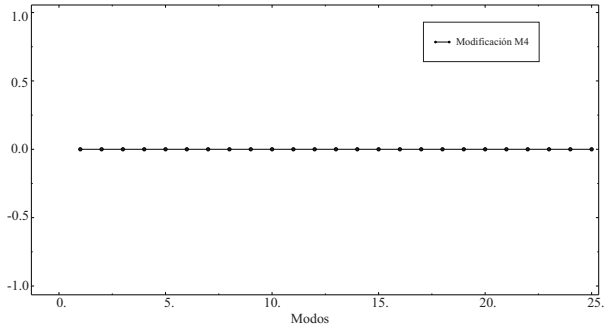


Figura 5: Relación de amortiguamiento (caso M4)

5. CONCLUSIÓN

En el presente trabajo ha efectuado un estudio comparativo de estabilidad dinámica del disco de freno del vehículo Citroën C4 utilizando el método de elemento finito mediante la plataforma ABAQUS/Standard. Se realizan diversas modificaciones geométricas para predecir el fenómeno de inestabilidad. La modificación M1, no se produce inestabilidad dinámica en el rango de frecuencias extraído, con la desventaja que el disco es completamente sólido. La modificación M2 corresponde a la geometría estándar del disco de freno del vehículo Citroën C4 y en el rango de frecuencias extraído se detecta en el modo 7 (1952 Hz) inestabilidad del sistema. La modificación M3 posee una disposición geométrica distinta de las aletas, atenuando parcialmente el fenómeno de inestabilidad. Ahora bien, con la última modificación geométrica M4 se rigidiza la estructura aumentando el numero de aletas de ventilación, eliminando así el fenómeno de inestabilidad del sistema. Esto es oportuno en tanto que la distribución geométrica implementada (M4) es la adecuada, ya que permite conservar las propiedades estructurales y también se suprime el fenómeno de inestabilidad del sistema.

REFERENCIAS

[1] SHIN K, BRENNAN MJ, OH JE, HARRIS CJ. *Analysis of disc brake noise using a two-degree-of-freedom model.* J Sound Vibrat 2002;254:837–48.

[2] GUAN DH, HUANG JC. *The method of feed-in energy on disc brake squeal*. J Sound Vibrat 2003;261:297–307.

A MODEL FOR DYNAMIC ANALYSIS OF MAGNETO-ELECTRO-ELASTIC BEAMS WITH CURVED GEOMETRY

José M. Ramírez[♭] and Marcelo T. Piovan[♭,†]

[♭]*Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional F.R.B.B, 11 de Abril 461, B8000LMI, Argentina, (ramirezjose,mpiovan)@frbb.utn.edu.ar, http://www.utn.frbb.edu.ar*
[†]*CONICET*

Abstract: In this article a dynamic analysis on structures featured with special materials is performed. The structure consists of a curved sandwich beam with ceramic-metallic materials whose elastic behavior can be modified by influence of electro-magnetic fields. The interest for this type of magneto-electro-elastic (MEE) couplings is associated to the construction of imbedded sensors and actuators in high performance aero spatial structures. The problem of coupled elastic, electric and magnetic fields is briefly explained and the 3D motion equations for curved members are introduced. A finite element approach is employed in order to calculate natural frequencies and forced vibrations of sandwiched MEE curved beams.

Keywords: *magneto-electro-elastic beams, functionally graded materials, finite elements.*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

In recent years, simple structures constructed with piezoelectric and magnetostrictive materials are employed in engineering applications for sensing and actuation uses. Hence studies of dynamics of this kind of structures are an important task. An interesting variety of models MEE structures has been introduced principally for piezoelectric and piezomagnetic plates and shells [1-3]. In these articles the 3D static behavior of multilayered MEE strips and plates has been presented. The structures where subjected to sinusoidally distributed magnetic, electric and mechanic loads. Dynamic counterparts of aforementioned works have been performed by Chen et al. [4] and Pan and Heyliger [5] among others. Wu and Lu [6] and Tsai et al. [7] among others studied dynamics responses of shells and plates appealing to 3D formulations.

All previously mentioned research articles are devoted to rectangular plates and shells or at least shells with slight geometrical complications, i.e. for example simply or doubly curved profile. According to the bibliographical review carried out, it is important to remark that articles related to static or dynamic behavior of curved MEE beams are absent. Thus, the scope of this research is directed toward offering some contributions in the mechanics of curved MEE beams. Thus, in this article, a 3D model representing the mechanics of layered curved MEE beams is presented. The coupled magnetic, electric and elastic equations are solved within the context of the finite element method in order to characterize the static behavior, free and forced vibration behavior of the curved beams.

2 THEORY AND MODELS FOR MAGNETO-ELECTRO-ELASTIC MATERIALS

2.1 MODEL DESCRIPTION AND CONSTITUTIVE EQUATIONS

Figure 1 shows a sketch of a curved multilayered beam composed by linear piezoelectric and magnetostrictive materials. The coordinate referencing is based on a typical circumferential system whose origin is located at the geometric center of the cross-section, i.e. point **O**. The beam domain is bounded such that $x \in [-b/2, b/2]$, $y \in [-h/2, h/2]$ and $z \in [0, L]$, whereas R is the curvature radius, assumed constant in this work, L is the length of the arc with radius R and subtended angle α. The linear constitutive equations of the considered MEE material are given by:

$$\sigma_{ij} = c_{ijkl}\varepsilon_{kl} - e_{ijm}E_m - q_{ijm}H_m \tag{1}$$

$$D_i = e_{ikl}\varepsilon_{kl} + \eta_{im}E_m + d_{im}H_m \tag{2}$$

$$B_i = q_{ikl}\varepsilon_{kl} + d_{im}E_m + \mu_{im}H_m \tag{3}$$

where, in contracted tensorial form, σ_{ij} and ϵ_{kl} are the stress and strain components, respectively; D_i and B_i are the electric displacements and magnetic fluxes, respectively; E_m and H_m are the components of the electric field and the magnetic field, respectively; c_{ijkl} are elastic coefficients; η_{im} are the dielectric coefficients; μ_{im} are the magnetic permeability coefficients; e_{ijm} are the piezoelectric coefficients; q_{ikl} are the piezomagnetic coefficients and d_{im} are the magnetoelectric coefficients. All sub-indexes are such that $\{i, j, k, l, m\} \in \{1, 2, 3\}$. Indexes are related to co-ordinate variables such that $\{1, 2, 3\} \equiv \{x, y, z\}$.

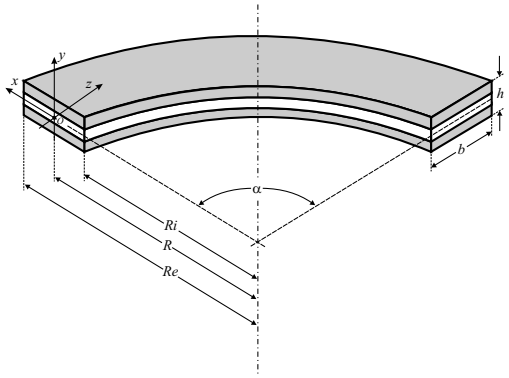


Figure 1: Curved and layered MME beam.

The vector forms of the electric and magnetic fields are defined in terms of the gradient (symbolized with the operator $\bar{\nabla}$) of the electric potential Φ, and the gradient of the magnetic potential Ψ , respectively, i.e.:

$$\begin{aligned} \bar{E} &= -\bar{\nabla}\Phi \\ \bar{H} &= -\bar{\nabla}\Psi \end{aligned} \tag{4}$$

The strain components can be defined in terms of displacements $\{u_1, u_2, u_3\} \equiv \{u_x, u_y, u_z\}$ as:

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{5}$$

The structure should verify the internal dynamic equilibrium given by the following equations:

$$\frac{\partial \sigma_{ij}}{\partial x_i} + X_j = \rho\frac{\partial^2 u_i}{\partial t^2} \tag{6}$$

where, ρ and t are the mass density and time variable respectively, whereas X_j are the volume forces.

The field equations of electrostatics and magnetostatics of a curved MEE beam in absence of electric charge density and magnetic charge density are given by:

$$\begin{aligned} \bar{\nabla} \cdot \bar{E} &= 0 \\ \bar{\nabla} \cdot \bar{H} &= 0 \end{aligned} \tag{7}$$

where $\bar{\nabla} \cdot (\bullet)$ is the divergence operator.

Thus, the problem of a MEE solid in a 3D formulation is governed by the three PDE given in Eq. (eqn:06) and the two PDE given in Eq. (eqn:07). Finally, the system of five coupled PDE is completely defined by imposing the appropriate mechanic and electromagnetic boundary conditions and initial conditions. These equations are case dependent and will be furnished in the analysis of illustrative cases.

2.2 FINITE ELEMENT MODELING AND ILLUSTRATIVE EXAMPLES

The coupled PDE system presented in the previous section can be solved numerically with the aid of a general purpose finite element solver for second order partial differential equations [8]. Due to space reasons the development of the finite element equations for this MEE problem is not furnished. The interested reader may follow the common bibliographical references on finite element analysis [9]. Nevertheless, employing the conventional steps of the finite element method to Eq. (6)-(7) it is possible to arrive at:

$$\mathbb{M}_{uu}\ddot{\mathbf{U}} + \mathbb{K}_{eq}\mathbf{U} = \mathbf{F} \tag{8}$$

where $\mathbb{M}_{uu}$ is the global mass matrix, $\mathbb{K}_{eq}$ is the global equivalent stiffness matrix, $\ddot{\mathbf{U}}$ is the global vector of accelerations, $\mathbf{U}$ is the global vector of displacements and $\mathbf{F}$ is the global vector of forces (which can be time-dependent or not). The matrix $\mathbb{K}_{eq}$ is obtained by means of standard condensation techniques which eliminate the electric and magnetic variables. The Eq. (8) can be employed to solve three cases: the general transient dynamic problem, the eigenvalue problem by eliminating any active force and adopting harmonic displacements and finally the static problem by eliminating time dependence.

The first example corresponds to a test of the numerical procedure. Then, taking into account that the present study is based on a general 3D formulation, an especial limit case which has analytical solutions [6], can serve as a test. The present 3D model of curved beam can be reduced to a rectangular plate by imposing the limiting condition $R \to \infty$. The geometrical properties of the plate are such that $L = b = 1\ m$, $L/2h = 10/3$, whereas the elastic, piezoelectric, dielectric and magnetic properties of $CoFe_2O_4$ and $BaTiO_3$ can be found in Ref. [10]. Models of more than 3000 tetrahedral finite elements of quadratic interpolation have been prepared to do the calculations. The boundary conditions for this example are given as follows:

$$\begin{array}{l} \sigma_{22} = \sigma_{12} = \sigma_{23} = \Psi = \Phi = 0, \text{on} \quad y = \pm h/2 \\ \sigma_{33} = u_1 = u_2 = \Psi = \Phi = 0, \text{on} \quad z = 0, z = L \\ \sigma_{11} = u_3 = u_2 = \Psi = \Phi = 0, \text{on} \quad x = \pm b/2 \end{array} \tag{9}$$

Table 1 shows the frequency coefficients $\bar{\omega}_i = L\omega_i\sqrt{\rho/c_{1111}}$ for two different cases of materials, where ω_i is the circular frequency measured in [rad/seg]. As it is possible to see for natural vibration problems, the correlation between analytical solutions [6] and the present numerical results is quite good.

Type of material	Approach	$\bar{\omega}_1$	$\bar{\omega}_2$	$\bar{\omega}_3$
$CoFe_2O_4$	Analytic [6]	1.2523	2.3003	3.8314
	FEM	1.2469	2.3000	3.8492
$BaTiO_3$	Analytic [6]	1.0212	1.9747	3.3905
	FEM	1.0211	1.9730	3.4005

Table 1: Test of the present numerical approach with available analytical solutions.

In the second example the natural vibration and static behavior of layered MEE curved beams are analyzed. The curved beam is composed by three layers as shown in Fig. 1. The upper domain is made of $CoFe_2O_4$, the lower domain is made of $BaTiO_3$, and the inner domain has properties varying functionally from $CoFe_2O_4$ to $BaTiO_3$ according to the rule $\mathcal{P} = \mathcal{P}_M\left[(y+1)/2\right]^{\kappa} + \mathcal{P}_P\left[1 - \left[(y+1)/2\right]^{\kappa}\right]$, where $\mathcal{P}$, $\mathcal{P}_P$, and $\mathcal{P}_M$ denote a generic (Density, Young's Modulus, etc) graded property, the generic property of piezoelectric and the generic property of magnetostrictive materials, respectively; whereas κ means the material property gradient index; the subscripts P and M denote the piezoelectric and magnetostrictive materials, respectively. The geometric measures of the curved beam are $L = R = 1\ m$ and $b/h = 5$ for dynamic analysis and $2L/\pi = R = 1\ m$ and $b/h = 3$ for static analysis. In the static analysis, the beam is subjected to a radial constant stress $p_R = -100\ N/m^2$ applied through $\mathbf{F}$ at $x = b/2$.

Three dimensional models of more than 6000 tetrahedral finite elements of cubic interpolation have been prepared in order to perform the calculations. The boundary conditions of this example are, for the electromagnetic part $\Phi = \Psi = 0$ in the whole domain and the mechanic conditions, called clamped-free (CF) and clamped-clamped (CC), can be described as:

$$CF \rightarrow \begin{cases} u_i(x,y,0) = 0, \sigma_{ij}(x,y,L) = 0 \\ \sigma_{ij}(\pm b/2, \pm h/2, z) = 0 \end{cases}, CC \rightarrow \begin{cases} u_i(x,y,0) = u_i(x,y,L) = 0 \\ \sigma_{ij}(\pm b/2, \pm h/2, z) = 0 \end{cases} \quad (10)$$

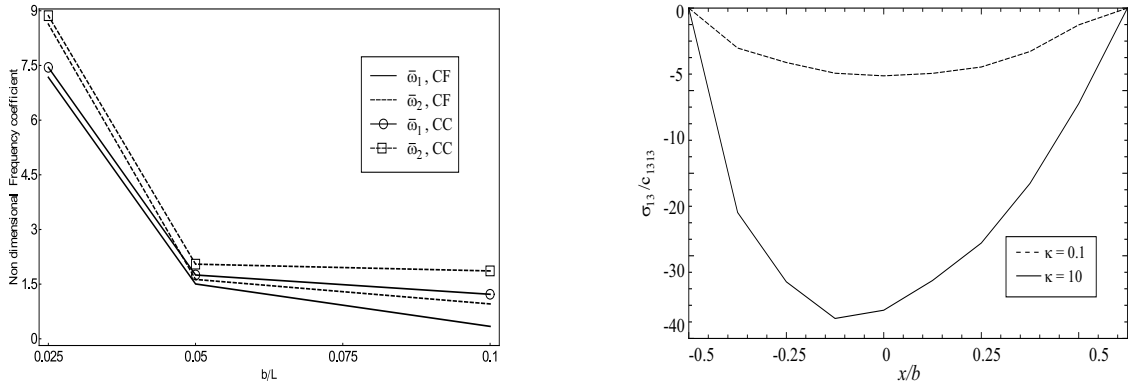


Figure 2: (a) Natural frequencies of MEE curved beams (b) Ratio of shear stress.

In Figure 2 some results of the finite element calculation are presented. Thus, Figure 2(a) shows the variation of the first two natural frequency coefficients $\bar{\omega}_i$ for the CF and CC boundary conditions with the material property gradient index $\kappa = 0.1$. Figure 2(b) shows the variation of the shear stress σ_{13} (normalized with respect to elastic coefficient c_{1313}) along the radius in the intersection of planes $y = 0$ and $z = L/2$, for different material gradient indexes.

3 CONCLUSIONS

In this work some new interesting studies on the mechanics of curved MEE beams have been furnished. The mechanics of the curved beam has been modeled in the context of a 3D approach. The motion equations have been solved with the finite element method. These results together with analytical solutions, in the fewer cases available in the literature, are important benchmarks for testing 1D models of curved MEE beams to be farther developed. These 1D models, due to their inherent computational low cost, should be useful tools to analyze stochastic problems associated with the uncertainties in material properties, piezo-eletro-magnetic loads, and environmental conditions among others.

ACKNOWLEDGMENTS

Authors want to recognize the support of Universidad Tecnológica Nacional and CONICET.

REFERENCES

[1] P. HEYLINGER, *Exact solutions for simply supported laminated piezoelectric plates*, J. Appl. Mech. 64 (1997), pp. 299-306.

[2] E. PAN, AND P. HEYINGER, *Exact solutions for magneto-electro-elastic laminates in cylindrical bending*, Int. J. Solids Struct. 40 (2003), pp.6859-6876.

[3] E. PAN, *Exact solution for simply supported and multilayered magneto-electro-elastic plates*, J. Appl. Mech. 68 (2001), pp. 608-618.

[4] J. CHEN, H. CHEN, E. PAN AND P. HEYLINGER, *Modal analysis of magneto-electro-elastic plates using the state-vector approach*, J. Sound Vib., 304 (2007), pp.722-734.

[5] E. PAN AND P. HEYLINGER, *Free vibration of simply supported and multilayered magneto-electro-elastic plates*, J. Sound Vib., 252 (2002), pp.429-442.

[6] C-P. WU AND Y-C. LU, *A modified Pagano method for the 3D dynamic responses of functionally graded magneto-eletro-elastic plates*, Composite Structures 90 (2009), pp. 363-372.

[7] Y-H. TSAI, C-P. WU AND Y-S. SYU, *Three-dimensional analysis of doubly curved functionally graded magneto-electro-elastic shells*, Eur. J. Mech. A/Solids 27 (2008), pp. 79-105.

[8] G. BACKSTROM, *Deformation and vibration by finite element analysis*, Studentlitteratur, Sweden 1998.

[9] K-J. BATHE, *Finite Element procedures in Engineerign Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, USA 1982.

[10] B. BIJU, N. GANESAN AND K. SHANKAR, *Response of multiphase magneto-electro-elastic sensors under harmonic mechanical loading*, Ent. J. Eng. Sci. Tech. 1(1) (2009),pp.216-227.

MODELO DE AGENTES PARA UN MERCADO FINANCIERO

Juan José M. Martínez ♭

♭*Centro de Investigaciones en Economía Teórica y Matemática Aplicada - Escuela de Economía y Negocios - Universidad Nacional de San Martín, Caseros 2241, 1650 San Martín, Argentina, jjmm@unsam.edu.ar, www.unsam.edu.ar*

Resumen: Se propone un modelo de agentes heterogéneos interactuantes para simular la dinámica de un mercado de capitales. Dicho modelo se basa en un mecanismo de competencia-complementariedad entre dos tipos de agentes definidos por sus estrategias de decisión fundamentadas en observables -información accesible a cada uno de ellos- y con racionalidad total o acotada según sea su tipo. Cada agente puede intercambiar cantidades arbitrarias definidas por sus expectativas y disponibilidades. Un proceso de "negociación"determina finalmente la magnitud de su participación y establece el precio de transacción. Incerteza en el precio fundamental, adaptación según beneficio, participación proporcional a las expectativas y frustración de las mismas, son puntos clave de este modelo.

Palabras clave: *Mercados Financieros, Equilibrio, Hechos Estilizados, Modelo de Agentes, Simulación*
2000 AMS Subject Classification: 68U20 91G99

1. INTRODUCCIÓN

En los últimos años ha habido un gran interés en el desarrollo de los llamados Modelos Basados en Agentes (MBA), destinados a la reproducción y comprensión de los Hechos Estilizados (HE) observados en las series de tiempo financieras ([4] y [7]). El modelo más simple de series de tiempo de precios es el paseo aleatorio (PA) introducido por Louis Bachelier en 1900[1].

La disponibilidad de grandes cantidades de datos ha puesto de manifiesto un conjunto de desviaciones sistemáticas al PA que son los HE, relativamente comunes a todos los mercados.

1.1. HECHOS ESTILIZADOS

Los principales HE, discutidos en detalle en [2],[3] y [6], son los siguientes: Ausencia de autocorrelación lineal, distribución de los retornos no gaussiana -colas que se pueden aproximar por una ley de potencia- y volatilidad agrupada -según Mandelbrot, [5]: l*os grandes cambios tienden a ser seguidos por grandes cambios, de cualquier signo, y los pequeños cambios tienden a ser seguidos por los pequeños cambios-*.

Estos HE han sido interpretados en términos de diversos modelos estocásticos fenomenológicos que permiten, en algunos casos, una estimación del riesgo más allá del modelo de Black y Scholes -que corresponde a los simples PA-. Para construir un marco conceptual, necesario para entender la dinámica del mercado, es importante añadir los siguientes elementos:Auto-organización y no estacionariedad con escalas de tiempo.

1.2. LINEAMIENTOS DEL MODELO

En este trabajo se propone un MBA, que incluye los siguientes elementos:

Precio de referencia: Todos los agentes tienen como referencia un precio $p_i^*(t)$, que es su apreciación del precio fundamental $p_F(t)$ -derivado del análisis económico del valor de las acciones-. La *imprecisión* en la evaluación se debería a información erronea, parcial, o mal interpretada por cada agente. Este *ruido* se supone no tendencioso. $p_i^*(t)$ es una medida -en el sentido físico- de $p_F(t)$.

Disponibilidades limitadas: Cada agente dispone de una cantidad de acciones $Q_i\,(t)$ y de capital $C_i\,(t)$ que evoluciona en el tiempo segun intercambien $q_i\,(t)$ acciones a precio $p(t)$.

Agentes: Dos tipos de agentes, en competencia-complementariedad, conforman la población. Los Fundamentalistas; racionales y de efecto estabilizador. Los Interaccionistas; adaptativos y dinamizadores del mercado.

Comportamiento frente a los precios: Cada agente genera un indicador cuantitativo, $W_i(t)$, medida de su preferencia. Esta señal se compara con la diferencia entre el precio de referencia que percibe cada cada uno y el precio de mercado. Esta conparación define la medida de su participación en la transacción.

Mecanismo de búsqueda de equilibrio: Frente a diferencias entre las espectativas de los agentes se lleva a cabo un proceso de búsqueda del equilibrio -con posible frustración parcial de las mismas-. Así el cambio de precio es modelado como una respuesta endógena del mercado al desequilibrio oferta-demanda de acuerdo con un ajuste walrasiano.

2. EL MODELO

2.1. GENERALIDADES

Dado un mercado de capitales con N agentes que negocian un activo, sea el conjunto de agentes $\Phi = \{i \in \mathbb{N} \,/\, i \leq N\}$ y sea una sucesión de *tiempos de transacción*, $\{\tau_t\}_{t\geq 0}$, $t \in \mathbb{N}$ - tal que en cada uno de ellos se realiza una transacción a un *precio de transacción* $p(t)$, en equilibrio oferta-demanda. Cada uno intercambia -compra, vende o no negocia-, a dicho tiempo, una cantidad $q_i(t) \geq 0$ - limitada por sus disponibilidades- del bien. Esto es descripto por un *vector de decisiones*

$$\overrightarrow{s}\,(t) = (s_1\,(t)\,, s_2\,(t)\,, .., s_N\,(t)) \quad / \quad s_i(t) = \begin{cases} -1 \text{ vende} \\ 0 \quad \text{inactivo} \\ +1 \text{ compra} \end{cases}$$

y un *vector de cantidades*: $\overrightarrow{q}\,(t)$ / $\overrightarrow{s}\,(t) * \overrightarrow{q}\,(t) = \sum_{i=1}^{N} s_i\,(t)\,q_i\,(t) = 0$

Para hacer tales transacciones los agentes disponen de capital para comprar y acciones para vender. Luego de la transacción se tendrá un *vector de capitales* $\overrightarrow{C}\,(t)$ y un *vector de cantidades disponibles* $\overrightarrow{Q}\,(t)$.

Entre los tiempos τ_{t-1} y τ_t , (t-ésimo *intervalo de negociación*, con los $\Delta\tau = \tau_t - \tau_{t-1}$), no hay movimientos de compra-venta y se realiza el "*tatonnement*"que llevará a una nueva transacción a tiempo τ_t ; un mecanismo de búsqueda de equilibrio -Sección 3- actúa de *subastador walresiano.*

2.2. AGENTES FUNDAMENTALISTAS

Su estrategia de intercambio, con propensión al riesgo de grandes desvios, es vender o comprar si el precio excede o decae en más de un $\kappa_i(t)$,$100\,\%$. Es decir

$$s_i^{(k)}(t) = \begin{cases} -1 & \text{si } \frac{p^{(k)}(t)}{p_i^*(t)} > 1 + \kappa_i(t) \\ +1 & \text{si } \frac{p^{(k)}(t)}{p_i^*(t)} < 1/(1+\kappa_i(t)) \\ 0 & \text{en otro caso} \end{cases} \Rightarrow W_i^{(k)}(t) = -\,sg\left[p^{(k)}(t) - p_i^*(t)\right].\lambda.\ln\left[1+\kappa_i(t)\right]$$

El valor de $\kappa_i(t)$ resulta de la evaluación de riesgos, criterios de inversión y otras consideraciones, supuestamente racionales, de cada agente.

2.3. AGENTES INTERACCIONISTAS

Los interaccionistas consideran que las desviaciones entre el precio de negociación $p^{(k)}(t)$ y el precio de referencia $p_i^*(t)$ son una medida de la *asimetría* del mercado: $\ln\left(p^{(k)}(t)/p_i^*(t)\right) = \frac{1}{\lambda}.\frac{1}{N}.\widehat{\sum_{i=1}^{N} s_i^{(k)}}(t)$ estimando la asimetría real, $\frac{1}{N}.\sum_{i=1}^{N} s_i^{(k)}(t)$, no conocida por el agente. Para simplificar, consideremos que la preferencia de acción del agente es $W_i(t) = \omega_i^1.\underbrace{\frac{1}{n_i}.\sum_{j\in\Omega_i} s_j(t-1)}_{(a)} + \omega_i^2.\underbrace{\widetilde{s_i(t)}}_{(b)} / \omega_i^1 + \omega_i^2 = 1$

(a): Asimetría de los vecinos al agente i-ésimo en el periodo anterior: muestra la tendencia o campo local; (b): $\widetilde{s_i(t)}$ es una variable aleatoria -en principio $\widetilde{s_i(t)} \sim N\left(0, \sigma_i^2\right)$- : componente no racional de la decisión del agente; ω_i , ponderaciones.

Este tipo de agentes tienen aversión a los grandes desvíos **y** toman su decisión comparando el comportamiento local con el del mercado en conjunto según

$$s_i^{(k)}(t) = \begin{cases} +1 & \text{si } W_i(t) > \nu_i(t).\lambda.\ln\left(p^{(k)}(t)/p_i^*(t)\right) > 0 \\ -1 & \text{si } W_i(t) < \nu_i(t).\lambda.\ln\left(p^{(k)}(t)/p_i^*(t)\right) < 0 \\ 0 & \text{en otro caso} \end{cases} \quad /\nu_i(t) = \begin{cases} +1 & \text{actua con la mayoría} \\ -1 & \text{actua con la minoria} \end{cases}$$

donde $\nu_i(t) = \begin{cases} \text{si } \mod_{t_0}(t) \neq 0: & \nu_i(t-1) \\ \text{si } \mod_{t_0}(t) = 0: & \begin{cases} \nu_i(t-1) & \text{si } \overline{M}_i(t) \geq \overline{M}_i(t-t_0) \\ -\nu_i(t-1) & \text{si } \overline{M}_i(t) < \overline{M}_i(t-t_0) \end{cases} \end{cases}$,

siendo $\overline{M}_i(t) = \frac{1}{t_0} \sum_{j=0}^{t_0-1} [Q_i(t-j).p(t-j) + C_i(t-j)]$ el valor promedio de su cartera entre las últimas t_0 transacciones.

2.4. CANTIDADES

Un indicador $e_i^{(k)}(t) \geq 0$ dará la magnitud del interés de cada agente en la operación, medida de la *intensidad* de su expectativa de compra-venta, que se definirá proporcional a la separación entre las preferencias y la barrera λ. $\left|\ln\left(p^{(k)}(t)/p_i^*(t)\right)\right|$; y tal que $e_i^{(k)}(t) \leq 1$ ($\eta_i(t)$ es un factor de normalización): $e_i^{(k)}(t) = \eta_i(t) \,.\, \left|s_i^{(k)}(t)\right| \,.\, \left||W_i(t)| - \lambda.\left|\ln\left(p^{(k)}(t)/p_i^*(t)\right)\right|\right|$

La cantidad que cada agente desea intercambiar en la negociación se tomará proporcional a $e_i^{(k)}(t)$:[1]

$$q_i^{(k)}(t) = \begin{cases} \left\lfloor Q_i(t-1).e_i^{(k)}(t) \right\rfloor & \text{si} \quad s_i^{(k)}(t) = -1 \\ \left\lfloor \left(C_i(t-1)/p^{(k)}(t)\right).e_i^{(k)}(t) \right\rfloor & \text{si} \quad s_i^{(k)}(t) = +1 \\ 0 \quad \text{si} \quad s_i^{(k)}(t) = 0 \end{cases}$$

3. BÚSQUEDA DEL EQUILIBRIO

El proceso se da en dos etapas.

Inicialización: El precio inicial *-precio de inicialización-* de la negociación será $p^{(0)}(t) = p(t-1)$ -condición de continuidad del precio-. Los agentes, en consecuencia, toman una decisión inicial de comprar, vender o no accionar (los $s_i^{(0)}(t)$) con cierto nivel de convicción sobre la conveniencia de la operación (los $e_i^{(0)}(t)$) a partir de dicho precio. Para los vectores así obtenidos queda definido el exceso de demanda $\overrightarrow{s^{(0)}}(t) * \overrightarrow{q^{(0)}}(t) = \sum_{i=1}^{N} s_i^{(0)}(t).q_i^{(0)}(t)$.

Negociación: Si las cantidades ofertadas y demandadas son iguales, resulta el equilibrio con $\overrightarrow{s}(t) = \overrightarrow{s^{(0)}}(t)$, $\overrightarrow{q}(t) = \overrightarrow{q^{(0)}}(t)$ y $P(t) = P^{(0)}(t)$; en caso contrario, comienza un proceso con sucesivos precios de negociación: $\Delta\left[\ln p^{(k)}(t)\right] = \varepsilon_k(t).\overrightarrow{s^{(k)}}(t) * \overrightarrow{q^{(k)}}(t)$

El exceso de demanda es una funcional de los estados $\overrightarrow{s}(t)$ accesibles según $p^{(k)}(t)$. Así, no hay condiciones necesarias o suficientes de equilibrio sino que resulta en una sucesión $\left\{\left|\overrightarrow{s^{(k)}}(t) * \overrightarrow{q^{(k)}}(t)\right|\right\}_{0 \leq k \leq m}$ que se resuelve en un valor mínimo *-mejor acuerdo-*, eventualmente cero, con un número finito, m, de pasos.

Llegado al precio de mejor acuerdo $p(t) = p^{(m)}(t)$, el exceso remanente –mínimo en el valor absoluto del exceso de demanda- se resuelve con una *frustración* de los excedidos en sus cantidades ofertadas o demandadas, el *coeficiente de satisfacción* $\xi(t)$: $\xi(t).\sum_{s_i^{(m)}(t)=1} q_i^{(m)}(t) - \sum_{s_i^{(m)}(t)=-1} q_i^{(m)}(t) = 0$

4. SIMULACIÓN

Las simulaciones fueron llevadas a cabo con un programa desarrollado a tal efecto. Este se encuadra en una geometría doblemente cilíndrica de autómatas celulares. El vector de estado $\overrightarrow{s}(t)$ de dimensión N es mapeado a una matriz de estado: $M_S(t) \in A_S^{DxD}/N = D^2 \wedge A_S = \{-1, 0, 1\} \Leftrightarrow s_k(t) = [M_S(t)]_{ij} \wedge k = i + D.j/1 \leq k \leq N, 1 \leq i, j \leq D$; de igual forma se mapean los otros vectores, como ser $\overrightarrow{q}(t)$ por $M_q(t)$. Esta última matriz se representa gráficamente como un cuadro de DxD con un código de colores tal que el azul se corresponde con la compra, el rojo con la venta y el blanco con la no intervención. Los tonos de azul y rojo son proporcionales a las cantidades normalizadas a las máximas. La vecindad geométrica del i-ésimo elemento está, en principio, asociada a la vecindad de agentes Ω_i definida anteriormente.

Son parámetros configurables de la simulación:

[1] Aquí "$\lfloor\rfloor$"denota a la *función piso* definida por $\lfloor x \rfloor = \{y \in \mathbb{Z}/x \in \mathbb{R} \wedge y \leq x < y + 1\}$.

a) N (número de agentes,cuadrado perfecto) y *NF %* (fracción de agentes fundamentalistas). Fijada la fracción, éstos se distribuyen aleatoriamente en el mapeado.

b) $P_F(t)$ (precio fundamental). Se puede mantener constante o con distintos tipos de variaciones durante la corrida. En particular se simulan saltos aleatorios que se pueden configurar en cantidad y con magnitud aleatoria en un rango prefijado.

c) t (número de pasos de la simulación). No hay límite para ello. En general las corridas se han hecho entre 2000 y 10000 pasos.

d) σ_α (desvío estándar del ruido en el precio fundamental). Los $\alpha_i(t)$ se distribuyen normalmente con media cero y desvío σ_α. Estos valores se distribuyen aleatoriamente para cada tipo de agente - y en forma independiente un tipo del otro-.

e) Las $\overrightarrow{Q}(0)$ (cantidades iniciales) y los $\overrightarrow{C}(0)$ (capitales iniciales) se distribuyen aleatoriamente según normales de parámetros μ_Q, σ_Q, μ_C y σ_C ,fijadas independientemente para cada tipo de agente.

f) Para los agentes interaccionistas se elije λ (factor de escala), las dos constantes de peso $\omega_i^j(t)$ que en principio se fijan iguales para todos los agentes y constantes en el tiempo. También el número de vecinos geométricos n_i, igual para todos los agentes interaccionistas - en 4 u 8- y el t_0 (horizonte).

g) El estado $\overrightarrow{s}(0)$ se elije $s_i(0) = 0, \forall i \in \Phi$ o bien con valores aleatorios de distribución tricotómica con probabilidades arbitrarias (a priori iguales).

Como ejemplo tomemos $N = 1600$; $NF\,\% = 10$; $n_i = 4$; $\omega_i^1(t) = 0,75$; $\omega_i^2(t) = 0,25$; $P_F(t) = 100$ (precio fundamental constante); $t_0 = 5$. Este escenario es de baja densidad de fundamentalistas (10 %) y con interaccionistas que fundan su estrategia más teniendo en cuenta el comportamiento de sus vecinos en el periodo anterior (75 %, con memoria corta) que fundandose en sus propias y subjetivas razones (25 %, aleatoria).

En un somero análisis de la simulación hay puntos a destacar: i) Se observa el retorno clusterizado. ii) Etapas cuasi normales se identifican con actividad cercana a la media, alto coeficiente de satisfacción -cercano a 1- y agentes organizados en grupos de compra-venta donde las fronteras varían entre comprar, vender o no actuar . iii) Etapas de fuertes fluctuaciones -excess volatility puzzle- se identifican con alta actividad, caida en el coeficiente de satisfacción y agentes desorganizados y se asocian a periodos de alta incertidumbre. iv) En un análisis de la distribución de retornos se obtienen colas gruesas y alta curtosis . v) Por otro lado se observa que la volatilidad acumulada va según $P(|R_1(t)| > x) \simeq a.x^{-b}$ con $b = -3,02$; valor en el rango de lo observado (ver [3]).

5. CONCLUSIONES

Las simulaciones han permitido reproducir varios HE. La fracción de agentes resultó ser un parámetro crítico en la aparición de comportamientos no gaussianos -alejados de PA- con etapas caóticas. Por otro lado, dentro de los agentes interaccionistas, el término de campo medio se confirma como responsable de generar cluster de comportamiento coordinado –herding-. Además, las disponibilidades finitas, en algunos casos, dan lugar a altas concentraciones de capital y acciones en pocos agentes de una manera, aparentemente, aleatoria. En consecuencia, la exploración de escenarios, asociada a la configuración de parámetros permitiría identificar la importancia de ellos en mercados reales.

REFERENCIAS

[1] L. BACHELIER, *Théorie de la spéculation*. Annales Scientifiques de l'Ecole Normale Supérieure, III-17 (1900) pp.21–86.

[2] J. P. BOUCHAUD AND M. POTTERS, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, (2003).

[3] R. CONT, *Empirical properties of assets returns: stylized facts and statistical issues*. Quant. Finance, 1 (2001) pp.223–236.

[4] B. LEBARON, *Agent-based computational finance*, Handbook of computational economics Vol.2. North-Holland (2006**)**.

[5] B.B. MANDELBROT, *Fractals and Scaling in Finance*. Springer Verlag, NewYork, (1997).

[6] R. N. MANTEGNA AND H.E. STANLEY, *An Introduction to Econophysics: Correlation and Complexity in Finance*. Cambridge University Press, NewYork, NY, USA, (2000).

[7] E. SAMANIDOU, E. ZSCHISCHANG, D. STAUFFER AND T. LUX, *Agent-based Models of Financial Markets* (2007). Disponible en web: http://arxiv.org/PS_cache/physics/pdf/0701/0701140v1.pdf

UN MODELO DE DECISIÓN CON VOTACIÓN AMPLIADA

David L. La Red†, José I. Peláez‡ y Jesús M. Doña‡

†Dpto. de Informática, Universidad Nacional del Nordeste, Corrientes, Argentina, lrmdavid@exa.unne.edu.ar
‡Dpto. de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, España, {jignacio, jmdona}@lcc.uma.es

Resumen: En las democracias representativas el pueblo delibera y gobierna a través de sus representantes, los que son elegidos mediante sufragio universal de entre listas de candidatos. Esta forma de elección, es uno de los factores que está haciendo que los porcentajes de participación en las procesos electorales sean cada vez menores, ya que los ciudadanos se sienten limitados tanto en la amplitud como en la expresión de su voluntad. Una forma de mejorar la participación de los ciudadanos en los procesos democráticos es disponer de sistemas de votación que permitan realizar una votación referida a todos los candidatos. En este trabajo se propone un modelo lingüístico de votación y de recuento de votos denominada democracia representativa con votación ampliada que permite al ciudadano valorar a todos los candidatos de todas las listas.

Palabras claves: *toma de decisiones, computación con palabras, democracia representativa*

1. INTRODUCCIÓN

Numerosos estudios acerca de la democracia representativa han demostrado un aumento del desinterés y/o la desconfianza de los ciudadanos hacia los políticos y las instituciones políticas en general, y hacia la administración pública en particular [1], [4], [8], [10], [13]. Esto se ha dado tanto en los países de larga tradición democrática representativa [3], como en los países con democracias más jóvenes [11], donde existen altos niveles de descontento con este sistema de organización política, el que no tiene tanto que ver con los valores y el ideal democrático, sino más bien con el funcionamiento real e institucional de la democracia representativa [17].

En Estados Unidos (votar no es obligatorio) la participación en elecciones presidenciales pasó del 62,8% en 1960 al 48,9% en 1996 [16]. En las elecciones parlamentarias también se han registrado descensos en la participación, con un repunte en los últimos años, existiendo además una clara diferenciación entre los años en que las elecciones parlamentarias coinciden con las presidenciales y los años en que esta coincidencia no se da [9]. En la Unión Europea también se ha registrado una baja tasa de participación en las elecciones para el Parlamento Europeo, pasándose del 62% en 1979 al 43% en 2009.

Otra cuestión observado en las democracias representativas es la forma de realizar las votaciones, caracterizadas por los siguientes problemas: a) cuando se vota en una elección de concejales, diputados o senadores, se vota por una lista completa, es decir, el 100% del voto del ciudadano va a una lista en particular; b) el ciudadano no puede votar una lista parcial, no puede votar a una mezcla de candidatos de distintas listas para la misma categoría de cargos electivos; c) el ciudadano no puede indicar diferentes grados de preferencias para los distintos candidatos de la lista por la cual vota.

Una solución a los problemas mencionados debe permitir al ciudadano votar valorando a la totalidad de los candidatos, expresando su voluntad de manera sencilla y acorde a su manera de expresarse. En este sentido, la lógica borrosa se ha mostrado muy adecuada para resolver problemas de diferentes tipos, especialmente problemas de representación de información y Problemas de Toma de Decisiones (Decision Making Problem: DMP). El concepto de variables lingüísticas es ampliamente usado en aquellos problemas de toma de decisiones donde se dan valoraciones imprecisas (muy utilizadas en expresiones cotidianas) en una forma lingüística [5], [6], [7], [12], [14], [15], [18]. En este trabajo se propone un modelo lingüístico de votación y de recuento de votos denominada democracia representativa con votación ampliada que permite al ciudadano valorar a todos los candidatos de todas las listas. El trabajo se ha estructurado de la siguiente manera: en la sección 2 se presenta el modelo global de la que llamamos democracia representativa con votación ampliada; en la sección 3 se detalla el modelo de decisión propuesto; en la sección 4 se muestra un ejemplo; y finalmente se presentan las conclusiones.

2. DEMOCRACIA REPRESENTATIVA CON VOTACIÓN AMPLIADA

El modelo propuesto de Democracia Representativa con Votación Ampliada (DRVA) tiene por objetivo resolver los problemas de la votación tradicional: imposibilidad de votar a lista parcial,

imposibilidad de votar por candidatos de distintos partidos políticos para la misma categoría de cargos, imposibilidad de expresar distintos grados de preferencias. El proceso para llevar a cabo la DRVA comienza con el registro de los ciudadanos, a partir del cual la Autoridad Electoral (AE) producirá el Padrón Oficial de Ciudadanos habilitados para votar. A su vez, los partidos políticos efectúan la presentación de sus respectivas listas de candidatos ante la AE, que producirá las Listas Oficiales de Candidatos habilitados para ser elegidos. El día de las elecciones los ciudadanos habilitados votan expresando sus preferencias para la totalidad de los candidatos de las listas oficiales (votación ampliada); la votación podrá efectuarse de la manera tradicional o con medios electrónicos. Posteriormente la AE aplica un modelo de decisión para obtener la lista de candidatos elegidos (Figura 1).

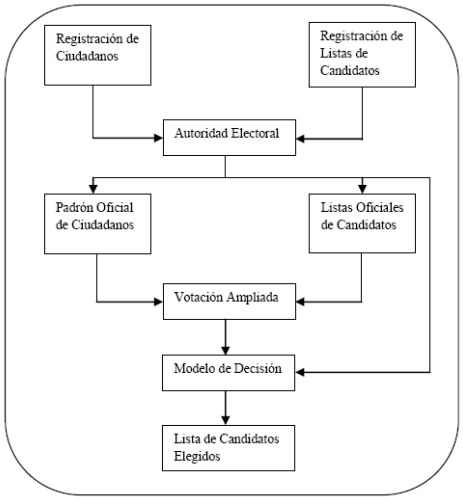


Figura 1: Modelo global de la democracia representativa con votación ampliada.

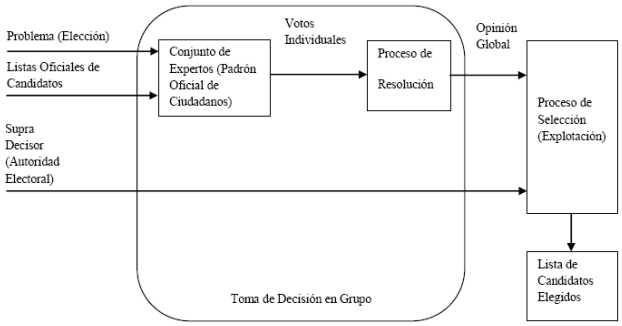


Figura 2: Modelo de decisión.

3. MODELO DE DECISIÓN

En el modelo de decisión propuesto se deben aplicar dos procesos para obtener una solución final: resolución y selección (Figura 2). El objetivo es combinar las opiniones individuales para producir una solución global satisfactoria. Para ello se define el modelo de democracia representativa con votación ampliada (DRVA), que aplica técnicas de conjuntos difusos para tratar con términos lingüísticos usados en las evaluaciones (votos) de los ciudadanos. Los ciudadanos evalúan a *todos* los candidatos de *todas* las listas usando lenguaje natural. Estas evaluaciones se combinan (agregan) en el proceso de resolución. Finalmente se aplica una fase de explotación para obtener una lista de candidatos elegidos, donde para resolver conflictos de igualdad de evaluación se usan las cardinalidades de las etiquetas lingüísticas de mayor intensidad de preferencia. El Modelo de Decisión se define en cuatro niveles:

Nivel 1: El escenario de decisión. Determinar las expresiones de los evaluadores y las jerarquías lingüísticas. Se supone que se tiene un Padrón Oficial de Ciudadanos $E = \{e_1, \ldots, e_m\}$, un conjunto de Listas Oficiales de Candidatos de los Partidos Políticos $P = \{p_1, \ldots, p_j\}$, cada una de ellas con un conjunto de candidatos $C = \{c_1, \ldots c_i\}$ a evaluar para cubrir *i* cargos electivos $(m \geq 1, j \geq 1, i \geq 1, n = i\,.\,j)$.

Nivel 2: Evaluación individual. Los ciudadanos individualmente expresan sus valoraciones para cada opción de acuerdo al método definido anteriormente: $V_{e_i} = \{v_{i1}, \ldots, v_{in}\}$ con $v_{ij} \in L$ donde L es el conjunto de etiquetas establecido previamente por la Autoridad Electoral (Tabla 1).

Etiquetas	**Semánticas**
BE: La mejor	0,75; 1; 1
GO: Buena	0,50; 0,75; 1
ME: Media	0,25; 0,50; 0,75
BA: Mala	0; 0,25; 0,50
WO: La peor	0; 0; 0,25

Tabla 1: Grupo de etiquetas lingüísticas.

Nivel 3: Evaluación global. La evaluación global se obtiene usando las evaluaciones individuales, y calculando un valor medio, en este caso se propone la utilización de la media aritmética del conjunto de etiquetas atribuido a cada candidato C_i de cada lista (partido) P_j, mediante el modelo computacional simbólico [2]:

$$\emptyset(C_iP_j) = \frac{v_{1ij} \oplus \ldots \oplus v_{tij}}{t} \ (t\text{: total de votos para } C_iP_j)$$

Nivel 4: Explotación. Esta fase transforma la información global acerca de los candidatos en un orden global de los mismos, desde el cual se obtiene la Lista de Candidatos Elegidos. Si varios candidatos han obtenido la misma valuación global, entonces la Autoridad Electoral (supra-decisor) los ordena según la cardinalidad obtenida por los candidatos en las etiquetas que indiquen mayor intensidad de preferencia. Se considera ganadores a los *i* primeros candidatos de la lista ordenada, ya que se deben cubrir *i* cargos.

4. EJEMPLO

Nivel 1. Se deben elegir 3 diputados; se presentan a las elecciones 3 partidos políticos, c/u con su lista de 3 candidatos; las respectivas listas son oficializadas por la Autoridad Electoral, al igual que le Padrón de votantes.

Nivel 2. Los ciudadanos votan de acuerdo a un conjunto de etiquetas lingüísticas L_l (l = 1,…, 5) cuya semántica se muestra en la Tabla 1. Los votos obtenidos se muestran en la Tabla 2. C_iP_j: significa candidato i del partido j (i=1, .., 3; j=1, …, 3); CD_k: significa conjunto de decisores (votantes) que han votado de igual manera (k=1, …, 8).

	Card	C_1P_1	C_2P_1	C_3P_1	C_1P_2	C_2P_2	C_3P_2	C_1P_3	C_2P_3	C_3P_3
CD_1	10	BE	WO	GO	BA	BE	GO	WO	WO	BE
CD_2	3	GO	ME	BE	WO	ME	WO	BA	GO	ME
CD_3	5	BA	BA	WO	BA	GO	BA	ME	BA	ME
CD_4	4	ME	BA	ME	BE	BA	GO	WO	BE	BA
CD_5	3	ME	BA	BE	ME	WO	BE	WO	GO	BE
CD_6	1	WO	BE	BA	BA	WO	ME	BE	BA	GO
CD_7	2	GO	WO	GO	BA	BE	BA	GO	GO	GO
CD_8	6	ME	ME	WO	BE	BA	GO	GO	ME	ME

Tabla 2: Votos obtenidos.

La cantidad de votos emitidos es la siguiente:

$$\sum_{k=1}^{k} \text{Card}_k = 10 + 3 + 5 + 4 + 3 + 1 + 2 + 6 = 34$$

Nivel 3. Se calculan las valoraciones globales obtenidas por cada candidato de cada partido (C_iP_j), según lo establecido en el proceso de resolución; se obtienen las agregaciones que se muestran en la Tabla 3.

C_1P_1	C_2P_1	C_3P_1	C_1P_2	C_2P_2	C_3P_2	C_1P_3	C_2P_3	C_3P_3
GO	BA	ME	ME	ME	ME	BA	BA	GO

Tabla 3: Valoraciones globales obtenidas por los candidatos.

Nivel 4. Los candidatos que han obtenido en la agregación la mejor valoración global son: GO: C_1P_1, C_3P_3, ME: C_3P_1, C_1P_2, C_2P_2, C_3P_2. Las cardinalidades obtenidas por los candidatos en la valoración de máximo nivel (BE) se muestran en la Tabla 4.

Candidato	Cardinalidad (BE)
C_1P_1	10
C_3P_1	6
C_1P_2	10
C_2P_2	12
C_3P_2	3
C_3P_3	13

Tabla 4: Cardinalidades obtenidas para el mayor nivel de preferencia.

Resulta la siguiente Lista de Candidatos Elegidos: 1) C_3P_3 (Candidato 3 del Partido 3); 2) C_1P_1 (Candidato 1 del Partido 1); 3) C_2P_2 (Candidato 2 del Partido 2).

5. CONCLUSIONES

Este trabajo presenta un nuevo modelo de votación para la democracia representativa llamado Democracia Representativa Con Votación Ampliada (DRCVA). La votación ampliada consiste en que los ciudadanos pueden votar por la totalidad de los candidatos de todos los partidos políticos que han oficializado listas, utilizando etiquetas lingüísticas para expresar sus valoraciones. En trabajos futuros se pretende modelar el proceso de agregación de la votación mediante operadores de mayoría y utilizar un sistema de valoración con diferentes dominios de expresión para los ciudadanos, de manera que cada uno pueda utilizar su propio dominio lingüístico de valoración.

REFERENCIAS

[1] S. CLIFT, *E-Democracy, E-Governance and Public Net-Work*, http://www.publicus.net/articles/edempublicnetwork.html. Consultado el 14/09/09, 2003.

[2] M. DELGADO, J. L. VERDEGAY & M. A. VILA, *On aggregation operations of linguistic labels*. International Journal of Intelligent Systems 8, pp 351-370, 1993.

[3] EUROPEAN COMISSION, *Eurobarometer 61, Spring 2004*: *public opinion in the European Union*, European Opinion Research Group EEIG, pp. B61, C20, 2004.

[4] M. HAGEN, *A Typology of Electronic Democracy. University of Giessen*. http://www.unigiessen.de/fb03/vinci/labore/netz/hag_en.htm. Consultado el 14/09/2009, 1997.

[5] F. HERRERA & L. MARTÍNEZ, *A 2-tuple fuzzy linguistic representation model for computing with words*. IEEE Transactions on Fuzzy Systems, 8(6): 746-752, 2000.

[6] F. HERRERA & L. MARTÍNEZ, *A model based on linguistic 2-tuples for dealing with multigranularity hierarchical linguistic contexts in multiexpert decision-making*. IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics. 31(2): 227-234, 2001.

[7] E. HERRERA-VIEDMA, L. MARTÍNEZ, F. MATA & F. CHICLANA, *A consensus support system model for group decision-making problems with multigranular linguistic preference relations*. IEEE Transactions on Fuzzy Systems, 13(5): 644-658, 2005.

[8] M. HUSE, *Renewing Management and Governance: New Paradigms of Governance?*, Journal of Management and Governance 7:211-221, 2003.

[9] INTERNATIONAL IDEA: INTERNATIONAL INSTITUTE FOR DEMOCRACY AND ELECTORAL ASSISTANCE. *Voter Turnout per Country*. http://www.idea.int. Consultado el 28/05/2010, 2010.

[10] A. JAIN, *Using the lens of Max Weber´s Theory of Bureaucracy to examine E-Government Research.*, System Sciences. Proceedings of the 37th Annual Hawaii International Conference on (pgs. 127-136), 2004.

[11] LATINOBARÓMETRO, *Informe-resumen Latinobarómetro 2004: una década de mediciones*, Latinobarómetro, pp. 4, 33, 2004.

[12] J. LU, G. ZHANG, D. RUAN & F. WU, *Multi-objective Group Decision Making: Methods, Software and Applications with Fuzzy Set Technology*. London: Imperial College Press, 2007.

[13] W. NISKANEN, *Cara y Cruz de la burocracia*. Espasa-Calpe. Madrid. España, 1980.

[14] J. I. PELÁEZ & J. M. DOÑA, *LAMA: A Linguistic Aggregation of Majority Additive Operator*, International Journal of Intelligent Systems 18, 809-820, 2003.

[15] J. I. PELÁEZ, J. M. DOÑA & J. A. GÓMEZ-RUIZ, *Analysis of OWA Operators in Decision Making for Modelling the Majority Concept.* Applied Mathematics and Computation. Vol. 186. Pages 1263-1275, 2007.

[16] R. PUTNAM, *Bowling alone: the collapse and revival of American*. Touchstone. Nueva York, 2002.

[17] J. SUBIRATS, *Los dilemas de una relación inevitable. Innovación Democrática y tecnologías de la información y de la comunicación*. http://www.democraciaweb.org/subirats.PDF (Consultado el 14/09/2009), 2002.

[18] L. ZADEH, *Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic*. Fuzzy Sets and Systems, 90:111-127, 1997.

THE FULL STRATEGY MINORITY GAME

Gabriel Acosta[♭], Inés Caridi[♭], Sebastián Guala[†] and Javier Marenco[†]

[♭] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, , Pabellón I, Ciudad Universitaria, (1428) Buenos Aires, Argentina, gacosta@dm.uba.ar, ines@df.uba.ar, tha authors are members of Conicet*

[†] *Instituto de Ciencias, UNGS, , J. M. Gutiérrez 1150, (1613) Los Polvorines, Argentina. , sguala@ungs.edu.ar, jmarenco@ungs.edu.ar*

Abstract: The Full Strategy Minority Game ($FSMG$) is an instance of the Minority Game (MG) which includes a single copy of every potential agent. In this work, we explicitly solve the $FSMG$ thanks to certain symmetries of this game. Furthermore, by considering the MG as a statistical sample of the $FSMG$, we compute approximated values of the key variable σ^2/N in the symmetric phase for different versions of the MG. We also show that the $FSMG$ verifies a strict period two dynamics (i.e., period two dynamics satisfied with probability 1) giving, to the best of our knowledge, the first example of an instance of the MG for which this feature can be analytically proved. Thanks to this property it is possible to give a simple way of computing the probability that a general instance of the MG verifies the period two dynamics.

Keywords: *Minority Game, Period Two Dynamics, Updating Rule*
2000 AMS Subject Classification: 91A99

1 INTRODUCTION

The Minority Game (MG) was introduced in 1997 by Challet and Zhang [1] in an attempt to catch essential characteristics of a competitive population in which an individual achieves the best result when she manages to be in the minority group. In the MG, there are N agents (usually odd), that at each step of the game must choose 0 or 1. Let $N_0(t)$ (resp. $N_1(t)$) be the number of agents choosing 0 (resp. 1) at the step t (note that $N_0(t) + N_1(t) = N$). The winners are those who happen to be in the minority group (i.e., the minimum between $N_0(t)$ and $N_1(t)$). The only information available for the agents is the system *state* $\mu \in \{0,1\}^m$, that is updated after each step of the game. In the classical version of the MG, μ is determined from the sequence of minority sides in the last m steps, although other kinds of *updating rules* can be found in the literature [2, 3]. Therefore, the number of possible states is $\mathcal{H} = 2^m$. Agents play using the so-called *strategies*. A strategy is a function that assigns a prediction (0 or 1) for each of the possible states. In this way, there are $\mathcal{L} = 2^{\mathcal{H}}$ different strategies. Each agent has s strategies at her disposal (we use $s = 2$ in this work), randomly chosen with replacement from the complete set of strategies at the beginning of the game (note that it is possible for an agent to have two identical strategies, and for two agents to have the same pair of strategies). At every step of the game, each of the strategies that correctly predicted the winning side is awarded a *virtual point*, regardless of use in that step. At each step, each agent plays what her best-performing strategy (in terms of virtual points) predicts. If the two strategies have the same number of virtual points, the agent randomly chooses one of them.

Since the Minority Game brings an elementary model of personal beliefs and subjectivities to be inductively confirmed or modified, the well-known and most widely extended application of this model is to financial markets dynamics where the binary option reduces to buy or sell. However, it can also be applied to little everyday behaviors such as election of two alternative highways, the best moment in the day or in the month to go to the supermarket (or cash machine) to avoid long queues, etc.

Given N agents, an *instance* of the MG is a particular assignment of strategies to the agents. We define a *configuration* $\mathcal{E}$ of the game to be a pair $\mathcal{E} = \{\mathcal{M}, I\}$, where $\mathcal{M} = \{\tilde{\mu}^1, \tilde{\mu}^2, ...\}$ is a sequence of states (generated by any updating rule) and I is an instance of the MG.

The observable $z = < (N_1 - N/2)^2/N >_{\mathcal{E}}$ is the most studied and instructive variable [5] in the MG. It measures the population's waste of resources by averaging on time and over different configurations $\mathcal{E}$, the quadratic deviation of the number of agents that chose a fixed side (for example N_1) from $N/2$. The notation

$\sigma^2/N = z$ is usual in the literature. One of the reasons why the MG has attracted so much attention is that for certain values of the parameters m, N, and s, the variable z is smaller than that obatined for a game in which each of the N agents randomly chooses between the two sides.

Calling $\alpha = \mathcal{H}/N = 2^m/N$, it is known that in the region given by $\alpha \ll 1$, *crowd effects* arise at some game steps. This behavior is related to a dynamics known as *Period Two Dynamics* (PTD), which was observed for the first time by Savit et al. [4]. To understand this dynamics, that plays an important role in the rest of the article, for each game step t we define the *parity array* $\mathcal{P}^t_{\mathcal{E}}$ to be an array of categorical variables recording the parity (odd or even) of the number of appearences of each state in the first $t-1$ game steps. More precisely, we have $\mathcal{P}^t_{\mathcal{E}} \in \{O, E\}^{\mathcal{H}}$ and if we identify any state μ with the integer number given by the binary expansion of μ plus one (so that μ can be thought as an integer ranging from 1 to $\mathcal{H}$), we have that $\mathcal{P}^t_{\mathcal{E}}(\mu) = O$ (resp. E) if μ has appeared an odd (resp. even) number of times in the first $t-1$ steps of the game. We assume $\mathcal{P}^1_{\mathcal{E}} = (E, E, \cdots, E)$, as at the beginning of the game any state has appeared zero times, hence an even number of times. We will drop t when referring to a generic time step, using $\mathcal{P}_{\mathcal{E}}(\mu)$ instead.

The PTD can be summarized in the following way: if at some time step, for some state μ we have $\mathcal{P}_{\mathcal{E}}(\mu) = O$, then in the next (and hence even) appearance of μ, the outcome of the game is *very likely* to be the opposite to that obtained in the previous appearance of μ. Broadly speaking, this dynamics is due to the fact that on even appearances of μ crowds of agents will move together to the side rewarded in the previous odd appearance. When crowds emerge in the game, their contribution to z is very important. Furthermore, crowd effects are the reason why z is a large number in this region, showing that fewer resources are allocated to the population as a whole.

In a little more than a decade, there have been many attempts from different backgrounds to give a formal framework to the game, and to analytically reproduce the results observed in computer simulations [5]. In [6] a mean field approach to the MG is presented and the standard deviation of the key variable $N_1 - N_0$ is computed by introducing a simplified framework. The key idea in [6] consists in defining a particular instance of the game so that, for any value of m, all possible strategies and all possible agents (each one represented by a possible pair of strategies in the case $s = 2$) take part in the game. We call this particular instance the *Full Strategy Minority Game* ($FSMG$). For $s = 2$, the number of *potential* agents is $\mathcal{N} = \binom{\mathcal{L}}{2} + \mathcal{L}$, where the first term represents all agents with two different strategies, and the second term represents the number of agents whose two strategies are identical. Thus, the number of agents of the $FSMG$ is a function of m, $\mathcal{N} = \mathcal{N}(m)$. Certain symmetries which appear only partially in the MG can be fully exploited in the $FSMG$, and this approach leads to interesting theoretical results in the PTD region.

Here, we show that the calculations given in [6] can be highly simplified and easily applied to other variants of the MG, as the above-mentioned MG_{rand} and MG_{per}. By taking advantage of its inherent symmetries, we analytically solve the $FSMG$. On the other hand we define the Strict Period Two Dynamics (SPTD) as a PTD with probability 1, and show that SPTD is the characteristic dynamics of the $FSMG$.

2 Analytical Results for different updating rules

If we take an arbitrary number N of agents it is clear that, in randomly generated instances I of the MG, the following may happen: *(1)* some potential agent may not participate in the game or *(2)* a multiple copy of the same agent may participate in the game. In the Full Strategy Minority Game both cases are excluded: by construction the number of agents in the $FSMG$ is set to $\mathcal{N}$ and a single copy of every potential agent is allowed. $\mathcal{S}_{\mathcal{H}}$ and $\mathcal{S}_{\mathcal{L}}$ (also known as the *Full Strategy Space*) denote the set of states and strategies respectively. Throughout this article, the symbol $\sharp$ stands for the cardinality of a set, hence we have $\sharp\mathcal{S}_{\mathcal{H}} = \mathcal{H}$, and $\sharp\mathcal{S}_{\mathcal{L}} = \mathcal{L}$. For a given state $\mu \in \mathcal{S}_{\mathcal{H}}$, the subset of strategies in $\mathcal{S}_{\mathcal{L}}$ that predict a certain outcome $\tilde{o}$ for the state μ is denoted by $\mathcal{S}_{\mathcal{L},\mu\to\tilde{o}}$. For an arbitrary outcome $\tilde{o} \in \{0, 1\}$ we will denote the opposite side by $\sim \tilde{o}$. It is clear that $\mathcal{S}_{\mathcal{L},\mu\to\tilde{o}} \cup \mathcal{S}_{\mathcal{L},\mu\to\sim\tilde{o}} = \mathcal{S}_{\mathcal{L}}$, and that $\sharp\mathcal{S}_{\mathcal{L},\mu\to\tilde{o}} = \sharp\mathcal{S}_{\mathcal{L},\mu\to\sim\tilde{o}} = \mathcal{L}/2$. This means that for each state μ, the number of strategies predicting $\tilde{o}$ and the number of strategies predicting $\sim \tilde{o}$ coincide. This symmetry together with the assumption of the SPTD allow us to make a remarkable analytic simplification of the game. Indeed, let us consider an arbitrary configuration $\mathcal{E}$ of the $FSMG$, and its state sequence $\mathcal{M}$ (note that there exists only one instance I for each m in the $FSMG$). Suppose that at step s a certain state $\tilde{\mu}^s = \mu$ occurs for the first (and hence odd) time, and call $\tilde{o}$ the winning side after the

voting round s. In that case the parity array verifies $\mathcal{P}^{s+1}(\mu) = O$ (since $\mathcal{P}^{l}(\mu) = E$ if $l \leq s$) and strategies belonging to the set $\mathcal{S}_{\mathcal{L},\mu\to\tilde{o}}$ are rewarded with a virtual point. Suppose now that at the time step s', $s' > s$, μ occurs for the second time (i.e., $\tilde{\mu}^{s'} = \mu$), then the SPTD implies that the winning side after the voting round s' will be $\sim \tilde{o}$, and thus exactly the *other half* of the strategies ($\mathcal{S}_{\mathcal{L},\mu\to\sim\tilde{o}}$) will correctly predict the minority side. Therefore, if we remove the point previously assigned to the set of strategies $\mathcal{S}_{\mathcal{L},\mu\to\tilde{o}}$ instead of adding a new point to strategies belonging to $\mathcal{S}_{\mathcal{L},\mu\to\sim\tilde{o}}$, the dynamics of the game will remain unchanged. Taking into account this rule, it is clear that the number of virtual points accumulated for any strategy, at any time step, ranges from 0 to $\mathcal{H}$. Remarkably we are able to show that the $FSMG$ necessarily *verifies* the SPTD, and hence the SPTD assumption can be removed in the previous argument. As far as we know this is the first example in the literature of a game enjoying this property.

We apply the $FSMG$ framework to different kinds of updating rules. Our two first examples are based on the exogenous updating rules found in MG_{rand} [2] and MG_{per} [3], finally we address the standard MG. Since the $FSMG$ verifies the SPTD and our calculations rely on the $FSMG$, *we expect to find good results only in the region of validity of the PTD*.

In the MG_{rand}, the present state μ_p is chosen at random (uniformly) from the whole set of states $\mathcal{S}_{\mathcal{H}}$. In this case we can show

$$\sigma^2_{MG_{rand}}/N = 1/4 + \frac{N-1}{8(1+2^{-\mathcal{H}})^2 2^{\mathcal{H}-1}} \sum_{n_0=1}^{\mathcal{H}} \left(1/2^{2n_o}\binom{2n_o}{n_o}\right)^2 \binom{\mathcal{H}-1}{n_o-1}. \tag{1}$$

which can be remarkably approximated by

$$\sigma^2_{MG_{rand}}/N \sim 1/4 + \frac{N}{4\mathcal{H}\pi}. \tag{2}$$

In Figure 1 we show the analytical result given by (1), the approximated expression (2), and the numerical results for the MG_{rand} with $N = 4001$.

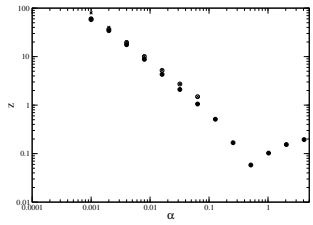


Figure 1: Filled circles show z as a function of α for the MG_{rand} for different values of m (from 2 to 14) and $N = 4001$. For each value of N and m, 100 runs have been performed, each one of $T = 100000$ time steps discarding the first 50000 steps. Empty circles show the analytical result (1), and star-shaped symbols show the approximated invariant (i.e., depending only on α) expression (2).

We also apply our calculations to the MG_{per} introduced in [3]. In this case the updating rule follows a periodic pattern of period $\mathcal{H}$ that runs over all the states. In fact, with our convention that identifies any state μ with its binary expansion plus one, the updating rule proposed in [3] follows the natural order $1 \to 2 \to 3 \cdots$ modulo $\mathcal{H}$. In this case we found

$$\sigma^2_{MG_{per}}/N = 1/4 + \frac{(N-1)}{8\mathcal{H}(1+2^{-\mathcal{H}})^2} \sum_{n_o=1}^{\mathcal{H}} \left(\frac{1}{2^{2n_o}}\binom{2n_o}{n_o}\right)^2. \tag{3}$$

In Figure 2 we show the agreement between (3) and numerical experiments for the MG_{per} with $N = 4001$. In much the same way as for the MG_{rand}, this expression can be highly simplified as

$$\sigma^2_{MG_{per}}/N \sim 1/4 + \frac{N}{8\mathcal{H}\pi}(\log \mathcal{H} + \gamma), \tag{4}$$

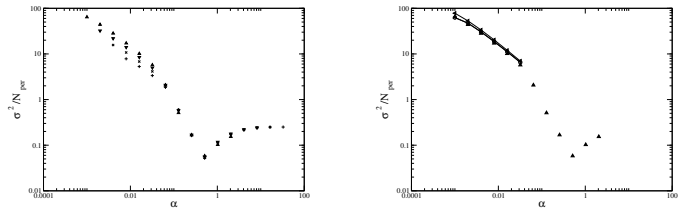


Figure 2: On the left, z as a function of α for the MG_{per} for different values of m (in the range from 2 to 14) and $N =:$ + symbols for $N = 501$; $\times$ $N = 1001$; ▼ $N = 2001$; ▲ $N = 4001$. For each value of N and m we perform 100 runs, each one of $T = 100000$ time steps discarding the first 50000 steps. On the right, the MG_{per}, ▲ with $N = 4001$, empty circles are given by (3), and ◀ show the approximated expression (4).

where $\gamma = 0.57\ldots$ is the constant of Euler-Mascheroni. Equation (4) is essentially the same obtained in [7] by means of a different approach.

Finally, we turn our attention to the standard MG. In this case certain conditional probability is numerically obtained from the De Bruijn diagram corresponding to each m [8], using this together with our analytic results for the $FSMG$ we obtain the approximations given in Figure 3 (Center). The $FSMG$ also allow us to analytically approximate the PTD, see Figure 3 (Right).

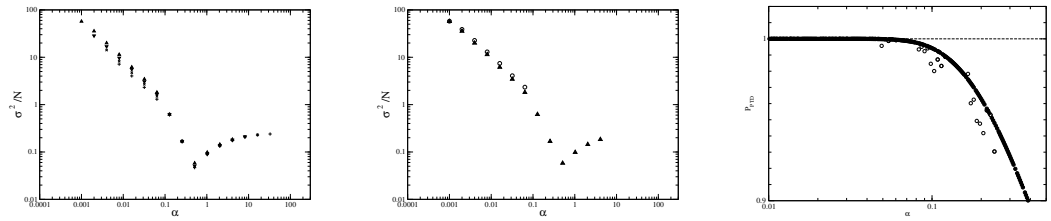


Figure 3: **Left**: z as a function of α in the MG for different values of m (in the range from 2 to 14) and $N =:$ + $N = 501$; $\times$ $N = 1001$; ▼ $N = 2001$; ▲ $N = 4001$. For each value of N and m we performed 100 runs, each one of $T = 100000$ time steps discarding the first 50000 steps. **Center**: Numerical MG (▲), and analytical result (○) for $N = 4001$. **Right**: Probability P_{PTD} for PTD to take place in even occurences of the states as a function of α. In filled circles: the analytic case (by using the approximation $n_o = \mathcal{H}/2$). Empty circles: results of the simulation of the MG. In the simulations we have computed P_{PTD} as follows: at each even occurrence of each state, if after the poll the PTD is not fulfilled (i.e., the minority side agrees with that obtained after the previous –odd– occurrence of the same state), then we consider that this step does not contribute to P_{PTD}. At the same time, present state and virtual points are assigned as if the PTD had not failed. This way we compute the probability of breaking the PTD for *the first time in the game*. The empty circles show the average of 100 runs of 100000 steps each, for $N = 501$, 531, 561, 591, 621, 651, 681, 711, 741, 771 and $m = 2, \ldots, 7$

REFERENCES

[1] D. Challet, Y. C. Zhang, Emergence of cooperation and organization in an evolutionary game, Physica A 246 (1997) 407.

[2] A. Cavagna, Irrelevance of memory in the minority game, Phys. Rev. E 59 (1999) R3783.

[3] SS. Liaw, C.Liu, The quasi-periodic time sequence of the population in minority game, Physica A 351 (2005) 571.

[4] R. Manuca, Y. Li, R, Riolo, R. Savit, The structure of adaptive competition in minority game, Physica A 282 (2000) 574.

[5] D. Challet, M. Marsili, Y. C. Zhang, Minority Games, Oxford University Press (2005).

[6] I. Caridi, H. Ceva, Minority game: a mean-field-like approach, Physica A 317 (2003) 247.

[7] S.S Liaw, Ch. Hung, Ch. Liu, Three phases of the minority game, Physica A 374 (2007) 359.

[8] D. Challet, M. Marsili, Relevance of memory in minority games, Phys Rev. E 62 (2000) 1862.

ESTUDIO DE LA VARIACIÓN EN LA COMPLEJIDAD DE REDES QUE EVOLUCIONAN EN EL TIEMPO

L. Catalano y A. Figliola

Universidad Nacional de General Sarmiento, Instituto del Desarrollo Humano, J. M. Gutierrez 1150, Los Polvorines, Pcia. de Bs. As., Argentina, www.ungs.edu.ar

Resumen: La evolución temporal de las redes sociales es un tema de interés en la actualidad. El objetivo del presente trabajo es estudiar las repercusiones que tiene el estado inicial de una red en el futuro de la misma luego de aplicarle un regla de evolución ya conocida (modelo de Barabási-Albert). Se realizaron pruebas partiendo de redes iniciales con diferentes características claramente definidas tanto en relación a la distribución de las conexiones como al tamaño de la red y calculando un coeficiente que permite medir la complejidad (*Medium articulation*). Luego de observar los resultados se concluyó que, sin importar el tamaño inicial de la red, dependiendo sólo de la distribución de las conexiones, la misma conserva ciertas particularidades que preserva con el correr del tiempo.

Palabras clave: *redes sociales, redes de pequeño mundo, redes complejas.*
2000 AMS Subject Classification: 91D30 - 05C82

1. INTRODUCCIÓN

Podemos pensar al estudio de las redes complejas como aquel que aborda los temas de la intersección entre la teoría de grafos y la mecánica estadística, que produjo una nueva disciplina. En los últimos años, la aparición de redes sociales, tales como Internet, World Wide Web, Facebook, Twitter y otras, provocó un verdadero salto cualitativo en la investigación en torno a ellas, ya que, por un lado se dispone de un verdadero laboratorio "real" donde se pueden observar y verificar los modelos que produce esta nueva rama de la ciencia y por otro, se empieza a producir una necesidad concreta en la resolución de problemas, que impulsa su estudio.

En cuanto al origen de esta nueva interdisciplina, se pueden mencionar tres trabajos pioneros y a la vez fundacionales: el de Watts y Strogatz sobre redes de "pequeño mundo" (*small-world networks*) [7], los trabajos de Barabási sobre modelos de escala libre [1], [2] y el paper de Girvan and Newman sobre la identificación de las estructuras comunitarias presentes en muchas redes [4].

La caracterización tradicional de las redes se orientó en un principio a sus cualidades topológicas, por lo que la bibliografía existente utiliza cuantificadores para caracterizar y obtener información sobre la distribución de las conexiones o nodos, tales como: *distrubución de grado, camino más corto, clustering coefficient* (ver para más detalles [5]). La bibliografía actual también describe modelos de crecimiento de redes a partir de una configuración inicial.

En suma, existe en la actualidad un gran interés por modelizar el comportamiento de las redes sociales, tanto en forma estática, como dinámica a partir de una evolución temporal. Este último punto es el menos explorado y en el cual nos centramos para este trabajo. No obstante, para estudiar la evolución de las redes necesitaremos poder caracterizar ciertas particularidades y atributos de ellas en forma estática. Una primera diferenciación que se puede hacer entre las redes estáticas es la de redes con conexiones pesadas o no pesadas y direccionadas o no direccionadas, cualquiera sea la situación, son representadas con la matriz de adyacencia.

En los últimos años surgieron nuevas formas de caracterizar a las redes mediante el cálculo de medidas de complejidad y entropía. T. Wilhelm y J. Hollunder [6] propusieron en los ultimos años una medida de complejidad que se retomó para este trabajo. Dicha medida consiste en calcular un coeficiente llamado *medium articulation* (MA) sobre la red en cuestión, luego calcularlo para una red con igual cantidad de nodos y conexiones construida en forma aleatoria (MA_r); una vez obtenidos estos datos, se dice que la red es compleja si y solo si $MA > MA_r$. En el mismo trabajo se hace otra caracterización de las redes, a partir de la manera en que se distribuyen sus conexiones: *redes democráticas* y *redes dictatoriales*. La manera de caracterizarlas numéricamente es similar a la realizada para estudiar la complejidad, pero utilizando otro

coeficiente para la comparación, *coeficiente de redundancia* (R). Escencialmente, las redes dictatoriales son aquellas en las que algunos pocos nodos ocupan roles preponderantes en la red (en relación con la cantidad de conexiones que tienen con otros nodos); por el contrario, las redes democráticas son aquellas en las cuales todos los nodos poseen una distribución de conexiones homogénea [ver Figura 1].

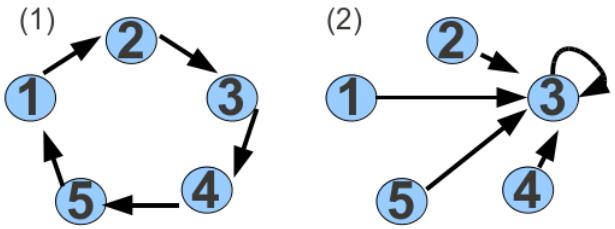


Figura 1: (1) Ejemplo de red democrática. (2) Ejemplo de red dictatorial.

Consideraremos ahora el aspecto dinámico de las redes. En los últimos años del siglo XX, Barabási popularizó el término de *conexión preferencial* a la hora de pensar en la evolución de redes. Este idea propone que en una red, la probabilidad de que un nuevo nodo se conecte con uno ya existente no es uniforme, sino que está dada por:

$$P_i = \frac{k_i}{\sum_j k_j}$$

Donde P_i es la probabilidad de que el nuevo nodo se conecte con el nodo i y k_l es el grado (cantidad de conexiones) del nodo l ya existente en la red.

De esta manera se obtienen redes que comparten algunas características con ciertas redes sociales reales (Scale Free, [3]). En este trabajo estudiamos cómo evolucionan las redes en función de alguna característica de su estado inicial.

2. Descripción de la obtención de los datos

Trabajamos con redes cuyas conexiones son direccionadas y no pesadas [6]. El estudio fue realizado con redes iniciales de diferentes tamaños con igual número de nodos y conexiones (3, 9, 15, 21 y 27 nodos y conexiones). En cada caso se utilizaron tres tipos de redes iniciales, uno puramente democrático, uno puramente dictatorial y uno aleatorio; los dos primeros pueden visulizarse en la Figura 1, el tercero consiste en construir redes con n cantidad de nodos y n conexiones distribuidas en forma aleatoria con la única restricción de que no queden nodos desconectados. Luego, se procedió a hacer evolucionar a cada una de ellas de la siguiente manera: se agregaron 200 nodos de a uno por vez utilizando el modelo de evolución de BA [2], calculando MA en cada paso para obtener así, una sucesión de 200 datos de MA para cada una de las redes. Debido a que el algoritmo que se utiliza para agregar nodos es probabilístico, la experiencia se repitió diez veces obteniendo un conjunto de datos promedio para cada caso. Con el objetivo de obtener una medida de complejidad, también se calculó en cada paso MA_r sobre redes generadas en forma aleatoria con tamaño y cantidad de conexiones similares a las de los casos a comparar.

3. Organización y análisis de los datos

Como descripción general de los datos obtenidos, se puede observar que los gráficos de evolución de MA tiene un comportamiento similar para todos los tamaños y tipos de redes iniciales. Dicho coeficiente aumenta hasta llegar a un máximo y luego decrece lentamente. No se observa un detenimiento completo en el decrecimiento (ver Figura 2).

Es importante recordar que el hecho de que MA descienda al evolucionar la red no implica que la compejidad disminuya, ya que esta surge de la comparación con MA_r. Estudiaremos cuáles de las siguientes cuestiones están ligadas a las características de las redes iniciales:

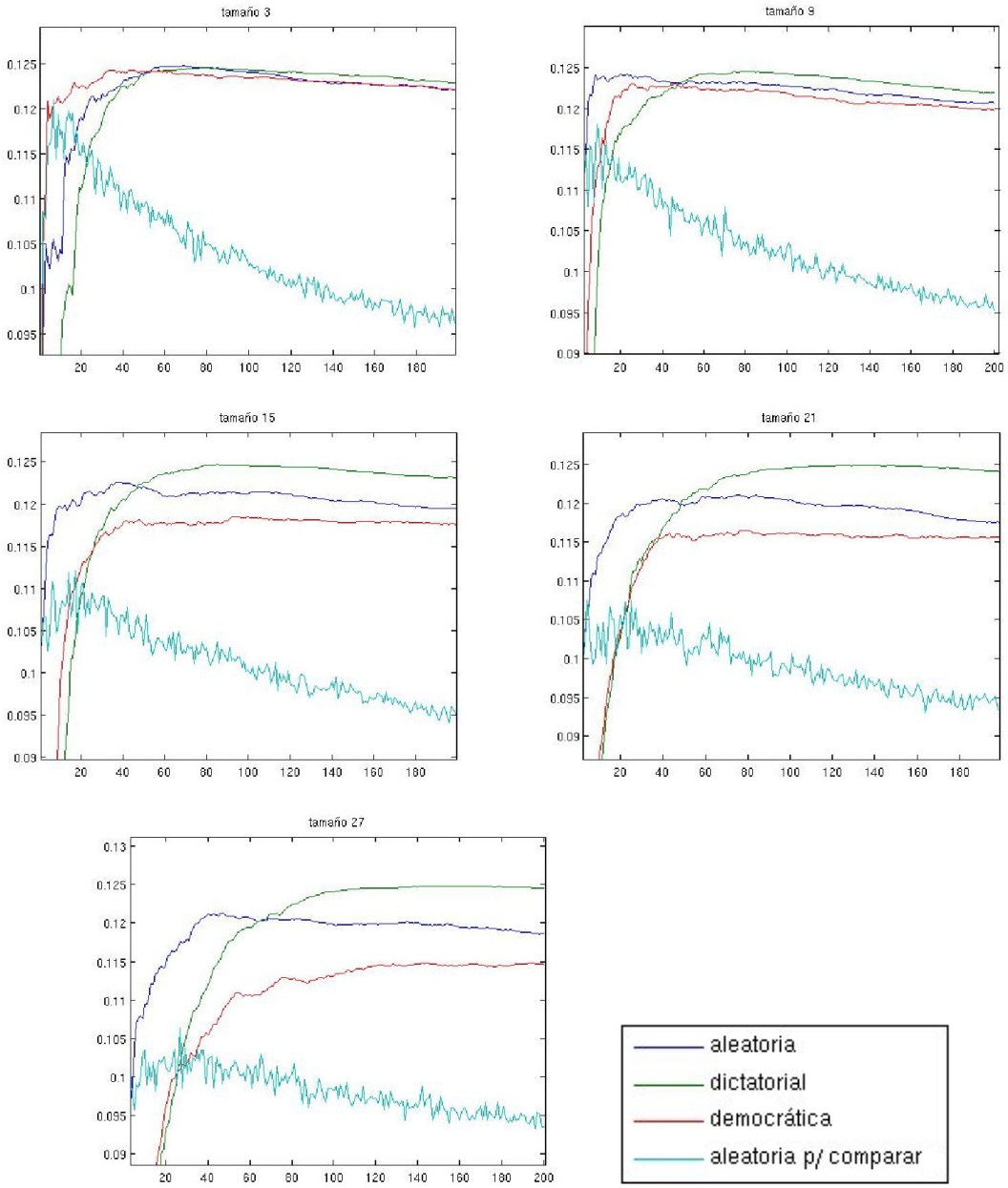


Figura 2: Gráficos de MA vs cantidad de nodos de la red. El tamaño de la red indicada sobre los gráficos refiere al tamaño de la red al comenzar la evolución (tamaño inicial). Las tres primeras indicaciones de la leyenda hacen referencia los estados iniciales de las redes en cada caso; la cuarta indicación refiere a MA_r.

- el momento (cantidad de nodos agregados) en que se alcanza el máximo valor de MA.
- el valor máximo que alcanza el coeficiente MA.
- el valor de MA alcanzado luego de agregar una gran cantidad de nodos.

4. CONCLUSIONES

En los gráficos de la Figura 2 se puede observar como, independientemente del tamaño inicial de la red, las mismas conservan ciertas características aún luego de largos períodos de tiempo. En todos los casos, las redes que parten de un estado inicial puramente democrático adquieren valores de MA menores que en los

tamaño	aleatoria	dictatorial	democrática
3	0.1225	0.1233	0.1225
9	0.1210	0.1224	0.1202
15	0.1198	0.1235	0.1178
21	0.1182	0.1245	0.1156
27	0.1192	0.1247	0.1146

Tabla 1: Valor de MA promedio en las últimos 50 pasos de evolución para redes con las características y tamaños iniciales indicadas en la primer fila y columna respectivamente.

tamaño	aleatoria		dictatorial		democrática	
	x	y	x	y	x	y
3	69	0.1248	80	0.1246	35	0.1243
9	20	0.1242	81	0.1245	26	0.1230
15	37	0.1225	85	0.1247	92	0.1185
21	75	0.1211	136	0.1249	78	0.1165
27	46	0.1213	161	0.1248	145	0.1149

Tabla 2: Valor de abscisa (x) y ordenada (y) del máximo valor de MA para redes con las características y tamaños iniciales indicadas en la primera fila y columna respectivamente.

casos de redes puramente dictatoriales. Además, las redes que inicialmente parten con una configuración aleatoria generalmente obtienen valores intermedios (ver Tabla 1).

También pueden obtenerse algunas conclusiones a partir de la observación de la Tabla 2. En la misma se puede observar cómo el valor de MA se mantiene relativamente estable para los diferentes tamaños iniciales en las redes dictatoriales; por el contrario, en los otros dos casos (redes inicialmente aleatorias y democráticas) se observa que a partir de redes de mayor tamaño, el valor de MA disminuye; la variación es más notoria en el caso de redes democráticas. Con respecto a la cantidad de nodos agregados en el que cada caso llega al máximo valor de MA se observa que, tanto en el caso de redes inicialmente dictatoriales como democáticas, éste aumenta mientras mayor sea el tamaño inicial de la red; no ocurre lo mismo en el caso de la red aleatoria.

Finalmente, como conclusión más general, las redes conservan gran cantidad de información inicial luego de realizar un proceso de evolución con el modelo BA. Aún después de largos períodos de tiempo, ciertas características prevalecen en las redes independientemente del tamaño que hayan tenido inicialmente, estas dependen sólo de sus características topológicas.

REFERENCIAS

[1] A.-L. BARABÁSI AND R. ALBERT. *Emergence of scaling in random networks*,Science, 286 (1997), pp. 509512.

[2] A.-L. BARABÁSI, H. JEONG, Z. NÉDA, E. RAVASZ, A. SCHUBERT AND T. VIECSEK. *Evolution of the social network of scientific collaborations*, Physica A, 311 (2002), pp. 590–614.

[3] L. DA F. COSTA, F. A. RODRIGUEZ, G. TRAVIESO, P. R. VILLAS BOAS, *Characterization of Complex Networks: A Survey of measurements*, Advances in Physics, Volume 56 (2008), pp. 167-242.

[4] M. GIRVAN AND M. E. J. NEWMAN, *Community structure in social and biological networks*, Proceedings of the National Academy of Science USA, 99(12):78217826, 2002.

[5] M. MITCHELL, *Artificial Intelligence* 170 (2006) 11941212.

[6] T. WILHELM AND J. HOLLUNDER, *Information theoretic description of networks*, Physica A 385 (2007), pp.385–396.

[7] D. J. WATTS AND S. H. STROGATZ, *Collective dynamics of small-world networks*, Nature, 93-6684 (1998), pp. 440442.

SISTEMA EXPERTO DIFUSO PARA EL PRONÓSTICO Y DIAGNÓSTICO DE DESÓRDENES TEMPOROMANDIBULARES UTILIZANDO ANÁLISIS FACTORIAL Y ELEMENTOS FINITOS

Alberto Hananel Baigorria†

†Facultad de Ingeniería, Universidad Católica Santo Toribio de Mogrovejo, Av. Panamericana Norte # 855 Chiclayo, Perú, ahananel@usat.edu.pe, www.usat.edu.pe

Resumen: El presente trabajo muestra cómo a través de la simplificación del examen odontológico para la detección de desórdenes temporomandibulares mediante la construcción de un sistema implementado en Visual Basic, es posible utilizar los datos obtenidos de un conjunto de pacientes, para su posterior Análisis Factorial en SPSS, desembocando en la obtención de once nuevos grupos del avance de la enfermedad, los mismos que se utilizaron como consecuentes de las reglas lógicas difusas creadas, permitiéndole este sistema al odontólogo usuario el ingreso de síntomas y signos de un paciente arbitrario, para la obtención del índice de Fricton, la sugerencia del estado aproximado del avance de la enfermedad mediante la toolbox fuzzy de MATLAB y finalmente de la creación dinámica de un prototipo semidiseñado en 3D de la mandíbula en SOLIDWORKS, para su análisis biomecánico mediante elementos finitos, y la visualización de su respectiva simulación del avance de la enfermedad y sus repercusiones

Palabras claves: *desórdenes temporomandibulares, análisis factorial, elementos finitos, lógica difusa*

1. INTRODUCCIÓN

El diagnóstico de los desórdenes temporomandibulares (DTM) ha sido un tema controversial desde que se planteó esta patología [1]. El funcionamiento de la articulación temporomandibular (ATM) ha sido sujeto de muchos estudios desde hace más de un siglo, y todavía, es aún un problema en discusión [2]. A pesar de las diferencias metodologías de diversos estudios epidemiológicos parece haber un consenso en que los signos y síntomas de los DTM son comunes en la población general. Es por estas razones que son necesarios los Índices para que estos permitan a los investigadores categorizar la severidad de un problema en un individuo, examinar la incidencia del problema en una población específica, medir la efectividad de una terapia dada, y finalmente, sus estrategias de prevención [3].

Ninguno de los sistemas ha podido llegar al grado de reflejar a los DTM como una entidad clínica[4], por lo que este estudio pretendió y pretende entre uno de sus cometidos, mediante el empleo de la técnica estadística multivariante del Análisis Factorial, relacionar un gran número de variables, obtenidas mediante el examen clínico con un número más pequeño de variables o factores que permitan simplificar y sistematizar el examen clínico de los pacientes para el diagnóstico de los DTM, utilizando para la recolección de la información en un paciente, técnicas de programación embebidas de agradable interfaz como las sugeridas por Visual Basic.

Además, otro de sus objetivos, persiguió el uso correcto de la experiencia del odontólogo para la creación de un motor de inferencia basado en Lógica Difusa, dado que, los criterios para diagnosticar y pronosticar la enfermedad y su avance, se basan, si bien es cierto, en estudios realizados, pero también en dicha experiencia y en su intuición. Es por ello que se pretendió recoger también este conocimiento para una correcta escala de cada uno de los factores que sean recogidos vía análisis factorial, de modo de no sólo sea una descripción puramente estadística, sino con el ingrediente de la inteligencia artificial.

Finalmente el estudio, ha pretendido no sólo quedarse con estos resultados, sino aplicar estos conocimientos dentro de una plataforma de CAD/CAE (diseño e ingeniería asistidas por computador) para obtener representaciones gráficas tridimensionales que simulen las zonas comprometidas por las fuerzas o factores que ocasionan el desorden en la mandíbula, para su posterior pronóstico de la sintomatología y tratamiento de dicho desorden.

2. PLAN DE ANÁLISIS

Se ingresaron los datos recolectados de los test registrados y archivados de los ochentaicinco pacientes que cumplieron con el criterio de inclusión en el sistema creado en Visual Basic, guardando los resultados de este ingreso en una nueva base de datos de Access exportable a Excel, así como la obtención del índice de Fricton y otras variables de importancia.

Se desarrolló el Análisis Factorial con los campos apropiados de los datos almacenados y se validó dicho análisis estadístico mediante la prueba de esfericidad de Barlett y la medida de adecuación de la muestra de Kaiser-Meyer-Olkin (KMO).

Con un conjunto de expertos en desórdenes temporomandibulares se crearon las reglas lógicas del sistema difuso teniendo como parte de sus consecuentes, los factores obtenidos en el Análisis Factorial, y su grado de membresía con la combinación del resultado proporcionado por la toolbox fuzzy de MATLAB y el resultado final del índice de Fricton almacenado ya en la base de datos.

Se construyó una representación CAD 3D del sistema estomatognático involucrado de un adulto joven que cumple con los criterios de inclusión, con medidas acorde con la bibliografía consultada en el programa SOLIDWORKS 2010.

Se evaluaron los datos de un sujeto de la muestra procediendo al diagnóstico de la enfermedad utilizando el sistema experto difuso creado.

3. EJECUCIÓN

Como se mencionó anteriormente, el sistema genera automáticamente dos archivos de Access, uno con toda la documentación de todos los pacientes que han pasado por el criterio de inclusión y otro con todos los datos de todos los pacientes. Para efectos del trabajo de investigación se maneja un archivo de Access especial distinto incluso a estos dos referidos, con el criterio de inclusión cumplido. Esta base de datos generada no contiene todas las variables analizadas. Si bien se tiene registrado todos los ítems en la otra base, de acuerdo a la opinión del experto sólo se han considerado treinta variables, las más significativas, las cuales han pasado por pruebas especiales como, pues para continuar con el estudio, se ha procedido a obtener el índice KMO el cual resultó adecuado; es obvio que no todas las variables ingresadas son las apropiadas, pues una de ellas puede no necesariamente guardar relación con las restantes, es por eso que se tienen que depurar algunas con la opinión del experto y con ensayos previos utilizando el criterio del análisis factorial.

Después del desarrollo del análisis factorial, aparecen los indicadores para construir otras nuevas variables ocultas las cuales son procesadas en Visual Basic bajo ciertos criterios, y agregadas a la base de datos. En el caso de la tesis se ligaron todas las variables ingresadas con los once factores obtenidos del análisis factorial para la obtención de las variables ocultas. Finalmente Visual Basic después de todo el proceso visible y oculto, devuelve en un nuevo formulario, los resultados de la evaluación de veintiún variables en total, los cuales son posteriormente leídos por la toolbox de lógica difusa de MATLAB, dado que también todos estos mismos resultados son automáticamente generados en un archivo de Excel.
Una vez que MATLAB termina el procesamiento de la información devuelve un archivo en Excel con los resultados del diagnóstico del avance de la enfermedad, cuyos resultados a su vez son transferidos a Visual Basic para su mejor visualización y su respectivo guardado en la misma base.

De acuerdo al análisis factorial se obtuvieron once factores, por tanto este último archivo de Excel referido contiene once evaluaciones. Dado que el estudio se fijó con un tamaño muestral fijo, que en su caso corresponde a ochentaicinco, el estudio es general para individuos que cumplen con los criterios evaluados por lo que eso significa que los resultados del análisis de esa base de datos construida se pueden utilizar para la evaluación de un paciente nuevo cualquiera, sea o no de la muestra, pues finalmente uno de los objetivos de esta investigación fue el uso del sistema y de sus resultados para cualquier individuo. Por

lo tanto, se sigue que para un paciente arbitrario y analizado, el sistema generado en Visual Basic finalmente mostrará un formulario con los tres resultados más importantes además de su recomendación:

-El índice de Fricton obtenido de la sistematización del mismo, que es lo único que se podía obtener antiguamente acerca del índice de Fricton por el camino clásico.
-La pertenencia de los resultados del paciente a cada uno de los factores que miden el avance de la enfermedad, herramienta nueva y documentada.
-El video que Solidworks determine del diagnóstico y pronóstico del avance de la enfermedad en formato asequible al paciente.
-La "receta" final del médico que consta de sus observaciones e información estructurada o no estructurada, y con el posible tratamiento a seguir.

Dado que los resultados son presentados al paciente se infiere que estos resultados provenientes del sistema experto difuso, sirven por tanto, al odontólogo para la toma de un mejor diagnóstico y pronóstico del paciente, y al paciente para el posible entendimiento de su situación con respecto a la enfermedad temporomandibular que en la actualidad posee y cómo le podría afectar a futuro la misma, para determinar si opta o no por un tratamiento, pues finalmente es dicho paciente quien toma esa decisión en función a diversos factores.

4. RESULTADOS

Se generó una interfaz en Visual Basic para la recolección de la información. (Figura 1)
Se simplificó, desarrolló, analizó y postularon los criterios relacionados al diagnóstico de desórdenes temporomandibulares en sujetos adultos jóvenes.
Se simplificaron los criterios diagnósticos articulares, musculares, oclusales de mayor relación a la presencia de desórdenes temporomandibulares.
Se analizó y definieron las dimensiones que agrupan a los factores y criterios articulares, musculares y oclusales para el diagnóstico de desórdenes temporomandibulares mediante el uso de la técnica estadístico-multivariante del Análisis Factorial. (Figura 1)
Se creó una base de conocimiento tratada con lógica difusa, dada la opinión de los expertos en desórdenes temporomandibulares y su respectiva implementación en el sistema a construir. (Figura 2)
Se obtuvieron simulaciones tridimensionales con análisis biomecánico mediante elementos finitos de las zonas comprometidas por los factores que ocasionan el desorden en la mandíbula. (Figura 3)
El sistema experto difuso construido pudo simplificar el examen clínico para procurar dar un mejor pronóstico y diagnóstico de los desórdenes temporomandibulares de un paciente dado. (Figura 4)

AGRADECIMIENTOS

A los doctores Daniel Paredes Ruiz y Arturo Kobayashi Shinya por los aportes odontológicos y al matemático Flabio Gutierrez Segura por su importante asesoría en la investigación.

5. REFERENCIAS

[1] FRICTON JR, SCHIFFMAN EL. *The craneomandibular index: Validity.* J Prosthet Dent 1987; 58(2):222-228.
[2] FRICTON JR, SCHIFFMAN EL. *Rehability of a craniomandibular index.* J Dent Res 1986; 65(11):1359-1364.
[3] SCHMIDT-KAUNISAHO, HILTUNEN K, AINAMO A. *Prevalence of symptoms of craniomandibular disorders in a population of elderly inhabitants in Helsinki, Finland*. Acta Odontol Scand 1994; 52:135-139.
[4] WADHWA L, UTREJA A, TEWARL A. *Study of clinical sings and symptoms of temporomandibular of function in subjects with normal occlusion, untread, and treated malocclusions.* Am J Orthod Dentofac Orthop 1993; 103 (1): 54-61.

6. FIGURAS

	A	B	C	D	E	F	G	H	I	J	K
1	FACTOR 1		0.5	Desorden inflamatorio articular con desorden funcional mandibular							
2	FACTOR 2		0.5	Desorden crónico degenerativo articular con desorden muscular mandibular por probable parafunción							
3	FACTOR 3		0.5	Desorden degenerativo crónico con desorden muscular y funcional mandibular, con alteración de la estabilidad oclusal							
4	FACTOR 4		0.5	Desorden crónico de los músculos de la nuca							
5	FACTOR 5		0.5	Desorden crónico oclusal en movimientos excéntricos (Lateralidad Lado de no Trabajo y Protusiva)							
6	FACTOR 6		0.5	Desorden crónico articular de desarreglo cóndilo-disco (cierre)							
7	FACTOR 7		0.5	Desorden crónico con sintomatología muscular cervical							
8	FACTOR 8	0.83666667		Alteración crónica oclusal en movimientos excéntricos (Lateralidad Lado de Trabajo)							
9	FACTOR 9	0.74307439		Desorden crónico con desorden muscular masticatorio							
10	FACTOR 10		0.5	Desorden crónico articular de desarreglo cóndilo-disco (apertura, cierre, lateralidad)							
11	FACTOR 11		0.5	Desorden crónico degenerativo con alteración de la estabilidad oclusal							

Figura 1: Sistema de recolección de información y procesamiento de resultados

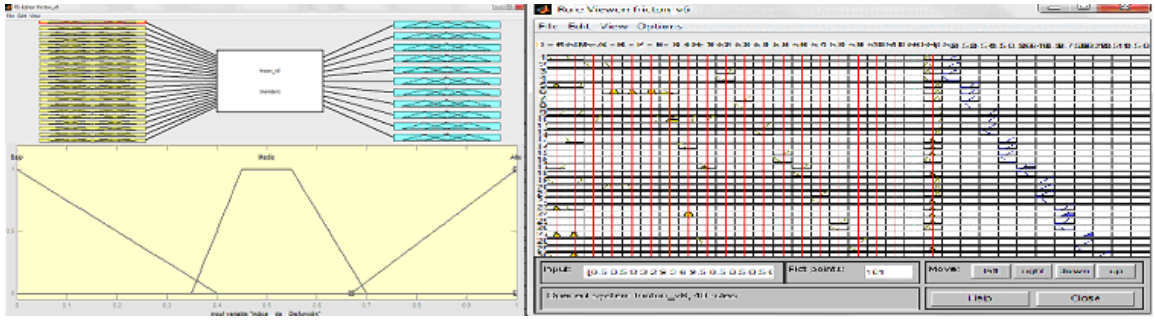


Figura 2: Utilización de la toolbox fuzzy de MATLAB para la construcción del sistema experto difuso

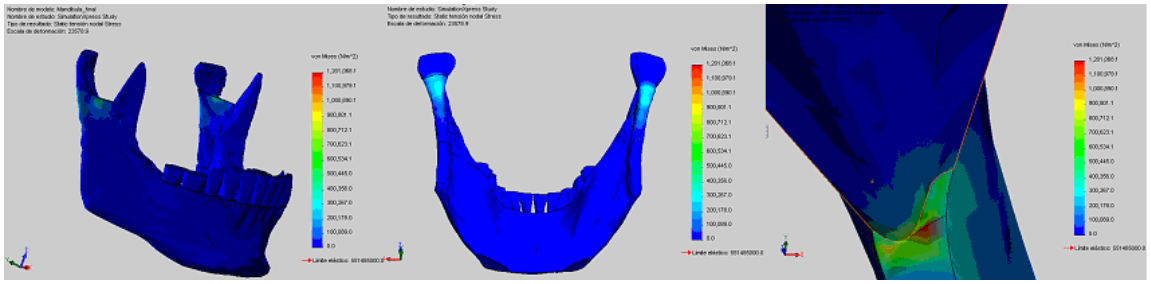


Figura 3: Simulaciones obtenidas mediante elementos finitos para un paciente arbitrario

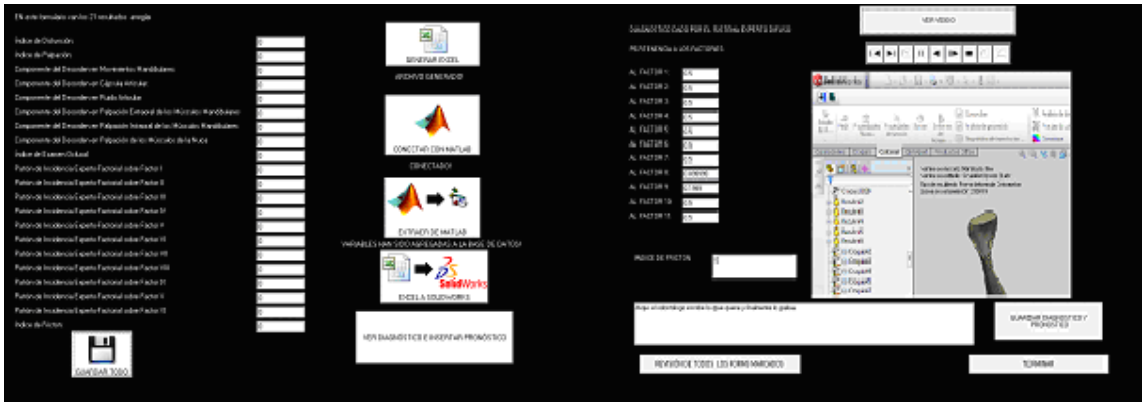


Figura 4: Resultados obtenidos por el sistema experto difuso para un paciente arbitrario

ELECTROSEISMIC MONITORING OF CO_2 SEQUESTRATION: A FINITE ELEMENT APPROACH

Fabio I. Zyserman[*,♭], Juan E. Santos[*,♭,†] and Patricia M. Gauzellino[♭]

[*] *CONICET*
[♭] *Departamento de Geofísica Aplicada, Fac. de Cs. Astronómicas y Geofísicas, Universidad Nacional de La Plata, Paseo del Bosque s/n, B1900FWA - La Plata, Argentina*
[†] *Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, Indiana, 47907-2067, USA; santos@math.purdue.edu*

Abstract: The capture and geological storage of CO_2 is currently being used as a means of reducing CO_2 emissions into the atmosphere. However, substantial research efforts are now underway on predicting the injected CO_2 behaviour, possible migration and how its presence can affect the host reservoir. In this work a frequency-domain finite element procedure for two dimensional electroseismic modeling in poroelastic fluid-saturated media is presented, and its capability as a potential monitoring tool is analysed. The model involves the simultaneous solution of Biot's equations of motion and Maxwell's equations, coupled via an electrokinetc coefficient and employing absorbing boundary conditions at the artificial boundaries. Maxwell equations are discretized in space using Nedelec's mixed finite element spaces, whilst for Biot' s equations a nonconforming element for each component of the solid displacement and the vector part of the Raviart-Thomas-Nedelec of zero order for the fluid displacement vector are used. The case of compressional and vertically polarized shear waves coupled with the transverse magnetic polarization (PSVTM-mode) is used to test this potential prospecting technique as a tool for CO_2 injection monitoring.

Keywords: *Electroseismic Modeling, Poroelasticity, CO_2 sequestration, Finite element methods.*

2000 AMS Subject Classification: 78M10,35Q61,35Q35

1 PROBLEM STATEMENT

The injection of large amounts of man-produced CO_2 in depleted oil wells below the sea floor and in other apropriate geological formations has been used, for several years now, as a means of reducing the carbon dioxide emissisons into the atmosphere. For example, CO_2 is being injected in the Sleipner field in the North Sea since 1996 at a rate of 0.85 Mt per year [1], and also beneath the Sahara desert, at In Salah in Algeria [2].

Scientists from different areas have been studying this topic, and a still open problem is to determine the behaviour of the gas once set into the reservoir. Will it remain stable? Will it migrate, and make its way back to the surface?. How the stored CO_2 can be efficiently monitored is still a topic of intense reseach.

In this work, some preliminary studies are shown that suggest the usefulness of electroseismics as a possible CO_2 storage monitoring tool.

In order to begin with the modeling of the stated problem, consider the following set of equations, to be solved in a 2D-rectangular domain $\Omega = \Omega^a \cup \Omega^B$ where Ω^a and Ω^B are associated with the air and subsurface (disjoint) parts of Ω, respectively, and Ω^B encloses an isotropic porous solid saturated by a compressible viscous fluid. Denote with u^s and $\widetilde{u}^f$ the averaged displacement vectors of the solid and fluid phases, respectively; and let $u^f = \phi(\widetilde{u}^f - u^s)$ be the average relative fluid displacement per unit volume of bulk material, where ϕ denotes the effective porosity. Assuming an $e^{+i\omega t}$ temporal dependence, and setting $u = (u^s, u^f)$, where $u^s = (u^s_x(x,z,\omega), u^s_z(x,z,\omega))$ and $u^f = (u^f_x(x,z,\omega), u^f_z(x,z,\omega))$, the coupled electromagnetic-poroviscoelastic equations [3] -simplified for the PSVTM-mode as proposed in [4]- for the electric and magnetic fields $E = (E_x(x,z,\omega), E_z(x,z,\omega))$ and $H = H_y(x,z,\omega)$ and the displacement

vectors u^s and u^f, can be stated in the space-frequency domain as follows:

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} + i\omega\mu H_y = -J_m^s \text{ in } \Omega, \tag{1}$$

$$\sigma(E_x, E_z) - (-\frac{\partial H_y}{\partial z}, \frac{\partial H_y}{\partial x}) = 0 \text{ in } \Omega, \tag{2}$$

$$-\omega^2\rho_b u^s - \omega^2\rho_f u^f - \nabla\cdot\tau = 0 \text{ in } \Omega^B, \tag{3}$$

$$-\omega^2\rho_f u^s - \omega^2 g_0 u^f + i\omega\frac{\eta}{\kappa_0}u^f + \nabla p_f = \frac{\eta}{k_0}L_0 E \text{ in } \Omega^B, \tag{4}$$

In these equations, σ and μ have the usual meaning within the context of electromagnetism; while J_m^s is an external electromagnetic source. Also, $\rho_b = \phi\rho_f + (1-\phi)\rho_s$, where ρ_s and ρ_f denote the mass densities of the solid grains composing the solid matrix and the saturant fluid respectively. In (3) $\tau = \tau(u)$ denotes the stress ; κ_0 denotes the permeability and η stands for the fluid viscosity . Finally, the coupling between electromagnetic and mechanical processes is represented in the above equations by the coefficient L_0.
Attached to these equations the following boundary conditions are used:

$$\beta(1-i)E\cdot\chi + H_y = 0, \text{ on } \Gamma, \tag{5}$$

$$\tau\cdot\nu = 0\ ,\ p_f = 0 \text{ on } \Gamma^{t,B}, \qquad -\mathcal{G}(u) = i\omega\mathcal{D}\mathcal{S}(u) \qquad \text{on } \Gamma^{l,b,r,B}., \tag{6}$$

where $\mathcal{G}(u) = (-\tau(u)\nu\nu, \tau(u)\nu\chi, p_f)^t$, $\mathcal{S}(u) = (u^s\cdot\nu, u^s\cdot\chi, u^f\cdot\nu)$, ν is the unit outwards normal to the boundary and χ is a unit vector tangent to the boundary. Here Γ is the boundary of the whole domain, whilst $\Gamma^{t,B}$ stands for top portion of the boundary of the subsurface; the meaning of $\Gamma^{s,B}, s = l, b, r$ can be easily inferred.
In [5] it is shown that $\mathcal{D}$ in equation (6) is a 3×3 positive definite matrix depending on the subsurface properties, $\mathcal{D} = \mathcal{R}^{\frac{1}{2}}\mathcal{S}^{\frac{1}{2}}\mathcal{R}^{\frac{1}{2}}$, where $\mathcal{S} = \mathcal{R}^{-\frac{1}{2}}\mathcal{M}^{-\frac{1}{2}}\mathcal{R}^{-\frac{1}{2}}$; matrices $\mathcal{R}$ and$\mathcal{M}$ depend on the Biot equations parameters. Finally, the electromagnetic source is $J_m^{ext} = -i\omega\mu SI(\omega)\delta(x-x_f)\delta(z-z_f)\breve{y}$ is a magnetic dipole of infinite length centered in (x_f, z_f) [6]; here S is the area of the current loop and the electric current is $I(\omega)$.

2 FINITE ELEMENTS METHOD

As Maxwell's equations are decoupled from Biot's equations [4], for each frequency ω the numerical procedure is naturally separated in a first part where the secondary electromagnetic fields are obtained, and a second part where Biot's equations are solved. In order to approximate Maxwell's equations an iterative hybridized mixed domain decomposed finite element procedure is implemented [7, 8]. The main concept underlying this method is to split the problem in a collection of small ones whose individual solution can be easily computed. The structure of the algorithm makes it specially appropriated for being implemented in parallel architectures. Next we describe the finite element spaces employed. Denote by $(\cdot,\cdot)$ the inner product in an element, and by $\langle\cdot,\cdot\rangle$ the inner product on the boundary of an element. Further, denote by $\Omega_{\mathtt{j}}$ the elements of the finite element partition of the domain Ω, assumed to coincide with the finite element partition. Let $\Gamma_{\mathtt{jk}}$ denote the common boundary between the adjacent elements $\Omega_{\mathtt{j}}$ and $\Omega_{\mathtt{k}}$, and $B_{\mathtt{j}}^a$ be the intersection of $\Gamma_{\mathtt{j}}$ with the computational domain. As approximating mixed finite element spaces for the electric and magnetic fields consider

$$V^h = \left\{E^h \in L^2(\Omega) : E^h|_{\Omega_{\mathtt{j}}} \in P_{0,1}\times P_{1,0}\right\}, \qquad W^h = \quad \left\{H^h \in L^2(\Omega) : H^h|_{\Omega_{\mathtt{j}}} \in P_{0,0}\right\}.$$

Here $P_{1,0}$ denotes a polynomial of degree less or equal 1 in x and less or equal 0 in z. These spaces have five degrees of freedom associated with each element, four to the electric field and one to the magnetic field, respectively.
To approximate each component of the solid displacements a nonconforming finite element space is used; while the fluid displacements are approximated by the vector part of the Raviart-Thomas-Nedelec space of

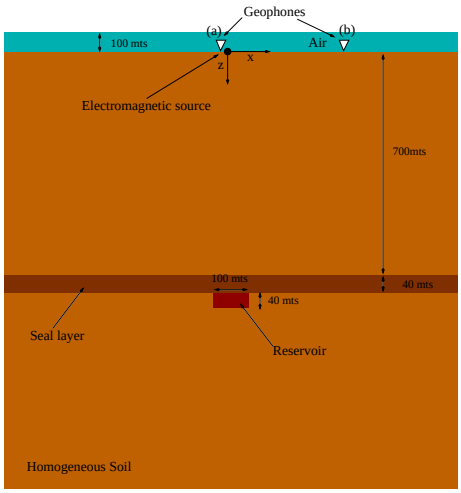


Figure 1: The model

zero order. More specifically set $\theta(x) = x^2 - \frac{5}{3}x^4,\ R = [-1,1]^2$ and

$$\varrho^L(x,z) = \frac{1}{4} - \frac{1}{2}x - \frac{3}{8}(\theta(x)-\theta(z)), \varrho^R(x,z) = \frac{1}{4} + \frac{1}{2}x - \frac{3}{8}(\theta(x)-\theta(z)), \tag{7}$$

$$\varrho^B(x,z) = \frac{1}{4} - \frac{1}{2}z + \frac{3}{8}(\theta(x)-\theta(z)), \varrho^T(x,z) = \frac{1}{4} + \frac{1}{2}z + \frac{3}{8}(\theta(x)-\theta(z)); \tag{8}$$

defining $\mathcal{Y}(R) = \text{Span}\{\varrho^L, \varrho^R, \varrho^B, \varrho^T\}$. Also, if $\varphi^L(x) = -1+x$, $\varphi^R(x) = x$, $\varphi^B(z) = -1+z$, $\varphi^T(z) = z$, set $\mathcal{Z}(R) = \text{Span}\left\{(\varphi^L(x),0),(\varphi^R(x),0),(0,\varphi^B(z)),(0,\varphi^T(z))\right\}$, and the finite element spaces $\mathcal{Y}^h_{\mathrm{j}}$ and $\mathcal{Z}^h_{\mathrm{j}}$ are defined as usual by scaling and translating to the element Ω_{j}. Notice that in each domain of the finite element partition there exist twelve unknowns, four for each solid displacement component, and two for each component of the fluid displacement. For details on the finite element algorithm and the corresponding error estimates we refer to [7, 8, 4].

3 Numerical example

In this example a reservoir with with different CO_2 concentrations is simulated. The model, as depicted in Fig. 1, comprises an homogeneous Earth; at 700 m depth a 40 m thick low permeability layer is located, and just below it, a 20 m $\times$ 100 m CO_2 storage region. The source, as explained previously, is an (in the y-axis) infinite solenoid located on the surface, generating a Ricker-wavelet with central frequency of 20 Hz, with a duration of 0.15 s. This model is useful to observe that a variation in the CO_2 concentration in the storage region, and consequently a reduction of in the electrical conductivity of this zone [1], can be observed in the seismic traces recorded on the surface. In Fig. 2 traces of the x-component and z-component of the acceleration for different CO_2 saturations are recorded at the geophone nearest to the electromagnetic source. It can be seen that, although the reservoir is not resolved by the traces -because of the relative low peak frequency of the source-, the different concentrations of carbon dioxide do influence the shape of the traces.

Notice that the point at which traces of Fig. 2 are recorded, the z-component of the electromagnetic field is near to its minimum, and the x-component is near to its maximum. In Fig. 3 it is shown how the traces change when recorded at a point to the right of the source, beyond the reservoir rightmost boundary, compared to one recorded near the source, at point (a) in Fig. 1. The numerical results obtained in this example show that electroseismics can be considered as a monitoring tool, and further research is needed, considering different electromagnetic sources with different frequency spectra, sensitivity analysis to different parameters, etc.

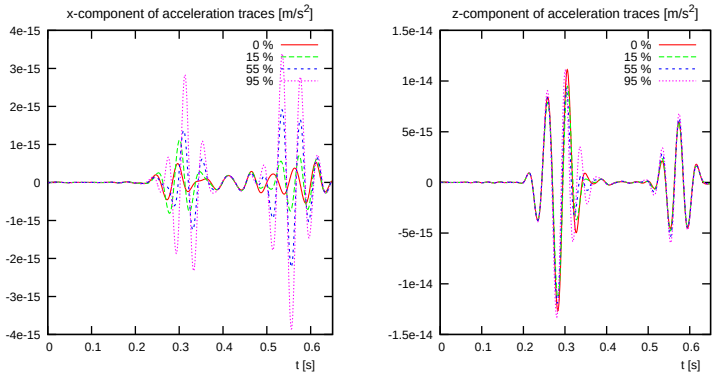


Figure 2: Traces of both components of the acceleration for different CO_2 concentrations

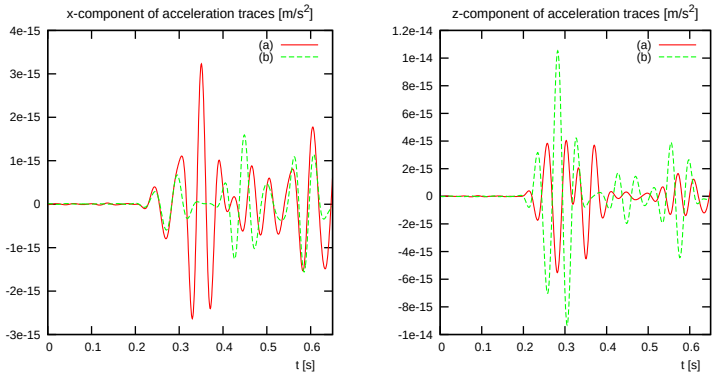


Figure 3: Traces of both components of the acceleration recorded at locations (a) and (b) of Fig.1, for a CO_2 concentration of 95%

ACKNOWLEDGMENTS

The results presented here have been obtained in the CO2ReMoVe project, which envisages the development of technologies and procedures for monitoring and verifying geological CO_2 storage. The financial support of the European Commission and the industrial consortium involved is greatly appreciated.

Partial financing was also received from CONICET, PIP 112-200801-00952 and by Agencia Nacional de Promoción Científica y Tecnológica PICT 03-13376.

REFERENCES

[1] M.ELLIS, *The Potential of Controlled Source Electromagnetic Surveying in CO2 Storage Monitoring*, SEG Expanded Abstracts 29, 843-847, 2010.

[2] P. RINGROSE, M. ATBI, D. MASON, M. ESPINASSOUS, O. MYHRER, M. IDING, A. MATHIESON AND I. WRIGHT, *Plume development around well KB-502 at the In Salah CO_2 storage site*, First Break, 27, 85–89, 2009.

[3] S. R. PRIDE, *Governing equations for the coupled electromagnetics and acoustics of porous media*, Physics Review B, 50, (1994), pp 15678.

[4] F. ZYSERMAN, P. GAUZELLINO AND J. E. SANTOS, *Finite element modeling of SHTE and PSVTM electroseismics*, Journal of Applied Geophysics, 72, 79-91, 2010.

[5] J. E. SANTOS, J. DOUGLAS, JR., M. E. MORLEY, AND O. M. LOVERA, *Finite element methods for a model for full waveform acoustic logging*, IMA Journal of Numerical Analysis, 8, (1988), pp 415.

[6] WARD, S., HOHMANN, G. W., *Electromagnetic theory for geophysical applications. In: Nabiguian, M. N. (Ed.), Electromagnetic Methods in Applied Geophysics.* Vol. 3 of Investigations in Geophysics. SEG, pp. 131–311, 1987.

[7] SANTOS, J. E., *Global and domain-decomposed mixed methods for the solution of Maxwell's equation with application to magnetotellurics.* Numerical Methods for Partial Differential Equations 14, 263–280, 1998

[8] ZYSERMAN, F. I., SANTOS, J. E., *Parallel finite element algorithm for three–dimensional magnetotelluric modelling. Journal of Applied Geophysics* 44, 337–351, 2000.

SIMULACIÓN DE FOTOCONDUCTIVIDAD PERSISTENTE EN ÓXIDOS SEMICONDUCTORES Y DETERMINACIÓN DE DISTRIBUCIÓN DE TRAMPAS

Silvina C. Real[†], Mónica C. Tirado[‡] y David Comedi[†‡]

[†]*Área Matemática Aplicada, Dpto. de Matemática, FACET, Universidad Nacional de Tucumán, Tucumán, Argentina, sreal@herrera.unt.edu.ar*
[‡]*Laboratorio de Propiedades Dieléctricas de la Materia, Dpto. de Física, FACET, Universidad Nacional de Tucumán, Tucumán, Argentina, mtirado@herrera.unt.edu.ar*
[†‡]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y Laboratorio de Física del Sólido, Dpto. de Física, FACET, Universidad Nacional de Tucumán, Tucumán, Argentina, dcomedi@herrera.unt.edu.ar*
PROYECTO NANO-UNT: http://www.herrera.un.edu.ar/nano

Resumen: El fenómeno de fotoconductividad persistente es común en muchos semiconductores. Las trampas de portadores (electrones y huecos) en semiconductores son importantes pues las propiedades optoelectrónicas dependen de ellas. En primer lugar en este trabajo se trata el problema directo de simulación del efecto de fotoconductividad persistente debido a distribución de trampas de huecos del tipo gaussiana. Posteriormente se aplica en forma numérica un método para la resolución del problema inverso que consiste en la determinación de la densidad de trampas en óxidos semiconductores a partir de mediciones del decaimiento temporal de la fotoconductividad persistente. Esto se lleva a cabo usando un modelo propuesto en la literatura de Reemisión- Recombinación, con la ayuda de transformada de Laplace.

Palabras claves: *Fotoconductividad, Nanoestructuras, ZnO, Densidad de Estados*
2010 AMS Subjects Classification: 82D80- 82D37

1. INTRODUCCIÓN

Los semiconductores presentan generalmente fotoconductividad, o sea el cambio de su conductividad eléctrica cuando se lo ilumina con fotones de energía próxima o mayor que la energía de banda prohibida del semiconductor. El motivo de esto es la excitación de fotoelectrones a la banda de conducción (BC) y fotohuecos a la banda de valencia (BV). Muchos semiconductores presentan, además fotoconductividad persistente (FCP) [1,2], o sea, la persistencia de la conductividad eléctrica alterada en relación a la conductividad de equilibrio incluso después de haber interrumpido la iluminación. Los motivos de este efecto son variados y en muchos casos ha originado controversia en la literatura. En los óxidos de banda ancha, como por ejemplo ZnO y TiO_2, se ha sugerido que el fenómeno de FCP se debe a la presencia de estados electrónicos profundos en la banda prohibida que atrapan los fotohuecos, inhibiendo y retrasando el proceso de recombinación electrón-hueco [2,3].

El objetivo de este trabajo es, por un lado, explorar la posibilidad de explicar dentro de este modelo, a través de una simulación numérica simple, el fenómeno de FCP en óxidos semiconductores. Por otro lado, se realiza el intento de determinar la distribución en energía de trampas de huecos en nanoestructuras de ZnO a partir de las curvas experimentales de FCP (decaimiento temporal de la corriente eléctrica luego de interrumpir la iluminación), aplicando el método de la transformada de Laplace [4,5] al modelo de reemisión-recombinación.

2. FUNDAMENTOS DEL MODELO FISICO

La suposición básica es que los fotohuecos caen a trampas profundas, caracterizadas por una función densidad de estados *g(E)*, mientras que los fotoelectrones en la BC dominan la conductancia [2,3] (Fig. 1). Después de interrumpir la iluminación, los electrones sólo pueden recombinar con aquellos huecos liberados de las trampas. La reemisión de huecos a la BV ocurre con un tiempo característico τ que depende exponencialmente de la energía *E* de la trampa [1],

$$\tau = \left(\frac{1}{\upsilon}\right) e^{\frac{E}{KT}}$$

Donde υ es la frecuencia de intento de escape, K es la constante de Boltzmann y T la temperatura absoluta. La conductividad es proporcional a la corriente eléctrica medida cuando se aplica un voltaje constante, la cual a su vez es proporcional a la densidad de electrones en la BC, cuya dependencia temporal puede escribirse como:

$$n(t) = n_0 + \int g(E) e^{-\frac{t}{\tau(E)}} dE$$

donde n_o es la densidad de electrones en la BC en equilibrio (antes de la iluminación).

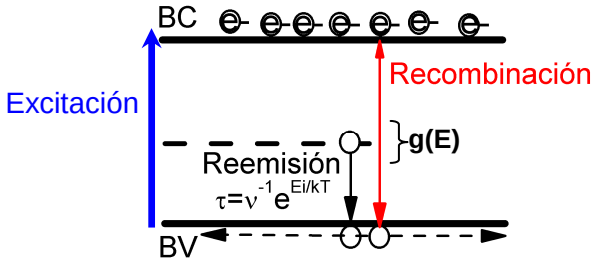


Figura 1: Diagrama esquemático de bandas de energía en un semiconductor y de los procesos de excitación de electrones y huecos y reemisión de huecos desde trampas.

3. PROBLEMA DIRECTO: SIMULACIÓN DE LA DENSIDAD ELECTRÓNICA A PARTIR DE UNA DISTRIBUCIÓN DE TRAMPAS GAUSSIANA

Se analiza a continuación qué tipo de variación temporal de corriente predice el modelo, aproximando *g(E)* a una campana gaussiana (Fig. 2 y 4) como lo sugerido por otros autores (por ejemplo [6]). Se varían la posición y ancho de la gaussiana, y se estudian sus efectos sobre la corriente simulada, suponiendo que ésta es proporcional a *n(t)* (Fig. 3 y 5).

(a) Efecto de la posición de la centroide de la distribución (ancho y altura fijos):

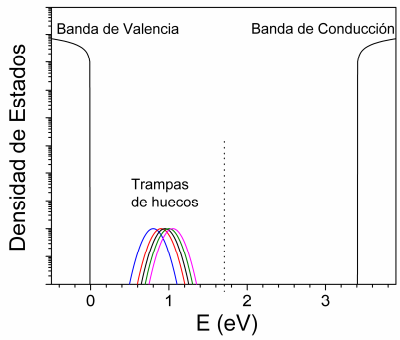


Figura 2: Densidad de estados del modelo

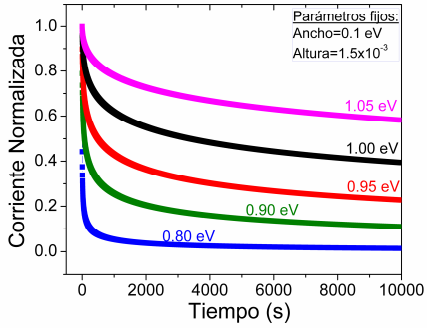


Figura 3: Corriente normalizada

Las curvas de corriente normalizada son muy sensibles a la variación de la posición del pico de la densidad de estados. Esto es debido a la dependencia exponencial del tiempo característico de reemisión con la energía de la trampa.

(b) Efecto del ancho de la distribución (posición fija en 0.9 eV y área constante)

Para ancho muy fino de 0.01 eV << 0.9 eV, la distribución gaussiana se aproxima a una Delta de Dirac. La corriente obtenida es por lo tanto muy próxima a una función exponencial [1] con un $\tau=10^{-13}exp(0.9/KT)=123$ *s*. Cuánto más ancha es la distribución, mayor es la contribución de los estados más profundos en la banda prohibida. Por eso aumentan las contribuciones relativas a la corriente normalizada para tiempos largos, con un τ muy grande (pendiente chica). Por otro lado, también aumenta la contribución de estados rasos (E<0.9 eV), lo que hace que la corriente caiga rápidamente (τ chico) para tiempos cortos.

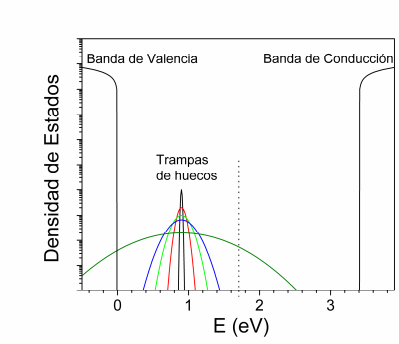


Figura 4: Densidad de estados del modelo

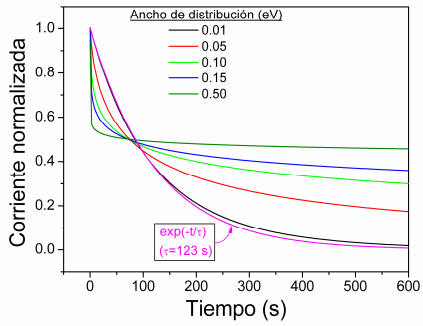


Figura 5: Corriente normalizada

Las corrientes se cruzan para un tiempo del orden de 100 s, del mismo orden de magnitud que el τ promedio (123 s). Ese valor de τ divide el eje de tiempos en dos regiones, la de tiempos cortos (t<<123 s) y la de tiempos largos (t>>123 s).

4. Problema Inverso: Determinación de la Densidad de Trampas a Partir del Decaimiento Temporal de la Corriente Eléctrica Medida

Aplicando Transformada de Laplace [4,5] a *n(t)*, la cual se indicará con $\hat{n}(s)$, se encuentra:

$$\hat{n}(s) = \frac{n_0}{s} + \int g(E) \frac{1}{s + \frac{1}{\tau}}\, dE$$

que puede llevarse a la expresión:

$$\frac{d\left(s\,\hat{n}(s)\right)}{d\left(\ln s\right)} = s \int g(E) h(s,E)\, dE$$

$$h(s,E) = \frac{\upsilon \exp\left(-\frac{E}{kT}\right)}{\left[s + \upsilon \exp\left(-\frac{E}{kT}\right)\right]^2}$$

Diversos autores [3,5] usan la siguiente aproximación:

$$h(s,E) \approx \left(\frac{kT}{s}\right) \delta(E - E_0) \qquad con \quad E_0 = kT \ln\left(\frac{\upsilon}{s}\right)$$

De este modo la función densidad de estados puede escribirse como:

$$g(E_0) = \frac{1}{kT} \frac{d\left(s\,\hat{n}\right)}{d \ln s}$$

expresando la variable "*s*" en términos de la energía de la trampa *(E_0)*.

El cálculo de *g(E)* se llevó a cabo en forma numérica, con un programa escrito en MATLAB.

A continuación se determina numéricamente *g(E)* para dos casos experimentales (Fig. 6) donde una misma muestra de nanoestructuras de ZnO fue medida en condiciones atmosféricas (aire) y luego mantenida en una cámara evacuada (vacío) [7]. En la Fig. 7 se muestran los resultados correspondientes a las densidades de estados calculadas a partir de los datos experimentales.

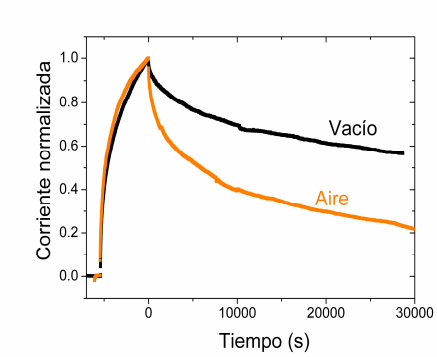


Fig. 6**:** (a) Corriente normalizada en función del tiempo. Después de iluminar durante 5100 s una muestra de ZnO nanoestructurada con fotones de energía 3.1 eV, en t=0 s se apaga la fuente de luz. El decaimiento es más rápido cuando la muestra está en aire [7].

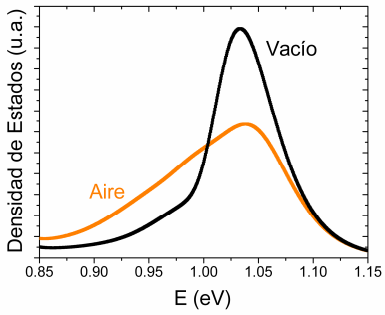


Fig. 7**:** Distribuciones de trampas calculadas a partir de los datos de la Fig. 6. La distribución se ensancha hacia energías menores en el caso de la muestra en aire, probablemente indicando contribuciones en ese rango de energías de trampas debidas a adsorbatos atmosféricos.

5. CONCLUSIONES

Se exploró el modelo de reemisión-recombinación para explicar la FTP en óxidos semiconductores. La posición en energía del centroide de la distribución de trampas determina sensiblemente el tiempo característico de la FCP, mientras que la curvatura de la corriente en función del tiempo depende fuertemente del ancho de la distribución. La FTP experimental en ZnO nanoestructurado puede ser interpretada como debida a una densidad de trampas de huecos centrada en ~1.03 eV. La distribución de trampas deducida para la medición hecha con la muestra en aire está ensanchada hacia las bajas energías en comparación con la distribución determinada para la muestra en vacío.

AGRADECIMIENTOS

Este trabajo fue realizado con apoyo del Proyecto CIUNT 26/E419. Se agradece a Paulo Di Carlo por su ayuda en la digitalización de los datos experimentales y a Nadia Celeste Vega por los datos experimentales de la Fig. 6 y discusiones.

REFERENCIAS

[1] R.H. BUBE, *Photoconductivity of Solids*, John Wiley and Sons, Inc. NY, 1960.

[2] D.COMEDI, S.P. HELUANI, M. VILLAFUERTE, R.D. ARCE, R.R. KOROPECKI, *Power-law photoconductivity time decay in nanocrystalline TiO_2 thin films*, J. Phys.: Condens. Matter 19 (48) (2007), pp. 486205(1)-486205(10).

[3] S.A. STUDENIKIN, N. GOLEGO, M. COCIVERA, *Density of band-gap traps in polycrystalline films from photoconductivity transients using an improved Laplace transform method*, J. Appl. Phys. 84 (1998), pp.5001-5004.

[4] M.N. LEVIN, A.E. AKHKUBEKOV, A.V. TATARINTSEV, *Determination of Radiation-Defect Symmetry by High-Resolution Laplace DLTS*, Bulletin of the Russian Academy of Sciences: Physics, 72 (11) (2008), pp. 1584-1588.

[5] H. NAITO, M. OKUDA, *Simple analysis of transient photoconductivity for determination of localized-state distributions in amorphous semiconductors using Laplace transform.* J. Appl. Phys. 77 (7) (1995), pp.3541-3542.

[6] W-J CHO, *Influences of Trap States at Metal/Semiconductor Interface on Metallic Source/Drain Schottky-Barrier MOSFET*, J. Semicond. Technol. Sci. **7** (2007), pp. 82-87.

[7] N.C. VEGA, M. TIRADO, D. COMEDI, *Conductancia y fotoconductancia de arreglos de nanohilos de ZnO semiorientados verticalmente*, 95ª Reunión Nacional de Física, Asociación Física Argentina, Malargüe, Mendoza, Argentina, del 28 de setiembre al 1 de octubre de 2010 (poster).

APLICACIÓN DEL MÉTODO DE LOS ELEMENTOS FINITO AL FENÓMENO DEL GOLPE DE ARIETE

Alicia E. Carbonell†, Irma M. Benitez‡, Liliana E. Gimenez†‡, Mauricio C. Friedrich‡‡
†U.T.N, Facultad Regional Paraná, alielecarbo@gmail.com, www. frp.utn.edu.ar.
‡U.T.N, Facultad Regional Paraná - U.A.D.E.R., Facultad de Ciencia y Tecnología, manuelabenitez@frp.utn.edu.ar, www. frp.utn.edu.ar.
†‡U.T.N., Facultad Regional Paraná – U.N.E.R. Facultad de Ingeniería, lilianagimenez@yahoo.com.ar, www.frp.utn.edu.ar
‡‡U.T.N., Facultad Regional Paraná – U.N.E.R. Facultad de Ingeniería, mclfriedrich@hotmail.com, www.frp.utn.edu.ar

Resumen: El Fenómeno del Golpe de Ariete es un fenómeno de imprescindible consideración en el diseño y operación de acueductos bajo presión, que se presenta ante una variación brusca de las condiciones de flujo, por el cierre o la obturación de válvulas. Se aborda la resolución desde distintas disciplinas, involucrándose Mecánica de los fluidos, Matemática avanzada y Computación con fines educativos. El modelo real se basa en las ecuaciones de balance de masa (continuidad), cantidad de movimiento y de energía. Se hacen distintos supuestos para lograr un modelo dinámico simple, que tengan una correlación aceptable con los resultados teóricos y experimentales de la bibliografía. Se presentan los fundamentos de la resolución con el método de elementos finitos y se contrasta con la implementación computacional del método de las características y de diferencias finitas. Se simula el fenómeno implementando computacionalmente con software libre los métodos numéricos mencionados.

Palabras claves: *Golpe de Ariete, Métodos numéricos, Implementación Computacional*
2000 AMS Subjects Classification: 35L20- 00A72

1. INTRODUCCIÓN

Se estudia un modelo característico de flujo no permanente de líquidos, el golpe de ariete, correspondiente a un movimiento oscilatorio de fluido en una sección confinada de una cañería. Las variables de interés del fluido de variación ondulatoria de gran magnitud son la velocidad y presión.

En el análisis de este fenómeno participan aspectos propios de la dinámica de los fluidos, como de la estática del sólido y de la resistencia de materiales, lo que lo convierte en un ejemplo típico de integración de conocimientos en áreas de la ingeniería y ameritan el trabajo interdisciplinario. Además, las aplicaciones derivadas del desarrollo de sistemas de protección de las instalaciones, como de la optimización de recursos en los materiales utilizados ofrecen un campo tan vasto de posibilidades, que permiten lograr soluciones con novísimas metodologías de cálculo numéricos y consecuentes simulaciones.

Los modelos numéricos permiten entender de manera eficiente el comportamiento de la velocidad y la presión en el período transitorio e ir agregando otras variables en función de complejizar los estudios. Se pueden variar las condiciones de frontera para diferentes casos, como cierres rápidos o lentos, e ir incorporando los términos que se han despreciado. Estas posibilidades generan una flexibilidad inigualable para obtener resultados de las variables del fenómeno.

2. EL GOLPE DE ARIETE APLICANDO ELEMENTOS FINITOS

Se toma como referencia en los cálculos lo que se define como “cierre instantáneo”.

El estudio del problema y su solución fueron desarrollados inicialmente por Joukowsky [4] y Allievi [1], aunque su desarrollo analítico, similar al de la cuerda vibrante de D´Alembert, lo presentan recién en 1913. Años después, Schnyder [10] desarrollan resoluciones por método gráfico.

Joukowsky, presentó públicamente sus primeras conclusiones en 1898 sobre la hipótesis que denominó “teoría elástica” del fenómeno. Se basa en los principios de conservación de la energía, masa y cantidad de movimiento, admitiendo la deformabilidad de los líquidos ante gradientes elevados de presión, y una supuesta elasticidad del material de la tubería. La energía cinética que posee la masa fluida en el instante previo al cierre instantáneo se transforma luego de éste, por una parte en energía de deformación de la tubería que se expande por el incremento de la presión interna, y por la otra en trabajo de compresión

del fluido por el mismo efecto. En cuanto a los resultados de la aplicación de esta teoría, se puede señalar que se aproximan sensiblemente a los que arroja la experiencia, no obstante puede ocurrir en el comportamiento real que en la fase inicial y cuando la presión original sea suficientemente elevada pueden comprobarse valores de sobrepresión de entre un 20% a un 30% mayores que los obtenidos por cálculo. Esta diferencia podría adjudicarse a que, con elevados gradientes de presión el material del conducto reacciona retardadamente con el aumento brusco de la carga.

En la actualidad se han hecho numerosas investigaciones, la mayoría usa el método de las características combinado con las diferencias finitas o elementos finitos. Se puede citar los trabajos en FEM de Kochupillai et al.[6], Karra [5] y Tijsseling[7] que presentan un estudio basado en la interacción del fluido y la estructura de la cañería.

Sobre la base de las hipótesis de la teoría elástica, las ecuaciones que dan lugar al modelo tienen como incógnitas a la presión y la velocidad de un elemento de fluido. El modelo resulta equivalente a un par de ecuaciones diferenciales en derivadas parciales hiperbólicas y desacopladas, [13][8], de la forma:

$$\frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{en } \Omega \text{ x } \mathbb{R}^+ , \tag{1}$$

$$u = 0 \qquad \text{en } \partial\Omega \text{ x } \mathbb{R}^+ \text{ (condiciones Dirichlet)}, \tag{2}$$

$$u(\cdot\,;\,0) = u_0, \quad \frac{\partial u}{\partial t}(\cdot\,;\,0) = w_0 \text{ en } \Omega. \tag{3}$$

El problema a simular se basa en un sistema formado por un tanque de suministro agua que tiene una altura de $H_{cot} = 40$ m y una cañería horizontal, que tiene 600 m desde la inserción al tanque a la válvula de cierre instantáneo. En las condiciones elegidas el agua tiene una densidad $\rho = 1000 \frac{kg}{m^3}$, con una velocidad de perturbación en el fluido de $a = 1200 \frac{m}{s}$ (definida por $a = \frac{K/\rho}{1+\frac{K D}{E e}}$, donde K, módulo de elasticidad volumétrico del fluido, E, módulo de elasticidad del material, e, espesor de la pared del conducto, D, diámetro de la cañería). Cuando la válvula está abierta el fluido circula libremente por la cañería con una velocidad: $V_0 = \sqrt{2.g.H_{cot}}$ y una presión: $P_0 = \rho.g.H_{cot}$; g, aceleración de la gravedad[8].

En el momento del cierre instantáneo de la válvula, considerado como el tiempo $t = 0$, se produce un aumento brusco de presión en la válvula, que según la teoría elástica es $\Delta P = \rho.a.V_0$, que se propagará como la amplitud de la onda de sobrepresión en la cañería en los instantes siguientes y dará lugar al movimiento oscilatorio estudiado. La velocidad del fluido en la válvula, para tiempos $t > 0$, es cero, porque se supone totalmente obturada.

Considerando fricción cero, el fenómeno oscilatorio tiene un ciclo de tiempo $T = \frac{4L}{a}$, en las secuencias que siguen al cierre rápido de la válvula[9] [12]. La presión teórica esperada en la válvula y en el punto medio de la cañería es la especificada en la Figura 1.

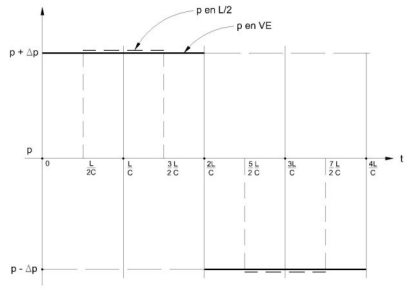


Figura 1: Característica de la presión en función del tiempo

Para aplicar el método de elementos finitos que permite resolver de manera aproximada la ecuación diferencial (1) con las condiciones de borde e iniciales (2) y (3), se hace una partición del dominio espacial, intervalo [0 , L], en n partes: $0 = x_0 < x_1 < \ldots < x_n = L$, $\Delta x_i = x_i - x_{i-1}$ con $i = 1, \ldots, n$. Esto es, se hace una semidiscretización del dominio. Sea S_h el espacio de funciones continuas en [0,L], lineales en cada subintervalo y que valen cero en los extremos del intervalo. Para cada i = 1,…,n-1 se usan las funciones de base de S_h tales que $\phi_i(x_j) = \delta_{ij}$, [3]. La forma débil del problema es:

$$\left(\phi_i(x), \frac{\partial^2 u}{\partial t^2}\right) + a^2 \left(\frac{\partial \phi_i(x)}{\partial x}, \frac{\partial u}{\partial x}\right) = 0, \ \forall \phi_i(x) \in S_h. \quad (4)$$

La aproximación discreta de la solución del problema débil de la ecuación (4), [14] [11]se expresa como:

$$u_h = \sum_{j=1}^{n} \alpha_j(t)\, \phi_j(x). \quad (5)$$

El problema débil discretizado es:

$$\left(\phi_i(x), \frac{\partial^2 u_h}{\partial t^2}\right) + a^2 \left(\frac{\partial \phi_i(x)}{\partial x}, \frac{\partial u_h}{\partial x}\right) = 0, \quad \forall \phi_i(x) \in S_h. \quad (6)$$

Para discretizar la derivada segunda respecto del tiempo se usa el método backward de diferencias finitas, que tiene error de truncamiento menor al de la aproximación de diferencias centras. Se elige un paso constante $\Delta t = k$, y con valores de tiempo $t = tn = n\,k$, con $n \in \mathbb{Z}^+$. A la aproximación de $u(t_n)$ se llama $U^n = U_h^n \in S_h$. En cada nivel de tiempo U_h^n se expresa como en (5). El método de Euler es incondicionalmente estable para cualquier h y Δt. Considerando la ecuación (6) para i = 1,…, n-1, resulta:

$$\left(\phi_i(x), \frac{U^n - 2U^{n-1} + U^{n-2}}{\Delta t^2}\right) + a^2 \left(\frac{\partial \phi_i(x)}{\partial x}, \frac{\partial U_h^n}{\partial x}\right) = 0. \quad (7)$$

La ecuación (7) da lugar al sistema matricial a resolver. El equivalente matricial:

$$[M]\,[\alpha^n] + \Delta t^2 a^2 [K]\,[\alpha^n] = 2[M]\,[\alpha^{n-1}] - [M]\,[\alpha^{n-2}], \quad (8)$$

donde:

$[M] = \left(\phi_i(x), \phi_j(x)\right)_{i,j}$; $[K] = \left(\frac{\partial \phi_i(x)}{\partial x}, \frac{\partial \phi_j(x)}{\partial x}\right)_{i,j}$; los valores iniciales, para t = 0, de la función solución y su derivada son datos; se usan los valores proyectados en el espacio $L_2\,(\Omega)$. Los valores de la función permiten obtener $[\alpha^0]$, y las derivadas, los valores virtuales $[\alpha^{-1}]$ en un tiempo anterior al tiempo cero.

Los resultados obtenidos por elementos finitos se comparan con los de los métodos de diferencias finitas y el método de líneas presentados en un trabajo anterior [2].

El trabajo se ha planteado de manera interdisciplinaria con los especialistas en diseño de instalaciones y los calculistas, a fin de asegurar la factibilidad de los resultados. En cuanto los resultados teóricos simulados con los métodos numéricos estudiados con las simulaciones de este caso simplificado, la práctica de la ingeniería indica que debajo del cero absoluto de presión no se puede bajar, por lo que no se ajustan al modelo físico pero si al teórico. La diferencia con la realidad no se produce por la elección de los métodos numéricos sino de las simplificaciones hechas en el modelo físico a simular.

Si bien las sobrepresiones consecuentes pueden significar un enorme riesgo sobre la integridad de la conducción, el mayor riesgo de colapso se da como consecuencia del vacío en la fase negativa, por el espesor de pared de la cañería. Las pérdidas económicas ameritan el mejoramiento de las simulaciones y la consideración de modelos físicos por otros factores que no se han considerado en este trabajo como: viscosidad del fluido, respuesta inmediata del material a la carga súbita, límite de vacío y presión de vapor del líquido, cierre de la válvula en tiempo finito, rápido y no precisamente instantáneo, elasticidad imperfecta del conducto, sin despreciar otros que pudieran mencionarse pero que son sin duda de mucha menor importancia[2].

3. CONCLUSIONES

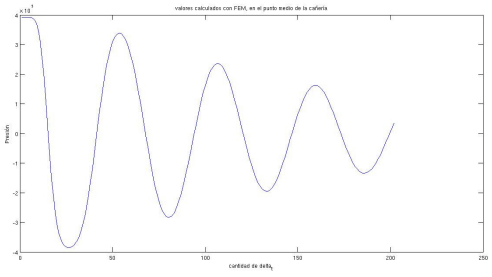


Figura 2: Valores de la presión, calculados con FEM, en el punto medio de la cañería

En cuanto a los resultados logrados con la aplicación propuesta en este proyecto puede señalarse lo siguiente:

1. La ilustración dinámica obtenida con elementos finitos muestra con claridad el fenómeno armónico que predice la teoría basada en hipótesis sencilla de elasticidad perfecta tanto del material como del fluido intervinientes.
2. La frecuencia concordante y el amortiguamiento de la amplitud en el tiempo producto de la corrección de las hipótesis iniciales corroboran la aproximación del comportamiento del modelo teórico con el del sistema real.
3. No se ha modificado en las hipótesis iniciales la limitación de la depresión en el tubo a un valor próximo al de la tensión de vapor y eso ha sido deliberadamente obviado hasta tener un comportamiento teórico que se adecue al real, lo que ha sucedido. Es por lo tanto una faceta de desarrollo inmediato próximo.
4. Se puede concluir de trabajos anteriores que el método de las características permite obtener resultados más parecidos a los teóricos que los obtenidos con elementos finitos o diferencias finitas.

REFERENCIAS

[1] L. ALLIEVI, *Teoría del colpo d'ariete – nota 1: esposizione generale del metodo; nota 2: il colpo d'ariete in chiusura; nota 3: il colpo d'ariete in apertura; nota 4: contraccolpi di ritorno a regime; nota 5 fenomeni di risonanza,* Atti dell'Associazione Elettrotecnica Italiana, Vol.17, (1913), pp.127-150; pp 861-900; pp.1127-1145; pp.1235-1253.

[2] A.CARBONELL, M. BENITEZ, L.GIMENEZ Y M. FRIEDRICH, *Aproximándonos al golpe de ariete mediante el método de elementos finitos*, VI CAEDI "Formando al Ingeniero del siglo XXI", Salta, (2008). ISBN 978-987-633-011-4

[3] C. JOHNSON, *Numerical solution of partial differential equations by the finite element method,* Studentlitteratur, Lund, Sweden, 1995.

[4] N. JOUKOWSKY, *On the hydraulic hammer in water supply pipes,* Memoires de l'Académie Imperials des Sciences de St.-Petersburg, Series 8, 9(5), (1898), pp1-71.

[5] C. KARRA ET AL., *Finite element analysis of a fluid structure interaction in flexible pipe line,* African Journal of Science and technology, Science and Engineering series 8, Nª 1, pp. 63-70.

[6] J. KOCHUPILLAI ET AL. *A new finite element formulation base on the velocity of floor for water hammer problems,* Elsevier, India, 2004

[7] D. LESLIE AND A. TIJSSELING, *Travelling discontinuities in water hammer theory: attenuation due to friction,* In Proceedings of the 8th International Conference on Pressure Surges, BHR Group, The Hague, 2000.

[8] R.MOTT, *Mecánica de Fluidos Aplicada*, CUARTA EDICIÓN, Ed. Prentice-Hall, Mexico, 2006.

[9] M. POTTER, WIGGERT, D. C., *Mecánica de Fluidos,* Tercera Edición, Ed. Thomson, Mexico, 2002.

[10] O. SCHNYDER, (1929). *Water hammer in pumping pipelines,* Wasserkraft undWasserwirtschaft 94(22): 271-273; 94(23): 283-286.

[11] A.SCHMIDT, A. ET AL. *Design of Adaptative Finite Element Software*. *The finite element toolbox ALBERTA*. Springer Berlin Heidelberg, Germany, 2005.

[12] V. STREETER Y B. WYLIE, *Mecánica de Fluidos*, Ed. Mcgraw-Hill, México, 2000.

[13] B. WYLIE AND V. STREETER, Fluid *Transients in Systems*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA, 1993.

[14] O. ZIENKIEWICZ AND R. TAYLOR, *El Método de los Elementos Finitos*, Ed. Mcgraw-Hill, Mexico, 1994.

NUMERICAL METHODOLOGY TO MODEL AND MONITOR CO2 SEQUESTRATION

Gabriela B. Savioli[♭,†] and Juan E. Santos[♭]

[♭]*Laboratorio de Ingeniería de Reservorios, IGPUBA and Departamento de Ingeniería Química, Facultad de Ingeniería, Universidad de Buenos Aires,, Av. Las Heras 2214 Piso 3 C1127AAR Buenos Aires, Argentina, gsavioli@di.fcen.uba.ar*

[†]*CONICET, Facultad de Cs Astron. y Geofísicas, Universidad Nac. de La Plata, Paseo del Bosque S/N, La Plata, B1900FWA, Argentina, and , Department of Mathematics, Purdue University, 150 N. University St., West Lafayette, IN 47907, U.S.A, santos@fcaglp.unlp.edu.ar*

Abstract: In this work we introduce a methodology to model and monitor CO_2 sequestration. For that purpose we integrate numerical simulators of CO_2-brine flow and seismic wave propagation. The simultaneous flow of brine and CO_2 is modeled applying the Black-Oil formulation for two phase flow in porous media, which uses the PVT data as a simplified thermodynamic model. Seismic monitoring is modeled using Biot's equations of motion describing wave propagation in fluid-saturated poroviscoelastic solids. Two numerical examples of CO_2 injection and time-lapse seismics are analyzed. In the first one a sealed shale layer causes CO_2 accumulation beneath it. In the second one the seal has a fracture that allows CO_2 migration; therefore a second accumulation region appears beneath the top of the formation. Both cases show the capability of the proposed methodology to identify the horizontal and vertical saturation distribution of CO2.

Keywords: *Numerical Modeling, Poroelasticity, CO_2 Sequestration, Finite element methods.*
2000 AMS Subject Classification: 76S05 - 86A15 - 76M10

1 INTRODUCTION

CO_2 sequestration is a means of mitigating the greenhouse effect [1]. Geologic sequestration involves injecting CO_2 into a target geologic formation at depths typically greater than 1000 m where pressure and temperature are above the critical point for CO_2 (31.6 C, 7.38 MPa). Saline aquifers are a good alternative as storage sites due to their large volume and their common occurrence in nature. Nevertheless, very little was known about the effectiveness of CO_2 sequestration over very long periods of time. In this way, numerical modeling of CO_2 injection and seismic monitoring is an important tool to understand the behavior of CO_2 after injection and to make long term predictions in order to prevent CO_2 leaks from the storage into the atmosphere. The first commercial CO_2 injection project is that of the Sleipner field in the Utsira Sand aquifer (North Sea) [1]-[2]. CO_2 separated from natural gas produced at Sleipner is currently being injected into the Utsira Sand, a saline aquifer some 26000 km^2 in area. Injection started in 1996 and is planned to continue for about twenty years, at a rate of about one million tonnes per year.

Time-lapse seismic surveys aim to demonstrate storage integrity. Recent papers [3]-[4] successfully apply seismic modeling for monitoring the spatio-temporal distribution of CO_2 using synthetic generated CO_2 saturation fields. Instead, in this work we employ numerical simulations of CO_2 injection; therefore saturation fields are obtained as a result of the simultaneous flow of CO_2 and brine in porous media.

The final objective is to test that underground storage is a safe and verifiable technology in the long term.

2 THE BLACK-OIL FORMULATION OF TWO-PHASE FLOW IN POROUS MEDIA

The simultaneous flow of brine and CO_2 is described by the well-known Black-Oil formulation applied to two-phase, two component fluid flow [5]. In this model, CO_2 may dissolve in the brine but the brine is not allowed to vaporize into the CO_2 phase. This formulation uses, as a simplified thermodynamic model, the following PVT data, determined using the Hassanzadeh's correlations [6]-[7]: R_s: CO_2 solubility in brine; B_{CO2}: CO_2 formation volume factor, and B_b: brine formation volume factor. The nonlinear system of partial differential equation is,

$$\nabla \cdot (\underline{k}(\frac{k_{rCO2}}{B_{CO2}\mu_{CO2}}(\nabla p_{CO2} - \rho_{CO2} g \nabla D) + \frac{R_s k_{rb}}{B_b \mu_b}(\nabla p_b - \rho_b g \nabla D))) + q_{CO2} = \frac{\partial\Big[\phi\left(\frac{S_{CO2}}{B_{CO2}} + \frac{R_s S_b}{B_b}\right)\Big]}{\partial t}, \tag{1}$$

$$\nabla \cdot (\underline{k}(\frac{k_{rb}}{B_b \mu_b}(\nabla p_b - \rho_b g \nabla D)) + q_b = \frac{\partial\Big[\phi(\frac{S_b}{B_b})\Big]}{\partial t}. \tag{2}$$

The unknowns are the fluid pressures p_{CO2}, p_b and saturations S_{CO2}, S_b for the CO_2 and brine phases. The parameters k and ϕ are the absolute permeability and porosity respectively. Also, for $\beta = CO2, b$, the funcions $k_{r\beta}$, μ_β and ρ_β are the relative permeability, viscosity, and density of the β-phase, respectively.

Two algebraic equations relating the saturations and pressures, complete the system:

$$S_b + S_{CO2} = 1, \qquad p_{CO2} - p_b = P_C(S_b), \tag{3}$$

where P_C is the capillary pressure.

The solution of the Black-Oil fluid-flow model was obtained employing the public domain software BOAST [8], which solves the differential equations using IMPES, a finite difference technique [9].

3 Biot's Equations of Motion

Let us consider a 2D isotropic fluid-saturated porous material Ω. The oscillatory motion of Ω at the angular frequency ω subject to external sources $F^{(s)}$ and $F^{(f)}$ obeys Biot's equation of motion [3]

$$-\omega^2 \rho_b u^{(s)} - \omega^2 \rho_f u^{(f)} - \nabla \cdot \sigma(u) = F^{(s)} \tag{4}$$

$$-\omega^2 \rho_f u^{(s)} - \omega^2 \widetilde{g} u^{(f)} + i\omega b u^{(f)} + \nabla p_f(u) = F^{(f)}. \tag{5}$$

The unknowns are $u^{(s)}$ and $u^{(f)}$, the time Fourier transforms (FT) of the averaged displacement vectors of the solid and fluid phases, respectively. Also, ρ_f and ρ_b denote the densities of the single-phase fluid and the bulk material, $\widetilde{g}$ and b are mass and viscous coupling coefficients, σ_{ij} is the FT of the stress tensor of the bulk material and p_f is the FT of the fluid pressure. See [4] for the definition of the variables involved in (4)-(5).

Biot's equations were solved with the finite element method, employing a 2D non-conforming finite element space for each component of the solid displacement vector and the vector part of the Raviart Thomas Nedelec space of zero order for the fluid displacement [10].

4 Numerical Experiments

4.1 Idealized model of the Utsira formation

To test the proposed methodology, we consider an idealized geometrical and physical domain consisting of 5 regions as shown in Figure 1. The upper 100 m is region Ω_1, a sand of permeability 60 mD and porosity 0.32. Ω_2 is a sealed shale 2 m thick, the top of the Usira formation. Regions Ω_3 and Ω_5 are the Utsira formation, of permeability 1000 mD and porosity 0.37. We assume that Ω_4 is either a shale seal layer or a fractured shale seal layer within the Utsira sand. The medium was excited with a compressional point source located at x= 400 m, z= 710 m.

The Biot model assumes a single-phase fluid, therefore effective fluid density, viscosity and bulk modulus were obtained using the properties of the CO_2 and brine weighted by the corresponding saturations computed by the fluid-flow simulator.

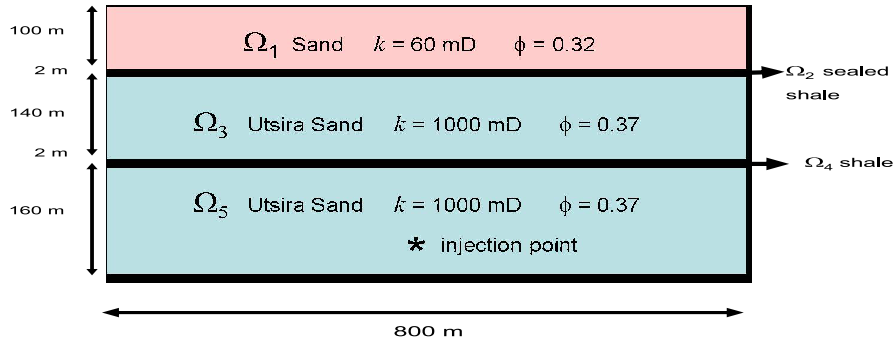


Figure 1: Idealized model of the Utsira formation. The injection point is located at x= 400 m, z= 1060 m

4.2 Injection Modeling

CO_2 saturation distribution after 5 years of injection is shown in Figure 2 when Ω_4 is a seal layer (left plot) and when Ω_4 is a fractured seal layer (right plot). Results are computed applying the BOAST simulator. In the presence of a fracture, part of the injected CO_2 has migrated through it Therefore, two regions of CO_2 accumulation can be observed in the right plot: one beneath Ω_4 and the other beneath the top of the Usira formation. When Ω_4 is sealed only the accumulation beneath Ω_4 occurs (left plot).

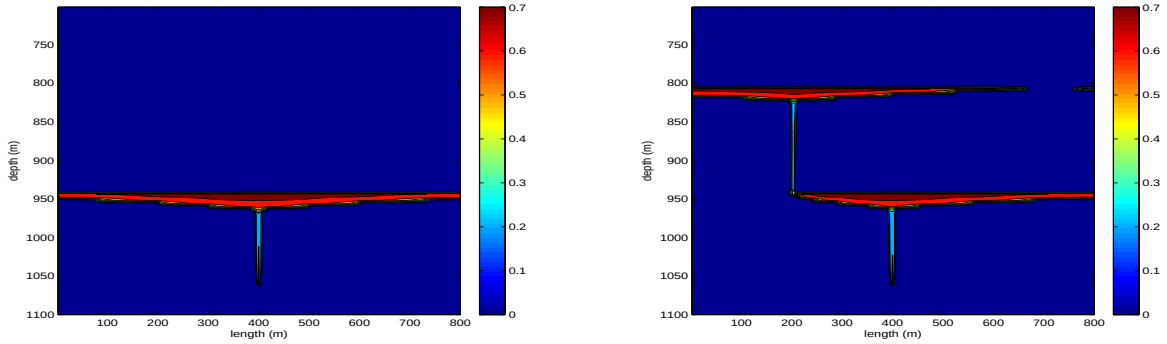


Figure 2: CO_2 saturation distribution after 5 years of injection

4.3 Seismic Monitoring

Time histories measured near the surface before and after 5 years of CO_2 injection are shown in Figures 3: before CO_2 injection (left plot); after 5 years of CO_2 injection when Ω_4 is a seal layer (center plot) and when Ω_4 is a fractured seal layer (right plot) The first reflection in all figures is due to the direct wave coming from the point source located at x= 400 m, z= 710 m. The second reflection in the center and right plots is generated by the CO_2 accumulations below the thin shale layer at depth z = 940 m. In the right plot a third reflection appears, this is due to the accumulation below the top of the Utsira.

Figure 4 displays traces of the vertical component of the particle velocity of the solid phase before the injection (black curve), after 5 years of CO_2 injection when Ω_4 is a seal layer (red curve) and when Ω_4 is a fractured seal layer (green curve). The strong arrival at about 250 ms corresponds to a reflection due to the CO_2 accumulation beneath the seal. This is detected by both curves, the red and the green one. Besides, the green curve also detects the CO_2 accumulation beneath the top of Utsira, by a reflection at about 200 ms.

5 Conclusions

In this work we integrate numerical simulators of CO_2-brine flow and seismic wave propagation to model and monitor CO_2 storage in saline aquifers. Numerical examples show the effectiveness of this metodology

to detect CO_2 accumulations. Therefore, it constitutes an important tool to analyze storage integrity, provide early warning should any leakage occur, and monitor the migration and dispersal of the CO2 plume.

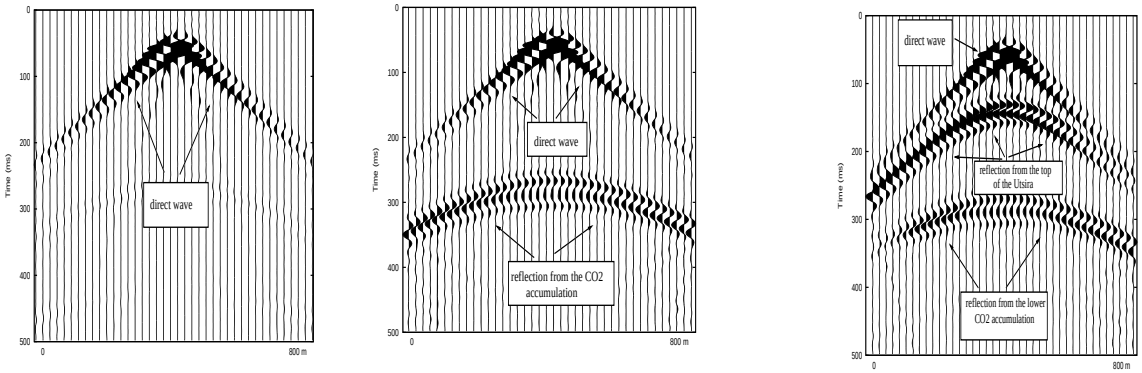


Figure 3: Time histories measured near the surface before and after 5 years of CO_2 injection

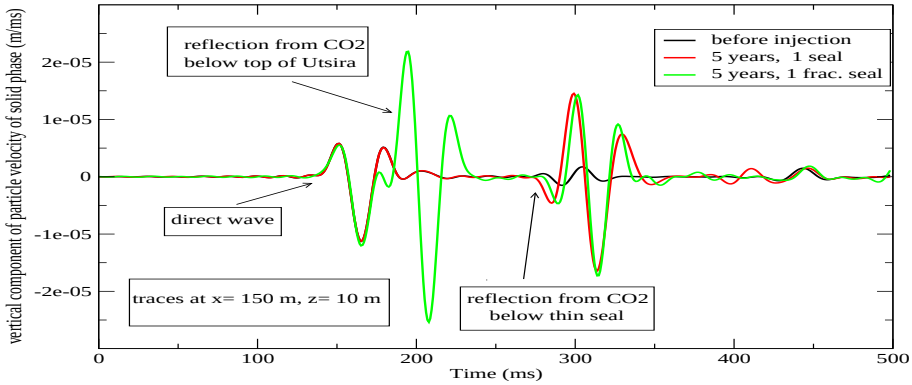


Figure 4: Traces of particle velocity of the solid phase before and after 5 years of CO_2 injection for both cases of Ω_4

REFERENCES

[1] R. ARTS, A. CHADWICK, O. EIKEN, S. THIBEAU, AND S. NOONER, *Ten years of experience of monitoring CO_2 injection in the Utsira Sand at Sleipner, offshore Norway*,First break, 26 (2008), pp.65-72.

[2] A. CHADWICK, R. ARTS, AND O. EIKEN, *4D seismic quantification of a growing CO_2 plume at Sleipner, North Sea* in Petroleum Geology: North West Europe and Global Perspectives - Proc. 6th Petroleum Geology Conference, Dore A G and Vincent B (Eds), Geological Society, London, (2005), 1385-1399.

[3] J. SANTOS, J. RUBINO, AND C. RAVAZZOLI, *Modeling mesoscopic attenuation in a highly heterogeneous Biot's medium employing an equivalent viscoelastic model* in Procceedings of the 78th Annual International Meeting SEG, Las Vegas, (2008), 2212-2215.

[4] J. CARCIONE, S. PICOTTI, D. GEI, AND G. ROSSI, *Physics and Seismic Modeling for Monitoring CO_2 storage*, Pure and Applied Geophysics, 163 (2006), pp. 175-207.

[5] K. AZIZ, AND A. SETTARI, *Petroleum Reservoir Simulation*, Elsevier Applied Science Publishers, Great Britain, 1985.

[6] H. HASSANZADEH,M. POOLADI-DARVISH, A. ELSHARKAWY, D. KEITH, AND Y. LEONENKO, *Predicting PVT data for CO_2-brine mixtures for black-oil simulation of CO_2 geological storage*, International Journal of Greenhouse Gas Control, 2 (2008), pp.65-77.

[7] N. SPYCHER, AND K. PRUESS, *CO_2-H_2O mixtures in the geological sequestration of CO_2. II. Partitioning in chloride brines at 12-100 C and up to 600 bar*, Geochim. Cosmochim, Acta 69, 13 (2005), pp.3309-3320.

[8] J. FANCHI *Principles of Applied Reservoir Simulation*, Gulf Professional Publishing Company, Houston, Texas, 1997.

[9] G. SAVIOLI, AND M. S. BIDNER *Simulation of the oil and gas flow toward a well - A stability analysis*, Journal of Petroleum Science and Engineering, 48 (2005), pp.53-69.

[10] J. SANTOS, AND D. SHEEN, *Finite element methods for the simulation of waves in composite saturated poroviscoelastic materials*, SIAM J Numer Anal., 45 (2007), pp.389-420.

CORRECCIÓN GEOMÉTRICA DE LA POSICIÓN DE REFLECTORES GEOLÓGICOS USANDO MIGRACIÓN SÍSMICA

Saúl Becerra Ospina, Hernán Estrada B y Jorge M. Ruíz V

Departamento de Matemáticas, Universidad Nacional de Colombia Bogotá D.C. Colombia, www.unal.edu.co

Resumen: En el procesamiento convencional de datos de reflexión sísmica se asume que los reflectores geológicos se encuentran en la mitad de cada par fuente receptor, lo cual no es necesariamente cierto, conduciendo a una localización errada de los reflectores geológicos. En este trabajo, mediante experimentación numérica, se estudia la técnica de migración basada en la ecuación de onda escalar, para corregir la posición de los puntos de reflexión y lograr su correcta localización. En una primera parte se simulan secciones sísmicas apiladas observadas en la superficie, posteriormente se utiliza la migración en tiempo inverso para obtener la posición exacta de los reflectores y lograr imágenes de la configuración geológica escogida a priori. Se analiza la estabilidad para la discretización espacial y temporal y también las condiciones de frontera ficticias para representar límites computacionales no reflectantes y modelar el terreno como un dominio espacial semiinfinito.

Palabras clave: *Experimentación numérica, Migración en tiempo inverso.*
2000 AMS Subject Classification: 86A22 - 86A20

1. INTRODUCCIÓN

Conocer la configuración geológica subsuperficial es un interesante reto para la ciencias puras y aplicadas. La geofísica enmarca las técnicas diseñadas para modelar el subsuelo y una de sus ramas más importantes, dedicada a la búsqueda de materiales de interés económico, es la prospección. La sísmica de reflexión tiene amplia aceptación en ámbitos industriales, ya que ofrece una buena resolución, que se traduce en imágenes exactas de las interfaces entre estratos geológicos (también denominados reflectores geológicos) y en la óptima estimación de parámetros físicos de los materiales, útiles para la caracterización de yacimientos de minerales como petróleo [5].

La técnica de migración es un problema propuesto hace más de cincuenta años que consiste en corregir la posición de los reflectores geológicos en imágenes obtenidas a partir del análisis convencional de datos sísmicos [1]. Puede realizarse geométricamente, aplicando los principios de la física óptica, o mediante la solución de la ecuación de onda en el dominio frecuencia-espacio o tiempo-espacio. Como los esfuerzos para realizar la migración no son pocos, era considerada como superflua dentro del procesamiento de datos de reflexión. Sin embargo, en los últimos años esta percepción ha cambiado y ahora no es vista únicamente como método de mejoramiento de imágenes, sino también, como una poderosa herramienta dentro del proceso de inversión sísmica.

En este trabajo se modela computacionalmente el sistema físico que gobierna la propagación de ondas sísmicas con aproximación acústica y se aplica al estudio de la migración basada en la ecuación de onda escalar en el dominio tiempo-espacio. El modelamiento numérico de este sistema presenta varios retos, como el costo computacional, la dispersión numérica y condiciones no reflectantes sobre las fronteras computacionales para modelar la tierra como un espacio bidimensional semiinfinito.

2. MODELO MATEMÁTICO

Los medios elásticos soportan la propagación de dos tipos de ondas: ondas de presión o ondas P y ondas de cizalla o ondas S [4]. El campo de desplazamientos generado por este tipo de ondas es ortogonal, lo cual posibilita observar de manera independiente ondas P de las ondas S. Esto permite aproximar la propagación de ondas sísmicas mediante un modelo acústico.

Los levantamientos sísmicos de reflexión son de escala local, por lo tanto la superficie terrestre se puede asumir como plana, permitiendo para el problema bidimensional, seleccionar como sistema de referencia

[1] El artículo pionero es [1]

un plano cartesiano, donde uno de sus ejes coincide con la superficie terrestre, mientras el otro se escoge ortogonal al terreno y su sentido positivo hacia el interior de la Tierra.

Los impactos mecánicos que perturban el sistema pueden modelarse mediante una función fuente $S(t)$ que tiene la siguiente forma

$$S(\mathbf{x},t) = s(t)\delta(\mathbf{x}-\mathbf{x_r}), \tag{1}$$

donde δ es la delta de Dirac, $\mathbf{x} = (x,z)$ es el vector posición, $\mathbf{x_r} = (x_r, z_r)$ son las coordenadas donde hay fuentes y

$$s(t) = A\,\mathrm{sen}(2\pi f t)e^{-2\pi f t^2}. \tag{2}$$

De acuerdo con la discusión anterior, un modelo matemático de un levantamiento sísmico es

$$\begin{aligned}
\frac{1}{(c(x,z))^2}\frac{\partial^2}{\partial t^2}u(x,z,t) &= \nabla^2 u(x,z,t) + S(x,z,t), \quad -\infty \le x \le \infty, \quad 0 \le z \\
u(x,z,0) &= 0 \qquad (3)\\
u_t(x,z,0) &= 0 \qquad (4)\\
u(x,z,t) &= 0, \quad z<0, \qquad t\in[0,T],
\end{aligned}$$

donde u es el campo de desplazamiento y $c(x,z)$ es la función de velocidad de propagación de la onda. La condición inicial establece que el medio esta en reposo, mientras la siguiente condición indica que no hay propagación de las ondas en la atmósfera.

Para la migración se utiliza el mismo modelo planteado en (3), pero para un campo $\widetilde{u}(x,z,\tau)$, donde se reemplaza el tiempo usual t por el tiempo inverso $\tau = T - t$, siendo T el tiempo total de observación. A este método se le denomina migración en tiempo inverso (RTM, por sus siglas en inglés). La función fuente, ahora corresponde a una sección sísmica apilada y la condición inicial significa que la observación en campo se realiza hasta que se disipen las ondas sísmicas generadas.

3. Modelamiento numérico

Teniendo en cuenta que la geometría del problema (3) es simple, la solución se puede aproximar con el siguiente esquema de diferencias finitas de segundo orden en espacio y tiempo [3]

$$U_{i,j}^{n+1} = 2U_{i,j}^n - U_{i,j}^{n-1} + \left(\frac{c_{i,j}\Delta t}{\Delta h}\right)^2 \left[U_{i+1,j}^n + U_{i-1,j}^n + U_{i,j+1}^n + U_{i,j-1}^n - 4U_{i,j}^n\right] + S_{i,j}^n, \tag{5}$$

donde $c_{i,j}$ es la velocidad de propagación en el punto (x_i, y_j), $\Delta h = \Delta x = \Delta z$ el espaciado de la grilla, $S_{i,j}^n$ es

$$S(x_i, z_j, t_n) = s(t_n)\delta(x_i - x_k, y_i - y_h), \tag{6}$$

(x_k, y_h) son las coordenadas de los puntos de la grilla que se consideran fuentes.

Para RTM la solución se puede aproximar por un esquema análogo a (5)

$$\widetilde{U}_{i,j}^{n+1} = 2\widetilde{U}_{i,j}^n - \widetilde{U}_{i,j}^{n-1} + \left(\frac{c_{i,j}\Delta t}{\Delta h}\right)^2 \left[\widetilde{U}_{i+1,j}^n + \widetilde{U}_{i-1,j}^n + \widetilde{U}_{i,j+1}^n + \widetilde{U}_{i,j-1}^n - 4\widetilde{U}_{i,j}^n\right] + \widetilde{S}_{i,j}^n. \tag{7}$$

La función $\widetilde{S}_{i,j}^n$, corresponde a los datos sísmicos observados. Adicionalmente, se debe tener presente la definición de τ, por lo cual se cumple que

$$\widetilde{u}(x,z,\tau) = u(x,z,T-\tau), \tag{8}$$

o para el sistema discretizado

$$\widetilde{U}(i\Delta x, j\Delta z, n\Delta\tau) = U(i\Delta x, j\Delta z, T - (N_t - n)\Delta\tau). \tag{9}$$

Se observa en (9), que la migración se debe realizar hasta un $n = n_{max} < N_t$, de tal manera que se propague la onda en tiempo inverso hasta $\tau < T$ y que permita discriminar los reflectores geológicos, ya que $\tau = T$ es equivalente a $t = 0$ en (3), lo que significa que el medio se encuentra en reposo.

4. CONDICIONES DE FRONTERA ABSORBENTES

El dominio espacial para la solución de (3) es el semi-plano $\Omega_\infty = \{(x, z) : 0 \leq z\}$. En otros términos, se trata de un dominio infinito. Sin embargo, la memoria computacional es limitada y por lo tanto, una solución numérica, solo es posible para un dominio finito, por lo tanto se tiene una frontera computacional. El problema es imponer condiciones adecuadas para evitar reflexiones sobre dichas fronteras. Existen métodos refinados para modelar este tipo de fronteras (por ejemplo, [2]), que requieren una importante capacidad de almacenamiento o de computo. En este trabajo se modelan las fronteras no reflectantes con el bien conocido método de Reynolds, ampliamente utilizado en aplicaciones de migración sísmica. Es una aproximación de condición de frontera no reflectante local, tanto espacial como temporal, por lo que no implica un costo adicional de almacenamiento en memoria y tampoco se incrementa el tiempo de computo. El modelo consiste en factorizar el operador diferencial de la ecuación de onda acústica.

5. RESULTADOS NUMÉRICOS

Las aproximaciones numéricas se calcularon con un programa de computo escrito en lenguaje C++. Tiene un diseño orientado a objetos, lo cual permite manipularlo y extenderlo fácilmente. El ejemplo seleccionado es el de un perfil compuesto de dos capas geológicas con forma sinclinal, ver Figura 1. Este modelo permite ilustrar claramente el objetivo de la migración. Con (5) se simula la sección sísmica apilada y con (7) se realiza la migración en tiempo inverso.

6. CONCLUSIONES Y RECOMENDACIONES

De acuerdo con los experimentos numéricos realizados, el método de migración basado en la ecuación de onda es exacto para corregir la posición de los reflectores geológicos. Sin embargo, requiere de un modelo de velocidad en profundidad conocido lo cual constituye una gran desventaja.

Según los resultados de dispersión numérica, el número de canales (receptores) es crucial en el modelamiento de datos de campo, dado que se relaciona directamente con el tamaño de paso de la malla espacial. Esto tiene implicaciones sobre los costos de levantamientos de campo, por esto, es interesante estudiar métodos de densificación de datos de reflexión.

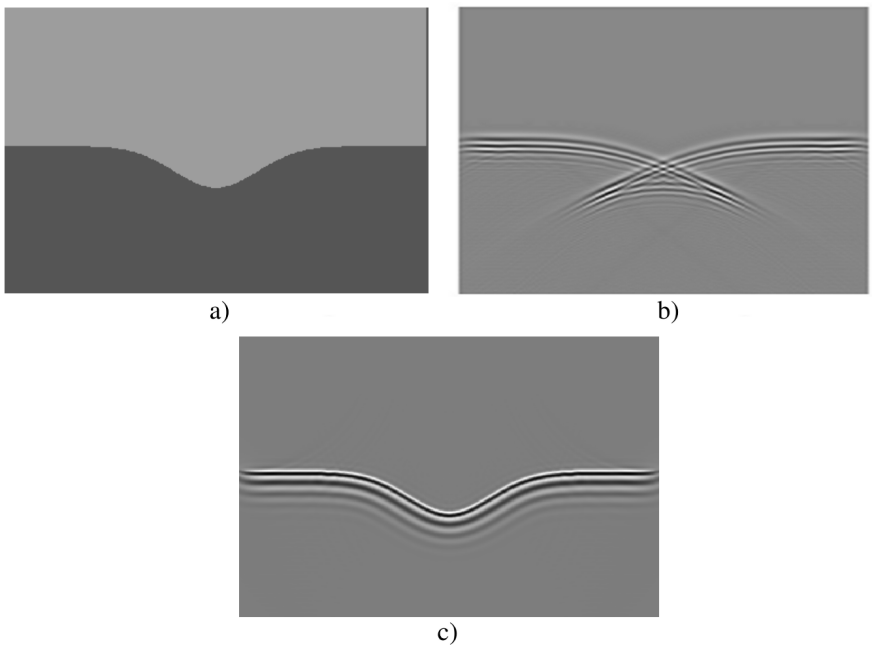


Figura 1: Modelo de perfil geológico. a) Modelo de velocidad $c(x, z)$. b) Sección sísmica apilada. c) Imagen corregida con RTM.

REFERENCIAS

[1] HAGEDOORN J. G., *A process of seismic reflection interpretation.*, Geophysical Prospecting. Vol 2 (1954), pp. 85-127.

[2] GROTE M. J. Y KIRSCH C, *Nonreflecting boundary condition for time-dependent multiple scattering.*, Journal of Computational Physics. (2007), pp. 41-62.

[3] LEVEQUE R. J., *Finite Difference Methods for Ordinary and Partial Differential Equations.* SIAM, 2007.

[4] SADD M. H., *Elasticity Theory, Applications, and Numerics.* Elsevier, 2005.

[5] VEEKEN P. C. H., *Seismic stratigraphy, besin analysis and reservoir characterisation*, Elseiver, 2007.

DISEÑO ÓPTIMO DE PLANTAS DE TRATAMIENTO DE AGUAS RESIDUALES

Cecilia I. Stoklas† y Víctor H. Cortínez†‡

†Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca. 11 de abril 461, B8000LMI Bahía Blanca, Argentina. stoklas@frbb.utn.edu.ar
‡CONICET, vcortine@frbb.utn.edu.ar, www.frbb.utn.edu.ar

Resumen: En el presente trabajo se describe el desarrollo de un modelo computacional para la localización óptima de tuberías de descarga de plantas de tratamiento de efluentes urbanos en un cuerpo de agua, como también para la determinación del grado de purificación adecuado, buscando cumplir en forma óptima con objetivos ambientales y económicos. Para ello se emplea una solución numérica mediante el método de elementos finitos de un modelo hidrodinámico bidimensional, basado en la teoría de aguas poco profundas, acoplado con ecuaciones bidimensionales de transporte de contaminantes del tipo difusión-advección. El problema de optimización se formula mediante una combinación de una técnica de búsqueda aleatoria con un método basado en gradiente. Para la formulación de las restricciones del problema se hace uso de funciones de influencia obtenidas convenientemente a partir del problema adjunto del modelo de transporte. Tales ecuaciones fueron implementadas en el programa de simulación por elementos finitos FlexPDE.

Palabras claves: *Modelo Hidrodinámico, Transporte de Contaminantes, Optimización, Método Adjunto.*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

El vertido de aguas servidas en cuerpos de agua aledaños a grandes ciudades es una situación seria de contaminación ambiental, ya que el nivel de tales contaminantes comúnmente supera las posibilidades de autodepuración de los sistemas naturales, implicando la necesidad de efectuar un tratamiento adecuado. En efecto, el vertido de efluentes cloacales se realiza en general mediante una tubería submarina cuya salida se encuentra a cierta distancia de la costa. La contaminación provocada alcanza niveles que pueden medirse con diversos indicadores, entre ellos los Coliformes fecales (CF), encontrándose presentes únicamente en efluentes cloacales y no en efluentes industriales. El exceso de dicha concentración con respecto a los valores permitidos puede generar impactos negativos en el ecosistema.
Es viable, desde el punto de vista práctico, definir áreas de resguardo ambiental dedicadas a diferentes usos, como toma de agua potable, zonas de pesca, zonas de recreación, etc., pudiendo plantearse el problema de la siguiente manera: se desea establecer la localización de la salida de la tubería de los efluentes urbanos tratados en plantas de depuración, así como el grado de tratamiento necesario para garantizar un nivel de contaminación tolerable en las zonas previamente aludidas procurando que el costo económico sea el mínimo posible.
A efectos de resolver tal problema se ha desarrollado una herramienta computacional basada en la formulación de las ecuaciones hidrodinámicas en aguas someras [1] combinadas con las ecuaciones de advección-difusión de transporte de contaminantes [2]. Tales ecuaciones, con sus correspondientes condiciones de borde e iniciales, son implementadas en el programa de elementos finitos FlexPDE [3], obteniéndose el régimen de corrientes y el transporte de sustancias dentro del cuerpo de agua.
Esta solución numérica brinda la distribución espacial y temporal del contaminante CF, pudiéndose determinar las correspondientes concentraciones en las áreas de resguardo predefinidas. Este tipo de simulación requiere la solución del problema de transporte para cada conjunto de prueba de las variables de diseño dadas por las coordenadas de las tuberías de descarga y de los coeficientes de reducción de los caudales vertidos. Un enfoque alternativo consiste en determinar los coeficientes de influencia del problema de transporte, que relacionan las concentraciones causadas en las zonas protegidas con vertidos unitarios de las tuberías de descarga. Una manera eficiente de obtener tales coeficientes se realiza obteniendo la solución mediante el método de elementos finitos (EF) del problema adjunto asociado al problema de transporte. Tal enfoque solo requiere realizar tantas simulaciones numéricas como zonas protegidas existan. Los resultados pueden utilizarse para el cálculo de las concentraciones en dichas zonas,

para localizaciones arbitrarias de las tuberías de descarga, así como para las reducciones de los correspondientes vertidos, mediante la utilización de una fórmula muy simple [4,5].
El problema de optimización se plantea a partir de las fórmulas obtenidas de los coeficientes de influencia. La estrategia de optimización adoptada se basa en una búsqueda aleatoria para las coordenadas de las tuberías de descarga combinadas con una técnica basada en gradiente para los coeficientes de reducción.
Se muestra que tal metodología es muy eficiente desde el punto de vista computacional y provee un adecuado marco para la comprensión del costo y beneficio resultante de la decisión técnica a adoptar.

2. FORMULACIÓN MATEMÁTICA

2.1. MODELO HIDRODINÁMICO

Para la modelización en aguas poco profundas se pueden considerar hipótesis que simplifican el problema estudiado, sin alterar la esencia del mismo [1]. Esto permite plantear el problema en un dominio bidimensional, de fácil procesamiento, aliviando la carga computacional a la hora de emplear el modelo.

Finalmente, si se considera que el régimen es estacionario, las ecuaciones de continuidad y de movimiento pueden ser reordenadas obteniendo la siguiente expresión que modela el movimiento del fluido en un cuerpo de agua poco profundo:

$$\frac{\partial}{\partial x}\left(\frac{\rho g H^2}{\alpha}\frac{\partial H}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{\rho g H^2}{\alpha}\frac{\partial H}{\partial y}\right)+\frac{\partial}{\partial x}\left(\frac{I_x H}{\alpha}\right)+\frac{\partial}{\partial y}\left(\frac{I_y H}{\alpha}\right)=\frac{\partial}{\partial x}\left(\frac{\tau_{sx} H}{\alpha}\right)+\frac{\partial}{\partial y}\left(\frac{\tau_{sy} H}{\alpha}\right), \quad (1)$$

siendo ρ la densidad, g la gravedad, H la profundidad, I_x e I_y las pendientes del fondo, τ_{sx} y τ_{sy} la tensión de corte en la superficie generadas por viento y α la fricción de fondo (cuya expresión depende de H). Las condiciones de borde para este problema son:

$$\mathrm{Q_n}=-\frac{\rho g H^2}{\alpha}\frac{\partial H}{\partial n}+\frac{\tau_n H}{\alpha}-\frac{I_n H}{\alpha}, \quad (2)$$

donde Q_n representa el flujo normal saliente al borde.

2.2. MODELO DE TRANSPORTE DE SUSTANCIAS DISUELTAS

La evolución de la concentración de una sustancia en un cuerpo de agua se plantea utilizando la ecuación de transporte de masa [2] cuya forma bidimensional es:

$$U\frac{\partial C}{\partial x}+V\frac{\partial C}{\partial y}=\frac{\partial}{\partial x}\left(K_x\frac{\partial C}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\frac{\partial C}{\partial y}\right)+T_f-T_s \quad (3)$$

En esta ecuación C*(x,y)* representa la concentración de la sustancia, U y V las velocidades promediadas verticalmente en la dirección x e y respectivamente, K_x y K_y los coeficientes de dispersión y T_f y T_s los términos de tasas de generación y degradación de sustancia introducida en el medio. Como condición de frontera se asume, considerando que los bordes están alejados de la zona de descarga, que la derivada con respecto a la normal al borde es nulo:

$$\frac{\partial C}{\partial n}=0 \quad (4)$$

2.3. PROBLEMA ADJUNTO

Asociado a la ecuación (3), se puede encontrar un problema adjunto [4] que permite obtener los coeficientes de influencia G_{ij} por cada zona de resguardo ambiental C_i, cuyas condiciones de borde son similares a las consideradas en el modelo de transporte de sustancias disueltas. Dicho cálculo permite determinar la concentración en las zonas protegidas de acuerdo a la siguiente expresión:

$$\mathrm{C_i}=\sum_{i=1}^{N}\mathrm{Q_j}\,\mathrm{G_{ij}}\,\beta_j, \quad (5)$$

siendo Q_j la tasa de emisión de la descarga j y β_j su coeficiente de reducción. Estos últimos miden el grado de purificación y varían entre 1 (sin reducción) y 0 (reducción máxima).

Es importante notar que para la obtención de los coeficientes de influencia G_{ij} solamente se necesita resolver, mediante el método de EF, tantos problemas como zonas protegidas se consideren. Como éstas son limitadas en número, tal estrategia presupone un notable ahorro de tiempo computacional.

2.4. Problema de Diseño Óptimo

Para encontrar la mejor solución al problema planteado, se define matemáticamente una Función Objetivo CT que representa los costos de construcción y operación, la cual debe minimizarse sin violar las restricciones ambientales en las áreas protegidas.

El problema consiste en encontrar las coordenadas de salida de descarga para cada fuente (x_j, y_j), como así también los factores de reducción en las tasas de vertido β_j, de tal manera de minimizar la función CT:

$$\begin{aligned} \mathrm{CT} = & \sum_{j=1}^{N} C_D\left[Q_j\left(1-\beta_j\right)\right] + C_L L_j \\ & \textit{sujeto a} \\ & 0 \leq \beta_j \leq 1 \\ & C_i \leq C_{admisible} \ , \end{aligned} \tag{6}$$

donde L_j es el largo de la tubería, C_D y C_L son los costos asociados a la descarga y al largo de tubería respectivamente, C_i es la concentración de CF en la zona protegida i-ésima y $C_{admisible}$ es la concentración máxima permitida.

2.5. Método de Solución

Para encontrar la solución óptima se desarrolló un modelo computacional de optimización, que permite automatizar la búsqueda mediante un proceso que converge a la solución óptima.

En primer lugar se evalúa las velocidades de la corriente a partir de la solución mediante EF del problema hidrodinámico (1). Con estos resultados se determinan los coeficientes de influencia resolviendo tantos problemas adjuntos asociados a (3) como zonas protegidas existan mediante EF. Luego es posible formular las restricciones mediante las fórmulas (5).

Para resolver el problema (6) se usa una búsqueda aleatoria de (x_j , y_j). Para cada conjunto de prueba de tales variables se determinan los coeficientes β_j a partir de un método basado en gradiente, que permite minimizar una expresión cuadrática equivalente al primer término de la función de costo (6), considerando las correspondientes restricciones. Con estos valores de las variables de diseño se calcula la función objetivo CT. Se repite el proceso en forma iterativa buscando el valor mínimo de CT. Con los valores óptimos y haciendo uso de las velocidades hidrodinámicas se resuelve el problema (3) mediante EF para visualizar los resultados. Este algoritmo fue implementado en el sistema FlexPDE que permite programar cada subproblema, correrlos en secuencia y transferir los datos entre los mismos.

3. Ejemplos Numéricos

A modo de ejemplo, se trata de un cuerpo de agua (río) sobre el cual se vierten efluentes cloacales, previamente depurados de manera parcial en tres plantas de tratamientos, a través de emisarios submarinos, encontrándose aguas debajo de dichas descargas dos zonas de resguardo ambiental.

En la figura 1a se muestra la distribución de velocidades de un río de 20 kilómetros de largo con una profundidad media de 8 metros. Dichas velocidades son transferidas al modelo de transporte de contaminante para obtener los valores de las concentraciones en las dos zonas protegidas (indicadas con cruces), provenientes de las tres descargas de las plantas de tratamiento (círculos amarillos), ubicadas en lugares arbitrarios. Como se puede apreciar en la figura 1b, dichas concentraciones superan el valor máximo permitido de CF (1000 NMP/100ml).

De acuerdo a este escenario es que se demuestra la necesidad de desarrollar un modelo computacional para la localización óptima de las tuberías de descarga de las plantas de tratamiento, así como la determinación del grado de purificación adecuado.

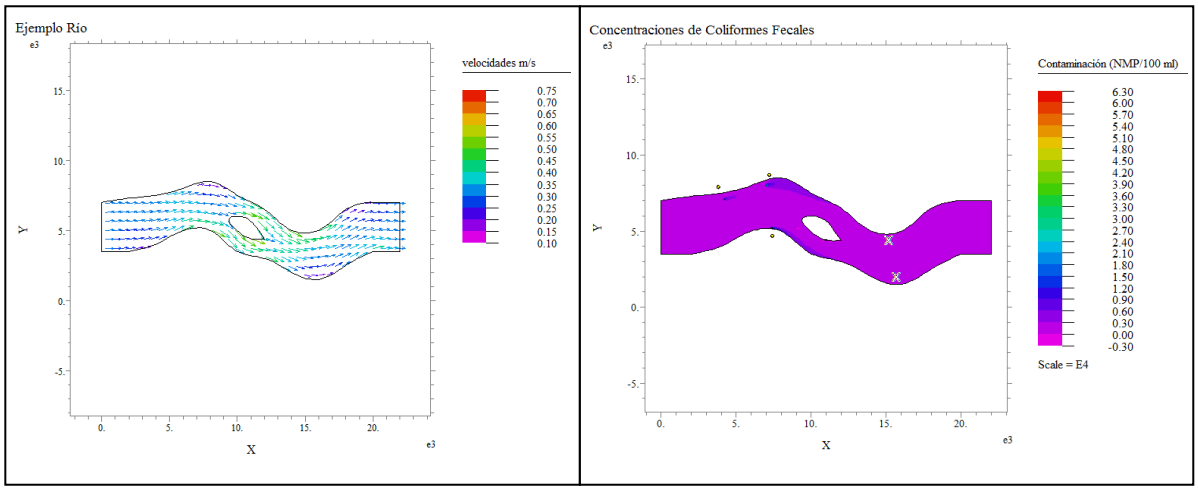


Figura 1a y 1b

En la figura 2 se indican las coordenadas de localización óptima de las descargas de las tuberías de las tres plantas de tratamiento (X11, Y11, X22, Y22, X33, Y33), los largos de los emisarios submarino (L1, L2, L3), los porcentajes de reducción (B11, B22, B33), los niveles de concentración en las dos zonas protegidas (Cp1, Cp2) y los costos constructivos y de operación asociados a estos indicadores (CT1), siendo los primeros calculados en función de la longitud de la tubería submarina y los últimos en función del porcentaje de depuración de CF.

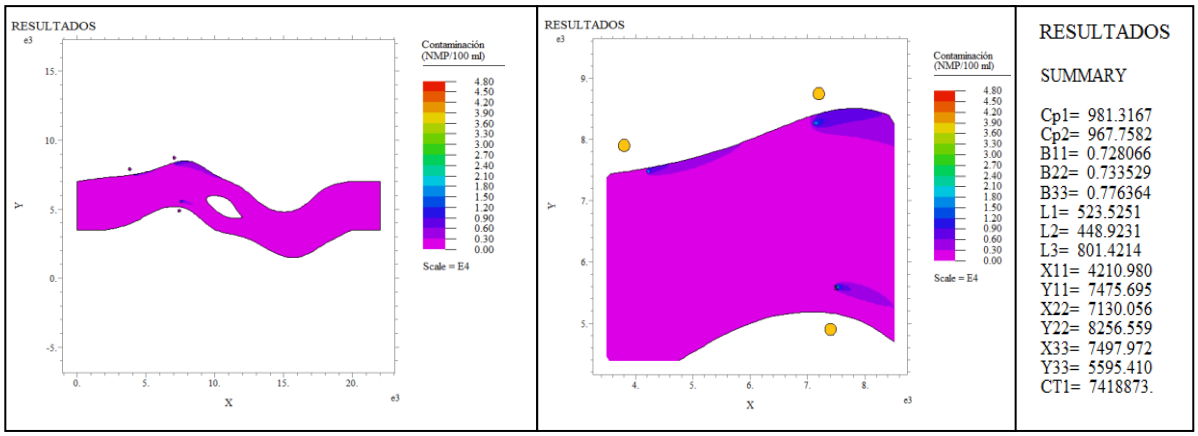


Figura 2

AGRADECIMIENTOS

Los autores desean agradecer a la Universidad Tecnológica Nacional y a la ANPCyT.

REFERENCIAS

[1] O. C. ZIENKIEWICZ. *El método de los elementos finitos*. Ed. Reverté, 1980.

[2] N.D. KATOPODES, M. PIASECKI. *Site and size optimization of contaminant sources in surface water systems.* Journal of environmental engineering. Vol. 122, No.10 (1996), pp. 917-923.

[3] FLEXPDE. *Manual de FlexPDE, version 6.08,* Copyright © 2009 PDE Solutions Inc.

[4] Y. SKIBA. *On a method of detecting the industrial plants which violate prescribed emission rates.* Ecological Modelling, Vol. 159 (2003), pp. 125-132.

[5] L. J. ALVARES-VÁZQUEZ, A. MARTINEZ, C. RODRÍGUEZ AND M.E. VÁZQUEZ-MÉNDEZ, *Mathematical model for optimal control in wastewater discharges: the global performance.* Comptes rendus, Vol. 328 (2005), pp. 327-336.

NUMERICAL ANALYSIS OF THE DRIVETRAIN BEHAVIOR OF A LARGE HORIZONTAL–AXIS WIND TURBINE

Cristian Gebhardt[1], Sergio Preidikman[1,2] y Julio Massa[1,2]

[1]*Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, cgebhardt@efn.uncor.edu, http://www.efn.uncor.edu*

[2]*Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Campus Universitario, Ruta Nacional 36 Km. 601, 5800 Río Cuarto, Argentina. Tel/Fax: 0358-4676246, spreidikman@ing.unrc.edu.ar, http://www.ing.unrc.edu.ar*

Abstract: An aeroelastic model that captures the interaction between aerodynamics and drivetrain dynamics is used to study the behavior of a large horizontal-axis wind turbine in the starting initial regime, considering different laws of brake releasing. An aerodynamical model based on the general three dimensional version of the non–linear and unsteady vortex-lattice method is used to compute the aerodynamic loads and their evolution in time domain. Moreover, the vorticity distribution in, and the position of, the wakes are obtained as part of the solution considering multiple interactions among blades, wakes, hub, nacelle, supporting tower, ground and land-surface boundary layer. In addition a model for the drivetrain is developed by considering the flexibility of the high speed shaft which connects the gear box and the generator. For the inter-model combination, a strong interaction numerical scheme was used. This scheme is based on the fourth order Hamming's predictor-corrector method which allows to compute directly the solution in the time domain.

Key words: drivetrain performance, large horizontal–axis wind turbines, unsteady and non-linear aerodynamics, aeroelasticity

1. INTRODUCTION

The models and the interaction scheme were implemented in a computational tool; and by using it, the behavior of the wind turbine in the starting initial regime is investigated considering different laws of brake releasing. The capability to simulate these phenomena is one of the novel aspects of the present effort.

2. THE AERODYNAMICAL MODEL

Consider a 3D incompressible flow of an inviscid fluid generated due to the unsteady motion of the rotor blades. The absolute velocity of a fluid particle which occupies the position $\mathbf{R}$ at instant t is denoted by $\mathbf{V}(\mathbf{R},t)$. Since the flow is irrotational outside the boundary layers and the wakes, the velocity field can be expressed as the gradient of a total velocity potential $\Phi(\mathbf{R},t)$ as

$$\mathbf{V}(\mathbf{R},t) = \nabla\Phi(\mathbf{R},t) \tag{1}$$

The spatial/temporal evolution of $\Phi(\mathbf{R},t)$ is governed by the continuity equation for incompressible flows:

$$\nabla^2\Phi(\mathbf{R},t) = 0 \tag{2}$$

A set of boundary conditions (BCs) must be added. The location of the body's surface is known as a function of time, and the normal component of the fluid velocity is prescribed on this boundary. The first BC requires the normal component of the velocity of the fluid relative to the body to be zero at the boundaries of the body, this is commonly called the '*non–penetration or impermeability*' BC (on the surface of the solid surface):

$$(\mathbf{V}-\mathbf{V}_S)\cdot\mathbf{n} = (\nabla\Phi-\mathbf{V}_S)\cdot\mathbf{n} = 0 \tag{3}$$

where $\mathbf{V}_S$ is the velocity of the boundary surface S, and $\mathbf{n}$ is the unit normal vector. In general, they can vary in space and time. A regularity condition at infinity must also be imposed. This second BC requires that the flow disturbance, due to the motion of the body (or bodies) through the fluid, should diminish far from the body. This is usually called the regularity condition at infinity and is given by

$$\lim_{\|\mathbf{R}\|_2\to\infty}\|\mathbf{V}(\mathbf{R},t)\|_2 = \lim_{\|\mathbf{R}\|_2\to\infty}\|\nabla\Phi(\mathbf{R},t)\|_2 = \|\mathbf{V}_\infty\|_2 \tag{4}$$

where $\mathbf{V}_\infty$, is the non–perturbed free stream velocity and $\|\ \|_2$ denotes the vector 2–norm.

Since the disturbance velocity field is computed according to the Biot–Savart law, the regularity condition at infinity is satisfied identically. For incompressible potential flows, the velocity field is determined from the continuity equation, and hence, it may be established independently of the pressure. Once the velocity field is known, the pressure is calculated from the unsteady Bernoulli equation. Moreover, since the speed of sound is assumed to be infinite, the influence of the BCs is immediately radiated across the whole fluid region; therefore, the instantaneous velocity field is obtained from the instantaneous BCs. In addition to the BCs, the Kelvin–Helmholtz theorems and the unsteady Kutta condition are used to determine the strength and position of the wakes.

The integral representation of the velocity field $\mathbf{V}(\mathbf{R},t)$ in terms of the vortex field $\boldsymbol{\Omega} = \nabla \times \mathbf{V}$, is an extension of the well–known Biot–Savart law. For 3D flows, it takes the following form:

$$\mathbf{V}(\mathbf{R},t) = \frac{1}{4\pi} \iint_{S(\mathbf{R}_0,t)} \frac{\boldsymbol{\Omega}(\mathbf{R}_0,t)\times(\mathbf{R}-\mathbf{R}_0)}{\|\mathbf{R}-\mathbf{R}_0\|_2^2} \, dS(\mathbf{R}_0,t) \tag{5}$$

where $\mathbf{R}_0$, is a position vector on the compact region $S(\mathbf{R}_0,t)$ of the fluid domain. The integrand in the surface integral (5) is zero wherever $\boldsymbol{\Omega}(\mathbf{R},t)$ vanishes. Thus, the region where the flow is irrotational does not contribute to $\mathbf{V}$ anywhere. $\mathbf{V}$ can be evaluated explicitly at each point, i.e., independently of the evaluation of $\mathbf{V}$ at neighboring points. As a consequence of this feature, which is absent in finite difference methods, the evaluation of $\mathbf{V}$ can be confined to the viscous region; the vorticity distribution in the viscous region determines the flow field in both, the viscous and inviscid regions.

In order to formulate the non–penetration BC given by Equation (3) it is convenient to divide the total velocity potential $\Phi(\mathbf{R},t)$ into three parts, the first one due to the bound–vortex sheet Φ_B, the second one due to the free–vortex sheet Φ_W and the last one due to the free stream Φ_∞. Hence, Eqn. (3) can be rewritten as:

$$\left(\nabla\Phi_B + \nabla\Phi_W + \nabla\Phi_\infty - \mathbf{V}_S\right) \cdot \mathbf{n} = 0 \tag{6}$$

The proposed model allows determining the aerodynamic loads acting on the LHAWT. The non-linear, unsteady and fully 3D model is based on a very well-known technique: the Unsteady Vortex-Lattice Method (UVLM). This method is a generalization of the familiar 'vortex-lattice method', which is widely used to compute steady and incompressible flows. This technique considers aerodynamic non-linearities associated to high angles of attack, static deformations, vorticity dominated flows, and unsteady behavior.

Using the UVLM it is possible to estimate, in the time domain, the aerodynamic loads acting on each blade, the vorticity distribution associated to the vortex sheets bounded to the blades, and the vorticity distribution and shape of the wakes initiated in the trailing edges and tips of each one of the blades. It is also possible to take into account all the aerodynamic interactions among the LHAWT's components [1,2,3]. A visualization of a functioning LHAWT, and the space evolution of the wakes, is presented in Figure 1.

Because the vorticity in the wakes at a given time was generated on, and shed from, the blades at an earlier time, the flow field is history-dependent and therefore, the current distribution of vorticity on the surface of the turbine depends to some extent on the previous distributions of vorticity. The vorticity distribution in, and the shape of, the wakes are determined as part of the solution; hence, the history of the motion is stored in the wakes. We say that the wakes are the 'historians' of the flow. As time passes and the vorticity in the wakes convect far downstream, its associated velocity field does not have any appreciable influence on the flow around the blades; thus, the historians have a fading memory. In the numerical method, this means that only the wakes near to the blades are important; the rest can be safely neglected. The input of the aerodynamic model is the free stream, which can vary in time and space. These data can be synthetic or obtained from experiments.

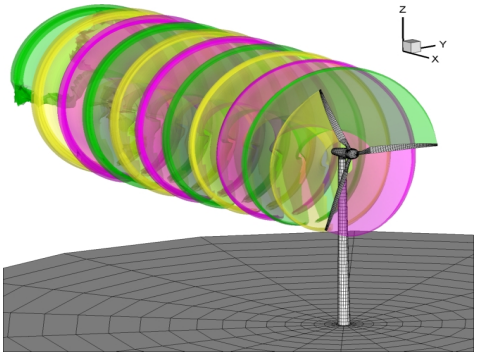


Figure 1: Evolution of the wakes emanating from a LHAWT

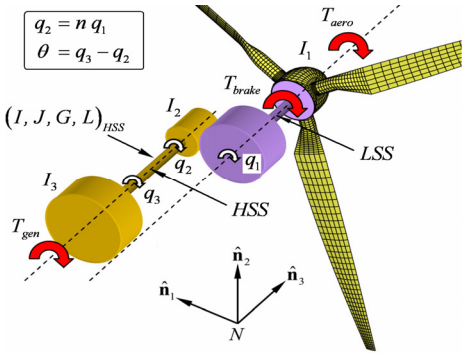


Figure 2: Reduced dynamic representation of the drivetrain

3. THE DRIVETRAIN MODEL

The aerodynamic torque (T_{aero}) on the drivetrain of a LHAWT (Figure 2) varies continuously over the time due to the unsteady and non-linear characteristics of the complex aerodynamics. The variations are directly transferred to the dynamic mechanical transmission system. The electrical generator rotates at a relatively high angular speed ($\dot{q}_3$) compared to the rotor ($\dot{q}_1$). In the drivetrain, a low speed shaft (LSS) in the rotor side is connected to a high speed shaft (HSS) in the electrical generator side by using a gearbox. Actual wind turbines are continuous systems

with an infinitely number of degrees of freedom (DOF), but a detailed investigation with a reduced number of DOF is, as a rule, entirely sufficient for analyzing its dynamic behavior. Consequently, it is sufficient to look at a mathematical model that reflects the relevant features of the actual system as accurately as possible. Any restrictions on the relative motion (linkages) between the bodies are modelled with joints with specific properties. Such mechanical systems are described mathematically by coupled ordinary differential and algebraic equations.

In general a simulation model must satisfy the following requirements: *i*) the model must represent the conditions of the actual system as accurately as possible. *ii*) the connection between the actual system and the reduced model should be noticeable at each point. *iii*) it should be possible to obtain the system parameters, on which the model is based, from the technical documents or from the actual system with sufficient accuracy.

A reduced 2-DOF model with simple inertia representation of the drivetrain and containing the gear box is shown in Figure 2. In this model the rotor, gears and LSS are considered rigid bodies while HSS is flexible, and the resulting equations of motion become :

$$\left[\mathbf{M}\right]\{\ddot{\mathbf{q}}\}+\left[\dot{\mathbf{M}}+\mathbf{D}\right]\{\dot{\mathbf{q}}\}+\left[\mathbf{K}\right]\{\mathbf{q}\}=\{\mathbf{T}\} \tag{7}$$

where $\mathbf{M}$ and $\dot{\mathbf{M}}$ are the mass matrix and its first time derivative, $\mathbf{D}$ and $\mathbf{K}$ are the damping and stiffness matrices, respectively. The generalized coordinate vector is $\mathbf{q} = (q_1, q_3)$; q_1 represents the azimuth angle of the rotor, q_2 is related to q_1 by the transmission relation (n) of the gear box and q_3 represents the angular position of the generator (the torsional deformation θ of the HSS is given by the difference between q_3 and q_2), and $\mathbf{T}$ is the generalized torques vector which takes in account the contribution coming from: the aerodynamics, the control systems, and the brake.

4. COMBINING THE MODELS

In the current investigation, the aerodynamics and dynamics are treated as the elements of a single dynamical system. All the governing equations are solved simultaneously and interactively in the time domain. The methodology is based on a fourth order predictor-corrector method developed by Hamming [4]. In the late nineties this method was adapted and expanded to solve fluid-structure interaction problems [5]. An iterative scheme was developed to account for the interaction among the aerodynamic loads, drivetrain dynamics and control systems. The interaction scheme proposed, based on the above mentioned model, to solve all the governing equations in the time domain is presented in Figure 3.

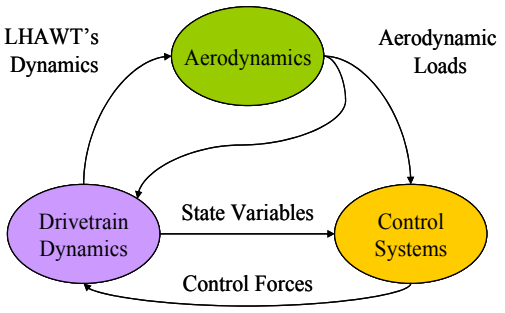


Figure 3: Combining the models

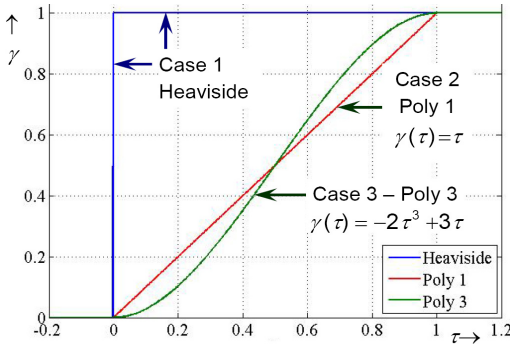


Figure 4: Laws of brake releasing

5. RESULTS AND DISCUSSION

The analysis was carried out for a standard 3 blades LHAWT with a rotor diameter of 78 m (tower height 80 m) considering the following wind conditions: wind speed at 10 m of altitude 10 m/s (reference for the land-surface boundary layer); wind speed at the hub 12,8 m/s, flat terrain with very low building The total number of elements of the mesh for the aerodynamical model are 3952, including 456 for each blade, 864 for the hub, 496 for the nacelle, 882 for the tower and 342 for the ground.

In the present effort, the case of study is focused on the impact of the non-linear and unsteady aerodynamic loads over the drivetrain at starting initial regime. In order to reach this target, three laws of brake releasing were proposed and further investigated ($T_{brake}(t)$ in Figure 2). The first one contains a step or Heaviside releasing law, the second is modelled by a first order polynomial $\gamma(\tau)=\tau$, and the third is modelled by using a third order polynomial law $\gamma(\tau) = -2\tau^3 + 3\tau$, where $\tau = (t/t_{rel})$ and t_{rel} is the reference time at which the brake is completely released (see Figure 4).

5.1. COMPARISONS FOR THE THREE LAWS OF BRAKE RELEASING

The results obtained from the computational tool are discussed for each of the braking law used in the present study. In order to make a consistent comparison among the various case studied, the results ($\theta, \dot{\theta}$) are plotted as a function of the rotor azimuth angle q_1. A comparison study had been carried out for the three cases considered in the present investigation. In Figure 5 the torsional deformation angle (θ) of the HSS for the three cases studied are plotted as a function of rotor azimuth angle q_1. It can be observed that the laws of brake releasing do not have a significant effect on the angular velocity of the rotor.

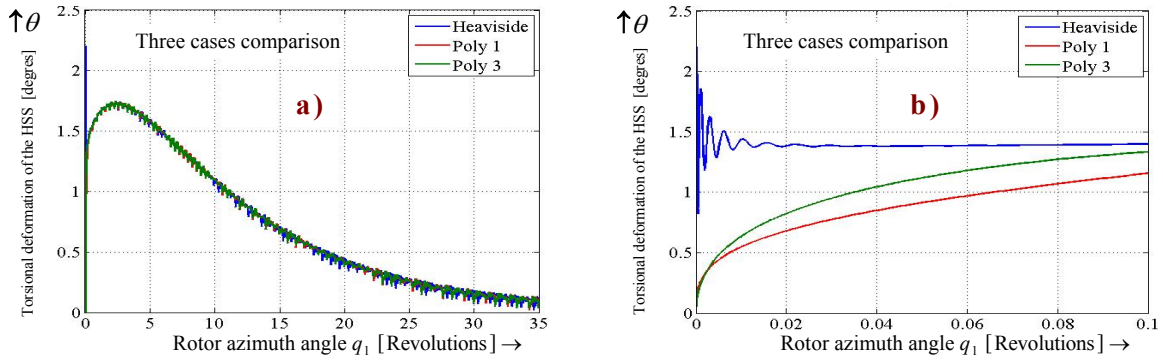


Figure 5: a) Torsion angle θ of the HSS for the all cases. b) Zoom for time $t < 0{,}1$

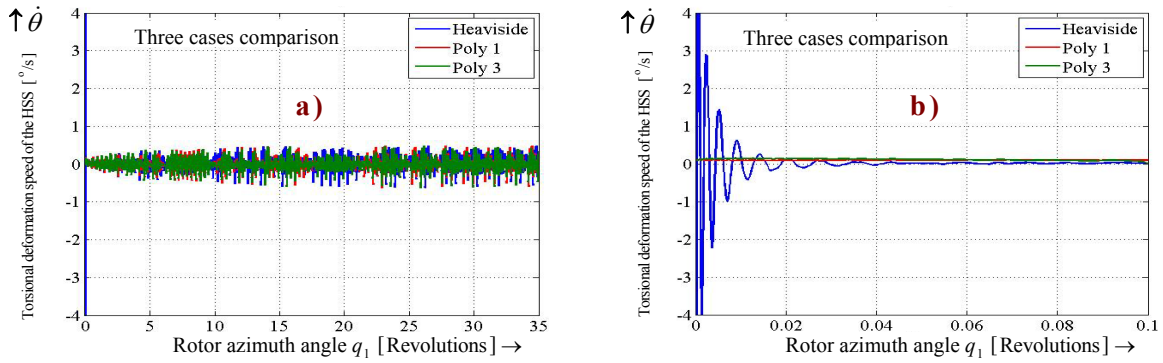


Figure 6: a) Torsional deformation speed $\dot{\theta}$ of the HSS for the all cases. b) Zoom for time $t < 0{,}1$

In Figure 6 the torsional deformation speed of the HSS ($\dot{\theta}$) is plotted as a function of rotor azimuth angle q_1. A shock response, as an impact load on the HSS, can be observed in the initial stage of releasing. Later on, the fluctuations decrease very rapidly. These initial fluctuations can have a considerable impact on fatigue life but this topic is not the subject of the present study. This phenomenon demands further investigation.

CONCLUSIONS

In this present work, a novel methodology has been developed to study the dynamic effect of the aerodynamic loads on the drivetrain of a LHAWT. In this event some concluding remarks can be drawn: *i*) the angular speed of the HSS shows the same trend irrespective of the used law of brake releasing; *ii*) the chosen law of brake releasing shows a considerable affect on the torsion angle of the HSS at the initial regime of the LHAWT start up; and *iii*) the transfer of transient fluctuations on the system behavior from the rotor to the drivetrain system is considerable but this effect does not occur in the other direction.

REFERENCES

[1] C.G. GEBHARDT, S. PREIDIKMAN, J.C. MASSA AND A. DELLA BARCA, *Interacciones aerodinámicas no-lineales e inestacionarias en turbinas eólicas de eje horizontal y de gran potencia*. Mecánica Computacional 28 (2009), pp. 1489-1505.

[2] S. PREIDIKMAN, C.G. GEBHARDT; A.T. BREWER AND B.A. ROCCIA, *Aeroservoelastic analysis of large horizontal-axis wind turbines: A new methodology*. Proceedings of the 11th Pan-American Congress of Applied Mechanics, Foz do Iguaçu, 2010.

[3] C.G. GEBHARDT, S. PREIDIKMAN AND J.C. MASSA, *Numerical simulations of the aerodynamic behaviour of large horizontal-axis wind turbines*. International Journal of Hydrogen Energy (2010); doi:10.1016/j.ijhydene.2009.12.089

[4] B. CARNAHAN, H.A LUTHER AND J.O. WILKES, *Applied numerical methods*. John Wiley & Sons, 1969.

[5] S. PREIDIKMAN, *Numerical simulations of interactions among aerodynamics, structural dynamics, and control systems*. Ph.D. Thesis, Virginia Polytechnic Institute and State University, 1998.

DISEÑO ACÚSTICO ÓPTIMO DE RECINTOS INDUSTRIALES MEDIANTE EL USO DE UN META-MODELO

Martín E. Sequeira† y Víctor H. Cortínez†‡

†Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca, 11 de Abril 461, 8000, Bahía Blanca, Argentina, martins@frbb.utn.edu.ar
‡Consejo Nacional de Investigaciones Científicas y Tecnológicas, Argentina, vcortine@frbb.utn.edu.ar

Resumen: El control acústico es una disciplina de interés creciente en el ambiente laboral industrial. A los efectos de diseñar soluciones acústicas, es necesario efectuar predicciones de los niveles sonoros que tendrán lugar una vez implementadas las mismas. Un nueva técnica para abordar la acústica de recintos, basada en un modelo de difusión, fue propuesta hace unos años [J. Picaut *et al*., Acustica 83 (1997)]. En tal sentido, en el presente artículo se propone un enfoque que combina la estructura teórica del modelo acústico clásico con el modelo de difusión, para estimar las variaciones del campo sonoro en recintos industriales considerando diversas características acústicas. Finalmente, el modelo resultante, se combina con la técnica de simulated annealing para desarrollar un diseño acústico óptimo.

Palabras claves: *ruido industrial, diseño óptimo, modelo de difusión*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

Para realizar el diseño acústico en recintos industriales, es posible utilizar simples formulaciones analíticas que describen esencialmente relaciones entre los niveles sonoros en el ambiente específico y las potencias de las fuentes generadoras, a través de un coeficiente de influencia. La utilización de dichos modelos implica, muchas veces, una reducción de la precisión en función de la elección de las variables involucradas en la determinación de dicho coeficiente, relacionado directamente con la complejidad de la situación considerada. A este coeficiente se lo conoce como nivel de propagación sonora (*SP*) y describe la manera en que las características geométricas y acústicas del recinto afectan la variación del nivel de presión sonora, en función de la distancia desde la fuente, independientemente del nivel de potencia sonora de la misma.

El modelo de difusión acústica, propuesto por Picaut *et al*. [1], permite calcular la distribución del campo reverberante no uniforme en recintos, extendiendo la teoría de campo difuso. Las ecuaciones derivadas pueden ser fácil y rápidamente resueltas mediante el método de los elementos finitos. Dicho modelo se ha aplicado con aceptable precisión, para predecir los niveles sonoros en diversos ambientes acústicos interiores [2,3,4]. Recientemente, se han realizado diferentes comparaciones entre este modelo y la técnica de trazado de rayos considerando recintos con geometrías complejas, mostrando buenos resultados [5].

En este trabajo, se emplea el modelo de difusión acústica para obtener los valores de *SP* considerando distintas configuraciones extremas de absorción en las superficies del recinto. Luego, los valores de *SP* para las condiciones intermedias de absorción se determinan a partir de una interpolación lineal. Dichos valores se incorporan dentro de un modelo analítico clásico para predecir las variaciones del campo sonoro contemplando todas las configuraciones de absorción posibles. El meta-modelo resultante, se utiliza entonces, para realizar un enfoque de diseño óptimo consistente en definir el tratamiento acústico de menor costo económico en un recinto industrial multi-fuente, procurando que los niveles sonoros globales no superen cierto valor límite establecido. Se propone utilizar el método de *simulated annealing* conjuntamente con el meta-modelo aludido para implementar la optimización del diseño acústico.

2. FORMULACIÓN DEL PROBLEMA DE DISEÑO

El nivel de presión sonora está relacionado con las potencias suministradas por las fuentes sonoras mediante el nivel de propagación sonora *SP*, el cuál depende, principalmente, de las dimensiones del recinto, de las coordenadas de la fuente sonora y de los coeficientes de absorción α de cada superficie interior. En este trabajo, los valores de *SP* se determinan a partir del modelo de difusión acústica. Para ello, se combinan las distintas configuraciones entre los valores de coeficientes de absorción mínimos (αmin) en todas las superficies con los coeficientes de absorción máximos (αmax) adoptados para cada una de ellas. Los valores

de *SP* para los α intermedios en cada superficie se obtiene a partir de una interpolación lineal entre los valores extremos previamente obtenidos. De esta manera, el nivel *SP*, en decibeles, queda establecido como:

$$SP_{j\alpha_f}(\mathbf{r}) = SP_{j\alpha\min_f}(\mathbf{r}) + \sum_{i=1}^{Ns} \frac{(SP_{j\alpha\min_{if}}(\mathbf{r}) - SP_{j\alpha\max_{if}}(\mathbf{r}))}{\Delta\alpha_{if}} (\alpha_{if} - \alpha\min_{if}) , \quad (1)$$

donde $SP_{j\alpha_f}(\mathbf{r})$ es el nivel de propagación sonora en cualquier punto del recinto **r** debido a la fuente *j* para la banda de octava *f* considerando la contribución de cada coeficiente de absorción α_{if} en cada superficie *i*, siendo *Ns* el número total de superficies. A partir de esto, el nivel de presión sonora, en decibeles, se expresa:

$$Lp_{j\alpha_f}(\mathbf{r}) = (Lw_{jf} - PT_{jf}) + SP_{j\alpha_f}(\mathbf{r}) , \quad (2)$$

donde $Lp_{j\alpha f}(\mathbf{r})$ es el nivel de presión sonora en cada punto receptor del recinto bajo estudio, a causa de la fuente *j* para la banda *f*, Lw_{jf} es el nivel de potencia sonora de cada fuente *j* para la banda *f* y PT_{jf} es la pérdida de transmisión sonora generada por un posible encapsulamiento sobre la fuente *j* para la banda *f*.

A los efectos de implementar el correspondiente enfoque de optimización, se consideran como variables de diseño los valores de pérdida de transmisión (*PT*), que podrán variar en función de tipo de encapsulamiento adoptado para cada fuente y los coeficientes de absorción (α_i), que dependerán del tipo de material absorbente utilizado en cada superficie i. Todos los parámetros restantes se presumen conocidos.

En consecuencia y asumiendo que el costo de cada material absorbente es proporcional al área de cada superficie que se propone tratar y que el costo de cada encapsulamiento depende del grado de pérdida de transmisión que posee cada cerramiento, el problema de optimización se formula de la siguiente manera:

$$\begin{aligned} &min \quad C = \sum_{i=1}^{Ns} C_i S_i + \sum_{j=1}^{M} \overline{C}_j \\ &s.a.: \quad Lp_{total}(\mathbf{r}) = 10\log_{10}\left[\sum_{f=1}^{Nf}\sum_{j=1}^{M} 10^{\frac{Lp_{jf}(\mathbf{r})}{10}}\right] \leq 85 \text{ decibeles} , \end{aligned} \quad (3)$$

donde *C* es la función objetivo (costo de instalación), S_i es el área de cada superficie *i*, *M* es el cantidad de fuentes sonoras *j* plausibles de acondicionar acústicamente mediante un encapsulamiento, C_i se define como el costo por unidad de superficie, en función del tipo de material absorbente utilizado, para tratar la superficie *i* y $\overline{C}_j$ se define como el costo por unidad, según la pérdida de transmisión de cada tipo de cerramiento, para encapsular la máquina *j*. El nivel de presión sonora total Lp_{total} (**r**), se obtiene considerando, en cada punto receptor, la contribución de cada fuente sonora *j* en cada banda de octava *f*, donde *Nf* es el número total de bandas consideradas.

3. MODELO DE DIFUSIÓN ACÚSTICA

Este modelo describe la distribución del campo reverberante no uniforme en recintos a partir de una analogía matemática entre la propagación del sonido en recintos con superficies reflectantes difusivas y la difusión de partículas de un medio gaseoso en un fluido difusivo. A partir de esto, es posible obtener, para una posición **r** = (x, y, z) dentro de un recinto con volumen *V*, la densidad de energía sonora estacionaria $u(\mathbf{r})$, correspondiente al campo reverberante, como la solución del siguiente sistema de ecuaciones [2,4]:

$$D\ \nabla^2\ u(\mathbf{r}) - \sigma u(\mathbf{r}) + w(\mathbf{r}) = 0, \quad (4)$$

$$D\frac{\partial u(\mathbf{r})}{\partial n} + A_X c u(\mathbf{r}) = 0. \quad (5)$$

La ecuación (4) se aplica en el volumen del recinto donde ∇^2 es el operador laplaciano, *D* es el coeficiente de difusión acústica, σ es un término de absorción volumétrica y *w* representa la potencia sonora por unidad de volumen generada por la ó las fuentes acústicas consideradas. En el caso de recintos vacíos, $D = \lambda c/3$ donde se considera la morfología del recinto, de superficie interior *S*, a través de la expresión clásica

del camino libre medio $\lambda = 4V/S$ y $\sigma = mc$, siendo m el coeficiente de absorción del aire y c la velocidad del sonido. La ecuación (5) corresponde a las condiciones de borde sobre las superficies interiores del recinto, siendo A_X el factor de absorción el cual puede adopta diferentes expresiones a los fines de modelar cada superficie con el rango completo de posibles coeficientes de absorción [2,4].

Finalmente, el nivel de propagación sonora en cada punto receptor $SP_{jf}(\mathbf{r})$ para cada fuente j en la banda f, se obtiene a partir de la solución numérica estacionaria de $u(\mathbf{r})$ y queda establecido como:

$$SP_{j\alpha f}(\mathbf{r}) = 10\log_{10}\left\{\rho c\left[\int w_{jf}/(4\pi r^2)\,\mathrm{dV_S} + cu_{jf}(\mathbf{r})\right]/P_{ref}^2\right\}, \tag{6}$$

donde r expresa la distancia entre la fuente j y el receptor $\|\mathbf{r} - \mathbf{r}_{sj}\|$, siendo $\mathbf{r}_{sj}$ la posición de la fuente sonora j, ρ la densidad del aire, $P_{ref} = 2 \times 10^{-5}$ Pa y $w_{jf} = W_0\delta(\mathbf{r} - \mathbf{r}_{sj})$, donde $W_0 = 10^{-12}$ vatios.

4. MÉTODO DE OPTIMIZACIÓN: SIMULATED ANNEALING

Es una técnica heurística de optimización combinatoria basada en una generación aleatoria de soluciones factibles cuya principal característica es la de evitar convergencia local en problemas de gran escala. La función que determina y controla el descenso de la temperatura (T), juega un rol fundamental en la eficiencia del método. En este trabajo, a diferencia del esquema geométrico habitualmente considerado, se utiliza una nueva estrategia [6] para disminuir la temperatura cuyo principal objetivo es logra reducir la temperatura más rápidamente al inicio del algoritmo, evitando aceptar en un comienzo la mayoría de las soluciones factibles y en consecuencia reducir el elevado costo inicial. El esquema general del algoritmo ha sido presentado por los autores en otro trabajo [7].

5. RESULTADOS NUMÉRICOS

El enfoque de optimización se implementó en el entorno Matlab®. Previamente, se utilizó el software comercial Flex-PDE® para resolver las ecuaciones correspondientes al modelo de difusión acústica a partir del método de elementos finitos. Se adoptó, como opción de diseño, no implementar tratamiento acústico ó utilizar tres tipos diferentes de calidades de material absorbente comerciales sobre el cielorraso y las paredes interiores. Los coeficientes de absorción utilizados son: 0.07, 0.08, 0.08 y 0.09 para superficies sin tratar, 0.15, 0.35, 0.45 y 0.5 para superficies con absorción baja, 0.25, 0.41, 0.5 y 0.55 para superficies con absorción media y 0.35, 0.6, 0.7 y 0.75 para superficies con absorción alta, para las bandas de octava de 250, 500, 1000 y 2000, respectivamente. Por otro lado, se consideraron dos tipos de calidades de encapsulamiento sobre las fuentes, determinados a partir de los valores de pérdida de transmisión sonora (PT) de cerramientos comúnmente utilizados en estos casos. En este caso se eligió también, como opción de diseño, no implementar el encapsulamiento de la fuente. Los valores de PT utilizados son: 0, 0, 0 y 0, para fuentes sin encapsular, 15,20, 24 y 29, para cerramientos con baja PT y 22,30, 34 y 35, para cerramientos con alta PT, para las bandas de octava de 250, 500, 1000 y 2000, respectivamente. El costo económico de cada tipo de tratamiento acústico (paneles absorbentes ó cerramientos) se estimó con relación al correspondiente a 1 m^2 de panel con baja absorción.

Se consideró un recinto de 4 m de altura, con una variación acentuada en su geometría, donde se dispusieron 5 fuentes sonoras puntuales omnidireccionales, con una altura de 1 m cada una, cuyos niveles de potencia sonora son: 90, 91, 90 y 87 para la fuente 1 (S1), 86, 89, 92 y 92 para la fuente 2 (S2), 83, 86, 89 y 89 para la fuente 3 (S3) y fuente 4 (S4) y 92, 89, 85 y 80 para la fuente 5 (S5), para las bandas de octava de 250, 500, 1000 y 2000, respectivamente. Se adoptaron 31 receptores, ubicados a una altura de 1.5 m, a los efectos de implementar las restricciones durante el proceso de optimización. En la Figura 1 se muestra el esquema del recinto junto a la ubicación de las fuentes sonoras y los receptores mencionados. Se eligió el cielorraso y 5 paredes como superficies factibles a tratar acústicamente (ver Figura 1).

Los resultados obtenidos, en función del tipo de tratamiento acústico sobre cada fuente sonora, fueron los siguientes: PT media para la fuente 1 y sin encapsulamiento para el resto de las fuentes. En el caso de las superficies, el tratamiento acústico resultante consistió en: absorción baja sobre el cielorraso y las superficies 1 y 4 y sin tratamiento sobre las superficies 2, 3 y 5. En la Figura 2, se muestra la evolución de la función objetivo frente al número de iteraciones. Se aprecia que alrededor de las 750 iteraciones se alcanza la solución óptima, con un valor del costo de instalación ($C = 7060$) muy cercano al mínimo.

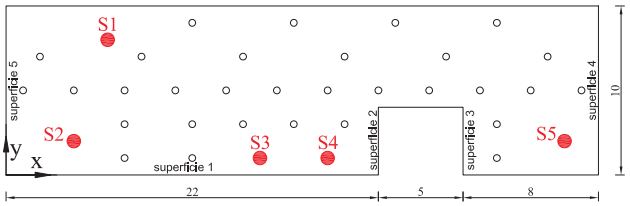


Figura 1: Esquema del recinto con la ubicación de las fuentes y puntos receptores (O)

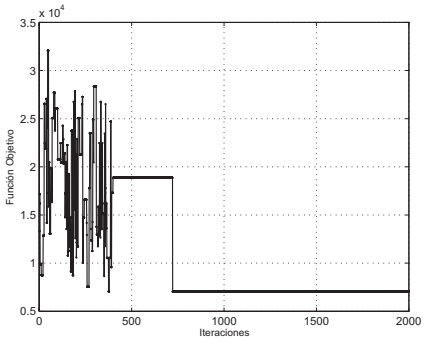


Figura 2: Evolución de la función objetivo durante el proceso de optimización

6. CONCLUSIONES

Se formuló un enfoque de optimización para realizar el diseño acústico en recintos industriales multi-fuente, procurando minimizar los costos de implementación de soluciones técnicas. La utilización del meta-modelo acústico presentado, permitió predecir con adecuada eficiencia el nivel *SP* (y por consiguiente la distribución de los niveles sonoros) con errores máximos promedios del orden de los 2.5 dB. El diseño propuesto, fue resuelto satisfactoriamente mediante la aplicación de la técnica de optimización simulated annealing modificada, la cual mejora la velocidad de convergencia con respecto al enfoque clásico.

Por último, el tiempo de cálculo empleado durante la aplicación de este enfoque fue de aprox. 460 segundos (400 y 60 segundos para determinar los valores de *SP* y realizar la optimización, respectivamente). Esto presenta una ventaja significativa, en el contexto de diseño presentado, debido al gran número de simulación que es necesario efectuar durante dicho proceso.

AGRADECIMIENTOS

Este trabajo ha sido auspiciado por la Secretaría de Ciencia y Tecnología de la Universidad Tecnológica Nacional. Los autores agradecen a la UTN y al Dpto. de Ingeniería de la Universidad del Sur.

REFERENCIAS

[1] J. PICAUT, L. SIMON, AND J. POLACK, *A mathematical model of diffuse sound field based on a diffusion equation*, Acust. Acta Acust., 83 (1997), pp. 614–621.

[2] V. VALEAU, J. PICAUT AND M. HODGSON, *On the use of a diffusion equation for room-acoustic prediction*, J. Acoust. Soc. Am., 119 (2006), pp. 1504–1513.

[3] A. BILLON, V. VALEAU AND A. SAKOUT, *On the use of a diffusion model for acoustically coupled rooms*, J. Acoust. Soc. Am., 120 (2006), pp. 2043–2054.

[4] Y. JING, AND N. XIANG, *On boundary conditions for the diffusion equation in room-acoustic prediction: Theory, simulations and experiments*, J. Acoust. Soc. Am., 123 (2007), pp. 145–153.

[5] M. SEQUEIRA Y V. CORTÍNEZ. *Un modelo de difusión acústica para recintos: comparación con el método de rayos*, Mecánica Computacional, XXVIII (2009), pp. 163–179, ISSN: 1666-6070.

[6] M. VIDAL, *Un procedimiento heurístico parta un problema de asignación cuadrática*, Tesis de Magíster en Matemática, Dpto. de Matemática, Universidad Nacional del Sur, Bahía Blanca, Argentina, (2003).

[7] V. CORTÍNEZ Y M. SEQUEIRA, *Identificación de las condiciones acústicas en recintos industriales*, Mecánica Computacional, XXIX (2010), pp. 2155–2172, ISSN: 1666-6070.

MODELO MATEMÁTICO PARA LA EPIDEMIOLOGÍA DE LA TOXOPLASMOSIS USANDO DOS FUENTES IMPORTANTES DE TRANSMISIÓN

Carlos A. Peña-Rincón[a], Graciela Juez-Castillo[b]

[a]*Miembro del grupo de Matemáticas Aplicadas, IMA, Universidad Sergio Arboleda, Bogotá, Colombia,* *carlospena@ima.usergioarboleda.edu.co*
[b]*Miembro del Grupo de Investigación en Bioquímica y Biología Molecular de Parásitos, Universidad de los Andes, Bogotá, Colombia,* sheliusky@gmail.com

Resumen: *Toxoplasma gondii* es un parásito protozoario que en condiciones oportunistas afecta al ser humano por ejemplo en individuos inmunocomprometidos genera serios problemas de salud. La infección por *T. gondii* en los huéspedes intermediarios entre ellos el hombre se da por diferentes vías; transmisión de la madre al feto por vía placentaria debido a que durante el embarazo existe cierta probabilidad de que las infecciones agudas por *T. gondii* conlleven a enfermedades congénitas del neonato, transmisión de taquizoitos en trasplante de órganos, ingestión de bradizoitos presentes en carne poco cocida e ingestión de ooquistes presentes en agua destinada al consumo. El presente trabajo plantea un modelo matemático de simulación numérica para representar la dinámica epidemiológica de la Toxoplasmosis teniendo en cuenta como fuentes principales de transmisión, el contacto directo con gatos infectados y el consumo de carne de cerdo infectada y preparada con inadecuada cocción.

Palabras claves: *Modelo epidemiológico, Toxoplasmosis, vías de transmisión, simulación*.

1. INTRODUCCIÓN

Toxoplasma gondii (*T. gondii*) es un importante patógeno del ser humano y de otros animales vertebrados de sangre caliente que tiene como huésped definitivo a los gatos. Este parásito es de amplia importancia clínica por las enfermedades que causa en el feto, debido a que durante el embarazo existe cierta probabilidad de que las infecciones agudas por *T. gondii* conlleven a enfermedades congénitas del neonato [1]. Varios de los estudios clínicos que se han realizado demuestran que son diversas las manifestaciones clínicas que se pueden presentar en individuos que padecen toxoplasmosis congénita, entre ellas; corioretinitis, microcefalia, hidrocefalia, encefalomielitis, retardo mental, hepatoesplenomegalia, eritroblastocis [2]**.** Este patógeno también tiene la capacidad de reactivar la infección en individuos inmunocomprometidos por ejemplo en pacientes con algún tipo de cáncer [3], trasplante de órganos [2], con SIDA [3] causando serios problemas que podrían llegar a ser fatales para el individuo.

T. gondii predomina fuertemente debido a sus medios eficaces de difusión y a la capacidad de resistir al sistema inmunológico. Tiene un ciclo vital muy complejo que incluye etapas sexuales y asexuales que dependen del desarrollo biológico de este parásito [2]. El aumento en el interés sobre el estudio de la Toxoplasmosis en humanos radica principalmente en las vías de transmisión de éste parásito las cuales están determinadas por el consumo de alimentos contaminados tales como: frutas, aguas, verduras y carnes mal cocidas [2]. Algunos estudios reportan que en seres humanos una de las fuentes de mayor importancia en la transmisión de la infección por *T. gondii* es el consumo de carnes de diferentes especies infectadas con quistes del parásito [4]. En Colombia se reporta que esta fuente de infección puede ser la responsable del 25% de casos de Toxoplasmosis en mujeres en periodo de gestación [5]. El cerdo se caracteriza por ser la especie en la que se ha encontrado una mayor prevalencia de este patógeno, considerando el consumo de su carne en condiciones de mala preparación como el factor más importante de transmisión de la infección [6]. Un estudio realizado con muestras de carne de 3 especies destinadas al consumo humano provenientes de algunas ciudades de Colombia, reporta que la prevalencia de *T. gondii* en este tipo de alimento presenta un porcentaje considerable, el 52,7% del total de las muestras analizadas fueron positivas con *T. gondii* [6], sin embargo la carne de cerdo fue la especie que mayor porcentaje de detección del parásito presentó con 70% de positividad, mientras que la carne de res mostró un 48,3% y la carne de pollo un 40% [6]. Estos resultados sugieren que las tres clases de carnes representan un factor de transmisión de *T. gondii* de alto riesgo para la población humana. Estudios reportan que la alta prevalencia del parásito en el cerdo se debe a diferentes causas, entre ellas, las deficientes actividades de higiene en sacrificio y manipulación de la carne en los lugares de expendio [7] y las condiciones abiertas del sitio de crianza del animal, pues en su mayoría son criados en lugares de la zona rural donde fácilmente tienen contacto con otros animales como roedores, moscas y heces de gatos [8], siendo éste último los de mayor riesgo debido a que se considera el huésped definitivo de *T. gondii* y el cual libera en sus heces la forma infestiva del parásito que puede sobrevivir a temperaturas de 50°C por largo tiempo [3]. Teniendo en cuenta la alta prevalencia de este parásito en varias regiones de Colombia, este trabajo plantea un modelo matemático sobre la epidemiología de la Toxoplasmosis involucrando dos fuentes importantes responsables de la transmisión de la infección en la población humana de Colombia, el contacto directo con gatos infectados y el consumo de carne de cerdo infectada y preparada con inadecuada cocción.

2. MODELO MATEMÁTICO

El modelo matemático que se plantea es un complemento a modelos matemáticos presentados anteriormente [9] y [10]. El modelo de este trabajo se basa en las siguientes hipótesis:

2.1. HIPÓTESIS

1. La población total N(t) es dividida en tres subpoblaciones:
   - **S (t) Susceptibles**: Individuos de la población quienes podrían ser infectados por *T. goondii*.
   - **I (t) Infectados**: Individuos de la población infectada por el parásito, quienes presentan los síntomas clínicos y los que no presentan la sintomatología.
   - **C (t) Controlados**: Individuos de la población infectada que han presentado los síntomas clínicos y a los cuales se les ha realizado un tratamiento. Dentro de esta subpoblación se encuentran: pacientes con algún tipo de cáncer [3], trasplante de órganos [2], con SIDA [3], niños con Toxoplasmosis congénita.

2. **β, β_c, γ, μ_c :** Tasas de transmisión
3. **d:** Tasa por muerte natural
4. **μ:** Tasa de nacimiento
5. **ε:** Tasa de muerte causada por *T. gondii*
6. **p:** Probabilidad de nacimiento sin *T. gondii*
7. **p_c:** Probabilidad de nacimiento de un gato a partir de un gato infectado
8. **p_0:** Probabilidad de nacimiento de un cerdo en lugares abiertos sin control de las condiciones alimenticias.
9. **β_f:** Tasa de consumo de carne de cerdo infectada con *T. gondii*, sin la adecuada cocción.
10. **β_p:** Parámetro de Malos hábitos de consumo de carne de cerdo infectada, en donde el valor de 1 significa el 100% de transmisión de la infección.

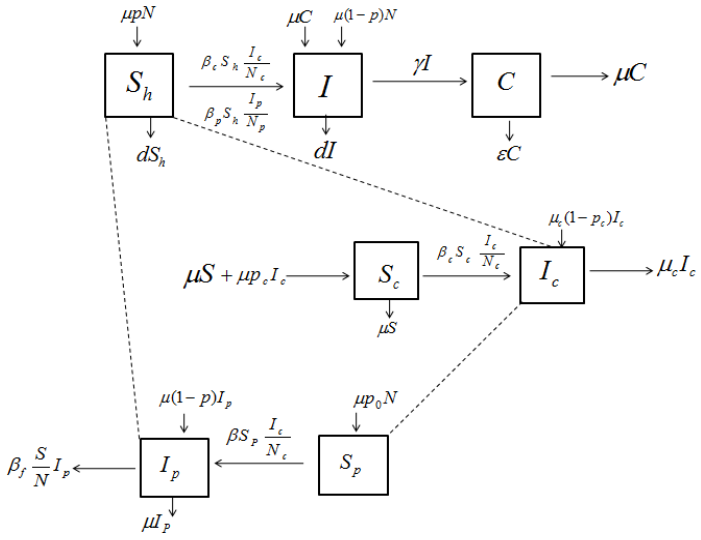


Figura 1: Modelo matemático de la epidemiología de la Toxoplasmosis

2.2. ECUACIONES DIFERENCIALES ORDINARIAS DEL MODELO EPIDEMIOLÓGICO

❖ Población Humana

$$\dot{S}(t) = p\mu N(t) - dS(t) - \beta_c S_h \frac{I_c}{N_c} - \beta_p S_h \frac{I_p}{N_p} \qquad \dot{I}(t) = \mu(1-p)N + \mu_h C_0 - dI - \gamma_h I_h + \beta_c S_h \frac{I_c}{N_c} + \beta_p S_h \frac{I_p}{N_p}$$

$$\dot{C}(t) = \gamma_h I_h - \mu_h C - \varepsilon C \qquad N(t) = S(t) + I(t) + C(t)$$

❖ Población de gatos

$$\dot{S}_c(t) = \mu p_c I_c - \beta_c S_c \frac{I_c}{N_c} \qquad \dot{I}_c(t) = \beta_c S_c \frac{I_c}{N_c} - dI_c \qquad N_c(t) = S(t) + I(t)$$

❖ Población de cerdos

$$\dot{S}_p(t) = \mu pN - \beta S_p \frac{I_c}{N_c} \qquad \dot{I}_p(t) = \beta S_p \frac{I_c}{N_c} - \beta_p S_h \frac{I_p}{N_p} \qquad N_p(t) = \dot{S}(t) + I(t)$$

3. ESCALANDO EL MODELO

$$N(t) = \mu\dot{N} - dN - \varepsilon C \qquad \frac{\dot{N}}{N} = \mu - d - \varepsilon\frac{C}{N}$$

Definimos: $s = \frac{S}{N}; \qquad i = \frac{I}{N}; \qquad c = \frac{C}{N}$

$$\boxed{\frac{\dot{N}}{N} = \mu - d - \varepsilon c} \qquad \boxed{\dot{s} = \frac{\dot{S}}{N} - s(\mu - d - \varepsilon c)} \qquad \boxed{\dot{i} = \frac{\dot{I}}{N} - \frac{I}{N}x\frac{\dot{N}}{N} = \frac{\dot{I}}{N} - i(\mu - d - \varepsilon c)}$$

$$\boxed{\dot{c} = \frac{\dot{C}}{N} - c(\mu - d - \varepsilon c)}$$

3.1. DINÁMICA DE LA POBLACIÓN HUMANA

$$\boxed{\dot{s} = p\mu - s(\mu - \varepsilon c) - \beta_c sB(t) - \beta_p sE(t)} \qquad \boxed{\dot{i} = \mu(1-p) + i(\varepsilon c - \gamma - \mu) + \mu c + \beta_c sB(t) + \beta_p sE(t)}$$

$$\boxed{\dot{c} = \gamma i - c(2\mu - d + \varepsilon + \varepsilon c)} \qquad \boxed{1 = s + i + c}$$

3.2. DINÁMICA DE LOS GATOS

$$\dot{S_c} = I_c\mu_c p - \beta S_c\frac{I_c}{N_c} \qquad \dot{I} = \beta S_c\frac{I}{N} - \mu I_c p \qquad A(t) = \frac{S_c}{N_c} \qquad B(t) = \frac{I_c}{N_c}$$

$$\boxed{\begin{aligned}\dot{A} &= B(t)\mu_c p - \beta A(t)B(t)\\ \dot{B} &= \beta A(t)B(t) - \mu B(t)p\end{aligned}}$$

3.3. DINÁMICA DE LOS CERDOS

$$\dot{S_p} = \mu pN - \beta Sp\frac{I_c}{N_c} \qquad P_s = \frac{\dot{S_p}}{N_p} \qquad E_s = \frac{\dot{I_p}}{N_p}$$
$$\dot{I}_p = \beta Sp\frac{I_c}{N_c} - p\mu I_p - \beta\frac{S_h}{N}I_p$$

$$\boxed{\begin{aligned}\dot{P_s} &= \mu p - \beta P_s(t)B(t)\\ \dot{E}_p &= \beta P_s(t)B(t) - E_p(t)p\mu - \beta s(t)E_p(t)\end{aligned}}$$

4. SIMULACIÓN NUMÉRICA

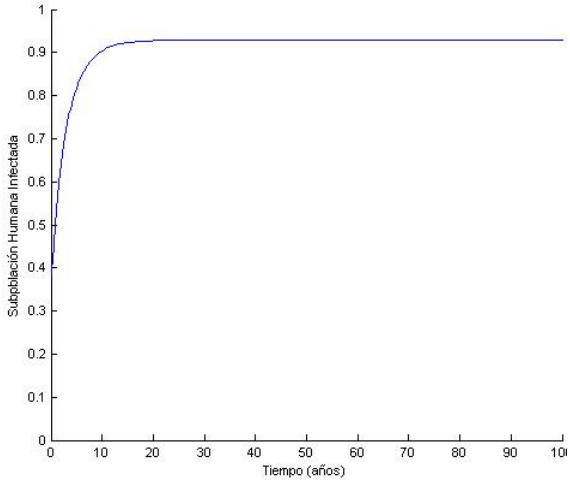


Figura 2: Dinámica de la Población Humana Infectada

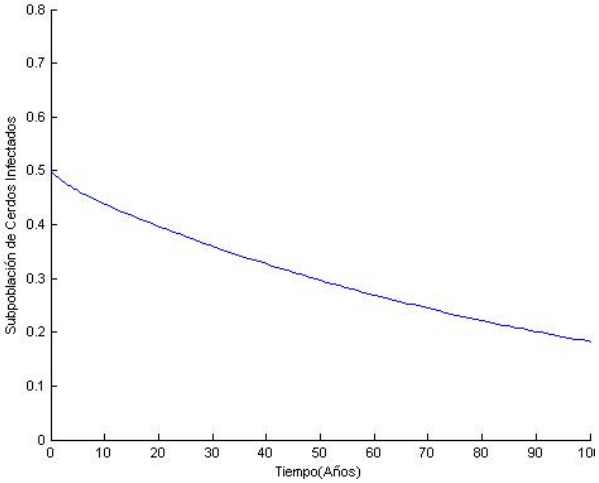


Figura 3: Dinámica de la Población de Cerdos Infectados

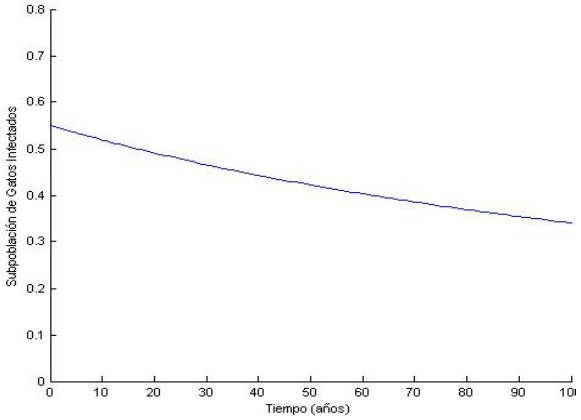


Figura 4: Dinámica de la Población de Gatos Infectados

Parámetros	Valor
p	035
μ	0.3
ε	0.2
β_c	0.01
β_p	0.0212
d	0.01
γ	0.03

Tabla 1: Parámetros Básicos del Modelo Matemático

Para este ambiente se asumen las siguientes condiciones iniciales: s(0)= 0.5253, i(0)= 0.36, c(0)= 0.1147, A(0)= 0.45, B(0)= 0.55, $P_s(0)$= 0.5, $E_p(0)$= 0.5

5. CONCLUSIONES

El presente trabajo consideró enriquecer el modelo matemático de la Toxoplasmosis en Colombia de modelos anteriores [9] y [10], asociando una importante fuente de transmisión de la infección determinada por el consumo de carne de cerdo infectada y preparada con inadecuada cocción [7]. Se presentaron nuevas ecuaciones diferenciales ordinarias no lineales. Según los resultados se muestra que hay una tendencia muy rápida en el aumento de la subpoblación de humanos infectados (Figura 2), debido a la presencia de la nueva fuente de transmisión que permite que los individuos se expongan fácilmente a la infección. Esta subpoblación humana infectada se satura debido a la disminución de las fuentes de transmisión de la infección (Figura 3 y 4). La simulación se obtuvo mediante el programa de Matcont.

AGRADECIMIENTOS

Damos gracias a la Dra. Barbara H. Zimmermann Directora del grupo de investigación BBMP de la Universidad de los Andes y al Dr. Reynaldo Núñez, Director del Departamento de Matemáticas, Universidad Sergio Arboleda. Bogotá.

6. REFERENCIAS

[1] C. GALLEGO, C. SAAVEDRA-MATIZ AND J.E. GÓMEZ-MARÍN. Direct genotyping of animal and human isolates of Toxoplasma gondii from Colombia (South America), 97 (2006), pp. 161-167.

[2] J.P. DUBEY, D.S. LINDSAY, C.A. SPEER. Structures of Toxoplasma gondii Tachyzoites, Bradyzoites, and Sporozoites and Biology and Development of Tissue Cysts. Clinical Microbiology, 11 (1998), pp. 267-299.

[3] D.K. HOWE, S. HONORÉ, F. DEROUIN AND L.D. SIBLEY. Determination of genotypes of Toxoplasma gondii strains isolated from patients with toxoplasmosis. J. Clin. Microbial, 35 (1997), pp. 1411-1414.

[4] G.J. PÉREZ, M.T. MORENO, C. BECERRA, C. MARTÍNEZ. The seroprevalence of human toxoplasmosis in Córdoba. Rev Sanid Hig Publica, 66 (1992), pp. 83-9.

[5] C.A. LÓPEZ, J. DÍAZ, J.E. GÓMEZ. Factores de riesgo en mujeres embarazadas, infectadas por *Toxoplasma gondii* en Armenia-Colombia. Rev Salud Pública, 7 (2005), pp. 180-190

[6] F. LORA, H.J. ARICAPA, J.E. PÉREZ, L.E. ARIAS, S.E. IDARRAGA, D. MIER, J.E. GÓMEZ. Detección de *Toxoplasma gondii* en carnes de consumo humano por la técnica de reacción en cadena de la polimerasa en tres ciudades del eje cafetero. Infection, 11 (2007), pp. 117-123

[7] S.A. MONTEALEGRE, Y.A. VALBUENA, L.J. CORTÉS, S.A FLÓREZ. Seroprevalencia de la toxoplasmosis y factores relacionados con las enfermedades transmitidas por alimentos en trabajadores de plantas de beneficio animal en cinco ciudades capitales de Colombia, 2008. Biomédica, 7 (2009), pp. 66-70

[8] C.H. WANG, J. KLIEBENSTEIN, A. HALLAM, J. ZIMMERMAN, V. DIDERERRICH, S. PATTON, C. FAULKNER, R. MCCORD, E. BUSH. Levels of Toxoplasma gondii in swine operations. [On line]: Swine Research Report 2000, Iowa State University, Department of Economics, (2000), pp. 198-201.

[9] D.F. ARANDA, R.J. VILANUEVA, A.J. ARENAS, G.C. GONZALEZ-PARRA. Mathematical modeling of toxoplasmosis disease in varying size populations. Computers and Mathematics with applications, 56 (2008), pp. 690-696

[10] G.C. GONZALEZ-PARRA, A.J. ARENAS, D.F. ARANDA, R.J. VILANUEVA, L. JÓDAR. Dynamics of a model of Toxoplasmosis disease in human and cat populations. Computers and Mathematics with applications, 57 (2009), pp.1692-1700

A NN-BASED AUTOREGRESSIVE MODEL THAT CONSIDERS THE ENERGY ASSOCIATED OF TIME SERIES FOR FORECASTING

C. RODRÍGUEZ RIVERO, J. PUCHETA, J. BAUMGARTNER, M. HERRERA, C. SALAS AND V. SAUCHELLI

Departments of Electrical and Electronic Engineering, Mathematics Research Laboratory Applied to Control (LIMAC), Faculty of Exact, Physical and Natural Sciences - National University of Córdoba, Velez Sarsfield Av. 1611 (University Campus) Córdoba, Argentina, cristian.rodriguezrivero@gmail.com, www.mathlabappliedcontrol.blogspot.com.

Departments of Electrical Engineering, Faculty of Technologies and Applied Sciences, National University of Catamarca, Catamarca, Argentina.

Abstract: In this work a neural network (NN) based autoregressive model for time series forecasting that takes into account the energy associated with the series is presented. The criterion for fitting comprises to yield time series values from forecasted time series area. These values are approximated by the NN to generate a primitive calculated as an area by the predictor filter. The NN output will tend to approximate the current value available from the series which has the same Hurst Parameter as the real time series. The approach is tested over a time series obtained from samples of the Mackey-Glass delay differential equations (MG) and serve to be applied for meteorological variables measurements such as soil moisture series, daily rainfall and monthly cumulative rainfall time series forecasting.

Keywords: Neural networks, time series forecast, area, Primitive, Hurst's parameter, Mackey-Glass.

1. INTRODUCTION

Natural phenomena prediction is a challenging topic, useful for control problems from agricultural activities and decision-making that helps the producer to decide. There are several approaches based on NN that face the rainfall forecast problem for energy demand purposes, for water availability and seedling growth by taking an ensemble of measurement points [1]. Here, the proposed approach is based on the classical NAR filter using time-lagged feed-forward neural networks, where the forecast data used from the MG benchmark equation is simulated by a Monte Carlo [2] approach. The number of filter parameters is put function of the roughness of the time series, in such a way that the error between the smoothness of the time series data and the forecasted data modifies the number of the filter parameters.

2. OVERVIEW OF THE MG EQUATION AND FBM

Samples of MG equation are used to model natural phenomena and have been implemented by different authors to perform comparisons between different techniques for forecasting and regression models [3]. Here we propose an algorithm to predict values of time series taken from the solution of the MG equation. The MG equation is explained by the time delay differential equation defined as

$$\dot{y}(t) = \frac{\alpha y(t-\tau)}{1+y^c(t-\tau)} - \beta y(t), \tag{1}$$

where α, β, and c are parameters and τ is the delay time. According as τ increases, the solution turns from periodic to chaotic. Equation (1) is solved by a standard fourth order Runge-Kutta integration step. By

setting the parameter β ranging between 0.1 and 0.9 the stochastic dependence of the deterministic time series obtained varies according to its roughness. The performance of the proposed method is tested with the *SMAPE* index and it is compared with a traditional NN based predictor.

Due to the random process, it is proposed to use the Hurst's parameter *H* in the learning process to modify on-line the number of patterns, of iterations and filter inputs. This *H* gives information about the roughness of a signal, and also to determine its stochastic dependence. The definition of the Hurst's parameter is defined by Mandelbrot through its stochastic representation

$$B_H(t) = \frac{1}{\Gamma\left(H+\frac{1}{2}\right)}\left(\int_{-\infty}^{0}\left((t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}}\right)dB(s) + \int_0^t (t-s)^{H-\frac{1}{2}}\,dB(s)\right) \tag{2}$$

where, $\Gamma(\cdot)$ represents the Gamma function

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx \tag{3}$$

and $0<H<1$ is called the Hurst parameter. The integrator B is a stochastic process, ordinary Brownian motion. Note, that *B* is recovered by taking $H=1/2$ in (2) and *B* is defined on some probability space (Ω, F, P). Thus, an fBm is a time continuous Gaussian process depending on the so-called Hurst parameter $0<H<1$. The ordinary Brownian motion is generalized to $H=0.5$, and its derivative is the white noise. The fBm is self-similar in distribution and the variance of the increments is defined by

$$Var\left(B_H(t) - B_H(s)\right) = v|t-s|^{2H} \tag{4}$$

where *v* is a positive constant.

This special form of the variance of the increments suggests various ways to estimate the parameter *H*. In fact, there are different methods for computing the parameter *H* associated to Brownian motion [4]. In this work, the algorithm uses a wavelet-based method for estimating H from a trace path of the fBm with parameter *H* [5].

3. PROBLEM STATEMENT

Motivations that have led to this study follows the closed-loop control scheme in [6]. The controller considers future meteorological conditions for designing the control law in the sense that the controller takes into consideration the actual state of the crop by a state observer and the meteorological variables, respectively. However, in this work only the controller's portion concerning with the prediction system is presented by using a benchmark of MG time series. The controller design is inspired on the one presented in [6]. The main contribution of this work is in the learning process, which employs the Levenberg-Marquardt rule and considers the long or short term stochastic dependence of passed values of the time series to adjust at each time-stage the number of patterns, of iterations, and the length of the tapped-delay line, in function of the Hurst's value, *H* of the time series. The NN-based nonlinear filter is applied to the time series obtained from MG to forecast the next 18 values out of a given historical data set of 102 values.

4. THE PROPOSED LEARNING PROCESS

The NN's weights are tuned by means of the Levenberg-Marquardt rule, which considers the long or short term stochastic dependence of the time series measured by the Hurst's parameter *H*. The proposed learning approach consists of changing the number of patterns, the filter's length and the number of iterations in function of the parameter *H* for each corresponding time series. The proposed criterion to modify the pair (i_t, N_p) is given by the statistical dependence of the time series $\{x_n\}$, supposing that it is an

fBm. The dependence is evaluated by the Hurst's parameter H, which is computed by a wavelet-based method [5]. Then, a heuristic adjustment for the pair (i_t,N_p) in function of H is proposed.

5. APPROXIMATION BY PRIMITIVE

It is considered to take the area resulting of integrating the time series data of MG equation. That primitive is obtained by considering each value of the time series its derivate as follows

$$\int_{t_k}^{t_{k+1}} y_t dt \cong y_t\left(t_{k+1} - t_k\right) \tag{5}$$

where y_t is the original time series value. The approximation area is assumed to be its periodical primitive

$$I_{t_n} = \int_{t_n}^{t_{n+p}} y_t dt = Y_t\Big|_{t_n}^{t_{n+p}}, n = 1,2,...N. \tag{6}$$

During the learning process, those primitives are calculated and serve as a new input to the NN. The prediction attempts to make the area of the forecasted times series equal to the primitive real area predicted. The real area is used in two instances. Firstly, from the real time series, an area is obtained and the H parameter from this time series is called H_A. On the other hand, the time series data is forecasted, so the H parameter from this time series is called H_S. After the training process is completed, both sequences $\{\{I_n\}, \{I_e\}\}$ and $\{\int\{y_n, y_e\}\}$ are compared, in accordance with the hypothesis, they should have the same H parameter. If the error between H_A and H_S is greater than a threshold θ, the value of l_x is either increased or decreased according to $l_x \pm 1$.

6. MAIN RESULTS

6.1 GENERATIONS OF AREAS FROM MG EQUATIONS

Primitives of time series are obtained from the MG equations (1) with parameters τ=100 and α=20, β=0.32 and β=1.6. This collection of coefficients was chosen to generate time series whose H parameters vary between 0 and 1. The chosen one was selected in accordance to its roughness. The performance measure of the filter is evaluated using the Symmetric Mean Absolute Percent Error (*SMAPE*) proposed in the most of metric evaluation, defined by

$$SMAPE_S = \frac{1}{n}\sum_{t=1}^{n} \frac{\left|X_t - F_t\right|}{\left(X_t + F_t\right)/2} \cdot 100 \tag{7}$$

where t is the observation time, n is the size of the test set, s is each time series, X_t and F_t are the actual and the forecasted time series values at time t respectively. The SMAPE of each series $\underline{\boldsymbol{s}}$ calculates the symmetric absolute error in percent between the actual X_t and its corresponding forecast value F_t, across all observations t of the test set of size n for each time series s.

6.2 COMPARATIVE RESULTS

The performance of the stochastic NN-based predictor filter is evaluated through the *SMAPE* index along the time series from MG solutions with β=1.6 shown in Table 1.

Table 1. Comparisons obtained by the proposed approach

Series	H_S	H_A
MG with β=1.6	0.6543	0.6423

The comparison was made between the deterministic approach [7] and the present forecasted time series. In addition, an area of a primitive value acquired of MG time series was incorporated in order to use the proposed approach. Furthermore, the *SMAPE* value is improved by order of 10^{-5} for a class of time series with high roughness of the signal, in this case with β=1.6 which is one of the worst condition for signal prediction.

7. CONCLUSIONS

In this work a NN-based autoregressive model for time series forecasting that considers the energy associated with the data series was presented. The learning rule proposed to adjust the NN's weights is based on the Levenberg-Marquardt method. Likewise, in function of the long or short term stochastic dependence of the time series evaluated by the Hurst parameter *H*, an on-line heuristic adaptive law was proposed to update the NN topology at each time-stage. The major result shows that the area predictor system supplied to time series has an optimal performance from several samples of MG equations, in particular, those whose *H* parameter has a high roughness of signal, which is assessed by H_S and H_A, respectively. This fact encourages us to be applied for meteorological variables measurements such as soil moisture series, daily rainfall and monthly cumulative rainfall time series forecasting when the observations are taken from a single point.

ACKNOWLEDGMENTS

This work was supported by National University of Córdoba (UNC), FONCYT-PDFT PRH N°3 (UNC Program RRHH03), SECYT UNC, National University of Catamarca, National Agency for Scientific and Technological Promotion (ANPCyT) under grant PICT-2007-00526 and Departments of Electrotechnics - FCEFyN of National University of Cordoba.

8. REFERENCES

[1] MASULLI, F., BARATTA, D., CICIONE, G., STUDER, L. *Daily Rainfall Forecasting using an Ensemble Technique based on Singular Spectrum Analysis.* In Proceedings of the International Joint Conference on Neural Networks IJCNN 01, pp. 263-268, vol. 1, IEEE, Piscataway, NJ, USA, 2001.

[2] BISHOP, C. *Pattern Recognition and Machine Learning*. Springer. Boston, 2006.

[3] VELÁSQUEZ HENAO, JUAN DAVID, DYNA, *Red. Pronóstico de la serie de Mackey glass usando modelos de regresión no-lineal*. Universidad Autónoma de Mexico. Campus Aragón. 2004.

[4] DIEKER, T. *Simulation of fractional Brownian motion*. MSc theses, University of Twente, Amsterdam, The Netherlands. 2004.

[5] FLANDRIN, P. *Wavelet analysis and synthesis of fractional Brownian motion*. IEEE Trans. on Information Theory, 38, pp. 910-917. 1992.

[6] PUCHETA, J., PATIÑO, H., SCHUGURENSKY, C., FULLANA, R., KUCHEN, B. *Optimal Control Based-Neurocontroller to Guide the Crop Growth under Perturbations. Dynamics Of Continuous*, Discrete And Impulsive Systems Special Volume Advances in Neural Networks-Theory and Applications. DCDIS A Supplement, Advances in Neural Networks, Watam Press, Vol. 14(S1), pp. 618—623. 2007.

[7] PUCHETA, J., HERRERA, M., SALAS C., PATIÑO, H.D., AND B. KUCHEN. *A Neural Network-Based Approach for Forecasting Time Series from Mackey-Glass Equations"*. In proc. Of the XIII Reunión de Trabajo en Procesamiento de la Información y Control ISBN 950-665-340-2. XII RPIC, organizado por el Laboratorio de Sistemas Dinámicos y Procesamiento de la Información, 16 al 18 de Setiembre de 2009 Rosario, Argentina. (2009).

UN MODELO COMBINADO CONTINUO-DISCRETO PARA EL DISEÑO DE AUTOPISTAS. IMPACTO AMBIENTAL.

Patricia N. Dominguez†, Víctor H. Cortínez‡§

† Departamento de Ingeniería, Universidad Nacional del Sur, Av. Alem 1253, 8000 Bahía Blanca, pdoming@uns.edu.ar
‡ *Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional (FRBB), 11 de abril 461 8000 Bahía Blanca, vcortine@frbb.utn.edu.ar*
§*CONICET*

Resumen: En este trabajo se presenta un modelo combinado continuo-discreto para el diseño de autopistas y el impacto sobre el medio ambiente. Se asume que los usuarios de la red de transporte se encuentran distribuidos en la ciudad y pueden optar por viajar a través de la red de calles o acceder a las autopistas en determinados puntos de la ciudad, de acuerdo a un criterio de costo mínimo de viaje. Se utiliza el clásico enfoque discreto para modelar el tráfico en las autopistas y un nuevo enfoque continuo para modelar el resto de la red de transporte.

Palabras claves: *modelo continuo-discreto, tráfico urbano, impacto ambiental, contaminación atmosférica.*
2000 AMS Subjects Classification: 90B20, 62P12

1 INTRODUCCIÓN

Los problemas de asignación de tráfico urbano se abordan habitualmente a través de dos enfoques. En el enfoque discreto, la red de tráfico se modela mediante un grafo dirigido $N(V,A)$ donde los arcos A representan tramos de las vías de circulación y los nodos V los puntos de intersección de las calles y los puntos de origen/destino de los viajes. Este modelo permite obtener en forma detallada el flujo vehicular y los tiempos de viaje en cada arco [1]. En el enfoque continuo, la red es aproximada a un continuo donde los conductores son libres de elegir sus rutas en un espacio bidimensional, se acepta que la variación en áreas cercanas es pequeña comparada con el sistema completo y en consecuencia las características del sistema de transporte pueden ser representadas con funciones matemáticas continuas [2]. El enfoque continuo permite visualizar rápidamente el comportamiento de la red con menor cantidad de datos y menor esfuerzo computacional. En el modelo que se desarrolla en este trabajo, se aprovechan las ventajas de ambos modelos representando la red urbana mediante un modelo continuo generalizado desarrollado en la referencia [2] y la red de autopistas de acuerdo al clásico modelo discreto. El efecto sobre el medio ambiente se calcula con un modelo de emisión y uno de transporte.

2 DESCRIPCIÓN DEL MODELO

Consideramos la ciudad como una región Ω delimitada por un borde exterior Γ a través del cual no ingresan ni salen viajes (Figura 1a). La demanda de viajes se asume distribuida en la ciudad. Se considera un solo punto de destino, el centro de la ciudad O_0. La red de transporte se aproxima mediante la superposición de un modelo continuo que representa las calles de la ciudad y una red discreta de autopistas. Ambos sistemas interactúan en los puntos de acceso a las autopistas O_n. La solución del problema continuo proporciona la demanda en los puntos de acceso a las autopistas (matriz origen-destino) del problema discreto. A su vez, la solución del problema discreto fija las condiciones de borde (tiempos hasta O_0) en los puntos O_n, del problema continuo. En este proceso iterativo, cuando se alcanza el equilibrio, los tiempos en los accesos, obtenidos en la solución de ambos problemas son los mismos.

2.1 SISTEMA DISCRETO DE TRÁFICO

Los patrones de flujo y tiempo de viaje en los arcos de una red discreta se pueden obtener resolviendo el problema de equilibrio de usuario [1]. Llamamos a este problema P1.

$$\min z(\mathrm{x}) = \sum_a \int_0^{x_a} t_a(\omega) d\omega \qquad (1a)$$

s.a.

$$\sum_k f_k^n = Q_n, \qquad n = 1, 2, ..., N, \qquad (1b)$$

$$f_k^n \geq 0, \qquad \forall k, n \;\; n = 1, 2, ..., N, \qquad (1c)$$

donde

$$x_a = \sum_n \sum_k f_k^n \delta_{a,k}^n, \qquad \forall a \in A. \qquad (1d)$$

f_k^n es el flujo en la ruta k entre el acceso n y O_0, $\delta_{a,k}^n$ es un indicador que relaciona arcos y rutas y vale 1 si a pertenece a la ruta k entre n y O_0 y vale 0 en caso contrario, $t_a=t_a(x_a)$ es la función de costo, determinada experimentalmente, asociada al tramo a y x_a es el flujo vehicular en dicho tramo. La demanda en cada punto de acceso Q_n es proporcionada, en este caso, por el modelo continuo. Llamamos $\overline{\mathbf{U}} = (\overline{U}_n, n = 0,1,...,N)$ a los tiempos mínimos de viaje desde los puntos de intercambio n hasta el centro, solución de P1. Este problema se puede escribir, en forma abstracta, de la siguiente manera [3]:

$$\overline{\mathbf{U}} = \mathrm{G}(\mathbf{Q}), \qquad \mathbf{Q} = (Q_n, n = 0,1,...,N). \qquad (2)$$

2.2 SISTEMA CONTINUO DE TRÁFICO

La distribución de usuarios de la red sobre la ciudad se supone continua y representada por una función de demanda por unidad de superficie $q(x,y) = D(u(x,y),\xi)$ donde $u(x,y)$ es el costo mínimo para usuarios localizados en $H(x,y)$ hacia alguno de los accesos a las autopistas o al centro de la ciudad y ξ es un parámetro de sensibilidad. Se considera una función de tiempo de recorrido del arco $t_a = t_{a_0}\left(1+\alpha_a\left((x_a+x_{Ra})/C_a\right)^{\varphi_a}\right)$, donde C_a es la capacidad del arco, medida en vehículos por hora, α_a y φ_a son coeficientes que tienen en cuenta las características específicas de la arteria considerada, t_{a_0} es el tiempo de recorrido a flujo libre y x_{Ra} es un flujo residual. El modelo continuo utilizado en este trabajo contempla la posible anisotropía de la red de transporte [2]. Se divide el dominio completo de la ciudad Ω en M celdas de área L_xL_y (Figura 1a), donde las longitudes L_x y L_y son pequeñas con respecto a las dimensiones de la ciudad. Se supone que en el área que se examina existe un sistema de calles paralelas (que forman un ángulo γ_a con respecto al eje de referencia horizontal) de tal manera que para cada calle en una dirección y sentido, existe otra en igual dirección y sentido contrario.

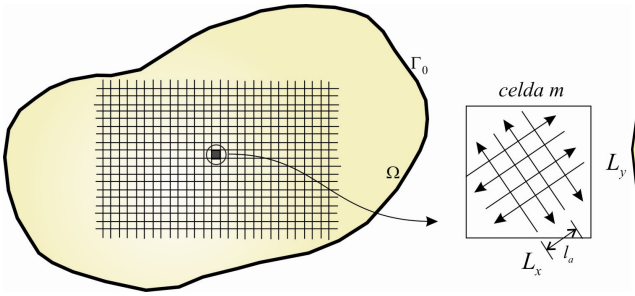


Figura 1a : Modelo continuo de tráfico urbano

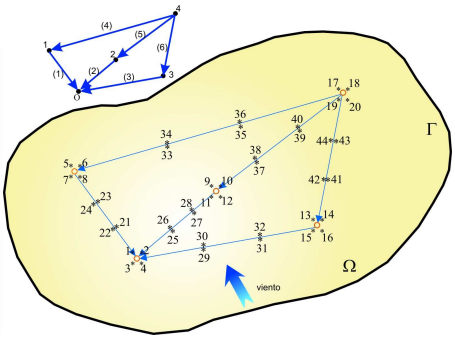


Figura 1b : Autopistas y receptores de CO.

A partir de la formulación dual del problema discreto, aplicando cálculo variacional, se obtiene la ecuación diferencial no lineal que permite resolver el problema de asignación de tráfico [2]:

$$\frac{\partial}{\partial x}\left(k_x\frac{\partial u}{\partial x}+k_{xy}\frac{\partial u}{\partial y}\right)+\frac{\partial}{\partial y}\left(k_{xy}\frac{\partial u}{\partial x}+k_y\frac{\partial u}{\partial y}\right)+q=0, \qquad \forall (x,y)\in\Omega, \qquad (3a)$$

donde $k_x = \sum_{a\in m} p_a(t_a)\frac{l_a^2\cos^2\gamma_a}{L_xL_y}$, $k_y = \sum_{a\in m} p_a(t_a)\frac{l_a^2\mathrm{sen}^2\gamma_a}{L_xL_y}$, $k_{xy} = \sum_{a\in m} p_a(t_a)\frac{l_a^2\cos\gamma\mathrm{sen}\gamma}{L_xL_y}$,

$$p_a = \frac{x_a(t_a)}{t_a}, \quad x_a(t_a) = \left(\frac{t_a - t_{a_0}}{\alpha_a t_{a_0}}C_a^{\varphi_a}\right)^{\frac{1}{\varphi_a}} - x_{Ra} \quad \text{y} \quad t_a = -l_a\left(\frac{\partial u}{\partial x}\cos\gamma+\frac{\partial u}{\partial y}\mathrm{sen}\gamma\right).$$

En el borde externo del dominio se debe cumplir:

$$f_x n_x + f_y n_y = 0, \qquad \forall (x,y)\in\Gamma, \qquad (3b)$$

donde $f_x = -\left(k_x \frac{\partial u}{\partial x} + k_{xy} \frac{\partial u}{\partial y} \right)$, $f_y = -\left(k_{xy} \frac{\partial u}{\partial x} + k_y \frac{\partial u}{\partial y} \right)$, mientras que n_x y n_y son las componentes del versor normal a la curva que define el contorno de la ciudad. Las otras condiciones de borde son: a) valor nulo de u en el centro de la ciudad, $u(x_o, y_o) = 0$ y b) valor conocido de u en los accesos a las autopistas $u(x_n, y_n) = U_n$. La solución del problema continuo, que llamamos P2, proporciona la demanda en los puntos de acceso a las autopistas, en función de los tiempos mínimos de viaje. En forma abstracta:

$$\mathbf{Q} = \mathrm{F}(\mathbf{U}), \qquad \mathbf{U} = (U_n, n = 0,1,...,N). \tag{4}$$

2.3 MODELOS DE EMISIÓN Y DE DISPERSIÓN

La emisión de contaminantes depende del tipo de vehículo y de las condiciones de circulación [4]. A su vez los contaminantes emitidos se dispersan en la atmósfera y la concentración de los mismos en determinada ubicación geográfica se puede obtener mediante la solución de la ecuación de transporte:

$$\frac{\partial (V_x C)}{\partial x} + \frac{\partial (V_y C)}{\partial y} + \frac{\partial (V_z C)}{\partial z} = \frac{\partial}{\partial x}\left(K_x \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial C}{\partial z} \right), \tag{5a}$$

donde C es la concentración de un contaminante específico, V_x, V_y y V_z son las componentes de la velocidad media del viento y, K_x, K_y y K_z son los coeficientes de dispersión turbulenta según las coordenadas x, y, z respectivamente. Las condiciones de borde que se aplican a la ecuación diferencial son las siguientes:

$$K_z \left. \frac{\partial C}{\partial z} \right|_{z=0} = -D, \; K_z \left. \frac{\partial C}{\partial z} \right|_{z=h} = 0, \; C \to 0 \;\; x, y \to \pm\infty, \tag{5b}$$

donde D es la emisión de la fuente en unidades de masa de contaminante por unidad de superficie.

3 ALGORITMO DE SOLUCIÓN.

De las ecuaciones (2) y (4)

$$\overline{\mathbf{U}} = \mathbf{G}(\mathbf{F}(\mathbf{U})). \tag{6}$$

La solución $\mathbf{U}^*$ es aquella que satisface simultáneamente las relaciones funcionales y las condiciones del equilibrio de usuario del problema combinado. El problema entonces, es encontrar $\mathbf{U}$ tal que las diferencias entre los costos de viaje en los accesos de los dos problemas (continuo, discreto) sean muy pequeñas. Esto se resuelve a través de un proceso iterativo hasta que en la iteración k, el error $\left|\mathbf{E}^k\right| < \varepsilon$, donde $\mathbf{E}^k = \mathbf{U}^k - \overline{\mathbf{U}}^k$. En este trabajo para resolver el problema P1 se utiliza el algoritmo de Frank-Wolfe [1] implementado en Matlab y para resolver el problema P2 se utiliza el software FlexPDE que resuelve ecuaciones diferenciales a derivadas parciales en un dominio espacial mediante el método de elementos finitos.

4 EJEMPLO DE APLICACIÓN

El ejemplo se desarrolla con los siguientes datos. Demanda: $q = 100\, veh/h/km^2$; $0,8q$; $1,3q$ y $0,7q$ en las zonas 2, 3, 4 y 5 respectivamente (Figura 2). Composición vehicular: automóviles particulares 80%, de los cuales 70% utiliza nafta y el resto gasoil; vehículos pesados 15% y motos 5%. Los factores de emisión de CO se obtienen de la guía EMEP/CORINAIR 2009 [4]. En el sistema discreto se adopta $t_a = t_{0a}\left(1 + 0,85\left(x_a / C_a\right)^5\right)$ y en el continuo $t_a = t_{0a}\left(1 + 0,15\left(x_a / C_a\right)^4\right)$. Los tiempos a flujo libre se calculan para una velocidad máxima en autopistas de 100 *km/h* y en las calles de 60 *km/h*. La capacidad de los tramos 1, 2 y 3 se fija en 10000 *veh/h*, en los tramos 4, 5 y 6, 6000 *veh/h* y en las calles 600 *veh/h*. Viento: 120º con dirección *O-E;* velocidad 3,17 *m/s* a 10*m* de altura, $V_x = v_x^* / 0,4(\ln(z/0,3) - \psi_1)$ (x en la dirección del viento), $\psi_1 = 2\ln((1+\phi^{-1})/2) + \ln((1+\phi^{-2})/2) - 2\mathrm{arctg}(\phi^{-1}) + \pi/2$, $v_x^* = 0,41\, m/s$, $\phi = (1 - 15z/L)^{-0.25}$, $L = -50m$. Coeficientes de dispersión $K_z = 0,4 v_x^* z(1 - 15z/6L)^{0,25}$, $K_y = 2K_z$ [5]. Los receptores de CO (Figura 1b) en las autopistas están colocados a 25 *m* de las mismas y los que rodean a los puntos de acceso a 100 *m* del borde que los delimita.

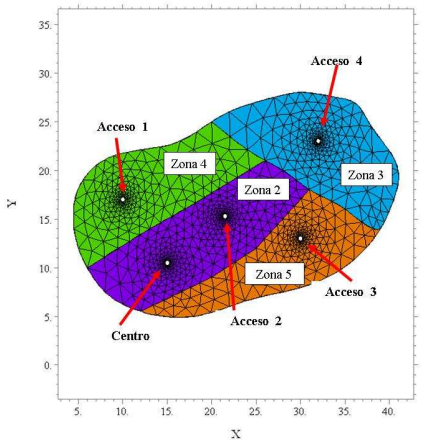


Figura 2: Zonas urbanas y accesos a las autopistas

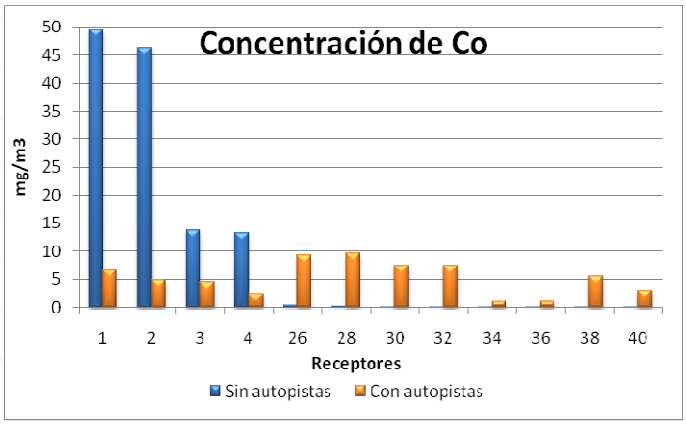


Figura 3: Concentración de CO a 1*m* de altura.

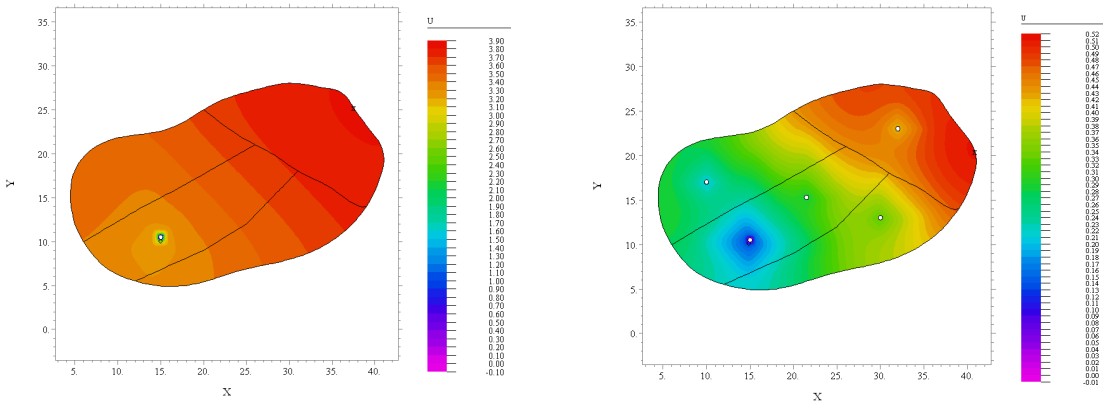


Figura 4: Tiempos de viaje hasta el centro de la ciudad en la red de tráfico sin y con autopistas

De los 61048 viajes que se generan, el 56,5% de los conductores eligen las autopistas, ingresando 16,1%, 13,7%, 17,3% y 9,4% del total de viajes por los accesos 1, 2, 3 y 4 respectivamente. El resto de los conductores se dirige al centro por las calles de la ciudad. A modo de ejemplo se muestran algunos valores representativos de concentración de CO y tiempo de viaje en las Figuras 3 y 4 respectivamente.

5 CONCLUSIONES

El modelo presentado permite estudiar el comportamiento de una red completa de tráfico urbano (calles y autopistas) y determinar la contaminación atmosférica originada por el flujo vehicular. Debido a la reducida cantidad de datos que requiere y la velocidad de cálculo, constituye una buena opción cuando es necesario aplicarlo reiteradamente, como por ejemplo en problemas de optimización. El diseño óptimo de autopistas sujeto a restricciones de diseño, de costo y ambientales, tanto acústicas como atmosféricas, es el tema a desarrollar en futuras investigaciones por parte de los autores.

AGRADECIMIENTOS

Este trabajo forma parte de un proyecto desarrollado en el CIMTA, SCyT Universidad Tecnológica Nacional, bajo la dirección del Dr. Víctor Cortínez.

REFERENCIAS

[1] Sheffi, Y., Urban *Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1984.

[2] Cortínez V., Dominguez, P. *Un modelo continuo anisótropo para el estudio del comportamiento del tráfico urbano congestionado*. Mecánica Computacional, Vol XXIX (2010), pp. 2173-2197.

[3] Wong, S.; Du, Y.; Ho, H.; Sun, L. *A simultaneous optimization formulation of a discrete/continuous transportation system*. Workshop on Theory and Practice of Transportation Science, China, (2002).

[4] Dominguez, P., Vidal, M., Cortínez V. *Diseño óptimo de redes de transporte urbano considerando aspectos medioambientales*. Mecánica Computacional, Vol. XXVIII (2009), pp.2599-2624.

[5] Elkamel, A., Fatehifar, E., Taheri, M., Al-Rashidi, M. y Lohi, A. *A heuristic optimization approach for Air Quality Monitoring Network design with the simultaneous consideration of multiple pollutants*. Journal of Environmental Management, Vol. 88 (2008), pp. 507-516.

ALGORITMOS PARA TRANSFERIR DATOS ENTRE GRILLAS AERODINÁMICAS Y MALLAS ESTRUCTURALES: UNA REVISIÓN DE ALTERNATIVAS PARA LA AEROELASTICIDAD COMPUTACIONAL

Mauro S. Maza†‡§, Sergio Preidikman†‡§ y Fernando G. Flores†‡

†CONICET – Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917, Buenos Aires, Argentina, mauro-maza@hotmail.com, www.conicet.gov.ar
‡Departamento Estructuras, FCEFyN, Universidad Nacional de Córdoba, Casilla de Correo 916, Córdoba, Argentina, fflores@efn.uncor.edu, www.efn.uncor.edu
§Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Ruta Nacional 36 Km 601, 5800 Río Cuarto, Argentina, spreidikman@ing.unrc.edu.ar, www.ing.unrc.edu.ar

Resumen: El principal objetivo de este artículo es presentar los resultados de una extensa revisión bibliográfica realizada con el fin de identificar y evaluar los métodos más comúnmente utilizados para transferir información entre grillas/mallas correspondientes a la dinámica de fluidos computacional (CFD) y mallas correspondientes a la dinámica de estructuras computacional (CSD). El problema de transferencia puede fácilmente transformarse en el factor que controla la precisión de la simulación aeroelástica e involucra usualmente: 1) la transferencia de desplazamientos, velocidades, y aceleraciones desde los nudos de la malla de CSD hacia los puntos nodales de la grilla de CFD; y 2) la transferencia de fuerzas/presiones desde los llamados puntos de control de la grilla de CFD hacia los nudos de la malla de CSD. El objetivo final es identificar los mejores candidatos para ser implementados en un código computacional de alta fidelidad que permita realizar simulaciones del comportamiento aeroservoelástico de generadores eólicos de gran potencia y de eje horizontal.

Palabras claves: *Fluid-solid interactions, Finite element methods, Boundary element methods, Vortex methods*
2000 AMS Subjects Classification: 74F10 - 74S05 - 76M15 - 76M23

1. INTRODUCCIÓN

Una tarea de gran importancia dentro del campo de la aeroelasticidad computacional (CAE) es la simulación numérica de problemas de interacción fluido-estructura (FSI). La principal dificultad radica en que las acciones aerodinámicas sobre un cuerpo flexible inmerso en un fluido, con movimiento relativo entre ellos, dependen de la forma, velocidad, y aceleración del cuerpo, mientras que estas tres dependen de las cargas aerodinámicas que el fluido ejerce sobre el cuerpo flexible. El problema aeroelástico puede ser tratado computacionalmente de dos maneras:

1. Resolviendo en forma conjunta y acoplada las ecuaciones que gobiernan al comportamiento del fluido y la estructura (esquema monolítico); o
2. Resolviendo en forma separada (con métodos separados para cada problema) las ecuaciones en ambos medios, utilizando un esquema de interacción entre ellos (esquema particionado).

Muchos autores creen que el enfoque apropiado para resolver el problema de FSI es el de plantear un esquema particionado. Las principales razones son las siguientes:

1. Las matrices involucradas en la solución con el esquema monolítico estarán mal condicionadas debido a las grandes diferencias de rigidez entre la estructura y el fluido; y
2. Con el esquema particionado se puede utilizar, para cada disciplina, la estrategia de solución que mejor se adapte.

En general, la respuesta de la estructura a las cargas aerodinámicas se calcula utilizando el método de elementos finitos (FEM), mientras que para las cargas aerodinámicas sobre la estructura se utilizan técnicas de la CFD. En todos los casos el dominio del problema es discretizado, apareciendo dos mallas o grillas. Sobre la estructura aparece una malla de elementos finitos, que denominaremos malla estructural (ME). El dominio fluido también se discretiza completamente, se denominará malla aerodinámica (MA) a la parte que se encuentra sobre el contorno del cuerpo definido en el modelo aerodinámico.

La interacción de los códigos se realiza transfiriendo información entre las mallas. Por un lado es necesario transferir variables cinemáticas calculadas con el FEM desde la ME a la MA. Por el otro, deben

llevarse las cargas calculadas con el código de CFD desde la MA a la ME. En la Figura 1 se observan la MA y la ME que representan a la misma ala, aunque una se utiliza en el código de CFD y la otra en el de CSD. En general las mallas tienen topologías muy diferentes. En este caso particular la MA resulta de la discretización de la superficie externa de un ala, mientras la ME representa el modelo del cajón de torsión.

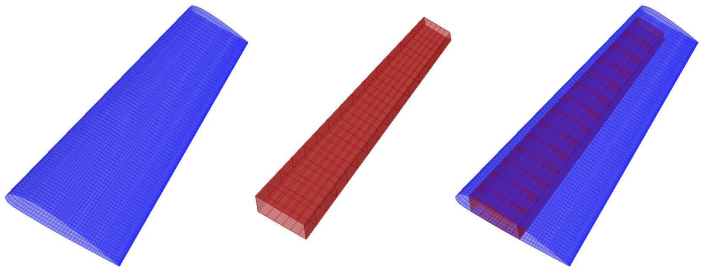


Figura 1: Diferencias entre la malla estructural y la malla aerodinámica.

2. MÉTODOS VARIACIONALES

Estos métodos hacen uso del Principio de los Trabajos Virtuales para asegurar que la energía adquirida (o entregada) por la estructura (excepto que exista amortiguamiento estructural) debe ser igual a la energía entregada (o adquirida) por el fluido. Para ello se fuerza a que el trabajo virtual realizado por la estructura sea igual al realizado por el fluido, esto es:

$$\delta W_E = \delta W_A \quad \Rightarrow \quad \mathbf{f}_E{}^T \delta \mathbf{u}_E = \mathbf{f}_A{}^T \delta \mathbf{u}_A \,, \tag{1}$$

donde $\mathbf{f}_E$ y $\mathbf{f}_A$ son vectores con cargas puntuales aplicadas sobre los nodos de la ME y de la MA, y $\delta\mathbf{u}_E$ y $\delta\mathbf{u}_A$ son vectores de desplazamientos virtuales nodales. En caso de utilizar un método de transferencia para los desplazamientos, tal que $\Delta\mathbf{u}_A = \mathbf{H}\Delta\mathbf{u}_E$, los desplazamientos virtuales deben ser compatibles, por lo que pueden interpolarse de la misma manera. Reemplazando $\delta\mathbf{u}_A = \mathbf{H}\delta\mathbf{u}_E$ en la Ec. (1) se tiene:

$$\mathbf{f}_E{}^T \delta \mathbf{u}_E = \mathbf{f}_A{}^T \mathbf{H}\, \delta \mathbf{u}_E \,. \tag{2}$$

Luego, simplificando los $\delta\mathbf{u}_E$ (debido a su arbitrariedad) y transponiendo ambos miembros se llega a:

$$\mathbf{f}_E = \mathbf{H}^T \mathbf{f}_A \,. \tag{3}$$

La Ec. (3) indica que, una vez calculada la matriz de transferencia **H** para interpolar los desplazamientos de la MA a partir de los de la ME, debe utilizarse su transpuesta para interpolar las fuerzas de la MA en la ME si se desea conservar la energía en el proceso de interacción.

2.1. INTERPOLACIÓN CON FUNCIONES BASE RADIALES

Una función radial $\varphi(r)$ es una función continua, de una variable escalar r, con un comportamiento radial respecto a un punto denominado *centro*, siendo r la distancia desde el centro hasta el punto donde se evalúa la función φ. En general se utiliza la distancia euclídea, aunque algunos autores han sugerido definir otras normas con propiedades que las hacen más adecuadas para ciertos problemas de interacción.

La teoría general de interpolación con funciones base radiales (RBF) la presentan Buhman [1] y Wendland [2]. Estos métodos han resultado muy exitosos para la interpolación de una función $s(\mathbf{x})$ a partir de valores conocidos en puntos discretos y ubicados de forma no estructurada $\mathbf{x}_i$, utilizando la función de interpolación:

$$s(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \varphi\left(\left\|\mathbf{x} - \mathbf{x}_i\right\|\right) + p(\mathbf{x}) \,. \tag{4}$$

En la Ec. (4), $p(\mathbf{x})$ es un polinomio en tres dimensiones (cuya utilización es opcional). Beckert y Wendland [3] utilizaron polinomios lineales para $p(\mathbf{x})$, de manera de recuperar exactamente traslaciones y rotaciones de cuerpo rígido, y de conservar la fuerza y el momento totales. Los coeficientes α_i se calculan imponiendo que la función interpolante devuelva los valores conocidos $s(\mathbf{x}_i) = s_i$ y que se cumpla que:

$$\sum_{i=1}^{N} \alpha_i\, q(\mathbf{x}) = 0 \,, \tag{5}$$

para todo polinomio $q(\mathbf{x})$ de grado menor o igual que $p(\mathbf{x})$, cuando se utiliza la parte polinómica en la Ec. (4).

Puede utilizarse una gran variedad de funciones. En la Tabla 1 se observan funciones utilizadas desde la década de 1970. El principal problema de las funciones crecientes es que los valores de $s(\mathbf{x})$ en puntos lejanos al centro $\mathbf{x}_i$ utilizado tiene más influencia, lo que tiende a suavizar variaciones locales de la función y a producir una matriz **H** llena, que hace más caro computacionalmente el método. La utilización de funciones decrecientes permite realizar una interpolación con mejor correspondencia con la física del problema.

Wendland [4] introdujo la función *Euclid's Hat* y las denominadas *Wendland's Functions*, que son decrecientes y de soporte compacto, lo que permite localizar mejor la interpolación y reducir el número de elementos no nulos en la matriz **H**. Wendland demostró que estas funciones, dado el número de dimensiones en el que se realizará la interpolación y especificada la continuidad deseada, poseen el menor grado posible entre las funciones radiales de soporte compacto definidas positivas. El que las funciones sean definidas positivas, aseguran la unicidad de la solución del problema de interpolación definido en la ecuación (4).

Funciones crecientes		*Funciones decrecientes*	
$\varphi(r) = (r^2 + k^2)^{1/2}$	Biharmonic-Multiquadrics	$\varphi(r) = (r^2 + k^2)^{-1/2}$	Inverse Multiquadrics
$\varphi(r) = r^2\ln(r^2)$	Infinite-Plate Spline	$\varphi(r) = (1 - r)^2$	Wendland (2D y 3D – C^0)
$\varphi(r) = r^2\ln(r)$	Thin-Plate Spline	$\varphi(r) = (1 - r)^4(4r + 1)$	Wendland (2D y 3D – C^2)

Tabla 1: Ejemplos de Funciones Base Radiales

2.2. Método de Elementos de Contorno

Chen y Jadic [5] proponen asimilar el problema de FSI a uno de mecánica de sólidos y resolverlo con un Método de Elementos de Contorno (BEM), utilizando la MA como el contorno de un sólido elástico, lineal y homogéneo y los nodos de la ME como un conjunto de puntos dentro de ese sólido.

Los BEM permiten resolver ecuaciones diferenciales lineales en derivadas parciales a partir de una formulación con integrales en el contorno. En problemas de elasticidad estática, permiten calcular la matriz de transferencia **H** que relaciona los desplazamientos de puntos en el contorno con los de puntos interiores.

Una ventaja es que las deformaciones en todos los sentidos están acopladas utilizando un criterio físico. La desventaja es que se obtiene una matriz **H** llena y no necesariamente simétrica, lo que aumenta el costo computacional.

2.3. Interpolación con Funciones de Forma

Este método hace uso del Inverse Isoparametric Mapping (IIM) para calcular las coordenadas locales de los nodos de la MA dentro de los elementos de la ME de manera eficiente. Si las topologías de las mallas son muy similares, se encuentran solapadas y resulta natural pensar a cada nodo de la MA como perteneciente a un elemento de la ME. Con las coordenadas locales de los nodos de la MA, los desplazamientos nodales en la ME y las funciones de forma del FEM, se puede obtener una matriz de transferencia **H** para calcular desplazamientos de los nodos de la MA.

Las transformaciones isoparamétricas utilizadas en el FEM permiten calcular de manera eficaz las coordenadas materiales de un punto a partir de sus coordenadas en el dominio computacional. Sin embargo, el cálculo inverso generalmente involucra la solución de sistemas de ecuaciones algebraicas no lineales, lo que puede realizarse con métodos iterativos de orden N^2 o N^3 (en dos y tres dimensiones respectivamente). Murti et ál. [6], [7] desarrollaron el IIM para realizar la transformación inversa, que resulta ser de orden N o N^2 (según se trate de dominios bi- o tridimensionales).

3. Técnicas basadas en el Método de Galerkin

3.1. Interpolación Consistente

Cebral y Löhner [8] sugieren interpolar presiones, y no fuerzas, sobre la ME. Para ello se parte de suponer que los campos de presiones en el modelo aerodinámico y en el estructural son iguales ($p_E(\mathbf{x}) = p_A(\mathbf{x})$). Suponiendo que el campo de presiones en cada modelo puede aproximarse utilizando los valores nodales de la presión y las funciones de forma correspondientes, y recurriendo al Método de Residuos Ponderados y al Método de Galerkin (con $W^i = N_E^i$) se llega a:

$$\int_\Gamma W^i(\mathbf{x})\, p_E(\mathbf{x})\, d\Gamma = \int_\Gamma W^i(\mathbf{x})\, p_A(\mathbf{x})\, d\Gamma\,, \tag{6}$$

$$\sum_{j=1}^{NE}\left(\underline{p}_E{}^j \int_\Gamma N_E{}^i N_E{}^j d\Gamma \right) = \sum_{k=1}^{NA}\left(\underline{p}_A{}^k \int_\Gamma N_E{}^i N_A{}^k d\Gamma \right) \qquad (i = 1,...,NE)\,, \tag{7}$$

$$\mathbf{M}_E\, \underline{\mathbf{p}}_E = \mathbf{M}_{AE}\, \underline{\mathbf{p}}_A\,, \tag{8}$$

Claramente $\mathbf{M}_E$ es la matriz de masa consistente del modelo estructural, la cual hay que invertir para poder calcular las presiones nodales en la ME. Utilizar la matriz de masa consistente produce una distribución de presiones poco suave. Para subsanar este problema se propone utilizar una técnica propia de los métodos de la Mecánica de Fluidos Computacional denominada Flux Corrected Transport (FCT).

En este método, la nueva ubicación de cada nodo de la MA es tal que se mantiene la posición relativa inicial del nodo respecto de la ME. Las velocidades, sin embargo, no se calculan a partir de la variación de la posición, sino realizando un planteo similar al utilizado para la interpolación de presiones, de manera que se conserve la energía. Para ello se parte de la igualdad $E_E(\mathbf{x}) = E_A(\mathbf{x})$, llegando a una expresión análoga a la Ec. (8):

$$\mathbf{M}_E\, \underline{\mathbf{E}}_E = \mathbf{M}_{AE}\, \underline{\mathbf{E}}_A\,, \tag{9}$$

donde $\underline{\mathbf{E}}_E$ y $\underline{\mathbf{E}}_A$ son los vectores con valores nodales de energía en la ME y la MA respectivamente, que se pueden escribir como función de los valores nodales de presión y velocidad utilizando el producto de Hadamard (representado con el símbolo ⊗), como:

$$\underline{\mathbf{E}}_E = \underline{\mathbf{p}}_E \otimes \underline{\mathbf{v}}_E \qquad , \qquad \underline{\mathbf{E}}_A = \underline{\mathbf{p}}_A \otimes \underline{\mathbf{v}}_A\,. \tag{10}$$

3.2. Refinamiento Común de Mallas

En estas técnicas basadas en el Método de Galerkin es necesario integrar numéricamente, sobre una superficie, productos de funciones de forma definidas en dominios esencialmente diferentes, y no está claro si debe utilizarse como superficie de integración la MA o la ME. Jiao y Heath [9] proponen hacer un refinamiento común de las mallas dato (en este caso la MA y la ME), de manera de obtener una nueva malla con elementos formados a partir de la subdivisión de elementos de las mallas dato, de tal manera que cada nuevo subelemento pertenezca completamente a un elemento de la ME y a uno de la MA.

De esta forma, si las $N_E{}^j$ y las $N_A{}^k$ son polinomios en los elementos de la ME y de la MA respectivamente, su producto es un polinomio en cada subelemento de la malla de refinamiento común y las integrales que dan la matriz $\mathbf{M}_{AE}$ de la Ec. (8) pueden calcularse exactamente sobre los subelementos con una regla de cuadratura adecuada.

Referencias

[1] M. Buhmann, *Radial Basis Functions,* Cambridge University Press, 2005.

[2] H. Wendland, *Scattered Data Approximation,* Cambridge University Press, 2005.

[3] A. Beckert y H. Wendland, *Multivariate interpolation for fluid-structure-interaction problems using radial basis functions,* Aerospace Science and Technology, 5 (2001), pp.125-134.

[4] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree,* Advances in computational Mathematics, 4 (1995), pp.389-396.

[5] P.C. Chen e I. Jadic, *Interfacing of fluid and structural models via innovative structural boundary element method,* AIAA Journal, 36 (1998), pp.282-286.

[6] V. Murti y S. Valliappan, *Numerical inverse isoparametric mapping in remeshing and nodal quantity contouring,* Computers & Structures, 22 (1986), pp.1011-1021.

[7] V. Murti, Y. Wang y S. Valliappan, *Numerical inverse isoparametric mapping in 3D FEM,* Computers & Structures, 29 (1988), pp.611-622.

[8] J.R. Cebral y R. Löhner, *Conservative load projection and tracking for fluid-structure problems,* AIAA Journal, 35 (1997), pp.687-692.

[9] X. Jiao y M.T. Heath, *Common-refinement-based data transfer between non-matching meshes in multiphysics simulations,* International Journal for Numerical Methods in Engineering, 61 (2004), pp.2420-2427.

MODELADO BASADO EN SUBDIVISIÓN: REFINAMIENTO

Diana Salgado[þ] y Liliana Castro[þ]

[þ]*Departamento de Matemática*
[þ]*Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica*
[þ]*Universidad Nacional del Sur, Bahía Blanca, B8000CPB, Argentina, dsalgado@uns.edu.ar, lcastro@uns.edu.ar*

Resumen: En este trabajo empleamos una técnica de subdivisión para calcular los puntos de control que subdividen a curvas polinómicas. Dados un polígono de control y la curva polinómica generada por ese polígono, mediante operaciones matriciales obtenemos polígonos a izquierda y a derecha, que aproximan a esa curva. Más precisamente, en este artículo aplicamos este método localmente, es decir, realizamos un refinamiento o subdivisión en alguno de los subpolígonos obtenidos.

Palabras clave: *refinamiento, multirresolución, subdivisión de curvas Beta-spline, computación gráfica*

1. INTRODUCCIÓN

Las escenas tridimensionales contienen modelos detallados con requerimientos altamente exigentes para una gran variedad de aplicaciones entre las cuales se pueden mencionar las basadas en Internet, las de visualización interactiva, los modelos 3D para ambientes virtuales complejos y los juegos multijugador entre muchas otras. Esto exige la transmisión y el rendering de modelos 3D para uso masivo lo que a su vez demandará mayor cantidad de gráficos 3D en la red. Esta situación motiva el desarrollo de modelos de superficies 3D y de volúmenes que satisfagan requerimientos tales como el uso efectivo del espacio en disco y del ancho de banda de la red así como también una reducción sustancial del tiempo de transferencia en la red. El principal objetivo dentro de esta línea de investigación es la obtención de mejores modelos de superficies 3D que soporten multirresolución y refinamiento bajo demanda.

2. MÉTODO DE MODELADO BASADO EN SUBDIVISIÓN

En los últimos años ha habido un rápido desarrollo de una teoría y de algoritmos para subdivisión básica de superficies y se ha alcanzado cierto nivel de madurez pero aún es necesaria mucha investigación.

La subdivisión permite crear una función mediante refinamientos repetidos de una función a trozos para producir una secuencia de funciones cada vez más detalladas que convergen a una función límite. Los métodos de subdivisión permiten generalizar el análisis por multirresoluciones asociado a las wavelets tradicionales y éste es uno de los motivos que hacen que estos métodos sean poderosos en computación gráfica ([5]).

Los métodos de subdivisión pueden ser usados para obtener una representación multirresolución si ésta puede ser evaluada local y progresivamente, permitiendo incrementar el nivel de detalle en regiones específicas. Esto hace posible la obtención de una máxima cantidad de detalle dentro de un tiempo de procesamiento acotado y/o la obtención de una mínima cantidad de detalle satisfaciendo una cota de error. Por otra parte, los métodos basados en multirresoluciones permiten visualizar un objeto con diferentes niveles de detalle bajo demanda del usuario o de la aplicación en particular; sin embargo, para ello es necesario que el método de representación permita extraer el detalle requerido. Es decir, que debe ser factible aplicar el método de subdivisión localmente para generar diferentes niveles de aproximación del objeto según sea el detalle a mostrar.

Hasta el momento, hemos trabajado en diversos métodos de subdivisión para generar curvas y superficies en forma paramétrica ([3],[4],[6]). En particular, hemos desarrollado un método de subdivisión para generar curvas y superficies Beta-spline cúbicas; el mismo está basado en polígonos de control a izquierda y a derecha. Actualmente estamos trabajando en la representación multirresolución para lo cual es necesario aplicar el método localmente ([2]). Comenzamos desarrollando el tema para la subdivisión de curvas, con la idea de realizar, en un futuro, la extensión al caso de superficies.

3. SUBDIVISIÓN DE CURVAS

La curva Beta-spline cúbica es una generalización de la B-spline cúbica uniforme, en la cual la propiedad de continuidad paramétrica de segundo grado (C^2) es reemplazada por la de continuidad geométrica (G^2) [3].

Un trozo de la curva Beta-spline cúbica está caracterizado por las funciones $B_0(t), ..., B_3(t)$ definidas por:

$$\begin{cases} B_0(t) = \frac{1}{\delta}[2\beta_1^3 - 6\beta_1^3 t + 6\beta_1^3 t^2 - 2\beta_1^3 t^3] \\ B_1(t) = \frac{1}{\delta}[(\beta_2 + 4\beta_1^2 + 4\beta_1) + (6\beta_1^3 - 6\beta_1)t - (3\beta_2 + 6\beta_1^3 + 6\beta_1^2)t^2 \\ \qquad +(2\beta_2 + 2\beta_1^3 + 2\beta_1^2 + 2\beta_1)t^3] \\ B_2(t) = \frac{1}{\delta}[2 + 6\beta_1 t + (3\beta_2 + 6\beta_1^2)t^2 - (2\beta_2 + 2\beta_1^2 + 2\beta_1 + 2)t^3)] \\ B_3(t) = \frac{1}{\delta}2t^3, \end{cases}$$

donde $\delta = \beta_2 + 2\beta_1^3 + 4\beta_1^2 + 4\beta_1 + 2$, β_1 es el parámetro de sesgo y β_2 es el parámetro de tensión, como se puede ver en [1].

Hemos desarrollado una técnica para calcular puntos de control que subdividen a curvas polinómicas, dadas como una combinación lineal de las funciones B_0, B_1, B_2 y B_3 y puntos de control determinados.

Hemos analizado el comportamiento de la subdivisión para distintos valores de los parámetros de sesgo β_1 y de tensión β_2 ([3],[4],[6]).

3.1. REFINAMIENTO. UN EJEMPLO: CURVA BETA-SPLINE CÚBICA

Continuando con esta línea de investigación, estamos trabajando en la aplicación del método arriba mencionado para representar el nivel de detalle de una curva y posteriormente poder extenderlo al caso de una superficie.

En principio consideramos un polígono de control formado por cuatro o más puntos y realizamos una subdivisión para obtener los respectivos subpolígonos a izquierda y a derecha (ver [7]). Luego seleccionamos uno de los subpolígonos obtenidos y refinamos localmente. La elección del subpolígono a subdividir depende de las necesidades del usuario o de la aplicación en particular.

Para ello fue preciso elaborar un algoritmo en MATLAB que permita refinar en algún lugar específico: alguno de los subpolígonos, es decir, poder aplicar el método localmente.

En la Figura 1 se puede apreciar un polígono de control formado por seis puntos y un paso de la subdivisión. En la Figura 2 se observa una subdivisión local en el segundo subpolígono de ese mismo polígono.

4. CONCLUSIONES Y TRABAJO FUTURO

En el ejemplo presentado aplicamos localmente un método de subdivisión que genera curvas Beta-spline cúbicas. Mostramos únicamente cómo se realiza la subdivisión, no graficamos la curva Beta-spline a la cual convergen los subpolígonos.

Como trabajo futuro, nos centraremos en la aplicación de este método para obtener superficies multirresolución, éste se llevará a cabo con el objeto de poder subdividir localmente una superficie dada, y así, poder representar el nivel de detalle de una superficie.

AGRADECIMIENTOS

Este trabajo fue parcialmente financiado con fondos del proyecto PGI 24/N020, SECyT, UNS.

REFERENCIAS

[1] B. BARSKY AND J. BEATTY, *Local control of bias and tension in Beta-splines*, ACM Transactions on graphics, 2(2)(1983), pp.109-134.

[2] L. BOSCARDÍN, G. PAOLINI, D. SALGADO, S. CASTRO AND L. CASTRO, *Nuevas Alternativas para el Modelado de Volúmenes*, XII Workshop de Investigadores en Ciencias de la computación (2010), pp. 312-316.

[3] L. CASTRO, S. CASTRO, S. KAHNERT, AND D. SALGADO, *Matrices de Subdivisión para curvas Beta-spline cúbicas*, Anales del V Workshop de Computación Gráfica, Imágenes y Visualización (XIII Congreso Argentino de Ciencias de la Computación), (2007), pp. 710-720.

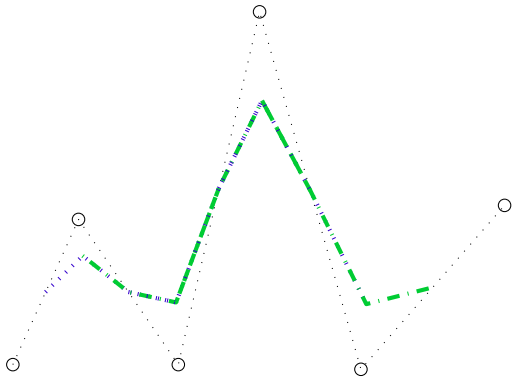


Figura 1: Polígono de control con seis puntos y subdivisi´on

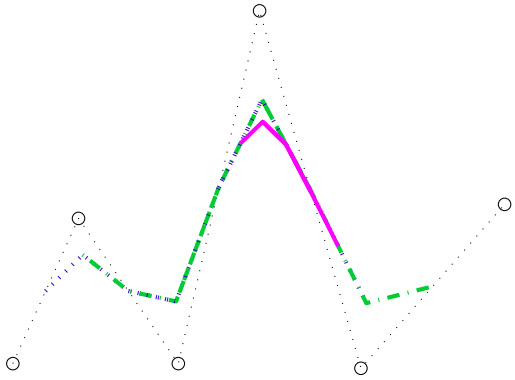


Figura 2: Refinamiento

[4] L. CASTRO, S. CASTRO AND D. SALGADO, *Subdivisión de superficies Beta-spline cúbicas*, Anales del VI Workshop de Computación Gráfica, Imágenes y Visualización (XIV Congreso Argentino de Ciencias de la Computación)(2008), pp. 1-8.

[5] C. CHUI AND J. DE VILLIERS, *Wavelets Subdivision Methods: GEMS for Rendering Curves and Surfaces* , CRC Press (August 23, 2010), 479 pages.

[6] D. SALGADO, L. CASTRO, S. CASTRO AND S. KAHNERT, *Subdivisión de curvas Beta-spline cúbicas*, Serie Mecánica computacional, Vol.XXVII (2008), pp. 3071-3080.

[7] R. GOLDMAN AND T. DEROSE, *Recursive subdivision without the convex hull property* , Computer Aided Geometric Design 3 (1986), pp. 247-265.

ESTIMACIÓN DE UN MARCO DE REFERENCIA CINEMÁTICO PARA LA ZONA DE DEFORMACIÓN ANDINA COLOMBIANA CON EL MÉTODO DE COLOCACIÓN POR CUADRADOS MÍNIMOS

Ana Milena Nemocón Romero[†], Saúl Becerra Ospina[♭] y Hernán Estrada B[♭]

[†]*Departamento de Física, Universidad Nacional de Colombia Bogotá D.C. Colombia, www.unal.edu.co*
[♭]*Departamento de Matemáticas, Universidad Nacional de Colombia Bogotá D.C. Colombia, www.unal.edu.co*

Resumen: Con la llegada de las técnicas geodésicas satelitales, fue necesario adoptar marcos de referencia modernos, que consideran las variaciones de las coordenadas producidas por la dinámica de la corteza terrestre mediante modelos globales de tectónica de placas. Sin embargo, en zonas de deformación es conveniente estimar campos continuos de velocidad que se ajusten a la dinámica local. En este trabajo, se presenta un modelo de interpolación usando colocación por cuadrados mínimos, para estimar un marco de referencia cinemático para la zona de deformación andina de Colombia. Los datos usados son las velocidades observadas de estaciones de rastreo permanente de las redes SIRGAS Y MAGNA-ECO.

Palabras clave: *Técnicas geodésicas, deformación, Colocación por cuadrados mínimos, Marco de referencia.*
2000 AMS Subject Classification: 86A60 - 86A30

1. INTRODUCCIÓN

En la superficie terrestre se pueden observar los efectos de la dinámica del planeta. Los sistemas de cordilleras, así como fracturas y fallas del terreno pueden explicarse con el fenómeno de la deriva continental. Con modelos de cinemática de placas geológicos y geofísicos, se puede explicar bastante bien el comportamiento tectónico. No obstante, algunos casos particulares como el extremo noroccidental de Sur América, donde convergen tres placas, Nazca (NA) y Caribe (CA) de tipo oceánicas y Sudamericana (SA) oceánica y continental, no se ajusta a un modelo de escala global [5]. Esta zona de contacto interplaca genera una zona de deformación considerada una microplaca según PB2002[1] denominada Bloque Norandino (ND). Estas deformaciones interplaca e intraplaca son modeladas por métodos físicos (Elemento Finito visco-elástico-plástico) o matemáticos [2].

Debido a las especiales condiciones tectónicas de Colombia, se presentan significativas variaciones en las coordenadas de estaciones. Por esta razón el Instituto Geográfico Agustín Codazzi IGAC, en la última década ha incrementando las estaciones geodésicas de rastreo continuo, red MAGNA-ECO[2], permitiendo obtener mediciones directas del desplazamiento de la corteza terrestre.

Este trabajo propone un modelo de interpolación para la zona de deformación Andina colombiana utilizando colocación por cuadrados mínimos, el cuál corresponde a un modelo horizontal continuo de velocidad. Las velocidades de entrada son obtenidas a partir del análisis de las soluciones semanales de 20 estaciones de funcionamiento permanente de la red geodésica nacional MAGNA-ECO y de las estaciones MARA (Maracaibo) y S061 (Quito) de la red regional SIRGAS, que por encontrarse en el bloque ND dentro de la orogenia andina fueron seleccionadas. Las velocidades horizontales son interpoladas para una grilla de $0,25^\circ \times 0,25^\circ$. El resultado se propone como marco nacional de referencia cinemático.

2. ESTIMACIÓN DE VELOCIDADES

Las coordenadas semanales de las estaciones permanentes son soluciones procesadas por IGS RNAAC-SIR[3]. El comportamiento en el tiempo de las coordenadas representa el cambio de posición de las estaciones permanentes debido a efectos dinámicos terrestres. Las velocidades de cada estación son estimadas con un

[1]PB2002, Modelo global de tectónica de placas, realizado por Peter Bird, 2002.

[2]MAGNA-ECO, Marco Geocéntrico Nacional de Referencia - Estaciones Continuas, densificación del Sistema de Referencia Geocéntrico para las Américas SIRGAS.

[3]IGS, International GNSS Service. RNAAC-SIR, Regional Network Associate Analysis Center-South America

modelo lineal para cada componente, Norte y Este. El modelo utilizado es

$$y(t_i) = a + bt_i, \text{ para } i = 1, 2, \ldots, m \tag{1}$$

donde t_i son las épocas de solución en unidades de años, $y(t_i)$ es la posición para la época t_i, b son las velocidades estimadas y a es la coordenada de la época t_0 [6].

3. COLOCACIÓN POR CUADRADOS MÍNIMOS

Colocación por cuadrados mínimos LSCM es desarrollada por Moritz H. [4] como un método para la determinación del campo de gravedad anómalo por mediciones geodésicas de diferentes clases. En LSMC se incluye un término estocástico denotado t, que permite incluir en la estimación de una señal efectos aleatorios que no son tenidos en cuenta en el método convencional de cuadrados mínimos LSM.

En el caso del análisis del campo gravitacional terrestre, Moritz muestra que la parte estocástica del campo puede ser considerada usando un tipo especial de funciones de covarianza sobre la esfera. Pero LSCM ya ha sido empleado en muchos y variados trabajos de geodesia y fotogrametría, en donde siempre se tienen mediciones en algunos puntos de observación (datos discretos) y a los cuales se les puede establecer una estructura espacial de correlación, con el objetivo de realizar predicción. En este trabajo se tienen velocidades de estaciones geodésicas, que se asumen como variables regionalizadas definidas por un proceso estocástico, y por su característica de continuidad espacial pueden interpolarse.

El vector de velocidades observadas es

$$V_{obs} = \left[v_{n_1}\ v_{e_1}\ \ldots, v_{n_q}\ v_{e_q}\right]^T, \tag{2}$$

donde v_{n_i} es la velocidad de la componente Norte de la $i-$ésima estación y v_{e_i} la Este. Similarmente el vector de señales a predecir en h puntos es

$$\widehat{V}_{pred} = [\widehat{v}_{n_1}\ \widehat{v}_{e_1}\ \ldots, \widehat{v}_{n_h}\ \widehat{v}_{e_h}]^T. \tag{3}$$

El modelo de predicción utilizado es

$$V_{pred} = C_{pq}C_{qq}^{-1}V_{obs}, \tag{4}$$

siendo C_{qq} la matriz de autocovarianza de los vectores de velocidad observados y C_{pq} la matriz de covarianza entre los vectores de velocidades observadas y predecidas. Los elementos de las matrices son obtenidos de funciones de covarianza isotrópicas determinadas empíricamente.

4. ANÁLISIS ESTRUCTURAL

Para realizar la predicción se requiere las matrices de covarianza cruzada entre velocidades a predecir y observadas, además de la matriz de autocovarianza; para esto se necesitan dos funciones de covarianza, una para cada componente Norte y Este. Entonces el modelo de predicción (4) requiere para su estimación una etapa preliminar que corresponde con el conocido análisis estructural que se realiza en geoestadística cuyo objetivo es determinar la dependencia espacial de la variable en estudio y se realiza sobre la información muestral.

Los covariogramas experimentales son calculados para clases de distancias esféricas, el estimador utilizado es

$$C(d) = \frac{\sum(z(x) - m)(z(x + d) - m)}{N}, \tag{5}$$

donde m es la media, N es el númeto de estaciones, z es la variable aleatoria, en este caso son las velocidades Norte v_n y Este v_e.

Con el covariograma experimental se ajusta un modelo de covarianza exponencial, cuya expresión se da por:

$$\gamma(d) = C(0)e^{-ad}, \tag{6}$$

con $C(0) = \sigma$ la varianza. Entonces para poder estimar las matrices de covarianza requeridas se debe estimar valores adecuados de los parámetros $C(0)$ y a.

5. Reducción de tendencia y Modelo de Velocidades

En la estimación de las velocidades para cada estación, se observa una tendencia positiva en la componente Norte y aunque en menor magnitud también hay tendencia positiva en la componente Este, lo que es consistente según el modelo PB2002[1] el cual sugiere un desplazamiento con tendencia noreste. Sin embargo la magnitud de las velocidades observadas son mayores a las calculadas por el polo de rotación propuesto por Bird. Para modelar por colocación es necesario remover la tendencia en los datos y de acuerdo con la teoría, la rotación del bloque norandino puede expresarse a través de las coordenadas geográficas de un polo de Euler (PE) y una velocidad angular:

$$\begin{aligned} \Phi = &\quad -2,432549^\circ \\ \Lambda = &\quad 132,8983627^\circ \\ \Omega = &\quad 0,4157318(^\circ/Ma). \end{aligned}$$

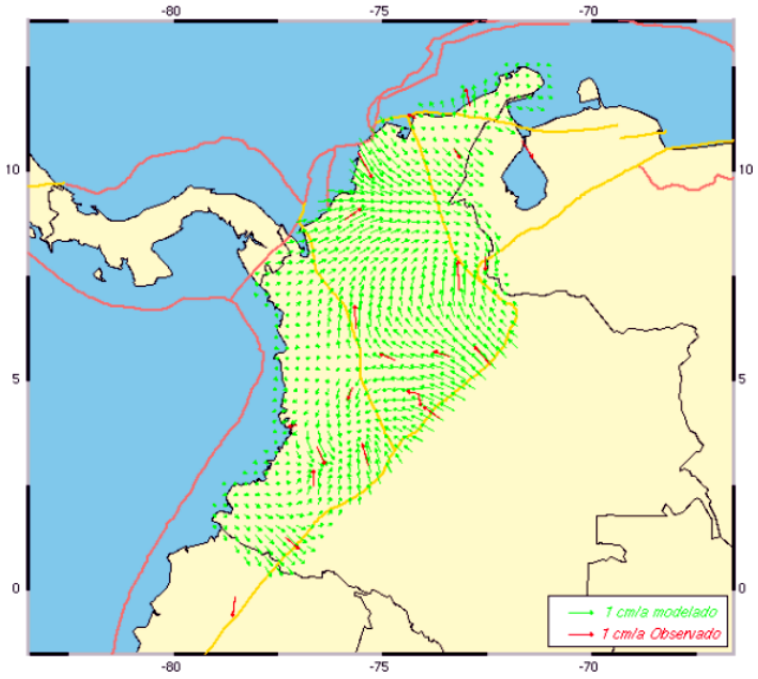


Figura 1: En rojo las velocidades residuales de las estaciones, datos de entrada. En verde velocidades predichas.

Teniendo las velocidades estimadas y las calculadas con PE, se encuentran finalmente las velocidades residuales de las 22 estaciones seleccionadas y se les aplica la interpolación del modelo de predicción y filtrado (4).

El resultado, corresponde al modelo de la Figura 1 donde se observan patrones de deformación que pueden relacionarse detalladamente con el Marco Geológico, U.S. Geological Survey (USGS) and Gabriel Paris, 2000 y con el estudio de Piedras-Girardot [3], el cuál discrimina las rocas en la deformación de la zona Andina, coincidiendo las variaciones de dirección y magnitud en las velocidades del modelo con las áreas blandas y rígidas de los materiales encontrados en este estudio.

Los tipos de interacciones entre las placas NA, SA, CA, y el bloque Panamá con el bloque ND, dan una explicación a las deformaciones que pueden observarse en el modelo de velocidades realizado. La zona de contacto de la placa SA con el bloque ND por ejemplo, presenta fallas inversas o de corrimiento (según Marco Geológico), este choque se refleja en las velocidades residuales estimadas con una fuerte dirección noreste para la parte alta y sureste para la baja.

6. CONCLUSIONES

La región norandina es una zona de deformación y si bien las estaciones de rastreo permanente muestran el comportamiento de algunos puntos discretos, se vuelve obsoleto asumir que en un levantamiento diferencial GPS apoyado en estaciones de funcionamiento continuo, la velocidad de los puntos rastreados son las mismas de las estaciones MAGNA-ECO utilizadas como estaciones base en el levantamiento.

Es recomendable estudiar modelos físicos provenientes de la mecánica de medios continuos, para analizar desde otro enfoque la deformación de la zona de estudio.

LSCM, permite incluir en un solo modelo la estimación de las dos componentes horizontales de la velocidad de un punto, por tal razón difiere de los métodos convencionales de predicción espacial como Kriging.

REFERENCIAS

[1] BIRD P., *An updated digital model of plate boundaries.*, Department of Earth and Space Sciences, University of California, Los Angeles, California 90095, USA, marzo de 2003.

[2] DREWES H., HEIDBACH O., *International Association of Geodesy Symposia Volume 128.*, chapter Deformation of the South American Crust Estimated from Finite Element and Collocation Methods, SpringerLink junio de 2006.

[3] MONTES C., HATCHER R. D., RESTREPO P., *Tectonic reconstruction of the northern Andean blocks: Oblique convergence and rotations derived from the kinematics of the PiedrasGirardot area, Colombia.*, Tectonophysics, Andean Geodynamics, Volume 399, Issues 1-4, 27 April 2005, Pages 221-250.

[4] MORITZ H., *Advanced Physical geodesy.*, Herbert Wichmann Verlag Karlsruhe, 1980.

[5] NAVA A., *La Inquieta Superficie Terrestre.*, Fondo de Cultura Económica, 2 edición, mexico 1998.

[6] NIKOLAIDIS R., *Observation of geodetic and seismic deformation with the global positioning system.*,Ph.D. thesis, University of California, San Diego, 2002.

MONOTONE AND NONMONOTONE TRUST-REGION-BASED-ON ALGORITHMS FOR THE LARGE UNCONSTRAINED MINIMIZATION PROBLEMS

M.C. MACIEL[♭], M.G. MENDONÇA[†] and A.B. VERDIELL[♭]

[♭]*Departamento de Mátematica, Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca, Argentina, immaciel@criba.edu.ar, averdiel@criba.edu.ar*

[†]*Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Comodoro Rivadavia, Argentina, mendonca@ing.unp.edu.ar*

Abstract: Two trust regions algorithms for unconstrained nonlinear optimization problems are presented: a monotone and a nonmonotone one. Both of them solve the trust region subproblem by the spectral projected gradient (SPG) method proposed by Birgin, Martínez and Raydán in 2000. SPG is a nonmonotone projected gradient algorithm for solving large-scale convex-constrained optimization problems. It combines the classical projected gradient method with the spectral gradient choice of steplength and a nonmonotone line search strategy. The simplicity and rapid convergence of this scheme fits nicely with globalization techniques based on the trust region philosophy, for large-scale problems. In the nonmonotone algorithm the trial step is evaluated by acceptance by a rule which can be considered a generalization of the well known fraction of Cauchy decrease condition and a generalization of the nonmonotone line search proposed by Grippo, Lampariello y Lucidi in 1986. Convergence properties and extensive numerical results are presented establishing the robustness and efficiency of the algorithms.

Keywords: *Projected gradients, nonmonotone line search, large scale problems, trust-region subproblems, spectral gradient method.*
2000 AMS Subject Classification: 49M07 - 49M10 - 65K - 90C06 - 90C20.

1 INTRODUCTION

In this work two algorithms for solving the unconstrained minimization problem

$$\min f(x) \qquad \text{subject to } \; x \in \mathbb{R}^n, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is at least continuously differentiable are disccused. Boths algorithms are based on the trust reion approach and the trust region subproblems are approximately solved by the spectral projected gradient method [1]. In the classical case (the monotone case) the trial step is chosen to satisfy the fraction of Cauchy deacrease condition

$$q_k(0) - q_k(s_k) \geq \sigma_0[q_k(0) - q_k(s_k^{\mathsf{CP}})], \tag{2}$$

for some $\sigma_0 \in (0, 1)$ and s_k^{CP} is the solution of the problem $\min\limits_{t\geq 0} q_k(x_k - tg_k)$ inside of the trust region of radius δ_k.

The nonmonotonicity of the TR-region algorithm is characterized by the way the trial step is evaluated [4, 7, 10]. In this work the following scheme is adoptes to evaluated the step

$$f_{k+1} \leq \max_{j=0,\ldots,N} f_{k-j} + \sigma \left(\nabla f_k^T s + \frac{1}{2} s_k^T \nabla^2 f_k s \right).$$

It is a generalization of the condtion (**??**) and the nonmonotone linesearch introduced by Grippo, Lampariello y Lucidi in 1986 [5].

2 THE ALGORITHMS

Algorithm 1 (Monotone algorithm)
Given the current iterate $x_k \in \mathbb{R}^n$, δ_k, $g_k = \nabla f_k$, $H_k \in \mathbb{R}^{n\times n}$, symmetric with $\|H_k\| \leq \beta$, if x_k is a stationary point of the problem, the algorithm terminates, otherwise, the following steps allow to obtain the next iterate, x_{k+1}.

Step 1. *(Compute a trial step)*
Find s as solution of the following trust region subproblem

$$\min q(s) \equiv \frac{1}{2} s^T H s + g^T s \;\; \textit{subject to} \;\; \|s\| \leq \delta_k, \tag{3}$$

by using the spectral gradient method described in [1].

Step 2. *(Evaluate the trial step)*
Define the quantities

$$\begin{aligned} ared_k(s_k, \delta_k) &= f(x_k) - f(x_k + s_k) \\ pred_k(s_k, \delta_k) &= f(x_k) - q_k(s_k) \\ \rho_k &= \frac{ared_k}{pred_k} \end{aligned} \tag{4}$$

and evaluate the step for acceptance as follows:
Given the constants $\eta_1, \eta_2, \alpha_1, \alpha_2$ such that $0 < \eta_2 < \eta_1 < 1$ and $0 < \alpha_1 < 1 < \alpha_2$,

- *If $\rho \geq \eta_1$ then $x_{k+1} \leftarrow x_k + s_k$ and choose $\delta_+ \in (\delta_k, \alpha_2 \delta_k]$.*
- *If $\eta_2 < \rho < \eta_1$ then $x_{k+1} \leftarrow x_k + s_k$ and choose $\delta_+ \in [\alpha_1 \delta_k, \delta_k]$.*
- *If $\rho \leq \eta_2$ then the step is rejected, $x_{k+1} \leftarrow x_k$ and reduce the trust region radius according with $\delta_{k+1} \in [\min\{\delta_{min}, \alpha_1 \delta_k\}, \delta_k)$.*

Step 3. *Update all the information and go to step 1.*

2.1 THE NONMONOTONE ALGORITHM

The mnonmonotone algorithm coincides with Algorithm 1 except at the evaluation of the trial step. Denoting $f_{max} = \max\limits_{0 \leq j \leq m(k)} f_{k-j}$ and $q_{max} = f_{max}$ it is describe below.

Algorithm 2

Step 2. *Evaluate the step and update the trust region radius*

$$\widetilde{\rho}_k = \frac{f_{max} - f(x_k + s_k)}{q_{max} - q_k(s_k)}$$

3 THEORETICAL DISCUSSION

We begin by stating the standard assumptions under which the well definition propierty and global convergence of the algorithm 2 is proved.

A1. There exists an open convex $\Omega \subseteq \mathbb{R}^n$ such that, for all k, $x_k, x_k + s_k \in \Omega$.

A2. $\nabla f \in Lip_\gamma(\Omega)$.

A3. The sequence of Hessian aproximation $\{H_k\}$ is uniformly bounded: there exists $\beta_0 \geq 0$ such that $\|H_k\| \leq \beta_0$ for all k.

A4. f is bounded below on the compact set $\Omega_0 = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$.

Let us introduce the amounts:

$$ared_k^{nm} = f_{max} - f(x_k + s_k) = \max_{0 \leq j \leq m(k)} f(x_{k-j}) - f(x_k + s_k),$$

is the descent of the objective function at k-th iteration with respect to the M previous values of f.

$$\begin{aligned} pred_k^{nm} &= q_{\ell(k)}(0) - q_k(s_k) = f(x_{\ell(k)}) - q_k(s_k) \\ &\geq f(x_k) - q_k(s_k) = q_k(0) - q_k(s_k) \\ &= pred_k, \end{aligned} \tag{5}$$

is the predicted decrease of the quadratic model $q_k(s)$ at the iteration k with respect to the value of the quadratic model around $x_{\ell(k)}$, where $\ell(k)$ is an integer such that: $k - m(k) \leq \ell(k) \leq m(k)$ and $f(x_{\ell(k)}) = f_{max} = \max_{0 \leq j \leq m(k)} f(x_{k-j})$. The ratio

$$\widetilde{\rho}_k = \frac{ared_k^{nm}}{pred_k^{nm}}$$

measures the agreement between the quadratic model an the objective function.

The following is a well known result dues to Powell [9]. A proof and comments can be found in [8] and [3].

Lemma 1 *If s_k satisfies the condition (2) then*

$$pred_k \geq \sigma_0 \|g_k\| \min\left\{\delta_k, \frac{\|g_k\|}{\|H_k\|}\right\}, \tag{6}$$

with $\sigma_0 > 0$.

Now we will present some technical results in order to obtain the well definition an convergence theorems.

Lemma 2 *Assuming (A1)-(A3) we have*

$$|ared_k^{nm} - pred_k^{nm}| \leq \beta_0 \|s_k\|^2. \tag{7}$$

Lemma 3 *Assuming (A1)-(A3), and $g_k \neq 0$ we have*

$$pred_k^{nm} \geq \|g_k\| \|s_k\| \frac{\sigma_0}{2} \min\left\{1, \frac{|\alpha_k|}{\beta}\right\}, \tag{8}$$

where α_k is the inverse Rayleigh quotient, i.e., $\alpha_k^{-1} = \dfrac{\langle s_{k-1}, s_{k-1}\rangle}{\langle s_{k-1}, y_{k-1}\rangle}$.

The following result guarantees the well definition of the algorithm.

Theorem 1 *Under assumptions (A1)-(A3), if the Algorithm 2 does not terminate at x_k, then it must have $\widetilde{\rho}_k \geq \eta_1$ after a finite number of reduction of the trust region radius.*

Finally we establish the global convergence theorem.

Theorem 2 *Assuming (A1)-(A4) we have*

$$\liminf_{k \to \infty} \|g_k\| = 0.$$

The proofs of Lemmas 2 and 3 and Theorems 1 and 2 can be found in [6].

4 NUMERICAL RESULTAS

ACKNOWLEDGMENTS

This work has been partially supported by Universidad Nacional del Sur, Project 24/069.

REFERENCES

[1] E. BIRGIN, J.M. MARTINEZ, and M. RAYDAN. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000.

[2] E. BIRGIN, J.M. MARTINEZ, and M. RAYDAN. Algorithm 813: Spg software for convex-constrained optimization. *ACM transactions on Mathematical Software* , 27:340–349, 2001.

[3] A. CONN, N.I.M. GOULD, and Ph. TOINT. *Trust-Region Methods*. SIAM-MPS, Philadelphia, Pennsylvania, 2000.

[4] N.Y. DENG, Y. XIAO, and F.J. ZHOU. A nonmonotonic trust-region algorithm. *Journal of Optimization Theory and Applications*, 76:259–285, 1993.

[5] L. GRIPPO, F. LAMPARIELLO, and S. LUCIDI. A nonmonotone line search technique for newton's method. *SIAM J. Numer. Anal.*, 23:707–716, 1986.

[6] M. DE G. MENDONÇA, M.C. MACIEL, and A.B. VERDIELL . Monotone and Nonmonotone Trust-Region-Based-on Algorithms for the Large Unconstrained Optimización Problems. Submitted to *Journal of Optimization Theory and Applications* , 2010.

[7] J. MO, K. ZHANG, and Z. WEI. A nonmonotone trust region method for unconstrained optimization. *Applied Mathemathics and Computation*, 171:371–384, 2005.

[8] J.J. MORE. Recent developments in algorithms and software for trust region methods. In A. Bachem, M. Grotschel, and B. Korte, editors, *Mathematical Programming: The state of the art* , pages 258–287. Springer-Verlag, 1984.

[9] M.J.D. POWELL. Convergence properties of a class of minimization algorithms. In O.L.Mangasarian, R.R. Meyer, and S.M. Robinson, editors, *Nonlinear Programming 2* , pages 1–27. Academic Press, New York, 1975.

[10] Ph.L. TOINT. Non-monotone trust-region algorithms for nonlinear optimization subject to convex constraints. *Mathematical Programming*, 77:69–94, 1997.

UN MÉTODO QUASI-NEWTON SIN DERIVADAS PARA RESOLVER SISTEMAS NO LINEALES INDETERMINADOS CON RESTRICCIONES DE COTAS EN LAS VARIABLES

N. Echebest[♭], M. L. Schuverdt[♭,†] y R. P. Vignau [♭]

[♭]*Departamento de Matemática. Facultad de Ciencias Exactas. UNLP, La Plata, Argentina*
[†]*CONICET, Argentina, www.conicet.gov.ar opti@mate.unlp.edu.ar, schuverd@mate.unlp.edu.ar, vignau@mate.unlp.edu.ar*

Resumen: En este trabajo se presenta una extensión del método Quasi-Newton para resolver sistemas indeterminados de ecuaciones no lineales sin derivadas, denominado DF-QNB en [2], para el caso en el que las variables deben satisfacer restricciones de cotas. El método propuesto utiliza la fórmula de Broyden de rango 1 para aproximar al Jacobiano y realiza una búsqueda lineal sin derivadas que es una modificación de la definida en [2] que combina la estrategia de Grippo, Lampariello y Lucidi [3] con la de Li y Fukushima [5]. Se presentan resultados de convergencia del método propuesto y experimentos numéricos.

Palabras clave: *Ecuaciones no lineales. Quasi-Newton. Búsqueda lineal sin derivadas.*
2000 AMS Subject Classification: 65H10

1. INTRODUCCIÓN

El conjunto factible de muchos problemas de programación no lineal usualmente se representa por un conjunto de ecuaciones no lineales, $F(x) = 0$, donde $F : \mathbb{R}^n \to \mathbb{R}^m$. Eventualmente también aparece la necesidad de encontrar un punto factible sujeto a satisfacer restricciones de cotas en las variables. En este trabajo consideramos el problema de determinar una solución $x^* \in \mathbb{R}^n$ que verifique $F(x) = 0$, sujeto a satisfacer $x^* \in \Omega = \{x \in \mathbb{R}^n \ : \ l \leq x \leq u\}$, siendo F una función continuamente diferenciable y $m \leq n$. Estamos interesados en sistemas para los cuales la matriz Jacobiana de F, denotada por $J(x)$, no está disponible o su cálculo requiere demasiado costo computacional. Problemas con tales características provienen de aplicaciones de la Ingeniería, la Física o la Economía.

El objetivo de este trabajo es extender la aplicabilidad del método DF-QNB [2], que resuelve sistemas no lineales indeterminados sin el cálculo de derivadas, al caso donde las variables tienen restricciones de cotas. En el método DF-QNB, la dirección de búsqueda se calcula como una solución aproximada de un sistema lineal utilizando la fórmula de Broyden de rango 1 como aproximación del Jacobiano. La búsqueda lineal sin derivadas utilizada en [2] es la definida en el algoritmo DF-SANE [4].

En este trabajo la dirección de búsqueda d_k en un iterado x_k, se calcula de manera semejante a la presentada en [2] con la salvaguarda de considerar direcciones tales que $x_k + d_k$ se mantenga en la caja Ω. Una vez calculada la dirección, la búsqueda lineal que se realiza es una combinación de la estrategia de Grippo, Lampariello y Lucidi [3] con la de Li y Fukushima [5].

La función de mérito que se utiliza es la usual para este problema, $f(x) = \frac{1}{2}\|F(x)\|^2$, $f : \mathbb{R}^n \to \mathbb{R}$, y la condición de descenso requerida en la búsqueda lineal es la siguiente

$$f(x_k + \alpha_k d_k) \leq \max_{0 \leq j \leq M-1} f(x_{k-j}) + \eta_k - \gamma \alpha_k^2 \|d_k\|^2 \tag{1}$$

donde M es un entero no negativo, $0 < \gamma < 1$ y $\sum_{k=0}^{\infty} \eta_k = \eta < \infty$, $\eta_k > 0$.

Como se puede observar la condición anterior no utiliza derivadas y genera un proceso de descenso no monótono [2].

A continuación se presenta el algoritmo básico de búsqueda lineal no monótona que se utiliza en el algoritmo principal para encontrar un nuevo punto en Ω en la dirección de búsqueda d_k a partir de x_k.

Algoritmo 1 *Búsqueda lineal no monótona*

Dados $d_k \in \mathbb{R}^n$, $0 < \tau_{min} < \tau_{max} < 1$, $0 < \gamma < 1$, $M \in \mathbb{N}$, $\{\eta_k\}$ *tal que* $\eta_k > 0$ *y* $\sum_{k=0}^{\infty} \eta_k = \eta < \infty$

Paso 1: Calcular $\overline{f}_k = \max\{f(x_k), \ldots, f(x_{max\{0,k-M+1\}})\}$

$\alpha = 1$

Paso 2:

Si $f(x_k + \alpha d_k) \leq \overline{f}_k + \eta_k - \gamma\alpha^2\|d_k\|^2$,

define $\alpha_k = \alpha$, $x_{k+1} = x_k + \alpha_k d_k$

sino

elige $\alpha_{new} \in [\tau_{min}\alpha, \tau_{max}\alpha]$,

$\alpha = \alpha_{new}$ *y va al Paso 2.*

A partir de este procedimiento se pueden demostrar los siguientes resultados:

Proposision 1 *El Algoritmo 1 está bien definido.*

Proposision 2 *Para todo* $k = 1, 2, \ldots$ *se considera* $U_k = \max\{f(x_{(k-1)M+1}), \ldots, f(x_{kM})\}$ *y se define* $\nu(k) \in \{(k-1)M + 1, \ldots, kM\}$ *el índice para el cual* $f(x_{\nu(k)}) = U_k$. *Entonces para todo* $k = 1, 2, \ldots$

$$f(x_{\nu(k+1)}) \leq f(x_{\nu(k)}) + \eta$$

donde $\eta = \sum_{i=0}^{\infty} \eta_i$. *Además* $\lim_{k\to\infty} \alpha^2_{\nu(k)-1}\|d_{\nu(k)-1}\|^2 = 0$.

Las demostraciones de las proposiciones previas son similares a las presentadas en [2].

De ahora en adelante consideramos el conjunto de índices,

$$K = \{\nu(1) - 1, \nu(2) - 1, \nu(3) - 1, \ldots\}. \tag{2}$$

Como dijimos previamente utilizamos la fórmula de Broyden de rango 1, para adaptar las matrices en el procedimiento iterativo cuando $s_k = x_{k+1} - x_k \neq 0$ e $y_k = F(x_{k+1}) - F(x_k)$:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{s_k^T\ s_k}. \tag{3}$$

Si $s_k = 0$, se define $B_{k+1} = B_k$.

A continuación presentamos el algoritmo DF-QNBC para resolver sistemas indeterminados de ecuaciones no lineales con cotas en las variables.

Algoritmo 2 *DF-QNBC*

Dados $x_0 \in \Omega$, $F(x_0)$, $0 < \gamma < 1$, $0 \leq \theta_0 < \bar{\theta} < 1$, $0 < \tau_{min} < \tau_{max} < 1$, $0 \leq \epsilon < 1$, $\Delta > 0$, $ind = 0$, $0 < imax \in \mathbb{N}$.

$k \leftarrow 0$.

Paso 1: Si $\|F(x_k)\| \leq \epsilon \max\{1, \|F(x_0)\|\}$ *termina el procedimiento.*

Paso 2: Cálculo de la matriz B_k*:*

Si $k = 0$ *o* $ind = imax$*, calcula* B_k *como una aproximación del* $J(x_k)$ *por diferencias finitas.*

Si $k > 0$ *e* $ind < imax$*, actualiza* B_k *usando (3).*

Paso 3: Cálculo de la dirección d_k*:*

Paso 3.1: Resuelve

$$B_k d + F(x_k) = 0, \quad x_k + d \in \Omega \qquad \|d\| \leq \Delta. \tag{4}$$

Si tal dirección d *existe, define* $d_k = d$*,* $\theta_{k+1} = \theta_k$ *y va al Paso 4.*

Paso 3.2: Halla una solución aproximada del problema

$$\min_{l - x_k \leq d \leq u - x_k} \|B_k d + F(x_k)\|.$$

Si d *satisface*

$$\|B_k d + F(x_k)\| \leq \theta_k \|F(x_k)\| \quad y \quad \|d\| \leq \Delta \tag{5}$$

define $d_k = d$*,* $\theta_{k+1} = \theta_k$*,* $ind = 0$*, va al Paso 4.*
En caso contrario, define $d_k = 0$*,* $x_{k+1} = x_k$*,* $\theta_{k+1} = \frac{\theta_k + \bar{\theta}}{2}$*,*
si $ind < imax$*,* $ind = ind + 1$*, va al Paso 5.*
Si $ind = imax$*, define* $\bar{\theta} = \frac{\bar{\theta}+1}{2}$*,* $ind = 0$*, va al Paso 5.*

Paso 4: Halla α_k *y* $x_{k+1} = x_k + \alpha_k d_k$ *usando el Algoritmo 1.*

Paso 5: Actualiza $k \leftarrow k + 1$ *y va al Paso 1.*

La idea principal del algoritmo propuesto es resolver el sistema lineal en forma precisa cuando es posible y si no es posible en forma aproximada de la manera considerada en [6].

En los siguientes teoremas se establecen las hipótesis necesarias para obtener los resultados de convergencia global.

Teorema 1 *Asumimos que el Algoritmo 2 genera una sucesión infinita* $\{x_k\}$*. Supongamos que existe* $k_0 \in \mathbb{N}$ *tal que, para todo* $k \geq k_0, \theta_k < \widehat{\theta} < 1$ *y*

$$\lim_{k \to \infty} \langle (B_k - J(x_k)) d_k, F(x_k) \rangle = 0 \tag{6}$$

entonces todo punto límite de $\{x_k\}_{k \in K}$ *es una solución del sistema* $F(x) = 0$ *con* $x \in \Omega$*, donde* K *está dado por (2).*

Teorema 2 *Asumiendo que en el Algoritmo 2 el parámetro* $\bar{\theta}$ *crece infinitas veces.*
Definiendo $K_* = \{k \in \mathbb{N} : \theta_{k+1} > \theta_k\}$ *y asumiendo que*

$$\lim_{k \in K_*} \|B_k - J(x_k)\| = 0, \tag{7}$$

entonces todo punto límite x^* *de la sucesión* $\{x_k\}_{k \in K_*}$ *es una solución del problema original o es un minimizador global de* $\|F(x^*) + J(x^*)(x - x^*)\|$ *sujeto a* $x \in \Omega$*.*

Un punto x^* que es un minimizador global de la función $\|F(x^*)+J(x^*)(x-x^*)\|$ sujeto a $x \in \Omega$ puede ser visto como la solución del problema de mínimos cuadrados

$$\min_{x\in\Omega} \|A(x-x^*)-b\| \tag{8}$$

donde $A = J(x^*)$ y $b = -F(x^*)$. La función lineal es un modelo afín de la función F alrededor de x^*. En general no podemos esperar encontrar $x^* \in \Omega$ tal que $F(x^*) = 0$ ya que este problema podría no tener solución. Tampoco podemos esperar encontrar $x^* \in \Omega$ tal que $F(x^*) + J(x^*)(x - x^*) = 0$ ya que este es un sistema de ecuaciones lineales indeterminado y $J(x^*)$ podría no tener rango completo. Debido a esto, parece razonable encontrar un minimizador global de (8) cuando el problema no tiene soluciones.

Finalmente, para analizar el desempeño del método propuesto desde el punto de vista práctico, se implementó el Algoritmo 2 en lenguaje FORTRAN y se seleccionó un conjunto de problemas de la literatura para analizar su comportamiento. Tales resultados muestran un comportamiento promisorio del algoritmo presentado.

REFERENCIAS

[1] J.E. DENNIS, R.B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs, 1983.

[2] N.ECHEBEST, M.L. SCHUVERDT, R. P. VIGNAU, *Two derivative-free methods for solving underdetermined nonlinear systems of equations.* por aparecer en Computational and Applied Mathematics, 2010.

[3] L. GRIPPO, F. LAMPARIELLO, S. LUCIDI, *A nonmonotone line search technique for Newton's method.* SIAM Journal on Numerical Analysis 23 (1986), pp.707-716.

[4] W. LA CRUZ, J.M. MARTÍNEZ, M. RAYDAN, *Spectral residual method without gradient information for solving large-scale nonlinear systems of equations.*. Mathematics of Computation 75 (2006), pp. 1429-1448.

[5] D.H. LI, M. FUKUSHIMA, *A derivative-free line search and global convergence of Broyden -like method for nonlinear equations*. Optimization Methods and Software 13 (2000), pp. 181-201.

[6] J.M. MARTÍNEZ, *Quasi-Inexact Newton methods with global convergence for solving constrained nonlinear systems*. Nonlinear Analysis 30 (1997), pp. 1-7.

PRIMAL SUPERLINEAR CONVERGENCE RESULTS FOR SOME NEWTONIAN METHODS

D. Fernández[♭], A. F. Izmailov[†] and M. V. Solodov[‡]

[♭]*FaMAF, Universidad Nacional de Córdoba, Medina Allende s/n, 5000 Córdoba, Argentina, dfernandez@famaf.unc.edu.ar, www.famaf.unc.edu.ar*

[†]*VMK Faculty, OR Department, Moscow State University, MSU, Uchebniy Korpus 2, Leninskiye Gory, 119991 Moscow, Russia, izmaf@ccas.ru, www.cs.msu.su*

[‡]*IMPA–Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, 22460-320 Rio de Janeiro-RJ , Brazil, solodov@impa.br, www.impa.br*

Abstract: As is well known, Q-superlinear or Q-quadratic convergence of the primal-dual sequence generated by an optimization algorithm does not, in general, imply Q-superlinear convergence of the primal part. Primal convergence, however, is often of particular interest. For the sequential quadratic programming (SQP) algorithm, local primal-dual quadratic convergence can be established under the assumptions of uniqueness of the Lagrange multiplier associated to the solution and the second-order sufficient condition. At the same time, previous primal Q-superlinear convergence results for SQP required strengthening of the first assumption to the linear independence constraint qualification. We show that this strengthening of assumptions is actually not necessary. Our study is performed for a general perturbed SQP framework which covers, in addition to SQP and quasi-Newton SQP, the linearly constrained (augmented) Lagrangian (LCL) methods and the sequential quadratically constrained quadratic programming (SQCQP) methods.

Keywords: *Newton methods, sequential quadratic programming, linearly constrained Lagrangian methods*
2000 AMS Subject Classification: 90C30, 90C33, 90C55, 65K05

1 INTRODUCTION

Consider the mathematical programming problem

$$\begin{array}{ll}\text{minimize} & f(x)\\ \text{subject to} & h(x)=0,\ g(x)\le 0,\end{array} \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$, $h:\mathbb{R}^n\to\mathbb{R}^l$ and $g:\mathbb{R}^n\to\mathbb{R}^m$ are twice differentiable near the point of interest $\bar{x}\in\mathbb{R}^n$, and their second derivatives are continuous at $\bar{x}$. Stationary points of problem (1) and the associated Lagrange multipliers are characterized by the Karush–Kuhn–Tucker (KKT) optimality system

$$\frac{\partial L}{\partial x}(x,\,\lambda,\,\mu)=0,\quad h(x)=0,\quad \mu\ge 0,\quad g(x)\le 0,\quad \langle\mu,\,g(x)\rangle=0, \tag{2}$$

where $L(x,\,\lambda,\,\mu)=f(x)+\langle\lambda,\,h(x)\rangle+\langle\mu,\,g(x)\rangle$ is the Lagrangian of problem (1). We shall consider the class of algorithms for solving (1) (or (2)) described by a *perturbed sequential quadratic programming* (pSQP) framework. Specifically, for the given primal-dual iterate $(x^k,\,\lambda^k,\,\mu^k)$, the next iterate $(x^{k+1},\,\lambda^{k+1},\,\mu^{k+1})$ in the pSQP framework must satisfy the following relations in the variables $(x,\,\lambda,\,\mu)\in\mathbb{R}^n\times\mathbb{R}^l\times\mathbb{R}^m$:

$$\begin{array}{c}\frac{\partial L}{\partial x}(x^k,\,\lambda^k,\,\mu^k)+\frac{\partial^2 L}{\partial x^2}(x^k,\,\lambda^k,\,\mu^k)(x-x^k)+\,h'(x^k)^{\mathrm{T}}(\lambda-\lambda^k)+g'(x^k)^{\mathrm{T}}(\mu-\mu^k)+\omega_1^k=0,\\ h(x^k)+h'(x^k)(x-x^k)+\omega_2^k=0,\\ \mu\ge 0,\quad g(x^k)+g'(x^k)(x-x^k)+\omega_3^k\le 0,\quad \langle\mu,\,g(x^k)+g'(x^k)(x-x^k)+\omega_3^k\rangle=0,\end{array} \tag{3}$$

where $\omega_1^k\in\mathbb{R}^n$, $\omega_2^k\in\mathbb{R}^l$, and $\omega_3^k\in\mathbb{R}^m$ are perturbation terms defining specific algorithms. In particular, when $\omega_1^k=0$, $\omega_2^k=0$, and $\omega_3^k=0$, then (3) is precisely the KKT system of the basic SQP subproblem:

$$\begin{array}{ll}\text{minimize} & f(x^k)+\langle f'(x^k),\,x-x^k\rangle+\frac{1}{2}\left\langle H_k(x-x^k),\,x-x^k\right\rangle\\ \text{subject to} & h(x^k)+h'(x^k)(x-x^k)=0,\ g(x^k)+g'(x^k)(x-x^k)\le 0,\end{array} \tag{4}$$

with the choice

$$H_k = \frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k,\, \mu^k). \tag{5}$$

It is worth emphasizing that for some forms of perturbations, the pSQP framework includes algorithms which may not be modifications of SQP per se, in the sense that subproblems of those algorithms are not even quadratic. The point is that iterates of all the methods in the considered class can be related to a perturbation of SQP given by (3) a posteriori.

Let $\mathcal{M}(\bar{x}) = \{(\lambda,\, \mu) \in \mathbb{R}^l \times \mathbb{R}^m \mid (\lambda,\, \mu) \text{ satisfies (2) for } x = \bar{x}\}$ be the set of Lagrange multipliers associated with $\bar{x}$. Thus $\bar{x}$ is a stationary point of problem (1) if $\mathcal{M}(\bar{x}) \neq \emptyset$. Let

$$A = A(\bar{x}) = \{i = 1,\, \ldots,\, m \mid g_i(\bar{x}) = 0\}, \quad N = N(\bar{x}) = \{1,\, \ldots,\, m\} \setminus A$$

be the sets of indices of active and inactive constraints at a stationary point $\bar{x}$ of problem (1). Also, let

$$A_+ = A_+(\bar{x},\, \bar{\mu}) = \{i \in A(\bar{x}) \mid \bar{\mu}_i > 0\}, \quad A_0 = A_0(\bar{x},\, \bar{\mu}) = A \setminus A_+$$

be the sets of indices of strongly and weakly active constraints, respectively. The linear independence constraint qualification (LICQ) at $\bar{x}$ consists of saying that the gradients of equality constraints together with the gradients of inequality constraints active at $\bar{x}$ form a linearly independent set in $\mathbb{R}^n$. Obviously, LICQ implies that the multiplier set $\mathcal{M}(\bar{x})$ is a singleton. The strict Mangasarian–Fromovitz constraint qualification (SMFCQ) consists of saying that the multiplier associated to $\bar{x}$ is unique. Thus SMFCQ is a weaker assumption than LICQ. The critical cone of problem (1) at its stationary point $\bar{x}$ is given by

$$C = C(\bar{x}) = \{\xi \in \mathbb{R}^n \mid h'(\bar{x})\xi = 0,\; g'_{A_+}(\bar{x})\xi = 0,\; g'_{A_0}(\bar{x})\xi \leq 0\}.$$

We say that the second-order sufficient condition (SOSC) is satisfied at $\bar{x}$ with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$ if

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x},\, \bar{\lambda},\, \bar{\mu})\xi,\, \xi \right\rangle > 0 \quad \forall\, \xi \in C \setminus \{0\}. \tag{6}$$

2 Primal superlinear convergence in pSQP framework

Recall that Q-superlinear/quadratic convergence of the primal-dual sequence does not automatically guarantee any Q-rate of convergence of the primal part of the sequence (e.g., [4, Exercise 14.8]). The issue of primal rate of convergence, assuming (or having established) primal-dual convergence, is thus studied separately, often in combination with quasi-Newton considerations.

Let π_C denote the Euclidean projection onto C. We obtain the following necessary conditions for superlinear primal convergence of pSQP iterates.

Proposition 1 *Let $\bar{x}$ be a stationary point of problem* (1) *with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$. Let $\{(x^k,\, \lambda^k,\, \mu^k)\}$ be convergent to $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$, and assume that for each k the triple $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies the system* (3) *with some ω_1^k, ω_2^k, and ω_3^k.*

If $\{(\omega_3^k)_N\} \to 0$ as $k \to \infty$ and the rate of convergence of $\{x^k\}$ is superlinear, then it holds that

$$\pi_C(-\omega_1^k) = o(\|x^{k+1} - x^k\| + \|x^k - \bar{x}\|), \tag{7}$$

$$\omega_2^k = o(\|x^{k+1} - x^k\| + \|x^k - \bar{x}\|), \tag{8}$$

$$(\omega_3^k)_{A_+} = o(\|x^{k+1} - x^k\| + \|x^k - \bar{x}\|). \tag{9}$$

The sufficient conditions for primal superlinear rate can be stated as follows.

Proposition 2 *Let $\bar{x}$ be a stationary point of problem* (1) *with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$ satisfying SOSC* (6)*. Let $\{(x^k,\, \lambda^k,\, \mu^k)\}$ be convergent to $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$, and assume that for each k the triple $(x^{k+1},\, \lambda^{k+1},\, \mu^{k+1})$ satisfies the system* (3) *with some ω_1^k, ω_2^k, and ω_3^k, and $\{\pi_C(-\omega_1^k)\} \to 0$, $\{\omega_2^k\} \to 0$, $\{\omega_3^k\} \to 0$ as $k \to \infty$.*

Then

$$\|x^{k+1} - \bar{x}\| = O(\|(\pi_C(-\omega_1^k),\, \omega_2^k,\, (\omega_3^k)_A)\|) + o(\|x^k - \bar{x}\|).$$

In particular, if (7)-(9) *and $(\omega_3^k)_{A_0} = o(\|x^{k+1} - x^k\| + \|x^k - \bar{x}\|)$ hold, then the rate of convergence of $\{x^k\}$ is superlinear.*

3 APPLICATIONS TO SPECIFIC METHODS

3.1 SQP AND QUASI-NEWTON SQP

The weakest assumptions under which Q-superlinear/quadratic primal-dual convergence of basic SQP (4)–(5) had been established in the literature are SMFCQ and SOSC [3]. Other results require, in addition to SOSC, the stronger LICQ (e.g., [4, Theorem 15.4]) or even strict complementarity (e.g., [15], [4, Theorem 15.2]). For SQP iterates given by (4), the following is known [4, Theorem 15.7] (see also [2, 3], [13, Theorem 18.5] for related statements). Assuming that the primal-dual sequence $\{(x^k,\, \lambda^k,\, \mu^k)\}$ converges and that LICQ and SOSC hold at the limit $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$, the rate of convergence of the primal sequence $\{x^k\}$ is superlinear if and only if the following Dennis–Moré-type condition holds:

$$\pi_C\left(\left(\frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k,\, \mu^k) - H_k\right)(x^{k+1} - x^k)\right) = o(\|x^{k+1} - x^k\|). \tag{10}$$

Note that the KKT system of SQP subproblem (4) is a special case of pSQP framework (3) with

$$\omega_1^k = \left(H_k - \frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k,\, \mu^k)\right)(x^{k+1} - x^k), \quad \omega_2^k = 0, \quad \omega_3^k = 0.$$

Then, the characterization of primal Q-superlinear convergence of both pure and quasi-Newton SQP methods readily follows from Proposition 1 and Proposition 2.

Theorem 1 *Let $\bar{x}$ be a stationary point of problem* (1) *with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$. Let $\{H_k\}$ be a sequence of $n \times n$ symmetric matrices, and let the sequence $\{(x^k,\, \lambda^k,\, \mu^k)\}$ be generated in the following way: for each k, x^{k+1} is a stationary point of problem* (4) *with $(\lambda^{k+1},\, \mu^{k+1})$ being an associated Lagrange multiplier. Suppose that $\{(x^k,\, \lambda^k,\, \mu^k)\}$ converges to $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$.*

If the rate of convergence of $\{x^k\}$ is superlinear, then the Dennis–Moré condition (10) *holds.*

Conversely, if $(\bar{\lambda},\, \bar{\mu})$ satisfies SOSC (6) *and the Dennis–Moré condition* (10) *holds, then the rate of convergence of $\{x^k\}$ is superlinear.*

3.2 LCL METHODS

This approach is adopted, in particular, in the MINOS software package [11]. Consider problem (1) with $g(x) = -x$, $x \in \mathbb{R}^n$, i.e.,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0,\ x \geq 0. \end{array} \tag{11}$$

Note that general inequality constraints can be reformulated as equality constraints introducing slack variables; see [14, 10, 6]. Subproblems of the LCL method for (11) consist in:

$$\begin{array}{ll} \text{minimize} & f(x) + \langle \lambda^k, h(x)\rangle + \dfrac{c_k}{2}\|h(x)\|^2 \\ \text{subject to} & h(x^k) + h'(x^k)(x - x^k) = 0,\ x \geq 0, \end{array} \tag{12}$$

where $c_k \geq 0$. The next primal iterate x^{k+1} is defined as a stationary point of (12). Taking an associated Lagrange multiplier $(\eta^k,\, \mu^{k+1})$, the next dual iterate is defined by $(\lambda^{k+1},\, \mu^{k+1}) = (\lambda^k + \eta^k,\, \mu^{k+1})$.

The sharpest local convergence result for LCL methods had been obtained in [8]. It affirms quadratic primal-dual convergence under SMFCQ and SOSC. Other results in the literature require the stronger LICQ and strict complementarity in addition [14, 10, 6]. To the best of our knowledge, no primal Q-rate of convergence has been previously available. Applying our general results for pSQP framework, we obtain the following theorem.

Theorem 2 *Let $\bar{x}$ be a stationary point of problem* (1) *with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$ satisfying SOSC* (6)*. Let $\{(x^k,\, \lambda^k,\, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ be a sequence generated in the following way: for each k, x^{k+1} is a stationary point of problem* (12) *with $c_k = c$, and $(\lambda^{k+1} - \lambda^k,\, \mu^{k+1})$ is an associated Lagrange multiplier. Suppose that $\{(x^k,\, \lambda^k,\, \mu^k)\}$ converges to $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$.*

Then the rate of convergence of $\{x^k\}$ is superlinear.

3.3 SEQUENTIAL QUADRATICALLY CONSTRAINED QUADRATIC PROGRAMMING

The SQCQP method for problem (1) is the following algorithm (see [1, 7, 16, 5]). For the current iterate x^k, the next primal iterate x^{k+1} is computed as a stationary point of the subproblem

$$\begin{aligned} \text{minimize} \quad & \langle f'(x^k),\, x - x^k\rangle + \tfrac{1}{2}\langle f''(x^k)(x - x^k),\, x - x^k\rangle \\ \text{subject to} \quad & h(x^k) + h'(x^k)(x - x^k) + \tfrac{1}{2}h''(x^k)[x - x^k,\, x - x^k] = 0, \\ & g(x^k) + g'(x^k)(x - x^k) + \tfrac{1}{2}g''(x^k)[x - x^k,\, x - x^k] \leq 0, \end{aligned} \tag{13}$$

and $(\lambda^{k+1},\, \mu^{k+1}) \in \mathbb{R}^l \times \mathbb{R}^m$ is defined as an associated Lagrange multiplier.

In the convex case, subproblem (13) can be cast as a second-order cone program [9, 12], which can be solved efficiently by interior-point algorithms. The primal superlinear convergence result in [1] refers to a trust-region version of SQCQP with exact values of second derivatives and assumes Mangasarian–Fromovitz constraint qualification (MFCQ) and a quadratic growth condition (a weak form of SOSC). Quadratic primal-dual convergence had been established in [8] under SMFCQ and SOSC. Superlinear primal convergence under Dennis–Moré-type conditions is shown in [5] (in a more general variational context) assuming LICQ and SOSC. We obtain the following.

Theorem 3 *Let $\bar{x}$ be a stationary point of problem* (1) *with $(\bar{\lambda},\, \bar{\mu}) \in \mathcal{M}(\bar{x})$ satisfying SOSC* (6)*. Let $\{(x^k,\, \lambda^k,\, \mu^k)\}$ be generated in the following way: for each k, x^{k+1} is a stationary point of* (13)*, and $(\lambda^{k+1},\, \mu^{k+1})$ is an associated Lagrange multiplier. Suppose that $\{(x^k,\, \lambda^k,\, \mu^k)\}$ converges to $(\bar{x},\, \bar{\lambda},\, \bar{\mu})$.*

Then the rate of convergence of $\{x^k\}$ is superlinear.

REFERENCES

[1] M. ANITESCU, *A superlinearly convergent sequential quadratically constrained quadratic programming algorithm for degenerate nonlinear programming*, SIAM J. Optim., 12 (2002), pp. 949–978.

[2] P.T. BOGGS, J.W. TOLLE, AND P. WANG, *On the local convergence of quasi-Newton methods for constrained optimization*, SIAM J. Control Optim., 20 (1982), pp. 161–171.

[3] J.F. BONNANS, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[4] J.F. BONNANS, J.CH. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *Numerical Optimization: Theoretical and Practical Aspects*, 2nd ed., Springer-Verlag, Berlin, 2006.

[5] D. FERNÁNDEZ AND M. SOLODOV, *On local convergence of sequential quadratically-constrained quadratic-programming type methods, with an extension to variational problems*, Comput. Optim. Appl., 39 (2008), pp. 143–160.

[6] M.P. FRIEDLANDER AND M.A. SAUNDERS, *A globally convergent linearly constrained Lagrangian method for nonlinear optimization*, SIAM J. Optim., 15 (2005), pp. 863–897.

[7] M. FUKUSHIMA, Z.-Q. LUO, AND P. TSENG, *A sequential quadratically constrained quadratic programming method for differentiable convex minimization*, SIAM J. Optim., 13 (2003), pp. 1098–1119.

[8] A.F. IZMAILOV AND M.V. SOLODOV, *Inexact Josephy–Newton framework for generalized equations and its applications to local analysis of Newtonian methods for constrained optimization*, Comput. Optim. Appl., 46 (2010), pp. 347–368.

[9] M.S. LOBO, L. VANDENBERGHE, S. BOYD, AND H. LEBRET, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.

[10] B.A. MURTAGH AND M.A. SAUNDERS, *A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints*, Math. Program. Study, 16 (1982), pp. 84–117.

[11] B.A. MURTAGH AND M.A. SAUNDERS, *MINOS* 5.0 *User's Guide*, Technical report SOL 83.20, Stanford University, Palo Alto, CA, 1983.

[12] Y.E. NESTEROV AND A.S. NEMIROVSKII, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM Publications, Philadelphia, 1993.

[13] J. NOCEDAL AND S.J. WRIGHT, *Numerical Optimization*, Springer-Verlag, New York, 1999.

[14] S.M. ROBINSON, *A quadratically convergent algorithm for general nonlinear programming problems*, Math. Program., 3 (1972), pp. 145–156.

[15] S.M. ROBINSON, *Perturbed Kuhn-Tucker points and rates of convergence for a class nonlinear-programming algorithms*, Math. Program., 7 (1974), pp. 1–16.

[16] M.V. SOLODOV, *On the sequential quadratically constrained quadratic programming methods*, Math. Oper. Res., 29 (2004), pp. 64–79.

CONVERGENCE TO THE OPTIMAL VALUE FOR BARRIER METHODS COMBINED WITH HESSIAN RIEMANNIAN GRADIENT FLOWS

Felipe Alvarez[♭] and Julio López[†]

[♭]*Departamento de Ingeniería Matemática, Centro de Modelamiento Matemático, Universidad de Chile, Av. Blanco Encalada 2120, Santiago, Chile, falvarez@dim.uchile.cl, www.dim.uchile.cl*
[†]*Departamento de Ingeniería Matemática, Universidad de Chile, Av. Blanco Encalada 2120, Santiago, Chile, jclopez@dim.uchile.cl, www.dim.uchile.cl*

Abstract: We consider the problem $\min_{x\in\mathbb{R}^n}\{f(x) \mid Ax = b,\ x \in \overline{C},\ g_j(x) \leq 0,\ j = 1,\ldots,s\}$, where $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ is a full rank matrix, $\overline{C}$ is the closure of a nonempty, open and convex subset C of $\mathbb{R}^n$, and $g_j(\cdot)$, $j = 1,\ldots,s$, are nonlinear convex functions. Our strategy consists firstly in to introduce a barrier-type penalty for the constraints $g_j(x) \leq 0$, then endowing $\{x \in \mathbb{R}^n \mid Ax = b, x \in C\}$ with the Riemannian structure induced by the Hessian of an essentially smooth convex function h such that $C = \text{int}(\text{dom}\, h)$, and finally considering the flow generated by the Riemannian penalty gradient vector field. Under minimal hypotheses, we investigate the well-posedness of the resulting ODE and we prove that the value of the objective function along the trajectories, which are strictly feasible, converges to the optimal value. Moreover, the value convergence is extended to the sequences generated by an implicit discretization scheme which corresponds to the coupling of an inexact generalized proximal point method with parametric barrier schemes.

Keywords: *Gradient flow, Hessian Riemannian metric, Bregman distance, Proximal algorithm.*
2000 AMS Subject Classification: 34G20, 34A12.

1 PRELIMINARIES

1.1 RIEMANNIAN GRADIENT

Let M be a smooth manifold and we denote by T_xM the tangent space to M at $x \in M$, a $\mathcal{C}^k$ metric on M, $k \geq 0$, is a family of scalar products $(\cdot,\cdot)_x$ on each T_xM such that $(\cdot,\cdot)_x$ depends in a $\mathcal{C}^k$ way on x. The pair $(M,(\cdot,\cdot)_x)$ is called a $\mathcal{C}^k$ Riemannian manifold. This structure permits to define a notion of gradient vector. Indeed, the gradient $\text{grad}f(x)$ of $f \in \mathcal{C}^1(M;\mathbb{R})$ at $x \in M$ is uniquely determined by:
(g_1) tangency condition: $\text{grad}f(x) \in T_xM$,
(g_2) duality condition: for all $v \in T_xM$, $df(x)v = (\text{grad}f(x), v)_x$.
Here $df(x) : T_xM \to \mathbb{R}$ denotes the differential of f at $x \in M$. If N is a submanifold of M then $T_xN \subset T_xM$ for all $x \in N$ so that the metric $(\cdot,\cdot)_x$ on M induces a metric on N by restriction. So, the corresponding gradient vector field of f restricted to N is $\text{grad}f|_N(x) = \Pi_x^{T_xN}\text{grad}f(x)$, where $\Pi_x^{T_xN} : T_xM \to T_xN$ is the $(\cdot,\cdot)_x$-orthogonal projection onto the linear subspace T_xN.

1.2 RIEMANNIAN GRADIENT FLOWS

Let $Q \subset \mathbb{R}^n$ be a nonempty, open and convex set. We denote by $\mathcal{S}^n_{++}$ the cone of positive definite symmetric $n \times n$ real matrices. Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous and convex function with effective domain $\text{dom}\, h = \{x \in \mathbb{R}^n \mid h(x) < +\infty\}$. We assume the following conditions:

$$(H_h;Q)\quad \begin{cases} (a) & Q = \text{int}(\text{dom}\, h). \\ (b) & h|_Q \in \mathcal{C}^2(Q;\mathbb{R}) \text{ and } \forall x \in Q, \nabla^2 h(x) \in \mathcal{S}^n_{++}. \\ (c) & \text{The map } x \mapsto \nabla^2 h(x) \text{ is locally Lipschitz continuous on } Q. \\ (d) & \forall \bar{x} \in \partial Q,\ \forall x^k \to \bar{x} \text{ with } x^k \in Q,\ \|\nabla h(x^k)\| \to +\infty. \end{cases}$$

Next, let us endow Q with the variable metric defined by

$$\forall v, w \in \mathbb{R}^n,\ (v,w)_x := \langle \nabla^2 h(x)v, w\rangle. \tag{1}$$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. For any $x \in Q$ and $v \in \mathbb{R}^n$ we have that

$$df(x)v = \langle \nabla f(x), v\rangle = \langle \nabla^2 h(x)\nabla^2 h(x)^{-1}\nabla f(x), v\rangle = (\nabla^2 h(x)^{-1}\nabla f(x), v)_x.$$

Thus the gradient with respect to the metric $(\cdot,\cdot)_x$ of f restricted to Q is given by

$$\text{grad}_h f(x) = \nabla^2 h(x)^{-1}\nabla f(x), \quad x \in Q. \tag{2}$$

Note that, given a $x \in Q$, the vector $-\text{grad}_h f(x)$ can be interpreted as that direction in $\mathbb{R}^n$ such that f decreases the most steeply at x with respect to the metric $(\cdot,\cdot)_x$, which motivates to consider the following dynamical system for the (local) minimization of f on Q:

$$\frac{du}{dt}(t) = -\text{grad}_h f(u(t)), \tag{3}$$

with initial condition $u(0) = x^0 \in Q$.

1.3 Problem

In this work, we treat a general mathematical programming problem of the type

$$(P) \qquad v(P) \equiv \min\{f(x) \mid Ax = b,\ x \in \overline{C},\ g_j(x) \le 0,\ j \in I\},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_j : \mathbb{R}^m \to \mathbb{R}$ for $j \in I := \{1,\dots,s\}$ are continuously differentiable convex functions. Set $G = \{x \in \mathbb{R}^n : g_j(x) < 0,\ j \in I\}$. Throughout this work, we assume:

$$(H_P) \qquad \begin{cases} (a)\ S(P),\ \text{set of optimal solutions of } (P), \text{ is nonempty and bounded.} \\ (b)\ \mathcal{F}^0 = \{x \in \mathbb{R}^n : Ax = b,\ x \in C\} \cap G \neq \emptyset \text{ (Slater's condition).} \end{cases}$$

1.4 Gradient flows for solving (P)

Let us take a function h_C satisfying $(H_h; C)$, and a *barrier-type* function $\theta : \mathbb{R} \to (0, +\infty]$ with $\text{dom}\,\theta = (-\infty, 0)$ such that:

$$(H_\theta) \qquad \begin{cases} (a)\ \theta : (-\infty,0) \to \mathbb{R} \text{ is smooth and convex.} \\ (b)\ \theta(s) > 0,\ \text{for all } s \in (-\infty,0),\ \text{with } \lim_{s\to 0^-}\theta(s) = +\infty. \\ (c)\ \theta'(s) > 0 \text{ with } \lim\limits_{s\to-\infty}\theta'(s) = 0 \text{ and } \lim\limits_{s\to 0^-}\theta'(s) = +\infty. \end{cases}$$

We have two alternatives to derive θ-based gradient flows on $\mathcal{F}^0$:
(A1) Riemannian gradients flow using the Hessian of the extended function given by

$$h_{C\cap G}(x) := h_C(x) + \sum_{i\in J}\theta(g_j(x)), \tag{4}$$

under second-order regularity conditions on θ and all g_j, $j \in J$ (at least $\mathcal{C}^2$).
(A2) Hybrid barrier-gradient flows using h_C and replacing the original objective function with the penalty approximate:

$$f_\varepsilon(x) = f(x) + \varepsilon\sum_{j\in I}\theta(g_j(x)/\varepsilon), \tag{5}$$

where $\varepsilon > 0$ is a scalar parameter which will ultimately go to 0.

In the first alternative (A1), we notice that the extended function $h_{C\cap G} : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ defined by (4) satisfies $(H_h; C\cap G)$, hence $\mathcal{F}^0$ can be endowed with the Riemannian structure induced by the Hessian of $h_{C\cap G}$:

$$\nabla^2 h_{C\cap G}(x) = \nabla^2 h_C(x) + \sum_{j\in I}\theta''(g_j(x))\nabla g_j(x)\nabla g_j(x)^\top + \sum_{j\in I}\theta'(g_j(x))\nabla^2 g_j(x). \tag{6}$$

The gradient flow corresponding to (3) is given in this case by

$$(GF; u^0) \qquad \begin{cases} \dfrac{du}{dt}(t) = -\text{grad}_{h_{C\cap G}} f(u(t)), \\ u(0) = u^0 \in \mathcal{F}^0. \end{cases}$$

Here $\text{grad}_{h_{C\cap Q}} f : \mathcal{F}^0 \to \mathbb{R}^n$ stands for the Riemannian gradient vector field of f restricted to $\mathcal{F}^0$, with respect to the Hessian metric given by (1) for $Q = C \cap G$ and $h = h_{C\cap G}$. Therefore

$$\text{grad}_{h_{C\cap G}} f(x) = \Pi_x^{\ker A} \nabla^2 h_{C\cap G}(x)^{-1} \nabla f(x).$$

On the other hand, the second alternative (A2) is inspired by previous work on the coupling of the Euclidean steepest descent method with penalty schemes [4, 7]. In this case, we consider the non-autonomous Cauchy problem

$$(B\text{-}GF; u^0) \qquad \begin{cases} \dfrac{du}{dt}(t) = -\text{grad}_{h_C} f_{\varepsilon(t)}(u(t)), \\ u(0) = u^0 \in \mathcal{F}^0, \end{cases}$$

where the vector field $\text{grad}_h f_\varepsilon : \mathcal{F}^0 \to \mathbb{R}^n$ stands for the gradient of f_ε restricted to $\mathcal{F}^0$ with respect to the Hessian Riemannian metric given by (1). Here $\varepsilon : [0, +\infty) \to (0, +\infty)$ is a continuously differentiable parameterization in time of the penalty scheme such that

$$(H_\varepsilon) \qquad \varepsilon(t) > 0, \quad \dot{\varepsilon}(t) \leq 0 \quad \text{and} \quad \lim_{t\to+\infty} \varepsilon(t) = 0.$$

2 Global existence and convergence to the optimal value

Under (H_P), $(H_{h_C}; C)$ and (H_θ), it follows from [2, Theorem 4.1] that the Cauchy problem $(GF; u^0)$ is well-posed. The next result establishes that the same holds for $(B\text{-}GF; u^0)$ under additional condition (H_ε).

Theorem 1 *Under (H_P), $(H_h; C)$, (H_θ) and (H_ε), the following statements hold:*

(i) The Cauchy problem $(B\text{-}GF; u^0)$ admits a unique $\mathcal{C}^1$ solution $u : [0, +\infty) \to \mathcal{F}^0$.

(ii) The mapping $t \mapsto f_{\varepsilon(t)}(u(t))$ is nonincreasing, the trajectory $\{u(t) \mid t \in [0, +\infty)\}$ is bounded, and $(\dot{u}, \dot{u})_u \in L^1([0, +\infty); \mathbb{R})$.

(iii) For all $a \in \mathcal{F}^0$ and for all $t > 0$

$$f_{\varepsilon(t)}(u(t)) \leq f(a) + \frac{1}{t}\left[D_h(a, x^0) - D_h(a, u(t)) + \sum_{j\in I} \theta(g_j(a)/\varepsilon_0) \int_0^t \varepsilon(s) ds \right], \tag{7}$$

where

$$D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle \geq 0. \tag{8}$$

Hence,

$$\lim_{t\to+\infty} f_{\varepsilon(t)}(u(t)) = \lim_{t\to+\infty} f(u(t)) = \min_{\mathcal{F}} f = v(P),$$

and every cluster point of $\{u(t) \mid t \to +\infty\}$ belongs to $S(P)$.

3 GENERALIZED BARRIER PROXIMAL POINT ALGORITHM

The purpose of this section is to provide one discrete version of the Theorem 1. For each $k = 1, 2, \ldots$, let generate a sequence $\{x^k\} \in \mathcal{F}^0$ satisfying

$$\frac{\nabla h(x^k) - \nabla h(x^{k-1}) + A^\top w^k}{\lambda_k} \in -\partial_{\zeta_k} f_{\varepsilon_k}(x^k);\ Ax^k = b, \tag{9}$$

for some $w^k \in \mathbb{R}^m$ and $\zeta_k \geq 0$ is a tolerance for the computation of approximate subgradients. The following result generalizes the estimate derived in [5, Lemma 3.3(iii)] and extends the value convergence result established in [5, Theorem 3.4] for the exact version of (9) without penalty parameters, i.e, $\varepsilon_k = \zeta_k = 0$, $\forall k$ and also with $A = 0$.

Theorem 2 *Let $\{x^k\} \subset \mathcal{F}^0$ be the sequence generated by the generalized barrier proximal point algorithm (9) with $\{\varepsilon_k\}$ being decreasing to 0. Set $\sigma_n = \sum_{k=1}^n \lambda_k$. If $\sum_{k=1}^\infty \zeta_k < \infty$, then the following statements hold:*

(i) The real sequence $\{f_{\varepsilon_n}(x^n)\}$ is convergent, the sequence $\{x^n\}$ is bounded and we have that $\sum_{k=1}^\infty \lambda_k^{-1} \langle \nabla h(x^k) - \nabla h(x^{k-1}), x^k - x^{k-1}\rangle < +\infty$.

(ii) For all $a \in \mathcal{F}^0$ and for all $n \geq 1$ we have

$$\begin{aligned} \sigma_n(f_{\varepsilon_n}(x^n) - f(a)) \ \leq\ & \sum_{j\in I} \theta(g_j(a)/\varepsilon_0) \sum_{k=1}^n \lambda_k \varepsilon_k + D_h(a, x^0) - D_h(a, x^n) \\ & - \sum_{k=1}^n \sigma_k \lambda_k^{-1} D_h(x^k, x^{k-1}) + \sum_{k=1}^n \sigma_k \zeta_k. \end{aligned} \tag{10}$$

(iii) If $\sigma_n \to +\infty$, then the sequence $\{f_{\varepsilon_n}(x^n)\}$, converges to $v(P)$, hence $\{f(x^n)\}$ does so and every cluster point of $\{x^n\}$ belongs to $S(P)$.

REFERENCES

[1] R. BELLAMN, AND W. KARUSH, *On a new functional transform in analysis: the maximum transform*, Bull. AMS, 67 (1961), pp.501-503.

[2] F. ALVAREZ, J. BOLTE AND O. BRAHIC, *Hessian Riemannian gradient flows in convex programming*, SIAM J. Control Optim., 43 (2004), no 2, pp. 477-501.

[3] F. ALVAREZ AND R. COMINETTI, *Primal and dual convergence of a proximal point exponential penalty method for linear programming*, Math. Prog., 93 (2002) no. 1, Ser. A, pp. 87-96.

[4] H. ATTOUCH AND R. COMINETTI, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519-540.

[5] G. CHEN AND M. TEBOULLE, *Convergence analysis of proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), no. 3, pp. 538-543.

[6] R. COMINETTI, *Coupling the proximal point algorithm with approximation methods*, J. Optim. Theory Appl., 95 (1995), pp. 581-600.

[7] R. COMINETTI AND M. COURDURIER, *Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm*, SIAM J. Optim., 13, (2002), no 3, pp. 745-765.

[8] O. GÜLER, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), no 2, pp. 403-419.

RESTAURACIÓN INEXACTA SIN DERIVADAS EN OPTIMIZACIÓN NO LINEAL.

M. B. Arouxét[♭], N.E. Echebest[♭] y E. Pilotta[†]

[♭] *Departamento de Matemática. Facultad de Ciencias Exactas. UNLP, Argentina, belen@mate.unlp.edu.ar, opti@mate.unlp.edu.ar*

[†] *FAMAF, Universidad Nacional de Córdoba – (CIEM) CONICET, Argentina, pilotta@famaf.unc.edu.ar*

Resumen: Los métodos de optimización sin derivadas (DFO) son necesarios para resolver problemas en los cuales las derivadas de la función objetivo y/o de las restricciones no están disponibles para su aplicación en un procedimiento. Este trabajo propone un algoritmo en la metodología DFO basado en el esquema de resolución del método de Restauración Inexacta (IR) propuesto por Martínez y Pilotta (2000) y atendiendo a sus posteriores modificaciones. En el algoritmo propuesto los subproblemas son apropiadamente modelizados con el objetivo de no emplear derivadas, usando interpolación polinomial lineal o cuadrática en una cantidad mínima de puntos en los que se conocen los valores funcionales. Se estudian las condiciones del decrecimiento requerido en las dos etapas del algoritmo para obtener convergencia global a puntos estacionarios del problema. Se presentan resultados numéricos obtenidos sobre problemas con dimensiones medianas.

Palabras clave: *Optimización sin derivadas, Restauración Inexacta, interpolación polinomial.*
2000 AMS Subject Classification: 06B10- 06D05.

1. INTRODUCCIÓN

Consideramos el problema de optimización con restricciones

$$\text{Min } f(x) \text{ s.a } x \in \Omega,\ C(x) = 0, \tag{1}$$

siendo el conjunto Ω convexo y cerrado, $\Omega \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, $C : \mathbb{R}^n \to \mathbb{R}^m$, funciones continuamente diferenciables. En el caso más simple Ω toma la forma $\Omega = \{x \in \mathbb{R}^n \mid L \leq x \leq U, L < U\}$. Asumimos que las derivadas de f y $C(x)$ no pueden ser calculadas. Esta situación frecuentemente ocurre en problemas donde los valores funcionales de $f(x)$ y $C(x)$ provienen de medidas físicas, quimicas o geofísicas, o son el resultado de complejos cálculos de simulación. En [1, 2, 3, 8] se describen diferentes situaciones de esta clase y el estado de avance en la resolución correspondiente.

Notación. $|\cdot|$ indica una norma en $\mathbb{R}^m$, $\|\cdot\|$ norma arbitraria en $\mathbb{R}^n$, $\|\cdot\|_2$ la norma Euclídea en $\mathbb{R}^n$.

En este tabajo proponemos la resolución sin derivadas del problema (1) aplicando el esquema algorítmico del método de Restauración Inexacta propuesto, por Martínez y Pilotta [4], originalmente con el uso de derivadas. Los pasos característicos de esta clase de metodología son los siguientes:

Paso de Restauración. Dado un iterado actual $x_k \in \mathbb{R}^n$, se calcula un punto y donde una función que mide la infactibilidad del problema decrece en relación al valor en x_k. El procedimiento a utilizar para obtener tal resultado puede ser el que resulte más conveniente para cada problema particular.

Paso de Minimización. Se calcula un punto z, sobre el conjunto tangente (aproximación al dominio factible) a través de y, en una región de confianza centrada en y, satisfaciendo que el valor de una función que mide la optimalidad mejore en z en relación al valor de la misma en y.

Se define el $x_{k+1} = z$ si el punto z del paso previo es "aceptable" de acuerdo a un criterio que combina factibilidad y optimalidad. En el caso contrario, se reduce el radio de la región de confianza considerada en y y se repite el Paso de Minimización.

El estudio sobre la requerida mejoría de la optimalidad en z se hace midiendo el decrecimiento de la función objetivo $f(x)$ en [4], mientras que en [5] se hace considerando el Lagrangiano. La aceptación o no de z como el nuevo iterado en ambos trabajos [4, 5] depende de una función que combina factibilidad y optimalidad.

En el presente trabajo, tanto para obtener el punto z como para aceptarlo o rechazarlo se sigue la estrategia propuesta en [4]. Para concretar la resolución, sin el cálculo de derivadas en las etapas enunciadas en el esquema IR previo, se utilizan modelos lineales obtenidos mediante interpolación funcional en $n+1$ puntos.

En la etapa de Restauración se aproxima la función $h(x) = |C(x)|^2$, en el iterado actual x_k. Un modelo $m_k(x_k + s)$ es creado para aproximar a $h(x)$ alrededor de x_k. El modelo requiere que se interpole a $h(x)$ en x_k y en un conjunto $\{y^1, y^2, \ldots, y^n\}$ de puntos muestrales adicionales, o sea,

$$m_k(x_k) = h(x_k), \qquad m_k(y) = h(y) \quad \text{para todo } y \in \{y^1, y^2, \ldots, y^n\} \tag{2}$$

Las condiciones de interpolación determinan un sistema lineal de ecuaciones cuyas incógnitas son los coeficientes del modelo. Para que el sistema lineal (2) esté bien definido, se debe asegurar que los puntos muestrales determinen que la matriz del sistema lineal sea no singular. Tal propiedad se denomina "condición geométrica" y el respectivo conjunto de puntos muestrales que satisfacen esta propiedad se denomina no degenerado. Para modelos lineales sólo se requieren $n + 1$ puntos muestrales y en un iterado actual x_k toma la forma $m_k(x_k + s) = h(x_k) + g_k^t s$, donde g_k es un vector en $\mathbb{R}^n$ a ser determinado. Como g_k tiene n componentes, para el conjunto de n satélites se impone las condiciones de interpolación $m_k(y^l) = h(y^l)$, $l = 1, \ldots, n$, las cuales pueden ser escritas como

$$g_k^t s^l = h(y^l) - h(x_k) \qquad l = 1, \ldots, n, \tag{3}$$

donde s^l es el desplazamiento de x_k a y^l, $y^l = x_k + s^l$ $l = 1, \ldots, n$. Se sigue de (3), que el modelo lineal está unívocamente determinado si y sólo si el conjunto de puntos muestrales $\Sigma_k := \{x_k\} \cup \{y^1, y^2, \ldots, y^n\}$ es tal que el conjunto

$$\{s^l \ : \ l = 1, \ldots, n\}$$

es linealmente independiente. Teniendo en cuenta la estrategia seguida por Marazzi y Nocedal en [6] para calcular al nuevo iterado, primero se selecciona un punto $y^{l_{out}}$, el punto satélite que está más lejos de x_k. Con el objetivo que el nuevo conjunto de puntos muestrales no resulte degenerado, después de incluir el nuevo punto calculado, se considera el subespacio de $(n-1)$-dimensión generado por los vectores de desplazamiento

$$\{s^l \ : \ l = 1, \ldots, n, \qquad l \neq l_{out}\} \tag{4}$$

correspondientes a los satélites que permanecen en el conjunto muestral. Se calcula $b_k \in \mathbb{R}^n$ vector normal a (4), y se pide que la magnitud del coseno del ángulo entre el nuevo paso s y el normal b_k no sea menor que una constante dada $\gamma \in (0, 1)$, o sea,

$$|b_k^t s| \geq \gamma \|b_k\|_2 \|s\|_2 \tag{5}$$

Así se calcula el paso de prueba s resolviendo

$$\begin{aligned} \min_s m_k(x_k + s) &= h(x_k) + g_k^t s \\ \text{sujeto a } ||s|| &\leq \Delta_k, \\ |b_k^t s| &\geq \gamma \|b_k\|_2 \|s\|_2. \end{aligned} \tag{6}$$

Para resolver tal problema es posible ignorar la restricción (5) y considerar s_{TR} la correspondiente solución. Si s_{TR} satisface (5), entonces $s = s_{TR}$ es la solución del subproblema (6). Si no, es fácil verificar que la solución óptima se genera considerando g_k y b_k. Rotando s_{TR} en el plano generado por s_{TR} y b_k se encuentran dos puntos para los cuales se satisface la restricción por igualdad, y se elige el de menor valor funcional del modelo [6]. Así se obtiene una solución global para el subproblema (6), la cual es única (excepto si $b_k^t g_k = 0$, cuando hay exactamente dos soluciones globales).

2. ALGORITMO IR-DFO

El algoritmo IR-DFO para resolver el problema (1), basado en el método IR de Martínez y Pilotta [4], consta de los siguientes pasos.

Paso I: Restauración. Dado $x_k \in \Omega$, $\eta > 0$, $\beta > 0$, $r_k \in [0, 1)$, Σ_k

Calcula $y \in \Omega$ a partir de la resolución del problema (6) tal que satisfaga

$$\begin{aligned} |C(y)| &\leq r_k|C(x_k)| \\ \|y - x_k\| &\leq \beta|C(x_k)|. \end{aligned}$$

Si no es posible obtener tal y, termina declarando falla al mejorar la factibilidad: **END**.

Calcula la aproximación lineal a f en y, $L_y(w) = f(y) + gf^t(w-y)$, interpolando los valores funcionales en $\{y\} \bigcup \{y^1, y^2, \ldots, y^n\}$

Genera $A(y)$ aproximación al Jacobiano de $C(x)$ en y.

$$\pi_y = \{w \in \Omega \mid A(y)(w - y) = 0\} \quad (7)$$

$$gf_{tan} = P(y - \eta gf) - y \quad (8)$$

- Si $C(y) = 0$ y $gf_{tan} = 0$, termina : **END**.

Paso II: Minimización Dado $\delta_y > 0$, $0 < \tau_1 < \tau_2$,

Calcula z utilizando el problema lineal (9) tal que $f(z) \leq f(y + tgf_{tan})$.

Halla un $z \in \pi_y$ tal que $f(z) \leq \max\{f(y + tgf_{tan}),\ f(y) - \tau_1\delta_y,\ f(y) - \tau_2\}$

Paso III: Elección del parámetro de penalización. Considera la función
$Pred(\theta) = \theta[f(x_k) - f(z)] + (1 - \theta)[|C(x_k)| - |C(y)|]$, con $\theta \in [0, 1]$.
Elige θ tal que $Pred(\theta) \geq \frac{1}{2}[|C(x_k)| - |C(y)|]$

Criterio de aceptación del punto z . Define $Ared(\theta) = \theta[f(x_k) - f(z)] + (1 - \theta)[|C(x_k)| - |C(z)|]$.

- Si $Ared(\theta) \geq 0.1\ Pred(\theta)$, define $x_{k+1} = z$, adapta Σ_{k+1} y va al Paso I.
  Si $Ared(\theta) < 0.1\ Pred(\theta)$, adapta $\delta_y \in [0.1\delta_y,\ 0.9\delta_y]$ y va al paso II.

Para calcular el punto z en el Paso II, se considera el problema:

$$\begin{aligned} \min_s L_y(y + s) &= f(y) + gf^t s \quad (9) \\ \text{sujeto a } ||s|| &\leq \delta_y, \\ A(y)(s) &= 0, \\ |b_k^t s| &\geq \gamma\|b_k\|_2\|s\|_2. \end{aligned}$$

La obtención del punto y en el Paso I y el punto z en el Paso II, basados en los problemas (6) y (9) respectivamente, requiere un proceso iterativo. Se describe sintéticamente el esquema de obtención del y a continuación.

El procedimiento para resolver el problema (9) es similar al previo aunque con las modificaciones necesarias para satisfacer la condición de II, y considerando que el conjunto de puntos satélites se centra en el y, $\Sigma_y = \{y\} \cup \{y^1, y^2, \ldots, y^n\}$.

Dado x_c, $\Sigma_c = \{x_c\} \cup \{y^1, y^2, \ldots, y^n\}$, Δ_c, $m(x_c + s)$:
$x^+ = x_c$, $\Sigma_+ = \Sigma_c$, $\Delta_+ = \Delta_c$.

(i) Halla el punto de Σ_+ más alejado del iterado actual: $y^{l_{out}}$.

(ii) Calcula el paso s que resuelve el subproblema y evalúa la función $h(x)$ en el punto hallado.

-Si $x^+ + s$ satisface la condición requerida en el Paso I, define $y = x^+ + s$, elimina el satélite $y^{l_{out}}$ e incluye el nuevo y al Σ_+. Termina.

Si no, hace una búsqueda en la dirección s:
- Si existe un $y = x^+ + ts$, $0 < t < 1$, que cumple la condición del Paso I, elimina $y^{l_{out}}$ como punto satélite en Σ_+ e incluye y. Termina.

- Si existe un $x^+ + \tau s$, tal que $h(x^+ + \tau s) < h(x_c)$, define $x^+ = x^+ + \tau s$, elimina $y^{l_{out}}$ del conjunto satélite e incluye x^+ y va al paso (iii).

- Si no, reemplaza el punto satélite $y^{l_{out}}$ por un punto conveniente $x^+ + \tau s$.

(iii) Adapta el modelo $m(x^+ + s)$, $\Sigma_+ = \{x^+\} \cup \{y^1, y^2, \ldots, y^n\}$, y Δ_+, va al paso (i).

En el trabajo completo se presentarán resultados que muestran la buena definición del algoritmo IR-DFO. También los resultados necesarios para establecer la convergencia global a puntos estacionarios del problema (1).

Se expondrán resultados numéricos obtenidos con una implementación del algoritmo IR-DFO en Fortran 77, utilizando el código MINOS para resolver los subpoblemas lineales requeridos en el Paso de Restauración y de Minimización.

Se han resuelto problemas test de la colección de Hock and Schittkowski [7] de dimensiones medianas, con m restricciones no lineales y n variables. Tales resultados muestran que el comportamiento de la actual implementación de IR-DFO es satisfactoria.

REFERENCIAS

[1] M.B. AROUXÉT, N. ECHEBEST AND E.A. PILOTTA, *Active-set strategy in Powell's method for optimization without derivatives*, To appear in Comp. Appl. Math. 2011.

[2] A. CONN AND K. SCHEINBERG AND L. VICENTE, *Introduction to derivative-free optimization*, MPS SIAM Series on Optimization, SIAM, 2009.

[3] M. A. DINIZ EHRHARDT, J.M.MARTÍNEZ AND L.G. PEDROSO, *Derivative-Free methods for nonlinear programming with general lower-level contraints*. To appear in Comp. Appl. Math. 2011.

[4] J.M. MARTÍNEZ, E.A. PILOTTA, *Inexact Restoration algorithms for constrained optimization*. Journal of Optimization Theory and Applications 104(2000), pp. 135-163.

[5] J. M. MARTÍNEZ, *Inexact-restoration method with lagrangian tangent decrease and new merit function for nonlinear programming*, Journal of Optimization Theory and Applications 111 (2001), pp. 39-58, 2001.

[6] M. MARAZZI, J. NOCEDAL, *Wedge trust region methods for derivative free optimization*, Mathematical Programming 91 (2002) 289–305.

[7] W. HOCK AND K. SCHITTKOWSKI, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Economics and Mathematical Systems, Vol. 187, Springer 1981.

[8] M. J. D. Powell, The BOBYQA algorithm for bound constrained optimization without derivatives, 2009, Cambridge NA Reports, 6.

ESTRATEGIA DE REGIÓN DE CONFIANZA PARA PROBLEMAS DE OPTIMIZACIÓN MULTIOBJETIVO *

Gabriel Aníbal Carrizo[♭] y Maria Cristina Maciel[†]

[♭]*CONICET-Departamento de Matemática, Universidad Nacional del Sur, Bahía Blanca, Argentina, gabrielanibal@gmail.com*

[†]*Departamento de Matemática, Universidad Nacional del Sur, Bahía Blanca, Argentina, immaciel@criba.edu.ar*

Resumen: En este trabajo se considera el problema de optimización multiobjetivo sin restricciones y se propone un algoritmo basado en la estrategia de región de confianza para optimización escalar. Dentro de las características más interesantes corresponde señalar que los resultados de convergencia obtenidos incluyen problemas no convexos.

1. INTRODUCCIÓN

Consideremos el problema de optimización multiobjetivo

$$\min_{x\in U} F(x) \tag{1}$$

con $U \subseteq \mathbb{R}^n$ y $F : \mathbb{R}^n \to \mathbb{R}^m$ dos veces continuamente diferenciable, $F(x) = \left(\begin{array}{ccc} f_1(x), & \ldots & f_m(x) \end{array} \right)^T$.

Para este problema los métodos propuestos en [4] y [5] son generalizaciones de los métodos de Newton y máximo descenso para optimización escalar. En ambos casos definen una dirección de descenso via la resolución de un subproblema de optimización con restricciones y utilizan la estrategia de globalización de búsqueda lineal. En ambos trabajos se asume la hipótesis de convexidad de las funciones objetivos, propiedad que es útil en la resolución de los subproblemas y en las condiciones de optimalidad. En este trabajo se considera el caso irrestricto: $U = \mathbb{R}^n$ y se proponen modificaciones al subproblema introduciendo la restricción de región de confianza.

2. CONCEPTOS DE OPTIMIZACIÓN MULTIOBJETIVO

Con propósito de es minimizar simultaneamente las k funciones objetivo, si no hay conflicto entre ellas la solución puede hallarse donde todas realizan sus mínimos, pero este no es el caso en general. Cuando no se da esa particularided es necesario establecer una preferencia entre puntos factibles, para ello se introduce el concepto de *optimo Pareto*:

Definicin 1 *Se dice que $x^\star$ es Pareto optimal si $\nexists y \in U$ tal que*

$$F(y) \leq F(x^\star) \text{ y } F(y) \neq F(x^\star).$$

y si restringimos este concepto a un entorno del punto el mismo se dice *localmente Pareto optimal*.

Un concepto estrechamente relacionado es el de *debilmente Pareto optimal*:

Definicin 2 *Decimos que $x^\star$ es debilmente Pareto optimal si $\nexists y \in U$ tal que*

$$F(y) < F(x^\star).$$

Para el resto del trabajo consideraremos una condición necesaria para que un punto sea localmente Pareto optimal, que es la de punto *crítico Pareto*:

Definicin 3 *Se dice que un punto $\bar{x} \in \mathbb{R}^n$ es crítico Pareto si*

$$Rg\left(\nabla F(\bar{x})^T\right) \cap \left(-\mathbb{R}^m_{++}\right) = \phi.$$

donde $Rg\left(\nabla F(\bar{x})^T\right) = \left\{\nabla F(\bar{x})^T v \ : \ v \in \mathbb{R}^n\right\}$.

*Este trabajo ha sido subsidiado por la Universidad Nacional del Sur, proyecto N° 24/L069.

Por simplicidad nos referiremos a estos puntos utilizando el término *crítico*.

En optimización escalar decimos que s es una dirección en x de descenso cuando $\exists t_0 > 0$ tal que $\forall t \in (0, t_0]$

$$f(x + ts) < f(x),$$

y, si f es continuamente diferenciable, es facilmente caraterizable mediante

$$\nabla f(x)^T s < 0.$$

Para generalizar esta idea en [4] se establece el concepto de dirección de descenso en el caso multiobjetivo:

Definicin 4 *Se dice s es una dirección de descenso para F en x si existe $t_0 > 0$ tal que*

$$F(x + ts) < F(x) \ \ \forall t \in (0, t_0]\,.$$

3. ALGORITMOS PREVIOS

En este trabajo se sonsidera el método de escalarizacion propuesto en [4, 5].

En pos de obtener, en caso de que exista, una dirección de descenso para F, a partir de x, en [5] se resuelve el problema

$$\begin{array}{ll} \text{mín} & f_x(v) + \frac{1}{2}\|v\|^2 \\ s.a. & x \in \mathbb{R}^n, \end{array}$$

donde

$$f_v(x) = \text{máx}\left\{(\nabla F(x)v)_i : \ 1 \leq i \leq m\right\}.$$

Como la función objetivo es propia, cerrada y fuertemente convexa, siempre tiene solución única [1]. Agregando una variable α se lo puede reformular de modo tal de evitar el inconveniente de la no diferenciabilidad:

$$\begin{array}{ll} \text{mín} & \alpha + \frac{1}{2}\|v\|^2 \\ s.a. & (\nabla F(x)v)_i \leq \alpha, \text{ para } 1 \leq i \leq m, \end{array} \tag{2}$$

que es un problema cuadrático con restricciones lineales de desigualdad, equivalente al anterior. Donde el mínimo es 0 si x es crítico.

De este modo se obtiene que la dirección hallada es una dirección de máximo descenso con la cual se realiza una búsqueda lineal a través de una regla tipo Armijo. Con este procedimiento se obtiene convergencia a puntos críticos, que en el caso convexo son debilmente Pareto optimales.

En el trabajo [4] se define como dirección de Newrton a s, solución de

$$\begin{array}{ll} \text{mín} & \text{máx}_{j=1,\ldots,m} \nabla f_j(x)^T s + \frac{1}{2} s^T \nabla^2 f_j(x)\, s \\ s.a. & s \in \mathbb{R}^n. \end{array}$$

Debido a que este problema es no diferenciable, se resuelve el siguiente problema equivalente

$$\begin{array}{ll} \text{mín} & t \\ s.a. & \nabla f_j(x)^T s + \frac{1}{2} s^T \nabla^2 f_j(x)\, s - t \leq 0 \ \ (0 \leq j \leq m) \\ & (t, s) \in \mathbb{R} \times \mathbb{R}^n. \end{array} \tag{3}$$

En el caso convexo la dirección s es de descenso y se puede hacer un análisis análogo al de [5], con la ventaja de obtener convergencia local q-quadrática.

4. DIFICULTADES DE LA NO-CONVEXIDAD

Al considerar funciones F tales que las f_i, $1 \leq i \leq m$, no sean convexas el problema 3 puede estar no acotado. Para solucionar este problema se puede acotar la variable v, obteniendose:

$$\begin{array}{ll} \text{mín} & t \\ & \nabla f_j(x)^T s + \frac{1}{2} s^T \nabla^2 f_j(x) s - t \leq 0 \ (0 \leq j \leq m) \\ s.a. & \|s\| \leq \Delta \\ & (t,s) \in \mathbb{R} \times \mathbb{R}^n. \end{array}$$

De este modo tenemos garantizada la existencia de solución y que esté acotada, pero s puede no ser de descenso si es una dirección de curvatura negativa, para ello imponemos las restricciones del problema 2:

$$\begin{array}{ll} \text{mín} & t \\ & \nabla f_j(x)^T s + \frac{1}{2} s^T \nabla^2 f_j(x) s - t \leq 0 \quad (0 \leq j \leq m) \\ s.a. & \nabla f_j(x)^T s - t \leq 0 \quad (0 \leq j \leq m) \\ & (t,s) \in \mathbb{R} \times \mathbb{R}^n. \end{array} \tag{4}$$

El subproblema 4 será el subproblema elegido para para el algoritmo utilizando región de confianza.

5. EL ALGORITMO PRINCIPAL

Los algoritmos que utilizan la estrategia de región de confianza establecen un modelo cuadrático alrededor del iterado actual para luego resolver el subproblema de minimizar dicho modelo en una región alrededor del punto. En nuestro caso definimos un modelo para cada una de las funciones objetivo:

$$q_j^k\left(x^k + s\right) = \nabla f_j\left(x^k\right)^T s + \frac{1}{2} s^T \nabla^2 f_j\left(x^k\right) s.$$

El siguiente algoritmo aproxima a un punto crítico del problema 1.

Algoritmo 1 *Dados* x^0 Δ_0, $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < \gamma_2 < 1$, $tol > 0$ *y* $k = 0$

1. *Calcular* $\nabla f_j\left(x^k\right)$, $\nabla^2 f_j\left(x^k\right)$ *y resolver el subproblema*

$$\begin{array}{ll} \text{mín} & t \\ & \nabla f_j\left(x^k\right)^T s + \frac{1}{2} s^T \nabla^2 f_j\left(x^k\right) s - t \leq 0 \quad (0 \leq j \leq m) \\ s.a. & \nabla f_j\left(x^k\right)^T s - t \leq 0 \quad (0 \leq j \leq m) \\ & \|s\| \leq \Delta_k \\ & (t,s) \in \mathbb{R} \times \mathbb{R}^n. \end{array}$$

2. *Si* $|t| < \epsilon$ *terminar.*

3. *Calcular*

$$\rho_k^j = \frac{f_j\left(x^k\right) - f_j\left(x^k + s_k\right)}{-q_j^k\left(x^k + s_k\right)}. \tag{5}$$

4. *Ajustar el radio de la región de confianza*

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & si \quad \rho_k^j \geq \eta_2 \forall j. \\ [\gamma_2 \Delta_k, \Delta_k) & si \quad \rho_k^j \geq \eta_1 \forall j \ y \ \exists l \ / \rho_k^l \leq \eta_2. \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k) & si \quad \exists l \ / \rho_k^l \leq \eta_1. \end{cases}$$

5. *Si* $\rho_k^j \geq \eta_1 \forall j$ $x^k + 1 = x^k + s_k$, *si no* $x^k + 1 = x^k$.

6. $k = k + 1$.

A diferencia de los esquemas estándar de región de confianza aquí se definen m cocientes entre reducciones reales y predichas, uno por cada función. A partir de ellos se modifica el tamaño de la región de confianza, aumentandose si todos los cocientes sean muy satisfactorios, disminuyéndose en caso que haya uno no satisfactorio y manteniéndose en otro caso.

6. CONVERGENCIA

La prueba de la convergencia comparte características con la prueba usual para el caso escalar, ver [3].

Teorema 1 *Asumiendo que* $\max \left\|\nabla^2 f_j(x)\right\| < \kappa$. *Si* $\left|\nabla f_j\left(x^k\right) s\right| \neq 0$ *y*

$$\Delta_k \leq \frac{\left|\nabla f_j\left(x^k\right) s\right|(1-\eta_2)}{2\kappa\|s\|} \forall j.$$

Entonces la iteración k *es muy buena y*

$$\Delta_{k+1} \geq \Delta_k.$$

Teorema 2 *Supongamos que* $\exists \bar{k}$ *tal que* $\frac{\left|\nabla f_j\left(x^k\right)^T s_k\right|}{\|s_k\|} > \bar{k}\ \forall k\ \forall j$ *entonces* $\exists \tilde{k}$ *tal que*

$$\Delta_k > \tilde{k}$$

Teorema 3 *Supongamos que hay una cantidad finita de iteraciones exitosas, entonces* $x^k = x^\star$ *para todo* k *lo suficientemente grande y* $x^\star$ *es un punto crítico.*

Teorema 4

$$\lim \inf_{k\to\infty} \left|\nabla f_j\left(x^k\right)^T s_k\right| = 0.$$

$$\lim \inf_{k\to\infty} \left|q_j^k(s_k)\right| = 0.$$

Luego lo que se obtiene es que la existe una subsucesión convergente a un punto crítico.
Las demostraciones de los teoremas 1, 2, 3 y 4 pueden verse en [2].

REFERENCIAS

[1] D. BERTSEKAS, *Convex analysis and optimization*, Athena Scientific, Belmont (2003).

[2] G.A. CARRIZO, M.C. MACIEL, *A trust-region-based on algorithm for nonconvex unconstrained multiobjective optimization*, en preparación.

[3] A.R. CONN, N.I.M. GOULD, P.L. TOINT, *Trust-Region Methods*, MPS-SIAM (2000).

[4] J. FLIEGE, L.M. GRAÑA DRUMMOND, B.F. SVAITER, *Newton's method for multiobjective optimization*, SIAM J. Optim. 20, (2009) 602-626.

[5] J.FLIEGE, B.F. SVAITER, *Steepest descent methods for multicriteria optimization*, Mathematical methods of operations research, Vol 51 (2000), No 3, 479-494.

SOBRE LA CONVERGENCIA DE UN ALGORITMO NEWTON PARA EL PROBLEMA DE OPTIMIZACIÓN MATRICIAL

María Gabriela Eberle[♭] y María Cristina Maciel[♭]

[♭]*Departamento de Matemática, Universidad Nacional del Sur, Av. Leandro N. Alem 1253, 8000 Bahía Blanca, Argentina, gabriela.eberle@uns.edu.ar, immaciel@criba.edu.ar, www.uns.edu.ar*

Resumen: El problema de optimización matricial definido con orden parcial de Löwner es resuelto para el caso cuadrático empleando un algoritmo basado en el método de Newton clásico. Fliege, Graña y Svaiter [1] han utilizado la misma técnica para resolver el problema de optimización multiobjetivo definido con el orden de Pareto, y han probado los correspondientes resultados de convergencia. Para el caso cuadrático ambos problemas son equivalentes [5], y este trabajo se propone demostrar la convergencia del algoritmo propuesto explotando la citada equivalencia.

Palabras clave: *Orden de Löwner, Orden de Pareto, Método de Newton.*

1. MÉTODO DE NEWTON PARA FUNCIONES CUADRÁTICAS MATRICIALES

En esta contribución se considera el problema

$$\begin{cases} \min\limits_{s.a} & \mathbf{Q}(X) = X^T A X + B^T X + X^T B + C \\ & X \in \mathcal{S}, \end{cases} \tag{1}$$

donde A, B y $X \in \mathbb{R}^{n\times n}$.

El conjunto factible está dado por el subespacio de matrices simétricas, dotado con el orden parcial de Löwner y la norma de Frobenius. Dadas A y $B \in \mathcal{S}$, se dice que A es mayor ó igual que B en el sentido de Löwner si:

$$A \succeq B \qquad \Longleftrightarrow \qquad A - B \ \ es \ \ semidefinida \ \ positiva.$$

De manera similar, se define la correspondiente relación en sentido estricto. Cuando dos matrices estén relacionadas en este sentido se dirá que dichas matrices son comparables.

Definición 1 *Dados $X_\star \in \mathcal{S}$ y $\delta > 0$, se llama entorno de centro $X_\star$ y radio δ al conjunto de todas las matrices simétricas comparables con $X_\star$ tales que*

$$\mathcal{E}(X_\star, \delta) = \left\{ X \in \mathcal{S} \ : \ \left\| X - X_\star \right\|_{\mathrm{F}} < \delta \right\}.$$

Definición 2 *Sea $f : \mathcal{S} \longrightarrow \mathcal{S}$, se dice que $X_\star$ es un minimizador local (minimizador local en sentido estricto) para f si $(\mathbf{Q}(X) - \mathbf{Q}(X_\star))$ es semidefinida (definida) positiva, para toda $X \in \mathcal{E}(X_\star, \delta) \subset \mathcal{S}$.*

Se pretende generar un algoritmo basado en el método de Newton, por lo tanto se requieren los diferenciales de Frechet primero y segundo de $Q(X)$:

$$D\mathbf{Q}(X, \Delta) = XA\Delta + \Delta AX + B\Delta + \Delta B, \qquad D^2\mathbf{Q}(X, \Delta) = \Delta A \Delta.$$

El siguiente es el algoritmo propuesto, en el que se observan dos ciclos, uno externo y otro interno. Para cada iteración externa k hay que determinar Δ_k, lo que supone la resolución de una ecuación de Riccati. Empleando nuevamente el método de Newton para aproximar Δ_k, se genera un proceso iterativo que, en cada iteración interna l trata de resolver una ecuación de matricial de Sylvester.

Algoritmo 1.1 *(Método de Newton para minimización de* $\mathbf{Q}$*)*

Dada $X_0 \in \mathcal{S}$

Para $k = 0, 1, \cdots\cdots$, *hasta convergencia* **repetir**

$D^2\mathbf{Q}(X_k, \Delta_k) = -D\mathbf{Q}(X_k, \Delta_k)$

$\Longleftrightarrow$

$\mathbf{R}(\Delta_k) = \Delta_k A \Delta_k + {\alpha_k}^T \Delta_k + \Delta_k \alpha_k = 0$. *(Ec. de Riccati)*

siendo $\left\{ \ \alpha_k = \alpha(X_k) \right.$

(NE)

Dada ${\Delta_k}^0 \in \mathcal{S}$

Para $l = 0, 1, \cdots\cdots$, *hasta convergencia* **repetir**

$D\mathbf{R}({\Delta_k}^l) H^l = -\mathbf{R}({\Delta_k}^l)$

$\Longleftrightarrow$

(NI) ${\alpha_k}^{l^T} H^l + H^l {\alpha_k}^l = -\mathbf{R}({\Delta_k}^l)$.*(Ec. de Sylvester)*

siendo $\left\{ \ {\alpha_k}^l = {\alpha_k}^l({\Delta_k}^l) \right.$

Se resuelve por Kronecker para H^l

${\Delta_k}^{l+1} = {\Delta_k}^l + H^l$

proyectar $\Delta_k = \dfrac{\Delta_k{}^\star + \Delta_k{}^{\star T}}{2}$

$X_{k+1} = X_k + \Delta_k$

2. Relación con el problema de optimización multiobjetivo.

El problema de optimización multiobjetivo es de la forma

$$\left\{ \begin{array}{ll} \min\limits_{s.a} & F(x) \\ & x \in \mathcal{D}, \end{array} \right. \tag{2}$$

donde F, función multiobjetivo, está definida de $\mathbb{R}^m$ en $\mathbb{R}^k$ y $x = (x_1, x_2, \cdots, x_n)^T$, vector de decisión, pertenece a algún subconjunto de $\mathbb{R}^n$, región de factibilidad del problema. A la hora de optimizar, es fundamental el siguiente concepto

Definición 3 *(Vector Pareto Optimal) Sea* F *en las condiciones anteriores, se dice que* $x^\star \in \mathcal{D}$ *es Pareto Optimal si y sólo si no existe otro vector de decisión* $x \in \mathcal{D}$ *tal que* $F_i(x) \leq F_i(x^\star)$ *para todo* $i = 1, \cdots, k$ *y* $F_j(x) < F_j(x^\star)$ *al menos para un índice* j*.*

Definición 4 *(Vector Pareto Localmente Optimal) El vector de decisión* $x^\star \in \mathcal{D}$ *es Pareto Localmente Optimal si y sólo si existe* $\delta > 0$ *tal que* $x^\star \in \mathcal{D}$ *es Pareto Optimal en* $\mathcal{D} \cap \mathcal{E}(X_\star, \delta)$*.*

El objetivo de esta contribución es proponer un problema de la estructura anterior que sea equivalente a (1).

Sea vec es la transformación que aplicada a toda matriz apila sus columnas, de modo que

$$X \in \mathcal{S} \Longrightarrow vec(X) = x \in \mathbb{R}^{n^2}.$$

La transformación inversa

$$x \in \mathbb{R}^{n^2} \Longrightarrow vec^{-1}(x) = X \in \mathbb{R}^{n\times n} \Longleftrightarrow vec(X) = x \in \mathbb{R}^{n^2}$$

y la transformación $H : \mathbb{R}^{n^2} \longrightarrow \mathbb{R}^{n^2}$ dada por

$$H = \left(vec \circ \mathbf{F} \circ vec^{-1}\right)(x) = \left(h_1(x), h_2(x), \cdots h_{n^2}(x)\right),$$

siendo $x = (x_1, x_2, \cdots, x_{n^2}) = vec(X) = (X_{11}, X_{21}, X_{31}, \cdots, X_{nn})$.

Teniendo en cuenta que la aplicación que a cada matriz X asigna el vector $x = vec(X)$ y su inversa conservan las distancias, y llamando $vec(\mathcal{S})$ al subespacio que contiene a todas las vectorizaciones de matrices simétricas, el problema matricial (1) resulta equivalente al problema multiobjetivo

$$\begin{cases} \underset{s.a}{\text{mín}} & \left(vec \circ \mathbf{Q} \circ vec^{-1}\right)(x) \\ & x \in vec(\mathcal{S}), \end{cases} \tag{3}$$

tal como establecen los siguientes resultados:

Teorema 1 *Para cada $X \in \mathcal{S}$ sea $x = vec(X)$, $\mathbf{F} : \mathcal{S} \longrightarrow \mathcal{S}$ y H dada por $H = \left(vec \circ \mathbf{F} \circ vec^{-1}\right)$, $H : \mathbb{R}^{n^2} \longrightarrow \mathbb{R}^{n^2}$. Si $X_\star$ es un minimizador local de $\mathbf{F}$ en el sentido de Löwner entonces $x_\star$ es Pareto localmente optimal para H.*

El resultado que establece la recíproca del teorema anterior, no es válido para cualquier función $\mathbf{F}$, ya que la no negatividad de una matriz (es decir, todos sus elementos son no negativos) no necesariamente implica la propiedad de ser definida positiva [5].

Teorema 2 *Sea $\mathbf{Q}(X) = XAX + BX + XB$ con A simétrica y definida positiva y X, B simétricas. Sea $h = \left(vec \circ \mathbf{Q} \circ vec^{-1}\right)$ y $x_\star$ Pareto localmente optimal de H, entonces $X_\star = vec^{-1}(x_\star)$ es un minimizador de $\mathbf{Q}$ en el sentido de Löwner.*

3. Método de Newton para el problema Multiobjetivo.

El problema (2), ha sido resuelto por Fliege, Graña Drummond y Svaiter [1] empleando un algoritmo basado en el método de Newton. Demuestran que la *dirección de Newton en x*, $s(x)$, es la solución óptima del problema no diferenciable

$$\begin{cases} \underset{s.a}{\text{mín}} & \underset{j=1:m}{\text{máx}} & (\nabla F_j(x))^T s + \frac{1}{2} s^T \left(\nabla^2 F_j(x)\right)^T s \\ & s \in \mathbb{R}^{n^2}. \end{cases} \tag{4}$$

El valor óptimo para este problema es

$$\Theta(x) = \underset{s\in\mathbb{R}^n}{\text{ínf}} \underset{j=1:m}{\text{máx}} \quad (\nabla F_j(x))^T s + \frac{1}{2} s^T \left(\nabla^2 F_j(x)\right)^T s,$$

y es alcanzado en

$$s(x) = arg \underset{s\in\mathbb{R}^n}{\text{mín}} \underset{j=1:m}{\text{máx}} \quad (\nabla F_j(x))^T s + \frac{1}{2} s^T \left(\nabla^2 F_j(x)\right)^T s.$$

El punto clave de la propuesta en [1] es la equivalencia entre (4) y el siguiente problema

$$\begin{cases} \underset{s.a}{\text{mín}} & g(t,s) = t \\ & (\nabla F_j(x))^T s + \frac{1}{2} s^T \left(\nabla^2 F_j(x)\right)^T s - t \leq 0 \\ & (t,s) \in \mathbb{R} \times \mathbb{R}^n, \end{cases} \tag{5}$$

y planteadas para (5) las condiciones de Karush, Kuhn y Tucker, demuestran que existen multiplicadores de Lagrange, los cuales junto a $(\Theta(x), s(x))$ satisfacen las condiciones KKT siendo

$$s(x) = -\left(\sum_{j=1}^{m} \lambda_j(x)\nabla^2 F_j(x)\right)^{-1}\left(\sum_{j=1}^{m} \lambda_j(x)\nabla F_j(x)\right),$$

$$\Theta(x) = \sup_{\lambda\geq 0} \inf_{s\in\mathbb{R}^n} \mathcal{L}((t,s),\lambda) = \sup_{\lambda\geq 0,\ \sum_{j=1}^{m}\lambda_j = 1} \inf_{s\in\mathbb{R}^n} \sum_{j=1}^{m}\left(\left(\nabla F_j(x)\right)^T s + \frac{1}{2}s^T\left(\nabla^2 F_j(x)\right)^T s\right).$$

El algoritmo de Newton, es el siguiente

Algoritmo 3.1 (Algoritmo de Newton para Optimización Multicriterio)

- (Inicialización)
  Elegir $x_0 \in U$, $0 < \sigma < 1$, sea $k := 0$ y defina $\mathcal{J} = \left\{\frac{1}{2^n} \ : \ n = 0, 1, 2, \cdots\right\}$.
- a) Resolver (4) para obtener $s(x_k)$ y $\theta(x_k)$.
  b) Si $\theta(x_k) = 0$, parar. Caso contrario ir a c).
  c) (Búsqueda lineal) Elegir t_k como el mayor $t \in \mathcal{J}$ tal que

$$x_k + ts(x_k) \in U,$$

$$F_j(x_k + ts(x_k)) \leq F_j(x_k) + \sigma\theta(x_k)), \quad j = 1, \cdots, m.$$

  d) (Actualización) Definir $x_{k+1} = x_k + t_k s(x_k)$ y $k := k + 1$.

La prueba de convergencia puede ser consultada en [1].

4. Convergencia del algoritmo de minimización de una función cuadrática matricial con el orden de Löwner

Dado el problema (1), se quiere probar la convergencia del algoritmo que, basado en el método de Newton. resuelve el problema de minimización de una función cuadrática matricial con el orden de Löwner. Demostrado que en ciertas condiciones, el paso de Newton hallado para el problema multiobjetivo en el algoritmo (3.1) es la vectorización del paso generado por (1.1) y que por lo tanto es posible identificar la sucesión la sucesión de matrices generada por éste, con la sucesión de pasos obtenida vía el proceso iterativo diseñado por Fliege, Graña y Svaiter para (2), es posible probar la convergencia de la versión matricial de Newton.

Teorema 3 *En las hipótesis del Lema, la sucesión $\{\Delta_k\}_k$ generada por el algoritmo (1.1) converge q-cuadráticamente a un minimizador de $\mathbf{Q}(X)$ en el sentido de Löwner*

$$\left\|X_{k+1} - X_\star\right\|_{\mathtt{F}} \leq \eta\left\|X_k - X_\star\right\|_{\mathtt{F}}^2.$$

Referencias

[1] J. Fliege and L. M. Graña Drummond and B. F. Svaiter, *Newton's Method for Multiobjective Optimization*, SIAM Journal on Optimization, Vol. 20 (2009), pp. 602-626.

[2] K. M. Miettinen, *Nonlinear Multiobjective Optimization*, Kluwer's International Series, Stanford, California, (2002)

[3] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, (1992)

[4] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, (1991)

[5] M. G. Eberle and M. C. Maciel *Relación entre el Poblema de Optimización Matricial en el Sentido de Loewener y el Problema de Optimización Multicriterio*, Actas SIO 2009, Simposio Argentino de Investigación Operativa (2009), pp 84-97.

UN ALGORITMO DE LAGRANGIANO AUMENTADO CON DIFERENTES ESTRATEGIAS EN EL CÁLCULO DE LA INFORMACIÓN DE SEGUNDO ORDEN

Graciela M. Croceri[♭], Karina Navarro Alvarez[†] y Graciela N. Sottosanto[♭]

[♭]*Depto. de Matemática, Universidad Nacional del Comahue, Santa Fe 1400, 8300 Neuquén, Argentina, gcroceri@uncoma.edu.ar, gsottos@uncoma.edu.ar*
[†]*Universidad Nacional de la Patagonia Austral, Unidad Académica Caleta Olivia, División Ciencias Exactas, knavarro@uaco.unpa.edu.ar*

Resumen: En este trabajo se introducen dos técnicas a fin de mejorar la performance de un método iterativo existente para resolver el problema de minimización con restricciones de igualdad. El algoritmo, desarrollado originalmente por M.C.Maciel y G.N.Sottosanto, está basado en la minimización secuencial del Lagrangiano aumentado y combina un método de gradiente conjugado y región de confianza. Posteriormente, fue aplicado a la resolución de problemas de cuadrados mínimos por G.Croceri, M.C.Maciel y G.Sottosanto donde se incorporó una aproximación secante estructurada del tipo BFGS para aprovechar la estructura típica del problema de cuadrados mínimos.
En esta contribución adicionamos dos técnicas basadas en el uso de la información de la iteración actual y de los pasos previos para actualizar la información de segundo orden y construir el modelo cuadrático de la función aumentada de Lagrange. El algoritmo se aplica a la resolución de problemas de cuadrados mínimos no lineales con restricciones.

Palabras clave: *Lagrangiano aumentado, cuadrados mínimos, actualizaciones secantes*
2000 AMS Subject Classification: 90C30 - 90C53

1. INTRODUCCIÓN

En este trabajo se introducen dos técnicas a fin de mejorar la performance de un método iterativo existente para resolver el problema de minimización con restricciones de igualdad. El algoritmo está basado en la minimización secuencial del Lagrangiano aumentado y combina un método de gradiente conjugado y región de confianza. Fue desarrollado, originalmente, por M.C.Maciel y G.N.Sottosanto [7] y aplicado a la resolución de problemas de cuadrados mínimos por G.Croceri, M.C.Maciel y G.Sottosanto [2] donde se incorporó una aproximación secante estructurada (SBFGS) del tipo BFGS que aprovecha la estructura típica del problema de cuadrados mínimos.
En esta contribución adicionamos dos técnicas basadas en el uso de la información de la iteración actual y de los pasos previos para actualizar la información de segundo orden y construir el modelo cuadrático de la función aumentada de Lagrange. El algoritmo se aplica a la resolución de problemas de cuadrados mínimos no lineales con restricciones de igualdad. Las actualizaciones de la información de segundo orden están inspiradas en los trabajos de J. Eriksson [5] y Z. Wei, G. Li y L. Qi [9] que denominaremos Broyden con interpolación (BwI) y cuasi Newton modificadas (MqN), respectivamente.

2. EL PROBLEMA

El problema de cuadrados mínimos no lineales con restricciones de igualdad es

$$\begin{aligned} \min_{s.a} \quad & \frac{1}{2}\|F(x)\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2 \\ & c_j(x) = 0, \;\; j = 1, \ldots, p, \end{aligned}$$

la función residual $F : \mathbb{R}^n \to \mathbb{R}^m$ es, generalmente, no lineal, $f_i(x)$ denota la i-ésima componente de la función F y $m \geq n$.
El Lagrangiano aumentado asociado a este problema es

$$L_\mu(x, \lambda) = \frac{1}{2}\|F(x)\|_2^2 + \lambda^T c(x) + \frac{1}{2\mu}\|c(x)\|_2^2. \tag{1}$$

Su vector gradiente es

$$\nabla L_\mu(x,\lambda) = \nabla F(x)F(x) + \nabla c(x)\left(\frac{1}{\mu}c(x) + \lambda\right). \tag{2}$$

La matriz Hessiana del Lagrangiano aumentado es

$$\begin{aligned}\nabla^2 L_{\mu_k}(x,\lambda) &= \nabla F(x)\nabla F(x)^T + \sum_{i=1}^{m} f_i(x)\nabla^2 f_i(x) + \frac{1}{\mu}\nabla c(x)\nabla c(x)^T \\ &\quad + \frac{1}{\mu}\sum_{i=1}^{p} c_i(x)\nabla^2 c_i(x) + \sum_{i=1}^{p}\lambda\nabla^2 c_i(x).\end{aligned}$$

La estructura de esta matriz es importante en la formulación de cualquier problema de cuadrados mínimos. En ese sentido, consideramos $\nabla^2 L_{\mu_k}(x,\lambda) = G_\mu(x) + S_\mu(x,\lambda)$ donde

$$G_\mu(x) = \nabla F(x)\nabla F(x)^T + \frac{1}{\mu}\nabla c(x)\nabla c(x)^T,$$

$$S_\mu(x,\lambda) = \sum_{i=1}^{m} f_i(x)\nabla^2 f_i(x) + \sum_{i=1}^{p}\frac{1}{\mu}c_i(x)\nabla^2 c_i(x) + \sum_{i=1}^{p}(\lambda_k)_i\nabla^2 c_i(x).$$

Como $\nabla F(x)$ y $\nabla c(x)$ están disponibles, ya sea analíticamente o usando diferencias finitas, el término $G_\mu(x)$ está también disponible cuando se calcula la matriz Hessiana.

3. EL ALGORITMO

En cada iteración externa se encuentra una solución aproximada x_{k+1} de

$$\min_x L_{\mu_k}(x,\lambda_k), \tag{3}$$

donde λ_k es el estimado actual del vector de multiplicadores de Lagrange y $\mu_k > 0$ es el parámetro de penalización. Este parámetro, al igual que el estimado del multiplicador se actualiza al final de cada iteración externa. El subproblema irrestricto se resuelve usando el método de gradiente conjugado que tiene incorporada una estrategia de región de confianza. Por lo tanto, el paso se obtiene como solución de

$$\min_{\|s\|_2 \le \delta_k} Q_k(s,\lambda_k,\mu_k),$$

donde Q_k es un modelo cuadrático del Lagrangiano aumentado alrededor de x_k y $\delta_k > 0$ es el radio de la región de confianza.

Si la reducción de la función $L_{\mu_k}(x,\lambda_k)$ en el minimizador del modelo cuadrático es suficiente, se acepta el correspondiente paso de prueba como nuevo iterado, de lo contrario, se reduce el radio de la región de confianza siguiendo esquemas estándar [1].

Claramente, el método para resolver el subproblema irrestricto define iteraciones internas en cada iteración externa del algoritmo de penalización.

Con respecto al criterio de parada, dada una sucesión $\{\epsilon_k\}$ que converge a cero, cada iteración interna termina en $(x^\star)_{k+1}$ como solución aproximada a (3) si

$$\|\nabla L_{\mu_k}((x_\star)_{k+1},\lambda_k)\|_2 < \epsilon_k.$$

En el algoritmo principal, un iterado $(x_\star)_{k+1}$ y su multiplicador asociado se declara *próximo* a un punto de Karush, Kuhn and Tucker $(x_\star,\lambda_\star)$ del problema si

$$\|\nabla l((x_\star)_{k+1},\lambda_k)\|_2 \le \epsilon_1, \quad \|c((x_\star)_{k+1})\|_2 \le \epsilon_2,$$

donde $l(x,\lambda) = \frac{1}{2}\|F(x)\|_2^2 + \lambda^T c(x)$ es la función de Lagrange y ϵ_1, ϵ_2 son constantes positivas. Si alguna de estas condiciones falla, el parámetro de penalización y el multiplicador se actualizan.

3.1. ACTUALIZACIÓN SECANTE ESTRUCTURADA (SBFGS)

Tradicionalmente un método secante para el problema de minimización sin restricciones consiste en un método iterativo $x_+ = x + s$ donde en cada paso se actualiza una matriz $B_+ = B + \Delta(s, y, B, v)$ tal que $Bs = -g$, g es el gradiente de la función objetivo, $B_+ s = y$ con $y = g(x_+) - g(x)$ y la escala $v = v(s, y, B)$, para el caso de actualizaciones BFGS, toma la forma

$$v = y + \left(\frac{y^T s}{s^T B s}\right)^{1/2} Bs.$$

La matriz B_+ se interpreta como una aproximación de la matrix Hessiana de la función objetivo en el nuevo iterado.
La estrategia que se presentó en [2] para actualizar la matriz Hessiana en una forma estructurada sigue las ideas de J.E Dennis, H.J. Martínez y R.A. Tapia [3]. En nuestro caso, consideramos $B^s = G_\mu(x_+) + A$ donde A es una aproximación de $S_\mu(x, \lambda)$, y la actualización secante de B^s es

$$B_+ = B^s + \Delta(s, y^s, B^s, v(s, y^s, B^s)).$$

Como para cualquier v resulta $\Delta(s, y^s, B^s, v) = \Delta(s, y^\#, A, v)$, donde $A_+ s = y^\#$ definimos

$$B_+ = G_\mu(x_+) + A + \Delta(s, y^\#, A, v(s, y^s, B^s)).$$

En nuestro algoritmo de Lagrangiano aumentado resulta

$$y^\# = (\nabla F(x_+) - \nabla F(x))\, F(x_+) + (\nabla c(x_+) - \nabla c(x)) \left(\frac{1}{\mu_k} c(x_+) + \lambda_k\right),$$

mientras que

$$y^s = \nabla L(x_+, \lambda_k, \mu_k) - \nabla L(x, \lambda_k, \mu_k).$$

3.2. ACTUALIZACIÓN DE BROYDEN CON INTERPOLACIÓN (BWI)

En [5] se propone una familia de actualizaciones de Broyden con estrategias de interpolación, que aprovecha las buenas propiedades geométricas del problema de cuadrados mínimos sin restricciones y cuando son incorporadas a un algoritmo permiten acelerar su convergencia. En este trabajo, la estrategia BwI se adapta para el problema de minimización del Lagrangiano aumentado. La aproximación de la parte $S_\mu(x, \lambda)$ de la matrix Hessiana se efectua en dos etapas: en la primera es posible aplicar una actualización tipo BFGS o bien una actualización simétrica de rango uno (SR1) y en la segunda se efectua una corrección de rango uno basada en información de los dos pasos previos. La corrección modifica el factor de escala v y permite aproximar la derivada segunda de la función objetivo en la dirección del paso $s_p = x_k - x_{k-2}$.

3.3. ACTUALIZACIÓN CUASI NEWTON MODIFICADA (MQN)

En [9] los autores propusieron modificar la actualización cuasi Newton original $B_+ s_k = y_k$ por otra $B_+ s = \tilde{y}$ en la cual $\tilde{y}$ se toma como la suma de y_k más $A_k s_k$ para alguna matriz A_k que se construye en base a información de la matriz Hessiana de la función objetivo. En este trabajo aplicamos esa construcción a la parte $S_\mu(x, \lambda)$ de la matriz Hessiana del Lagrangiano aumentado. De esta manera, la matriz A se construye usando información de la función Lagrangiano aumentado y de su grandiente en el punto anterior y en el actual y la información del paso previo.

4. RESULTADOS NUMÉRICOS

El algoritmo, con las diferentes actualizaciones de la matriz Hessiana, ha sido codificado en Matlab. A fin de evaluar la performance, se han resuelto inicialmente un grupo de 38 problemas test tomados de Hock y Schittkowski [6] y Schittkowski [8]. Se ha trabajado tanto con problemas de cuadrados mínimos con restricciones de igualdad y desigualdad. En el caso de estos últimos, para restricciones de la forma $g_i(x) \leq 0$,

$i \in I$ se han tranformado al caso de igualdad, agregando variables de holgura en la forma $g_i(x) + z_i^2 = 0$, $i \in I$.

En la figura 1 (a) y (b), se muestran los resultados numéricos obtenidos por medio de la técnica de *performance profiles* [4] graficados para diferentes rangos a fin de observar distintas áreas de interés. Se representa la performance basada en el tiempo de ejecución para las tres actualizaciones SBFGS, BwI y MqN. En el caso de la actualización BwI se implementó la primera etapa tanto con la técnica BFGS como SR1 resultando la primera en menor tiempo de ejecución.

En el caso (a) se muestra el desempeño en el intervalo $[0, 144]$. Se observa que, si bien MqN y SBFGS son ambas competitivas, la actualización MqN es la más eficiente, en particular para aquellos problemas que requieren mayor tiempo de ejecución. En la figura (b) el intervalo de tiempo se ha reducido a $[0, 20]$, en ella es claro que la actualización SBFGS supera a MqN.

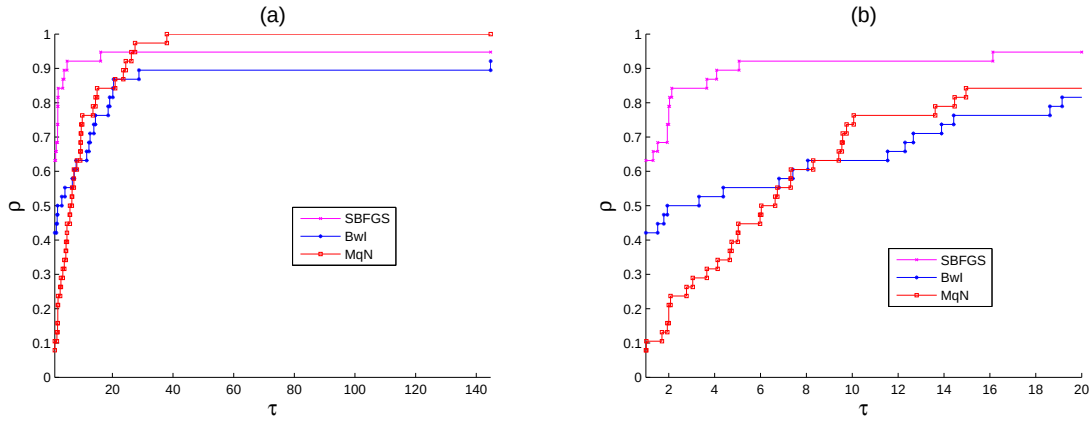


Figura 1: Performance Profile para (a) $[0, 144]$, (b) $[0, 20]$.

AGRADECIMIENTOS

Este trabajo ha sido realizado con el apoyo de la Universidad Nacional del Comahue, Proyecto E081/09.

REFERENCIAS

[1] A.R. CONN, N.I.M. GOULD AND P.L. TOINT, *Trust-Region Methods*, MPS-SIAM Series on Optimization, 2000.

[2] G.M. CROCERI, M.C. MACIEL AND G.N. SOTTOSANTO, *Augmented penalty algorithms based on BFGS secant approximations and trust regions*, Applied Numerical Mathematics, 57 (2007), pp. 320-334.

[3] J.E. DENNIS, H.J. MARTÍNEZ AND R.A. TAPIA, *Convergence theory for the structured BFGS secant methods with an application to nonlinear least squares*, Journal of Optimization Theory and Applications 61 (2) (1989), pp. 161-178.

[4] E. DOLAN, J. MORÉ, *Bechmarking optimization sofware with performance profiles*, Mathematical Programming 91 (2002), pp. 201-203.

[5] J. ERIKSSON, *Quasi-Newton methods for nonlinear least squares focusing on curvatures*, BIT, Vol. 39 (2) (1984), pp. 228-254.

[6] W. HOCK AND K. SCHITTKOWSKI, *Test examples for nonlinear programming codes*, M.Beckmann and H. P. Künzi, eds., Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1981.

[7] M.C. MACIEL AND G.N. SOTTOSANTO, *An augmented penalization algorithm for the equality constrained minimization problem*, Seleta do XXIV CNMAC Vol. 3(2) (2002), pp. 171-180.

[8] K. SCHITTKOWSKI, *More test examples for nonlinear programming codes*, M. Beckmann and W. Krelle, eds., Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1987.

[9] Z. WEI, G. LI AND L. QI, *New quasi-Newton methods for unconstrained optimization problems*, Applied Mathematics and Computation, 175 (2006), pp. 1156-1188.

CHARACTERIZATION OF THE NONEMPTYNESS AND BOUNDEDNESS OF SOLUTION SETS IN VECTOR OPTIMIZATION*

Felipe Lara[♭] y Fabián Flores-Bazán[♭]

[♭]*Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-mail:* felipelara@udec.cl, fflores@ing-mat.udec.cl

Resumen: Usando herramientas del análisis asintótico hemos desarrollado finas estimaciones para los conjuntos de soluciones eficientes y débilmente eficientes (y sus conos asintóticos) para problemas de optimización multiobjetivo en caso convexo y casiconvexo, además de estimaciones mediante escalarizaciones lineales del problema multiobjetivo original. Finalmente entregamos una nueva caracterización para el conjunto de soluciones eficientes para cualquier cono convexo cerrado sin ninguna hipótesis de convexidad.

Palabras clave: *Optimización multiobjetivo no convexa, eficiencia, débil eficiencia, conos y funciones asintóticas.*
2000 AMS Subject Classification: 90C25-90C26-90C29-90C30

1. INTRODUCCION

Sea $K \subseteq \mathbb{R}^n$ un conjunto convexo cerrado no vacío y sea $F : K \to \mathbb{R}^m$ una función vectorial donde $F(x) = (f_1(x), f_2(x), ..., f_m(x))$, con $f_i : K \to \mathbb{R}$, para todo $i \in \{1, 2, ..., m\}$, estamos interesados en estudiar el siguiente problema de minimización multiobjetivo,

$$\min_{x \in K} \; F(x) \tag{P}$$

Claramente la noción de óptimo se ve distorcionada en esta nueva realidad, pues un punto puede ser mínimo para una función, pero de máximo para otra, en general, es significativamente díficil encontrar puntos que sean de mínimo(máximo) para cada una de las funciones objetivos.

Lo anterior produce que existan muchos conceptos de solución asociados al problema (P), como ideales, propiamente eficientes, eficientes y débilmente eficientes. Una solución propiamente eficiente es una solución eficiente que elimina las compensaciones no acotadas entre los objetivos. Esta noción fue introducida por Geoffrion en [7], nosotros utilizaremos un noción equivalente dado por Benson en [1].

Estos conceptos de solución en optimización multiobjetivo están asociados a un cono, dado $P \subseteq \mathbb{R}^m$ un cono convexo cerrado, decimos que $\overline{x} \in K$ es solución,

a) ideal, si $F(x) - F(\overline{x}) \in P$, para todo $x \in K$. Denotaremos al conjuto de todas las soluciones ideales por I.

b) propiamente eficiente, si $\overline{\text{cone}}(F(K) - F(\overline{x}) + P) \cap (-P) = \{0\}$. Donde cone$(A)$ denota el menor cono que contiene a A. Denotaremos al conjunto de todas las soluciones propiamente eficientes por E_P.

c) eficiente, si $F(x) - F(\overline{x}) \notin -P \backslash l(P)$, para todo $x \in K$, donde $l(P) = P \cap (-P)$. Denotaremos al conjunto de todas las soluciones eficientes por E.

d) débilmente eficiente, si $\operatorname{int} P \neq \emptyset$ y $F(x) - F(\overline{x}) \notin -\operatorname{int} P$, para todo $x \in K$. Denotaremos al conjunto de todas las soluciones débilmente eficientes por E_W.

Es claro de la definición que $I \subseteq E_P \subseteq E \subseteq E_W$.

*Basado en el artículo "Inner and outer estimates for solution sets and their asymptotic cones in vector optimization", submitted to Optimization Letters.

Resulta a veces complicado calcular los conjuntos E_P, E o E_W de manera directa, es por esto que es importante realizar aproximaciones para estos conjuntos, con conjuntos definidos a partir de los puntos de óptimo para problemas escalarizados.

Consideremos la función $h_q : K \to \mathbb{R}$, dada por $h_q(.) = \langle q, F(.)\rangle$, para todo $q \in P^*$, donde P^* denota el cono polar positivo de P, y consideremos su conjunto de mínimos $\operatorname{argmin}_K h_q$.

Definamos el siguiente problema de minimización escalar, dado $p \in K$ y $q_0 \in \operatorname{int} P^*$,

$$\min_{x \in K_p} h_{q_0}(x) \tag{P_p}$$

donde $K_p = \{x \in K : F(x) - F(p) \in -P\}$.

Finalmente recordemos que dado un conjunto $K \subseteq \mathbb{R}^n$, se define su cono asintótico K^∞ como,

$$K^\infty = \{v \in \mathbb{R}^n : \ \exists \ \{x_n\}_{n\in\mathbb{N}} \subseteq K, \ \exists \ t_n \downarrow 0, \ t_n x_n \to v\}.$$

Si $K = \emptyset$, entonces por convención tenemos que $K^\infty = (\emptyset)^\infty = \{0\}$.

Para cualquier función $g : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$, su función asintótica esta dada por g^∞, donde $g^\infty : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ y es tal que,

$$\operatorname{epi}(g^\infty) = (\operatorname{epi}(g))^\infty.$$

donde $\operatorname{epi}(f) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : \ f(x) \le t\}$.

2. Resultados principales

Lema 2.1. Sea P un cono convexo cerrado pointed con $\operatorname{int} P \neq \emptyset$, luego,

$$\operatorname{argmin}_K \langle q_0, F(x)\rangle \subseteq E_P \subseteq E \subseteq \bigcup_{p\in K} \operatorname{argmin}_{K_p} \langle q_0, F(x)\rangle \subseteq E_W, \ \ \forall \ q_0 \in \operatorname{int} P^*.$$

Teorema 2.1. Sea P un cono convexo cerrado y h_q convexa, para todo $q \in J$,

a) Si $E \neq \emptyset$ y

$$\sup_{x\in E} \ \{\sup_{q\in J} \ h_q(x)\} \ < \ +\infty \tag{1}$$

entonces,

$$E^\infty = \bigcap_{q\in J} \{v \in K^\infty : \ h_q^\infty(v) \le 0\} = R_J.$$

b) Si $\operatorname{int} P \neq \emptyset$, $E_W \neq \emptyset$ y

$$\sup_{x\in E_W} \ \{\sup_{q\in J} \ h_q(x)\} \ < \ +\infty \tag{2}$$

entonces

$$E_W^\infty = \bigcap_{q\in J} \{v \in K^\infty : \ h_q^\infty(v) \le 0\} = R_J.$$

Prueba. Solo probaremos a), la demostración de b) es análoga.

($\supseteq$) Consideremos primero el siguiente cono, el cual coincide con el cono R_J cuando $J = P^*$ y cada componente es convexa,

$$R_P \doteq \bigcap_{y\in K}\bigcap_{\lambda>0}\{v \in K^\infty : \ F(y+\lambda v) - F(y) \in -P\}.$$

Sea $v \in R_P$ y $\overline{x} \in E$, demostraremos que $\overline{x} + tv \in E$, para todo $t > 0$. Dado $y \in K$, se tiene que,

$$F(y) - F(\overline{x}+tv) \ = \ F(y) - F(\overline{x}) + F(\overline{x}) - F(\overline{x}+tv),$$

entonces,

$$F(y) - F(\overline{x}) + F(\overline{x}) - F(\overline{x}+tv) \ \in \ (-P\backslash l(P))^c + P \ \subseteq \ (-P\backslash l(P))^c,$$

es decir, $\overline{x} + tv \in E$, para todo $t > 0$ y por lo tanto $v \in E^\infty$.

($\subseteq$) Sea entonces $v \in E^\infty_W, ||v|| = 1$, luego existe $\{x_n\}_{n\in\mathbb{N}} \subseteq E$, con $\|x_n\| \to +\infty$, tal que $\dfrac{x_n}{\|x_n\|} \to v$. Por (1) tenemos que existe $M > 0$ tal que, $h_q(x_n) \ \leq \ M$, para todo $n \in \mathbb{N}$ y $q \in J$.

Luego, dado $q \in J$ arbitrario, se tiene por definición de función asintótica que,

$$h^\infty_q(v) \ \leq \ \liminf_{n\to\infty} \frac{1}{\|x_n\|} h_q(\frac{\|x_n\|x_n}{\|x_n\|})$$

$$= \ \liminf_{n\to\infty} \frac{h_q(x_n)}{\|x_n\|} \ \leq \ \liminf_{n\to\infty} \frac{M}{\|x_n\|} \ = \ 0,$$

entonces $h^\infty_q(v) \leq 0$, para todo $q \in J$ y por lo tanto $v \in \bigcap_{q\in J}\{v \in K^\infty : \ h^\infty_q(v) \leq 0\}$, es decir,

$$E^\infty \subseteq \bigcap_{q\in J}\{v \in K^\infty : \ h^\infty_q(v) \leq 0\}.$$

Teorema 2.2. Sea P un cono polihédrico y sea h_{q_i} una función convexa y continua, para todo $i \in \{1, 2, ..., k\}$, consideremos las siguientes afirmaciones,

a) $E \neq \emptyset$ y acotado.

b) $R_P = \{0\}$ y la hipótesis (1) se tiene para $J = \{1, 2, ..., k\}$.

Si int $P \neq \emptyset$,

c) $R_P = \{0\}$ y la hipótesis (2) se tiene para $J = \{1, 2, ..., k\}$.

d) $E_W \neq \emptyset$ y compacto.

e) $\operatorname{argmin}_K h_{q_i} \neq \emptyset$ y compacto, para todo $i \in \{1, 2, ..., k\}$.

f) $R_W = \{0\}$.

luego se tiene que,

$$\text{a)} \Leftrightarrow \text{b)} \Leftarrow \text{c)} \Leftrightarrow \text{d)} \Leftrightarrow \text{e)} \Leftrightarrow \text{f)}.$$

Prueba. a) $\Leftrightarrow$ b) y c) $\Leftrightarrow$ d), consecuencia directa del Teorema 2.1, junto con el Lema 5.5 y Teorema 5.8 en [5], mientras que c) $\Rightarrow$ b) es producto de que $E \subseteq E_W$.

Finalmente, d) $\Leftrightarrow$ e) $\Leftrightarrow$ f) es el teorema 5.1 en [6].

Teorema 2.3. Sea P un cono cerrado, luego tenemos que

a) $\overline{x} \in E$ si y sólo si existe $p \in K$ tal que $\overline{x} \in \operatorname{argmin}_{K_p} h_q$ para todo $q \in P^*$, es decir,

$$E = \bigcup_{p \in K} \bigcap_{q \in P^*} \operatorname{argmin}_{K_p} h_q. \tag{3}$$

b) Si P es polihédrico y pointed, entonces dado $q_0 \in \operatorname{int} P^*$ se tiene que, $\overline{x} \in E$ si y sólo si existe $p \in K$ tal que $\overline{x} \in \operatorname{argmin}_{K_p} h_{q_0}$, es decir,

$$E = \bigcup_{p \in K} \operatorname{argmin}_{K_p} h_{q_0}, \ \ \forall \ q_0 \in \operatorname{int} P^*. \tag{4}$$

Prueba. a) $\Rightarrow$) Sea $\overline{x} \in E$ y supongamos que para todo $p \in K$ existe $q_p \in P^*$ tal que $\overline{x} \notin \operatorname{argmin}_{K_p} h_{q_p}$. Tomando $p = \overline{x}$ tenemos que existe $q_0 \in P^*$ y $z \in K_{\overline{x}}$ tal que $h_{q_0}(z) < h_{q_0}(\overline{x})$, lo cual implica que,

$$F(z) - F(\overline{x}) \notin P \tag{5}$$

Como $z \in K_{\overline{x}}$ se tiene que,

$$F(z) - F(\overline{x}) \in -P \tag{6}$$

Finalmente, de (5) y (6) se tiene que $F(z) - F(\overline{x}) \in (-P) \backslash P$, contradiciendo el hecho de que $\overline{x} \in E$.

$\Leftarrow$) Sea $p \in K$ tal que $\overline{x} \in \operatorname{argmin}_{K_q} h_q$, para todo $q \in P^*$ y supongamos por el contrario que $\overline{x} \notin E$, entonces existe $z \in K$ tal que,

$$F(x) - F(\overline{x}) \in (-P) \backslash P.$$

Luego $z \in K_p$ y existe $q_0 \in P^* \backslash \{0\}$ tal que $h_{q_0}(z) < h_{q_0}(\overline{x})$, lo cual es una contradicción.

b) Consecuencia directa del Lema 3.1 en [8].

REFERENCIAS

[1] Benson, H. *An Improved Definition of Proper Efficiency for Vector Minimization with Respect to Cones*, Journal of Mathematical Analysis and Applications, Vol. 71, (1979) pp. 232-241.

[2] Deng. S. *Boundedness and Nonemptiness of the Efficient Solution Sets in Multiobjective Optimization*, Journal of Optimization Theory and Applications, Vol 144, (2010) pp. 29-42.

[3] Deng, S. *On the efficient Solution in Vector Optimization*, Journal of Optimization Theory and Applications, Vol. 96, (1998) pp. 201-209.

[4] Flores-Bazán, F. *Ideal, weakly efficient solutions for vector optimization problems*, Mathematical Programing, Vol. 93, (2002) pp. 453-475.

[5] Flores-Bazán, F. *Existence Theory for Finite-Dimensional Pseudomonotone Equilibrium Problems*, Acta Applicandae Mathematicae, Vol. 77, (2003) pp. 249-297.

[6] Flores-Bazán, F. and Vera, C. *Characterization of the Nonemptyness and Compacteness of Solution Sets in Convex and Nonconvex Vector Optimization*, Journal of Optimization Theory and Applications, Vol. 130, (2006) pp. 185-207.

[7] Geoffrion, A. *Proper Efficient and Theory of Vector Optimization*, Journal of Mathematical Analysis and Applications, Vol. 22, (1968) pp. 618-630.

[8] Huang, X.X. and Yang, X.Q. *On Characterizations of Proper Efficiency for Nonconvex Multiobjective Optimization*, Journal of Global Optimization, Vol. 23, (2002) 213-231.

BILEVEL OPTIMIZATION FOR THE DESIGN OF DISTILLATION COLUMNS

Ana Friedlander[♭] and Esdras P. Carvalho[†]

[♭]*Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13081-970 Campinas, SP, Brazil, friedlan@ime.unicamp.br*
[†]*Department of Mathematics, State University of Maringá, 87020-900 Maringá, PR, Brazil, epcarvalho@uem.br*

Abstract: In this work, phase changes in vapor-liquid equilibrium systems are considered. The mathematical model of the separation process is presented. The number of phases and the phase equilibria on each tray are determined by minimizing the Gibbs free energy, allowing a variable number of phases on each tray. The minimization problems are embedded within a larger problem that minimizes the operating cost of the column, creating a bilevel optimization problem. The constraints can be identified as mass, component, energy balances and bounds on flows variables and temperatures. The bilevel optimization problem is solved via an inexact restoration strategy, and the problems at each level are solved by using adequate algorithms, according to their structures and characteristics.

Keywords: *bilevel optimization, simulation, separation process, inexact restoration*
2000 AMS Subject Classification: 90C30 - 90C90

1 INTRODUCTION

An interesting class of problems is the design of multi-stage processes involving chemical equilibrium. Optimization of such systems requires the identification of the phases present in each stage as part of the optimization process. Phase equilibrium problems are characterized by implicit discontinuities in which different sets of equations are applicable depending on the number of phases present at equilibrium. We can pose the phase identification problem as one of minimizing Gibbs free energy. The design of chemical equilibrium based processes can be modeled as follows [8]:

$$\begin{array}{ll} \min\limits_{u,v} & F(u,v) \\ \text{s. t.} & G(u) = 0,\ u \in U \\ & \min\limits_{v} f(u,v) \\ & \text{s. t.} \quad g(u,v) = 0,\ v \in V \end{array} \tag{1}$$

where $\Omega = U \times V$, $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ are multidimensional bounded boxes, $F : \mathbb{R}^{n+m} \to \mathbb{R}$, $G : \mathbb{R}^n \to \mathbb{R}^q$, and $g : \mathbb{R}^{n+m} \to \mathbb{R}^p$. All functions are assumed to be continuous and twice differentiable in Ω. The essential feature of a bilevel problem is that a subset of its variables is required to solve another optimization problem, parameterized by the remaining variables, called the lower-level problem [1]. In chemical equilibrium based process design, the outer objective will correspond to the reboiler heat input, which account for main part of operating cost, and the outer level constraints can be identified as mass, component, energy balances and bounds on flow variables and temperatures. The inner optimization problem is one of minimizing Gibbs free energy [8].

Bilevel programming problems have been studied since the seventies, and in the last years some surveys and bibliographic reviews appeared. History, applications, algorithms, theoretical questions and almost all relevant references can be found in [3, 9]. We subdivide the algorithms for solving bilevel programming problems into three classes, following Dempe [3]: (i) algorithms that solve them globally, (ii) methods that find stationary points or points that satisfy some local optimality condition, and (iii) heuristics. Also inside these classes there are problems with particular structures for which special algorithms have been designed. In [3] extensive bibliography is given that covers most of the significant work done in the last years [1]. In this work, based on the recent work of Fischer and Friedlander [4], we present one algorithm, that belongs to the second class in Dempe's classification, for the solution of bilevel programming problems.

The key idea is to reformulate the problem (1) as a single-level problem subject to optimality conditions of the lower-level problem and to apply an inexact restoration approach to the resulting problem. The inexact restoration method deals separately with feasibility and optimality at each iteration. In the feasibility stage, called restoration phase, it seeks for a more feasible point. In the optimality phase, after computing an inexactly restored point, the new iterate (in an approximation of the tangent set) is obtained by means of a single line-search procedure that only involves a penalty function. If the new iterate is not accepted, the size of the trust region is reduced. The user is free to use any algorithm in each phase, making the choice of problem-oriented solvers possible. In this work a new general scheme for inexact restoration methods is introduced. This algorithm differs from previous methods, in which the tangent phase needs both a line search based on the objective function (or its Lagrangian) [1, 2, 6, 7] and a confirmation based on a penalty function or a filter decision scheme. Besides its simplicity the new scheme enjoys some nice theoretical properties. In particular, a key condition for the inexact restoration step could be weakened [4]. The work is organized as follows. In Section 2 the mathematical model of a distillation column is presented. In Section 3 a summary of the fundamentals of the inexact restoration method is presented. In Section 4 we describe the algorithm based on the recent work of Fischer and Friedlander [4].

2 MATHEMATICAL MODEL

Consider a distillation column having N trays or stages, with reboiler (as stage $i = 1$) and total condenser (as stage $i = N$), and j components in the feed. Feed enters on tray S and the column is operated at a pressure P and we neglect pressure drop across the column. F is the feed flow rate; $L_i(V_i)$, flow rate of liquid(vapor) leaving tray i; T_i, temperature of tray i; H_F, feed enthalpy; $H_{L,i}(H_{V,i})$, enthalpy of liquid(vapor) leaving tray i; x_F, feed composition; $x_{i,j}(y_{i,j})$, molefraction of j in liquid(vapor) leaving tray i; ξ_i, relaxation parameter; $s_{L/V,i}$, slack variables; A_j, B_j, C_j, Antoine coefficients; D, distillate flow rate; and $Q_R(Q_C)$, heat load on reboiler(condenser). We assume that ideal vapor-liquid equilibrium holds and that there is no mass transfer resistance. The governing mass balances, equilibrium relations, mole fraction summations and energy balances (MESH equations) for modeling distillation are as follows:

Total mass balances:

$$
\begin{array}{l}
L_1 + V_1 - L_2 = 0 \\
L_i + V_i - L_{i+1} - V_{i-1} = 0 \;\; i = 2, \ldots, N+1, i \neq S+1 \\
L_i + V_i - L_{i+1} - V_{i-1} - F = 0 \;\; i = S+1 \\
L_{N+2} + D - V_{N+1} = 0
\end{array}
$$

Enthalpy balances:

$$
\begin{array}{l}
L_1 H_{L,1} + V_1 H_{V,1} - L_2 H_{L,2} - Q_R = 0 \\
L_i H_{L,i} + V_i H_{V,i} - L_{i+1} HL, i+1 - V_{i-1} HV, i-1 = 0 \;\; i = 2, \ldots, N+1 \\
L_i H_{L,i} + V_i H_{V,i} - L_{i+1} H_{L,i+1} - V_{i-1} H_{V,i-1} - F H_F = 0 \;\; i = S+1 \\
V_{N+1} H_{V,N+1} - (L_{N+2} + D) H_{L,N+2} - Q_C = 0
\end{array}
$$

Molefraction balances

$$
\sum_{j=1}^{m} y_{i,j} - \sum_{j=1}^{m} x_{i,j} = 0 \;\; i = 1, \ldots, N+2
$$

Relaxed phase equilibrium

$$
y_{i,j} - \frac{\xi_i}{P} exp(A_j + \frac{B_i}{C_j + T_i}) x_{i,j} = 0 \;\; j = 1, \ldots, m, \; i = 1, \ldots, N+2
$$

Bounds

$$
\begin{array}{l}
s_{L,i}, s_{V,i}, V_i \geqslant 0 \;\; i = 1, \ldots, N+1 \\
L_i, T_i \geqslant 0, \;\; i = 1, \ldots, N+2 \\
D, Q_R, Q_C \geqslant 0 \\
0 \leqslant y_{i,j}, x_{i,j} \leqslant 1 \;\; j = 1, \ldots, m, \; i = 1, \ldots, N+2
\end{array}
$$

Component mass balances

$$
\begin{array}{l}
L_1 x_{1,j} + V_1 y_{1,j} - L_2 x_{2,j} = 0 \;\; j = 1, \ldots, m \\
L_i x_{i,j} + V_i y_{i,j} - L_{i+1} x_{i+1,j} - V_{i-1,j} y_{i-1,j} = 0, \;\; i = 2, \ldots, N+1, i = S+1 \\
L_i x_{i,j} + V_i y_{i,j} - L_{i+1} x_{i+1,j} - V_{i-1} y_{i-1,j} - F x_{F,j} = 0 \;\; j = 1, \ldots, m, \; i = S+1 \\
(L_{N+2} + D) x_{N+2,j} - V_{N+1} y_{N+1,j} = 0 \;\; j = 1, \ldots, m
\end{array}
$$

Our objective is to minimize the reboiler heat input which accounts for a major portion of operating costs. We are also interested in identifying the phases and the phase equilibria on each tray by minimizing the Gibbs free energy (δ), i. e.

$$
\begin{array}{ll}
\min & Q_R \\
\text{s. t.} & \min \sum_{j=1}^{n+2} \delta_j \\
& \text{s. t.} \quad \text{MESH equations}
\end{array}
\tag{2}
$$

3 THE INEXACT RESTORATION METHOD

Let us consider the optimization problem

$$\begin{array}{ll} \min & F(x) \\ \text{s. t.} & H(x) = 0, \ \ x \in \Omega. \end{array} \tag{3}$$

with given functions $F : \mathbb{R}^n \to \mathbb{R}$, $H : \mathbb{R}^n \to \mathbb{R}^m$ and a given compact and convex set $\Omega \subset \mathbb{R}^n$. The functions F and H are assumed to be at least continuous on Ω. Inexact restoration methods (IR) are motivated by the bad behavior of feasible methods in the presence of strong nonlinearities [1]. A modern approach of these methods for Nonlinear Programming began with the algorithm of Martínez and Pilotta [7]. The common features of this and other IR methods are the following [4]:

(i) Given the current iterate $x^k \in \mathbb{R}^n$, an intermediate more feasible point y^k is computed using an arbitrary procedure which, in practice, is chosen according to the problem characteristics. This is the Restoration Phase of the method.

(ii) A trial point z is computed on the "tangent set" that passes through y^k , in such a way that an optimality measure improves at z with respect to y^k.

(iii) If the point z is acceptable for a criterion that combines feasibility and optimality, one defines the new iterate $x^{k+1} = z$. Otherwise, the trial point z is chosen in a smaller trust region around y^k.

The optimality improvement (ii) involved in the choice of z can be done by a line search with respect to the objective function [5, 7] or the Lagrangian [6]. The acceptability of z in (iii) depends, in [6, 7], on a function that combines feasibility and optimality. In [5] the acceptability of z was decided on the basis of a filter strategy. As tools for describing the algorithm, detailed on Page 4, we will make use of functions $h : \Omega \to [0, \infty)$ and $\Phi : \Omega \times [0, 1] \to \mathbb{R}$. The function h is defined as $h(x) = \max ||H(x)||_\infty$ for all $x \in \Omega$ and the penalty function is defined as $\Phi(x, p) := pF(x) + (1 - p)h(x)$ for all $p \in [0, 1]$ and $x \in \Omega$. We are now going to describe a simple frame for an inexact restoration algorithm. This frame allows the user to apply several concrete methods within both the restoration phase and the optimization phase. For an accurate global convergence analysis see [4].

4 THE INEXACT RESTORATION METHOD FOR BILEVEL PROBLEMS

In order to simplify as much as possible the notation in this work we will define the procedure used to solve the bilevel problem considering the following:

$$\begin{array}{ll} \min\limits_{(u,v)\in\Omega} & F(u, v) \\ \text{s. t.} & \min\limits_{y} f(u, v) \\ & \text{s. t.} \quad g(u, v) = 0, \ v \geqslant 0. \end{array} \tag{10}$$

From problem (10) we define the function C that takes into account feasibility and optimality of the lower-level problem,

$$C(u, v, \mu, \lambda) = \begin{bmatrix} \nabla_v f(u, v) + \nabla_v^t g(u, v)\mu - \lambda \\ g(u, v) \\ v_1\lambda_1 \\ \vdots \\ v_{n_v}\lambda_{n_v} \end{bmatrix}. \tag{11}$$

Therefore, the problem (10) results in:

$$\begin{array}{ll} \min\limits_{(u,v)\in\Omega} & F(u, v) \\ \text{s. t.} & C(u, v, \mu, \lambda) = 0, \ v \geqslant 0. \end{array} \tag{12}$$

Considering $x = (u, v, \mu, \lambda)$ the problem (12) reduces to the problem (3). Therefore, the Algorithm 1 can be applied to solve the problem (10).

Algorithm 1 IR Method

Let $r \in [0,1)$ and β, γ, $\bar{\gamma}$, $\tau > 0$ be fixed.

Step 0: *Initialization.*

Choose $x^0 \in \Omega$ and $p_0 \in (0,1)$. Set $k := 0$.

Step 1: *Inexact restoration.*

Compute $y^k \in \Omega$ so that

$$h(y^k) \leqslant rh(x^k) \tag{4}$$

$$F(y^k) \leqslant F(x^k) + \beta h(x^k). \tag{5}$$

Step 2: *Penalty parameter.*

Determine $p_{k+1} \in \{2^{-i}p_k | i \in 0,1,2,\ldots\}$ as large as possible so that

$$\Phi(y^k, p_{k+1}) - \Phi(x^k, p_{k+1}) \leqslant \frac{1}{2}(1-r)(h(y^k) - h(x^k)). \tag{6}$$

Step 3: *Search direction for optimization.*

Compute $d^k \in \mathbb{R}^n$ so that $y^k + d^k \in \Omega$ and

$$F(y^k + td^k) \leqslant F(y^k) - \gamma t||d^k||^2, \tag{7}$$

$$h(y^k + td^k) \leqslant h(y^k) + \bar{\gamma}t^2||d^k||^2 \tag{8}$$

holds for all $t \in [0, \tau]$.

Step 4: *Line search.*

Determine $t_k \in \{2-i | i \in \mathbb{N}\}$ as large as possible so that

$$\Phi(y^k + t_k d^k, p_{k+1}) - \Phi(x^k, p_{k+1}) \leqslant \frac{1}{2}(1-r)(h(y^k) - h(x^k)). \tag{9}$$

Step 5: *Update.*

Set $x^{k+1} := y^k + t_k d^k$ and $k := k+1$. Go to Step 1.

REFERENCES

[1] R. ANDREANI, S. L. C. CASTRO, J. CHELA, A. FRIEDLANDER, S. A. SANTOS, *An inexact-restoration method for nonlinear bilevel programming problems*, Comput. Optim. Appl., 43 (2009), pp. 307-328.

[2] F. E. BUFFO, M. C. MACIEL, *Problema de optimizacin en dos niveles para el diseño óptimo de procesos de separación*, XVI Congreso sobre Métodos Numéricos y sus Aplicaciones, Córdoba, Argentina, Octubre 2-5, (2007).

[3] S. DEMPE, *Annottated bibliography on bilevel programming and mathematical problems with equi- librium constraints*, Optimization, 52 (2003), pp. 333-359.

[4] A. FISCHER, AND A. FRIEDLANDER, *A new line search inexact restoration approach for nonlinear programming*, Comput Optim Appl., 46 (2010), pp. 333-346.

[5] C. C. GONZAGA, E. W. KARAS, M. VANTI, *A globally convergent filter method for nonlinear programming*, SIAM J. Optim., 14 (2003), pp. 646-669.

[6] J. M. MARTÍNEZ, *Inexact restoration method with Lagrangian tangent decrease and new merit function for nonlinear programming*, J. Optim. Theory Appl., 111 (2001), pp. 39-58.

[7] J. M. MARTÍNEZ, AND E. A. PILOTTA, *Inexact restoration algorithms for constrained optimization*, J. Optim. Theory Appl., 104 (2000), pp. 135-163.

[8] A. U. RAGHUNATHAN, AND L. T. BIEGLER, *Mathematical programs with equilibrium constraints (MPECs) in process engineering*, Comp. & Chem. Eng., 27 (2003), pp. 1381-1392.

[9] L. N. VICENTE, AND P. H. CALAMAI, *Bilevel and multilevel programming: A bibliography review*, J. Global Optim., 5 (1994), pp. 1-23.

UN PROBLEMA DE EQUILIBRIO HIDROTÉRMICO CON RESTRICCIONES DE RED

L.A. Parente[♭], P.A. Lotito[†] y A.J. Rubiales[†]

[♭] *CONICET, Universidad Nacional de Rosario, Rosario, Argentina.*
[†] *CONICET, Universidad Nacional del Centro de la Prov. de Buenos Aires, Tandil, Argentina.*

Resumen: Estudiamos un problema de planeamiento a corto plazo de la generación eléctrica en un sistema hidrotérmico con restricciones en la red de distribución. Proponemos un modelo basado en el equilibrio de Nash-Cournot, obteniendo una inclusión monótona con un operador de estructura separable que permite aplicar un esquema proximal de descomposición. Describimos nuestra implementación y presentamos resultados numéricos preliminares.

Palabras clave: *Equilibrio de Nash-Cournot, Inclusiones Variacionales, Descomposición*
2000 AMS Subject Classification: 90C30 - 90C33

1. INTRODUCCIÓN

Consideramos un sistema de producción de energía en un mercado competitivo oligopólico, compuesto por unidades térmicas e hidráulicas de *almacenamiento de bombeo*. Estas centrales poseen dos reservorios a diferentes niveles, con la capacidad de generar electricidad en períodos de alta demanda o bombear agua cuando la demanda es baja, aumentando la capacidad de generación en períodos de alto consumo. El problema de planeamiento a corto plazo de un sistema de este tipo fue estudiado en [7] sin considerar restricciones de red. En el presente trabajo consideramos las restricciones físicas de la red de distribución con un modelo de corriente continua. Asumimos conocidos los coeficientes relacionados con las funciones de demanda y las características técnicas de la red y las plantas, como también el volumen total de agua a ser utilizado en el horizonte de planeamiento, producto de un estudio previo a largo plazo (véase [7]).

2. PROBLEMA DE EQUILIBRIO HIDROTÉRMICO CON RESTRICCIONES DE RED

El modelo consta de $\mathcal{I}$ centrales térmicas distribuidas en M compañías (según conjuntos de índices $\mathcal{C}_m^{Th}$) y $\mathcal{J}$ centrales hidráulicas distribuidas en N compañías (según conjuntos de índices $\mathcal{C}_n^{H}$). Las centrales y los centros de consumo están interconectados mediante una red compuesta por barras (nodos k, ℓ, etc.) y líneas (arcos (k,ℓ), etc.) con restricciones en la capacidad de transporte de energía. En un horizonte discreto de T períodos estudiamos el equilibrio de Nash-Cournot derivado de querer maximizar el beneficio individual de cada compañía. Las variables son la producción térmica $x = (x_{it})$, con restricciones de caja, y la producción hidroeléctrica $y = (y_{jt})$, que además presenta la restricción de igualdad $\sum_t y_{jt} = y_j^{TOT}$, que expresa el monto total de agua a ser utilizado en el horizonte de planeamiento. Como variable auxiliar, consideramos el flujo de energía en las líneas $w = (w_{k\ell t})$, con restricciones de caja.

Los beneficios de las compañías térmicas e hidráulicas en cada período están dados por el producto de las respectivas producciones por el precio de mercado,

$$Ben_m^{Th} = \sum_{i\in\mathcal{C}_m^{Th}} \sum_{t=1}^{T} \Big(x_{it}p_t - c_i^{Th}(x_{it}) \Big), \qquad m = 1, ..., M, \tag{1}$$

donde $c_i^{Th}(x_{it})$ es un costo de generación cuadrático y

$$Ben_n^{H} = \sum_{j\in\mathcal{C}_n^{H}} \sum_{t=1}^{T} f_j(y_{jt})p_t, \qquad n = 1, ..., N, \qquad \text{con } f_j(s) = \begin{cases} s, & \text{if} \quad s \geq 0, \\ \alpha_j s, & \text{if} \quad s < 0, \end{cases} \tag{2}$$

donde la función seccionalmente lineal f_j es usada para representar la diferencia entre bombear ($y_{jt} < 0$) y generar ($y_{jt} > 0$). El coeficiente de eficiencia $\alpha_j > 1$ indica que la energía utilizada para bombear agua es mayor que la energía generada por el mismo volumen de agua.

Las variables están acopladas a su vez por la restricción lineal

$$\sum_{i\in\mathcal{B}_k^{Th}} x_{it} + \sum_{j\in\mathcal{B}_k^{H}} y_{jt} - d_{kt} = \sum_{\ell\in\mathcal{B}_k^{B}} w_{k\ell t}, \qquad \forall k,t, \tag{3}$$

que responde a las restricciones físicas de un modelo de red de corriente continua, siendo respectivamente $\mathcal{B}_k^{Th}$, $\mathcal{B}_k^{H}$,$\mathcal{B}_k^{B}$ los conjuntos de plantas térmicas, plantas hidráulicas y barras conectadas con la barra k, y d_{kt} la demanda en la barra k en el período t. Las demandas se relacionan con el precio de mercado p a través de una función de demanda afín $d_{kt}(p) = D_{kt} - a_{kt}p$, para ciertos coeficientes D_{kt}, a_{kt}, con lo cual el precio está dado por la función de demanda inversa $p_t(d_{kt}) = \frac{1}{a_{kt}}(D_{kt} - d_{kt})$, o bien $p_t(d) = \frac{1}{a_t}(D_t - d_t)$, siendo $D_t = \sum_k D_{kt}$, $a_t = \sum_k a_{kt}$ y $d_t \sum_k d_{kt}$. Como la demanda total es suplida por todos los jugadores, coincide con la producción total, es decir $\sum_{j=1} y_{jt} + \sum_{i=1} x_{it} = d_t$ y entonces el precio de mercado es $p_t = \frac{1}{a_t}\left(D_t - \sum_{j=1} y_{jt} - \sum_{i=1} x_{it}\right)$, los beneficios (1) y (2) son cuadráticos y la restricción (3) resulta

$$\sum_{i\in\mathcal{B}_k^{Th}} x_{it} + \sum_{j\in\mathcal{B}_k^{H}} y_{jt} - D_{kt} + \frac{a_{kt}}{a_t}\left(D_t - \sum_{j=1}^{\mathcal{J}} y_{jt} - \sum_{i=1}^{\mathcal{I}} x_{it}\right) = \sum_{\ell\in\mathcal{B}_k^{B}} w_{k\ell t}. \tag{4}$$

Siendo $\mathcal{K}_m^{Th}$ y $\mathcal{K}_n^{H}$ los conjuntos factibles de las compañías térmicas m e hidroeléctricas n, respectivamente, el problema de equilibrio de Nash-Cournot con restricciones de red consiste en determinar la terna (x^*, y^*, w^*) a fin de satisfacer

$$Ben_m^{Th}(x^*, y^*) = \max_{x_m\in\mathcal{K}_m^{Th}} Ben_m^{Th}(x_m, x^*_{/m}, y^*), \qquad m = 1, ..., M, \tag{5}$$

$$Ben_n^{H}(x^*, y^*) = \max_{y_n\in\mathcal{K}_n^{H}} Ben_n^{H}(x^*, y_n, y^*_{/n}), \qquad n = 1, ..., N, \tag{6}$$

y la restricción (4), donde $x_m = (x_{it})_{i\in\mathcal{C}_m^{Th}}$, $x_{/m} = (x_{it})_{i\notin\mathcal{C}_m^{Th}}$, $y_n = (y_{jt})_{j\in\mathcal{C}_n^{H}}$ and $y_{/n} = (y_{jt})_{j\notin\mathcal{C}_n^{H}}$.

3. El Método de Resolución

En nuestro problema de equilibrio, los problemas (5) no presentan mayor dificultad pues son problemas diferenciables cuadráticos con restricciones de caja, pero las funciones f_j hacen que los problemas (6) no sean diferenciables en el origen, y además presentan restricciones de igualdad. Abordaremos la falta de diferenciabilidad en cero mediante la descomposición $y = y_+ - y_-$, donde

$$(y^+)_{jt} = \begin{cases} y_{jt}, & \text{si} \quad y_{jt} \geq 0, \\ 0, & \text{si} \quad y_{jt} < 0, \end{cases} \qquad (y^-)_{jt} = \begin{cases} 0, & \text{si} \quad y_{jt} \geq 0, \\ -y_{jt}, & \text{si} \quad y_{jt} < 0, \end{cases}$$

para $j = 1, .., \mathcal{J}$ and $t = 1, .., T$. Tomando $z = (y_+^\top, y_-^\top)^\top$, la complementariedad entre y_+ e y_- introduce una restricción no lineal, pero puede demostrarse que los puntos de equilibrio del problema relajado, sin considerar esta restricción, verifican la complementariedad (véase [7, Proposition 1]), con lo cual no es necesario introducirla en la formulación. De esta manera, los problemas (6) expresados en z resultan diferenciables cuadráticos con restricciones de caja e igualdad. Además, los gradientes de las funciones de beneficio, tanto térmicas como hidráulicas, resultan funciones lineales. De forma similar a lo hecho en [7], podemos ahora asociar las variables con restricciones de caja x y w en un conjunto de restricciones $\mathcal{K}_1$ y las variables con restricciones de caja e igualdad z en un conjunto de restricciones $\mathcal{K}_2$. Asimismo, asociamos los gradientes de los beneficios térmicos en una función lineal en x, $Q^{Th}x + b^{Th}(z)$, y los gradientes de los beneficios hidráulicos en una función lineal en z, $Q^H z + b^H(x)$, para adecuadas matrices Q^{Th}, Q^H y funciones vectoriales lineales b^{Th}, b^H. Las restricciones de red lineales (4) pueden asociarse en una restricción lineal de la forma $Ax + Bw + Cz - d = 0$ para un adecuado vector d y apropiadas matrices A, B y C. Considerando las condiciones de optimalidad y siguiendo a [2], el problema de equilibrio de Nash-Cournot puede formularse como una inecuación variacional de la forma

$$\text{hallar } \xi^* \in \Omega \text{ tal que } \langle \Phi(\xi^*), \xi - \xi^* \rangle \geq 0, \quad \forall \xi \in \Omega, \tag{7}$$

donde

$$\Omega = \{((x,w),z) \in \mathcal{K}_1 \times \mathcal{K}_2 \mid Ax + Bw + Cz - d = 0,$$
$$\Phi((x,w),z) = (Q^{Th}x + b^{Th}(z), Q^H z + b^H(x))$$

Asociando un multiplicador de Lagrange a la restricción de igualdad en la definición de Ω, reemplazamos (7) por la inecuación variacional

$$\text{hallar } (\xi^*, \lambda^*) \in \mathcal{K}_1 \times \mathcal{K}_2 \times \mathbb{R}^s \text{ tal que } \langle \Psi(\xi^*, \lambda^*), (\xi, \lambda) - (\xi^*, \lambda^*) \rangle \geq 0, \quad \forall (\xi, \lambda) \in \mathcal{K}_1 \times \mathcal{K}_2 \times \mathbb{R}^s, \tag{8}$$

donde

$$\Psi(\xi, \lambda) = \begin{pmatrix} Q^{Th}x + b^{Th}(z) - (A,B)^\top \lambda \\ Q^H z + b^H(x) - C^\top \lambda \\ Ax + Bw + Cz - d \end{pmatrix}. \tag{9}$$

Siguiendo [4, section4.2], la inecuación variacional (8) resulta equivalente a la inclusion variacional

$$0 \in T((x,w),z,\lambda) = F((x,w),z,\lambda) \times [G((x,w),z,\lambda) + H(z,\lambda)], \tag{10}$$

con

$$F((x,w),z,\lambda) = \begin{pmatrix} Q^{Th}x + b^{Th}(z) - (A,B)^\top \lambda \\ 0 \end{pmatrix} + N_{\mathcal{K}_1}(x,w),$$

$$G((x,w),z,\lambda) = \begin{pmatrix} Q^H z + b^H(x) - C^\top \lambda \\ Ax + Bw + Cz - d \end{pmatrix},$$

$$H(z,\lambda) = \begin{pmatrix} N_{\mathcal{K}_2}(z) \\ 0 \end{pmatrix},$$

donde $N_{K_1}(\cdot)$ y $N_{K2}(\cdot)$ son los conos normales para los conjuntos $\mathcal{K}_1$ y $\mathcal{K}_2$, respectivamente. La estructura del operador T resulta adecuada para la aplicación del esquema proximal de descomposición de [4], y es fácil ver que las hipótesis [4, Assumptions A1-A5] son satisfechas. Dicho esquema se basa en los procedimientos proximales de métrica variable desarrollados en [6], que poseen condiciones de convergencia de naturaleza constructiva y de verificación computacional factible. El método descompone el problema en las variables (x,w) y (z,λ), separando en un primer paso vez la suma $G + H$ en (z,λ) mediante un procedimiento de tipo *avance-retroceso*, y procediendo luego a un paso proximal para (x,w). En nuestro caso la ejecución del paso de separación resulta equivalente a calcular

$$\begin{array}{rcl} \hat{z}^k & = & \text{Proy}_{\mathcal{K}_2}\left(z^k - c_k\left[Q^H z^k + b^H(x^k) - C^\top \lambda^k\right]\right) \\ \hat{\lambda}^k & = & \lambda^k - Ax^k - Bw^k - Cz^k - d \end{array}$$

La segunda igualdad es una simple operación, y la primera es una proyección sobre la intersección de un hiperplano y una caja, que puede aproximarse con el grado de precisión deseado con algoritmos de orden $O(T)$ (véase [3]).

A continuación, la ejecución del paso proximal en su forma exacta nos da, para cierto $c_k > 0$ y cierta matriz simétrica positiva definida U_k,

$$0 \in \frac{1}{c_k} U_k \begin{pmatrix} \hat{x}^k - x^k \\ \hat{w}^k - w^k \end{pmatrix} + \begin{pmatrix} Q^{Th}x + b^{Th}(z) - (A,B)^\top \lambda \\ 0 \end{pmatrix} + N_{\mathcal{K}_1}(x,w). \tag{11}$$

Con una adecuada elección de la matriz U_k, y simple trabajo algebraico el paso proximal toma la forma de una proyección ortogonal sobre $\mathcal{K}_1$, y siendo este conjunto una caja, se puede calcular mediante una fórmula explícita.

Además, eligiendo parámetros adecuados, se pueden satisfacer las condiciones necesarias (véase [6, Proposition 3.1]) para que las nuevas iteraciones resulten

$$\begin{aligned} x^{k+1} &= \hat{x}^k \\ w^{k+1} &= \hat{w}^k \\ z^{k+1} &= \hat{z}^k - Q^h(\hat{z}^k - z^k) - b^H(\hat{x}^k) + b^H(x^k) + C^\top(\lambda^k - \hat{\lambda}^k) \\ \lambda^{k+1} &= \hat{\lambda}^k - A\hat{x}^k - B\hat{w}^k - C\hat{z}^k + d \end{aligned}$$

En la siguiente sección presentamos algunos resultados numéricos preliminares.

4. Resultados Numéricos

Consideramos 3 problemas con un número creciente de unidades, líneas y barras, donde los datos son variaciones del ejemplo IEEE BUS 30 presentado en [5]. Se consideró un planeamiento de un día con 24 períodos (uno por hora) con variaciones en la demanda esperada en cada período (en el primer problema se consideraron 12 períodos). Los coeficientes de la Función de Demanda Inversa se calculan previamente con una técnica de punto de anclaje (demanda esperada-costo térmico marginal) a partir de un coeficiente de elasticidad, siguiendo el procedimiento de [1]. Cada problema se corrió 2 veces, con variaciones en las capacidades de las líneas. Como era de esperar, con menores capacidades en las líneas se observó menor bombeo por parte de las hidráulicas y menores beneficios de las compañías. La siguiente tabla da una noción del desempeño del algoritmo.

Problema	Tamaño	N^o Iter.	Tiempo Comp. (s)
1.1	168	544	15.32
1.2	168	521	14.16
2.1	552	802	76.11
2.2	552	810	77.14
3.1	1464	1312	892.66
3.2	1464	1301	891.83

Referencias

[1] M.S. Arellano,*Market power in mixed hydro-thermal electric systems*, Econometric Society 2004 Latin American Meetings 211, (2004).

[2] F. Facchinei y J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Series in Operations Research, Spinger, 2003.

[3] P.A. Lotito, *Issues in the implementation of the DSD algorithm for the traffic assignment problem*, European Journal of Operational Research, 175 (2006), pp.1577–1587.

[4] P.A. Lotito, L.A. Parente y M.V. Solodov, *A class of variable metric decomposition methods for monotone variational inclusions*, Journal of Convex Analysis, 16 (2009), pp.857–880.

[5] A. Maiorano, Y.H. Song y M. Trovato, *Dynamics of noncollusive oligopolistic electricity markets*, Proc. IEEE Power Eng. Soc. Winter Meeting, Singapore (2000), pp.838–844.

[6] L.A. Parente, P.A. Lotito y M.V. Solodov, *A class of inexact variable metric proximal point algorithms*, SIAM Journal on Optimization, 19 (2008), pp.240–260.

[7] L.A. Parente, P.A. Lotito, F.J. Mayorano, A.J. Rubiales y M.V. Solodov, *The hybrid proximal decomposition method applied to the computation of a Nash equillibrium for hydrothermal electricity markets*, Optimization and Engineering, (2011), en prensa.

THE FACE PROJECTION METHOD IN LINEAR PROGRAMMING

by

Ezio Marchi(*) & Martin Matons(**)

(*) IMASL - Universidad Nacional de San Luis
San Luis, Argentina
emarchi@speedy.com.ar

(**) Universidad Tecnológica Nacional, Facultad Regional Mendoza
Mendoza, Argentina
mmatons@unsl.edu.ar

Abstract: in this short paper we introduce a new method for linear programming which is an interior one which works on face projections.

Keyworks: face projection, linear programming, interior point.

Introduction

The landmark in the subject of allocation of resources and its optimization has been introduced by Kantorovich [3]. He came up with the technique of linear programming. After that, during the 40´s decade, Dantzig[1] has introduced the famous simplex method. This still plays an extreme important role in the actual days, even though Klee and Minty [6] provided an example where the simplex method is not efficient. Independently, Khachiyan [4] after several years introduced another method consisting of a novel algorithm for LP. Finally, Karmakar [4] was the first mathematician introducing a new polynomial time algorithm for LP. This is related with interior points methods.

In this short paper we present a novel method, called the face projection method. This, in the first step works as an interior method. But it considers the gradient or the objective function until it reaches the first hyperplane. At this point we consider the cone of the normal vectors of all hyperplanes containing the point. If the objective vector belongs to such a cone, the problem is solved and it reaches a global maximum. If not, we project the objective vector onto the linear manifold which is obtained by the intersection of the hyperplanes mentioned above. It turns out that if this happens and the objective vector is not zero, the objective function at this last projection is strictly greater than before. Therefore, we have derived a new method since the problem is finite. By the last remark the method cannot enter into a cycle.

The Face Projection Method

Consider a linear programming problem given by:

$$\begin{aligned} &\max cx \\ &Ax \le b \end{aligned} \qquad (1)$$

where A is a $m \times n$ matrix, $b \in R^m$, $c \in R^n$ and $x \in R^n$. Let $X = \{x \in R^n / Ax \le b\}$ which is considered to be non-empty, convex and compact. Let ∂X be the boundary of X. Consider an interior point $x_0 \in X - \partial X$. At this interior point x_0, the objective function

is cx_0. Therefore consider the ray $x_0 + \lambda' c$ with $\lambda' \in R$. By hypothesis there are negative and positive λ's such that the point $x_0 + \lambda' c$ reaches ∂X. Consider:

$$\bar{\lambda}' = \min_{\lambda_i > 0} \{a_i (x_0 + \lambda_i c) = b_i\}$$

The point $x_0 + \bar{\lambda}' c$ belongs to an hyperplane H_1 which has some region in the boundary. At that point we have

$$c(x_0 + \bar{\lambda}' c) = cx_0 + \bar{\lambda}' cc > cx_0$$

since $cc > 0$, $\bar{\lambda}' > 0$ and $c \neq 0$. Thus, the objective function has increased the value. At the point $x_0 + \bar{\lambda}' c \in \partial X$ consider all the hyperplanes H_i, $i = r_1, ..., r_s$, containing such a point. Take the cone of normal vectors a_i, $i = r_1, ..., r_s$, where a_i are the components of the incidence matrix A. If c belongs to the cone of H_i, that is $c = \sum_{j=1}^{s} \mu_j a_j$ with $\mu_j \geq 0$, then since $\bar{x} = x_0 + \bar{\lambda}' c$ is the point that belongs to the π_i hyperplanes

$$c(x_0 + \bar{\lambda}' c) = c\bar{x} = \sum_{j=1}^{s} \mu_{r_j} a_{r_j} \bar{x} = \sum_{j=1}^{s} \mu_{r_j} b_{r_j} \geq \sum_{j=1}^{s} \mu_{r_j} a_{r_j} x = cx$$

for all $x \in X$. Therefore the point $\bar{x}$ is a global solution of the linear programming.

On the other hand from a geometrical point of view it happens that the hyperplane H_c with normal c separates all the convex polyhedron X. In the case that c does not belongs to the cone, then at least one μ_{r_i} is negative and therefore H_c does not separate the set X.

Let π_i, $i : r_i, .., r_s$ the subspaces with normal vectors a_i at $x_0 + \bar{\lambda}' c$. Then consider

$$\bigcap_{i=1}^{s} \pi_i = \text{gen}\{z_1, ..., z_s\}$$

where the z_i determine a orthonormal base. Then the projection of c onto the face $\bigcap_{i=1}^{s} \pi_i$ may be written as

$$\text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c) = \sum_{i=1}^{s} \frac{z_i c}{z_i z_i} z_i$$

and the objective function on it is

$$c\left(x_1 + \text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c) \right).$$

Consider the point $x_1 + \bar{\lambda}^2 \text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c)$, where

$$\bar{\lambda}^2 = \min_{\lambda_j > 0} \left\{ \lambda_j : a_j \left(x_j + \lambda_j \text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c) \right) = b_j, j \notin I_{x_1} \right\}$$

where I_{x_1} is the set of i such that $i \neq r_1, ..., r_s$. By finiteness of the number n of hyperplanes and by construction, it is clear that, $\bar{\lambda}^2 > 0$. This is because X is bounded. Therefore

$$c\left(x_1 + \bar{\lambda}^2 \text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c) \right) = cx_1 + \bar{\lambda}^2 c \sum_{i=1}^{s} \frac{z_i c}{z_i z_i} z_i =$$

$$cx_1 + \overline{\lambda^2} c \sum_{i=1}^{s} \frac{|z_i|^2 |c|^2}{|z_i|} \cos^2\theta_i = cx_1 + \overline{\lambda^2} |c|^2 \sum_{i=1}^{s} \cos^2\theta_i$$

where the cosine is determined by z_i and c. The term with the cosines is strictly greater that zero since at least one $\cos^2\theta_i$ is strictly greater than zero, and $\overline{\lambda^2} > 0$ and $|c| > 0$ if $c \neq 0$. Thus we obtain another point $x_2 = x_1 + \overline{\lambda^2} \text{proj}_{\bigcap_{i=1}^{s} \pi_i}(c)$, where the objective function is strictly greater than in x_1. In this way we repeat the same procedure at x_2 and then by a finiteness argument we obtain the global maximum in a finite number of steps. Therefore, it converges. Thus, the face projection method proposed by us takes a form, since, at each projection the value of the objective function strictly increases.

If $x_0 \in \partial X$, we are in the case that the point belongs to an hyperplane and therefore we are in the inductive step described above.

Comment

We have implemented several examples in low dimensional problems and our method is much better than the simplex method.

We would like to say that the purpose of the context of this note is to present a new method in linear programming without considering the algorithm complexity. We would like to say that in this way there are many facts to be considered. As for example, the comparison with other methods, as well as the relation with the Dantzig-Wolfe Theorem. Surely the geometrical part of it appears to be interesting.

Acknowlegments

We are indebted with Prof. Bill Cooper, Prof. Juan Enrique Martinez Legaz and Prof. Martin Groetschel for some positives comments.

Bibliography

[1] Dantzig, G.B.: Linear Programming and Extension. Princeton University Press. Princeton, MJ (1963)

[2] Grötschel, M. L. Lovász and A. Schrijver: The Ellipsoid Method and its Consequences in Combinational Optimization. Combinatoria I (1981) 169-197

[3]Kantorovich, L.V. (1939): Mathematics Methods of Organizing and Planning Production, Management Sciences, Vol. 6, No 4 (1960) pp.366-422

[4] Karmarkar, N. A.: New Polynomial-Time Algorithm for Linear Programming. Combinatorica 4 (1984) 373-395

[5] Khachigan L.G.: A Polynomial Algorithm in Linear Programming. Doklady Akademii Mank SSSR 244:s (1979) 1093-1096, translate in Soviet Mathematics Doklady 20: 1 (1979), 191-194.

[6] KleeV. And Minty G. L.: How good is the Simplex Algorithm? In Inequalities III. Ed. Shiska. Acad. Press, New York, 1972, 159-179

[7] Luenberger, D: Linear and Nonlinear Programming, Kluwer Academic Publishers, 2003

SOLVING THE SEGMENTATION PROBLEM FOR THE 2010 ARGENTINE CENSUS WITH INTEGER PROGRAMMING

Flavia Bonomo[♭,‡], Diego Delle Donne[♭,†], Guillermo Durán[⋆,*,‡] and Javier Marenco[♭,†]

[♭]*Depto. de Computación, FCEN, Universidad de Buenos Aires, Argentina.*
[†]*Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina.*
[⋆]*Depto. de Matemática, FCEN, Universidad de Buenos Aires, Argentina.*
[*]*Depto. de Ingeniería Industrial, FCFM, Universidad de Chile, Chile.*
[‡]*CONICET, Argentina.*

Abstract: One of the most challenging tasks within the planning of a demographic census is to partition each census track into sets of homes such that each census taker visits exactly one set from this partition. In this work we introduce the *home segmentation problem*, which consists in designing such a partition subject to specific constraints. We present an integer programming-based algorithm for this problem, and we report the application of this algorithm for the 2010 census in the main province in Argentina.

Keywords: *demographic census, integer programming, home segmentation.*

1 INTRODUCTION

A *demographic census* is the process of acquiring information about the people and households within a country. The main objective is to provide statistical housing information as well as demographic, economic, and social data about the inhabitants of a country. A typical census takes place within one specific day, so planning a census can be a complex enterprise.

During the planning stages of a demographic census, a crucial task is to assign a set of homes to each census taker. In this context this process is called *home segmentation* and, depending on the specific constraints, it can give rise to hard combinatorial problems. When this partitioning is performed at a higher level, e.g., to reassign political borders within a county, this problem is related to *redistricting problems* [2, 3, 5, 6, 7, 8, 9]. There exist many computational software tools to solve redistricting problems (see [4] for an open source package implemented in R). However redistricting constraints usually differ from the requirements for a home segmentation, so specific models and algorithms must be developed for the latter. We are not aware of previous works on this specific problem.

In this work we describe the home segmentation problem for the 2010 census in Argentina, and we present an integer programming-based algorithm for this problem. We report computational results for the main province in Argentina. The present work was a two-month project within the census planning activities for this province.

2 PROBLEM DESCRIPTION AND MODEL FORMULATION

The Argentine territory is divided into *provinces* and each province is divided into *counties*. For censal purposes, each county is partitioned into sets of connected blocks called *census tracks*. Each census track has approximately 300 houses and, depending on the population density, between 1 and 50 blocks.

Given a census track, the *home segmentation problem* asks for a partition of the track into disjoint sets of connected *block sides* (or fractions of block sides) called *segments*, such that every home in the track belongs to exactly one segment. Each segment will be attended by one census taker. Empty block sides (i.e., sides with no houses) must nevertheless be visited by census takers, so each empty block side must belong to exactly one segment.

2.1 SEGMENTATION RULES

For a given census track, a group of connected homes represents a *valid segment* if it contains between 32 and 40 homes and its length does not exceed a prespecified value L, which depends on the track density. We say that a segment is *exceeded* if it contains more than 40 houses or its length exceeds L. From a block side it is allowed to cross the street, within the same segment, to the sides in adjacent blocks. However, some street crossings are more desirable than others in order to get segments as compact as possible. Figure 1 illustrates the admissible adjacent sides for a given side **b** as well as the *adjacency level* for each side, the most desirable adjacencies having level 0. For $\delta \in \{0, 1, 2\}$, a *δ-connected segment* is a segment whose sides are connected by adjacencies of level less or equal to δ. Census takers must not cross avenues, railroads or rivers.

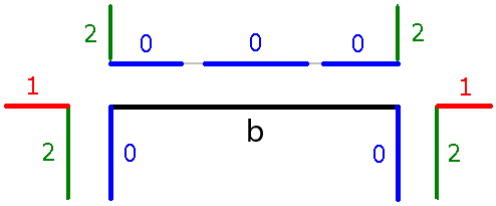


Figure 1: Adjacency levels for a block side

Block sides must not be splitted into more than one segment. However, this rule can be relaxed if there is no feasible solution without splitting sides (e.g., when a side has more than 40 homes). Block sides with zero houses must also be visited by a census taker. Finally, it is desirable for a segment to be as compact as possible, e.g., a segment consisting of a complete block is preferred over a segment with two sides from one block and two sides from a contiguous block.

2.2 INTEGER PROGRAMMING MODEL

For every census track $\mathcal{T}$, a straightforward integer programming formulation for the home segmentation problem can be given by considering the set of all feasible segments $\mathbb{S}_\mathcal{T}$ of $\mathcal{T}$. For $s \in \mathbb{S}_\mathcal{T}$, we introduce a binary variable x_s representing whether s is part of the solution or not. In order to maximize the compactness of the segmentation we define the *valuation* of each segment $s \in \mathbb{S}_\mathcal{T}$ to be $val(s) = 10^r$ with $r = \frac{sides(s)}{blocks(s)}$, where $sides(s)$ resp. $blocks(s)$ is the number of sides resp. blocks in s.

Let $\mathbb{H}$ be the set of houses in $\mathcal{T}$, and let $\mathbb{B}_0$ be the set of empty sides in $\mathcal{T}$. For $h \in \mathbb{H}$, define $S_h \subseteq \mathbb{S}_\mathcal{T}$ to be the set of segments including the home h and, for $l \in \mathbb{B}_0$, define $L_l \subseteq \mathbb{S}_\mathcal{T}$ to be the set of segments including the side l. With these definitions, the proposed model for the home segmentation problem is given by:

$$\max \sum_{s \in \mathbb{S}_\mathcal{T}} val(s)\, x_s$$

$$\sum_{s \in S_h} x_s = 1 \qquad \forall h \in \mathbb{H}$$

$$\sum_{s \in L_l} x_s = 1 \qquad \forall l \in \mathbb{B}_0$$

$$x_s \in \{0, 1\} \qquad \forall s \in \mathbb{S}_\mathcal{T}$$

3 SEGMENTATION ALGORITHM

The proposed algorithm first generates the set of single-block segments, called the *base segments*; for each block, every possible segment is generated without splitting block sides. If the integer programming model with these segments is infeasible, more segments are generated by combining the base segments

from each pair of adjacent blocks in order to produce two-block feasible segments. Adding these segments, the model is tested again for feasibility. On each new iteration, existing segments are combined with base segments from adjacent blocks in order to produce broader segments. This process iterates until a feasible solution is found or a prespecified limit on the maximum number of blocks is reached. Algorithm 1 illustrates this procedure, taking also the maximum allowable adjacency level $\delta \in \{0, 1, 2\}$ as a parameter.

Algorithm 1 Segmentation for a given adjacency level δ

$S_b \leftarrow \{\}$ *// base set*
for each block q **do**
  $S_b \leftarrow S_b \cup \{$not exceeded segments from $q\}$
end for
$i \leftarrow 1$
$S_1 \leftarrow S_b$
Test the IP model using valid segments from S_1
while no solution obtained **and** $i <$ generation limit **do**
  $S_{i+1} \leftarrow S_i$
  for each $(s_i, s_b) \in S_i \times S_b$ **do**
    if $s_i \cup s_b$ is a not exceeded δ-connected segment **then**
      $S_{i+1} \leftarrow S_{i+1} \cup \{(s_i \cup s_b)\}$
    end if
  end for
  Test the IP model using valid segments from S_{i+1}
  $i \leftarrow i + 1$
end while

In order to obtain compact solutions, Algorithm 1 is executed for $\delta = 0, 1, 2$ in sequence, until a solution is found. If this procedure fails to find a feasible solution, we split blocks sides and use fractions of block sides to build the base segments. The algorithm is executed again for every adjacency level and this process is performed until a solution is found or a limit in the size of the fractions is reached. If upon completion the procedure still finds no feasible solution, the overall algorithm ends suggesting the user to relax the parameters for the census track.

In low populated census tracks, feasible segments may span over several blocks yielding an intractable number of segments within the proposed algorithm. Due to this fact, we added a parameter to impose a minimum number of homes that a block must have in order to split the block while generating base segments, and a second parameter specifying a minimum number of homes for base segments. If a base segment does not reach this number of homes then it is arbitrarily attached to adjacent base segments until such "aggregated segment" meets this parameter. A preprocessing step classifies census tracks as *urban*, *semi-urban*, and *semi-rural*, and a different set of parameters is used for each of these three categories.

4 RESULTS AND CONCLUSIONS

The Buenos Aires Province is the main province in Argentina both in terms of population and total area. The Province has some 5,000,000 homes distributed among 16,691 non-rural census tracks. The segmentation process for the previous demographic census in Argentina (which took place in 2001 having 15% less census tracks) was performed manually, demanding 25 full-time operators for a period of 30 days in Buenos Aires Province (approximately 6000 man-hours).

Our computational tool was applied in order to solve the home segmentation problem for Buenos Aires in the 2010 census, being this the first time the province uses an automatic segmentation tool for a census. As a result, 96% of the census tracks of Buenos Aires could be automatically segmented within approximately 320 hours of processing time (e.g. less than 24 hours in a cluster of 15 PCs). The remaining 4% of

the census tracks (i.e., about 600 tracks) were segmented with this computational tool by manually relaxing some segmentation constraints, and eventually by a manual procedure when such approach failed. The present work was accomplished within the two months assigned for this project in a timely manner, and the obtained results have been described, in the words of the Provincial Director of Statistics, Karina Angeletti, as a "complete success, as 95% of the households were censed and every county could be covered by this census" [10].

Preliminary experimentation showed that by using the proper parameter set for a census track, the segmentation is done in less than a second, while using an incorrect configuration the process may take up to an hour of computation for a single census track as the number of valid segments may rise up to tens of thousands. The track classification by their population density allowed us to properly handle almost every instance. Moreover, besides the optimization given by the objective function of the integer programming model, the secuential generation of the segments helped to obtain solutions according to the prespecified preference order."The use of this computational tool allowed an homogeneous segmentation with uniform compactness criteria, unlike the manual segmentation which highly depends on the operator decisions"[1], stated Fernando Aliaga, GIS design responsible for the census in Buenos Aires.

REFERENCES

[1] ALIAGA, F., *Personal communication*, November 2010.

[2] ALTMAN, M., *Is Automation the Answer: The Computational Complexity of Automated Redistricting*, Rutgers Computer and Law Technology Journal, **23**(1), (1997), 81–141.

[3] ALTMAN, M., MACDONALD, K., AND MCDONALD, M.P., *From Crayons to Computers: The Evolution of Computer Use in Redistricting*, Social Science Computer Review, **23**(3), (2005), 334–346.

[4] ALTMAN, M. AND MCDONALD, M.P., *Bard: Better Automated Redistricting*, Journal of Statistical Software, **31**(3), (2009).

[5] BOZKAYA, B., ERKUT, E., AND LAPORTE, G., *A tabu search heuristic and adaptive memory procedure for political districting*, European Journal of Operational Research, **144**(1), (2003), 12–26.

[6] FLESICHMANN, B. AND PARASCHIS, J.N., *Solving a large scale districting problem: a case report*, Comput. Oper. Res., **15**(6), (1988), 521–533.

[7] GARFINKEL, R.S. AND NEMHAUSER, G.L., *Optimal Political Districting by Implicit Enumeration Techniques*, Management Science, **16**(8), (1970), B495–B508.

[8] HELBIG, R.E., ORR, P.K., AND ROEDIGER, R.R., *Political redistricting by computer*, Commun. ACM, **15**(8), (1972), 735–741.

[9] HESS, S.W., WEAVER, J.B., SIEGFELDT, H.J., WHELAN, J. N., AND ZITLAU, P.A., *Nonpartisan Political Redistricting by Computer*, Operations Research, **13**(6), (1965), 998–1006.

[10] LA VOZ DE TANDIL (2010), *Se censó más del 95% de las viviendas en la provincia* (in spanish). Retrieved November 15, 2010. `http://www.lavozdetandil.com.ar/ampliar_nota.php?id_n=20090`.

Modelos de Programación Mixta Lineal-Entera para el Predespacho de Máquinas Térmicas

Juan Manuel Alemany*, Fernando Magnago* y Diego Moitre*

*GASEP - Dpto. de Electricidad y Electrónica - Fac. de Ing. - Universidad Nacional de Río IV, Córdoba, Argentina - jalemany@ing.unrc.edu.ar - www.ing.unrc.edu.ar/grupos/gasep

Resumen: La técnica de optimización denominada Ramificación-Cota utilizada para la Programación Lineal Entera Mixta, fue una de las primeras propuestas de solución para el predespacho de máquinas térmicas. No obstante ello, la aplicación práctica a sistemas reales es reciente, gracias a los novedosos avances algorítmicos y computacionales. El objetivo de este trabajo es presentar una revisión de distintos modelos mixtos lineal-entero para el predespacho térmico clásico, exhibiendo diferentes opciones relacionadas con la función objetivo y las principales restricciones del problema que permitan la simulación de sistemas reales tales como el Sistema Argentino de Interconexión.

Palabras clave: *GAMS, Mixed Integer Linear Programming, Short-Term Thermal Unit Commitment*
2000 AMS Subject Classification: 00000 - 00000

1. Introducción

El problema del predespacho puede definirse como el proceso de programar la producción de electricidad de unidades generadoras sobre un horizonte de tiempo predeterminado de manera tal de minimizar el costo de producción. Matemáticamente puede definirse como un problema de programación mixta lineal-entera (MILP) [1] basado en Ramificación y Corte (Branch and Cut, B&C).

La investigación de soluciones algorítmicas para problemas MILP comenzó a inicios de 1960 con el desarrollo de dos métodos [3]; el algoritmo de los planos cortantes (Cutting Planes) [5] y el método de ramificación y cota (Branch and Bound) [6], aunque sus aplicaciones en sistemas de gran escala fueron limitadas. En los últimos veinte años surgieron innovaciones importantes para los algoritmos MILP [2], avances que fueron implementados en programas comerciales [3]. Las mejoras más destacadas pueden agruparse en pre-procesamiento, heurísticas y gestión avanzada de datos [7]. Algunas de las principales características son:

- Cortes
- Prerresolución
- Selección de variables
- Heurísticas
- Prerresolución de nodos

En la actualidad es posible resolver con relativa facilidad, problemas que unos pocos años atrás eran prácticamente irresolubles [7]. Adicionalmente, la gran mayoría de programas comerciales incorporan características que son particularmente útiles para el modelado en el área de generación eléctrica como la utilización de variables semi-continuas [3].

La característica principal del modelo MILP es su naturaleza lineal. El algoritmo de solución basado en B&C resuelve sistemáticamente un conjunto de programas lineales (LP) [2]. Esto ofrece algunas bondades pero también desventajas [3]. La principal ventaja está relacionada con la técnica LP como algoritmo de optimización, cuya madurez lo convierten en una opción muy confiable y robusta. La principal desventaja está relacionada con la linealización de algunas partes del modelo general de predespacho, siendo esto último discutible dependiendo del caso. En el balance, el método MILP ofrece características muy beneficiosas, en tanto y en cuanto, permite construir modelos detallados e incorporar restricciones complejas con facilidad, flexibilidad y modularidad [4]. Sumado a lo último, el modelo MILP permite acoplar el predespacho térmico con el hidráulico y la red de transmisión [8].

Consecuentemente, el objetivo de este trabajo consiste en revisar y evaluar la aplicación de las nuevas tendencias de modelado MILP, basado en B&C, para el predespacho en instancias de escala real.

Este trabajo esta organizado de la siguiente manera: en la Sección 2 se desarrolla la revisión de los componentes básicos del modelo MILP; en la Sección 3 se presenta la simulación ilustrativa con el sistema de potencia argentino; finalmente en la Sección 4 se plasman las conclusiones.

2. Modelos MILP para predespacho

El problema genérico del Predespacho puede expresarse como [9]:

Minimizar el costo operacional del sistema

Sujeto a:

Restricciones		
De la unidad	De la central	Del sistema
Rampas de transición	Combustibles	Balance energético
Tiempos de servicio	Cuadrilla	Reservas
Capacidad	Estado	Red

2.1. Funcion objetivo

La función objetivo a minimizar representa la sumatoria de todos los costos de producción más los respectivos costos de arranques.

Las funciones de costo de producción dependen de las características tecnológicas de generación y de los requerimientos del mercado. La función de costo más genérica es no lineal no convexa. En el modelo MILP esto puede formularse como aproximaciones lineales por tramos para curvas convexas [10] y/o no convexas [11]. Una formulación alternativa basada en cortes de perspectiva (Perspective Cuts) se propone en [12].

Los costos de arranque de las unidades también dependen de la tecnología de generación y de los requerimientos del mercado. El costo de arranque genérico es de carácter exponencial. En el modelo MILP esto puede formularse con una aproximación escalonada como se presenta en [11]. En [13] se utiliza una variable binaria por cada escalón considerado. En [14] la formulación sólo depende de las variables en/fuera de servicio de las unidades. En [15] se definen los tres estados térmicos de arranque más comunes de un generador, arranque en caliente, tibio y frío como estados incrementales.

2.2. Restricciones

El conjunto de restricciones globales esta constituido por las restricciones de demanda, reserva y red, algunos modelos también suelen incluir combustibles y emisiones [13]. En un sistema multinodal la oferta de generación, la demanda y la reserva son distribuidas, para lo cual existen modelos DC y AC [27] aplicables a este tipo de predespacho con red de transmisión.

El conjunto clásico de restricciones locales esta constituido por las siguientes restricciones:

- Mínimos tiempos en/fuera de servicio
- Límites de potencia activa
- Rampas de transición
- Relaciones lógicas

Las unidades generadoras poseen limitaciones técnicas relacionadas a la tecnología inherente, los denominados tiempos de servicio son unas de ellas. En la referencia [11] se propone un complejo modelo compuesto por seis restricciones que cubren las horas iniciales, intermedias y finales del horizonte de programación. Otro modelo constituido por sólo dos restricciones puede encontrase en [17]. En [13] se presenta un modelo similar al de [17] con un par de ecuaciones extras. En [10] se presenta una formulación basada

sólo en las variables de estado de los generadores. Por último en [18] se desarrolla una demostración matemática de la formulación eficiente que allí se propone.

En el trabajo [19] pueden encontrarse las restricciones de límites de potencia activa para las tecnologías de generación térmica, hidráulica y bombeo, expresadas como una sola restricción. En [11] se detalla una formulación dependiente de las transiciones de potencia activa, la cual se basa en la formulación de [19].

En la referencia [19] se establece una de las primeras formulaciones para restricciones de rampa en el predespacho MILP, allí las transiciones de potencia son restringidas con límites fijos. En [20] se propone un nuevo método para incorporar rampas de transición, el cual flexibiliza la selección de la tasa de transición entre varias opciones. El trabajo [21] presenta un detallado modelo de transiciones de potencia de una unidad térmica durante los procesos de arranque/parada e incrementos/decrementos de potencia. En [22] se describe un modelo dinámico de rampas donde los límites pueden ser constantes, escalonados o funciones lineales por tramos.

La formulación del predespacho térmico clásico esta compuesta tanto por variables continuas como binarias. Las variables binarias se utilizan para representar estados y se encuentran relacionadas entre sí por restricciones del tipo lógicas que definen el conjunto factible. Inicialmente la relación lógica entre variables binarias se presentó en [23]. Posteriormente el conjunto de restricciones lógicas fue ampliado en la referencia [24]. En la publicación [25] se analiza la dominancia de estas restricciones.

3. Simulaciones ilustrativas

Para las simulaciones con un sistema de escala real se utiliza el parque generador térmico argentino con un modelo DC de la red de transmisión. Los datos del sistema pueden encontrarse en [26]. El modelo aplicado posee las características de Tabla 1. Las estadísticas relativas a las simulaciones pueden leerse en Tabla 2. El modelo fue implementado y simulado en GAMS/CPLEX.

Tabla 1: Modelo aplicado al SEP argentino.

Modelo		Referencia
Función objetivo		
	Costos de producción	[10]
	Costos de arranque	[14]
Restricciones		
Globales	Balance energético	[1]
	Combustibles	[13]
	Red DC 500Kv *	[1]
	* Sin descomposición	
Locales	Límites de potencia	[19]
	Tiempos de servicio	[9, 18]
	Lógicas	[25]
	Ciclos combinados	[28]

Tabla 2: Estadísticas de simulación.

Ecuaciones	135,275	Variables	45,583
No ceros	497,267	Variables binarias	5,827
Opciones por defecto			
MIP solución:	30150825	52653 iter.	503 nodos
Gap relativo:	0.87 %	Tiempo usado:	133.4 seg.
Sin simetría CPLEX			
MIP solución:	30094419	52975 iter.	503 nodos
Gap relativo:	0.68 %	Tiempo usado:	131.8 seg.
Sin presolve			
MIP solución:	30171517	72420 iter.	580 nodos
Gap relativo:	0.93 %	Tiempo usado:	171.4 seg.
Branch & Bound			
MIP solución:	30183830	87205 iter.	1550 nodos
Gap relativo:	0.99 %	Tiempo usado:	171.5 seg.
Branch & Cut			
MIP solución:	30144154	87892 iter.	1760 nodos
Gap relativo:	0.84 %	Tiempo usado:	200.4 seg.
Sin heurística de nodo			
MIP solución:	30147826	143032 iter.	2226 nodos
Gap relativo:	0.85 %	Tiempo usado:	264.6 seg.
Con heurística ramificación local			
MIP solución:	30182583	129506 iter.	1513 nodos
Gap relativo:	0.96 %	Tiempo usado	343.5 seg.
Sin heurística RINS			
MIP solución:	30103487	48283 iter.	2333 nodos
Gap relativo:	0.70 %	Tiempo usado:	422.6 seg.

El resultado más significativo observable en Tabla 2, se relaciona a los reducidos tiempos de cómputo con un sistema de escala real y sin la utilización de técnicas de descomposición que desacoplen la generación de la red. La diversidad en la tolerancia de convergencia - Gap, umbral que fue fijado en 1 % - reside en la característica de búsqueda del algoritmo, por lo que no es relevante concluir al respecto. De las soluciones presentadas en Tabla 2, se observa que la tolerancia más baja corresponde a la simulación *Sin simetría CPLEX* que circunstancialmente corresponde con el menor tiempo de cómputo. Sin embargo, el segundo menor valor de tolerancia corresponde al mayor tiempo de cómputo. Los resultados demuestran como los desarrollos teóricos y las mejoras implementadas en los últimos aos en esta metodología, han hecho posible

la utilización de la misma en problemas de escala real. Es destacable que la segunda solución en rapidez corresponda a la simulación con opciones por defecto.

4. Conclusiones

El propósito de este trabajo fue presentar una revisión de modelos de predespacho térmico basados en MILP y presentar resultados de simulación con un sistema real. El desarrollo de este trabajo comprobó la existencia de diferentes formulaciones para el problema y demostró mediante simulaciones que la metodología MILP puede ser una alternativa factible para la resolución del predespacho. La reaparición de la metodología como alternativa viable, es en gran parte debido a las significativas mejoras en algoritmos de optimización y es probable que los recientes desarrollos permitan usar esta técnica más efectivamente en la industria eléctrica. Sin embargo, el mayor desafío será la aplicación eficiente en instancias de gran escala.

Referencias

[1] A. Wood, B. Wollenberg, *Power Generation, Operation and Control*, John Wiley and Sons, Inc. NY 1996.

[2] R. Bixby, M. Fenelon, Z. Gu, E. Rothberg, R. Wunderling, *MIP: Theory and practice-closing the gap*, ILOG CPLEX, System Modelling and Optimization: Methods, Theory and Applications, 2000.

[3] B. Hobbs, M. Rothkopf, R. O'Neill, H. Chao, *The Next Generation of UC Models*, Springer 2001.

[4] A. Ott, *Evolution of computing requirements in the PJM market: Past and future*, PES General Meeting, IEEE, July 2010.

[5] R. Gomory, *An Algorithm for the Mixed-Integer Problem*, Technical Report, The Rand Corporation, 1960.

[6] A. Land, A. Doig, *An Automatic Method of Solving Discrete Programming Problems*, Econometrica, July 1960.

[7] R. Bixby, *The Latest Advances in MIP Solvers*, Spring School on Combinatorial Optimization in Logistics, Université de Montréal, May 2010.

[8] W. Sifuentes, A. Vargas, *Short-term hydrothermal coordination considering an AC network modeling*, International Journal of Electrical Power & Energy Systems, July 2007.

[9] J. Alemany, *Programación óptima de la operación de unidades de generación térmica de electricidad en el corto plazo*, Tesis de Maestría, Dpto. E&E-FI-UNRC, Agosto 2009.

[10] M. Carrion, J. Arroyo, *A computationally eficient MILP formulation for the thermal UC problem*, Power Systems, IEEE Transactions on, Aug. 2006.

[11] J. Arroyo, A. Conejo, *Optimal response of a thermal unit to an electricity spot market*, Power Systems, IEEE Transactions on, Aug 2000.

[12] A. Frangioni, C. Gentile, F. Lacalandra, *Tighter Approximated MILP Formulations for UC Problems*, Power Systems, IEEE Transactions on , Feb. 2009.

[13] L. Tao, M. Shahidehpour, *Price-based UC: a case of LR versus MIP*, Power Systems, IEEE Transactions on, Nov. 2005.

[14] M. Nowak, W. Rmisch, *Stochastic LR Applied to Power Scheduling in a Hydro-Thermal System under Uncertainty*, Annals of Operations Research, Springer Netherlands, 2000.

[15] R. Naidoo, *A MIP formulation of generator startup costs*, International Power Engineering Conference, IPEC 2007.

[16] S. Vemuri, L. Lemonidis, *Fuel constrained UC*, Power Systems, IEEE Transactions on, Feb 1992.

[17] G. Chang, Y. Tsai, C. Lai, J. Chung, *A practical MILP based approach for UC*, PES General Meeting, 2004.

[18] D. Rajan, S. Takriti, *Minimum up/down polytopes of the UC problem with start-up costs*, IBM Research Report, 2005.

[19] T. Dillon, K. Edwin, H. Kochs, R. Taud, *Integer Programming Approach to the Problem of Optimal UC with Probabilistic Reserve*, Power Apparatus and Systems, IEEE Transactions on, Nov. 1978.

[20] C. Wang, S. Shahidehpour, *Optimal generation scheduling with ramping costs*, Power Systems, IEEE Transactions on, Feb 1995.

[21] J. Arroyo, A. Conejo, *Modeling of start-up and shut-down power trajectories of thermal units*, Power Systems, IEEE Transactions on, Aug. 2004.

[22] L. Tao, S. Shahidehpour, *Dynamic Ramping in UC*, Power Systems, IEEE Transactions on, Aug. 2007.

[23] L. Garver, *Power Generation Scheduling by Integer Programming-Development of Theory*, Power Apparatus and System, Transactions of the AIEE, April 1962.

[24] J. Muckstadt, R. Wilson, *An Application of MIP Duality to Scheduling Thermal Generating Systems*, Power Apparatus and Systems, IEEE Transactions on, Dec. 1968.

[25] K. Edman, R. O'Neill, S. Oren, *Analyzing valid inequalities of the generation UC problem*, Power Systems Conference and Exposition, IEEE/PES, March 2009.

[26] www.cammesa.com.ar

[27] H. Pinto, F. Magnago, S. Brignone, O. Alsac, B. Stott, *SCUC: Network Modeling and Solution Issues*, Power Systems Conference and Exposition, IEEE/PES, Oct-Nov 2006.

[28] J. Alemany, D. Moitre, H. Pinto, F. Magnago, *Short-term Scheduling of Combined Cycle Units Using MILP Solution*, Electric Power Systems Research, March 2010, Under Review.

UNA HEURÍSTICA PARA LA ASIGNACIÓN ÓPTIMA DE FRECUENCIAS EN REDES CELULARES.

Esteban Carranza†, Mercedes Carnero‡ y José Hernández †‡

†Departamento de Telecomunicaciones,Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Argentina, hcarranza@ing.unrc.edu.ar
‡Grupo de Optimización – Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Argentina, {mcarnero | jlh}@ing.unrc.edu.ar

Resumen: El enorme crecimiento de las redes de sistemas de comunicación celulares ha provocado un interés cada vez mayor en la operación óptima de dichos sistemas. Uno de los problemas más importantes involucrados es la asignación de frecuencias o canales a solicitudes de llamadas en una celda de la red. Este es un problema de optimización combinatoria del tipo NP completo y ha sido tratado mediante heurísticas teles como redes neuronales, algoritmos genéticos, recocido simulado, etc. En este trabajo se propone un enfoque basado en un algoritmo híbrido combinando estrategias de asignación exhaustivas con un algoritmo evolutivo.

Palabras claves: *Problema de asignación de canales, Redes móviles celulares, Algoritmos Evolutivos*

1. INTRODUCCIÓN

La importancia de los sistemas de comunicación móvil ha crecido en los últimos tiempos no sólo por la disponibilidad de recursos que estos sistemas proveen sino por la alta conectividad y portabilidad que involucran. Esto trae consigo nuevos desafíos por cuanto la demanda no sólo ha aumentado en cantidad de usuarios sino en tipos de servicio con nuevos requerimientos. La facultad de un sistema para satisfacer estos requerimientos incluye la capacidad de asignar los recursos disponibles en forma óptima. La asignación de canales a solicitudes de llamadas entrantes, es un claro e importante ejemplo de manejo óptimo de recursos. Este problema, conocido en la literatura como C*hannel Assignment Problem* (CAP), consiste básicamente en determinar una distribución eficiente de un conjunto de canales disponibles a cada celda demandada a la vez que se satisfacen tanto restricciones de tráfico como de compatibilidad electromagnética. Estas últimas pueden, en general, agruparse en [1]:
Restricciones Cocanal: el mismo canal no puede ser asignado simultáneamente a ciertos pares de celdas.
Restricciones de canal adyacente: canales muy próximos en frecuencia no pueden ser asignados simultáneamente a celdas con cierta proximidad geográfica.
Restricciones Co-sitio: canales asignados simultáneamente en la misma celda deben guardar cierta separación en frecuencia.
Los esquemas más comúnmente usados para resolver el CAP consideran la asignación fija de canales (FCA) o la asignación dinámica de canales (DCA). En el primer caso un conjunto de canales está permanentemente asociado a cada celda en la red. Si bien es un esquema simple, no presenta flexibilidad. En el segundo caso todos los canales se mantienen en un conjunto central, de modo tal que todas las celdas tienen acceso a todos los canales. A medida que las llamadas ingresan en una celda se les asigna un canal dependiendo de determinadas condiciones [2], [3].
En un esquema DCA, cuando ingresa una nueva llamada, la red entera debería ser susceptible de rearmar la asignación previa de canales de forma tal de disminuir la probabilidad de bloqueo manteniendo la calidad de servicio.

2. FORMULACIÓN DEL PROBLEMA

Considerando un sistema con n celdas y z canales (frecuencias) disponibles, cada posible asignación de canales se describe por medio de una matriz binaria **F** de n filas y z columnas. En esta notación cada elemento f_{ij} será igual a uno si el j-ésimo canal está asignado a la i-ésima celda y será cero en caso contrario.

La demanda del sistema se representa mediante un vector **D** de dimensión n donde cada componente d_i representa el número de canales requeridos en la celda i.

Las restricciones electromagnéticas se simbolizan a partir de una matriz simétrica **C** de dimensión *nxn*, cuyos elementos c_{ij} describen la mínima separación (medida en cantidad de canales) entre canales utilizados simultáneamente en las celdas *i* y *j*.

El problema del mínimo número de canales está resuelto para un vector **D** de demandas. Sin embargo, esto supone un ordenamiento determinado de llamadas para el cual es posible utilizar dicho número mínimo de canales.

En un escenario real, esto puede no ocurrir. En general, para un ordenamiento arbitrario el número necesario de canales puede ser mayor que el mínimo determinado en función del vector de demandas.

Sin embargo, si se realiza una reasignación de las llamadas en curso, se podría estar en condiciones de alojar la nueva llamada, aunque deba tolerarse una disminución en la calidad de servicio la cual será menor mientras menos reasignaciones se realicen.

En este trabajo se propone atacar el problema de asignación de llamadas considerando un conjunto de escenarios diferentes para una misma condición de tráfico esperado y representado por el vector de demanda. Para ello se propone la exploración de técnicas evolutivas para su resolución.

Al generarse una solicitud de asignación para una nueva llamada, que no pueda ser atendida en la situación de asignaciones preexistente sin violar las restricciones impuestas, se resolverá el problema cambiando el esquema de las asignaciones anteriores.

Dados una matriz de asignación $\mathbf{F}_{\text{inicial}}$ y una celda *c*, demandante, se busca $\mathbf{F}_{\mathbf{salida}}$ tal que a partir de ella sea posible asignar el tráfico solicitado por *c*, respetando las restricciones indicadas por **C** y, además que la cantidad de cambios de asignación entre $\mathbf{F}_{\mathbf{inicial}}$ y $\mathbf{F}_{\mathbf{salida}}$ sea mínima, al tiempo que la cantidad de canales a utilizar permanece acotada. Matemáticamente la función objetivo puede ser expresada como sigue:

$$
\begin{aligned}
&min \quad N_{reasign} = \frac{G\left(Xor\left(\mathbf{F}_{inicial}, \mathbf{F}_{salida}\right)\right)}{2} \\
&s.a. \\
&z < z_m \\
&\sum_{j=1}^{z} f_{i,j} = d_i \quad para \quad i = 1,\ldots,n \\
&\left|p-q\right| \geq c_{ij} \quad para \quad p,q = 1,\ldots,z \; y \; i,j = 1,\ldots,n \\
&\qquad\qquad tal\ que \quad f_{pi} = f_{qj} = 1 \quad y \\
&f_{i,j} \in \{0,1\} \quad para\ i = 1,\ldots,n \; y \; j = 1,\ldots,z
\end{aligned}
\tag{1}
$$

Donde **G**(**M**) es una función que cuenta la cantidad de unos en una matriz binaria **M**.

La primera restricción que debe satisfacerse es no sobrepasar una cantidad de canales prefijados z_m. La segunda restricción asegura la satisfacción de la demanda del sistema mientras que la tercera asegura la inexistencia de interferencias entre dos comunicaciones cualesquiera. La cuarta restricción simplemente fuerza a que la matriz **F** sea binaria.

3. ALGORITMO DE ASIGNACIÓN

Cuando se solicita una llamada sobre una celda dada, el sistema realiza una asignación de un canal utilizando una estrategia de asignación exhaustiva (Frequency Exhaustive Assignment, FEA). Esta estrategia trabaja bien y rápido cuando el sistema no está fuertemente demandado. Pero, a medida que la cantidad de solicitudes aumenta, la cantidad de canales disponibles disminuye y la capacidad de asignarlos sin violar las restricciones impuestas decae significativamente. En muchos casos, el resultado es el bloqueo de la llamada entrante.

El algoritmo propuesto en este trabajo parte de este escenario en el cual se considera una llamada que habría sido bloqueada utilizando otra estrategia, a la cual se intentará asignarle un canal para que la misma pueda ser cursada. La forma de hacerlo es revisar y modificar el esquema de asignación realizado en el momento de la solicitud. La redistribución de canales a celdas trae aparejado problemas de calidad de

servicio, por lo que es necesario minimizar la cantidad de reasignaciones a realizar. Esta es la misión del algoritmo evolutivo híbrido propuesto. El esquema completo de asignación se muestra en la figura 1.

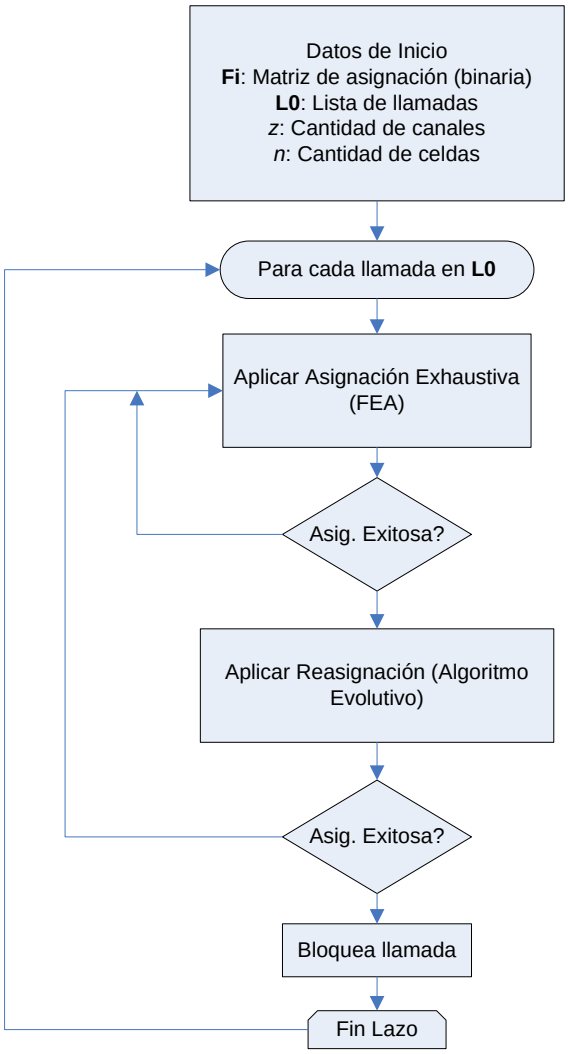


Figura 1: Algoritmo de asignación de canales

Los datos con los cuales cuenta el algoritmo son los siguientes: una lista de llamadas **L0** en la cual cada elemento $l0_i$ representa la celda en la cual ingresa la llamada *i*. Una matriz de asignación inicializada como matriz nula, la cual se irá modificando en el proceso. También se cuenta con los valores de cantidad de canales disponibles y celdas en el sistema.

El algoritmo intenta asignar, a medida que las llamadas ingresan, los canales desde el conjunto de disponibles. Esto lo realiza utilizando una estrategia FEA. Cuando una llamada no puede ser asignada mediante dicha estrategia, el algoritmo evolutivo trata de lograrlo modificando el esquema de asignación.

4. EXPERIMENTOS Y RESULTADOS

Se implementó el algoritmo mencionado para un caso de 25 celdas y 73 canales con un vector de demanda de 167 llamadas posibles. Se tomó una muestra aleatoria de 20000 ordenamientos diferentes de llamadas obtenidos de manera tal que verifican la demanda de tráfico **D.**

En el diseño del algoritmo genético se utilizó una población sembrada de matrices de asignación como individuos. Dichas matrices verificaron la condición de tráfico impuesta aunque no necesariamente la restricción de compatibilidad electromagnética. Se utilizó una probabilidad de cruzamiento de 0.8 y una de mutación de 0.01, el método de selección fue torneo binario.

Se midió la eficiencia del algoritmo en comparación con la estrategia FEA, contabilizando porcentaje de casos en los que hubo bloqueo de llamadas y porcentaje de éxito en las aplicación de la estrategia de reasignaciones. En este último caso se midió además la cantidad de reasignaciones realizadas para lograr incorporar la nueva llamada. Los resultados se muestran en la tabla 1.

Casos exitosos FEA [%]	Bloqueos [%]	Reasignaciones exitosas [%]	Nro de reasignaciones promedio	Rango de Nro de Reasignaciones
27.21	3.49	69.30	37.5	18

Tabla 1: Resultado de la aplicación del algoritmo de asignación de canales

Aproximadamente un 73% de las veces es necesario recurrir al algoritmo evolutivo de reasignación. Este resuelve exitosamente el 95% de los casos que se le presenta con un número de reasignaciones pormedio de 37.5 reasignaciones con un máximo de 47 y un mínimo de 29 reasignaciones.

5. CONCLUSIONES

Se ha planteado una estrategia basada en un algoritmo evolutivo híbrido que permite asignar canales en redes celulares con independencia del ordenamiento de las llamadas. Esto se realiza a través de reasignaciones de llamadas ya iniciadas y a sabiendas de que la calidad del servicio se afecta. No obstante, se logra evitar el bloqueo de la llamada entrante debido a la falta de canales disponibles. En virtud que la cantidad de canales es un recurso limitado con el que cuenta el operador de servicio, la posibilidad de ubicar la demanda con independencia del orden de arribo de llamadas es un logro de importante impacto económico.

Por otra parte, el método propuesto, permite prever un escenario a medida que las llamadas van ingresando al sistema. Esta previsión resulta en un acomodamiento de las asignaciones de frecuencias que permite incrementar la probabilidad de alojar mayor cantidad de llamadas con menor cantidad de canales disponibles.

REFERENCIAS

[1] S. GHOSH, A. KONAR, A. NAGAR, *Dynamic Channel Assignment Problem in Mobile Networks using Particle Swarm Optimization*, Second UKSIM European Symposium on Computer Modeling and Simulation, 2008, pp. 64-69

[2] D. GÖZÜPEK, G. GENC¸, C. ERSOY, *Channel Assignment Problem in Cellular Networks:A Reactive Tabu Search Approach.* ISCIS METU 2009, pp 298-303.

[3] SEYED ALIREZA GHASEMPOUR SHIRAZI, *A new Hybrid Method for Channel Assignment Problems in Cellular Radio Networks*, Sixth International Conference on Wireless and Mobile Communications, 2010, pp 461-465.

DESIGN AND PRODUCTION PLANNING OF MULTIPRODUCT BATCH PLANTS UNDER UNCERTAINTY

Susana Moreno† and Marcelo Montagna‡

† Planta Piloto de Ingeniería Química, PLAPIQUI – (CONICET - UNS), Camino La Carrindanga km 7, 8000 Bahía Blanca, Argentina, smoreno@plapiqui.edu.ar
‡ INGAR – Instituto de Desarrollo y Diseño (CONICET - UTN), Avellaneda 3657, S3002 GJC Santa Fe, Argentina, mmontagna@santafe-conicet.gov.ar

Abstract: A two-stage stochastic multiperiod LGDP (Linear Generalized Disjunctive Programming) model has been developed for the overall production planning and design of multiproduct batch plants. Both problems are encompassed considering uncertainty in product demands represented by a set of scenarios. The design variables are modeled as here-and-now decisions which are made before the demand realization, while the production planning variables are delayed in a wait-and-see mode to optimize in the face of uncertainty. The proposed model determines the structure of the batch plant (both kinds of unit duplication, in series and in parallel) and the unit sizes, together with the production planning decisions (e.g. quantities to be produced, policy of inventory, purchases of raw materials, and sales of products) in each time period within each scenario. Also, this model allows the incorporation of new equipment items at different periods. The objective is to maximize the expected net present value of the benefit. In order to assess the advantages of the proposed formulation, an extraction process that produces oleoresins is solved.

Key words: Design and planning; Uncertain demands; Units in series; Multiperiod

1. INTRODUCTION

This work is focused on multiproduct batch plants where several products are produced following the same sequence of processing stages. Usually, at the stage of conceptual design of a batch plant, there are process parameters which are subject to considerable uncertainty. Stochastic programming deals with optimization problems whose uncertain parameters are modeled either by continuous probability distributions or by a finite number of scenarios. This last approach has been considerably exploited in the literature [1-2] and it is adopted here for describing uncertainty in product demands. In this first work, the purpose is to generate a representative set of scenarios that are both optimistic and pessimistic within a risk analysis framework [3].

The goal of this paper is to propose a scenario-based approach for the simultaneous design and production planning of multiproduct batch plants under uncertain demands over a multiperiod context. From the design perspective, both kinds of unit duplications, in series and in parallel are considered. A two-stage stochastic model is proposed, where capacity expansion is admitted. New in parallel units working out-of-phase can be added in different time periods. First-stage decisions consist of design variables that allow determining the batch plant structure. Second-stage decisions consist of planning variables to determine the production, purchases, and inventories of raw materials and products for each period throughout the time horizon under each scenario, given the plant structure decided at the first-stage. Generalized disjunctive programming (GDP) has been employed in order to formulate the multiperiod stochastic linear model.

2. PROBLEM DESCRIPTION

Consider a multiproduct batch plant with a set P of operations that processes a set I of products over a time horizon H, which is divided into $t = 1, 2,\ldots, NT$ specified time periods H_t, not necessarily of the same length. Each operation p can be performed by different configurations of units in series. Let H_p denote the set of possible configurations of units in series h for each operation p. The selected configuration of units in series can be also duplicated in parallel operating out-of-phase. The duplication of units in parallel can be different in each time period t allowing the capacity expansion of the plant. Let M_p be the set $\{1, 2,\ldots, M_p^U\}$ of possible number of equal units that can be allocated in parallel in each operation p. Thus, in each time period t, m identical sets of units in parallel operate out-of-phase. The variable N_{pt} represents the number of sets of units in parallel in operation p at each time period t. This value is modified taking into account that $g \in G_p = \{0, 1,\ldots, M_p^U\}$ set of units in parallel can be added in every time period t. Also, the design problem involves the selection of equipment sizes for batch units in each operation p, among a set

$SV_p = \{v_{p1}, v_{p2}, \ldots, v_{p,r_p}\}$ of available discrete sizes. The basic data for representing the operations are the size factors SF_{ipt} and processing time pt_{ipt} required for each product i in each operation p at every time period t. Product demands are uncertain and can be represented by a set of scenarios S. Each scenario has a known probability p_s that reflects the likelihood of each scenario to take place with $\sum_{s\in S} p_s = 1$. Also, these scenarios are described through lower and upper bounds on product demand levels in each time period d_{its}^L and d_{its}^U. The structural option of duplicating of units in series selected in operation p not only affects the operation itself but also the rest of the operations of the process. To maintain the formulation with fixed size and time factors the yield for all the configurations in series in a given operation is assumed to be constant, through appropriate size factors values. In consequence, the size factor for product i in operation p remains equal regardless of the selected configuration of units in series. However, each configuration of units in series h has a different operation time pt_{ipht}. For more details see Moreno and Montagna [5].

In every scenario s, production planning decisions allow to determine at each period t and for each product i, the amount to be produced q_{its}, the number of batches n_{its}, and the total time T_{its} to produce product i. Furthermore, at the end of every period t, the levels of both final product IP_{its} and raw material inventories IM_{its} are obtained. The total sales QS_{its}, the amount of raw material purchased C_{its}, and the raw material to be used for the production RM_{its} of product i in each time period t are determined with this formulation. In this model, it is assumed that each product requires a unique raw material that it is not shared by other products. This assumption is valid for the oleoresins plant example solved below. However, more sophisticated transformation processes can be easily incorporated. If time periods are equal, wastes due to the expired product shelf life PW_{its} and due to the limited raw material lifetime RW_{its} are also added in the formulation. Also, late deliveries ϑ_{its} that take place in each period are determined.

3. LINEAR GENERALIZED DISJUNCTIVE PROGRAMMING MODEL

$$\max ENPV = \sum_{s\in S}\sum_{t\in T}\sum_{i\in I} p_s\ np_{it}\ QS_{its} - \sum_{s\in S}\sum_{t\in T}\sum_{i\in I} p_s \left\{ \begin{array}{l} \kappa_{it}\ C_{its} + \left[\varepsilon_{it} \left(\frac{IM_{i,t-1,s} + IM_{its}}{2} \right) H_t + \sigma_{it} \left(\frac{IP_{i,t-1,s} + IP_{its}}{2} \right) H_t \right] \\ + \left(wp_{it}\ PW_{its} + wr_{it}\ RW_{its} \right) + \left(co_{it}\ q_{its} + cp_{it}\ \vartheta_{its} \right) \end{array} \right\} - \sum_{t\in T}\sum_{p\in P} CE_{pt} \quad (1)$$

$$\underset{h\in H_p}{\vee} \left[\begin{array}{c} Z_{ph} \\ \underset{k\in SV_p}{\vee} \left[\begin{array}{c} W_{phk} \\ n_{its} \geq \left(\frac{SF_{ipt}}{v_{pk}} \right) q_{its} \quad \forall i\in I, t\in T, s\in S \\ CO_p = h\left(\alpha_p\ v_{pk}^{\beta_p}\right) \end{array} \right] \end{array} \right] \quad \forall p\in P \quad (2)$$

$$\underset{m\in M_p}{\vee} \left[\begin{array}{c} Y_{phmt} \\ N_{pt} = m \\ T_{its} \geq \frac{pt_{ipht}}{m} n_{its} \quad \forall i\in I, \forall s\in S \end{array} \right] \quad \forall t\in T, \forall p\in P \quad (3)$$

$$\underset{g\in G_p}{\vee} \left[\begin{array}{c} X_{pgt} \\ N_{pt} = N_{p,t-1} + g \\ CE_{pt} = g\ CO_p\ \gamma_{pt} \end{array} \right] \quad \forall t\in T, \forall p\in P \quad (4)$$

$$Z_{ph} \Leftrightarrow \left(\underset{m\in M_p}{\vee} Y_{phmt} \right) \quad \forall h\in H_p, p\in P, t\in T \quad (5)$$

$$IP_{its} = IP_{i,t-1,s} + q_{its} - QS_{its} - PW_{its} \quad \forall i\in I, t\in T, s\in S \quad (6)$$

$$IM_{its} = IM_{i,t-1,s} + C_{its} - RM_{its} - RW_{its} \quad \forall i \in I, t \in T, s \in S \tag{7}$$

$$IP_{its} \leq \sum_{\tau=t+1}^{t+\chi_i} QS_{i\tau s} \quad \forall i \in I, t \in T, s \in S \tag{8}$$

$$IM_{its} \leq \sum_{\tau=t+1}^{t+\zeta_i} RM_{i\tau s} \quad \forall i \in I, t \in T, s \in S \tag{9}$$

$$\vartheta_{its} \geq \vartheta_{i,t-1,s} + d^L_{its} - QS_{its} \quad \forall i \in I, t \in T, s \in S \tag{10}$$

$$RM_{its} = F_{it}\, q_{its} \quad \forall i \in I, t \in T, s \in S \tag{11}$$

$$\sum_i T_{its} \leq H_t \quad \forall t \in T, s \in S \tag{12}$$

The objective function maximizes the expected net present value (*ENPV*) over a set of scenarios *S*. The economic criterion in Eq. (1) is calculated by the probabilistic average of the difference between the revenue due to product sales and the overall costs in each scenario, with the latter consisting of the cost of raw materials, inventory costs, operating cost, penalty cost for late delivery, and the capital investment cost. Parameters np_{it}, κ_{it}, ε_{it}, σ_{it}, wp_{it}, wr_{it}, co_{it} and cp_{it} are the corresponding cost coefficients for each term.

Structural decisions are modeled through disjunctions in Eqs. (2)-(5). In Eq. (2) the Boolean variable Z_{ph} is true when configuration of units in series *h* is selected in unit operation *p* and is false in the opposite case. Variable W_{phk} is true when discrete unit size *k* is selected to carry out operation *p* with configuration *h*. Thus, constraints into this disjunction correspond to the sizing equation for the units and the equipment cost for this alternative CO_p. the in every operation can be. Duplication of units in parallel working out-of-phase is added by disjunctions in Eq. (3). Here, Boolean variable Y_{phmt} is true when there are *m* identical units in parallel in operation *p* with configuration *h* at time period *t*. Each term of these disjunctions includes constraints that determine the number of set of units in parallel N_{pt} at each period *t* and the total time to produce each product T_{its}. Disjunctions (4) are associated with the discrete choice of the units to be added at each time period in every operation. Here, Boolean variable X_{pgt} is true if *g* units in parallel are added at time period *t* in operation *p* with configuration *h*. The first constraint in these disjunctions determines the number of units in each period *t* considering the number in the previous one plus the units aggregated in the corresponding time period. The last constraint determines the expansion cost in each time period, CE_{pt}. Finally, logical constraint (5) establishes that *m* units in parallel operating out-of-phase will be selected in operation *p* with *h* units in series at time period *t*, if and only if, the configuration in series *h* is selected to carry out operation *p*.

The following planning constraints involve the second-stage variables explicitly associated with each demand scenario. Eqs. (6) and (7) state the inventory level of each product and raw material, respectively. When the lengths of time periods are equal, Eqs. (8) and (9) guarantee that the stock of both raw materials and products in each period cannot be used after the next ζ_i or χ_i time periods, respectively. Eq. (10) calculates the late delivery. The mass balance (11) determines the amount of raw material necessary for the production of product *i* in each period where the parameter F_{it} is the process conversion of product *i* in period *t* assuming that only one main raw material is used for the production of *i*. Eq. (12) ensures that the total time required for producing all products cannot exceed the period length.

4. NUMERICAL EXAMPLE

Consider the multiproduct batch plant that produce three oleoresins, namely, sweet bay (A), pepper (B) and thyme (C) oleoresins, is considered. All the products are manufactured via the following batch operations: (1) extraction, (2) expression, (3) evaporation, and (4) blending. A time horizon of 3 years has been considered, with six equal time periods *t* each 6 months long (3000 h). The process data for this example are shown in Table 1 and prices of raw materials and final products affected by seasonal variation are given in Table 2. There are 4 possible configurations of units in series for the operation 1, with a countercurrent arrangement. Thus, processing times for each product take smaller values as the number of units in series grows. All the operations can be duplicated in parallel up to 3 sets of units operating out-of-phase. A set of 5 discrete sizes is provided to select process units for each operation. Demand for the products is uncertain and 3 scenarios with probabilities $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$ are considered. In the first time period all the scenarios show the same upper demand, i.e., 16500, 12500, and 18500 kg for product A, B, and C, respectively. It is assumed that product demand for each period will increase in

comparison with the present conditions. Thus, the expected growth rates per period in each scenario are 25%, 15%, and 5%, respectively. Minimum product demands in each period for all scenarios are assumed as 50% of maximum product demands.

Table 1. Process data.

	Size factors, S_{ipt} (L/kg)				Processing time, t_{ipt} (h)							Conversion Factor
i	1	2	3	4	1 (h_1)	1 (h_2)	1 (h_3)	1 (h_4)	2	3	4	F_{it}
A	20	15	12	1.5	25.95	9.28	5.35	3.47	1.0	2.5	0.5	11.11
B	23	15	12	1.5	39.46	9.76	5.55	3.59	2.0	1.5	2.0	11.11
C	30	20	17	1.5	27.93	9.41	5.41	3.51	1.0	2.0	1.0	15.87

Table 2. Economic data.

	Costs of raw materials, κ_{it} (\$/kg)						Prices of products, np_{it} (\$/kg)					
Period	1	2	3	4	5	6	1	2	3	4	5	6
A	1.50	2.20	1.58	2.31	1.65	2.43	36.00	38.00	37.80	39.90	39.69	41.90
B	2.50	2.50	2.63	2.63	2.76	2.76	40.00	40.00	42.00	42.00	44.10	44.10
C	1.00	0.80	1.05	0.84	1.10	0.88	37.00	35.00	38.85	36.75	40.79	38.59

The optimal solution of this example was obtained in 22.75 CPU s. This solution has *ENPV* of \$5,910,794.32. The example involves 264 discrete variables, 911 continuous variables, and 4019 constraints. Figure 1 shows the optimal structure of the batch plant obtained in every time period and the unit sizes. In this figure, units in dotted line are included in different time periods. In period 1, there is only one unit in all operations except in operation 1 which has 3 units in series. Later, in period 3, a set of 3 units in series is incorporated in the operation of extraction.

5. CONCLUSION

In this work, a two-stage stochastic LGDP model has been derived to address the design and production planning of multiproduct batch plants in presence of demand uncertainty. The model takes into account both structural decisions of duplicating units in series and in parallel selecting the unit's dimensions from a set of available discrete sizes. Moreover, the proposed approach allows the incorporation of parallel units in different time periods. The performance of the proposed formulation has been assessed through two examples dealing with a batch plant that produces vegetable extracts, particularly oleoresins.

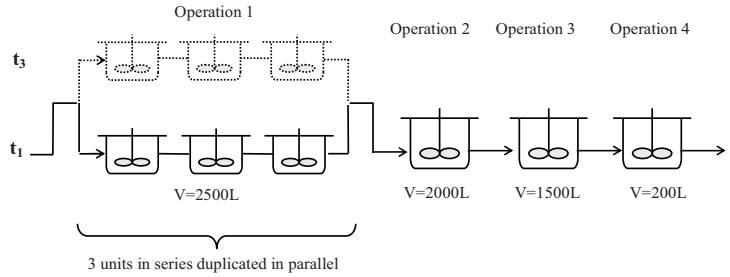


Figure 1: Optimal structure of the plant

REFERENCES

[1] S. SUBRAHMANYAM, J.F. PEKNY, G.V. REKLAITIS, *Design of batch chemical plants under market uncertainty*, Ind. Eng. Chem. Res., Vol. 33 (1994), pp. 2688-2701.

[2] J. CUI, S. ENGELL, *Medium-term planning of a multiproduct batch plant under evolving multi-period multi-uncertainty by means of a moving horizon strategy*, Comput. Chem. Eng., Vol. 34 (2010), pp. 598-619.

[3] J.H.VANSTON, W.P. FRISBIE, S.C. LOPREATO, D.L. POSTON, *Alternate scenario planning*, Technol. Forecast. Social Change, Vol. 10 (1977), pp.159-180.

[4] J.M. MULVEY, *Generating scenarios for the Towers Perrin Investment System*, Interfaces Vol. 26 (1996) pp. 1-15.

[5] M.S. MORENO AND J.M. MONTAGNA, *Optimal simultaneous design and operational planning of vegetable extraction processes*, Trans IChemE, Part C, Food Bioprod. Proc., Vol. 85 (2007), pp. 360-371.

DUALIDAD Y PROPIEDADES TIPO LIPSCHITZ EN OPTIMIZACIÓN

Marco A. López†, Andrea B. Ridolfi‡ y Virginia N. Vera de Serio†‡

† Dep. of Statistics and Operations Research, University of Alicante, Spain; Honorary Research Fellow in the Graduate School ofInformation Technology and Mathematical Sciences at University of Ballarat, Australia, marco.antonio@ua.es
‡ CONICET, Facultad de Ciencias Aplicadas a la Industria, Universidad Nacional de Cuyo, Mendoza, Argentina, abridolfi@gmail.com
†‡ Facultad de Ciencias Económicas, I.C.B., Universidad Nacional de Cuyo, Mendoza, Argentina, vvera@uncu.edu.ar

Resumen: En este trabajo estudiamos la estabilidad de los conjuntos factibles del problema dual, en optimización lineal en dimensión infinita con infinitas restricciones lineales y una restricción cónica adicional. Para ello, aplicamos algunos conceptos de la teoría de diferenciación generalizada, entre ellos el de Coderivada, para la aplicación a valores-conjunto de los conjuntos factibles. Analizamos también propiedades tipo Lipschitz de ésta aplicación y obtenemos cotas para el valor de la cota exacta Lipschitz.

Palabras claves: *programación semi infinita, dual, optimización convexa*
2000 AMS Subjects Classification: 90C34 - 90C48 - 90C31 - 49J53 - 46A20

1. INTRODUCCIÓN

Nuestro problema de optimización lineal perturbado es el siguiente:

$$P(b,c^*): \sup\left\langle \bar{c}^* + c^*, x\right\rangle$$
$$s.a.\ \left\langle a_t^*, x\right\rangle \le \bar{b}_t + b_t,\ t\in T,$$
$$x\in Q,$$

donde T es un conjunto de índices arbitrario, posiblemente infinito, Q es un cono convexo cerrado en un espacio real de Banach X, $\bar{b}_t,\ t\in T,$ son números reales y $a_t^*,\ t\in T$, y $\bar{c}^*$ están en el dual topológico de X, denotado por X^*. Considerando perturbaciones $c^*\in X^*$ y $b=(b_t)_{t\in T}\in \ell_\infty(T)$, donde $\ell_\infty(T)$ es el espacio real de Banach de todas las funciones acotadas en T con la norma supremo: $\|b\|_\infty = \sup\{b_t : t\in T\}$. (Omitiremos el subíndice ∞ en el símbolo de la norma). El conjunto $\{a_t^*,\ t\in T\}\subset X^*$ es fijo, arbitrario y acotado con la norma dual en X^* dada por: $\|x^*\| = \sup\{\langle x^*, x\rangle : \|x\|\le 1\}$, (usaremos indistintamente $\|\ \|$ para la norma en X y la correspondiente norma dual en X^*). Observemos que, para cada $x\in X$, $\left\langle a_{(.)}^*, x\right\rangle \in \ell_\infty(T)$.

El problema primal $P(b,c^*)$ tiene un problema dual asociado, también perturbado, llamado $D(b,c^*)$, definido por:

$$D(b,c^*): \inf\left\langle \mu, \bar{b}+b\right\rangle$$
$$s.a.\ A^*\mu \in \bar{c}^* + c^* - Q^0,$$
$$\mu \ge 0,$$

donde $\mu\in \ell_\infty(T)^*$, $A: X\to \ell_\infty(T)$ es el operador lineal definido por $Ax := \left\langle a_{(.)}^*, x\right\rangle$, (acotado y continuo debido a la acotación de $\{a_t^*,\ t\in T\}$), $A^*: \ell_\infty(T)^*\to X^*$ es el operador adjunto de A (es decir $\left\langle A^*\mu, x\right\rangle = \left\langle \mu, Ax\right\rangle$ para cada $\mu\in \ell_\infty(T)^*$, y cada $x\in X$) y Q^0 es el cono dual de Q, dado por $Q^0 := \{q^*\in X^* : \left\langle q^*, q\right\rangle \le 0$ para todo $q\in Q\}$.

Llamaremos P y D a los problemas $P(b,c^*)$ y $D(b,c^*)$ no perturbados respectivamente (es decir $b=0$ y $c^*=0$). D se puede considerar dentro del modelo de dualidad desarrollado por Kretschmer en [14] y tiene un nivel medio de generalidad entre los aportes en Duffin [9] y Borwein [2]. Anderson y Nash dan detalles de esta teoría en [1, Chapter 3]. Últimamente ha aumentado el interés de la estabilidad en optimización, muchos prefieren trabajar con soluciones estables en vez de soluciones óptimas inestables. En [6],[7] y [8] se estudia la estabilidad cualitativa. En cuanto a la perspectiva cuantitativa, ver [4], [3], etc. El resumen reciente [15], nos da una idea amplia de lo hecho en los últimos quince años tanto cualitativa como cuantitativamente. Las referencias que nos inspiraron son, Cánovas et al. [5], Ioffe y Sekiguchi [12], y el reciente preprint [11].

Nuestro objetivo es obtener caracterizaciones de la estabilidad Lipschitziana de las soluciones factibles del problema dual, establecido en dimensión infinita. No requerimos que X sea reflexivo. Si bien en el problema primal obtuvimos una fórmula para el módulo Lipschitz asociado, aquí no lo presentaremos, sin embargo daremos cotas para el módulo Lipschitz del problema dual cuya situación es un poco más compleja. Para lograrlo, usaremos una potente herramienta del análisis variacional, como lo es la noción de coderivada, la cual es aplicada a problemas de optimización y control (ver [13], [16]). En [5] se aplicaron para analizar la estabilidad del problema primal en programación semi- infinita. .

2. NOTACIONES Y DEFINICIONES

Para un subconjunto $\Omega \subset Z$ de un espacio de Banach Z, denotamos por $\operatorname{conv}\Omega$, $\operatorname{cone}\Omega$, $\operatorname{int}\Omega$, y $\operatorname{cl}\Omega$, a la cápsula convexa, cápsula cónica, interior y clausura de Ω , respectivamente (los dos últimos con respecto a la topología de la norma). La clausura en la topología débil estrella de un subconjunto $\Phi \subset Z^*$ la denotamos por $\operatorname{cl}^*\Phi$. En el caso $Z=\ell_\infty(T)$, es sabido que, (ver [10]), hay un isomorfismo isométrico entre $\ell_\infty(T)^*$ y $ba(T)$, el espacio de medidas $\mu: 2^T \to IR$, acotadas y aditivas. La norma dual en $\ell_\infty(T)^*$ es la variación total, (para $\mu \geq 0$ tenemos $\|\mu\| = \mu(T)$).

Dada una aplicación conjunto valuada $M: Z \to Y$ denotamos su dominio, gráfico y aplicación inversa (respectivamente) por $\operatorname{dom}M$, $\operatorname{gph}M$ y M^{-1}. Si $(\bar{z},\bar{y}) \in \operatorname{gph}M$, la ***coderivada*** de M en $(\bar{z},\bar{y})$ (*normal coderivative* en [16]) es la aplicación homogénea positiva $D^*M(\bar{z},\bar{y}): Y^* \to Z^*$ definida por:

$$D^*M(\bar{z},\bar{y})(y^*) := \{z^* \in X^* : (z^*,-y^*) \in N((\bar{z},\bar{y}),\operatorname{gph}M)\}, \quad y^* \in Y^*,$$

donde $N((\bar{z},\bar{y}),\operatorname{gph}M)$ es el cono normal al $\operatorname{gph}M$ en $(\bar{z},\bar{y})$ definido en [16, p.4]. El módulo de la coderivada está dado por

$$\|D^*M(\bar{z},\bar{y})\| := \sup\{\|z^*\| : z^* \in D^*M(\bar{z},\bar{y})(y^*), \ \|y^*\| \leq 1\}.$$

Si Z e Y son espacios normados y $(\bar{z},\bar{y}) \in \operatorname{gph}M$, se dice que M es ***Tipo Lipschitz*** alrededor de $(\bar{z},\bar{y})$ (*locally Lipschitz-like* in [16]) con módulo $\ell \geq 0$, si existen entornos U de $\bar{z}$ y V de $\bar{y}$ tal que

$$M(z) \cap V \subset M(u) + \ell\|z-u\|B_Y, \quad \text{para todo } z,u \in U$$

donde B_Y es la bola unitaria cerrada en el espacio Y. El ínfimo de éstos módulos ℓ se llama ***cota exacta Lipschitz*** de M alrededor de $(\bar{z},\bar{y})$ y se denota por $\operatorname{lip}M(\bar{z},\bar{y})$. Esta cota admite la siguiente representación:

$$\operatorname{lip}M(\bar{z},\bar{y}) = \limsup_{(z,y)\to(\bar{z},\bar{y})} \frac{dist(y,M(z))}{dist(z,M^{-1}(y))},$$

donde $dist(x,\phi) = \infty$ y por convención $0/0 := 0$, $\infty/\infty = \infty$. Escribimos $\operatorname{lip}M(\bar{z},\bar{y}) = \infty$, si M no es Tipo Lipschitz alrededor de $(\bar{z},\bar{y})$. Esta cota equivale a la inversa de la tasa de suryección de M^{-1} alrededor de $(\bar{z},\bar{y})$ (ver [12, p.256]).

3. ESTABILIDAD LIPSCHITZIANA PARA LA APLICACIÓN FACTIBLE DEL DUAL

Para estudiar la estabilidad del dual, consideraremos la aplicación factible a valores-conjunto (o multifunción) $F : X^* \to \ell_\infty(T)^*$, definida como: $F(c^*) = \left\{\mu \in \ell_\infty(T)^* : A^*\mu \in \bar{c}^* + c^* - Q^0, \text{ y } \mu \geq 0\right\}$.

Reformulamos el problema dual perturbado de manera que nos permita aplicar resultados conocidos:

$$\begin{aligned} D(b,c^*): \quad & \inf \left\langle \mu, \bar{b} + b \right\rangle \\ s.a. \quad & \left\langle \mu, Aq \right\rangle \geq \left\langle \bar{c}^*, q \right\rangle + \left\langle c^*, q \right\rangle, \quad q \in \tilde{Q}, \\ & \left\langle \mu, p \right\rangle \geq -1, \qquad\qquad p \in \ell_\infty(T)_+, \end{aligned}$$

donde $\tilde{Q}$ es cierto conjunto cerrado acotado que no contiene el vector nulo y genera el cono *Q*. De aquí en adelante consideraremos un cierto $\tilde{Q}$ fijo.

Definición 1 *F satisface la condición Strong Slater en* $c^* \in X^*$, *si existe* $\mu \in \ell_\infty(T)$, $\mu \geq 0$ *tal que*

$$\inf_{q \in \tilde{Q}} \left\{ \left\langle \mu, Aq \right\rangle - \left\langle \bar{c}^* + c^*, q \right\rangle \right\} > 0 .$$

Tal μ *se llama punto Strong Slater de F en c*.*

Esta definición es independiente de la elección del conjunto cerrado acotado $\tilde{Q}$, además $0 \notin \text{cl}(\text{conv}\tilde{Q})$ cuando *F* cumple la condición Strong Slater en $c^* \in X^*$, lo cual implica que el cono Q tiene vértice.

Definimos, al igual que en [5], el conjunto característico de $F(c^*)$ relativo a $\tilde{Q}$, como el subconjunto convexo de $\ell_\infty(T) \times IR$ dado por: $C_D(c^*) := \text{conv}\left(\left\{\left(Aq, \left\langle \bar{c}^* + c^*, q \right\rangle\right) : q \in \tilde{Q}\right\} \cup \left\{(p,-1) : p \in \ell_\infty(T)_+\right\}\right)$.

Podemos estudiar la estabilidad con respecto a la consistencia mediante la aplicación *F* observando que un problema dual $D(b,c^*)$, es estable consistente si y sólo si $c^* \in \text{int}(\text{dom}F)$. Una aplicación del clásico teorema de Robinson-Ursescu (ver [16]) indica que esta condición es equivalente a que *F* sea Lipschitz-like alrededor de (c^*, μ) para todo $\mu \in F(c^*)$. Además probamos que un problema consistente $D(b,c^*)$, satisface la condición Strong Slatter si y sólo si $(0,0) \notin \text{cl}^* C_D(c^*)$, y si además $0 \notin \text{cl}(\text{conv}\tilde{Q})$, esto es equivalente a que $c^* \in \text{int}(\text{dom}F)$.

La siguiente condición se necesita para obtener una estimación de la norma de la coderivada, la cual nos servirá para estimar la cota exacta Lipzchitz de la aplicación factible dual.

(A): $\tilde{Q}$ es un subconjunto cerrado generador de Q tal que existan números reales positivos *r* y *R*; y un $\bar{x}^* \in X^*$, $\|\bar{x}^*\| = 1$, que cumplan que $0 < r \leq \left\langle \bar{x}^*, q \right\rangle \leq \|q\| \leq R$ para todo $q \in Q$. (Notar que en este caso el cono Q tiene vértice. Por ejemplo, se puede tomar cualquier base compacta $\tilde{Q}$ de Q).

Teorema 2 *Supongamos que Q satisface la condición (A) y que* $\bar{\mu} \in F(0)$. *Entonces*

(i) *Si* $\bar{\mu}$ *es un punto Strong Slater de F en 0; entonces* $\|D^* M(\bar{z}, \bar{y})\| = 0$.

(ii) *Si* $\bar{\mu}$ *no es un punto Strong Slater de F en 0; entonces* $\|D^* M(\bar{z}, \bar{y})\| > 0$ *y*

$$r.\sup\left\{\left\|b^{**}\right\|^{-1} : \left(b^{**}, \left\langle \bar{\mu}, b^{**} \right\rangle\right) \in cl^* C_D(0)\right\} \leq \|D^* M(\bar{z}, \bar{y})\| \leq R.\sup\left\{\left\|b^{**}\right\|^{-1} : \left(b^{**}, \left\langle \bar{\mu}, b^{**} \right\rangle\right) \in cl^* C_D(0)\right\}$$

Aplicando este Teorema, algunos resultados de [16], y siguiendo los pasos de [5] podemos obtener los mismos resultados, reemplazando $\text{lip}M(\bar{z}, \bar{y})$ en lugar de $\|D^* M(\bar{z}, \bar{y})\|$. Así, podemos estimar la diferencia entre las constantes $\text{lip}M(\bar{z}, \bar{y})$ y $\|D^* M(\bar{z}, \bar{y})\|$, y también obtenemos un supuesto que nos garantiza la igualdad entre ambas constantes. La situación del dual es más compleja que la del problema

primal donde la igualdad entre ambas constantes siempre se cumple, (ver Proposición 5 en [11]). Esta estimación la describimos en el próximo Corolario.

Corolario 3 *(i) Sea* $\overline{\mu} \in F(0)$ *. Si Q satisface la condición (A) y F satisface la condición Strong Slater en 0, entonces existen constantes r y R,* $0 < r \le R$ *, que dependen solamente del cono Q, tal que*

$$0 \le \operatorname{lip} M(\overline{z}, \overline{y}) - \left\| D^* M(\overline{z}, \overline{y}) \right\| \le (R - r).\max\left\{ \left\| b^{**} \right\|^{-1} : \left(b^{**}, \left\langle \overline{\mu}, b^{**} \right\rangle \right) \in cl^* C_D(0) \right\}$$

(ii) Si $F^{-1}(\ell_\infty(T)^*_+)$ *tiene interior no vacío con respecto a la topología de la norma en X*, y* $\{q \in Q : \|q\| = 1\}$ *es w*-cerrada en X**, entonces*

$$\operatorname{lip} M(\overline{z}, \overline{y}) = \left\| D^* M(\overline{z}, \overline{y}) \right\| .$$

Notar que la cota superior para $\operatorname{lip} M(\overline{z}, \overline{y}) - \left\| D^* M(\overline{z}, \overline{y}) \right\|$ depende de las constantes obtenidas del cono Q y del conjunto característico $C_D(0)$ correspondiente a cierto conjunto generador cerrado $\widetilde{Q}$. En la prueba de (ii) se aplica [12, Proposición 5] y resultados obtenidos en aplicaciones perfectamente regular.

AGRADECIMIENTOS

Agradecemos a MICINN de España, Grant MTM2008-06695-C03-01/03, ARC Discovery Project DP110102011 of Australia, y SECTyP-UNCUYO, Argentina.

REFERENCIAS

[1] E. ANDERSON, P. NASH, *Linear Programming and Infinite-Dimensional Space*, Wiley and Sons, 1987.
[2] J. M. BORWEIN, *Semi-infinite programming duality: How special is it?*, in A.V. Fiacco and K.O. Kortanek (eds.), Semi-Infinite Programming and Applications, Springer-Verlag, Berlin,1983.
[3] M.J. CÁNOVAS, A. HANTOUTE, M.A. LÓPEZ, J. PARRA, *Lipschitz modulus in convex semi-infinite optimization via d.c. functions*, ESAIM Control Optim. Calc. Var.,15 (2009), pp.763-781.
[4] M.J. CÁNOVAS, D. KLATTE, M.A. LÓPEZ, J. PARRA, *Metric regularityin convex semi-infinite optimization under canonical perturbations*, SIAM J.Optim., 18 (2007), pp.717-732.
[5] M.J. CÁNOVAS, M.A. LÓPEZ, B.S. MORDUKHOVICH, J. PARRA, *Variational Analysis in Semi-infinite and Infinite Programming, I: Stability of Linear Inequality Systems of Feasible Solutions*, SIAM J. Optim., 20, no. 3,(2009), pp.1504-1526.
[6] M.J. CÁNOVAS, M.A. LÓPEZ, J. PARRA, M.I. TODOROV, *Stability and well-posedness in linear semi-infinite programming*, SIAM J. Optim., 10 (1999), pp.82-99.
[7] N. DINH, M.A. GOBERNA, M.A. LÓPEZ, *On the stability of the feasible set in optimization problems*, SIAM J. Optim., 20 (2010), pp.2254-2280.
[8] N. DINH, M.A. GOBERNA, M.A. LÓPEZ, *On the stability of the optimal value and the optimal set in optimization problems*, Preprint, Dpt. of Statistics and Operations Research, Alicante University, 2010.
[9] R.J. DUFFIN, *Infinite Programs. Linear inequalities and related systems*, Annals of Mathematics Studies, 38 (1956), pp.157-170.
[10] N. DUNFORD, J.T. SCHWART, *Linear Operators Part I: General Theory*,Wiley, New York, 1988.
[11] A.D. IOFFE, *On stability of solutions to systems of convex inequalities*, CRM Preprints, 984 (2010).
[12] A.D. IOFFE, Y. SEKIGUCHY, *Regularity estimates for convex multifunctions*,Math. Program., Ser. B, 117 (2009), pp.255-270.
[13] D. KLATTE, B. KUMMER, *Nonsmooth Equations in Optimization. Regularity, Calculus, Methods and Applications*, Kluwer, Dordrecht, 2002.
[14] K.S. KRETSCHMER, *Programmes in paired spaces*, Canad. J. Math., 13 (1961), pp. 221-238.
[15] M.A. LÓPEZ, *Stability in Infinitely Constrained Optimization: A Personal Tour*, preprint, Dpt. Of Statistics and Operations Research, Alicante University, 2010.
[16] B.S. MORDUKHOVICH, *Variational Analysis and Generalized Diferentiation I:Basic Theory*, Springer, Berlin, 2006.

DESARROLLO DE UN MODELO MATEMÁTICO SIMPLE PARA EL DISEÑO DE UNA PLANTA DE CAPTURA DE DIÓXIDO DE CARBONO

Néstor H. Rodríguez[1], Sergio Mussati [1 2] and Nicolás J. Scenna[1 2]
[1]*CAIMI, UTN Facultad Regional Rosario, Zeballos 1341, S2000BQA, Rosario, Argentina*
[2]*INGAR/CONICET Instituto de Desarrollo y Diseño, Avellaneda 3657 C.P. 3000 Santa Fe, Argentina*
E-mail: nestorhugo_r@yahoo.com

Resumen: Este trabajo tiene como principal objetivo presentar un modelo simple, del tipo lineal con pocas ecuaciones y variables que permita determinar el diseño preliminar de los principales equipos del proceso de captura de CO_2 y su motivación obedece a resultados obtenidos y presentados previamente. Ciertamente, el modelo propuesto fue derivado a partir de un minucioso análisis de resultados óptimos obtenidos utilizando un modelo más riguroso y detallado, desarrollado con anterioridad. Así, a partir de dicho análisis se desarrolló un modelo no-fenomenológico que involucre únicamente los principales parámetros y variables del proceso de manera tal de poder representar sencillamente el espacio de soluciones optimas para una dada especificación del gas a tratar (caudal, composición y temperatura).

Como ilustración, el modelo es aplicado a un caso de estudio y los resultados obtenidos son comparados con aquéllos determinados en forma rigurosa. Finalmente, se detallan las conclusiones y tareas futuras.

Palabras claves: diseño-parametrización-*optimización- simulación- captura de CO_2*
2000 AMS Subjects Classification: 68U20 - 90C31

1. INTRODUCCIÓN. DESCRIPCIÓN DEL PROCESO

El uso extendido de los combustibles fósiles (líquidos, sólidos y gaseosos) ha contribuido al incremento de la concentración de gases generadores de efecto invernadero (GEI) principalmente dióxido de carbono. No obstante, dada la eficiencia y abundancia relativa de dichos combustibles, en la etapa de transición hacia nuevos paradigmas energéticos más sustentables, son imprescindibles para suministrar la energía necesaria para el desarrollo humano. Es por esta razón que se investigan numerosas variantes para procesos de captura de CO_2 acoplados a plantas de generación termoeléctricas. La alternativa más estudiada es la captura postcombustión, es decir, la absorción del CO_2 de los gases de combustión antes de ser expulsado a la atmósfera. El CO_2 puro así obtenido se lo acondiciona para su transporte y/o disposición final según se adopte como destino para el mismo. Esta alternativa es la más atractiva ya que puede ser integrada fácilmente a las plantas de generación de energía eléctrica en operación sin la necesidad, prácticamente, de incluir equipos adicionales.

La Fig. 1 muestra un proceso convencional de captura de CO_2 por absorción química mediante el uso de una solución acuosa de amina, en nuestro caso DEA (dietanolamina). Como puede observarse, el gas de proveniente de la combustión se pone en contacto en contracorriente con una solución de DEA que lo despoja de la mayor parte del CO_2. El gas desprovisto de CO_2 se libera a la atmósfera mientras que la solución de DEA con el CO_2 disuelto se envía a una segunda columna en donde se recupera la solución de amina separándola del CO_2, el cual se obtiene en forma prácticamente pura y se dispone apropiadamente. La solución de amina se recircula a la primera columna cerrando el ciclo de absorción – desorción o de captura.

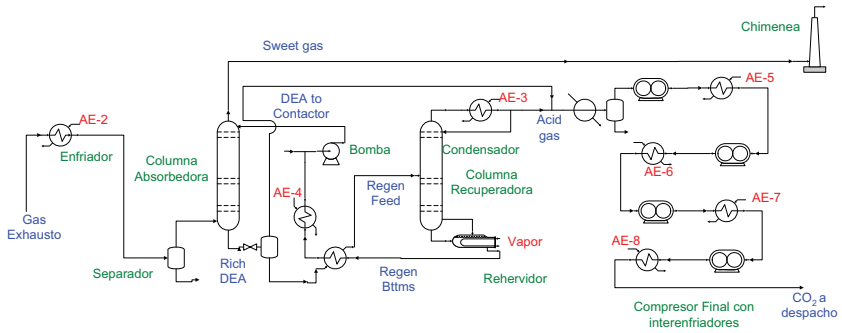


Figura 1: Flowsheet del proceso de captura

2. PLANTEO DEL PROBLEMA DE DISEÑO ÓPTIMO

En este trabajo abordamos el diseño óptimo económico del proceso, el cual simbólicamente puede formularse de la siguiente manera:

$$\min f(x,y)$$
$$\text{s.t.} \quad h_m(x,y) = 0, \quad m \in M \quad ; \quad g_k(x,y) \le 0, \quad k \in K$$
$$x \in X \le R^n \quad ; \quad y \in Y = \{0,1\}$$

siendo f(x,y) la función objetivo que representa el costo total del proceso (inversión más costos de operación). Por su parte, $h_m(x,y)$ son restricciones de igualdad y representan a las restricciones que modelan cada uno de los equipos-componentes de la planta (balances de materia, energía, y eventualmente cantidad de movimiento). Por otro lado, $g_k(x,y)$ son las funciones de desigualdad que se corresponden con las restricciones naturales que surgen de la operación del proceso, tales como la prevención de cruces de temperaturas, el nivel fijado para la recuperación de CO_2 deseados, y otros parámetros del problema.

Entre las principales variables se pueden mencionar: flujo de la solución de amina, carga de CO_2, presiones de operación de la columna absorbedora y desorbedora y temperatura de entrada de la amina a la columna de destilación.

3. DESARROLLO DEL MODELO SIMPLIFICADO.

Como fue mencionado al comienzo, en trabajos previos se formularon modelos rigurosos y detallados para la optimización del proceso de captura considerando distintas especificaciones de gas (caudal, composición, temperatura) y diferentes tipos de aminas [Rodríguez y col. 2009, 2010]. Dichas optimizaciones fueron realizadas con el simulador de procesos HYSYS, y para tal fin se propuso un método de optimización adecuado que permitió garantizar la convergencia de los modelos desarrollados, y así superar las dificultades computacionales que surgieron por la presencia de restricciones altamente no-lineales. Se estudiaron distintas soluciones de amimias: DEA+agua, MDEA+agua y sus mezclas (DEA+MDEA+agua).

Aquí en este trabajo, por simplicidad y a fin de reducir la magnitud del problema, se asume conocido el tipo y composición del solvente.

La idea básica fue establecer un vínculo entre las principales variables del proceso con los parámetros más importantes del mismo de manera tal de poder conducir un diseño preliminar del proceso en forma sencilla. Así, según se describe en la Fig. 2 el desafío fue "condensar" en forma simple, a partir de soluciones óptimas rigurosas y detalladas, toda la información necesaria para estimar dimensiones y costos de equipos para aquellas condiciones especificadas por el usuario.

A continuación se presentan los principales resultados obtenidos previamente e hipótesis de trabajo utilizados para derivar el modelo matemático.

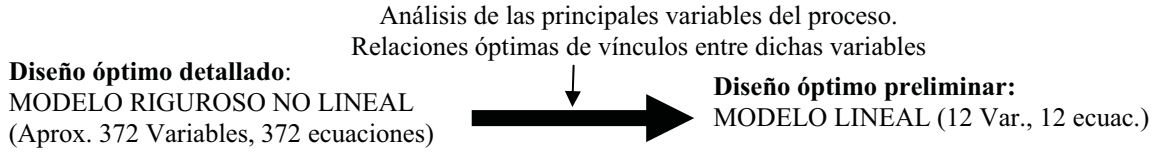


Figura 2: Pasos en la derivación de modelo

La Figura 3 ilustra la distribución del costo total de inversión de los distintos equipos para el proceso correspondiente a la Fig. 1 y para las siguientes especificaciones:

Gas: caudal 28,0 [MMmol/h] composición, 74,26 % N_2; 7, 74 % O_2; 12,00 % de H_2O y 6,00 % de CO_2 (base molar)

Amina: Temperatura, 35 ºC; cadual, 34,11 [MMmol/h], composición, 40 % [másico] de DEA en agua.

Se puede observar claramente que los equipos que mayor incidencia presentan en el costo total de inversión son: compresores, absorbedor, LR-HEX y la unidad regeneradora de amina. Similares distribuciones de costos se verificó para otras especificaciones, no abordados en este trabajo.

Por su parte, la Fig. 4 indica los diámetros óptimos que corresponden a las columnas de absorción y desorción para distintos flujos de gas y de CO_2 total a tratar, obtenidos a partir de simulaciones rigurosas. Según se puede observar claramente en dichas figuras, para un determinada cantidad total de CO_2 presente en el gas a tratar, el diámetro depende fuertemente del caudal de gas a tratar, mientras que permanece

prácticamente invariante para distintas cantidades totales de CO_2 y manteniendo fijo el caudal de gas a tratar. Exactamente lo opuesto podemos concluir para la regeneradora.

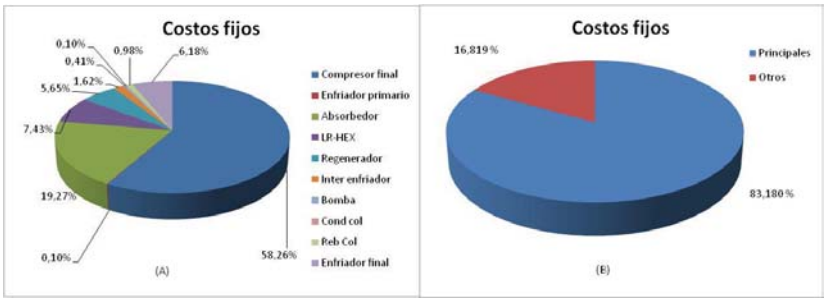


Figura 3: Costo fijo total por equipo

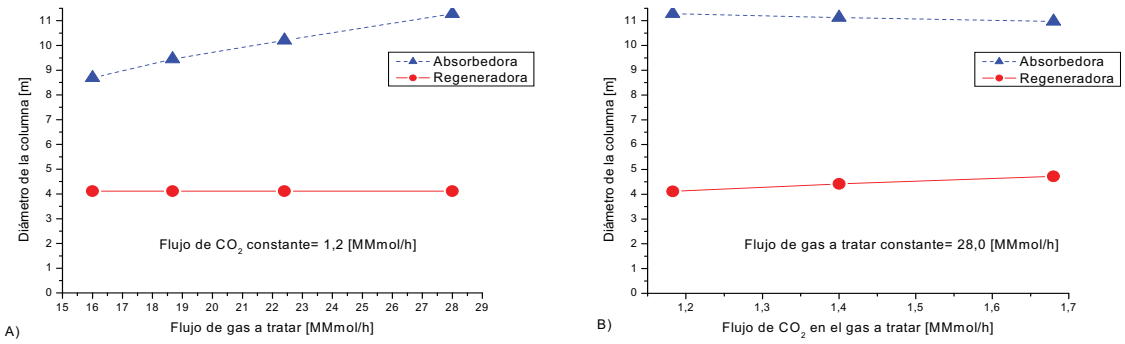


Figura 4: Diámetro de las columnas en función de flujo de gas y cantidad total de CO_2

A partir de esto, es posible aproximar los diámetros de la absorbedora y regeneradora en función del caudal de gas a tratar y del flujo total de CO_2 respectivamente, según se indica a continuación:

$$D_{Columna\ Absorbedora}(m) = 0{,}2119 * F_{fluegas}\ (\text{MMmol/h}) + 5{,}3992$$

$$D_{Columna\ Recuperadora}(m) = 1{,}2202 * F_{CO2}\ (\text{MMmol/h}) + 2{,}6856$$

Luego, se procedió del mismo con el resto de las variables más importante del proceso, obteniéndose así, sus respectivas aproximación en función del caudal de gas a tratar ó caudal total de CO_2 según corresponda. Por ejemplo, las siguientes son las aproximaciones obtenidas para el consumo de serv. aux. de calefacción y de enfriamiento, electricidad demandada por los compresores, y algunos costos de inversión:

$$C_{Columna\ Abosrbedora}\ (MM\$us/año) = 0{,}0478726 \times F_{fluegas}\ (\text{MMmol/h}) + 0{,}349322$$

$$C_{Columna\ Recuperadora}\ (MM\$us/año) = 0{,}136841 \times F_{CO2}\ (\text{MMmol/h}) + 0{,}239478$$

$$C_{Compresor}\ (\text{MM\$us/año}) = 2{,}63719 \times F_{CO2}\ (\text{MMmol/h}) + 0{,}463350$$

$$Q_{Condensador}\ [\text{MW}] = -0{,}00786 \times F_{fluegas}\ (\text{MMmol/h}) + 0{,}59130 \times F_{CO2}\ (\text{MMmol/h}) - 0{,}04504$$

$$Q_{Reboiler}\ [\text{MW}] = -0{,}29631 \times F_{fluegas}\ (\text{MMmol/h}) + 40{,}85331 \times F_{CO2}\ (\text{MMmol/h}) - 1{,}30053$$

$$Q_{Compresor}\ [\text{MW}] = -0{,}23980 \times F_{fluegas}\ (\text{MMmol/h}) + 15{,}87251 \times F_{CO2}\ (\text{MMmol/h}) - 1{,}41595$$

3. RESULTADOS Y VALIDACIÓN

En esta sección se presentan los resultados obtenidos para un caso de aplicación específico. Un ingeniero de procesos responsable de las emisiones gaseosas de una planta generadora de electricidad,

desea estimar el costo, los servicios auxiliares de calefacción y enfriamiento y las dimensiones de las columnas de absorción y desorción necesarios para absorber químicamente el 95 % de CO_2 presente en el gas de combustión cuyo caudal y composición (contenido de $CO_{2)}$ son, respectivamente,=$25x10^6$ [moles/h] y 4,8 % [molar]. Se dispone, para la absorción, de una solución acuosa de DEA al 40 % [peso] .

Variable de interés	Simulación Rigurosa	Modelo Simplificado
Costo total [MMus$/año]	25,333	24,671
Costo Fijo [MMus$/año]	3,642	3,700
Costo Operativo [MMus$/año]	21,691	20,973
Costo de la Absorbedora [Mmus$]	1,467	1,546
Diámetro de la Absorbedora [m]	10,363	10,697
Costo de la Regeneradora [Mmus$]	0,364	0,404
Diámetro de la Regeneradora [m]	3,962	4,150
Costo de Compresores [Mmus$]	3,701	3,628
Electricidad consumida por compresores [MW]	11,511	11,636
Energía consumida en el condensador, [MW]	0,463	0,468
Energía consumida en el reboiler [MW]	40,291	40,316

Tabla 1: Caso de estudio, valores comparativos con la optimización rigurosa

De los resultados presentados en la Tabla1, se puede concluir claramente que el diseño preliminar puede llevarse a cabo en forma sencilla y con bastante exactitud al que se obtiene empleando un modelo detallado y riguroso del proceso completo.

4. CONCLUSIONES

Mediante el uso exhaustivo de un simulador comercial de procesos (Hysys®) se obtuvieron datos de diseños óptimos para distintas especificaciones de gas a tratar. Estos datos se emplearon para desarrollar un modelo simplificado que permita abordar el diseño preliminar del proceso de captura de CO_2 (post-combustión). El modelo desarrollado fue satisfactoriamente validado comparando los valores pre-dichos con aquéllos obtenidos en forma rigurosa y estima con aceptable exactitud las dimensiones de los distintos equipos y sus respectivos costos asociados, consumos de servicios auxiliares (calefacción y enfriamiento) como así también valores de flujos de las principales corrientes. Así, éste permite rápida y sencillamente obtener valores preliminares sin la necesidad de cálculos complejos.

En trabajos futuros, se propondrá ampliar el rango de aplicabilidad con respecto a la composición y caudal total de gas. Para esto, posiblemente sea necesario desarrollar modelos no lineales (aunque aun simplificados) ó en su defecto proponer modelos lineales por tramos. También, se propondrá considerar el grado de recuperación deseado como así la composición de la mezcla de aminas variables del modelo lo que otorgará al modelo mayor flexibilidad y completitud.

AGRADECIMIENTOS

Al centro de Aplicaciones Informáticas y Modelado en Ingeniería (CAIMI), Universidad Tecnológica Nacional Facultad Regional Rosario y al Instituto de Desarrollo y Diseño (INGAR/CONICET) de la ciudad de Santa Fe por sus soportes técnicos y económicos al presente trabajo.

REFERENCIAS

[1] N. H. Rodríguez, S. F. Mussati, N. J. Scenna *"Modelado y optimización de una Planta de captura de Dióxido de carbono mediante el uso de un simulador secuencial comercial"*. MACI, 2 (2009),333-336

[2] N. H. Rodríguez, S. F. Mussati, N. J. Scenna, "*Optimization of post-combustion CO_2 process using DEA–MDEA mixtures"*. Chemical Engineering Research and Design (2010) En prensa.

OPTIMIZACIÓN DINÁMICA DE INTERCAMBIADORES DE CALOR CRIOGÉNICOS CON Y SIN CAMBIO DE FASE

Juan I. Laiglecia, P.Hoch y M. Soledad Diaz†

†Planta Piloto de Ingeniería Química (Universidad Nacional del Sur - CONICET), Camino La Carrindanga Km 7, 8000 Bahía Blanca, Argentina, sdiaz@plapiqui.edu.ar

Resumen: En este trabajo se formulan modelos dinámicos para intercambiadores de calor de carcasa y tubo, contracorriente, con y sin cambio de fase con un enfoque orientado a ecuaciones. El problema resultante de la formulación de los balances de energía dinámicos en el intercambiador en contra-corriente, es transformado en un sistema de ecuaciones diferenciales ordinarias (ODE) mediante la aplicación del método de líneas (Schiesser, 1991). Las ecuaciones algebraicas del modelo están conformadas por la ecuación de estado de Soave-Redlich-Kwong (SRK) para la predicción de las propiedades termodinámicas, ecuaciones hidráulicas y de diseño. El problema se ha implementado en un ambiente de AMPL. En este entorno, el problema de optimización dinámica se discretiza en el tiempo por colocación ortogonal sobre elementos finitos, transformándolo en un problema no lineal a gran escala, que se resuelve en IPOPT mediante un método de punto interior que emplea técnicas de Programación Cuadrática Sucesiva en espacio reducido.

Palabras claves: *Optimización Dinámica, Intercambiadores de Calor Criogénicos, Condensación Parcial*

1. INTRODUCCIÓN

El análisis dinámico de intercambiadores provee información sobre las respuestas transitorias sujetas a diversos tipos de perturbaciones. Con respecto a la complejidad, los modelos dinámicos de intercambiadores pueden ser de parámetros concentrados o distribuidos. En el primer caso, se consideran las temperaturas de los fluidos como funciones del tiempo. En el caso de modelos a parámetros distribuidos, se consideran variaciones de las variables de estado no solo en el tiempo, sino también en el espacio. En ambos casos, el modelo resultante es un sistema de ecuaciones diferencial-algebraico que representa las ecuaciones de continuidad y balances de momento y energía.

Correa y Marchetti (1987) desarrollan un modelo de simulación dinámica multi-celda, que describe el comportamiento de intercambiadores de calor multipaso con deflectores en los estados transitorios. Zinemanas et al. (1984) proponen un algoritmo para la simulación de intercambiadores horizontales o verticales de carcasa y tubos con cambio de fase y uno o más componentes.

En este trabajo se formulan modelos dinámicos para intercambiadores de calor de carcasa y tubo, contracorriente, con y sin cambio de fase con un enfoque orientado a ecuaciones. En el modelo se formularon los balances dinámicos de energía y masa, correlaciones hidráulicas y predicciones termodinámicas rigurosas mediante la ecuación de estado SRK para el cálculo del equilibrio. Los balances de energía en los intercambiadores de calor nos generan un problema de parámetros distribuidos el cual es transformado en un sistema de ecuaciones diferenciales ordinarias mediante la discretización espacial aplicando el método de líneas. Este problema de optimización dinámica, el cual incluye ecuaciones tanto diferenciales como algebraicas (DAE), ha sido resuelto mediante el enfoque simultáneo. Donde este enfoque, propone discretizar el sistema DAE mediante puntos de colocación sobre elementos finitos resultando en un problema no lineal a gran escala. La solución del problema provee los perfiles de variables de control y estado en espacio y tiempo, así como también la condensación parcial en el intercambiador de calor.

2. MODELADO MATEMÁTICO

2.1. INTERCAMBIADOR DE CALOR CRIOGÉNICO SIN CAMBIO DE FASE

Se han formulado los balances de energía en elementos diferenciales de volumen del lado de la carcasa y de los tubos. Estos balances dan lugar a un sistema de ecuaciones diferenciales a derivadas parciales (PDE) hiperbólico de primer orden.

Balance en los tubos:

$$\frac{\partial T_t(z,t)}{\partial t} + v_t \frac{\partial T_t(z,t)}{\partial z} = \frac{h_t * A_{sup}}{\rho_t * Cp_t * A_t * L}\left(T_s(z,t) - T_t(z,t)\right)$$

Balance en la carcaza:

$$\frac{\partial T_s(z,t)}{\partial t} - v_s \frac{\partial T_s(z,t)}{\partial z} = \frac{h_s * A_{sup}}{\rho_s * Cp_s * A_s * L}\left(T_t(z,t) - T_s(z,t)\right)$$

El PDE se ha transfprmado en un sistema ODE mediante la aplicación del Método de Líneas. El modelo incluye las ecuaciones algebraicas que incorporan la dependencia de la densidad, el factor de compresibilidad y velocidad, respecto de la temperatura para cada celda del intercambiador.

$$\rho_{j,i} = \frac{MW * P}{z_{j,i} * R * T_{j,i}} \quad ; \quad v_{j,i} = \frac{F_j}{\rho_{j,i} * A_j} \quad ; \quad z_{j,i} = a_j + b_j * T_{j,i}$$

i = 1,…, número de celdas; j = tubo (t), carcasa (s)

2.2 INTERCAMBIADOR DE CALOR CRIOGÉNICO CON CAMBIO DE FASE

El modelo de un intercambiador con cambio de fase consta de dos partes: un submodelo de flujo bifásico y otro de flujo monofásico. Se formulan balances de masa para la fase líquida y la fase vapor, balances de energía y de cantidad de movimiento. A continuación, se presentan los balances ya discretizados espacialmente:

Balance de masa fase vapor

$$\frac{dM_{V,i}}{dt} = V_{i-1} - V_i + m_{LV,i}$$

donde $m_{LV,i}$ corresponde al transporte de masa interfacial del líquido a la fase vapor en la celda i.

Balance de masa fase líquida

$$\frac{dM_{L,i}}{dt} = L_{i-1} - L_i - m_{LV,i}$$

Como ambas fases se encuentran en equilibrio termodinámico en cada instante de tiempo, se requiere sólo un balance de energía para el sistema líquido-vapor, en el que E se refiere a la energía interna.

$$\frac{dE_i}{dt} = L_{i-1} * h_{i-1} + V_{i-1} * H_{i-1} - L_i * h_i - V_i * H_i + Q_i^t$$

El perfil de presión a lo largo del intercambiador de calor ha sido calculado en cada celda asumiendo estado estacionario mediante el método de Bell-Delaware.

Las siguientes ecuaciones algebraicas completan el modelo y el cálculo del equilibrio líquido-vapor.

Energía interna:

$$E_i = M_{V,i} H_i + M_{L,i} h_i$$

Relación de equilibrio para el componente j ($K_{i,j}$):

$$K_{i,j} = \frac{\phi_{i,j}^L}{\phi_{i,j}^V}$$

$$y_{i,j} = K_{i,j} x_{i,j}$$

Ecuaciones de suma de fracciones molares:

$$\sum_j y_{i,j} - \sum_j x_{i,j} = 0$$

Entalpía de vapor (H_i) y de líquido (h_i):

$$H_i = H_i^{ideal} - \Delta H_i$$

$$H_i^{ideal} = \sum_{j=1}^{nc} H_{i,j}^{ideal}(T_i) y_{i,j}$$

$$h_i = h_i^{ideal} - \Delta h_i$$

$$h_i^{ideal} = \sum_{j=1}^{nc} h_{i,j}^{ideal}(T_i) x_{i,j}$$

donde ΔH_i y Δh_i son entalpías residuales, que se calculan con la ecuación de estado SRK (Soave, 1970).

Para evitar el cruce de temperatura, se incluyeron restricciones adicionales en el problema de optimización. El planteo del modelo dentro de un enfoque simultáneo de simulación y optimización dinámica permite el manejo directo de estas restricciones de camino a lo largo del horizonte de tiempo.

3. CASO DE ESTUDIO

Como se observa en la Figura 1, el fluido frío circula por los tubos de los intercambiadores e ingresa al primer intercambiador criogénico HE1, donde intercambia calor con el gas de alimentación que condensa parcialmente. En el segundo intercambiador HE2, el gas natural circulante por la carcasa, se enfría en contracorriente con el gas residual que sale del HE1. La válvula a la salida del primer intercambiador de calor permite re direccionar parte del gas residual sin intercambiar calor en HE2. De esta forma se controla la temperatura de salida del gas natural.

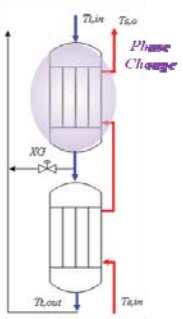


Figura 1. Configuración de los intercambiadores de calor

4. OPTIMIZACIÓN DINÁMICA DEL SISTEMA DE INTERCAMBIADORES CRIOGÉNICOS

El objetivo es minimizar el transitorio para lograr un cambio de condición operativa de la temperatura de salida del gas parcialmente condensado de los intercambiadores criogénicos (T_S) a un valor deseado (T_{SP}). El problema de optimización dinámica queda formulado de la siguiente manera:

$$min \int_0^{tf} \left(Ts - T_{SP}\right)^2 dt$$

$$st.$$

$$\{Modelo\ DAE\ Intercambiadores\}$$

$$0 \le x_{bypass} \le 1.$$

$$Ts_k - Tt_{k+1} \ge \Delta T_{min}; k = 1,...Ncells1 - 1$$

$$Ts_k - Tt_{IN1} \ge \Delta T_{min}; k = Ncells1$$

$$Ts_j - Tt_{j+1} \ge \Delta T_{min}; j = 1,...Ncells2 - 1$$

$$Ts_j - Tt_{IN} \ge \Delta T_{min}; j = Ncells2$$

$$Ts_{m,IN} \ge Ts_{m,1}; m = 1,2$$

$$Ts_{m,i} \ge Tsm_{,i+1}; i = 1,...,Ncellsm - 1; m = 1,2$$

$$Tt_{m,i} \ge Tt_{m,i+1}; i = 1,...Ncellsm - 1; m = 1,2$$

$$Tt_{m,Ncellsm} \ge Tt_{m,IN}; m = 1,2$$

$$z(t=0) = z_0; y_L \le y \le y_U; z_L \le z \le z_U$$

donde se imponen restricciones para impedir el cruce de temperaturas en cada celda del intercambiador 1 y 2 (Ncells1 y Ncells2), respectivamente. También se imponen restricciones adicionales para asegurar la monotonicidad en las temperaturas de las celdas. La función objetivo integral se ha tratado como una ecuación diferencial adicional:

$$\frac{dz_1}{dt} = (T_S - T_{SP})^2$$

$$z_1(0) = 0$$

El problema se ha implementado en un ambiente de AMPL, integrado al programa IPOPT (Cervantes et al., 2000; Biegler et al., 2002). Las derivadas parciales primeras y sus expresiones analíticas se incluyen en el código AMPL. Con una discretización del horizonte de tiempo con 8 elementos finitos y dos puntos de colocación, se obtiene un problema NLP con 15630 variables algebraicas. El problema se resolvió en 50 iteraciones, partiendo de un valor inicial del parámetro de barrera de 0.01, con un tiempo de cómputo de 700 s. Los resultados obtenidos se muestran en las siguientes figuras.

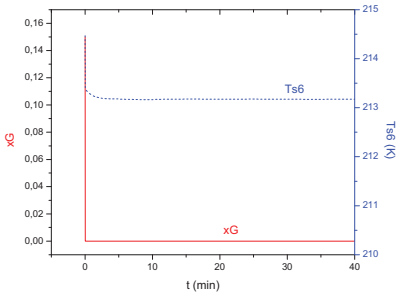


Figura 2. Perfil óptimo de la fracción bypass y temperatura de salida de la carcaza del HE1.

La Figura 2 muestra que se requiere un cierre total instantáneo de la válvula del bypass (de 15% a 0%), produciendo una respuesta de primer orden en la temperatura de salida del segundo intercambiador, alcanzándose el estado estacionario luego de los 35 minutos.

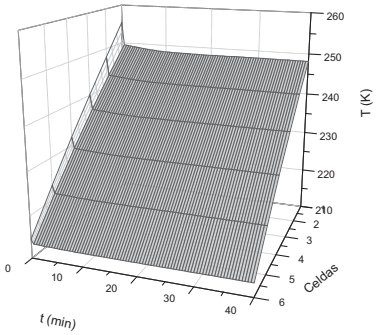


Figura 3. Perfiles temporales y espaciales de temperatura del fluido que condensa en la carcasa

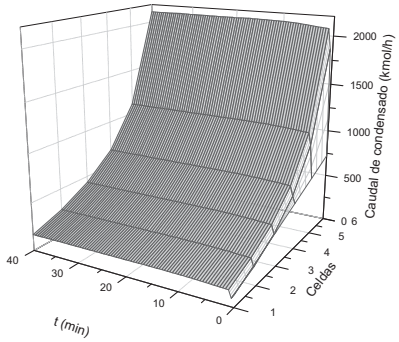


Figura 4. Perfil espacial y temporal del caudal de condensado en HE2.

REFERENCIAS

[1] K. J. BELL-DELAWARE METHOD FOR SHELL SIDE DESIGN, PETRO/CHEM ENG. 1 (1960) PP. 26-40

[2] L. T. BIEGLER, A. CERVANTES, A. WATCHER, *Advances in simultaneous strategies for dynamic process optimization*, Chem. Eng. Sci., 57, 575-593, 2002.

[3] L. T. BIEGLER, A. CERVANTES, *Large-scale DAE optimization using simultaneous nonlinear programming formulations*, AIChE J., 44, 1038-1050, 1998.

[4] D. CORREA, J. MARCHETTI, *Dynamic simulation of shell-and-tube heat exchangers*, Heat Transfer Eng., 8, 50-59, 1987.

[5] A. RAGHUNATHAN, M. S. DÍAZ, L. T. BIEGLER, *An MPEC formulation for dynamic optimization of distillation operations*, Comput. Chem. Eng., 28, 421-434, 1994.

[6] D. ZINEMANAS, D. HASSON, E. KEHAT, *Simulation of heat exchangers with change of phase*, Comput. Chem. Eng., 8, 367-375, 1984.

CADENA DE SUMINISTRO DE BIODIESEL. FORMULACIÓN MILP MULTIPERÍODO

Facundo Iturmendi†, Federico Andersen†, Susana Espinosa‡ y M. Soledad Diaz†

†Planta Piloto de Ingeniería Química (PLAPIQUI), Universidad Nacional del Sur – CONICET, Camino La Carrindanga km. 7, 8000 Bahía Blanca, Argentina, fiturmendi@plapiqui.edu.ar
‡Universidad Nacional del Comahue, Neuquén, Argentina

Resumen: Los combustibles alternativos presentan una opción diferente para afrontar el creciente problema del consumo de energía. La necesidad de utilizar materias primas que no compitan con cultivos que se destinan a la alimentación, ha conducido a la investigación hacia fuentes alternativas de combustible, como es el caso de la jatropha, que se encamina a ser una alternativa prometedora para una cadena de suministro de biodiesel que resulte sustentable. En este sentido, y con la intención de modelar y optimizar el funcionamiento de la cadena de suministro del biodiesel en Argentina, es que se ha propuesto plantear el complejo problema a través de un modelo MILP multiperíodo, considerando soja, girasol y jatropha curcas como materias primas.

Palabras claves: biodiesel*, optimización de cadena de suministro, formulación multiperíodo MILP*

1. INTRODUCCIÓN

El consumo de energía ha tenido un crecimiento continuo desde 1990. Este comportamiento conduce a problemas graves como el aumento de las emisiones de los gases de efecto invernadero, el agotamiento de los recursos naturales y fuentes de energía no renovables, el aumento de los precios de derivados del petróleo, la acumulación de productos no biodegradables, entre otros. Estos problemas fomentan la investigación sobre fuentes de energía alternativas. La investigación de la cadena de suministro del biodiesel en Argentina, su diseño y la planificación de la misma está respaldada por numerosas razones. Una de ellas surge de la ley 26.093/08, que establece la obligatoriedad para la comercialización del diesel con un corte, como mínimo, del 5 % de biodiesel. La demanda doméstica asociada es de alrededor de 800.000 [tn] y la demanda internacional de biodiesel es del orden de 2.000.000 [tn]. Para satisfacer esta demanda, para incentivar el agregado de valor a los productos nacionales y con el objetivo de que la cadena de suministro sea lo suficientemente grande como para abastecer otros mercados potenciales, se han reducido las tasas de retenciones a la exportación de biodiesel en comparación con las tasas de retención sobre el aceite. De esta forma resultaría razonable que sea más rentable exportar biodiesel que aceite. Por otra parte, la necesidad de considerar materias primas alternativas tiene como objetivo una producción sostenible, evitando la utilización de materias primas que presenten una competencia con la alimentación.
Estas características llevan a diseñar y planificar una eficiente cadena de suministro teniendo en cuenta los recursos disponibles y potenciales expansiones. En este trabajo se formula un problema lineal mixto entero multiperíodo (MILP) para representar la cadena de suministro desde la producción de materia prime hasta la distribución en las destilerías.

2. FORMULACIÓN DEL PROBLEMA

Para satisfacer la demanda nacional e internacional de biodiesel se ha desarrollado un modelo optimizando la utilización de los recursos existentes y definiendo la conveniencia de inversión en nuevos mercados. En el modelo se han utilizado tres cultivos (soja, girasol y jatropha) y se han desarrollado las cadenas de suministros desde sus semillas hasta la producción de biodiesel con sus respectivos subproductos. En las figuras 1 y 2 se muestra la cadena de suministro modelada y la superestructura de tecnologías respectivamente.

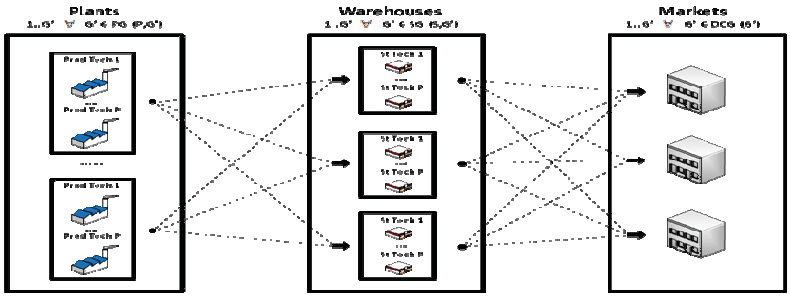


Figura 1: Cadena de suministro de 3 etapas

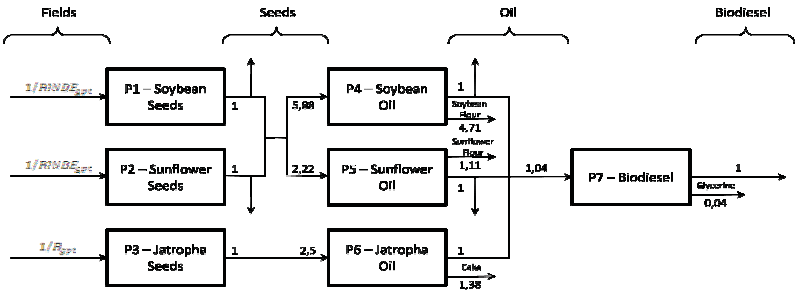


Figura 2: Topología de la red de tecnologías

El problema MILP multiperíodo resultante se basa en una formulación state task network (STN), que se representa, en forma general como:

$$min \quad Z = a^T y + b^T x \qquad \text{Función Objetivo}$$

$$s.t. \quad Ay + Bx \leq d \qquad \text{Restricciones}$$

$$y \in \{0,1\}^m, x \geq 0$$

2.1. BALANCES DE MASA

$$\sum_{s \in SS(i,s)} SW_{igstf-1} + PT_{igtf} + PU_{igt} + \sum_{g'} Q_{ig'gtf} = \sum_{s \in SS(i,s)} SW_{igstf} + DP_{igtf} + \sum_{g'} Q_{igg'tf} \qquad \forall\, i,g,tf \qquad \textbf{(1)}$$

La ecuación 1 representa el balance de masa para cada región g, cada producto i y cada período de tiempo tf. Los términos involucrados representan la producción (PT_{igtf}), almacenamiento (SW_{igstf}), transporte interzonal ($Q_{ig'gtf}$), compras (PU_{igt}) y ventas (DP_{igtf}).

$$PR_{igptf} = RHO_{ip} \sum_{i' \in MP(i',p)} PR_{i'gptf} \qquad \forall\, i,g,p,tf \qquad \textbf{(2)}$$

La ecuación 2 representa el balance de masa en las plantas conectando entradas y salidas afectadas por el rendimiento adecuado.

2.2. VARIABLES BINARIAS

CAPACIDAD DE PRODUCCIÓN

$$CP_{gpt} = CP_{gpt-1} + \sum_{z} CEP_{gptz} \qquad \forall\, g,p,t \tag{3}$$

$$LBQP_{gpz}\, NP_{gptz} \leq CEP_{gptz} \leq UBQP_{gpz}\, NP_{gptz} \qquad \forall\, g,p,t,z \tag{4}$$

La capacidad de producción se definió teniendo en cuenta la capacidad existente y las posibles expansiones. Además se limitó el tamaño de la expansión brindando la posibilidad que la misma pueda ser de un tamaño pequeño, mediano o grande.

La variable NP_{gptz} es una variable entera que indica cuántas plantas de tecnología de producción *p* deben instalarse en la región *g*, en el año *t*. El subíndice *z* considera la economía de escala que pasa a jugar un rol importante cuando la inversión de capital es grande.

Estas mismas ecuaciones se aplicaron para las instalaciones de almacenaje, generando otra variable binaria análoga a NP_{gptz}, que indica la cantidad de almacenes que deben instalarse.

TRANSPORTE

Se ha asociado una variable binaria al transporte entre diferentes regiones, es decir a las posibles conexiones entre nodos adyacentes. Teniendo en cuenta que el problema consta de 129 nodos, la cantidad de conexiones factibles es de 16731 por período de tiempo considerado. Debido a la comprensión del problema se pudo reducir esta cantidad de variables binarias a sólo 1649 por período de tiempo, eliminando las conexiones inexistentes y considerando sólo las realmente factibles.

Por otro lado, para disminuir aún más la cantidad de variables binarias, se ha modelado el envío entre zonas como parte del cumplimiento de contratos anuales. Esto implica que si el contrato es aceptado ($X_{igg't} = 1$), el transporte del producto *i* entre las zonas *g* y *g'* puede ocurrir en cualquiera de los períodos que conforman el año *t*. De esta forma las variables binarias se extienden sobre el dominio del set de tiempo en años y no en períodos.

Esta característica ha permitido reducir la cantidad de variables binarias de 1649 $\left[\frac{\#\, Vars\ Bin}{período}\right]$ a 1649 $\left[\frac{\#\, Vars\ Bin}{año}\right]$. Esta particularidad redunda en una reducción a la mitad de las variables binarias originales cuando el largo del período es de 6 meses, a una tercera parte de la cantidad de variables binarias original cuando el largo del período es de 3 meses, y asi siguiendo.

O sea que, debido al entendimiento del problema se ha logrado reducir la cantidad de variables binarias de 351351 (tomando períodos de 4 meses) a 11543. Esta reducción lograda significa una cantidad de variables binarias de, aproximadamente, el 4 % de la cantidad original de variables binarias.

Las ecuaciones 5 y 6 indican los límites de producto *i* que pueden transportarse entre 2 zonas en un año determinado. El límite inferior está dado por la capacidad de un camión, en tanto que el límite superior fue calculado con una cantidad promedio de camiones por día que entran a la zona en cuestión.

La ecuación 7 es necesaria para brindar información lógica al modelo, evitando los envíos innecesarios (si se envía producto *i* de la región *g* a *g'*, entonces que no se envíe de *g'* a *g*).

$$\sum_{tf \in TPS(tf)} Q_{igg'tf} \leq QUP_{igg't}\, X_{igg't} \qquad \forall\, \{i,g,g'\} \in IGG(i,g,g')\,,\ t \tag{5}$$

$$\sum_{tf \in TPS(tf)} Q_{igg'tf} \geq QLO_{igg't}\, X_{igg't} \qquad \forall\, \{i,g,g'\} \in IGG(i,g,g')\,,\ t \tag{6}$$

$$X_{igg't} + X_{ig'gt} \leq 1 \qquad \forall\, \{i,g,g'\} \in \{IGG(i,g,g')\ AND\ IGG(i,g',g)\}\,,\ t \tag{7}$$

2.3. FUNCIÓN OBJETIVO

Se utilizó la formulación de valor presente neto como función objetivo. En la ecuación 8 se distinguen los términos de flujo anual de efectivo, valor de mercado al final del período y deudas de capital por cuotas de préstamos no devengadas.

$$NPV = \sum_{t} \frac{CF_t}{(1+IR)^{t-1}} + \frac{MV}{(1+IR)^{t_F-1}} - \frac{DC_{t_F}}{(1+IR)^{t_F-1}} \quad \textbf{(8)}$$

3. RESULTADOS

Se ha resuelto el modelo utilizando Jatropha como materia prima y sin utilizarla a fin de comparar los resultados de la utilización de materias primas alternativas. Cómo puede advertirse en las figuras 3 y 4, el biodiesel producido a partir de soja resulta notablemente menor cuando se utiliza Jatropha que cuando no se utiliza. Este hecho permite que haya más disponibilidad de aceite de soja en el mercado oleaginoso. Ya que se ha logrado disminuir su demanda para producción de biocombustible. Se ha utilizado el solver CPLEX del software GAMS con un procesador Intel Core 2 Duo P8600 2.40 GHz y una memoria RAM de 3 Ghz, para obtener el rendimiento detallado abajo. La cantidad de ecuaciones y variables detalladas, corresponden a un horizonte de tiempo de 7 años con períodos de 6 meses.

Computational Details	
NPV [MM USD]	35269.1
Single equations	51883
Single variables	53241
Discrete variables	13566
Relative Gap	0.0006
CPU Time	0’ 26.118’’

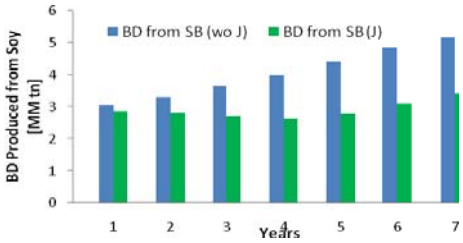


Figura 3: Cantidad de Biodiesel producido a partir de soja (considerando y no considerando Jatropha como materia prima)

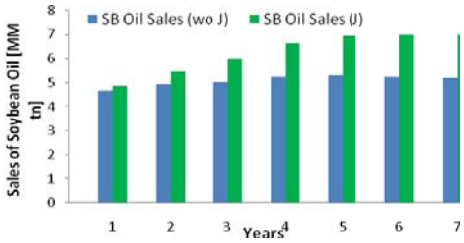


Figura 4: Cantidad de aceite de soja vendido (considerando y no considerando Jatropha como materia prima)

REFERENCIAS

[1] F. ANDERSEN, F. ITURMENDI, S. ESPINOSA AND M. S. DIAZ, *Optimal Planning of Biodiesel Supply Chain in Argentina with Alternative Oil Sources*, Aiche Annual Meeting, 7-12 Noviembre de 2010, Salt Lake City, USA, 2010.

[2] G. GUILLÉN-GOSÁLBEZ, AND I. E. GROSSMANN, *Optimal design and planning of sustainable chemical supply chains under uncertainty*, Aiche J., 55, 99-121, 2009.

[3] G. GUILLÉN-GOSÁLBEZ, F. D. MELE, AND I. E. GROSSMANN, *A Bi-Criterion Optimization approach for the design and planning of hydrogen supply chains for vehicle use*, Aiche J., 56, 650-667, 2009.

[4] C. MOLINA, *Biocombustibles: una oportunidad para el agro, una oportunidad para Argentina*, OSDE Foundation Dissertation, 2007.

[5] K. OPENSHAW, *A review of Jatropha curcas: an oil plant of unfullled promise. Biomass and Bioenergy 19*, 2000.

[6] E. P. SCHULZ, M. S. DIAZ, AND J. A. BANDONI, *Supply chain optimization of large scale continuous processes,* Computers & Chemical Engineering, 29, 1305–1316, 2005.

DESARROLLO DE UN MODELO MATEMÁTICO DISCRETO/CONTINUO PARA EL DISEÑO DE COLUMNAS DE DESTILACIÓN

Juan I. Manassaldi†, Nicolás Scenna†‡ y Sergio F. Mussati†‡

†CAIMI, Centro de Aplicaciones Informáticas al Modelado en Ingeniería, UTN Facultad Regional Rosario, Zeballos 1341, S2000BQA, Rosario, Argentina, juanmanassaldi@yahoo.com.ar, nscenna@santafe-conicet.gov.ar
‡INGAR /CONICET Instituto de Desarrollo y Diseño, Avellaneda 3657 C.P. 3000 Santa Fe, Argentina, mussati@santafe-conicet.gov.ar

Resumen: Este trabajo aborda el diseño óptimo de columnas de destilación utilizando una potente herramienta como es la programación matemática. Se propone un modelo matemático discreto/continuo de optimización basado en una superestructura que embebe todas las configuraciones atractivas y factibles tecnológicamente. El modelo propuesto permite determinar el número óptimo de etapas de equilibrio, como así también la ubicación de la alimentación y las condiciones óptimas de operación para alcanzar la separación deseada a un mínimo costo total anual. El modelo se implementó en GAMS y los resultados obtenidos fueron satisfactoriamente validados con aquéllos arrojados por simuladores comerciales y específicos de procesos químicos e industriales, como por ejemplo HYSYS. Un caso de aplicación es presentado para mostrar la capacidad del modelo y discutir los resultados.

Palabras claves: *destilación, MINLP, optimización*
2000 AMS Subjects Classification: 90C10 – 90C30

1. INTRODUCCIÓN

El estudio de la síntesis y el diseño óptimo de columnas de destilación han sido siempre de gran interés en la industria petroquímica debido a la alta inversión necesaria y costos de operación involucrados por el consumo de servicios auxiliares de calefacción y de enfriamiento. El creciente uso de las herramientas informáticas y la aplicación de técnicas de programación matemática han permitido desarrollar modelos matemáticos no sólo para reproducir procesos complejos como el de destilación sino optimizar la configuración y determinar las condiciones de operación óptimas que minimicen costos y/o maximicen eficiencias.

Sargent y Gaminibandara[1], pioneros en el tema, basaron su trabajo en la búsqueda de la ubicación optima del plato de alimentación para una determinada cantidad de etapas fijas. Como se observa en la Figura 1(a), los autores propusieron dividir la corriente de alimentación según la cantidad de platos candidatos, y luego mediante restricciones apropiadas seleccionar un único plato de alimentación óptimo.

Viwanathan y Grossmann[2] ampliaron el problema considerando la cantidad de platos además de la ubicación de la alimentación y plantearon la superestructura ilustrada en la Figura 1(b). Los autores formularon un problema del tipo mixto entero no lineal (MINLP) el cual variables binarias seleccionaban la ubicación del reflujo y de la entrada del vapor proveniente del rehervidor. En trabajos posteriores[4][5] los autores reconocen dificultades numéricas para la resolución mediante esta estrategia.

Años más tarde, Yeomans y Grossmann[3] aplicaron técnicas de programación disyuntiva (GDP) y propusieron otro tipo de superestructura que se indica en la Figura 1(c). Básicamente introdujeron el concepto de plato activo o inactivo, donde el primero cumple todas las ecuaciones MESH (Mass, Equilibrium, Sum and Heat) mientras que el segundo únicamente es un proceso de entrada-salida sin transferencia de materia.

El objetivo de este trabajo es presentar una nueva superestructura que combine las ventajas de aquéllas propuestas por Sargent y Gaminibandara (Fig. 1(a)) y Yeomans y Grossmann (Fig. 1(c)). La Figura 1(d) muestra la superestructura propuesta, la cual es representada por un modelo del tipo mixto entero no lineal capaz de optimizar simultáneamente la cantidad de platos, la ubicación de la alimentación (decisiones discretas), las condiciones de operación y perfiles de composición (decisiones continuas).

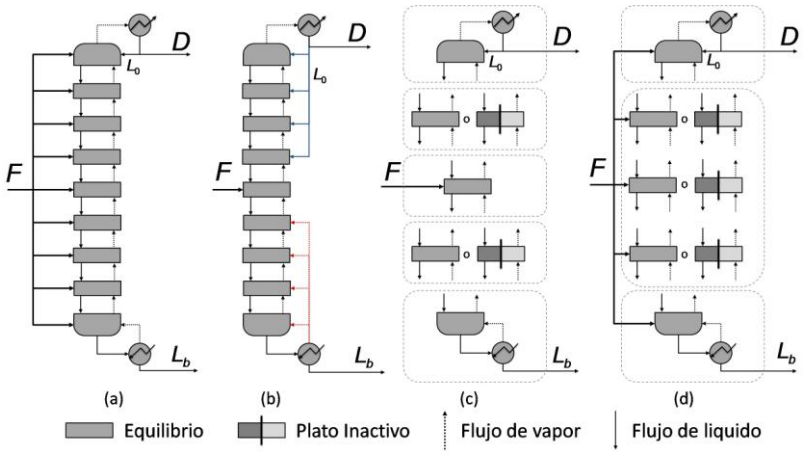


Figura 1: Superestructuras según diferentes autores

2. MODELO MATEMÁTICO

El modelado matemático fue desarrollado por platos y en forma compacta para lo cual fue necesario definir un índice (*p*). Básicamente cada plato que forme parte de la columna de destilación debe cumplir las mencionadas ecuaciones MESH. Estas no son más que simples balances de materia y energía en donde la única complejidad obedece al cálculo de las condiciones de equilibrio ya que las restricciones asociadas a dichos cálculos son, en su mayoría, altamente no lineales. A diferencia del enfoque tradicional de modelo plato a plato, donde se asume fija la configuración (numero de platos y ubicación de la alimentación), en este trabajo se presenta una modificación que nos permite trabajar con la estructura variable y así seleccionar los platos óptimos "reales" y eliminar aquéllos innecesarios. Precisamente para lograr esto último (selección y eliminación) utilizamos las variables binarias s_p y f_p que determinan la existencia del plato y la ubicación de la alimentación respectivamente. Por ejemplo, cuando s_2 toma el valor de 1 indica que el plato 2 existe y las corrientes que lo abandonan deben estar en equilibrio. Por el contrario, si s_2=0 la etapa 2 no existe y tanto el liquido como el vapor la atraviesan sin modificar sus propiedades (composición, temperatura y caudal). Observamos en la Figura 2 las corrientes que ingresan y egresan de cada plato, donde F'_p representa el posible ingreso de la alimentación.

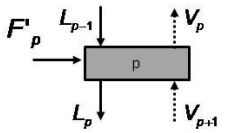


Figura 2: Esquema de las corrientes que ingresan y egresan de cada plato

Mediante el siguiente grupo de restricciones logramos que, a partir de la variable binaria s_p, se cumplan o no las ecuaciones de equilibrio al abandonar el plato.

$$y_p^i \le K_p^i x_p^i + (1 - s_p)M \quad ; \quad y_p^i \ge K_p^i x_p^i - (1 - s_p)M \qquad (1)\ ;\ (2)$$

Debemos asegurar que si un plato no existe, las corrientes deben atravesarlo sin mezclarse. Precisamente, las siguientes restricciones son impuestas a la corriente líquida para evitar el mezclado.

$$L_p \le L_{p-1} + s_p M \quad ; \quad L_p \ge L_{p-1} - s_p M \qquad (3)\ ;\ (4)$$

Así, si para un plato s_p = 1, (1) y (2) obligan a satisfacer la siguiente igualdad: $y_p^i = K_p^i x_p^i$; mientras que las restricciones (3) y (4) permite que el caudal de líquido a la entrada y salida del plato sean distintos y por ende se produzca la separación. En todas, *M* representa un parámetro de valor numérico alto. Del mismo modo puede procederse con las demás variables (*V, x, y, H y h*).

Como se trata de una columna de simple entrada se debe asegurar la elección de un único plato de alimentación, lo que se logra imponiendo las siguientes restricciones:

$$\sum_{p=1}^{b} f_p = 1 \qquad (5)$$

$$F'_p - f_p F \le 0 \qquad (6)$$

También debe asegurarse que un plato eliminado no pueda ser elegido como de alimentación. Esta consideración se logra relacionando las variables binarias f_p con s_p de la siguiente manera:

$$f_p \leq s_p \tag{7}$$

Por último, la siguiente restricción asegura la no repetitividad de soluciones imponiendo que la eliminación de etapas (en caso de ser necesario) comience a partir de la segunda y así sucesivamente, recordando que la primera y la última etapa estarán siempre presentes.

$$s_p \leq s_{p+1} \; ; \; 1 < p < \; b \tag{8}$$

3. IMPLEMENTACIÓN Y ESTRATEGIA DE RESOLUCIÓN DEL MODELO

El modelo considera un total de 2368 ecuaciones y 1846 variables y fue implementado en GAMS (General Algebraic Modelling System). El algoritmo SBB (Simple Branch and Bound) fue utilizado como revolvedor MINLP, y es uno de los tantos algoritmos disponibles en GAMS.

Se implementó una estrategia de resolución que garantizó siempre la convergencia del modelo. Ciertamente el procedimiento consistió, en primer lugar, en resolver el modelo MINLP pero considerando la presencia de todos los platos y asumiendo el lugar de la alimentación (configuración fija) para lo cual se fijaron en 1 los valores de las variables binarias s_p asociadas a los platos y del mismo modo se fijó la ubicación de la alimentación. Luego, dichos valores fueron liberados, pudiendo adoptar ahora valores 0 ó 1, y nuevamente el modelo MINLP se resuelve adoptando como inicialización la solución obtenida previamente (configuración fija). Así, la última solución obtenida por el procedimiento indica la cantidad de platos óptimos, ubicación de la alimentación y perfiles de composición dentro de la columna. En resumen, la solución obtenida a estructura fija resulta sumamente valiosa para inicializar el modelo a estructura variable y eliminar los platos en forma adecuada. Este procedimiento se implementó en forma sistemática.

4. DISCUSIÓN DE RESULTADOS

El modelo fue resuelto para distintos casos de estudio y por cuestiones de limitación de espacio solo se presenta un ejemplo. Para una corriente de proceso con caudal molar de 100 Kmol/hr y de composición equi-molar de n-heptano y n-octano se desea obtener una corriente con pureza superior al 99% (base molar). La naturaleza de la mezcla permite asumir, como principal hipótesis, comportamiento ideal en fase liquida y vapor. El problema de optimización consiste en determinar las condiciones de operación y el dimensionamiento de la columna para alcanzar las especificaciones de pureza a un mínimo costo total.
Grossman[4] sugiere una expresión simplificada del costo total anual que solamente tiene en cuenta el calor intercambiado y el numero de platos de la columna:

$$min: \quad Costo\,[usd/año] = Q_{reb} + 0.2Q_{cond} + 80N \tag{9}$$

La Tabla 1 y Figura 3 reportan los valores óptimos de las variables más importantes.

Q_{cond}	N	d_c	Q_{reb}	R	*alimentación*
1618 kW	20 platos	1.8 m	1643 kW	2.43	plato Nº 10

Tabla 1: Resultados

Como se puede apreciar la Figura 3(a) muestra los perfiles óptimos de composición en el líquido a lo largo de la columna.

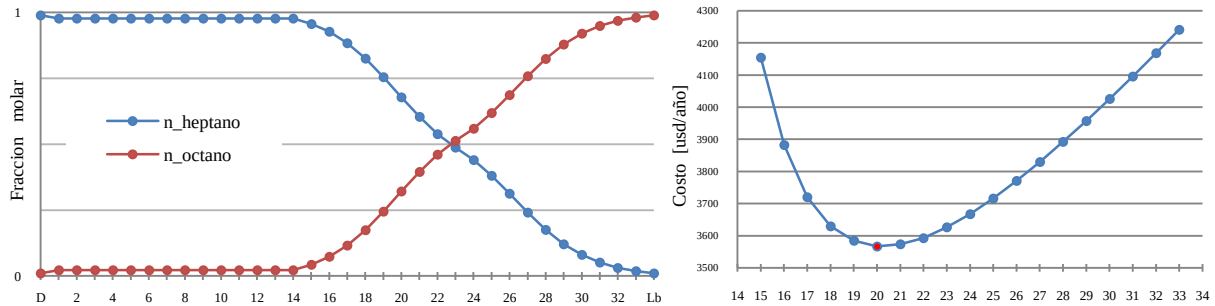


Figura 3: (a) perfiles de composición; (b) resolución paramétrica

Se observa claramente que para los platos eliminados, a partir del 2 hasta el 14, los perfiles dentro de la columna son constantes indicando que en dichos platos no existe separación alguna. El numero óptimo de platos resultó 20.

A modo de comparación y validación de resultados, el mismo problema fue resuelto con el mismo modelo pero ahora en forma paramétrica en el numero de platos, fijando los valores de las variables binarias según corresponda. La Figura 3(b) muestra la variación del costo total con la cantidad total de platos en donde se observa claramente que para 20 etapas la función objetivo alcanza el valor mínimo. Este resultado permite comprobar que el modelo MINLP propuesto (estructura variable) selecciona correctamente las etapas necesarias, eliminando el resto.

5. CONCLUSIONES

En este trabajo se presentó un modelo matemático del tipo MINLP que permite optimizar en forma simultánea el número total de platos, posición de la alimentación y condiciones de operación de una columna destilación. También se propuso un esquema sistemático y efectivo de inicialización de las variables de manera tal de asegurar la convergencia del modelo y obtener la mejor solución óptima. No obstante, es importante destacar que debido a la presencia de restricciones altamente no-lineales y no-convexas no se puede garantizar el hallazgo de soluciones óptimas globales.

Como tarea de investigación para próximos trabajos se propone extender el modelo desarrollado para abordar el diseño de destilación azeotrópica. También, se pretende extender el modelo al caso de trenes de columnas de separación para tratar mezclas compuesta por más de tres componentes, siendo el principal objetivo determinar la secuencia óptima de separación.

Seguramente, los modelos a desarrollar aumentarán considerablemente de tamaño y sus complejidades serán mayores no sólo desde el punto de vista de la formulación matemática sino de su resolución.

NOMENCLATURA

Variables

x /y	Fracción molar en el liquido/vapor	F'	Posible alimentación al plato **[kgmol/hr]**
L/V	Flujo molar de la fase liquida/vapor **[kgmol/hr]**	R	Relación de reflujo
D	Flujo molar de destilado **[kgmol/hr]**	N	Número total de platos
F	Flujo molar de alimentación **[kgmol/hr]**	d_c	Diámetro de la columna **[m]**
K	Constante de equilibrio	$Q_{reb/cond}$	Potencia en el rehervido/condensador **[kW]**

Variables binarias

s	Existencia del plato
f	Plato seleccionado como de alimentación

Subindices

p	Plato genérico	d	destilado
b	Ultimo plato (rehervidor)	0	reflujo

Superindices

i	compuesto de la mezcla

AGRADECIMIENTOS

Los autores del trabajo agradecen a CONICET (Consejo Nacional de Investigaciones Científicas y Tecnológicas), ANCyPT (Agencia Nacional de Promoción Científica y Tecnológica) y a la Universidad Tecnológica Nacional, Facultad Regional Rosario por el apoyo financiero recibido.

REFERENCIAS

[1] SARGENT, R. W. H. & GAMINIBANDARA, K., *Optimal design of plate distillation columns*, L. C. W. Dixon (Ed.), Optimization in action New York: Academic Press, 1976.

[2] VISWANATHAN, J. & GROSSMANN, I. E., *An alternative MINLP model for finding the number of trays required for a specified separation objective*, Computer and Chemical Engineering, 1993.

[3] YEOMANS, H. & GROSSMANN, I. E., *Disjunctive programming models for the optimal design of distillation columns and separation sequences*, Industrial and Engineering Chemical Research, 2000.

[4] CABALLERO, J. A. & GROSSMANN, I. E., *Hybrid Simulation-Optimization Algorithms for Distillation Design*, ESCAPE-20, 2010.

[5] GROSSMANN, I. E., AGUIRRE, P. A. & BARTTFELD, M., *Optimal synthesis of complex distillation columns using rigorous models*, Computers and Chemical Engineering, 2005.

UN MÉTODO DE CALIBRACIÓN PARA EL FLUJO TRANSITORIO EN CANALES EMPLEANDO UNA TÉCNICA ESTOCÁSTICA DE OPTIMIZACIÓN GLOBAL

Julia V. Martorana†, Víctor H. Cortínez‡

†Centro de Investigaciones en Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional (FRBB), 11 de Abril 461, 8000, Bahía Blanca, vcortine@frbb.utn.edu.ar
‡Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Rivadavia 1917, 1033, Capital Federal

Resumen: En el presente artículo se propone una metodología para la estimación del coeficiente de rugosidad en canales, formulada como un problema de optimización en el cual la función objetivo corresponde a la diferencia entre valores de tirantes medidos y calculados. Estos últimos se obtienen mediante la aproximación por diferencias finitas de las ecuaciones que describen la dinámica de un canal de irrigación. Por otra parte a los efectos de resolver el problema de optimización planteado se hace uso del método de recocido simulado. Se muestra que tal enfoque es suficientemente robusto y expeditivo desde el punto de vista computacional.

Palabras claves: *Canales, Calibración, Optimización, Coeficiente de Rugosidad.*

1. INTRODUCCIÓN

La simulación computacional basada en un modelo hidrodinámico adecuado, se ha convertido en una técnica de gran ayuda para establecer la estrategia de operaciones de compuertas de un canal de riego. De esta manera se evitan importantes pérdidas de agua y sus efectos colaterales (salinización de suelos, rotura de canales, falta de agua en alguna toma y exceso en otras, etc.).

La dinámica de un canal de irrigación puede ser descripta matemáticamente mediante las ecuaciones de Saint Venant. Tal modelo se expresa mediante dos ecuaciones diferenciales parciales que corresponden a la ecuación de continuidad (conservación de la masa) y la ecuación de movimiento (segunda ley de Newton) para las aguas del canal. Para poder resolver tales ecuaciones debe hacerse uso de métodos numéricos. Entre los métodos más utilizados pueden mencionarse el método de diferencias finitas, en particular, el esquema de Preissmann (1960) ([1] y [2]) y el método de los elementos finitos. [3]

El éxito de la simulación computacional de la hidráulica de canales, será dependiente de cuan bien la realidad sea representada por el modelo matemático. Esto dependerá en general de los valores que adopten varios parámetros del modelo que no siempre son conocidos de antemano con suficiente precisión. Ejemplos de tales parámetros son los coeficientes de rugosidad y coeficientes de infiltración. Entonces antes de proceder al diseño computacional de manejo del canal es necesario obtener tales parámetros de manera tal que se reproduzcan mediante el modelo computacional los valores medidos. Tal proceso se denomina calibración-verificación. Varias investigaciones sobre este tópico se han realizado en los últimos quince años. [4] [7]

En este trabajo se propone una metodología de calibración de las ecuaciones de Saint Venant aplicadas a canales de riego, basada en la evaluación de la diferencia en un sentido de mínimos cuadrados entre valores de la profundidad en diferentes puntos del canal medidos en forma directa contra valores determinados computacionalmente. Tal diferencia dependerá de los parámetros a determinar (coeficientes de rugosidad, coeficientes de infiltración, etc.). El problema de optimización así formulado será resuelto mediante una combinación de un método de simulación basado en el método de diferencias finitas en combinación con una técnica estocástica de optimización global denominada "Método de recocido simulado". Tal metodología se implementó computacionalmente en el ambiente de programación MATLAB. Se muestra la eficiencia de tal metodología para diferentes casos. Se discuten diferentes aspectos relacionados con la aplicación práctica de tal técnica de calibración tales como la cantidad de estaciones de medición necesarias, paso temporal de medición, etc.

2. ECUACIONES GOBERNANTES

La dinámica de un canal de irrigación puede describirse matemáticamente mediante el sistema de ecuaciones para flujo unidimensional no permanente de aguas superficiales (ecuaciones de Saint Venant). Este sistema esta formado por dos ecuaciones diferenciales parciales. Estas ecuaciones son la ecuación de continuidad (conservación de la masa):

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \tag{1}$$

y la ecuación de momento (segunda ley de Newton)

$$\frac{\partial Q}{\partial t} + \frac{\partial (Q^2/A)}{\partial x} + gA\frac{\partial y}{\partial x} + gAS_f - gAS_0 = 0 \tag{2}$$

donde $A(x,t)$ es la sección del canal (m^2); $Q(x,t)$ es el caudal (m^3/s); $y(x,t)$ es la profundidad del flujo (m); S_f es la pendiente de fricción; S_0 es la pendiente de fondo del canal; x es la distancia a lo largo del canal (m); y t es el tiempo (s). La pendiente de fricción, S_f esta dada por la ecuación de Manning:

$$S_f = \frac{n^2 Q^2 P^{4/3}}{A^{10/3}} \tag{3}$$

donde n es el coeficiente de rugosidad de Manning y P es el perímetro mojado de la sección (m). Las condiciones de borde consisten en la entrada de un determinado caudal en el extremo aguas arriba del canal y un vertedero aguas abajo del mismo:

$$Q(0,t) = Q_{up}(t); \qquad Q(L,t) = Q_v(y) \tag{4}$$

Las condiciones iniciales iniciales están dadas por:

$$y(x,0) = y_0(x); \qquad Q(x,0) = Q_0(x) \tag{5}$$

Para la resolución de las ecuaciones (1) y (2) debe hacerse uso de métodos numéricos. En este caso se utilizó el método de diferencias finitas, en particular el esquema de Preissmann (1960). [1] [2]

3. FORMULACIÓN DEL PROBLEMA

El problema inverso se formula como una optimización no lineal del modelo con el coeficiente de rugosidad (n) como variable a determinar. La función objetivo se plantea como el error cuadrático medio entre los valores estimados y los valores medidos de tirante de flujo. Esto es:

$$n_i = argmin\, F(n_i)$$
$$F(n_i) = \sqrt{\frac{\sum_{i=1}^{a}\sum_{j=1}^{M}\left(\frac{y_s(i,j) - y_m(i,j)}{y_m(i,j)}\right)}{a.M}} \tag{6}$$

donde y_s es el tirante estimado en diferentes puntos de medición; y_m es el tirante observado en los puntos análogos, a es la cantidad de estaciones de medición y M es la cantidad de mediciones temporales.

La dependencia con respecto a n_i viene dada por el hecho de que los valores calculados de y_i así lo hacen de acuerdo a las ecuaciones de Saint Venant. El problema así formulado corresponde a uno de variables continuas. Sin embargo, con el fin de reducir los costos computacionales de los experimentos numéricos se discretizó el conjunto de posibles soluciones generadas por el método de optimización. De esta forma quedó determinado un conjunto acotado (n_{min} = 0,008; $n_{máx}$ = 0,070) y discreto. La cantidad de puntos se eligieron de tal manera que entre dos valores consecutivos la diferencia entre tirantes no fuese apreciable.

4. MÉTODO DE RECOCIDO SIMULADO

El método de recocido simulado es una técnica heurística de optimización combinatoria basada en la generación aleatoria de soluciones factibles cuya principal característica es evitar la convergencia local en problemas de gran escala.

El algoritmo para el problema de optimización comienza fijando un valor inicial aleatorio como solución posible y, a partir de éste, genera nueva soluciones Para cada iteración, el algoritmo evalúa si el valor de la función objetivo correspondiente a ese punto produce un valor menor que el anterior. El punto es aceptado como nueva solución, si minimiza la función objetivo. Si no lo hace, su aceptación queda determinada por una distribución probabilística. [5] [6]

5. RESULTADOS

La evaluación del funcionamiento del modelo de optimización propuesto para la estimación del coeficiente de rugosidad se realizó mediante ejemplos de problemas supuestos en un canal abierto. Estos ejemplos incluyen el flujo en un canal individual caracterizado varios valores de coeficientes de rugosidad.

Los datos resultantes de la observación se simularon resolviendo las ecuaciones gobernantes (1) y (2) para valores de rugosidad que se asumen como los coeficientes reales del canal. Estos datos de tirantes de flujo simulados se utilizaron luego en el modelo de optimización para estimar los coeficientes.

De esta forma, es posible evaluar el comportamiento del método propuesto en diferentes casos que incluyen la variación de la ubicación de las estaciones de observación y la desviación de los datos debido a errores en las mediciones.

5.1. CARACTERÍSTICAS DEL PROBLEMA

Se estudia el comportamiento en un canal natural de sección trapezoidal con pendiente lateral 1H:1V y base de fondo de 1.5 m. La longitud de del canal es de 35000 m y la pendiente longitudinal tiene un valor de 0.0004. Las condiciones iniciales del flujo se asumen como conocidas. Como condición de borde aguas arriba del canal se elige un hidrograma de entrada. La variación de Q con respecto al tiempo está dada por la siguiente ecuación:

$$Q(t) = Q_b + \left(Q_p - Q_b\right) e^{-\frac{0.5(t-1)-t_p}{t_g - t_p}} (t-1)(0.5/t_p)^{t_p/(t_g - t_p)} \qquad Q(L,t) = Q_v(y) \tag{7}$$

donde Q_b es el caudal inicial (3,32 m^3/s); Q_p es el caudal pico (6 m^3/s); t_p es el tiempo del pico (8 hs) y t_g es el tiempo del centroide del hidrogama (10 hs)**.** Se obtuvieron datos de tirantes cada media hora y distribuidos espacialmente cada dos mil quinientos metros. El canal único se dividió en cuatro tramos con coeficientes de rugosidad diferentes. Los valores de los coeficientes reales son: $n_{r(1)}$=0.016; $n_{r(2)}$=0.0024; $n_{r(3)}$=0.032 y $n_{r(4)}$=0.040. La distribución de los mismos a lo largo del canal se muestra en la Figura 1. Los valores iniciales de los mismos para el modelo de optimización son iguales a n_{min}.

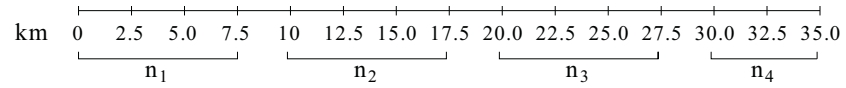


Figura 1: Distribución de los coeficientes de rugosidad

Se estimaron los coeficientes de rugosidad realizando mediciones de tirantes en cuatro, tres, dos y un punto del canal. La Tabla 1 muestra los coeficientes obtenidos. En la misma se puede observar que la precisión de la identificación disminuye a medida decrece la cantidad de estaciones de medición. El error presentado en la tabla está dado por la ecuación:

$$e_{n_i}(\%) = 100\,|n(i) - n_r(i)|/n_r(i) \tag{8}$$

En el caso de cuatro estaciones de medición, se obtuvieron los mismos resultados tomando mediciones de tirantes cada 1, 2 y 3hs.

Por otra parte se estudió la identificación de los coeficientes tomando mediciones de tirantes y de caudales. La Tabla 2 muestra los coeficientes obtenidos de esta forma midiendo sólo en dos puntos del canal. Además se muestra el aumento de precisión en la identificación cuando se introducen errores en las mediciones del 5% y 10%.

6. CONCLUSIONES

En el presente trabajo se aplicó una metodología para la identificación de coeficientes de rugosidad en canales de riego. La técnica aplicada consiste en una combinación de un método de optimización global y la solución aproximada por diferencias finitas de las ecuaciones que describen la dinámica de canales. El método de optimización aplicado se denomina “Método de recocido simulado” y se utilizó para minimizar una función objetivo planteada como la diferencia entre los tirantes estimados y los tirantes medidos. Estos tirantes se simulan a partir de coeficientes de rugosidad asumidos como reales.

Ptos. de medición x (km)	n	e_n (%)	Ptos. de medición x (km)	n	e_n (%)	Ptos. de medición x (km)	n	e_n (%)	Ptos. de medición x (km)	n	e_n (%)
2.5	0.0158	1.2	12.5	0.0196	22.5	15.0	0.0196	22.50	30.0	0.0235	46.88
12.5	0.0235	2.1	22.5	0.0274	14.2	30.0	0.0274	14.17		0.0158	34.17
22.5	0.0313	2.2	30.0	0.0351	9.7		0.0351	9.69		0.0390	21.88
30.0	0.0390	2.5		0.0468	17.0		0.0468	17.00		0.0468	17.00

Tabla 1: Identificación de *n* a partir de la medición de tirantes.

Datos:	y		y + Q		y (error 5%)		y + Q (error 5%)		y (error 10%)		y + Q (error 10%)	
Ptos. de med.	n	e_1 (%)	n	e_1 (%)	n	e_1 (%)	n	e_1 (%)	n	e_1 (%)	n	e_1 (%)
	0.0196	22.5	0.0158	1.2	0.0196	22.5	0.0158	1.2	0.0196	22.5	0.0158	1.2
15.0	0.0274	14.2	0.0235	2.1	0.0274	14.2	0.0235	2.1	0.0274	14.2	0.0235	2.1
30.0	0.0351	9.7	0.0313	2.2	0.0351	9.7	0.0313	2.2	0.0313	2.2	0.0313	2.2
	0.0468	17.0	0.0390	2.5	0.0506	26.5	0.0390	2.5	0.0468	17.0	0.0429	7.3

Tabla 2: Identificación de *n* a partir de la medición de tirantes y caudales.

Se analizaron diferentes escenarios en cuanto a la cantidad de estaciones de medición y a la calidad de los datos utilizados para la identificación. Se muestra que se obtienen buenos resultados con una estación de medición por coeficiente. Además, es posible mejorar la identificación con datos provistos de ruido tomando mediciones simultáneas de caudales y tirantes.

Así mismo, otros parámetros que intervienen en la dinámica de canales pueden ser estimados mediante esta metodología, como el coeficiente de infiltración o las condiciones iniciales.

REFERENCIAS

[1] CHANSON, H., *Environmental Hydraulics of Open Channel Flows*, Elsevier Butterworth-Heinemann, 2004.

[2] DING, Y., WANG, S.S.Y., *Optimal Control of Open-Channel Flow Using Adjoint Sensitivity Analysis*, Journal of Hydraulic Engineering, 132, (11), 2006.

[3] KEUNING, D.H., *Application of Finite Element Method to Open Channel Flow,* Journal of Hydraulics Division, ASCE, 102:459-467, HY4, 1976.

[4] RAMESH, R., BITHIN DATTA , MURTY BHALLAMUNDI, S., NARAYANA, A., *Optimal Estimation of Roughness in Open-Channel Flows,* Journal of Hydraulic Engineering, ASCE, 126, (4), 200.

[5] VIDAL, M.C., *Un procedimiento heuristic para un problema de asignación cuadrática,* Tesis de magíster en Matemática, Departamento de Matemática, Universidad Nacional del Sur, 2003.

[6] YANG, W.Y., CAO, W., CHUNG, T.S., MORRIS, J., *Applied numerical Methods using MATLAB,* Wiley-Interscience, 2005.

[7] YOST, S.A., KATOPODES, N.D., *Global Identification of Surface Irrigation Parameters,* Journal of Irrigation and Drainage Engineering, 124, (3), 1998.

IMPLEMENTACIÓN DE UNA HEURÍSTICA PARA LA ESTIMACIÓN DE UNA MATRIZ OD USANDO CIUDADSIM

Jorgelina Walpen[♭,†] y Elina M. Mancinelli[♭,†]

[♭]*FCEIA, Universidad Nacional de Rosario, Pellegrini 250, 2000 Rosario, Argentina*
[†]*CONICET, Argentina*
walpen@fceia.unr.edu.ar, elina@fceia.unr.edu.ar, www.fceia.unr.edu.ar/optimiz_control/

Resumen: Se estudia el problema de estimación de una matriz origen-destino asociada a un problema de afectación de tráfico a una red. Usando CiudadSim se implementa una heurística basada en la aproximación de derivadas parciales de una función objetivo respecto de la variación de la demanda.

Palabras clave: *afectación de tráfico, estimación de la matriz origen destino, solución numérica.*
2000 AMS Subject Classification:

1. INTRODUCCIÓN

La modelización matemática del tráfico requiere una gran cantidad de datos sobre la red vial y sobre las demandas de transporte, esto último dado por una matriz origen destino (OD). Consideramos el problema de estimar una matriz OD utilizando mediciones de flujo en ciertos arcos de la red y una matriz OD objetivo (por ejemplo una antigua matriz OD desactualizada u obtenida por otro medio).

2. FORMULACIÓN DEL PROBLEMA

La formulación matemática del problema de estimación de la matriz de demanda (DAP) se ubica en el marco de los MPECs, dado que se trata de un problema de optimización con restricciones de equilibrio. En este caso las restricciones corresponden a la versión determinística del equilibrio del usuario de Wardrop (DUE) parametrizado en la variable demanda. La demanda es la variable del nivel superior, donde tenemos el problema de ajuste de la matriz OD que combina información de anteriores estimaciones y mediciones actuales de la red. La formulación del problema DAP sobre el espacio de flujos por ruta, resulta:

$$\begin{aligned} \text{(DAP)} \quad & \text{mín } F(g,v) = \eta_1 F_1(g) + \eta_2 F_2(v) \\ & s.a.\ S(h)^t(h'-h) \geq 0,\ \forall h' \in \Omega(g), \\ & h \in \Omega(g),\ v = \Delta h,\ g \geq 0. \end{aligned}$$

donde la función F_1 mide la distancia entre g y la matriz target $\tilde{g}$ y F_2 mide la desviación entre el flujo asignado para la matriz g y el flujo medido sobre determinados arcos fijos $\tilde{v}$. Las métricas usualmente utilizadas son las de mínimos cuadrados, funciones de máxima entropía y máxima similitud, (ver[1]). Los parámetros η_1 y η_2 reflejan la confianza relativa en la información contenida en $\tilde{g}$ y $\tilde{v}$, v es el vector de flujos por arco, h vector de flujos por ruta y $S(h)$ vector de costos por ruta.

En una versión general de DAP para el cual sólo se piden hipótesis de continuidad sobre las funciones F_1, F_2 y $s(v)$ se prueba en [1] que el problema admite al menos una solución. En este trabajo consideramos el caso en el cual el problema de equilibrio (DUE) admite solución única en el espacio de flujos por arco. Será necesario entonces, hacer los siguientes supuestos sobre la red: – la red de transporte está fuertemente conectada, – las funciones de tiempo de viaje sobre las rutas son aditivas, – los tiempos de viaje sobre los arcos son separables, – la demanda d_{pq} es positiva para cada $(p,q) \in \mathcal{C}$, – la función de tiempo de viaje $c_a : \mathbb{R} \mapsto \mathbb{R}$ es positiva, continua y no decreciente para cada arco $a \in \mathcal{A}$.

Estas hipótesis garantizan la existencia de equilibrio (tanto en la variable de flujo por arco como la de flujo por ruta) y la unicidad de los tiempos de viaje asociados al equilibrio. Asumiendo además que: Cada función de tiempo de viaje t_a es estrictamente creciente, se tiene unicidad de solución de equilibrio dada en la variable de flujo por arco.

3. ANÁLISIS DE LA FUNCIÓN OBJETIVO

Para resolver este problema de optimización se lo puede reformular como un problema a un solo nivel a través de una función objetivo definida implícitamente, y usar un algoritmo de descenso basado en el método del gradiente proyectado. Será entonces necesario conocer la dependencia de la variación del flujo respecto de la variación de la demanda. En efecto, si consideramos como modelo de función objetivo una función del estilo:

$$F(g) = \eta_1 F_1(g) + \eta_2 F_2(v(g))$$

y procedemos al cálculo de su gradiente, obtenemos:

$$\nabla_g F(g) \quad = \quad \eta_1 \nabla_g F_1(g) + \eta_2 \nabla_g F_2(v(g)) \quad = \quad \eta_1 \nabla_g F_1(g) + \eta_2 \nabla_g v(g) \nabla_v F_2(v(g)) \tag{1}$$

$$\nabla_g F_1(g) = \left\{ \frac{\partial F_1(g)}{\partial g_i} \right\}_{i \in I} \text{ y } \nabla_v F_2(v(g)) = \left\{ \frac{\partial F_2(v(g))}{\partial v_a} \right\}_{a \in \tilde{A}}, \tag{2}$$

$$\nabla_g v(g) = \left\{ \frac{\partial v_a(g)}{\partial g_i} \right\}_{a \in A, i \in I} \tag{3}$$

donde los gradientes en (2) no presentan ninguna dificultad para su cálculo, cuando elegimos la función F dada por las distancias

$$F_1 = \frac{1}{2} \sum_{i \in I} (g_i - \tilde{g}_i)^2 \qquad\qquad F_2 = \frac{1}{2} \sum_{a \in \tilde{A}} (v_a(g) - \tilde{v}_a(g))^2,$$

donde $\tilde{A} \subset A$ comprende los arcos medidos.

En la matriz Jacobiana (3) no conocemos una fórmula explícita que relacione v y g y muestre cómo depende la primera de la segunda. Por lo tanto se hará una aproximación de la misma.

Implementamos la heurística local propuesta en [5] para la solución del problema general incorporando la versión del algoritmo DSD (ver [6]) disponible en [2] y descrita en [4], para el paso del cálculo de los flujos por ruta. Esto implica poder interpretar la información manejada por el DSD, la cual se encuentra económicamente almacenada, y utilizarla adecuadamente para seguir los pasos de la heurística.

4. IMPLEMENTACIÓN NUMÉRICA

La idea general de la propuesta es la de evaluar las derivadas direccionales de la función objetivo en el punto actual y obtener una dirección de descenso haciendo una proyección sobre el espacio definido por las restricciones de no negatividad.

El esquema de la heurística se describe a continuación:

0. *Inicialización.* Sea g^0 una aproximación inicial factible para la demanda. Resolver para g^0 el DUE. Poner $l = 0$.
1. *Cálculo de una dirección de descenso*. Encontrar una dirección r^l tal que $F(g) < F(g^l) + r^l(g - g^l)$ en un entorno de g^l $(g \neq g^l)$. Si $\exists$ el gradiente, $r^l = -\nabla_g F(g^l)$ es válida como dirección de descenso.
2. *Cálculo de una dirección de búsqueda.* Ajustar la dirección de descenso r^l con respecto a las condiciones de factibilidad de la demanda.
3. *Criterio de parada.*
4. *Búsqueda lineal*
   a. Encontrar una longitud de paso máxima ${\alpha^l}_{max} = \text{mín}\{+\infty, -{g^l}_i/\bar{r}_i^l : \bar{r}_i^l < 0, i \in I\}$.
   b. Encontrar α que minimiza $F(g^l + \alpha \bar{r}^l)$, $\alpha \in [0, {\alpha^l}_{max}]$.
5. *Actualización.* Poner $g^{l+1} = g^l + \alpha^l \bar{r}^l$ y v^{l+1} los flujos de equilibrio asociados a la demanda g^{l+1}. Poner $l = l + 1$ y volver al paso 1.

En el paso 1 se aproximan las derivadas direccionales. Se resuelve un problema cuadrático para cada par origen destino i a fin de obtener una aproximación de $\dfrac{\partial v}{\partial g_i}$, (ver [5]).

El problema cuadrático asociado al par OD $\bar{i}$ es el siguiente:

$$\min_{x} \quad \frac{1}{2} d^T S d$$

$$s.a. \;\; \Lambda x = e_{\bar{i}}, \;\; d = \Delta x. \tag{4}$$

donde el vector $x = (x_k)_{k \in K_{\bar{i}}}$ y $x_k = \dfrac{\partial h_k}{\partial g_{\bar{i}}}(g^l)$, el vector $e_{\bar{i}}$ expresa una unidad de cambio y el vector $d = (d_a)_{a \in A}$ es tal que $d_a = \dfrac{\partial v_a}{\partial g_{\bar{i}}}(g^l)$. La matriz constante S tiene por entradas $S_{ab} = \dfrac{\partial c_a}{\partial v_b}(g^l)$, $a, b \in A$. Por las hipótesis de separabilidad de los arcos, S se reduce a una matriz diagonal. La matriz Λ representa la incidencia parOD-ruta, y contempla solo las rutas efectivamente utilizadas. La matriz Δ representa la incidencia arco-ruta.

El gradiente de la función objetivo de (4) es fácil de calcular e involucra unas cantidades que pueden interpretarse como "costos por ruta"σ_k si consideramos los "costos por arco"$\mu_a = S_{aa} d_a$.

Los costos σ_k son definidos como: $\sigma_k = \sum_{a \in A} \mu_a \delta_{ak}$.

Desarrollando, tenemos:

$$\begin{aligned}
\sigma_k &= \sum_{a \in A} (S_{aa} d_a) \delta_{ak} \\
&= \sum_{a \in A} \frac{\partial c_a}{\partial v_a} \Big(\sum_{j \in K} x_j \delta_{aj} \Big) \delta_{ak} \\
&= \sum_{a \in A} \frac{\partial c_a}{\partial v_a} \Big(\sum_{j \in K_{\bar{i}}} \frac{\partial h_k}{\partial g_{\bar{i}}} \delta_{aj} \Big) \delta_{ak} \text{ con la matriz de proporción localmente constante} \frac{\partial h_k}{\partial g_{\bar{i}}} = \lambda_{k\bar{i}} \\
&= \sum_{a \in A} \frac{\partial c_a}{\partial v_a} \Big(\sum_{j \in K_{\bar{i}}} \lambda_{k\bar{i}} \delta_{aj} \Big) \delta_{ak}
\end{aligned}$$

La búsqueda lineal posterior es exacta y la subrutina para el problema cuadrático concluye con una aproximación del vector d.

El problema (4) puede interpretarse como el problema de asignar una unidad de flujo adicional en un par OD $\bar{i}$, de manera tal que los cambios en los costos de todas las rutas para un mismo par OD y para cada uno de ellos, asociados a la solución de equilibrio de la iteración actual, sean los mismos.

La aproximación d^* de la derivada direccional asociada al valor óptimo x^* representa la variación en los flujos por arco cuando una unidad adicional de demanda es asignada al par $\bar{i}$. El vector d^* conformará la columna $\bar{i}$ de la matriz (3).

El cómputo de estos costos requiere recorrer las rutas generadas para cada parOD durante el proceso de afectación con el DSD programado en [2].

Una vez obtenidas cada una de las columnas de la matriz (3) se calcula el gradiente y se proyecta sobre el espacio de las restricciones de no negatividad para las demandas:

$$\bar{r}_i^l := \begin{cases} r^l{}_i, & \text{si } g^l{}_i > 0 \text{ ó } g^l{}_i = 0 \; y \; r^l{}_i > 0, \\ 0, & \text{en otro caso,} \end{cases}$$

alcanzando finalmente la dirección de descenso para la solución global. Para la búsqueda lineal se utiliza una búsqueda tipo Armijo iniciada con una longitud de paso máxima, obtenida como se indica en el paso 4a, que garantice no negatividad en todas las demandas.

5. EJEMPLOS

Se presentan resultados numéricos para la red de Steenbrink.

La figuras 1 y 2 corresponden al caso de la red de Steenbrink con 2 demandas. Para esta cantidad de demandas fue posible graficar la función objetivo del problema del nivel superior y comparar los resultados obtenidos con la heurística (ver tabla en Figura 2).

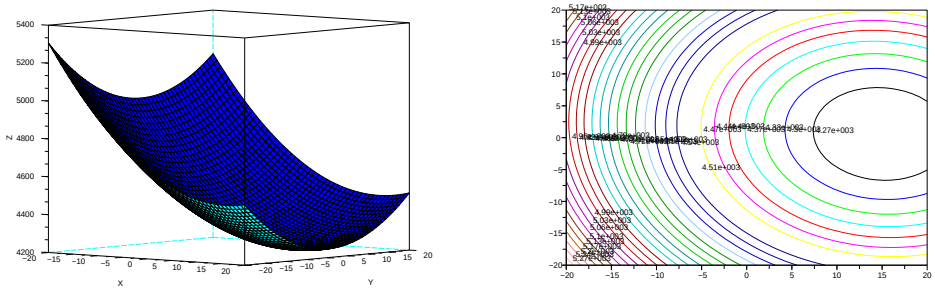


Figura 1: Gráfica y conjuntos de nivel de F para la red Steenbrink con 2 demandas

Iter	Valor de F	Valores de las demandas	
0	6739.4819	1055.	950.
1	5345.0507	1044.45	969.06685
2	4573.5087	1034.0055	985.88055
3	4276.9355	1023.6654	1000.2353
4	4267.6964	1022.6418	1001.355
5	4261.1628	1021.6191	1002.4545
6	4257.2454	1020.5975	1003.533
7	4255.9271	1019.5769	1004.5888
8	4256.7094	1018.5573	1005.6204
9	4260.0278	1017.5388	1006.6257
10	4265.5884	1016.5212	1007.6015

Figura 2: Resultados numéricos obtenidos con la heurística para la red Steenbrink con dos demandas y mediciones en 7 arcos.

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por los proyectos PIP 112-200801-00460 CONICET e ING272 UNR.

REFERENCIAS

[1] Chen Y., Florian M., (1996), *OD demand adjustment problem with congestion: Part I. Model Analysis and optimality conditions*, Advanced Methods in Transportation Analysis, Springer-Verlag, Berlin, pp.1-22.

[2] CIUDADSIM: http://www-rocq.inria.fr/metalau/ciudadsim/

[3] García-Ródenas R., Verastegui-Rayo D. (2006), *A column generation algorithm for the estimation of origin-destination matrices in congested traffic networks*, European Journal of Operational Research, ,pp.860-878.

[4] Lotito P., Mancinelli E., Quadrat J.P., Wynter L. (2003), *The Traffic Assignment Toolboxes of Scilab*, INRIA - Rocquencourt.

[5] Lundgren J.T., Peterson A. (2008), *A Heuristic for the Bilevel Origin-Destination Matrix Estimation Problem*, Transportation Research Part B: Methodological, 42, 4, pp.339-354.

[6] Patriksson M. (1994), *The Traffic Assignment Problem. Models and Methodes*, VSP BV, Utrecht.

[7] Patriksson M., *Sensitivity analysis of traffic equilibria* , Transportation Science 38, 3, pp. 258,281.

PARAMETRIZATION OF THE DOMAIN OF MAXIMAL ATTRACTION OF THE GUMBEL DISTRIBUTION

Aldo J. Viollaz† y Víctor F. Lazarte‡

†Facultad de Ciencias Económicas, Universidad Nacional de Tucumán, San Miguel de Tucumán,
ajviollaz@yahoo.com.ar, aviollaz@herrera.unt.edu.ar

‡ Facultad de Ciencias Exactas y Tecnología, Universidad Nacional de Tucumán, San Miguel de Tucumán,
vlazarte@herrera.unt.edu.ar

Summary: The main objective of this work is to propose a new approach to the problem of extremes estimation. The key idea is to parametrize the domain of maximal attraction of the Gumbel distribution through a semi-parametric family obtained applying the inverse Box-Cox transformation to a variable with a Gumbel distribution. This provides a model for each member of the Gumbel maximal attraction domain given as results a more accurate modeling of the population distribution function, faster convergence of the maxima estimates to their limits and better estimates of the population high level quantiles and other extremes statistics.

Key words: Box-*Cox transformation, parametrization Gumbel maximum attraction domain*.

1. INTRODUCTION

A classic result of the theory of the extremes states that the distribution of the maximum of a set of independent and identically distributed observations $X_1, \ldots, Xn$ of a random variable X with distribution function F, if does converge when n tends to infinity, can converges only to one of three possible limit distributions, one of which is the distribution of a Gumbel random variable that has an important role in this paper. It is a known fact that the convergence to its limit can be very slow. as Fisher and Tippett [4] showed us. See Galambos [3], Resnick [2], Embrechts et al [2] and Leadbetter et al [6]. In Galambos it is proved that the speed of convergence the distribution of the maximum of i.i.d. Gaussians observations is of 1/ln(n) order, whereas for the exponential distribution it is of l/*n* order. The fact is that in general the convergence of the distribution of the maximum to its limit is slow. What could it be an intuitive explanation of this phenomenon? The Gumbel distribution is max-stable, that is, the maximum distribution of i.i.d. Gumble observations conveniently normalized, coincides with the distribution of origin (also limit) for all n. This simple observation lets us infer that the speed of convergence of the maximum distribution will be much slower as further the sampled distribution be from the Gumbel distribution. In this work we propose to model the Gumbel maximal attraction domain through a family of distributions obtained applying an inverse Box-Cox transformation to a Gumbel distribution.

Let $X_1, \ldots, X_n$ be observations of a random variable X with distribution F_X. Let $Y = g(X)$ where g is a function strictly monotonous. If g^{-1} represents the inverse function of g and if $Y_i = g(X_i), i = 1, \ldots, n$ we have:

$$max(X_1, \ldots, X_n) = g^{-1}[max(Y_1, \ldots, Y_n)] \tag{1.1}$$

This equality allows us transform the observations into a set with better characteristics for the application of the statistics of extremes. Note that it is not necessary for the observations to be independent. On the other hand if $X_1, \ldots, X_n$ are i.i.d. according to F and F_Z is the distribution of $Z = max\{X_1, \ldots, X_n\}$ then $F_Z(z) = F(z)^n$. So the values of $F_Z(z) > \frac{1}{2}$ are determined by the values $F(z) > \left(\frac{1}{2}\right)^{\frac{1}{n}} = 0.9330$ if n=10, which shows that the transformation is heavily determined by the right tail of the distribution.

This work is structured as follows: In Section 2 we present the *potential Gumbel family* and it is proved that in a certain sense it is a sufficiently dense family in the *domain of maximal attraction of the Gumbel*

distribution. In Section 3 we consider approximations to the Gaussian, Weibull, Lognormal families through the potential Gumbel family. Section 4 is for some conclusions

2. POTENTIAL GUMBEL DISTRIBUTION

Definition. We say that a random variable *X* has a *potential Gumbel distribution* of parameters ξ, θ, λ, $-\infty < \xi < \infty,\ \theta > 0\ y\ \lambda > 0,$ if the random variable

$$Y = (X^{\lambda} - 1)/\lambda,\ \ X \geq 0. \tag{2.1}$$

has distribution function

$$F_Y(y) = \begin{cases} exp\left(-exp\left(-\frac{y-\xi}{\theta}\right)\right), & y \geq -\frac{1}{\lambda} \\ 0 \qquad\qquad\qquad\quad , & y < -\frac{1}{\lambda} \end{cases} \tag{2.2}$$

In the applications we will make a previous affin transformation over *X* so that the probability *X* be smaller than 0 would be practically negligible. Under this assumption we can consider, for practical effects, the random variable *X* be absolutely continuous with density function.

$$f_X(x) = \frac{1}{\theta} exp\left(-exp\left(-\frac{\frac{x^{\lambda}-1}{\lambda}-\xi}{\theta}\right)\right) exp\left(-\frac{\frac{x^{\lambda}-1}{\lambda}-\xi}{\theta}\right) x^{\lambda-1} \tag{2.3}$$

Fearn and Nebenzahl [7] and Nadarajah [10] proposed the use of power transformations for the quantiles approximations. Vanroelen [8] and Teguels and Vanroelen [9] have studied the effect of the transformations of Box-Cox type in the approximation of the maximum of random variables belonging to the *Frechet and Gumbel attraction domain*s. Here we show the importance of the p*otential Gumbel family* as a model of the *Gumbel attraction domain* and in the maximum estimation of observations belonging to that attraction domain.

Every distribution function of an absolute continuous variable with a range (0,∞) can be expressed as function of the hazard rate, $h(x) = f(x)/(1 - F(x))$, as:

$$\bar{F}(x) = 1 - F(x) = \exp\left(-\int_0^x h(t)dt\right) \tag{2.4}$$

Von Mises [1] found a representation of the of distribution functions of the Gumbel maximal attraction domain in the following way

$$\bar{F}(x) = c(x) \exp\left(-\int_z^x \frac{1}{a(t)} dt\right),\ z < x < \infty. \tag{2.5}$$

where z is a constant, $z < x < w = sup\{x: F(x) < 1\},\ a(x)$ a positive function, absolutely continuous with density $a'(x) \to 0$ when $x \to \infty$ and $c(x)$ is a function so that $c(x) \to c > 0$ when $x \to \infty$. If a function *F* has a representation (2.5) the auxiliary function *a(x)* can be chosen as the reciprocal of the hazard rate,

$$a(x) = h^{-1}(x) = \bar{F}(x)/f(x) \tag{2.6}$$

So the auxiliary function *a(x)* for the distribution F_X can be chosen equal to

$$a(x) = \frac{\bar{F}_X(x)}{f_X(x)} = \theta \bar{F}_X\left(\frac{\frac{x^{\lambda}-1}{\lambda}-\xi}{\theta}\right) / f_Y\left(\frac{\frac{x^{\lambda}-1}{\lambda}-\xi}{\theta}\right) x^{\lambda-1} \sim \theta x^{1-\lambda} \tag{2.7}$$

Where $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$.
Since $\lambda > 0,\ a'(x) \sim \theta(1-\lambda)x^{-\lambda} \to 0,$ when $x \to 0$. So the distribution *F* belongs to the Gumbel maximal attraction domain. Note that if $\lambda = 0,\ a'(x) = 1$ does not converge to 0 and therefore the distribution does not belong to the Gumbel domain any longer. It can be easily proved that in this case the transformed variable is $X = \exp(Y)$ which has a Frèchet distribution.
A general characteristic of the functions $a(x)$ is that $lim \frac{a(x)}{x} = 0$. That is $a(x)$ has to grow more slowly than *x,* and this precisely ocurrs with $a(x) = x^{1-\lambda},\ \lambda > 0$. The value $\lambda = 1$ gives us $a(x) = constant$

which corresponds to the Gumbel distribution. To large values of λ correspond large values *of* $a^{-1}$and so large values for its integral between *z* and *x*. Note that the family of functions

$$a(x) = \theta x^{1-\lambda},\ \lambda > 0, \tag{2.8}$$

has as boundary function the function $a(x) = \theta x$ which does not correspond to the family. Invoking the parsimony principle we assume that $a(x)$ are functions of regular variation. That is

$$a(x) = L(x)x^{p},\ p > 0 \tag{2.9}$$

Where $L(x)$ is a function of slow variation. In the expression (2.8) we replace the function $L(x)$ with a constant. Under the assumption (2.8), the family (2.8) seems to be rich enough to approximate the family to all the possible $a(x)$ functions.

3. Aproximations to the Normal, Lognormal, Weibull and Gamma Families

The potential Gumbel family is an adequate instrument to approximate the Normal, Lognormal, Weibull and Gamma families, among others. Equivalently we can approximate the transforms of these families through the Gumbel family. This approach is the core of the method which is explained in Section 1 to speed up the convergence of the maximum of a random sample of a random variable and to improve the approximation of the statistical models used in the analysis of the extremes, transforming the basic data so that their distribution function be near to the Gumbel distribution function.

Normal Distribution. We have the following relationship: $\bar{F}(x) = \bar{\Phi}(x) \sim \varphi(x)/x$. Therefore the corresponding hazard rate satisfies: $h(x) \sim x$. Moreover, the hazard rate of the potential Gumbel distribution is equal to $\theta^{-1}x^{\lambda-1}$. Should therefore take $\lambda = 2$ to approximate normal distribution for the potential Gumbel distribution.

Lognormal Distribution. A random variable *X* has lognormal distribution if the variable $Y = \ln(X)$ has normal distribution $N(\mu, \sigma^2)$. Then, taking $\mu = 0$ and $\sigma^2 = 1$ we have $F_X(x) = F_Y(\ln x)$, $f_X(x) = f_Y(\ln x)d(\ln x)/dx$ and the hazard rate satisfies: $h(x) = x^{-1}\varphi(ln\ x)/\bar{\Phi}(x) = ln\ x/x$. This function $h(x)$ is almost on the boundary $h(x) = 1/x$ of the functions h of the potential Gumbel family. In making the estimation of the parameters of the approximant potential Gumbel family would give a value (close to 0) for the exponent λ of the transformation.

Weibull Distribution. $\bar{F}(x) = exp(-(\gamma x)^{p})$, $f(x) = \gamma p\ exp(-(\gamma x)^{p})(\gamma x)^{p-1}$, $\gamma > 0, 0 < p < \infty$
Then the hazard rate satisfies: $h(x) = \gamma p\ (\gamma x)^{p-1}$. Therefore we must take $\lambda = p$.

Gamma Distribution. $f(x) = \eta\Gamma(\gamma)^{-1}(\eta x)^{\gamma-1}\ exp(-\eta x), \eta, \gamma > 0$. Taking limits for x tending to infinity and applying L'Hopital we have: $lim\frac{f(x)}{F(x)} = lim\frac{f'(x)}{f(x)} = \eta$.
Therefore we must take λ equal to one (corresponding to the identity transformation) as expected.

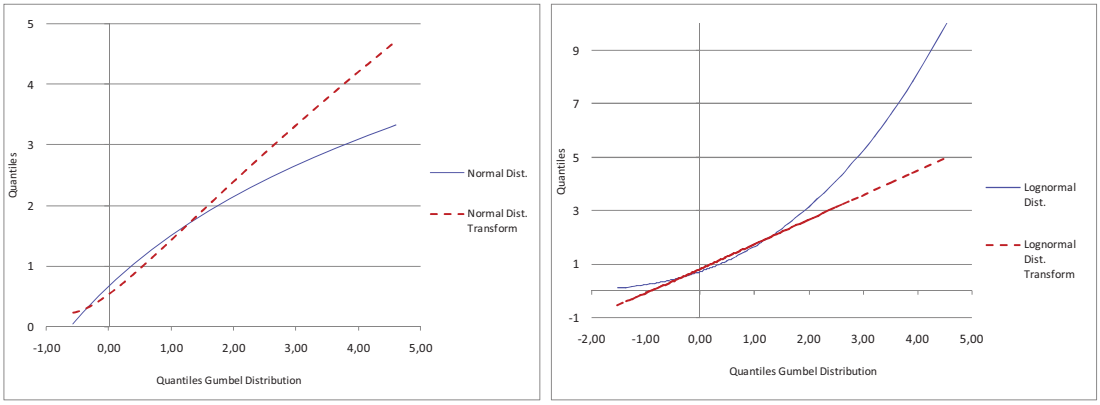
Figure 1 presents two cases that show the capacity of the potential Gumbel family to represent fuctions of the Gumbel maximal attraction domain.

4. Conclusions

In this work we propose to model the Gumbel maximal attraction domain through a family of distributions obtained applying an inverse Box-Cox transformation to a Gumbel distribution. Note that when we estimate the maximum taking the *Generalized Distribution of Extremes* as model, the Gumbel maximal attraction domain is represented by only one member: the Gumbel distribution, while the proposed family has infinite members distributed all over the Gumbel attraction domain. The proposed family is obtained from the von Mises representation of the functions of the Gumbel attaction domain and it appears to be dense enough for the applications.

In this paper we did not address the statistical problem of estimating the maximum and its higher percentiles. Note that the statistical estimation process will provide an estimate of the parameter λ thus

selecting an approximating family member. In a continuation of this work the statistical problem will be addressed in doctoral thesis of one the authors.



(a) (b)

Figure 1: (a) Q-Q Plot Normal distribution versus Gumbel distribution and Q-Q Plot transformed Normal distribution versus Gumbel distribution. (b) Q-Q Plot Logormal distribution versus Gumbel distribution and Q-Q Plot transformed Logormal distribution versus Gumbel distribution.

4. REFERENCES

[1] VON MISES, *La distribución de la plus grande de n valours. Selected papers II*. Am. Math. Soc. 271-294, 1936.

[2] S. I. RESNICK. *Extreme Values, Regular Variation, and Poin Process*, Springer-Verlag, 1987.

[3] J. GALAMBOS, *The Asypmtotic Theory of Extremes Order Statistics*, Second Edition, Robert E. Krieger Publishing Company. Malabar, Florida.1987.

[4] R. A. FISHER, *Limiting forms of the frecuence distributions of largest or smallest of a sample member*. Proc Cambride Philos, Soc. 24, 180-190, 1928.

[5] P. EMBRECHTS, C. KLUPPELBERG Y T. MIKOSCH. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

[6] M. R. LEADBETTER, G. LINDGREN Y H. ROOTZEN. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York Heidelberg Berlin. 1982.

[7] D.H. FEARN Y E. NEBENZAHL, *Using Power Transformations When Approximating Quantiles*. Comun. Statist. Teory Method 24(4) 1073-1093, 1995.

[8] G. VANROELEN, *The effect of transformations on second-orden regular variation*. Doctoral Thesis, Katholieke Univ. Leuven, 2003.

[9] J. TEUGELS Y G. VANROELEN, *Box-cox transformatios and heavy-tailed distributions*. J Appl. Prob. 41A, 213-27, 2004.

[10] S. NADARAJAH. *The exponentiated Gumbel distribution wiht climate application*. Environmetrics, 17(1), 13-23. John Wiley & Sons, Ltd. 2005.

ESTUDIO DE LA DISPONIBILIDAD DE UN SISTEMA UTILIZANDO CADENAS DE MARKOV AGRUPABLES

Fredy Cuenca

Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel, Perú, fcuenca@pucp.edu.pe

Resumen:

El rendimiento de un sistema puede verse afectado cuando éste es sometido a un entorno hostil. Una técnica muy utilizada para reducir el impacto del entorno hostil consiste en redundar los componentes críticos del sistema con la esperanza de que si algunos son afectados a causa de este entorno, los otros podrían continuar fucionando correctamente permitiendo que el sistema siga operando según lo esperado. En este trabajo, se estudiará la disponibilidad de un sistema cuyo componente crítico ha sido redundado de tal modo que el sistema en estudio puede ser modelado como un arreglo de dos componentes en paralelo. Se asumirá que las disponibilidades de los componentes pueden ser representadas como cadenas de Markov homogéneas e independientes. El objetivo del trabajo es identificar las situaciones en las que la disponibilidad del sistema puede ser representada como una cadena de Markov homogénea.

Palabras clave: *cadenas de Markov agrupables, sistemas tolerantes a fallas*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

El rendimiento de un sistema puede verse afectado cuando éste es sometido a un entorno hostil. Por ejemplo, el sistema eléctrico de un avión podría dejar de funcionar correctamente durante algunos momentos mientras el avión atraviesa una fuerte tormenta eléctrica. Una técnica muy utilizada para reducir el impacto del entorno hostil sobre un sistema es la denominada: técnica de redundancia de componentes. Esta consiste en redundar los componentes críticos del sistema con la esperanza de que si algunos son afectados a causa del entorno hostil, los otros podrían continuar fucionando correctamente permitiendo que el sistema siga operando según lo esperado.

En adelante, diremos que un sistema (componente) está disponible cuando éste está funcionando correctamente. En caso contrario, se dirá que el sistema (componente) no se encuentra disponible.

En este trabajo, se estudiará la disponibilidad de un sistema cuyo componente crítico ha sido redundado de tal modo que el sistema en estudio puede ser modelado como un arreglo de dos componentes en paralelo. Este estudio asume que los componentes se encuentran operando simultáneamente bajo un entorno hostil. Además, el sistema estará disponible cada vez que al menos uno de sus dos componentes se encuentre disponible. Formalmente, se dice que el sistema en estudio posee la configuración *1-out-of-2*.

En términos matemáticos, se utilizará el proceso estocástico $X^{(i)} = \{X_n^{(i)}\}_{n\in\mathbb{N}}, i = 1, 2$ para representar la disponibilidad del i-ésimo componente del sistema mientras éste es sometido a un entorno hostil. Estos procesos serán modelados como cadenas de Markov homogéneas e independientes definidas sobre el espacio de estados $\{0, 1\}$. La variable aleatoria $X_n^{(i)}$ tomará el valor 0 (ó 1) para indicar que el i-ésimo componente se encuentra disponible (no disponible) en el periodo n. Además, la distribución inicial del proceso $X^{(i)}$ se denotará como $\alpha^{(i)}$ y tomará el valor $(\alpha_1^{(i)}, \alpha_2^{(i)})$ lo cual significa que la probabilidad de que el i-ésimo componente haya estado disponible (no disponible) antes de ser sometido al entorno hostil era $\alpha_1^{(i)}$(ó $\alpha_2^{(i)}$). Finalmente, la matriz de probabilidad de transición (m.p.t.) $P^{(i)}$ que caracteriza a la cadena de Markov homogénea (CMH) $X^{(i)}$ viene dada por:

$$P^{(i)} = \begin{pmatrix} p_{1,1}^{(i)} & p_{1,2}^{(i)} \\ p_{2,1}^{(i)} & p_{2,2}^{(i)} \end{pmatrix}$$

donde $p_{1,1}^{(i)} + p_{1,2}^{(i)} = p_{2,1}^{(i)} + p_{2,2}^{(i)} = 1$.

Vale la pena mencionar que a pesar que las dos componentes realizan la misma tarea, sus m.p.t. podrían ser distintas, pues éstas representan la reacción de cada componente al entorno hostil. Obviamente, esta

reacción dependerá de muchas variables como: el fabricante y la fecha de fabricación de cada componente por ejemplo.

En cuanto al sistema, su estado puede ser modelado como el proceso estocástico $X = \{(X_n^{(1)}, X_n^{(2)})\}_{n\in\mathbb{N}}$ definido sobre el espacio de estados $\mathbb{E} = \{(0,0),(0,1),(1,0),(1,1)\}$. Obviamente, el estado del sistema depende de los estados de sus componentes (disponible / no disponible). Por otro lado, se sabe de [4] que cuando las CMH $X^{(i)}$ son independientes, el proceso X definido anteriormente es también una CMH con distribución inicial $\alpha = \alpha^{(1)} \otimes \alpha^{(2)} = (\alpha_1^{(1)} \cdot \alpha_1^{(2)}, \alpha_1^{(1)} \cdot \alpha_2^{(2)}, \alpha_2^{(1)} \cdot \alpha_1^{(2)}, \alpha_2^{(1)} \cdot \alpha_2^{(2)})$ y m.p.t. P donde:

$$P = P^{(1)} \otimes P^{(2)} = \begin{pmatrix} p_{1,1}^{(1)} \cdot p_{1,1}^{(2)} & p_{1,1}^{(1)} \cdot p_{1,2}^{(2)} & p_{1,2}^{(1)} \cdot p_{1,1}^{(2)} & p_{1,2}^{(1)} \cdot p_{1,2}^{(2)} \\ p_{1,1}^{(1)} \cdot p_{2,1}^{(2)} & p_{1,1}^{(1)} \cdot p_{2,2}^{(2)} & p_{1,2}^{(1)} \cdot p_{2,1}^{(2)} & p_{1,2}^{(1)} \cdot p_{2,2}^{(2)} \\ p_{2,1}^{(1)} \cdot p_{1,1}^{(2)} & p_{2,1}^{(1)} \cdot p_{1,2}^{(2)} & p_{2,2}^{(1)} \cdot p_{1,1}^{(2)} & p_{2,2}^{(1)} \cdot p_{1,2}^{(2)} \\ p_{2,1}^{(1)} \cdot p_{2,1}^{(2)} & p_{2,1}^{(1)} \cdot p_{2,2}^{(2)} & p_{2,2}^{(1)} \cdot p_{2,1}^{(2)} & p_{2,2}^{(1)} \cdot p_{2,2}^{(2)} \end{pmatrix}$$

El símbolo $\otimes$ representa el producto Kronecker.

Al igual que toda CMH, el proceso X puede ser caracterizado a partir su distribución inicial y su m.p.t. por lo que también se le denotará como la dupla (α, P).

En la siguiente sección se modelará la disponibilidad del sistema como un proceso estocástico Y (que depende de X) y se deteminarán las condiciones que deben cumplirse para que Y sea una CMH.

2. CADENA DE MARKOV AGRUPABLE

El objetivo de este trabajo no consiste en estudiar el estado del sistema descrito sino la disponibilidad del mismo. Si ésta disponibilidad pudiese ser modelada como una CMH, se podría hacer uso de la extensa teoría existente para analizar diversos aspectos de la misma.

Definición 1 *Proceso Agregado*

Dadas una CMH $X = (\alpha, P)$ definida sobre un espacio de estados $\mathbb{E} = \{1, ..., N\}$, y una partición $\mathbb{B} = \{B(1), ..., B(M)\}$ del conjunto $\mathbb{E}$. El proceso estocástico Y definido como:

$$Y_n = k \Leftrightarrow X_n \in B(k) \qquad donde \;\; k = 1, ..., M$$

se denominará proceso agregado asociado a X sobre $\mathbb{B}$ y se denotará por $agg(\alpha, P, \mathbb{B})$.

Para el problema en estudio, definamos la partición $\mathbb{B} = \{B(1) = \{(0,0),(0,1),(1,0)\}, B(2) = \{(1,1)\}\}$ del espacio de estados $\mathbb{E}$. A cada subconjunto de $\mathbb{B}$ se le denominará metaestado. Note que el metaestado $B(1)$ contiene las situaciones que indican que al menos un componente del sistema (y por ende el sistema mismo) está disponible mientras que el metaestado $B(2)$ contiene los estados que indican que ninguna componente del sistema está disponible. Entonces, el proceso agregado $Y = agg(\alpha, P, \mathbb{B})$ representará la disponibilidad del sistema, la cual estará definida sobre el espacio de estados $\mathbb{F} = \{1 = disponible, 2 = no\ disponible\}$. Un estudio previo que utiliza cadenas de Markov agrupables para modelar la disponibilidad de un sistema tolerante a fallas se encuentra en [6].

Es importante mencionar que a pesar que X es una CMH, nada se puede afirmar acerca de Y. Existen procesos agregados $agg(\alpha, P, \mathbb{B})$ que son CM no homogénea, CM de orden k (k ≥ 2) o incluso procesos no markovianos. (Se pueden ver ejemplos en [2], [3] y [5]). También puede ocurrir que $agg(\alpha, P, \mathbb{B})$ sea una CMH para un valor particular de α o que sea una CMH para cualquier vector estocástico α. En el primer caso, se dice X es débilmente agrupable y en el segundo caso, que X es fuertemente agrupable.

Debido a que la disponibilidad inicial α del sistema es desconocida, será conveniente concentrarse en el estudio de los procesos fuertemente agrupables.

El siguiente teorema [5] presenta las condiciones suficientes y necesarias que deberá satisfacer un proceso agregado $Y = agg(\alpha, P, \mathbb{B})$ para que sea una CMH para cualquier vector estocástico α.

Teorema 1 *Cadena de Markov Fuertemente Agrupable*

Dadas una CMH $X = (\alpha, P)$ definida sobre un espacio de estados $\mathbb{E} = \{1, ..., N\}$, y una partición $\mathbb{B} = \{B(1), ..., B(M)\}$ del conjunto $\mathbb{E}$. El proceso $Y = agg(\alpha, P, \mathbb{B})$ será una CMH para cualquier distribución inicial α sii $\forall B(i), B(j) \in \mathbb{B}$:

$$\sum_{p \in B(j)} P(X_1 = p | X_0 = k) \quad tiene\ el\ mismo\ valor\ \forall k \in B(i)$$

Este valor común se denotará por $\hat{p}_{i,j}$ y representa la probabilidad de transición desde el metaestado $B(i)$ hacia $B(j)$. La m.p.t. del proceso Y se denotará por $\hat{P} = (\hat{p}_{i,j})_{M \times M}$

Según el teorema anterior, la disponibilidad Y del sistema en estudio será una CMH para cualquier α siempre que se satisfagan:

$\hat{p}_{1,1} = p^{(1)}_{1,1}p^{(2)}_{1,1} + p^{(1)}_{1,1}p^{(2)}_{1,2} + p^{(1)}_{1,2}p^{(2)}_{1,1} = p^{(1)}_{1,1}p^{(2)}_{2,1} + p^{(1)}_{1,1}p^{(2)}_{2,2} + p^{(1)}_{1,2}p^{(2)}_{2,1} = p^{(1)}_{2,1}p^{(2)}_{1,1} + p^{(1)}_{2,1}p^{(2)}_{1,2} + p^{(1)}_{2,2}p^{(2)}_{1,1}$
$\hat{p}_{1,2} = p^{(1)}_{1,2}p^{(2)}_{1,2} = p^{(1)}_{1,2}p^{(2)}_{2,2} = p^{(1)}_{2,2}p^{(2)}_{1,2}$
$\hat{p}_{2,1} = p^{(1)}_{2,1}p^{(2)}_{2,1} + p^{(1)}_{2,1}p^{(2)}_{2,2} + p^{(1)}_{2,2}p^{(2)}_{2,1}$
$\hat{p}_{2,2} = p^{(1)}_{2,2}p^{(2)}_{2,2}$

Definición 2 *Se dirá que un componente (sistema) es perfecto sii su disponibilidad puede ser modelada como una CMH cuya m.p.t tiene la forma $P = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$.*

Se dirá que un componente (sistema) es casi-perfecto sii su disponibilidad puede ser modelada como una CMH cuya m.p.t tiene la forma $P = \left(\begin{smallmatrix} 1 & 0 \\ a & 1-a \end{smallmatrix}\right)$ para algún $a \in [0, 1]$.

Note que un componente (sistema) perfecto posee la siguiente característica: Primero, si el componente ya está operando correctamente entonces nunca fallará. Segundo, si antes del periodo hostil el componente no estaba disponible, éste corregirá inmediatamente el problema y pasará al estado disponible.

Por otro lado, un componente (sistema) casi-perfecto no posee la capacidad de autocorrección inmediata. Sólo se puede afirmar que la probabilidad que este tipo de componente (sistema) falle luego de estar disponible es nula.

Resolviendo los sistemas de ecuaciones mostrados anteriormente se demuestra que los casos donde la disponibilidad de un sistema *1-out-of-2* puede ser modelada como una CMH son los siguientes:

1. Las m.p.t. de los componentes del sistema deben poseer sus filas iguales. O sea, debe cumplirse que: $P^{(1)} = \left(\begin{smallmatrix} a & 1-a \\ a & 1-a \end{smallmatrix}\right)$ y $P^{(2)} = \left(\begin{smallmatrix} b & 1-b \\ b & 1-b \end{smallmatrix}\right)$ donde $a, b \in [0, 1]$. En esta situación, la m.p.t. de X quedaría como:

$$P = P^{(1)} \otimes P^{(2)} = \begin{pmatrix} ab & a(1-b) & (1-a)b & (1-a)(1-b) \\ ab & a(1-b) & (1-a)b & (1-a)(1-b) \\ ab & a(1-b) & (1-a)b & (1-a)(1-b) \\ ab & a(1-b) & (1-a)b & (1-a)(1-b) \end{pmatrix}$$

   y la m.p.t. de Y sería:

$$\hat{P} = \begin{pmatrix} a+b-ab & (1-a)(1-b) \\ a+b-ab & (1-a)(1-b) \end{pmatrix}$$

   El vector estacionario π de la CMH Y se puede obtener resolviendo el sistema $\pi\hat{P} = \pi$. Para la matriz $\hat{P}$ dada anteriormente, se verifica que el vector estacionario π sería: $(a + b - ab, (1 - a)(1 - b))$. Este estado sería alcanzado en el primer paso del proceso Y. Es importante mencionar que cuando la m.p.t. de una CMH posee todas sus filas iguales, dicha CMH se denomina cadena de Markov de orden cero y se caracteriza por su falta de memoria. Toda CMH de orden cero puede modelarse como un proceso de Bernoulli. En el ejemplo estudiado, la disponibilidad del sistema podría modelarse como un proceso Bernoulli con probabilidad de éxito $a + b - ab$.

2. Los dos componentes del sistema deben ser *casi-perfectos*. Es decir, la disponibilidad del sistema es una CMH si las m.p.t. de sus componentes tienen la siguiente forma: $P^{(1)} = \left(\begin{smallmatrix} 1 & 0 \\ a & 1-a \end{smallmatrix}\right)$ y $P^{(2)} = \left(\begin{smallmatrix} 1 & 0 \\ b & 1-b \end{smallmatrix}\right)$ donde $a, b \in [0,1]$. En esta situación, la m.p.t. de X quedaría como:

$$P = P^{(1)} \otimes P^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ b & 1-b & 0 & 0 \\ a & 0 & (1-a) & 0 \\ ab & a(1-b) & (1-a)b & (1-a)(1-b) \end{pmatrix}$$

y la m.p.t. de Y sería:

$$\hat{P} = \begin{pmatrix} 1 & 0 \\ a+b-ab & (1-a)(1-b) \end{pmatrix}$$

El vector estacionario π de la CMH Y se puede obtener resolviendo el sistema $\pi\hat{P} = \pi$. Para la matriz $\hat{P}$ dada anteriormente, podría ocurir que: $a + b - ab = 0$ en cuyo caso, el vector estacionario π sería $(1 - \alpha_2^{(1)}\alpha_2^{(2)}, \alpha_2^{(1)}\alpha_2^{(2)})$ que es la distribución inicial del proceso Y; pero si $a + b - ab \neq 0$, la distribución de Y_n convergirá a (1,0) cuando $n \to \infty$ pues $\alpha\hat{P}^n \to (1,0)$ cuando $n \to \infty$. En esta última situación, la convergencia hacia el estado estacionario podría no ser inmediata.

3. El sistema debe poseer al menos un componente *perfecto*. Es decir, la disponibilidad del sistema es una CMH si las m.p.t. de sus componentes tienen la siguiente forma: $P^{(1)} = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ y $P^{(2)} = \left(\begin{smallmatrix} a & 1-a \\ b & 1-b \end{smallmatrix}\right)$ donde $a, b \in [0,1]$. En esta situación, la m.p.t. de X quedaría como:

$$P = P^{(1)} \otimes P^{(2)} = \begin{pmatrix} a & 1-a & 0 & 0 \\ b & 1-b & 0 & 0 \\ a & 1-a & 0 & 0 \\ b & 1-b & 0 & 0 \end{pmatrix}$$

y la m.p.t. de Y sería:

$$\hat{P} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

La matriz $\hat{P}$ nos indica que el sistema sería un sistema perfecto: Nunca fallaría una vez que está disponible y corregiría inmediatamente cualquier problema que pudiese haber antes de ser sometido al entorno hostil. De hecho, el estado estacionario del proceso Y sería (1,0) lo que confirma que el sistema siempre estaría disponible. Es fácil verificar que se obtienen las mismas conclusiones si se asume que el componente perfecto es el segundo componente en lugar del primero.

3. CONCLUSIONES

En el presente trabajo se ha demostrado que sólo existen tres situaciones en las cuales la disponibilidad de un sistema que opera bajo la arquitectura *1-out-of-2* puede ser modelada como una CMH: Primero, cuando la disponibilidad de cada componente puede modelarse como un proceso de Bernoulli, en cuyo caso, la disponibilidad del sistema, también será un proceso de Bernoulli. Segundo, cuando el sistema posee dos componentes casi-perfectos, en cuyo caso, el sistema también será casi-perfecto. Tercero, cuando el sistema posee al menos un componente perfecto, en cuyo caso, el sistema también será perfecto.

Estos resultados podrían ser de utilidad para estudiar la disponibilidad de un sistema que opera bajo la arquitectura *1-out-of-N*.

REFERENCIAS

[1] J.G. KEMENY, AND J. L. SNELL, *Finite Markov Chains*, Springer-Verlag, 1976.

[2] G. RUBINO, AND B. SERICOLA, *On weak lumpability in Markov Chains*, J. Appl.Prob., 26 (1989), pp.446-457.

[3] G. RUBINO, AND B. SERICOLA, *A finite Characterization of Weak Lumpable Markov Processes. Part I: The Discrete Time Case*, Stoch. Proc. Appl., 38 (1991), pp. 195-204.

[4] O. KALLENBERG, *Foundations of Modern Probability*, New York: NY. Springer, 1997.

[5] L. GURVITS, AND J. LEDOUX, *Markov property for a function of a Markov chain: A linear algebra approach*, Linear Algebra and its applications, 404 (2005).

[6] JORGE R. CHÁVEZ-FUENTES, OSCAR R. GONZÁLEZ, AND W. STEVEN GRAY, *Transformations of Markov Processes in Fault Tolerant Interconnected Systems* , in Proceedings of American Control Conference (2009).

GEOMETRIC PROPERTIES OF PARTIAL LEAST SQUARES REGRESSION FOR APPLICATION TO PROCESS MONITORING

José L. Godoy[†], Jorge R. Vega[†,‡] and Jacinto L. Marchetti[†]

[†] *Instituto de Desarrollo Tecnológico para la Industria Química (CONICET-Universidad Nacional del Litoral), Güemes 3450, 3000, Santa Fe, Argentina*

[‡] *Facultad Regional Santa Fe, Universidad Tecnológica Nacional, Lavaisse 610, 3000, Santa Fe, Argentina*

{jlgodoy, jvega, jlmarch}@santafe-conicet.gov.ar, www.intec.unl.du.ar, www.frsf.utn.edu.ar

Abstract. A decomposition of the input and output variable spaces of an arbitrary stochastic process is obtained by partial least squares regression (PLSR) and their main geometric properties are derived. The proposed decomposition can be used for detecting and identifying faults or anomalies in complex processes due to the existing relationships between statistics on each determined measurement subspace and the fault type.

Key words: PLSR, space decomposition, multivariate process monitoring, fault detection indexes.
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCTION

Input / output variable correlations present in data collected from complex processes can be handled with a partial least square regression (PLSR) method. PLSR allows the determination of multivariable linear structures, while collinearities implicit in the data can automatically be overcome. Many multivariate process monitoring systems are based on a PLSR model that represents 'in-control' conditions. In such cases, a meaningful deviation of the variables from their expected trajectories serves for the detection and diagnosis of abnormal process behaviors. These data-driven methods use historic data (collected during normal operating conditions) to develop a latent-variables model able to effectively explain the common-cause variability of the input and output measurements. Monitoring of a stochastic process that includes collinear input and output variables can be performed through a model based on principal component analysis (PCA), which treats the data without differentiating outputs from inputs. In contrast, a PLSR model is closer to the intrinsic system structure because it allows the elimination of some undesired input variables from the original data sets (e.g., those interfering the regression model) [1].

Given a predictor matrix $\mathbf{X}=[\mathbf{x}_1 \ldots \mathbf{x}_N]'$ ($N\times m$) consisting of N samples with m variables per sample, and a response matrix $\mathbf{Y}=[\mathbf{y}_1 \ldots \mathbf{y}_N]'$ ($N\times p$) with p variables per sample, PLSR can be used to find a regression model between the measurement vectors $\mathbf{x}=[x_1 \ldots x_m]'$ and $\mathbf{y}=[y_1 \ldots y_p]'$, even when their correlation matrixes ($\mathbf{R}_\mathbf{x}$ and $\mathbf{R}_\mathbf{y}$) are both positive semi-definite (i.e. $\mathbf{X}$ and $\mathbf{Y}$ have collinear variables). The method projects $\mathbf{X}$ and $\mathbf{Y}$ onto correlated low-dimension spaces defined by a common (small) number of A latent variables. The NIPALS algorithm is normally used to perform PLSR with deflations on both data matrixes [2], with the implicit objective of finding the solution of the following problem:

$$\max_{\mathbf{w}_a,\mathbf{q}_a} \left(\mathbf{w}_a'\mathbf{X}_a'\mathbf{Y}_a\mathbf{q}_a\right) \quad s.t. \ \ \|\mathbf{w}_a\|=1, \|\mathbf{q}_a\|=1 \tag{1}$$

The classical NIPALS algorithm [2] provides an external, an internal, and a regression model. The external model decomposes $\mathbf{X}$ and $\mathbf{Y}$ in score vectors ($\mathbf{t}_a$ and $\mathbf{u}_a$), weight vectors ($\mathbf{p}_a$ and $\mathbf{q}_a$), and residual error matrices ($\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}_2$), as follows:

$$\mathbf{X}=\mathbf{T}\mathbf{P}'+\widetilde{\mathbf{X}}=\sum_{a=1}^{A}\mathbf{t}_a\mathbf{p}_a'+\widetilde{\mathbf{X}}=\hat{\mathbf{X}}+\widetilde{\mathbf{X}}, \tag{2}$$

$$\mathbf{Y}=\mathbf{U}\mathbf{Q}'+\widetilde{\mathbf{Y}}_2=\sum_{a=1}^{A}\mathbf{u}_a\mathbf{q}_a'+\widetilde{\mathbf{Y}}_2=\hat{\mathbf{Y}}^*+\widetilde{\mathbf{Y}}_2, \tag{3}$$

where $\mathbf{P}=[\mathbf{p}_1 \ldots \mathbf{p}_A]$ $(m\times A)$, $\mathbf{Q}=[\mathbf{q}_1 \ldots \mathbf{q}_A]$ $(p\times A)$, $\mathbf{T}=[\mathbf{t}_1 \ldots \mathbf{t}_A]$ $(N\times A)$, and $\mathbf{U}=[\mathbf{u}_1 \ldots \mathbf{u}_A]$ $(N\times A)$ are orthogonal by columns. Call $\mathbf{R}$ a generalized inverse of $\mathbf{P}'$ ($\mathbf{P}'\mathbf{R}=\mathbf{R}'\mathbf{P}=\mathbf{I}$); then, the prediction of $\mathbf{T}$ is directly obtained from Eq. (2), as: $\mathbf{T}=\mathbf{X}\mathbf{R}$, because the row space of $\widetilde{\mathbf{X}}$ belongs to the null-space of the linear transformation $\mathbf{R}'$, i.e. $\widetilde{\mathbf{X}}\mathbf{R}=0$ [3]. If $\mathbf{P}'=\mathbf{W}_A\mathbf{\Sigma}_A\mathbf{V}_A'$ is the compact singular value decomposition (SVD) of $\mathbf{P}'$, then $\mathbf{R}=(\mathbf{P}')^-=\mathbf{V}_A\mathbf{\Sigma}_A^{-1}\mathbf{W}_A'$ where $^-$ denote a generalized inverse [3]. Equivalently, from Eq. (3), $\mathbf{S}$ is a generalized inverse of $\mathbf{Q}'$ ($\mathbf{Q}'\mathbf{S}=\mathbf{S}'\mathbf{Q}=\mathbf{I}$), and since $\widetilde{\mathbf{Y}}_2\mathbf{S}=0$, then: $\mathbf{U}=\mathbf{Y}\mathbf{S}$.

For the internal model, $\mathbf{t}_a$ is linearly regressed against the y-score vector $\mathbf{u}_a$ (see Table 1), i.e.:

$$\mathbf{U} = \mathbf{TB} + \mathbf{H} = \hat{\mathbf{U}} + \mathbf{H}, \quad \mathbf{B} = diag(b_1 \ldots b_A) \tag{4}$$

where $b_1 \ldots b_A$ are the regression coefficients determined by minimization of the residual matrix $\mathbf{H}$. Then, the following **X**-**Y** regression model is obtained:

$$\mathbf{Y} = \mathbf{XRBQ}' + \mathbf{HQ}' + \widetilde{\mathbf{Y}}_2 = \widehat{\mathbf{Y}} + \widetilde{\mathbf{Y}}_1 + \widetilde{\mathbf{Y}}_2 \tag{5}$$

where $\widetilde{\mathbf{Y}}_2 = \mathbf{Y} - \mathbf{YSQ}'$ and $\widetilde{\mathbf{Y}}_1 = \mathbf{YSQ}' - \widehat{\mathbf{Y}}$. Also, $\widehat{\mathbf{X}} = \mathbf{XRP}'$ verifies:

$$\widehat{\mathbf{X}}\mathbf{RBQ}' = \mathbf{XRBQ}' = \widehat{\mathbf{Y}} \tag{6}$$

The A selection is determined by supervising the simultaneous deflation of $\mathbf{X}_a$ and $\mathbf{Y}_a$ to preclude interferences. Hence, though the X-deflated PLS-NIPALS algorithm is more appropriate for process monitoring [4], we propose an equivalent algorithm but with deflations of both matrices (see Table 1).

Table 1. (**X**, **Y**)-deflated PLS-NIPALS algorithm

Center the columns of **X**, **Y** to zero mean and scale them to unit variance. Set a=1 and $\mathbf{X}_1 = \mathbf{X}$, $\mathbf{Y}_1 = \mathbf{Y}$.

1. Set $\mathbf{u}_a$ and $\mathbf{t}_a^0$ equal to the maximum-variance column of $\mathbf{Y}_a$ and $\mathbf{X}_a$, respectively.
2. $\mathbf{w}_a = \mathbf{X}_a'\mathbf{u}_a / \left\|\mathbf{X}_a'\mathbf{u}_a\right\|$, $\left(\left\|\mathbf{w}_a\right\| = 1\right)$
3. $\mathbf{t}_a = \mathbf{X}_a'\mathbf{w}_a$,
4. $\mathbf{q}_a = \mathbf{Y}_a'\mathbf{t}_a / \left\|\mathbf{Y}_a'\mathbf{t}_a\right\|$, $\left(\left\|\mathbf{q}_a\right\| = 1\right)$
5. $\mathbf{u}_a = \mathbf{Y}_a'\mathbf{q}_a$, if $\left\|\mathbf{t}_a^0 - \mathbf{t}_a\right\| < \varepsilon$, go to step 6, else set $\mathbf{t}_a^0 = \mathbf{t}_a$ and return to step 2.
6. $\mathbf{p}_a^* = \mathbf{X}_a'\mathbf{t}_a / \left(\mathbf{t}_a'\mathbf{t}_a\right)$,
7. $\mathbf{p}_a = \mathbf{p}_a^* / \left\|\mathbf{p}_a^*\right\|$, $\mathbf{t}_a = \mathbf{t}_a \left\|\mathbf{p}_a^*\right\|$, $\mathbf{w}_a = \mathbf{w}_a \left\|\mathbf{p}_a^*\right\|$, $\left(\left\|\mathbf{p}_a\right\| = 1, \ \left\|\mathbf{w}_a\right\| \neq 1\right)$
8. $b_a = \mathbf{u}_a'\mathbf{t}_a / \left(\mathbf{t}_a'\mathbf{t}_a\right)$, (Inner regression)
9. $\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_a\mathbf{p}_a'$, $\mathbf{Y}_{a+1} = \mathbf{Y}_a - b_a\mathbf{t}_a\mathbf{q}_a'$, (Deflations)

Set a=a+1 and return to step 1. Stop when a>A.

2. PLS DECOMPOSITION OF THE INPUT AND OUTPUT SPACES

After synthesizing an *in-control* PLSR model, the measurement vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^p$ can be decomposed as described below.

Lemma 1. *Call* $\mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp}$ *(*$\mathbf{\Pi}_{\mathbf{Q}|\mathbf{S}^\perp}$*) the projector on the model subspace* $Span\{\mathbf{P}\} \subseteq \mathbb{R}^m$ *(*$Span\{\mathbf{Q}\} \subseteq \mathbb{R}^p$*), along the residual subspace* $Span\{\mathbf{R}\}^\perp$ *(*$Span\{\mathbf{S}\}^\perp$*). Then:*

$$\mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp} = \mathbf{PR}', \quad \mathbf{\Pi}_{\mathbf{R}^\perp|\mathbf{P}} = \mathbf{I} - \mathbf{PR}', \tag{7}$$

$$\mathbf{\Pi}_{\mathbf{Q}|\mathbf{S}^\perp} = \mathbf{QS}', \quad \mathbf{\Pi}_{\mathbf{S}^\perp|\mathbf{Q}} = \mathbf{I} - \mathbf{QS}', \tag{8}$$

where $^\perp$ *denotes the orthogonal complement of the subspace.*

Proof. The oblique projector onto $Span\{\mathbf{A}\}$ along $Span\{\mathbf{B}\}$ can be obtained by the following equation [3]:

$$\mathbf{\Pi}_{\mathbf{A}|\mathbf{B}} = \mathbf{A}\left(\mathbf{A}'\mathbf{\Pi}_{\mathbf{B}}^\perp\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{\Pi}_{\mathbf{B}}^\perp \tag{9}$$

where $\mathbf{\Pi}_{\mathbf{B}}^\perp$ is the orthogonal projector onto $Span\{\mathbf{B}\}^\perp$. Since $\mathbf{R}$ and $\mathbf{S}$ are full-column-ranked, then:

$$\mathbf{\Pi}_{\mathbf{R}^\perp}^\perp = \mathbf{\Pi}_{\mathbf{R}} = \mathbf{R}\left(\mathbf{R}'\mathbf{R}\right)^{-1}\mathbf{R}' = \mathbf{RR}', \tag{10}$$

$$\mathbf{\Pi}_{\mathbf{S}^\perp}^\perp = \mathbf{\Pi}_{\mathbf{S}} = \mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}' = \mathbf{SS}'. \tag{11}$$

Since $\mathbf{P}'\mathbf{R} = \mathbf{R}'\mathbf{P} = \mathbf{I}$ (or $\mathbf{Q}'\mathbf{S} = \mathbf{S}'\mathbf{Q} = \mathbf{I}$), then Eq. (9) and Eq. (10) [or Eq. (11)] yield: $\mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp} = \mathbf{PR}'$ (or $\mathbf{\Pi}_{\mathbf{Q}|\mathbf{S}^\perp} = \mathbf{QS}'$). Similarly, we have $\mathbf{\Pi}_{\mathbf{R}^\perp|\mathbf{P}} = \mathbf{I} - \mathbf{PR}'$ (or $\mathbf{\Pi}_{\mathbf{S}^\perp|\mathbf{Q}} = \mathbf{I} - \mathbf{QS}'$). □

Partially, Lemma 1 has already been proved (see [4]). Each oblique projector matrix, e.g. $\mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp}$, is idempotent (i.e. $\mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp}^2 = \mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^\perp}$), whose range is the subspace $Span\{\mathbf{P}\}$ and the null-space is the subspace $Span\{\mathbf{R}\}^\perp$. Then, in Eqs. (7, 8), each oblique projector acts as an identity matrix applied onto every vector belonging to its range. From Lemma 1, we propose the following theorem on the PLS decomposition.

Theorem. *Input and output variable spaces can be decomposed (by PLSR) in complementary oblique subspaces, with both modeled subspaces interrelated according to:*

$$\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}} \in \mathbb{R}^m,\ \hat{\mathbf{x}} = \mathbf{PR}'\mathbf{x} \in S_{MX} \equiv Span\{\mathbf{P}\},\ \tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I} - \mathbf{PR}')\mathbf{x} \in S_{RX} \equiv Span\{\mathbf{R}\}^{\perp} \quad (12)$$

$$\mathbf{y} = \hat{\mathbf{y}}^* + \tilde{\mathbf{y}}_2 \in \mathbb{R}^p,\ \hat{\mathbf{y}}^* = \mathbf{QS}'\mathbf{y} \in S_{MY} \equiv Span\{\mathbf{Q}\},\ \tilde{\mathbf{y}}_2 = \mathbf{y} - \hat{\mathbf{y}}^* = (\mathbf{I} - \mathbf{QS}')\mathbf{y} \in S_{RY} \equiv Span\{\mathbf{S}\}^{\perp} \quad (13)$$

$$\hat{\mathbf{y}}^* = \hat{\mathbf{y}} + \tilde{\mathbf{y}}_1,\ \ \hat{\mathbf{y}} = \mathbf{QBR}'\hat{\mathbf{x}} \in S_{MY},\ \tilde{\mathbf{y}}_1 = \hat{\mathbf{y}}^* - \hat{\mathbf{y}} = \mathbf{QS}'\mathbf{y} - \mathbf{QBR}'\mathbf{x} \in S_{MY} \quad (14)$$

Proof. Eqs. (12, 13) can be proved by taking into account that: (i) $Span\{\mathbf{I} - \mathbf{PR}'\} = Span\{\mathbf{R}\}^{\perp}$ and $Span\{\mathbf{I} - \mathbf{QS}'\} = Span\{\mathbf{S}\}^{\perp}$ (see Lemma 1); and (ii) the projections belong to complementary subspaces, because $rank(\mathbf{PR}'(\mathbf{I} - \mathbf{PR}')) = \dim(S_{MX}) + \dim(S_{RX}) = m$ and $rank(\mathbf{QS}'(\mathbf{I} - \mathbf{QS}')) = \dim(S_{MY}) + \dim(S_{RY}) = p$. Then, Eq. (14) is directly derived from Eq. (6). □

The **x** and **y** modeled projections are related to the latent spaces and between them, as follows:

$$\mathbf{t} = \mathbf{R}'\mathbf{x} = \mathbf{P}'\hat{\mathbf{x}},\quad \hat{\mathbf{u}} = \mathbf{Bt} = \mathbf{S}'\hat{\mathbf{y}},\quad \hat{\mathbf{x}} = \mathbf{Pt},\quad \hat{\mathbf{y}} = \mathbf{Q}\hat{\mathbf{u}}, \quad (15)$$

where $\mathbf{t} = [t_1 \cdots t_A]'$ and $\hat{\mathbf{u}} = [\hat{u}_1 \cdots \hat{u}_A]'$ are the latent coordinate vectors on the model hyper planes. Hence, their correlation matrixes are related through:

$$\mathbf{\Lambda} = (N-1)^{-1}\mathbf{T}'\mathbf{T} = diag(\lambda_1 \ldots \lambda_A),\qquad \mathbf{R}_{\hat{\mathbf{x}}} = (N-1)^{-1}\hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{P\Lambda P}' \quad (16a)$$

$$\mathbf{\Delta} = (N-1)^{-1}\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{B\Lambda B} = diag(\delta_1 \ldots \delta_A),\qquad \mathbf{R}_{\hat{\mathbf{y}}} = (N-1)^{-1}\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Q\Delta Q}' \quad (16b)$$

where λ_i and δ_i are the estimated variances of t_i and $\hat{u}_i$, respectively. Eqs. (15, 16) are directly derived from Eqs. (2, 3, 4, 11, 14).

For nonzero subspaces $S_{MX}, S_{RX} \subseteq \mathbb{R}^m$, the minimal angle θ_x between S_{MX} and S_{RX} is the number $0 \le \theta_x \le \pi/2$ that satisfies [5]: $\cos\theta_x = \max\left(\hat{\mathbf{x}}'\tilde{\mathbf{x}} / \|\hat{\mathbf{x}}\|\|\tilde{\mathbf{x}}\|\right)$, where $\hat{\mathbf{x}} \in S_{MX}, \tilde{\mathbf{x}} \in S_{RX}$. Notice that $\theta_x = 0 \Leftrightarrow S_{MX} \cap S_{RX} \neq 0$, and $\theta_x = \pi/2 \Leftrightarrow S_{MX} \perp S_{RX}$. The minimal angle θ_x (θ_y) between S_{MX} (S_{MY}) and S_{RX} (S_{RY}) satisfies [5]:

$$\sin\theta_x = 1 \Big/ \left\| \mathbf{\Pi}_{\mathbf{P}|\mathbf{R}^{\perp}} \right\|_2,\qquad \sin\theta_y = 1 \Big/ \left\| \mathbf{\Pi}_{\mathbf{Q}|\mathbf{S}^{\perp}} \right\|_2 \quad (17)$$

As a consequence from [4], it results that the closer θ_x/θ_y to 90°, the closer the **X**/**Y**-PLS decomposition (Eq. 12 or 13) to an equivalent **X**/**Y**-PCA decomposition. Figure 1 illustrates all the geometric properties mentioned in this work and the control limits on each input and output subspace (except for $\tilde{\mathbf{y}}_1$). Each measurement vector is decomposed and their projections are compared with the limits (see below).

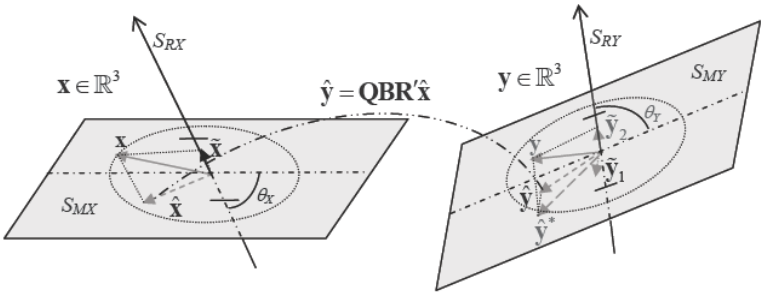


Figure 1: Induced PLS-decomposition with their relations and control limits.

3. FAULT DETECTION INDEXES BASED ON PLSR

The multivariate process monitoring strategy use statistical indexes on each subspace for fault detection. Based on the *in-control* PLSR model, we can analyze any future process behavior by projecting the new **x** and **y** measurements onto each subspace. Thus, for detecting a significant change in S_{MX}, the following Hotelling's T^2 statistic for **t** is defined:

$$T_t^2 = \left\| \mathbf{\Lambda}^{-1/2}\mathbf{t} \right\|^2 = \left\| \mathbf{\Lambda}^{-1/2}\mathbf{R}'\mathbf{x} \right\|^2 = \left\| \mathbf{\Lambda}^{-1/2}\mathbf{P}'\hat{\mathbf{x}} \right\|^2 \quad (18)$$

When a new special event (originally not considered by the in-control PLS model) is present in the process, the new observation $\mathbf{x}$ will move out of S_{MX}, i.e. in S_{RX}. The square error of prediction of $\mathbf{x}$ (SPE_X), or distance from $\mathbf{x}$-model, defined as:

$$SPE_X = \left\|\tilde{\mathbf{x}}\right\|^2 = \left\|\mathbf{\Pi}_{\mathbf{R}^\perp|\mathbf{P}}\mathbf{x}\right\|^2 \tag{19}$$

is used for detecting a change in S_{RX}. Similarly, the T^2 statistic for $\hat{\mathbf{u}}$, given by:

$$T_u^2 = \left\|\mathbf{\Delta}^{-1/2}\hat{\mathbf{u}}\right\|^2 = \left\|\mathbf{\Delta}^{-1/2}\mathbf{S}'\hat{\mathbf{y}}\right\|^2 \tag{20}$$

is used for detecting a change in S_{MY}.The distance from the regression model at S_{MY} is defined as:

$$SPE_{Y1} = \left\|\tilde{\mathbf{y}}_1\right\|^2 = \left\|\mathbf{QS}'\mathbf{y} - \mathbf{QBR}'\mathbf{x}\right\|^2 = \left\|\mathbf{Q}\left(\mathbf{u} - \hat{\mathbf{u}}\right)\right\|^2 \tag{21}$$

and the distance from the $\mathbf{y}$-model at S_{RY} is defined as:

$$SPE_{Y2} = \left\|\tilde{\mathbf{y}}_2\right\|^2 = \left\|\left(\mathbf{I} - \mathbf{QS}'\right)\mathbf{y}\right\|^2 \tag{22}$$

Frequently, $\mathbf{R}_{\hat{\mathbf{x}}}$ and $\mathbf{R}_{\hat{\mathbf{y}}}$ are singulars. Then, the generalized Mahalanobis' distance [6] for $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are:

$$D_{\hat{\mathbf{x}}} = \hat{\mathbf{x}}'\mathbf{R}_{\hat{\mathbf{x}}}^{-}\hat{\mathbf{x}} \qquad \text{and} \qquad D_{\hat{\mathbf{y}}} = \hat{\mathbf{y}}'\mathbf{R}_{\hat{\mathbf{y}}}^{-}\hat{\mathbf{y}}\ . \tag{23}$$

Proposition 1. *The metrics on* $\hat{\mathbf{x}}$, $\mathbf{t}$, $\hat{\mathbf{u}}$ *or* $\hat{\mathbf{y}}$ *are equivalents, i.e.:* $D_{\hat{\mathbf{x}}} = T_t^2 = T_u^2 = D_{\hat{\mathbf{y}}}$.

Proof. Since $\mathbf{Q}$ is orthonormal by columns, the property of the generalized inverse of a SVD yields: $\mathbf{R}_{\hat{\mathbf{y}}}^{-} = \left(\mathbf{Q\Delta Q}'\right)^{-} = \mathbf{Q\Delta}^{-1}\mathbf{Q}'$. Then, from Eq. (23): $D_{\hat{\mathbf{y}}} = \hat{\mathbf{y}}'\mathbf{R}_{\hat{\mathbf{y}}}^{-}\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{Q\Delta}^{-1}\mathbf{Q}'\hat{\mathbf{y}} = \hat{\mathbf{u}}'\mathbf{\Delta}^{-1}\hat{\mathbf{u}} = T_u^2$. Equivalently, by replacing $\mathbf{R}_{\hat{\mathbf{x}}}^{-} = \left(\mathbf{P\Lambda P}'\right)^{-} = \mathbf{P\Lambda}^{-1}\mathbf{P}'$ into Eq. (23), the distance results: $D_{\hat{\mathbf{x}}} = \hat{\mathbf{x}}'\mathbf{R}_{\hat{\mathbf{x}}}^{-}\hat{\mathbf{x}} = \hat{\mathbf{x}}'\mathbf{P\Lambda}^{-1}\mathbf{P}'\hat{\mathbf{x}} = T_t^2$. Furthermore, by combining Eqs. (15, 16, 20), $T_u^2 = \mathbf{t}'\mathbf{BB}^{-1}\mathbf{\Lambda}^{-1}\mathbf{B}^{-1}\mathbf{Bt} = T_t^2$ is obtained. From all these equalities, Proposition 1 is proven. □

4. APPLICATION AND FINAL DISCUSSION

Based on Proposition 1, the process output can be monitored through a PLSR-based statistic. The proposed strategy uses four non-overlapped metrics that completely cover the whole measurement space. Figure 2 shows the relation between variables and the location of each metric (dotted circles). The first two rows in Table 2 feature localized faults while the following cases represent process changes (or complex anomalies). By decomposing SPE_X (or SPE_Y) in their contributions, it is possible to discriminate between an external model change and a sensor fault. In fact, a process change (or an anomaly) involves a deviation of the current correlations from the model, thus increasing the value of a metric.

In summary, the proposed monitoring strategy is based on an input and output space PLS decomposition, which classifies the type of process fault or anomaly according to the statistic that triggers the alarm condition.

Table 2. Fault diagnosis based on alarmed index

Fault/Change	SPE_X	T_t^2	SPE_{Y1}	SPE_{Y2}
Sensor fault in **x**	×			
Sensor fault in **y**				×
X-external model	×			
Internal model			×	
Y-external model				×
cov($\hat{\mathbf{x}}$)		×		

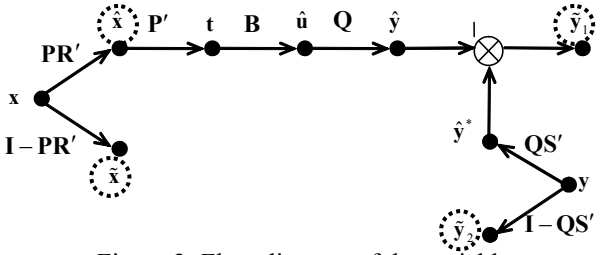


Figure 2: Flow diagram of the variables

REFERENCES

[1] S. WOLD, M. SJÖSTRÖM, AND L. ERIKSSON. Chem. Intell. Lab. Syst. Vol. 58 (2001), pp. 109-130.

[2] P. GELADI AND B. KOWALSKI. *Partial least-squares regression: A tutorial*. Anal. Chim. Acta, 185 (1986), pp. 1-17.

[3] C.D MEYER. *Matrix analysis and applied linear algebra*. SIAM, USA, 2000.

[4] L.G. GANG, S.J. QIN AND D. ZHOU. *Geometric properties of partial least squares for process monitoring.* Automatica, 46 (2010), pp. 204-210.

[5] I. IPSEN AND C. MEYER. *The angle between complementary subspaces.* Am. Math. Month, 102 (1995), pp. 904–911.

[6] MARDIA K. *Mahalanobis distances and angles. In Multivariate Analysis IV;* Krishnaiah P. Ed., North-Holland: Amsterdam, The Netherlands, pp. 495-511, 1977.

MEJORA EN LA PRECISIÓN DE MEDICIÓN DEL PARÁMETRO ÓPTICO DENOMINADO PMD MEDIANTE POST-PROCESADO MATEMÁTICO

Marcelo L. Gioda†, Fernando Corteggiano‡, Esteban H. Carranza†‡ y José L. Hernández‡‡

† Universidad Nacional de Río Cuarto, Argentina, mgioda@ing.unrc.edu.ar
‡ Universidad Nacional de Río Cuarto, Argentina, fcorteggiano@ing.unrc.edu.ar
†‡ Universidad Nacional de Río Cuarto, Argentina, hcarranza@ing.unrc.edu.ar
‡‡ Universidad Nacional de Río Cuarto, Argentina, jlh@ing.unrc.edu.ar

Resumen: La Dispersión por el Modo de Polarización (PMD o Polarization Mode Dispersion) es un fenómeno que afecta el desempeño de los sistemas de transmisión por fibra óptica de alta velocidad por lo que su medición correcta adquiere importancia en grandes empresas de telecomunicaciones. El modelo tradicional para la medición de la PMD mediante interferometría, se basa en la suposición que el patrón interferométrico coincide con la curva normal de Gauss. Un modelo alternativo propone promediar los patrones de interferometría para obtener dos réplicas simétricas de la función de densidad de probabilidad de la DGD (Diferencia del Retardo de Grupo) y, a partir de allí, obtener la PMD mediante una aproximación a dichas réplicas empleando la función de distribución de Maxwell. En este trabajo se propone efectuar un procedimiento de post-procesado matemático de los datos de las mediciones para aumentar la confianza en los valores de PMD.

Palabras claves: *PMD, interferometría, DGD, post-procesado*

1. INTRODUCCIÓN

La Dispersión por el Modo de Polarización (PMD) es un fenómeno que puede afectar seriamente el desempeño de los sistemas de transmisión por fibra óptica de alta velocidad. Las empresas de telecomunicaciones se ven obligadas a realizar mediciones precisas de la PMD pues un valor erróneo puede generar costos excesivos en equipos o nuevos cables.

La teoría subyacente, tradicionalmente empleada para realizar el cálculo de la PMD mediante un interferómetro, se basa en encontrar la curva de Gauss que mejor se ajusta al patrón interferométrico de intensidades ópticas al final de la fibra. El algoritmo matemático para el cálculo de la PMD, es el propuesto por la International Telecommunication Union ITU-T en la Recomendación G.650 Anexo II [1] (equivalente al FOTP-124 de la TIA [2]), que sigue al modelo tradicional. Sin embargo, los valores de las muestras que conforman el patrón interferométrico, en muchos casos se asemejan muy poco a la curva normal gaussiana. Existen publicaciones que proponen diferentes métodos y teorías [3,4], distintos al tradicional aplicados a la medición en campo de PMD.

Según una teoría alternativa, los promedios de los patrones de interferometría consisten en dos réplicas simétricas de la función de densidad de probabilidad del DGD (convolucionada con la función de autocorrelación de la fuente óptica), cuya función es una aproximación a la distribución de Maxwell, en lugar de la idealmente esperada, pero nunca vista, distribución de Gauss.

Aquí se propone post-procesar mediante un programa, desarrollado por los autores de este trabajo en Matlab©, las muestras de las mediciones en campo realizadas con un instrumento comercial y obtener ambas aproximaciones –la guassiana y la maxwelliana-, para luego comparar cuál es la que mejor se ajusta al patrón interferométrico obtenido. De esta forma es posible, mostrar que cuando la curva no se ajusta a la de Gauss (teoría tradicional) es más confiable y preciso el valor de PMD brindado por el post-procesamiento mediante una aproximación maxwelliana.

Durante el año 2005 se realizaron mediciones a nivel óptico de la red nacional de telecomunicaciones de Venezuela, obteniéndose los valores de PMD de los enlaces troncales de fibra óptica. El instrumento empleado, permitió guardar en archivos los 2243 valores de cada medición realizada. Dichos valores representan las muestras de la intensidad de la interferencia óptica en función del retardo temporal, es decir, el interferograma discretizado. Mediante un software, del mismo fabricante del instrumento, fue posible llevar los datos a un archivo de texto accesible desde otros programas. Así fue como se ha logrado acceder a los datos de las mediciones y aplicarles posteriormente algoritmos matemáticos que permitieron

comparar cuál curva –la gaussiana o la maxwelliana- se aproxima más al patrón de intensidades del interferómetro.

2. CÁLCULO DE PMD SEGÚN EL MODELO TRADICIONAL

El software de tratamiento matemático facilitó la programación de un procedimiento para acceder a los archivos de texto con los datos de cada medición y, en concordancia con la recomendación G-650 Anexo II de la ITU-T, programar un algoritmo con el fin de obtener el valor de PMD de acuerdo con el modelo tradicional.

Según la teoría en la que se basa el modelo tradicional, el patrón interferométrico debería coincidir con una curva de Gauss. Mayor coincidencia habrá mientras más aleatorio y cercano a infinito sea el acoplamiento de modos, menor la distancia de acoplamiento y mayor la longitud de la fibra. En ese caso, se puede calcular el valor de PMD como:

$$\Delta\tau = \sqrt{\frac{3}{4}}\sigma$$

Donde σ^2 es el segundo momento de la función gaussiana $e^{\frac{(t-C)^2}{2\sigma^2}}$ siendo t el adelanto/retardo temporal y C el centro de la curva de Gauss.

Una manera de encontrar σ es por aproximaciones sucesivas partiendo del conocimiento de σ_ε^2, que es el segundo momento del interferograma truncado. El interferograma se denomina truncado, cuando se ha eliminado el pico de autocorrelación, pues este pico no debe tenerse en cuenta para el cálculo de la PMD. La relación entre σ y σ_ε está dada por la siguiente aproximación:

$$\sigma_E = \frac{1}{2}\left\{\sqrt{\frac{\sum_{t=tjmin}^{tjl}(t-C)^2.e^{\frac{(t-C)^2}{2\sigma^2}}}{\sum_{t=tjmin}^{tjl}e^{\frac{(t-C)^2}{2\sigma^2}}}} + \sqrt{\frac{\sum_{t=tjr}^{tjmax}(t-C)^2.e^{\frac{(t-C)^2}{2\sigma^2}}}{\sum_{t=tjr}^{tjmax}e^{\frac{(t-C)^2}{2\sigma^2}}}}\right\}$$

Siendo *tjl* el límite temporal izquierdo del pico de autocorrelación; *tjr* el límite temporal derecho del pico de autocorrelación; *tjmin* el límite temporal a partir del cual los valores de intensidad ubicados a la izquierda están alejados más de 2 desviaciones standard de C (dentro de 2 desviaciones standard en una curva de Gauss se espera encontrar el 95% de los valores de las muestras) y *tjmax* el límite temporal a partir del cual los valores de intensidad ubicados a la derecha están alejados más de 2 desviaciones standard de *C*.

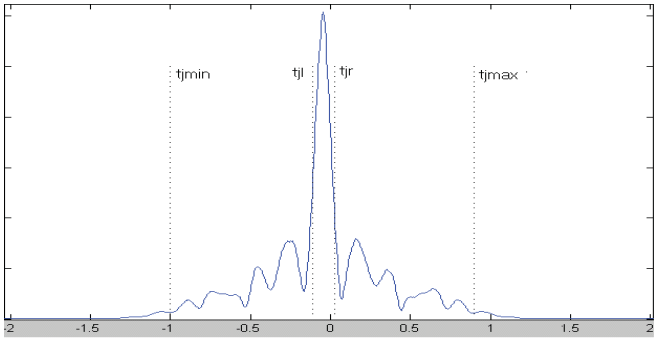


Figura 1: Patrón interferométrico con los 4 límites temporales tjmin, tjl, tjr y tjmax

Para encontrar el *tjmin* y el *tjmax* es necesario calcular entonces la desviación standard *S* así:

$$S = \frac{1}{2}\left\{\sqrt{\frac{\sum_{j=1}^{jl}(tj-C)^2 Ij}{\sum_{j=1}^{jl}Ij}} + \sqrt{\frac{\sum_{j=jr}^{N}(tj-C)^2 Ij}{\sum_{j=jr}^{N}Ij}}\right\}$$

El instrumento de medición permite exportar junto con los datos, la ubicación de los marcadores izquierdo y derecho, que determinan los valores temporales entre los cuales se encuentra el pico de autocorrelación,

asociado al tiempo de coherencia de la fuente. Los valores entre ambos marcadores no son tenidos en cuenta para el cálculo del PMD. Así: *jl* es el mayor indice *j* tal que $C\text{-}tj > \tau_c$; *jr* es el menor indice *j* tal que $tj\text{-}C > \tau_c$ con τ_c el tiempo de coherencia de la fuente.

En resumen, el proceso consiste en obtener los datos de Intensidad de cada muestra del interferograma y su correspondiente retardo/adelanto temporal, siendo *Ij* el valor de la Intensidad en el tiempo *tj* (con *j*=1...*N*) y *N* la cantidad de muestras, que para el caso del instrumento empleado para este trabajo es 2243. Encontrar el centro del interferograma *C*. Tomar la información de los marcadores izquierdos y derechos, para encontrar los instantes *tj* a ambos lados del pico de autocorrelación. Aplicar la ecuación para calcular el segundo momento S^2, y a partir de *S* los *tjmin* y *tjmax* ubicados a 2*S* del centro *C*. Calcular los segundos momentos σ_ε^2 y σ^2, del interferograma truncado (σ_ε) y de la curva de Gauss (σ) respectivamente. Finalmente, encontrar la PMD como: $\Delta\tau=\sigma\sqrt{(3/4)}$

3. CÁLCULO DE PMD MEDIANTE EL ALGORITMO DE MÁXIMA VEROSIMILITUD Y LA DISTRIBUCIÓN DE MAXWELL COMO PDF

Según la ITU G-663 [7] la relación entre la pdf con distribución de Maxwell y la PMD viene dada por:

$$P(\Delta\tau) = \frac{32.\Delta\tau_i^2}{\pi^2.<\Delta\tau>^3}.e^{\left(-\frac{4.\Delta\tau_i^2}{\pi.<\Delta\tau>^2}\right)}$$

siendo $<\Delta\tau>$ la PMD.

Operando matemáticamente sobre la ecuación anterior y aplicando el algoritmo de máxima verosimilitud, se puede escribir:

$$L[P(PMD)] = \prod_{i=1}^{N}[P(PMD)] \quad \text{donde} \quad \prod_{i=1}^{N}[P(PMD)] = \frac{32^N.\prod_{i=1}^{N}\Delta\tau_i^2}{\pi^{2N}.(PMD)^{3N}}.e^{\left(-\frac{4\sum_{i=1}^{N}\Delta\tau_i^2}{\pi.(PMD)^2}\right)}$$

Aplicando logaritmo natural a estas igualdades y derivando con respecto a PMD se tiene:

$$\frac{d(Ln\{L[P(PMD)]\})}{d(PMD)} = -\frac{3N}{PMD} + \frac{4\sum_{i=1}^{N}\Delta\tau_i^2}{\pi.(PMD)^3} = 0 \quad \Rightarrow \quad PMD = \sqrt{\frac{4\sum_{i=1}^{N}\Delta\tau_i^2}{3N\pi}}$$

Con lo cual la PMD puede ser calculada a partir de los valores de las muestras de interferometría que tienen una distribución de Maxwell (por ejemplo los valores ubicados a la derecha del pico de autocorrelación, es decir los que se encuentran entre tjr y tjmax). El modelo alternativo permite calcular un valor de la PMD diferente al tradicional, mediante la aproximación maxwelliana. Esta sólo se realiza para el lado derecho de los valores, sin tomar en cuenta el pico de autocorrelación. El lado izquierdo debe entenderse que es una réplica en espejo del lado derecho.

4. COMPARACIÓN DE RESULTADOS

Es posible comparar los valores de las muestras que forman la curva de interferometría, obtenidas por el instrumento en campo, con cada uno de los valores generados por la aproximación mediante una curva normal de Gauss (como lo lleva a cabo el método tradicional) y también con cada uno de los valores obtenidos por la aproximación mediante la distribución de Maxwell (propuesto por el nuevo modelo).

El porcentaje de error de ambas aproximaciones (gaussiana y maxwelliana) con respecto al patrón interferométrico obtenido en campo, permite apreciar cuán próximas son dichas distribuciones al patrón.

Hay que notar que ambos modelos tienen teorías distintas, que los sustentan, y que explican la relación entre dichas aproximaciones y la PMD, así como entre las curvas obtenidas en cada modelo y el patrón

interferométrico. Por lo que el análisis no es un caso trivial de encontrar la función, cualquiera que esta sea, que más se aproxime a la serie de valores obtenidos en el interferómetro, sino de encontrar el valor de PMD más verosímil, sustentado por una teoría óptica creíble, teniendo en cuenta que ninguna de las curvas (gaussiana y maxwelliana) coinciden exactamente con el patrón de interferometría.

Si bien la mayoría de los patrones coinciden en mayor medida con la curva de Gauss (ver Fig.2a), hay otros que son mejor aproximados con una curva de Maxwell (ver Fig.2b).

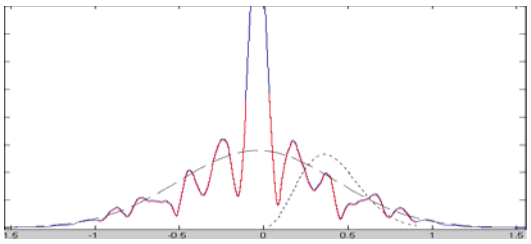


Fig.2a: Patrón con mayor coincidencia gaussiana.

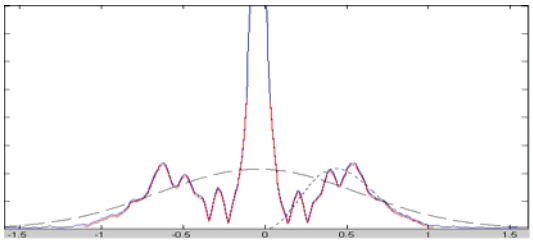


Fig.2b: Patrón con mayor coincidencia maxwelliana.

Linea de trazos largos => Aproximación gaussiana; Línea de trazos cortos => Aproximación maxwelliana

Como ejemplo, se brinda el enlace entre las localidades de Aragua de Barcelona – Anaco (Fig. 2a) la fibra medida (la 9) presentó un patrón más cercano a la curva de Gauss que a la de Maxwell, mientras que la fibra 13 del enlace Lecherias - Higuerote (Fig.2b) presentó una conformación del patrón interferométrico más próximo a la curva de Maxwell (modelo alternativo) que a la gaussiana (modelo tradicional).

El valor de PMD obtenido en el enlace Lecherías - Higuerote por el método tradicional fue de 0,5341 picosegundos mientras que el método alternativo (propuesto como más preciso en este caso, por su mayor coincidencia con la curva del interferómetro) fue de 0,4942 ps (una diferencia apreciable, mayor al 8%).

Cabe señalar que la FOTP-124, que sustenta al modelo tradicional, expresa: "La exactitud está relacionada con la capacidad de hacer coincidir precisamente el interferograma con la función gaussiana".

Una posible causa para la no coincidencia con la curva normal es que las fibras tengan un acoplamiento de modos ópticos mixto, poco aleatorio. La teoría subyacente del modelo alternativo indica que la aproximación maxwelliana, aún en ese caso, es válida. El tratamiento matemático brindado en la nueva teoría que apoya la aproximación mediante una distribución de Maxwell, es distinto para el caso de acoplamiento de modos aleatorio que determinístico. El tratamiento matemático es mucho más directo para el caso determinístico. Puede ser esa la explicación por la cual la nueva teoría se ajusta mejor a los interferogramas claramente no gaussianos, con acoplamiento de modo poco aleatorio.

5. CONCLUSIONES

Como conclusión de este estudio, se propone el empleo de un procedimiento, para post-procesar los archivos con las mediciones en campo, realizadas por los equipos que aplican el modelo tradicional para el cálculo de la PMD. El post-procesamiento permite aumentar la confianza en los valores de PMD obtenidos cuando el patrón interferométrico se aleja de la curva normal de Gauss y se aproximan a una distribución maxwelliana.

6. REFERENCIAS

[1] ITU-T REC. G-650.2 *Definición y métodos de prueba de los parámetros pertinentes de las fibras monomodo, Anexo II, Determinación del retardo de PMD a partir de un interferograma*

[2] TIA/EIA FOTP-124-A, *Polarization Mode Dispersion Measurement for Single-Mode Optical Fibers by Inteferometry.*

[3] M. ARTIGLIA, R. CAPONI, M. POTENZA, D. ROCCATO, M. SCHIANO, *Interferometer measurement of polarization mode dispersion statistics,* J. Lightwave Technology, Volume 20, Issue 8, Aug. 2002, Pag: 1374-1381

[4] N. CYR, *Polarization-Mode Dispersion Measurement: Generalization of the Interferometric Method to Any Coupling Regime,* J. Lightwave Technology, Volume 22, Issue 3, AMar. 2004, Pag: 794.

ANALISIS DE SENSIBILIDAD GLOBAL EN REDES DE BIOREACTORES

María Paz Ochoa, Patricia M. Hoch

Universidad Nacional del Sur – Departamento de Ingeniería Química – Planta Piloto de Ingeniería Química (PLAPIQUI – UNS - CONICET) – Avda. Alem 1253 - 8000 Bahía Blanca - ARGENTINA

Resumen: Este trabajo presenta un estudio de sensitividad global sobre una red de bioreactores para la producción de bioetanol a partir de azúcares de molasas y destilado de vinasas. Se consideran dos reactores en serie, el primero para la producción de biomasa. Se asocian distribuciones de probabilidad a cada uno de los parámetros inciertos y se determinan perfiles temporales para los índices de sensitividad para las principales variables diferenciales y algebraicas. Se aplica el método propuesto por Sobol' (1990) y se realizan las simulaciones estocásticas en el entorno gPROMS (PSEnterprise, 2009). Los resultados muestran la influencia de los parámetros, variable en el tiempo de operación de los fermentadores..

Palabras claves: *Análisis de sensitividad global, Redes de bioreactores, Sistemas DAE*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

Los modelos dinámicos de redes de bioreactores se formulan como un sistema de ecuaciones diferencial-algebraico, que surgen de balances de materia de sustrato, biomasa, biomasa muerta y producto. Las salidas del modelo se ven influenciadas en diferente medida por la incertidumbre en los parámetros de entrada. En este trabajo, se efectúa un análisis de sensitividad global a través de técnicas basadas en varianza para identificar los parámetros más influyentes, y cuáles de estos parámetros impactan en mayor medida en las salidas del modelo. El análisis de sensitividad global provee de información en las salidas del sistema cuando se explora simultáneamente el espacio completo de variación de los parámetros, muestreando desde la función de distribución asociada a cada parámetro de entrada y realizando repetidas simulaciones del modelo. Como ventaja adicional, los resultados son más realistas, ya que se pueden identificar interacciones entre los parámetros. El método es independiente del modelo, ya que no se requiere realizar la suposición de linealidad o aditividad (Saltelli et al., 2008)

2. ANÁLISIS DE SENSITIVIDAD GLOBAL EN FERMENTADORES

El método de Sobol' está basado en la misma descomposición de varianza que FAST, pero a través de la aplicación de los métodos de Monte Carlo en lugar del análisis espectral (Sobol´, 1990; Saltelli and Sobol´, 1995; Sobol´, 2001; Saltelli and Tarantola, 2002). La idea básica es que, dada una función y=f(x,t), donde y es la variable de estado diferencial o algebraica (por ejemplo concentración de biomasa o sustrato o velocidad de reacción), x es un vector de k parámetros de entrada del modelo (velocidad máxima de reacción, constante de saturación, etc.) y t es el tiempo, esta función se puede descomponer en términos de dimensión creciente (Sobol', 1990). Los índices de primer orden se pueden calcular como

$$S_i = \frac{V\left(E\left(y|x_i\right)\right)}{V\left(y\right)} = \frac{V_i}{V(y)} \quad \text{y} \quad S_i^{TOT} = \frac{E\left(V\left(y|x_{-i}\right)\right)}{V\left(y\right)} = \frac{V_i^{TOT}}{V(y)} \tag{1}$$

Existe un tercer índice, S_i^{int}, que muestra las interacciones entre los parámetros del modelo, definido como

$$S_i^{int} = S_i^{TOT} - S_i \tag{2}$$

Sobol´(2001) ha propuesto una metodología para calcular índices de sensitividad, basados en simulaciones Monte Carlo**,** con una cantidad mínima de evaluaciones de función. Los pasos principales para los casos dinámicos son los siguientes:

1. Generación de dos conjuntos diferentes y aleatorios de parámetros del modelo: $\xi=(\eta,\zeta)$ y $\xi'=(\eta',\zeta')$ en cada instante de tiempo. Cada matriz tiene dimensión N × k, donde N es el número de escenarios para el método Monte Carlo y k es el número de parámetros; η es un vector de dimensión N×1, que contiene N valores aleatorios para el parámetro xi cuyo índice de sensitividad se va a calcular, y ζ es una submatriz de dimensión N×(k-1) que contiene valores aleatorios para los k-1 parámetros restantes.

2- Generación de dos nuevas matrices por combinación de ξ y ξ', que se requieren para el cálculo de los valores esperados dependientes del tiempo de las variables de estado, las composiciones, (c0(t) representa E(c(t)|xi)) y las varianzas incondicionales (V(t)), en cada instante de tiempo t, así como las varianzas condicionales (V(t)i representan $V\left(E\left(c(t)|x_i\right)\right)$ y V(t)-i representan $V\left(E\left(c(t)|x_{-i}\right)\right)$):

$$c_0(t) = \frac{1}{N}\sum_{i=1}^{N} c(t,\xi_i) \tag{3}$$

$$V(t) = \frac{1}{N}\sum_{i=1}^{N} c^2(t,\xi_i) - c_0^2(t) \tag{4}$$

$$V(t)_i = \frac{1}{N}\sum_{i=1}^{N} c(t,\xi_i)c(t,\eta_i,\zeta_i') - c_0(t,\xi_i)c_0(t,\xi_i') \tag{5}$$

$$V(t)_{-i} = \frac{1}{N}\sum_{i=1}^{N} c(t,\xi_i)c(t,\eta_i',\zeta_i) - c_0^2(t) \tag{6}$$

3- Los perfiles de índices de sensitividad se calculan por las definiciones dadas previamente a lo largo de todo el horizonte de tiempo.

Los índices de sensitividad para cada parámetro se calculan siguiendo la aproximación de Sobol´ (2001), que emplea métodos de simulación Monte Carlo para el cálculo de perfiles temporales de las varianzas condicionales con respecto a los parámetros de entrada que tienen mayor impacto en el cálculo de las principales variables de estado, ya sea diferenciales o algebraicas.

El análisis de sensitividad global se efectuó en un sistema de ecuaciones diferencial-algebraico que representa una red de bioreactores batch para la producción de etanol (Corsano y col. 2004). El modelo tiene en cuenta balances de masa dinámicos para los tres bioreactores de la red, así como expresiones cinéticas y ecuaciones algebraicas adicionales para representar las conexiones entre los fermentadores. En el modelo se consideran 12 parámetros. El modelo de red de bioreactores se implementó en gPROMS (PSEnterprise, 2009). En este entorno, se generan dos conjuntos diferentes de parámetros aleatorios para k=12. Se considera un tamaño de muestra de N=1500 escenarios. Se realizaron las N(2k+1) simulaciones Monte Carlo y se calcularon las varianzas condicionales e incondicionales y los índices de sensitividad. De esta manera se calcularon los perfiles temporales de los índices de primer orden y totales para los 12 parámetros considerados.

Los índices de sensitividad calculados muestran cuáles son los parámetros del modelo que introducen mayor incertidumbre en el proceso de producción de etanol.

3. RESULTADOS NUMÉRICOS

En el presente estudio se ha aplicado Análisis de Sensitividad Global (GSA), como se describió en la sección anterior, a una red de dos fermentadores discontinuos para la producción de bioetanol a partir de azúcares de Molasas y Destilado de Vinasas. Dicha red consta de dos fermentadores en serie, el primero para producción de biomasa y el segundo para la producción de bioetanol a partir de Saccharomyces cerevisiae (Corsano et al. 2004). Se considera una concentración inicial de sustrato de 90 g/l a fin de evitar inhibición por sustrato.

El análisis de sensitividad global en un sistema diferencial algebraico a gran escala ha requerido un gran esfuerzo computacional. Como primer paso, se ha asociado una distribución de probabilidad normal a cada uno de los parámetros que se presentan en la Tabla 1, junto con su valor nominal y desviaciones estándar.

Como se puede ver en la Fig.1, para la concentración de biomasa, el índice de sensitividad más influyente es el correspondiente a la velocidad máxima de reacción, que a lo largo de todo el horizonte de tiempo

contribuye con más del cincuenta por ciento. Le sigue en importancia el índice correspondiente a la constante de saturación por sustrato.

La Fig. 2 muestra los índices de sensitividad para la concentración de sustrato; se puede apreciar que en este caso la velocidad máxima de reacción es también el parámetro más influyente. Pero a diferencia del caso anterior, su influencia es notablemente más importante a medida que se consume el sustrato. Le sigue en importancia el índice correspondiente a la constante de saturación por sustrato la concentración de azúcares de las molasas.

Tabla 1. Parámetros inciertos y sus distribuciones normales.

Parámetro	Nomenclatura	Valor nominal	Desv. Standard
Velocidad Máxima de Reacción (h^{-1})	Umax1	0.5	0.06
Constante de Saturación de Sustrato (g/l)	Ks	20	2.5
Velocidad de Muerte de Biomasa (h^{-1})	vdead	0.02	4.00E-003
Volumen fermentador 1(m^3)	V1	3.06	0.35
Concentración de azúcares en Vinasas (g/l)	SDV	10	0.2
Concentración de azúcares en Molasas(g/l)	Smolass	779	100

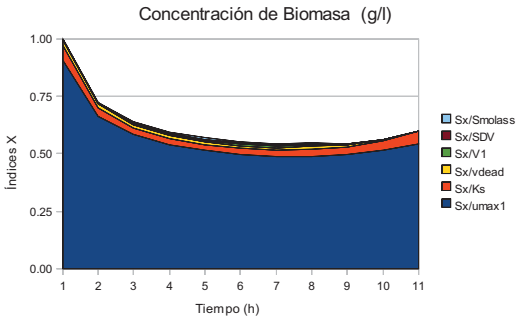


Figura 1. Índices de sensitividad de primer orden para la concentración de biomasa en el primer fermentador.

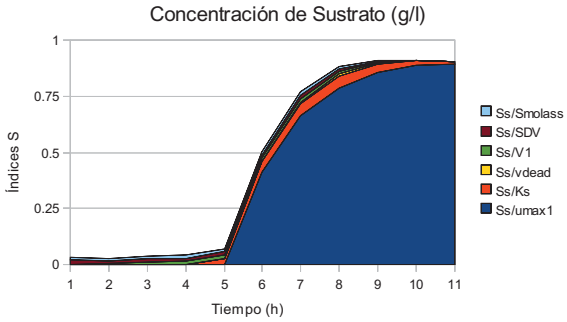


Figura 2. Índices de sensitividad de primer orden para la concentración de sustrato en el primer fermentador.

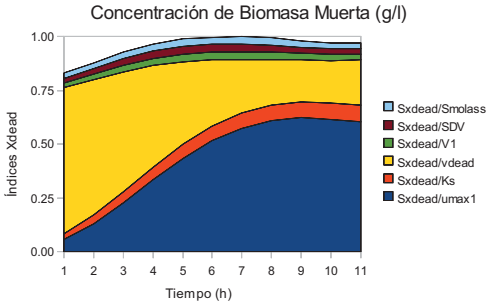


Figura 3. Índices de sensitividad de primer orden para la concentración de biomasa muerta en el primer fermentador.

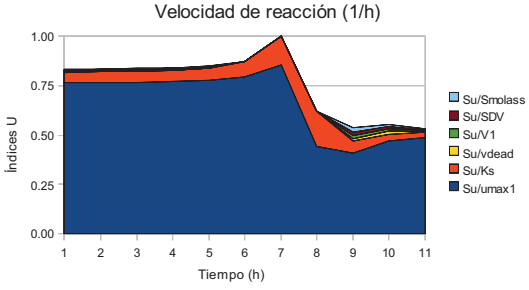


Figura 4. Índices de sensitividad de primer orden para la velocidad de reacción en el primer fermentador.

Como se puede apreciar en la Fig.3, que muestra los índices de sensitividad para la concentración de biomasa muerta, los parámetros dominantes a lo largo de todo el horizonte de tiempo son: la velocidad máxima de reacción, cuya influencia aumenta a lo largo del tiempo, a medida que avanza la reacción; y la velocidad de muerte de biomasa, cuya tendencia es opuesta.

Las Figs. 4 y 5 muestran los índices de sensitividad de primer orden para variables algebraicas del modelo: la velocidad de reacción en el primer fermentador y el coeficiente de rendimiento de biomasa en el fermentador 1. En el primer caso se puede observar que el mayor aporte está dado por el índice de sensitividad de la velocidad máxima de reacción, seguido nuevamente por el índice de la constante de saturación por sustrato. En cambio, en la Fig.5 se puede notar que los índices no varían con el tiempo, y que el mayor aporte está dado por la concentración de azúcares en el destilado de vinasas.

Las simulaciones estocásticas se llevaron a cabo en gPROMS (PSEnterprise, 2009), así como el cálculo de las varianzas condicionales e incondicionales y de los índices de sensitividad para cada instante de tiempo para un horizonte de 10.5 horas en el primer fermentador.

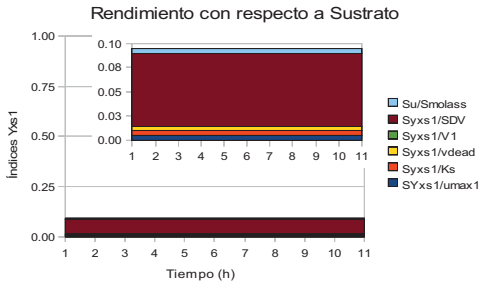


Figura 5. Índices de sensitividad de primer orden para el rendimiento de biomasa en el primer fermentador.

4. CONCLUSIONES

Se ha presentado un estudio de sensitividad global aplicado a fermentadores para la producción de bioetanol a partir de azúcares de molasas y destilado de vinasas. Se utilizó el método propuesto por Sobol´ (1990), en el que se obtienen los índices de sensitividad de primer orden y totales mediante simulaciones Monte Carlo. Se presentaron los perfiles temporales de los índices de sensitividad de primer orden para las variables de estado diferenciales y las principales algebraicas. Se puede observar que el parámetro más influyente en la mayor parte de los casos es la velocidad máxima de reacción, a lo largo de todo el horizonte de tiempo que corresponde con el período de operación del reactor. Los resultados obtenidos permiten realizar una clasificación de los parámetros más influyentes, para su posterior estimación basada en datos experimentales.

REFERENCIAS

[1] G. CORSANO, P.A. AGUIRRE, O.A. IRIBARREN, J.M. MONTAGNA, 2004, Batch Fermentation Networks Model for Optimal Synthesis, Design and Operation, Ind. Eng. Chem. Res. 43, 4211-4219.

[2] J. DI MAGGIO, J.C. DIAZ RICCI, M.S. DIAZ, 2010, Global Sensitivity Analysis in dynamic metabolic Networks, Computers and Chemical Engineering, 34, 2010, 770–781.

[3] PSENTERPRISE, gPROMS Introductory User guide, 2009, Process Systems Enterprise Ltd., London.

[4] A. SALTELLI, I.M. SOBOL', 1995, About the use of rank transformation in sensitivity analysis of model output. Reliability Engineering and System Safety 50 225-239.

[5] A. SALTELLI, S. TARANTOLA, 2002, On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. Journal of the American Statistical Association. 97 702–709.

[6] I. M. SOBOL', 1990, Sensitivity estimates for nonlinear mathematical models. Matematicheskoe Modelirovanie 2 112-118 in Russian, translated in English in Sobol', 1993, Mathematical Modelling and Computational Experiment 1 407-414.

[7] I.M. SOBOL', 2001, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and Computers in Simulation 55, 271-280.

TOWARDS A FOKKER-PLANCK DESCRIPTION OF SOME NON-MARKOV PROCESSES

Horacio S. Wio[♭], J. Ignacio Deza[†] and Roberto R. Deza[‡]

[♭]*IFCA (UC and CSIC), Avda. de los Castros, s/n, E-39005 Santander, Spain, wio@ifca.unican.es, www.ifca.unican.es/users/wio*
[†]*Dto. de Física, FCEyN-UNMdP, Deán Funes 3350, B7602AYL Mar del Plata, Argentina, juanignaciodeza@gmail.com*
[‡]*IFIMAR (UNMdP and CONICET), Deán Funes 3350, B7602AYL Mar del Plata, Argentina, deza@mdp.edu.ar, fisica.mdp.edu.ar/CV/rdeza/personal*

Abstract: Most of the advance in the detailed description of stochastic processes of interest to physics and applied mathematics has taken place on the shoulders of two giants: the central limit theorem and the Markov property. We are still lacking systematic analytical tools to deal with the host of highly interesting processes (e.g. financial timeseries) which are neither Gaussian nor Markovian. For a class of non-Markovian (and even non-Gaussian) stochastic processes which obey Langevin-like equations, we describe here a procedure to retrieve—by means of functional integration over their "phase space"—a consistent Fokker-Planck description which keeps nonetheless (through parameters) useful information on their true character. The method has been applied to systems submitted to colored-noise sources whose statistics can depart from Gaussian through a parameter, and is being presently applied to systems with inertia.

Keywords: *Non Markov process, Onsager–Machlup functional, Wiener path integral*
2000 AMS Subject Classification: 60J65 - 82C35 - 58D30

1 INTRODUCTION

1.1 THE WIENER PROCESS

The notion of functional (or "path") integration finds nowadays widespread application in the quantum realm—especially in quantum field theory—being associated to R. Feynman [1, 2]. Relatively few people realize that it was born almost thirty years before—even 12 years before the paper by Dirac [3] which inspired Feynman—within the context of stochastic processes, more precisely as a representation of a Wiener process $W(t)$ [4, 5, 6]. In fact, the Gaussian and Markovian character of $W(t)$ allows to rewrite its conditional pdf

$$P(W,t|W_0,t_0) = [2\pi D(t-t_0)]^{-1/2} \exp\left\{-\frac{[W-W_0]^2}{2D(t-t_0)}\right\} \tag{1}$$

in the form

$$P(W,t|W_0,t_0) = \int \mathcal{D}[W(\tau)] \exp\left[-\frac{1}{4D}\int_{t_0}^{t} d\tau \left(\frac{dW}{d\tau}\right)^2\right]. \tag{2}$$

D is the intensity of the (wild) stationary process $\xi(t)$ (called "white noise") of the increments of $W(t)$

$$\int_t^{t+\delta t} dt'\xi(t') = \Delta W(t), \quad \langle\xi(t)\rangle = 0, \quad \langle\xi(t)\xi(t')\rangle = 2D\,\delta(t-t'),$$

and

$$\mathcal{D}[W(\tau)] = \lim_{N\to\infty} \prod_{j=1}^{N} \frac{dW_j}{(4\pi\epsilon D)^{1/2}} \tag{3}$$

(with $\epsilon \equiv (t-t_0)/N \to 0$) is known as *the Wiener measure*. In fact, only by resorting to this context (through a Wick rotation) can a meaningful Feynman measure be defined [7].

1.2 THE ONSAGER–MACHLUP FUNCTIONAL

A Langevin-like equation such as $\dot{q} = f(q)+\xi(t)$ establishes (a monoparametric family of discretization-dependent) piecewise uniform mappings between the Markov processes $q(t)$ and $W(t)$, through which a path-integral representation can be written down for the conditional pdf $P(q,t|q_0,t_0)$ [8, 9, 10]

$$P(q,t|q_0,t_0) = \int \mathcal{D}[q(t)] \exp\left(\int_{t_0}^{t} ds \mathcal{L}[q(s),\dot{q}(s)]\right), \tag{4}$$

where

$$\mathcal{S}[q(t)] = \int_{t_0}^{t} ds \mathcal{L}[q(s),\dot{q}(s)] \tag{5}$$

is the *stochastic action*,

$$\mathcal{L}[q(s),\dot{q}(s)] = \frac{1}{4D}\left(\dot{q}(s) + f[q(s)]\right)^2 - \alpha \frac{df[q(s)]}{dq} \tag{6}$$

the *stochastic Lagrangian* (also called the *Onsager–Machlup functional*), and α the parameter of the above-mentioned family.

2 RETRIEVING A FOKKER–PLANCK EQUATION FOR NON-MARKOVIAN PROCESSES

The *Fokker–Planck equation* (FPE) obeyed by $P(q,t|q_0,t_0)$ [10],

$$\frac{\partial}{\partial t}P(q,t|q_0,t_0) = -\frac{\partial}{\partial q}\left[f(q)P(q,t|q_0,t_0)\right] + D\frac{\partial^2}{\partial q^2}P(q,t|q_0,t_0), \tag{7}$$

provides a handy and systematic tool for dealing with Markov processes. Instead, if either $\langle \xi(t)\xi(t')\rangle = 2D\,\delta(t-t')$ is not fulfilled ("colored noise" case) or the system is not extremely overdamped (so a Langevin-like equation cannot be written out), the process $q(t)$ becomes non-Markovian.

2.1 SYSTEMS DRIVEN BY COLORED NOISE

Several efforts have been addressed in the past towards retrieving a meaningful Fokker–Planck description of the "colored noise" case: from more phenomenologic ones, like the "unified colored-noise approximation" (UCNA) [11] or interpolation schemes [12] to systematic ones [13], based on path integration over the process' "phase space" [14]. The latter scheme has been successfully applied [15] even to processes driven by colored noises whose statistics can depart from Gaussian through a parameter [16].

2.2 SYSTEMS WITH INERTIA

Our main task in this paper is to illustrate the application of the method devised in Refs. [13] and [15] to systems with inertia. The philosophy behind the method follows closely the Kramers–Smoluchowski issue [17], namely that some processes which look non-Markovian in too few dimensions can be "unfolded" to suitably higher-dimensional processes which are indeed Markovian. We take as starting point the coupled system of stochastic differential equations (in the notation of Ref. [18])

$$\dot{x} = v(t), \tag{8}$$

$$\dot{v} = -\lambda v(t) - U'(x(t)) + \eta(t), \tag{9}$$

$$\dot{\eta} = \gamma[-V_q'(\eta(t)) + \xi(t)], \tag{10}$$

with $\langle \xi(t)\rangle = 0$ and $\langle \xi(t)\xi(t')\rangle = 2D\delta(t-t')$ as before. The potential of the drift force can be e.g. $U(x) = \frac{1}{2}ax^2 + \frac{1}{4}bx^4$, with $b > 0$.

We seek a path-integral representation of the transition probability $p \equiv P_{D,\gamma,q}(x_f, v_f, \eta_f, t_f \mid x_0, v_0, \eta_0, t_0)$, which obeys the following FPE:

$$\frac{\partial p}{\partial t} = \overbrace{-\frac{\partial}{\partial x}[v(t)p] - \frac{\partial}{\partial v}\{[-\lambda v(t) - U'(x(t))}^{L_a} + \overbrace{\eta(t)]p\}}^{L_i} + \overbrace{-\gamma\frac{\partial}{\partial \eta}[-V_q'(\eta(t))p] + D\gamma^2\frac{\partial^2}{\partial \eta^2}p}^{L_b}, \tag{11}$$

according to [18]. Now, the difficulties associated with the fact that this FPE has singular diffusion matrix can be circumvented by performing the functional integral in the *phase space* of the variables $x(t)$, $v(t)$ and $\eta(t)$ [14, 13, 15]:

$$p = \int_{x(t_0)=x_0, v(t_0)=v_0, \eta(t_0)=\eta_0}^{x(t_f)=x_f, v(t_f)=v_f, \eta(t_f)=\eta_f} \mathcal{D}\,[x(t)]\,\mathcal{D}\,[v(t)]\,\mathcal{D}\,[\eta(t)]\,\mathcal{D}\,[p_x(t)]\,\mathcal{D}\,[p_v(t)]\,\mathcal{D}\,[p_\eta(t)]\,\mathrm{e}^{\mathcal{S}_1(D,\gamma,q)}. \tag{12}$$

Here $p_x(t)$, $p_v(t)$ and $p_\eta(t)$ are respectively the canonical momenta conjugate to $x(t)$, $v(t)$ and $\eta(t)$, and $\mathcal{S}_1(D,\gamma,q)$ is the "stochastic action"

$$\begin{aligned}\mathcal{S}_1(D,\gamma,q) &= \int_{t_0}^{t_f} \mathrm{d}s \left\{ D\gamma^2 [i p_\eta(s)]^2 + i p_\eta(s)[\dot\eta(s) + \gamma V_q'(\eta(s))] \right. \\ &\left. + i p_x(s)[\dot x(s) - v(s)] + i p_v(s)[\dot v(s) + \lambda v(s) + U'(x(s)) - \eta(s)] \right\}. \end{aligned} \tag{13}$$

Performing successively the Gaussian integrations over $p_\eta(s)$, $p_x(s)$, $v(s)$, $p_v(s)$ and $\eta(s)$, we obtain

$$\eta(s) = \ddot x(s) + \lambda \dot x(s) + U'(x(s)), \quad t_0 \le s \le t_f, \tag{14}$$

and hence

$$\dot\eta(s) = \frac{\mathrm{d}^3 x}{\mathrm{d}s^3} + \lambda \ddot x(s) + \dot x(s) U''(x(s)), \tag{15}$$

so

$$p = \int_{x(t_0)=x_0}^{x(t_f)=x_f} \mathcal{D}\,[x(t)]\,\mathrm{e}^{\mathcal{S}_4(D,\gamma,q)}, \tag{16}$$

with

$$\mathcal{S}_4(D,\gamma,q) = -\frac{1}{4D\gamma^2} \int_{t_0}^{t_f} \mathrm{d}s \left[\frac{\mathrm{d}^3 x}{\mathrm{d}s^3} + \lambda \ddot x(s) + \dot x(s) U''(x(s)) + \gamma V_q'(\ddot x(s) + \lambda \dot x(s) + U'(x(s))) \right]^2. \tag{17}$$

This stochastic action provides a *non-Markovian* description, since it involves time derivatives higher than $\dot x(s)$. Clearly, if we want a *Markovian* description we must eliminate them (as in [15]). Moreover, since we want to retrieve from the path-integral representation a FPE for $P(x,t) \equiv P_{D,\gamma,q}(x,t \mid x_0,t_0)$, we perform on the Markov-approximated action a Kramers–Moyal-like approximation, by Taylor-expanding $V_q'(U'(x(s)) + \lambda \dot x(s))$ in powers of $\dot x(s)$ and eliminating powers $\dot x^n$ with $n > 2$. This yields

$$S_{\mathrm{FP}}(D,\gamma,q) = -\frac{1}{4D\gamma^2} \int_{t_0}^{t_f} \mathrm{d}s \,[\dot x(s) U''(x(s)) + \gamma V_q'(U'(x(s))) + \lambda \dot x(s) V_q''(U'(x(s)))]^2. \tag{18}$$

Following [15], we focus on a potential of the form $V_q(\eta) = \frac{\gamma D}{1-q} \ln[1 + H(\eta)]$, with $H(\eta) = \frac{(q-1)\eta^2}{2\gamma D}$, so the process η obeys q–statistics [16] (it may have compact support or infinite variance, depending on the range of q). Since $U'(x) = ax + bx^3$, we get

$$V_q'(\eta) = -\frac{\eta}{1 + H(\eta)} \quad \Rightarrow \quad \Phi(x) \equiv -\gamma V_q'(U'(x)) = \gamma \frac{ax + bx^3}{1 + \frac{(q-1)[ax+bx^3]^2}{2\gamma D}}, \tag{19}$$

$$V_q''(\eta) = -\frac{1 - H(\eta)}{[1 + H(\eta)]^2} \quad \Rightarrow \quad V_q''(U'(x)) = -\frac{1 - \frac{(q-1)[ax+bx^3]^2}{2\gamma D}}{\left[1 + \frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]^2}, \tag{20}$$

and

$$S_{\mathrm{FP}}(D,\gamma,q) = -\frac{1}{4D\gamma^2} \int_{t_0}^{t_f} \mathrm{d}s \,[\Psi(x(s))\dot x(s) - \Phi(x(s))]^2, \tag{21}$$

with

$$\Psi(x) \equiv ax + bx^3 - \gamma\lambda \frac{1 - \frac{(q-1)[ax+bx^3]^2}{2\gamma D}}{\left[1 + \frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]^2}. \tag{22}$$

3 THE FOKKER–PLANCK EQUATION

The FPE is then [15]

$$\partial_t P(x,t) = -\partial_x[A(x)P(x,t)] + \frac{1}{2}\partial_x^2[B(x)P(x,t)], \tag{23}$$

with

$$A(x) \equiv \frac{\Phi(x)}{\Psi(x)} = \frac{\gamma(ax+bx^3)\left[1+\frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]}{\gamma(ax+bx^3)\left[1+\frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]^2 - \gamma\lambda\left[1-\frac{(q-1)[ax(s)+bx(s)^3]^2}{2\gamma D}\right]} \tag{24}$$

and

$$B(x) \equiv \frac{D}{\Psi(x)^2} = D\left\{\frac{\left[1+\frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]^2}{\gamma(ax+bx^3)\left[1+\frac{(q-1)[ax+bx^3]^2}{2\gamma D}\right]^2 - \gamma\lambda\left[1-\frac{(q-1)[ax(s)+bx(s)^3]^2}{2\gamma D}\right]}\right\}^2, \tag{25}$$

which keeps information on the true nature of the process through the parameters γ and q.

4 CONCLUSIONS

The outlined calculation opens the door to the "brave new world" of nanosystems for which inertia is not negligible, like molecular motors and nanodevices in general.

ACKNOWLEDGMENTS

HSW acknowledges financial support from MICINN (Spain) through Project FIS2010-18023, and RRD from CONICET and UNMdP of Argentina.

REFERENCES

[1] R.P. FEYNMAN, *The Space-Time Formulation of Nonrelativistic Quantum Mechanics*, Rev. Mod. Phys. 20 (1948) pp.367–387.

[2] R.P. FEYNMAN AND A.R. HIBBS, *Quantum Mechanics and Path Integrals*, McGraw-Hill, New York, 1965.

[3] P.A.M. DIRAC, *The Lagrangian in Quantum Mechanics*, Physikalische Zeitschrift der Sowjetunion 3 (1933) pp.64–72.

[4] P.J. DANIELL, *Integrals in An Infinite Number of Dimensions*, Ann. Math. 2nd Series 20 (1919) pp.281–288.

[5] N. WIENER, *The average of an analytical functional*, Proc. Nat. Acad. Sci. USA 7 (1921) pp.253–260.

[6] N. WIENER, *The average of an analytical functional and the Brownian movement*, Proc. Nat. Acad. Sci. USA 7 (1921) pp.294–298.

[7] M. KAC, *On Distributions of Certain Wiener Functionals*, Trans. Am. Math. Soc. 65 (1949) pp.1–13.

[8] L. ONSAGER AND S. MACHLUP, *Fluctuations and irreversible processes*, Phys. Rev. 91 (1953) pp.1505–1512.

[9] S. MACHLUP AND L. ONSAGER, *Fluctuations and irreversible processes. II. Systems with Kinetic Energy*, Phys. Rev. 91 (1953) pp.1512–1515.

[10] H. S. WIO, *An Introduction to Stochastic Processes and Nonequilibrium Statistical Physics*, World Scientific, Singapore, 1994.

[11] P. JUNG AND P. HÄNGGI, *Dynamical systems: A unified colored-noise approximation*, Phys. Rev. A 35 (1987) pp.4464–4466.

[12] F. CASTRO, H. S. WIO, AND G. ABRAMSON, *Colored-noise problem: A Markovian interpolation procedure*, Phys. Rev. E 52 (1995) pp.159–164.

[13] H.S. WIO, P. COLET, L. PESQUERA, M.A. RODRÍGUEZ, AND M. SAN MIGUEL, *Path-integral formulation for stochastic processes driven by colored noise*, Phys. Rev. A 40 (1989) pp.7312–7324.

[14] F. LANGOUCHE, D. ROEKAERTS, AND E. TIRAPEGUI, *Functional Integration and Semiclassical Expansions*, Reidel, Dordrecht, 1982.

[15] M.A. FUENTES, H.S. WIO, AND R. TORAL, *Effective Markovian approximation for non-Gaussian noises: a path integral approach*, Physica A 303 (2002) pp.91–104.

[16] L. BORLAND, *Îto–Langevin equations within generalized thermostatistics*, Phys. Lett. A 245 (1998) pp.67–72.

[17] H. RISKEN, *The Fokker–Planck equation, 2nd ed.*, Springer, Berlin, 1989.

[18] M. RAHMAN, *Stationary solution for the color-driven Duffing oscillator*, Phys. Rev. E 53 (1996) pp.6547–6550.

COMPORTAMIENTO DEL PROBLEMA DE STEFAN A UNA FASE CUANDO EL NÚMERO DE BIOT TIENDE A CERO

Adriana C. Briozzo y Domingo A. Tarzia

Depto. Matemática - CONICET, F.C.E., Universidad Austral, Paraguay 1950, S2000FZF Rosario, ARGENTINA, E-mail: ABriozzo@austral.edu.ar, DTarzia@austral.edu.ar

Resumen: Se considera un problema de Stefan a una fase con una condición convectiva en el borde fijo, caracterizada por el coeficiente de transferencia $h > 0$ (directamente proporcional al número de Biot). Se estudia la convergencia de la temperatura y de la frontera libre cuando $h \to 0$. Se realiza el análisis matemático del problema fśico planteado en Naaktgeboren, Int. J. Heat Mass Transfer, 50 (2007), 4614-4622.

Palabras clave: *problema de Stefan, problema de frontera libre, ecuación integral de Volterra*
2000 AMS Subject Classification: 80A22 - 35R35 - 45D05 - 35C15

1. INTRODUCCIÓN

En este trabajo se considera el problema de frontera libre unidimensional (problema de Stefan a una fase) con una condición convectiva sobre la frontera fija $\xi = 0$ el cual consiste en determinar la temperatura $\theta = \theta(\xi, \tau)$ y la frontera libre $\xi = s(\tau)$ que satisfacen las siquientes condiciones

$$\begin{cases} (i)\ \rho c\theta_\tau - k\theta_{\xi\xi} = 0, & 0 < \xi < s(\tau), \tau > 0 \\ (ii)\ k\theta_\xi(0,\tau) = h\left[\theta(0,\tau) - g(\tau)\right], & \tau > 0 \\ (iii)\ \theta(s(\tau),\tau) = 0, & \tau > 0, \\ (iv)\ k\theta_\xi(s(\tau),\tau) = -\rho l \frac{ds}{d\tau}(\tau), & \tau > 0, \\ (v)\ \theta(\xi,0) = \varphi(\xi), & 0 \leqslant \xi \leqslant b \\ (vi)\ s(0) = b\ (b > 0) \end{cases} \tag{1}$$

donde $h > 0$ es el coeficiente de transferencia de calor, la temperatura inicial es $\varphi(\xi) \geq 0,\ 0 \leqslant \xi \leqslant b$, la temperatura del fluído exterior es $g = g(\tau) \geq 0,\ \tau > 0$ y se satisfacen las condiciones de compatibilidad $k\varphi'(0) = h\left(\varphi(0) - g(0)\right)$ y $\varphi(b) = 0$. El objetivo de este trabajo es estudiar el comportamiento de la solución $\theta = \theta_h(\xi,\tau),\ s = s_h(\tau)$ del problema (1) cuando $h \to 0$ realizando el análisis matemático del problema físico considerado en [6] donde se plantea el límite de la solución cuando el número de Biot (proporcional a h según (4)) tiende a cero.

La existencia y unicidad global de la solución del problema (1) está dada en [3]. En [10] fue probado que el comportamiento asintótico cuando $t \to +\infty$ del problema de frontera libre a una fase con una condición convectiva en el borde fijo es el mismo que para el caso en que la condición en el borde fijo $x = 0$ sea de temperatura. El comportamiento asintótico para el problema de Stefan a una fase con una condición de temperatura en el borde fijo fue estudiado en [1], [2] . Para el problema unidimensional de Stefan a dos fases el correspondiente comportamiento asintótico fue estudiado en [11]. En cambio, para el caso particular $g(\tau) = Const > 0$, y para un dominio multidimensional, el estudio del comportamiento asintótico fue obtenido utilizando la inecuación variacional parabólica en [8], [9].

2. RELACIONES INTEGRALES Y PROPIEDADES

Sobre las condiciones iniciales y los datos en el borde se asumen las siguientes hipótesis:
(i) Sea $\varphi = \varphi(\xi)$ una función positiva y con derivada seccionalmente continua y acotada.
(ii) Sea $g = g(\tau)$ una función positiva, seccionalmente continua y acotada.
(iii) Condiciones de compatibilidad: $g(0) > \varphi(\xi)$ en $(0,b),\ k\varphi'(0) = h\left(\varphi(0) - g(0)\right)$.
Si se define la siguiente transformación

$$u(x,t) = \frac{c}{l}\theta(\xi,\tau),\ x = \frac{\xi}{b}\ ,\ t = \frac{k}{\rho c b^2}\tau \tag{2}$$

entonces el problema (1) es equivalente al dado en variables adimensionales

$$\begin{cases} (i)\ u_t - u_{xx} = 0, & 0 < x < S(t),\ t > 0, \\ (ii)\ u_x(0,t) = H\left[u(0,t) - G(t)\right], & t > 0, \\ (iii)\ u(S(t),t) = 0, & t > 0, \\ (iv)\ u_x(S(t),t) = -\dot{S}\ (t), & t > 0, \\ (v)\ u(x,0) = \chi(x) \geq 0, & 0 \leqslant x \leqslant 1, \\ (vi)\ S(0) = 1 \end{cases} \tag{3}$$

donde

$$G(t) = \frac{c}{l} g(\frac{b^2 t}{\alpha}),\ H = \beta_i = b\frac{h}{k}\ \text{(número de Biot)}, \tag{4}$$

$$\chi(x) = \frac{c}{l}\varphi(b\xi),\ S(t) = \frac{1}{b} s(\frac{b^2 t}{\alpha}),\ \alpha = \frac{k}{\rho c}. \tag{5}$$

La solución $u = u_H(x,t)$ del problema de frontera libre tiene la siguiente representación integral [5], [11]

$$u_H(x,t) = \int_0^1 N(x,t;\xi,0)\chi(\xi)d\xi + \int_0^t N(x,t;S_H(\tau),\tau)V_H(\tau)d\tau \tag{6}$$

$$-H\int_0^t N(x,t;0,\tau)v_H(\tau)d\tau + H\int_0^t N(x,t;0,\tau)G(\tau)d\tau.$$

donde la nueva frontera libre $S = S_H(t)$ está dada por:

$$S_H(t) = 1 - \int_0^t V_H(\tau)d\tau, \tag{7}$$

las funciones V_H y v_H están definidas por

$$V_H(t) = u_x(S(t),t),\ \ v_H(t) = u_H(0,t) \tag{8}$$

y satisfacen las siguientes ecuaciones integrales de Volterra:

$$V_H\,(t) = 2\int_0^1 \chi'\,(\xi)\,G(S_H(t),t,\xi,0)d\xi - 2\int_0^t H[v_H(\tau) - G(\tau)]N_x(S_H(t),t,0,\tau)d\tau$$

$$+2\int_0^t V_H(\tau)N_x(S_H(t),t,S_H(\tau),\tau)d\tau \tag{9}$$

$$v_H\,(t) = \int_0^1 \chi\,(\xi)\,N(0,t,\xi,0)d\xi - \int_0^t H[v_H(\tau) - G(\tau)]N(0,t,0,\tau)d\tau \tag{10}$$

$$+\int_0^t V_H(\tau)N(0,t,S_H(\tau),\tau)d\tau.$$

Motivados por el problema planteado en [6], se estudia el comportamiento de la solución $u = u_H(x,t)$, $S = S_H(x,t)$ del problema de frontera libre(3) cuando $H \to 0$. Se prueba que la solución del problema (3) converge a la solución del problema de frontera libre parabólico (11)

$$\begin{cases} (i)\ u_{0t} - u_{0_{xx}} = 0, & 0 < x < S_0(t),\ t > 0, \\ (ii)\ u_{0_x}(0,t) = 0, & t > 0, \\ (iii)\ u_0(S_0(t),t) = 0, & t > 0, \\ (iv)\ u_{0_x}(S_0(t),t) = -\dot{S}_0\ (t), & t > 0, \\ (v)\ u_0(x,0) = \chi(x) \geq 0, & 0 \leqslant x \leqslant 1 \\ (vi)\ S_0(0) = 1 \end{cases} \tag{11}$$

cuando $H \to 0$.
La solución del problema (11) tiene la siguiente representación integral:

$$u_0(x,t) = \int_0^1 N(x,t;\xi,0)\chi(\xi)d\xi + \int_0^t N(x,t;S_0(\tau),\tau)u_{0_x}(S_0(\tau),\tau)d\tau \tag{12}$$

y la frontera libre es [4]

$$S_0(t) = 1 + \int_0^1 \chi(x)dx - \int_0^{S_0(t)} u_0(x,\tau)dx.$$

Las soluciones $u = u_H(x,t),\ S = S_H(x,t)$ y $u_0 = u_0(x,t),\ S = S_0(x,t)$ de los problemas (3) y (11) respectivamente, satisfacen las siguientes relaciones [7]:

Lema 1 *1.* $H\int_0^t [u_H(0,\tau) - G(\tau)]\,d\tau = 1 - S_H(t) - \int_0^{S_H(t)} u_H(x,t)dx + \int_0^1 \chi(x)dx$

2. $\int_0^{S_H(t)} x\ u_H(x,t)dx - \int_0^1 x\ \chi(x)dx = \frac{1}{2} - \frac{S_H^2(t)}{2} + \int_0^t u_H(0,\tau)d\tau$

3. $\int_0^{S_H(t)} u_H^2(x,t)dx - \int_0^1 \chi^2(x)dx + 2\int_0^t \int_0^{S_H(\tau)} u_{H_x}^2(x,\tau)dxd\tau \le H\int_0^t G^2(\tau)d\tau$

4. $S_0(t) = 1 - \int_0^{S_0(t)} u_0(x,t)dx + \int_0^1 \chi(x)dx$

5. $\int_0^{S_0(t)} x\ u_0(x,t)dx - \int_0^1 x\ \chi(x)dx = \frac{1}{2} - \frac{S_0^2(t)}{2} + \int_0^t u_0(0,\tau)d\tau$

6. $\int_0^{S_0(t)} u_0^2(x,t)dx - \int_0^1 \chi^2(x)dx + 2\int_0^t \int_0^{S_0(\tau)} u_{0_x}^2(x,\tau)dxd\tau = 0$

Lema 2 *Se tiene* $S_0(t) < S_H(t)$ *y* $u_H(x,t) \ge u_0(x,t)$ *para todo* $0 < x < S_0(t),\ t > 0,\ H > 0.$

Lema 3 *Si* $\int_0^t G(\tau)d\tau$ *es acotada, se tienen los siguientes límites:*

$$\lim_{H\to 0} S_H(t) = S_0(t),\ \ \lim_{H\to 0} u_H(0,t) = u_0(0,t)$$

para cada $t > 0$.

Prueba. Teniendo en cuenta las representaciones integrales (6) y (12) se tiene:

$$\begin{aligned} u_H(0,t) - u_0(0,t) &= -\int_0^t H[u_H(0,\tau) - G(\tau)]N(0,t,0,\tau)d\tau \\ &+ \int_0^t \dot{S}_H\,(\tau)\,[N(0,t;S_0(\tau),\tau) - N(0,t,S_H(\tau),\tau)]\,d\tau \\ &+ \int_0^t N(0,t,S_0(\tau),\tau)\left[\dot{S}_0\,(\tau) - \dot{S}_H\,(\tau)\right]d\tau = A_1 + A_2 + A_3. \end{aligned}$$

Se tienen las siguientes estimaciones:

$$A_1 = \int_0^t H[G(\tau) - u_H(0,\tau)]N(0,t,0,\tau)d\tau \le H\int_0^t \frac{G(\tau)}{\sqrt{\pi(t-\tau)}}d\tau = \frac{2H}{\sqrt{\pi}}\sqrt{t}\int_0^t G(\tau)d\tau,$$

$$|A_2| \le \left(\frac{6}{e}\right)^{3/2} \frac{1}{2\sqrt{\pi}} \left\|\dot{S}_H\right\|_{[0,t]} \|S_H - S_0\|_{[0,t]},$$

y

$$\begin{aligned} |A_3| &\le \int_0^t \left[|N_\xi(0,t,S_0(\tau),\tau)|\left|\dot{S}_0\,(\tau)\right| + |N_\tau(0,t,S_0(\tau),\tau)|\right]|S_0(\tau) - S_H(\tau)|\,d\tau \\ &\le \left\{\frac{1 + \int_0^1 \chi(x)dx}{2\sqrt{\pi}}\left(\frac{6}{e}\right)^{\frac{3}{2}}\left\|\dot{S}_0\right\|_{[0,t]} + \left(\frac{10}{e}\right)^{\frac{3}{2}}\frac{1}{2\sqrt{\pi}}\left[\frac{\|S_0\|_{[0,t]}^2}{2}t + \frac{t^2}{2}\right]\right\}\|S_0 - S_H\|_{[0,t]}. \end{aligned}$$

□

Teorema 1 *Si* $\int_0^t G(\tau)d\tau$ *es acotada para cada* $t > 0$*, se tiene*

$$\lim_{H\to 0} u_H(x,t) = u_0(x,t),$$

para cada $0 < x < S_0(t)$*,* $t > 0$*.*

Prueba. Teniendo en cuenta las igualdades 2) y 5) del Lema 1 se tiene

$$\begin{aligned} 0 \quad \leq \quad & \int_0^{S_0(t)} x\left[u_H(x,t) - u_0(x,t)\right]\, dx + \frac{S_H^2(t)}{2} - \frac{S_0^2(t)}{2} = \int_{S_0(t)}^{S_H(t)} x\, u_H(x,t)dx \\ & + \int_0^t \left[u_H(0,\tau) - u_0(0,\tau)\right] d\tau. \end{aligned}$$

Usando el lema anterior se obtiene

$$0 \leq \int_0^{S_0(t)} x\left[u_H(x,t) - u_0(x,t)\right]\, dx + \frac{S_H^2(t)}{2} - \frac{S_0^2(t)}{2} \leq \|u_H(.,t)\|_{[S_0(t),S_H(t)]} \left(\frac{S_H^2(t)}{2} - \frac{S_0^2(t)}{2}\right) +$$

$$+\frac{2H}{\sqrt{\pi}} t^{3/2} \int_0^t G(\tau)d\tau + \left(\frac{6}{e}\right)^{3/2} \frac{t}{2\sqrt{\pi}} \left\|\dot{S}_H\right\|_{[0,t]} \|S_H - S_0\|_{[0,t]} +$$

$$+\left\{\frac{1+\int_0^1 \chi(x)dx}{2\sqrt{\pi}} \left(\frac{6}{e}\right)^{\frac{3}{2}} \left\|\dot{S}_0\right\|_{[0,t]} + \left(\frac{10}{eb^{*2}}\right)^{\frac{3}{2}} \frac{1}{2\sqrt{\pi}} \left[\frac{\|S_0\|_{[0,t]}^2}{2} t + \frac{t^2}{2}\right]\right\}\, t\, \|S_0 - S_H\|_{[0,t]}.$$

□

REFERENCIAS

[1] J.R. CANNON AND C. D. HILL, *Remarks on a Stefan problem*, J. Math. Mech., 17 (1967), pp. 433-441.

[2] J.R. CANNON AND M. PRIMICERIO, *Remarks on the one-phase Stefan problem for the heat equation with the flux prescribed on the fixed boundary*, J. Math. Anal. Appl., 35 (1971), pp. 361-373.

[3] A. FASANO AND M. PRIMICERIO, *General Free-Boundary Problems for the Heat Equation*, I, J. Math. Anal. Appl., 57 (1977), pp. 694-723.

[4] A. FASANO AND M. PRIMICERIO, *New results on some classical parabolic free-boundary problems*, Quart. Appl. Math. 38 (1981), pp. 439-460.

[5] A. FRIEDMAN, *Free boundary problems for parabolic equations I. Melting of solids*, J. Math. Mech., 8 (1959), pp. 499-517.

[6] C. NAAKTGEBOREN, *The zero-phase Stefan problem*, Int. J. of Heat and Mass Transfer, 50 (2007), pp. 4614-4622.

[7] A. D. SOLOMON, V.ALEXIADES AND D. G. WILSON,*The Stefan problem with a convective boundary condition*, Quart. of Appl. Math., 40 (1982), pp. 203-217.

[8] D. A. TARZIA, *Sur le problemè de Stefan à deux phases*, C. R. Acad. Sci. Paris 288 A (1979), pp. 941-944.

[9] D. A. TARZIA, *Etude de l'inequation variationelle proposée par Duvaut pour le problème de Stefan à deux phases II*, Boll. Un. Mat. Italiana, 2 B (1983), pp. 589-603.

[10] D. A. TARZIA AND C. V. TURNER, *The asymptotic behavior for the one-phase Stefan problem with a convective boundary condition*, Appl. Math. Lett., 9 No.3 (1996), pp. 21-24.

[11] D. A. TARZIA AND C. V. TURNER, *The asymptotic behavior for the two-phase Stefan problem with a convective boundary condition*, Comm. Appl. Analysis, 7 (2003), N^o 3, pp. 313-334.

SOBRE LA RESOLUCIÓN DE UN PROBLEMA DE FRONTERA LIBRE A TRAVÉS DE UNA SUCESIÓN DE PROBLEMAS DE FRONTERA MÓVIL Y DE CAUCHY.

Luis T. Villa[#] y Angélica C. Boucíguez[##]

[#]*Facultad de Ingeniería, Universidad Nacional de Salta. Instituto de Investigaciones para la Industria Química (INIQUI) Av. Bolivia 5150 Salta, Argentina*, villal@unsa.edu.ar
[##]*Facultad de Ciencias Exactas. Universidad Nacional de Salta. Instituto de Investigaciones ene Energía No Convencional (INENCO) Av. Bolivia 5150 Salta, Argentina*. acbouciguez@gmail.com

Resumen: Se considera un modelo matemático descriptivo de un fenómeno de desorción de humedad libre durante el primer período o etapa de burbujeo en el proceso de freído por inmersión profunda de bastones de papa natural, consistente en un problema unidimensional de frontera libre (PVC) asociado a la ecuación parabólica de difusión de humedad en el interior de bastón sometido a freído. Se presenta una variante para obtener aproximaciones a la solución del PVC citado a través de la resolución de una sucesión de problemas de frontera móvil y de Cauchy, oportunamente asociados al PVC.

Palabras claves: *freído por inmersión, frontera móvil, frontera libre*
2000 AMS Subjects Classification: 35K20

1. INTRODUCCIÓN

En la industria alimenticia, resulta interesante analizar el proceso de freído por inmersión profunda de papa natural, en el que están presentes el transporte simultáneo de calor y materia, existiendo dos zonas con distintas propiedades, que por su carácter dinámico constituye una problema de frontera libre.

El transporte de calor es debido al aumento de temperatura en el proceso, mientras que el de materia se debe a la absorción de aceite y la desorción de humedad en la papa. Ambos procesos pueden estudiarse independientemente uno de otro [1] estando vinculados por la posición de la interfase. El proceso de transferencia de calor, es un problema de frontera libre que no puede reducirse a uno de frontera móvil, mientras que la transferencia de materia si lo permite; en virtud de ello, en este trabajo se analizará solo éste último.

Así, el perfil de concentración de humedad en el sólido puede tratarse como un problema de frontera libre, lo que permitirá evaluar la pérdida de peso por vaporización durante el proceso de cocción; dicho modelo fue formulado en [2], donde se señala la existencia de dos zonas, perfectamente demarcadas: corazón (RC) y corteza (RP), cuya separación está dada por la interfase $S(t)$. La primera, caracterizada por un contenido de humedad en condiciones de equilibrio termodinámico, se extiende desde el centro del bastón hasta la interfase y la segunda comprendida entre ésta última y la superficie en contacto con el aceite, es la zona donde ocurre la desorción de humedad libre y la migración hacia el contorno $x=R$, donde tiene lugar su vaporización, al entrar en contacto con el aceite caliente. La posición de la interfase, que varía en el tiempo, es precisamente la que se evalúa mediante un modelo de frontera libre. En la Figura 1, se presenta un esquema de la sección de un bastón prismático de longitud L y espesor $2R$, donde se produce la desorción de humedad libre en la papa. Se considera que $L>>2R$, de modo que el problema puede considerarse unidimensional.

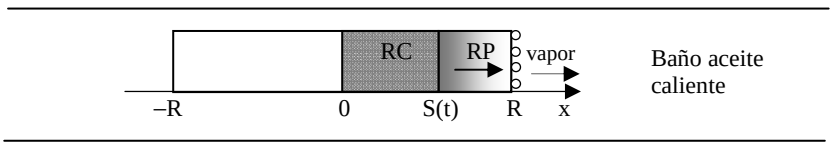


Figura 1: Esquema de la sección de papa

Donde:
x, *t* son las coordenadas espacial y temporal, respectivamente.
S=S(t), denota la posición instantánea del frente de desorción de humedad libre que se mueve desde el borde externo *x=R* (*S*(0)=*R*) hacia el centro *x*=0 del bastón.

2. MODELO MATEMATICO

El modelo matemático descriptivo del proceso se presenta en las ecuaciones (1) a (6)

Región central (RC)

$$C(x,t) = C_0, \quad 0 \le x \le S(t), \quad t > t_1 \tag{1}$$

Región periférica (PR)

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D(C)\frac{\partial C}{\partial x}\right), \quad S(t) < x < R, \quad t > t_1 \tag{2}$$

$$C(S(t),t) = C_0, \quad t > t_1 \tag{3}$$

$$-4AD\frac{\partial C}{\partial x}(R,t) = \omega_v, \quad t > t_1 \tag{4}$$

$$S(t_1) = R \tag{5}$$

Debiéndose tener presente que (5) es equivalente a la condición:

$$C(x,t_1) = C_0, \quad 0 \le x \le R \tag{6}$$

Siendo
A: Area superficial lateral del bastón (m^2)
t_1: Tiempo requerido por la etapa inicial (precalentamiento) del bastón de papa sometido a freído
C(x,t): Concentración volumétrica (kg/m^3) correspondiente a la humedad libre desorbida
C_0: Concentración volumétrica inicial para *C(x,t)*
D: Coeficiente de difusividad global de humedad desorbida (m^2/seg)
ω_v: Velocidad de vaporización (kg/seg)

Frontera libre

Representada con *S(t)*, determina la separación entre ambas regiones: CR y PR. Se utiliza el modelo cinético propuesto por Krokida et all [3] para obtener la expresión de la velocidad de vaporización ω_v y la ecuación diferencial descriptiva del comportamiento dinámico de la frontera libre *S=S(t).* Entonces se tiene:

$$-\omega_v = 4R^2 L(\rho_s - C_0)\frac{K}{60}(x_0 - x_e)exp(-Kt/60) \tag{7}$$

$$\frac{dS}{dt} = -\frac{R^2(\rho_s - C_0)\cdot(K_x/60)\cdot(X_0 - X_e)\cdot exp(-K_x t/60)}{(R + S(t))\cdot C_0 - \int_{S(t)}^{R} C(x,t)\cdot dx} \tag{8}$$

Las ecuaciones (1) a (8) constituyen el modelo matemático descriptivo para el proceso de desorción de humedad en el freído, que es un problema de frontera libre.

3. ESQUEMA PROPUESTO DE APROXIMACIONES PARA LA SOLUCION DEL PROBLEMA.

En primer término, es de notar que a la vista de (5) y (8) se puede escribir formalmente el siguiente problema de valor inicial o de Cauchy para la función $S=S(t)$:

$$\frac{dS}{dt} = F(t, S(t), \text{parámetros})$$
$$S(t_1) = R \tag{9}$$

Con F dada por el segundo miembro de (8) y con $C(x,t)$ una función en clase oportunamente regular, solución de (2) a (5)

Se comprueba que F es Lipschiziana, respecto de S, lo que provee una condición suficiente para la existencia de la solución de (9).

Entonces, a partir de (9) se genera mediante un proceso iterativo una sucesión de funciones $S_1=S_1(t)$, $S_2=S_2(t)$,..., $S_n=S_n(t)$,…que debe tener la propiedad de convergencia esperada de modo tal que:

$$\lim_{n\to\infty} S_n(t) = S(t), \qquad t > t_1 \tag{10}$$

con $S(t)$ solución de (9).

El proceso se inicia considerando $S_0(t)=R$, lo que también implica que $C(x,t)=C_0$, en $0\leq x\leq R$. Con estos valores insertados en (9) resulta el siguiente problema de Cauchy para $S_1=S_1(t)$:

$$\frac{dS_1}{dt} = -\frac{a^2\, exp(-bt)}{2RC_0} \qquad t > t_1$$
$$S_1(t_1) = R \tag{11}$$

Con los parámetros a^2 y b dados por:

$$a^2 = R^2(\rho_s - C_0)\frac{K}{60}(x_0 - x_e), \qquad b = \frac{K}{60} \tag{12}$$

De (11) se obtiene:

$$S_1(t) = R + \frac{a^2}{2RbC_0}[exp(-bt) - exp(-bt_1)], \qquad t > t_1 \tag{13}$$

Con $S_1=S_1(t)$,dada por (13) se conforma el siguiente problema de frontera móvil, para la función incógnita $C_1=C_1(x,t)$:

$$\frac{\partial C_1}{\partial t} = D\frac{\partial^2 C_1}{\partial x^2}, \qquad S_1(t) < x < R, \qquad t > t_1 \tag{14}$$

$$C_1(S_1(t),t) = C_0, \qquad t > t_1 \tag{15}$$

$$C_1(x,t_1) = C_0 \tag{16}$$

La solución de (14) – (16) provee C_1=C$_1$*(x,t)* que junto con S_1*(t)* permiten formular el siguiente problema de Cauchy, para la función incógnita S_2=S_2*(t):*

$$\frac{dS_2}{dt} = -\frac{a^2\, exp(-bt)}{(R + S_1(t))\cdot C_0 - \int_{S_1(t)}^{R} C_1(x,t)dx}, \qquad t > t_1$$

$$S_2(t_1) = R \tag{17}$$

Con S_2=S_2*(t)* solución de (17) se continúa formulando el correspondiente problema de frontera móvil análogo a (14) – (16), para la función C_2=C$_2$*(x,t)* y así sucesivamente hasta que el procedimiento lleva a la obtención de S_n=S_n*(t)* y C_n=C$_n$*(x,t),* aproximaciones a la solución de (1) – (8).

4. CONCLUSIONES

La resolución de problemas de frontera libre implica la obtención simultánea de la posición de la frontera y el perfil de desorción, lo que puede ser un proceso tedioso teniendo en cuenta las ecuaciones involucradas en el mismo. Si se trata de un problema de frontera móvil, al conocerse la ley con la que esta se desplaza, el problema se torna considerablemente más sencillo.

En virtud de estas consideraciones y de lo expresado en el trabajo sobre existencia de la solución propuesta, el modelo aquí presentado permite llevar el problema de frontera libre a uno de frontera móvil, facilitando su evaluación, agregándose como contrapartida, el hecho de tener que realizar un proceso iterativo, que con las herramientas de cálculo que hoy se cuentan no genera mayores inconvenientes.

REFERENCIAS

[1] L.VILLA, A. BOUCIGUEZ Y M.C. SANZIEL, *Un Problema Inverso de Stefan en Freído por Inmersión.* ENIEF – MACI 2007. Córdoba Argentina. 2 al 5 de octubre de 2007. Mecánica Computacional. Asociación Argentina de Mecánica Computacional. Vol. XXVI, pp. 2115 – 2125. ISSN 1666-6070.

[2] L.VILLA, J.C. GOTTIFREDI Y A. BOUCIGUEZ, *Some Considerations on a Simultaneous Heat and MassTransfer Food Process. Model Formulation.* Internacional Review of Chemical Engineering , en trámite de recepción.

[3] M. KROKIDA, V. OREOPOULOU Y Z. MAROULIS. *Water loss and oil uptake as a function of frying time.* Journal of Food Engineering 44 (2000) pp. 39-46.

EXISTENCIA Y UNICIDAD LOCAL DE UNA SOLUCIÓN CLÁSICA PARA EL PROBLEMA ACOPLADO DE CALOR Y MATERIA DURANTE LA SOLIDIFICACIÓN DE UN MATERIAL DE ALTO CONTENIDO DE AGUA

Roberto Gianni *, Domingo A. Tarzia †‡
* Dipartimento Me.Mo.Mat., Univ. di Roma "La Sapienza", Via Antonio Scarpa 16, 00161 Roma, Italia.
E-mail: verdandister@gmail.com
† Departamento de Matemática, Facultad de Ciencias Empresariales, Univ. Austral, Paraguay 1950, S2000FZF Rosario, Argentina.
‡ CONICET, Argentina.
E-mail: DTarzia@austral.edu.ar

Resumen: La solidificación de materiales con alto contenido de agua (alimentos, suelos y tejidos) implica dos procesos simultáneos de transferencia dentro del sistema: transferencia de calor por conducción (formación de hielo) y transferencia de masa por difusión (sublimación de la superficie de hielo). En Olguin-Salvadori-Mascheroni-Tarzia, Int. J. Heat Mass Transfer, 51 (2008), pp. 4379-4391, se propuso un modelo físico-matemático expresado como un problema de cambio de fase con dos fronteras libres. El objetivo del presente trabajo es el de obtener la existencia y unicidad de una solución clásica local en el tiempo para el correspondiente problema de frontera libre acoplado a dos fases en un adecuado espacio funcional.

Palabras clave: Problema de frontera libre, Problema de Stefan, Materiales con alto contenido de agua, Frente de sublimación, Frente de solidificación, Existencia y unicidad de solución clásica.

2000 AMS Subjects Classification: 35R35, 80A22, 35A07, 35K50.

1. INTRODUCCIÓN

Durante la solidificación del agua de materiales de alto contenido acuoso tales como suelos, tejidos animales o vegetales y alimentos, que no se encuentren cubiertos por un material impermeable y perfectamente adherido, ocurre simultáneamente la sublimación del hielo que se forma durante el proceso. La velocidad de ambos fenómenos (solidificación y sublimación) está determinada tanto por características del material (fundamentalmente su composición, estructura y forma), como por las condiciones de enfriamiento (temperatura, humedad y tipo de medio que rodea al material). El proceso de sublimación, aunque su magnitud es mucho menor que la de la congelación, determina aspectos fundamentales de la calidad final en el caso de alimentos y afecta la estructura y utilidad de los tejidos congelados. El modelado de estos procesos simultáneos es muy difícil debido a que los balances de materia y energía están acoplados, que existen dos frentes móviles de cambio de fase que se desplazan a velocidades muy diferentes y que las propiedades físicas involucradas son en la mayoría de los casos variables con la temperatura y el contenido de agua.

El proceso de congelación (sin sublimación) ha sido extensamente estudiado por [4], [13] y [14]. El sistema ha sido modelado en forma analítica por [14] y [18], y a través de métodos numéricos por [4], [14] y [15]. Teniendo en cuenta solamente la sublimación del hielo el proceso ha sido estudiado en [2], [3], [6], [11] y [16]. Debido a la no linealidad del problema, es dificultoso dar una solución analítica al mismo. En cambio, es factible resolverlo para sistemas idealizados o de composición y estructura simple. Una extensa bibliografía sobre problemas de frontera móvil y libre para la ecuación del calor-difusión fue dada en [19].

Al someter a un material de alto contenido acuoso a una temperatura inferior a su temperatura de solidificación, a la que se supone está inicialmente, se observan simultáneamente dos fenómenos: a) el líquido se congela; b) la superficie del hielo sublima. Por lo tanto, se pueden definir claramente tres zonas: una deshidratada, otra congelada y una tercera no congelada. La congelación comienza a partir de la superficie refrigerada, a una temperatura T_{if} que es menor que la temperatura de solidificación del agua pura, debido a la presencia de materiales disueltos y continúa a lo largo de una línea de equilibro (frontera libre) que es desconocida. Simultáneamente comienza la sublimación del hielo en la superficie congelada y aparece un frente de deshidratación cuya velocidad de avance debe determinarse también. Normalmente, esta velocidad es mucho menor que la velocidad del frente de congelación [1]. Por lo anterior, el problema consiste en resolver simultáneamente un problema de transferencia de calor (congelación) y un problema de transferencia de masa (pérdida de peso) y se desconocen los bordes que separan las zonas deshidratada de la congelada y ésta de la aún sin congelar.

Se considera un material semi-infinito con características similares a las de un gel muy diluido cuyas propiedades se suponen iguales a las del agua pura. El sistema tiene una temperatura inicial constante T_{if} y

al tiempo $t=0$ la superficie $x=0$ se expone a un medio exterior con temperatura constante T_s (< T_{if}) y coeficientes de transferencia de calor h_0 y de masa K_m constantes. Se asume que $T_s < T_0(t) < T_{if}$, $t>0$ donde $T_0(t)$ representa la temperatura de sublimación desconocida. Para calcular la evolución de la temperatura y del contenido de agua en el tiempo, se presenta en [17] el siguiente problema de frontera libre acoplado a dos fases: Hallar las temperaturas $T_d = T_d(x,t)$ (de la región deshidratada) y $T_f = T_f(x,t)$ (de la región congelada), la concentración $C_v = C_v(x,t)$ (de la región deshidratada), las dos fronteras libres $x = s_d(t)$ (frente de sublimación) y $x = s_f(t)$ (frente de congelación), y la temperatura $T_0 = T_0(t)$ en $x = s_d(t)$ que deben satisfacer las ecuaciones diferenciales y condiciones siguientes:

$$\rho_d c_d \frac{\partial T_d}{\partial t} = k_d \frac{\partial^2 T_d}{\partial x^2} \quad \text{en} \quad Q_{1T} \equiv \left\{(x,t): 0 < x < s_d(t), \quad 0 < t < T\right\}, \tag{1}$$

$$\varepsilon \frac{\partial C_v}{\partial t} = D \frac{\partial^2 C_v}{\partial x^2} \quad \text{en} \quad Q_{1T}, \tag{2}$$

$$\rho_f c_f \frac{\partial T_f}{\partial t} = k_f \frac{\partial^2 T_f}{\partial x^2} \quad \text{en} \quad Q_{2T} \equiv \left\{(x,t): s_d(t) < x < s_f(t), \quad 0 < t < T\right\}, \tag{3}$$

$$k_d \frac{\partial T_d}{\partial x}(0,t) = h_0\left[T_d(0,t) - T_s\right] \quad \text{sobre} \quad x=0, \quad 0<t<T, \tag{4}$$

$$D \frac{\partial C_v}{\partial x}(0,t) = k_m\left[C_v(0,t) - C_a\right] \quad \text{sobre} \quad x=0, \quad 0<t<T, \tag{5}$$

$$T_d\left(s_d(t),t\right) = T_f\left(s_d(t),t\right) = T_0(t) \quad \text{sobre} \quad x = s_d(t), \quad 0<t<T, \tag{6}$$

$$k_f \frac{\partial T_f}{\partial x}\left(s_d(t),t\right) - k_d \frac{\partial T_d}{\partial x}\left(s_d(t),t\right) = L_s m_s \frac{ds_d}{dt}(t) \quad \text{sobre} \quad x = s_d(t), \quad 0<t<T, \tag{7}$$

$$D \frac{\partial C_v}{\partial x}\left(s_d(t),t\right) = m_s \frac{ds_d}{dt}(t) \quad \text{sobre} \quad x = s_d(t), \quad 0<t<T, \tag{8}$$

$$C_v\left(s_d(t),t\right) = F\left(T_0(t)\right) \quad \text{sobre} \quad x = s_d(t), \quad 0<t<T, \tag{9}$$

$$T_f\left(s_f(t),t\right) = T_{if} \quad \text{sobre} \quad x = s_f(t), \quad 0<t<T, \tag{10}$$

$$k_f \frac{\partial T_f}{\partial x}\left(s_f(t),t\right) = m_f L_f \frac{ds_f}{dt}(t) \quad \text{sobre} \quad x = s_f(t), \quad 0<t<T, \tag{11}$$

$$s_d(0) = s_{0d}, \quad s_f(0) = s_{0f}, \tag{12}$$

$$T_d(x,0) = T_{0d}(x), \quad 0 \le x \le s_{0d}, \tag{13}$$

$$C_v(x,0) = C_{0v}(x), \quad 0 \le x \le s_{0d}, \tag{14}$$

$$T_f(x,0) = T_{0f}(x), \quad s_{0d} \le x \le s_{0f}. \tag{15}$$

Se asume que $T_s < T_0(t) < T_{if}$, $t>0$. Los coeficientes del problema son: C_a : concentración másica de vapor de agua en el aire; *c:* calor específico; *D:* coeficiente de difusión efectivo del agua; h_0*:* coeficiente de transferencia de calor; *k:* conductividad térmica; k_m : coeficiente de transferencia de masa; L_s*:* calor latente de sublimación del agua; L_f : calor latente de solidificación del agua; m_s : masa sublimada por unidad de volumen; ε*:* porosidad; ρ*:* densidad de masa; *T:* temperatura. El subíndice *f* se refiere a la zona congelada, el subíndice *d* a la zona deshidratada y el subíndice *o* a las condiciones iniciales. Se asumen las hipótesis siguientes:

$$H_1: \quad \rho_d, c_d, k_d, \varepsilon, D, \rho_f, c_f, k_f, h_0, k_m, L_s, M_s, T_{if}, C_a, T_s > 0, \quad T_s < T_{if}, \quad s_{0f} > s_{0d} > 0.$$

$$H_2: F(\eta) \in C^3(\mathbb{R}).$$

La condición (9) es una condición que generaliza la dada en [17] pues en el caso físico se tiene:

$$F(\eta) := \frac{M\,a\,e^{b-\frac{c}{\eta}}}{R\eta}, \tag{16}$$

donde los coeficientes b, M, a, R, c son constantes. Por otra parte, $C_v(s_d(t),t)$ representa la concentración de vapor de equilibrio a la temperatura $T_0(t)$, y la correspondiente presión de saturación se evalúa según [7]. In [17], se utiliza el método cuasi-estacionario y el sistema (1) – (15) se reduce a un sistema de ecuaciones diferenciales ordinarias acopladas par las fronteras libres $x = s_d(t)$ y $x = s_f(t)$ y la temperatura $T_0 = T_0(t)$. Estos resultados se usan para predecir las temperaturas $T_d(x,t)$ y $T_f(x,t)$, y la concentración C_v.

El objetivo del presente trabajo es el de obtener en la Sección II la existencia y la unicidad de una solución clásica local del problema de frontera libre acoplado a dos fases (1) – (15) en un adecuado espacio funcional. Se usa el trabajo fundamental [12]; otras referencias útiles sobre el tema son [5], [8], [9] y [10].

2. Existencia y Unicidad de una Solución Clásica Local

El sistema (1) – (15) es equivalente a uno nuevo en el cual las condiciones (8) y (9) son reemplazadas por:

$$k_f \frac{\partial T_f}{\partial x}(s_d(t),t) - k_d \frac{\partial T_d}{\partial x}(s_d(t),t) = \beta \frac{\partial C_v}{\partial x}(s_d(t),t) \text{ sobre } x = s_d(t),\ 0 < t < T \tag{8bis}$$

$$C_v(s_d(t),t) = F\left(T_d(s_d(t),t)\right) \quad \text{sobre} \quad x = s_d(t), \quad 0 < t < T \tag{9bis}$$

donde $\beta = L_s D$. Luego, se re-escribe el sistema (1)–(15) en una forma más conveniente a través de un sistema equivalente de ecuaciones diferenciales parabólicas en el mismo dominio cilíndrico proponiendo el siguiente cambio de coordenadas y de funciones (17) definido por:

$$y = \frac{x}{Ax + B} \quad , \quad t = t\,, \tag{17a}$$

$$\theta_d(y,t) = T_d(x,t) = T_d\left(\frac{B(t)y}{1 - A(t)y}, t\right), \quad \theta_f(y,t) = T_f(x,t) = T_f\left(\frac{B(t)y}{1 - A(t)y}, t\right), \tag{17b}$$

$$\mathrm{W}(y,t) = C_v(x,t) = C_v\left(\frac{B(t)y}{1 - A(t)y}, t\right), \tag{17c}$$

donde:

$$A = A(t) = \frac{2s_d(t) - s_f(t)}{2\left(s_d(t) - s_f(t)\right)}, \qquad B = B(t) = \frac{s_d(t)\,s_f(t)}{2\left(s_f(t) - s_d(t)\right)}. \tag{18}$$

Finalmente, se obtiene un nuevo sistema de ecuaciones diferenciales parciales (S), todas están definidas en el mismo dominio Ω_{1T}, introduciendo las nuevas funciones incógnitas:

$$u_1(z,t) = \theta_d(z,t), \quad u_2(z,t) = \theta_f(2 - z,t), \quad u_3(z,t) = W(z,t)\,. \tag{19}$$

Observación *La original temperatura desconocida* $T_0(t)$ *puede ser calculada por* $u_1(1,t)$ ó $\mathrm{u}_2(1,t)$.

Se asumen las siguientes hipótesis:

$$H_3: \quad T_{0d}, C_{0v} \in H^{2+\alpha}\left([0, s_{0d}]\right), \quad T_{0f} \in H^{2+\alpha}\left(\left[s_{0d}, s_{0f}\right]\right), \qquad \alpha \in (0,1),$$

H_4: se satisfacen las condiciones de compatibilidad de primer orden cuando se imponen condiciones de contorno de tipo Robin y las condiciones de compatibilidad de segundo orden cuando se imponen condiciones de contorno de tipo Dirichlet.

El problem (S) es equivalente al problema (1)–(15) en el sentido que toda solución clásica de (S) es una solución clásica (1) – (15) y recíprocamente. Para el sistema de ecuaciones (S) se prueba que existe una única solución clásica siempre que *T* sea elegido suficientemente pequeño.

Teorema *Bajo las hipótesis* $H_1 - H_4$ *existe un tiempo* $\hat{T} > 0$ *de manera que el problema (S) admite una única solución clásica en* $\Omega_{1\hat{T}}$, *i.e. existe un quintuple de funciones* $\left(u_1(z,t),\ u_2(z,t), u_3(z,t), s_d(t), s_f(t)\right)$ *tal que* $u_i \in H^{2+\beta}\left(\overline{\Omega}_{1\hat{T}}\right)$ $(i=1,2,3)$, $s_d, s_f \in H^{1+\beta/2}\left(\left[0,\hat{T}\right]\right), \forall \beta < \frac{\alpha}{2}$, *que satisfacen el problema* (S).

Prueba. Tanto para la existencia como para la unicidad de una solución clásica del problema (S) se utilizan, en forma reiterada, diferentes resultados de [12, pp. 601-616]. Para la existencia de solución local se utiliza además el método del argumento retardado. □

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por los Proyectos "Problemi matematici di diffusione in biologia" y "Modelli differenziali in matematica applicata" de la Sapienza Univ. di Roma para el primer autor, y por los Proyectos PIP Nº 0460 de CONICET-UA, Rosario, Argentina y Grant FA9550-10-1-0023 para el segundo autor.

REFERENCIAS

[1] L.A. CAMPAÑONE, *Transferencia de calor en congelación y almacenamiento de alimentos. Sublimación de hielo, calidad, optimización de condiciones de proceso*, Tesis Doctorado en Ingeniería, Univ. Nac. de La Plata, La Plata (2001).

[2] L.A. CAMPAÑONE, V.O. SALVADORI, AND R.H. MASCHERONI, *Weight loss during freezing and storage of unpackaged foods*, J. Food Eng., 47 (2001), pp. 69-79.

[3] L.A. CAMPAÑONE, V.O. SALVADORI, AND R.H. MASCHERONI, *Food freezing with simultaneous surface dehydration. Approximate prediction of weight loss during freezing and storage*, Int. J. Heat Mass Transfer, 48 (2005), pp. 1195-1204.

[4] A.C. CLELAND, *Food Refrigeration Processes. Analysis, Design and Simulation*, Elsevier, London (1990).

[5] A. FASANO, AND R. GIANNI, *Freezing of a two component liquid-liquid dispersion*, Nonlinear Anal., 1 (2000), pp. 435-448.

[6] M. FARID, *The moving boundary problems from melting and freezing to drying and frying of food*, Chem. Eng. Processing, 41 (2002), pp. 1-10.

[7] O. FENNEMA, AND L.A. BERNY, *Equilibrium vapour pressure and water activity of food at subfreezing temperature*, In Proc. IV Int. Congress of Food Sci. and Technology, 2 (1974), pp. 27-35.

[8] R. GIANNI, *Global existence of a classical solution for a large class of free boundary problems*, Nonlinear Diff. Eq. Appl., 2 (1995), pp. 291-321.

[9] R. GIANNI, AND P. MANNUCCI, *A free boundary problem in an absorbing porous material with saturation dependent permeability*, Nonlinear Diff. Eq. Appl., 8 (2001), pp. 219-235.

[10] R. GIANNI, AND P. MANNUCCI, *A filtration problem in a composite porous material with two free boundaries advances*, Adv. Math. Sci. Appl., 11 (2001), pp. 603-622.

[11] M. KOCHS, CH. KÖRBER, B. NUNNER, AND I. HESCHEL, *The influence of the freezing process on vapour transport during sublimation in vacuum-freeze-drying*, Int. J. Heat Mass Transfer, 34 (1991), pp. 2395-2408.

[12] O.A. LADYZENSKAJA, V.A. SOLONNIKOV, AND N.N. URAL'CEVA, *Linear and Quasilinear Equations of ParabolicTtype*, American Math. Society, Providence (1968).

[13] A.V. LUIKOV, *Systems of differential equations of heat and mass transfer in capillary porous bodies (review)*, Int. J. Heat Mass Transfer, 18 (1975), pp. 1-14.

[14] V.J. LUNARDINI, *Heat Transfer with Freezing and Thawing*, Elsevier, London (1991).

[15] R.H. MASCHERONI, AND A. CALVELO, *Relationship between heat transfer parameters and the characteristic damage variables for the freezing of beef*, Meat Sci., 4 (1980), pp. 267-285.

[16] J. D. MELLOR, *Fundamentals of Freeze Drying*, Academic Press, London, (1978).

[17] M.C. OLGUIN, V.O. SALVADORI, R.H. MASCHERONI, AND D.A. TARZIA, *An analytical solution for the coupled heat and mass transfer during the freezing of high-water content materials*, Int. J. Heat Mass Transfer, 51 (2008), pp. 4379-4391.

[18] E.A. SANTILLAN MARCUS, AND D.A. TARZIA, *Explicit solution for freezing of humid porous half-space with heat flux condition*, Int. J. Eng. Sci., 38 (2000), pp. 1651-1665.

[19] D.A. TARZIA, *A bibliography on moving-free boundary problems for the heat diffusion equation. The Stefan and related problems*, MAT - Serie A, 2 (2000), pp. 1-297. See http://web.austral.edu.ar/descargas/facultad-cienciasEmpresariales/mat/Tarzia-MAT-SerieA-2(2000).pdf

CONTROL ÓPTIMO EN UN PROCESO DE DESUBLIMACIÓN

Elina M. Mancinelli[♭,‡] y Eduardo A. Santillan Marcus[†,‡]

[†]*Dpto. de Matemática, FCE, Univ. Austral, Paraguay 1950, S2000FZF Rosario, Argentina*
[‡]*Dpto. de Matemática, FCEIA, Univ. Nacional de Rosario, Pellegrini 250, 2000 Rosario, Argentina*
[♭]*CONICET, FCEIA-UNR, Argentina*
elina@fceia.unr.edu.ar, esantillan@austral.edu.ar

Resumen: Se estudia un proceso de desublimación de humedad en un medio poroso finito a través de un enfoque analítico. Este proceso se describe como un problema acoplado de Stefan a dos fases en donde se controla la evolución de la frontera libre usando condiciones de temperatura. Se minimiza un funcional de costo apropiado. Se obtiene el sistema adjunto asociado.

Palabras clave: *frontera libre, control óptimo, problema adjunto*
2000 AMS Subject Classification: 35R35; 80A22; 49J20

1. INTRODUCCIÓN

En el presente trabajo se desarrolla una estrategia de optimización de la frontera libre en un problema de desublimación (congelamiento de humedad) de forma similar a lo establecido en [4]. Este modelo matemático de la transferencia de masa y calor en cuerpos porosos capilares fue establecido por Luikov [5]-[7]. El proceso se describe por un problema de Stefan a dos fases. Modelos similares fueron estudiados en [1], [2] y [9].
Nuestro objetivo es controlar la evolución de la frontera libre usando la temperatura en $x = 0$. El objetivo del control es que la frontera libre se pegue a una evolución determinada, y esta meta se logra minimizando un adecuado funcional de costo, en el que el error entre las gráficas de la frontera libre y una frontera libre límite deseada se reduce al mínimo.

2. MODELO FÍSICO

Consideraremos que la desublimación ocurre en un medio poroso acotado cerrado $\Omega = [0, 1]$, y se asume que la humedad se evapora completamente a una temperatura constante.
Para $t \in [0, T]$ el frente de evaporación está dado por $x = s(t)$ y así quedan establecidas dos regiones:

$$\Omega_T^1 = \{(x, t) : 0 < x < s(t), 0 < t < T\},$$

que es la región que ya está congelada, donde $u = u(x, t)$ es la temperatura, y

$$\Omega_T^2 = \{(x, t) : s(t) < x < 1, 0 < t < T\}$$

que es la región donde hay flujos acoplados de calor y humedad, siendo $v = v(x, t)$ la temperatura y $w = w(x, t)$ la humedad.

En la región de congelamiento Ω_T^1 no hay movimiento de humedad y la distribución de temperatura está descripta por la ecuación del calor

$$\partial_t u(x, t) = a_1 \partial_{xx} u(x, t),$$

donde a_1 es la difusividad termal en Ω_T^1.
La region Ω_T^2 es donde en el cuerpo poroso de capilares húmedos fluyen acoplados calor y humedad. El proceso está descripto por el sistema de Luikov:

$$\partial_t w(x, t) = a_m \partial_{xx} w(x, t), \quad \partial_t v(x, t) = a_2 \partial_{xx} v(x, t) + \varepsilon \frac{L c_m}{c_2} \partial_t w(x, t),$$

donde a_2 es la difusividad termal y a_m es la difusividad de humedad en Ω_T^2, δ es el coeficiente del gradiente termal y ε es el factor de conversión de fase de líquido a vapor).

Las distribuciones iniciales de temperatura y humedad están dadas por:

$$u(x,0) = \theta(x) \leq 0, \qquad v(x,0) = \phi(x) \geq 0, \qquad w(x,0) = \psi(x) > 0.$$

En $x = 1$, las distribuciones de temperatura y humedad satisfacen:

$$v(1,t) = h(t) \qquad\qquad w(1,t) = w_0(t).$$

Sobre el frente de congelamiento, existe una igualdad entre las temperaturas:

$$u(s(t),t) = v(s(t),t) = 0, \qquad\qquad w(s(t),t) = w_s.$$

El balance de calor y humedad establece que

$$k_1 \partial_x u(s(t),t) - k_2 \partial_x v(s(t),t) = -(1-\varepsilon)\rho_m L \dot{s}_t, \qquad \partial_x w(s(t),t) + \delta \partial_x v(s(t),t) = 0,$$

donde $k_i, i = 1, 2$ son las conductividades termales en Ω_T^i, ρ_m es la densidad del cuerpo poroso en Ω_T^2, y L es el calor latente de congelamiento. La posición inicial de la frontera libre es $s(0) = b > 0$.

3. Problema de optimización

Nuestro objetivo es controlar la frontera libre usando la temperatura $u_b(0,t)$. Como horizonte de control tomamos $t \in (0,T]$ para $T > 0$ finito. Separamos en dos partes: $u^b(0,t) = u^{b_0} + \beta u^{bc}$,
con una parte fija u^{b_0} (por ejemplo, una temperatura conocida experimentalmente) y una temperatura de control βu_{b_c}, donde β es una función de peso que permite adaptar la parte de control de la condición de frontera.
La condición de flujo en el borde donde se aplica el control está dada por

$$u(0,t) - \frac{k_1}{\alpha_1} \partial_x u(0,t) = u^b.$$

Entonces estamos en condiciones de enunciar que el funcional de costo del problema es:

$$\mathcal{J}(s, u_{b_c}) := \int_0^T (s(t) - \tilde{s}(t))^2 dt + \lambda_1 (s(T) - \tilde{s}(T))^2 + \lambda_2 \int_0^T \beta^2 u_{b_c}^2(t) dt$$

donde el primer término de este funcional modela el objetivo de nuestro problema de minimización, con λ_1 pesando la desviación de la frontera libre de la frontera libre deseada en el tiempo $t = T$, y el segundo término pesa el costo del control con λ_2 y puede servir como regularizador.
El problema se expresa como

$$\min_{s, u_{b_c}} \mathcal{J}(s, u_{b_c}),$$

sujeto a las siguientes condiciones: $[C1]$

$$\text{en } \Omega_T^1 \begin{cases} \partial_t u = a_1 \partial_{xx} u, \\ u(0,t) - \frac{k_1}{\alpha_1} \partial_x u(0,t) = u^b, , \\ u(x,0) = \theta(x) \leq 0, \end{cases} \qquad \text{en } \Omega_T^2 \begin{cases} \partial_t w = a_m \partial_{xx} v + \varepsilon \dfrac{L c_m}{c_2} \partial_t w, \\ \partial_t w = a_m \partial_{xx} w, \\ v(x,0) = \phi(x) \geq 0, \qquad w(x,0) = \psi(x) > 0, \\ v(1,t) = h(t), \qquad w(1,t) = w_0(t), \\ u(s(t),t) = v(s(t),t) = 0, \qquad w(s(t),t) = w_s, \\ s(0) = b > 0, \\ k_1 \partial_x u(s(t),t) - k_2 \partial_x v(s(t),t) = -(1-\varepsilon)\rho_m L \dot{s}, \\ \partial_x w(s(t),t) + \delta \partial_x v(s(t),t) = 0. \end{cases}$$

4. Problema desacoplado

Por conveniencia en el desarrollo de los cálculos, introducimos una nueva función z, que acopla a v y w en la región Ω_T^2, dada por

$$z(x,T) = v(x,T) + \frac{a_m}{a_2 - a_m} \varepsilon \frac{L c_m}{c_2} w(x,T)$$

y de esta forma luego de algunos cálculos elementales, las restricciones para nuestro problema en u, w y z se transforman en: $[C2]$

$$\text{en } \Omega_T^1 \begin{cases} \partial_t u = a_1 \partial_{xx} u, \\ u(0,t) - \frac{k_1}{\alpha_1}\partial_x u(0,t) = u^b, \\ u(x,0) = \theta(x) \leq 0, \\ u(s(t),t) = 0 \end{cases} \quad \text{en } \Omega_T^2 \begin{cases} \partial_t z = a_2 \partial_{xx} z, & \\ \partial_t w = a_m \partial_{xx} w, & \\ z(x,0) = \eta(x), & w(x,0) = \psi(x) > 0, \\ z(1,t) = \chi(t), & w(1,t) = w_0(t), \\ z(s(t),t) = \alpha w_s, & w(s(t),t) = w_s, \\ s(0) = b, & \\ k_1 \partial_x u(s(t),t) - k_2 \partial_x z(s(t),t) & \\ \qquad + k_2 \alpha \partial_x w(s(t),t) = -(1-\varepsilon)\rho_m L \dot{s}, & \\ \partial_x w(s(t),t) + \frac{\delta}{1-\delta\alpha}\partial_x z(s(t),t) = 0. & \end{cases}$$

5. Sistema adjunto

Ahora formalmente obtenemos la condición de optimalidad necesaria para nuestro problema de minimización utilizando el método de Lagrange. El funcional de Lagrange asociado al problema viene dado por:

$$\begin{aligned} L &:= L(s,u,w,z,u_{b_c},p_1,...,p_8) \\ &= J(s,u^{bc}) + \int_0^T\int_0^{s(t)}(\partial_t u - a_1\partial_{xx}u)p_1(x,T)dxdt + \int_0^T\int_{s(t)}^1(\partial_t z - a_2\partial_{xx}z)p_2(x,T)dxdt \\ &+ \int_0^T\int_{s(t)}^1(\partial_t w - a_m\partial_{xx}w)p_3(x,T)dxdt + \int_0^T((k_1\partial_x u - k_2\partial_x z + k_2\alpha\partial_x w + (1-\varepsilon\rho_m\dot{s}(t)))p_4)\,(s(t),t)dt \\ &+ \int_0^T u(s(t),t)p_5(s(t),t)dt + \int_0^T(z(s(t),t) - \alpha w(s(t),t))p_6(s(t),t)dt \\ &+ \int_0^T(u(0,t) - \tfrac{k_1}{\alpha_1}\partial_x u(0,t) - u^{b_0}(t) - \beta(t)u_{b_c}(t))p_7(0,t)dt + \int_0^T(\partial_x w(s(t),t) + \tfrac{\delta}{1-\alpha\delta}\partial_x z(s(t),t))p_8(s(t),t)dt \end{aligned}$$

donde u, w, z y s satisfacen

$$u(x,0) = \theta(x) \leq 0, \qquad z(x,0) = \eta(x), \qquad z(1,t) = \chi(t), \quad w(x,0) = \psi(x) > 0, \quad w(1,t) = w_0(t)$$

$$u^{b_0}(0) + \tfrac{k_1}{\alpha_1}\partial_x u(0,0) = \theta(0), \quad s(0) = \tilde{s}(0) = 0, \quad \beta(0) = 0.$$

Las funciones p_i, $i = 1,..,8$ son los multiplicadores de Lagrange asociado a las restricciones.
La condición de optimalidad de primer orden para nuestro problema viene dada por $\nabla L = 0$, y el sistema adjunto se define a través de $L_u\bar{u} = 0$, $L_z\bar{z} = 0$, $L_w\bar{w} = 0$, y $L_s\bar{s} = 0$, siendo:

$$\begin{aligned} L_u\bar{u} &= -\int_0^T\int_0^{s(t)}\bar{u}(\partial_t p_1 + a_1\partial_{xx}p_1)dxdt + \int_0^{s(t)}\bar{u}(x,T)p_1(x,T)dx + \int_0^T\bar{u}(0,t)(a_1p_1(0,t) - \tfrac{k_1}{\alpha_1}p_7(0,t))dt \\ &\quad + \int_0^T\partial_x\bar{u}(s(t),t)(k_1p_4(s(t),t) - a_1p_1(s(t),t))dt + \int_0^T\bar{u}(0,t)(p_7(0,t) - a_1\partial_x p_1(0,t)) \\ L_z\bar{z} &= \int_{s(t)}^1\bar{z}(x,T)p_2(x,T)dx - \int_0^T\int_{s(t)}^1\bar{z}(\partial_t p_2 + a_2\partial_{xx}p_2)dxdt - \int_0^T(a_2\partial_x\bar{z}(1,t)p_2(1,t)dt \\ &\quad + \int_0^T\partial_x\bar{z}(s(t),t)(a_2p_2(s(t),t) - k_2p_4(s(t),t) + \tfrac{\delta}{1-\alpha\delta}p_8(s(t),t))dt \\ L_w\bar{w} &= \int_{s(t)}^1\bar{z}(x,T)p_3(x,T)dx - \int_0^T\int_{s(t)}^1\bar{w}(\partial_t p_3 + a_2\partial_{xx}p_3)dxdt - \int_0^T(a_m\partial_x\bar{w}(1,t)p_3(1,t)dt \\ &\quad + \int_0^T\partial_x\bar{w}(s(t),t)(a_mp_3(s(t),t) - k_2\alpha p_4(s(t),t) + p_8(s(t),t))dt \end{aligned}$$

Estas condiciones nos llevan al siguiente sistema de ecuaciones adjuntas: ([C3])

$$\begin{aligned} &\partial_t p_1 + a_1\partial_{xx}p_1 = 0, && p_1(x,T) = 0, \\ &k_1p_4(s(t),t) - a_1p_1(s(t),t) = 0, && p_7(0,t) - a_1\partial_x p_1(0,t) = 0, \\ &a_1p_1(0,t) - \tfrac{k_1}{\alpha_1}p_7(0,t) = 0, && p_2(x,T) = 0, \\ &\partial_t p_2 + a_2\partial_{xx}p_2 = 0, && a_2p_2(s(t),t) - k_2p_4(s(t),t) + \frac{\delta}{1-\alpha\delta}p_8(s(t),t) = 0, \\ &p_2(1,t) = 0, && p_3(x,T) = 0, \\ &\partial_t p_3 + a_2\partial_{xx}p_3 = 0, && a_mp_3(s(t),t) - k_2\alpha p_4(s(t),t) + p_8(s(t),t) = 0, \\ &p_3(1,t) = 0. \end{aligned}$$

Agregamos al sistema adjunto la condición $\frac{a_2}{a_m}\,\alpha\,\partial_x p_2(s(t),t) + \partial_x p_3(s(t),t) = 0$ para obtener que el Lagrangeano tiene la siguiente forma:

$$L = \int_0^T(s(t) - \tilde{s}(t))^2dt + \alpha(s(T) - \tilde{s}(T))^2 + \beta\int_0^T u_b^2(t)dt - \int_0^b\theta(x)p_1(x,0)dx - \int_b^1\eta(x)p_2(x,0)dx +$$

$+\int_0^T a_2\chi(t)\partial_x p_2(1,t)dt - \int_b^1 \psi(x)p_3(x,0)dx + \int_0^T a_m w_0(t)\partial_x p_3(1,t)dt + \int_0^T (1-\varepsilon)\rho_m L\dot{s}(t)p_4(s(t),t)dt$
$-\int_0^T (u^{b_0}(t)+\beta(t)u_{b_c}(t))p_7(0,t)dt.$

Derivando respecto de s se tiene que

$$L_s\bar{s} = \int_0^T (s(t)-\tilde{s}(t))\bar{s}(t)dt + \lambda_2(s(T)-\tilde{s}(T))\bar{s}(T) + \int_0^T (1-\varepsilon)\rho_m L\dot{\tilde{s}}(t)p_4(s(t),t)dt$$

Integrando por partes y considerando que s no debe variar en $t=0$, pues $s(0)=\tilde{s}(0)=b$, surge que $\bar{s}(0)=0$. Además fijamos $L_s\bar{s}=0$ para las direcciones posibles de $\bar{s}$, en particular para aquellas tales que $\bar{s}(T)=0$ con lo que se obtiene que

$$\int_0^T \bar{s}(t)\left((s(t)-\tilde{s}(t)) - (1-\varepsilon)\rho_m L\partial_t p_4\tilde{s}\right)dt = 0,$$

y en consecuencia $$-(1-\varepsilon)\rho_m L\partial_t p_4(s(t),t)dt = s(t)-\tilde{s}(t), \quad \forall t\in(0,T).$$

Finalmente pidiendo $L_s\bar{s}=0, \forall\bar{s}$ y considerando lo ya obtenido se llega a

$$p_4(s(T),T) = \tfrac{\lambda_2}{(1-\varepsilon)\rho_m L}(\tilde{s}(T)-s(T)).$$

Integrando la penúltima ecuación respecto de t se tiene

$$-(1-\varepsilon)\rho_m L p_4(s(T),T) + (1-\varepsilon)\rho_m L p_4(s(t),t) = \int_t^T (\tilde{s}(\tau)-s(\tau))d\tau$$

Usando la última ecuación queda

$$p_4(s(t),t) = \tfrac{\lambda_2(\tilde{s}(T)-s(T))}{(1-\varepsilon)\rho_m L} + \tfrac{1}{(1-\varepsilon)\rho_m L}\int_t^T (\tilde{s}(\tau)-s(\tau))d\tau$$

Finalmente si asumimos que el problema de Stefan admite única solución para cada u_{b_c} y, en particular, consideramos a s como función de u_{b_c}, se puede definir el funcional de costo reducido

$$K(u_{b_c}) := \mathcal{J}(s(u_{b_c}), u_{b_c})$$

y se tiene que

$$K'(u_{b_c}) = L_{u_{b_c}}(s,u,w,z,u_{b_c},p_1,...,p_8).$$

En nuestro caso queda entonces que

$$K'(u_{b_c}) = \lambda_1\beta(0,t)^2 u_{b_c} - \beta\alpha_1 p_7(0,t).$$

En consecuencia, para un u_{b_c} dado, dicho gradiente puede calcularse resolviendo $([C2])$ para u y s y luego $([C3])$ para las variables adjuntas.
En nuestro caso la condición de optimalidad $\nabla L = 0$ es equivalente a $K'(u_{b_c})=0$.

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por los proyectos PIP 112-200801-00460 CONICET e ING272 UNR.

REFERENCIAS

[1] E. BOBULA, D. A. TARZIA, K. TWARDOWSKA, L. T. VILLA, *On a free-moving boundary diffusion problem in a catalytic gas-solid system with catalyst decay*, SIAM J. Appl. Math 60, 5 (2000), pp.1667-1685.

[2] J. R. CANNON, *The one-dimensional heat equation*, CA: Addison-Wesley, Menlo Park, 1984.

[3] M. HINZE, R. PINNAU, M. ULBRICH, S. ULBRICH, *Optimization with PDE constraints*, Mathematicval Modelling: Theory and Applications, Vol. 23, Springer, 2010.

[4] M. HINZE, S. ZIEGENBALG, *Optimal control of the free boundary in a two-phase Stefan problem*, Journal of Computational Physics 223 (2007), pp.657684.

[5] A. V. LUIKOV, *Heat and mass transfer in capillary-porous bodies*, Pergamon Press, Oxford, 1966.

[6] A. V. LUIKOV, *Analytical heat diffusion theory*, Academic Press, New York, 1968.

[7] A. V. LUIKOV, *Systems of differential equations of heat and mass transfer in capillary porous bodies*, Int. J. Heat Mass Transfer 18, (1975), pp.1-14.

[8] I. PAWLOW, *Optimal control of dynamical processes in two-phase systems of solid-liquid type*, Banach Center Publication 24 (1990), pp.293-319.

[9] E. A. SANTILLAN MARCUS, A. C. BRIOZZO, *On freezing of a finite humid porous medium with a heat flux condition*, Nonlinear Analysis 67 (2007), pp.1919-1937.

EL PROBLEMA DEL VALOR PROPIO INVERSO PARA CIERTA CLASE DE MATRICES

Leila Lebtahi y Néstor Thome

Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, España. {leilebep,njthome}@mat.upv.es

Resumen: Dados m vectores columna $x_1, \ldots, x_m$ en $\mathbb{C}^n$ dispuestos en una matriz $X \in \mathbb{C}^{n\times m}$ y m escalares $\lambda_1, \ldots, \lambda_m$ dispuestos en una matriz diagonal $D \in \mathbb{C}^{m\times m}$, el problema de encontrar una matriz $A \in \mathbb{C}^{n\times n}$ tal que $AX = XD$ se conoce con el nombre de problema del valor propio inverso en la Teoría del Análisis Matricial. En la literatura se ha abordado el estudio del problema del valor propio inverso suponiendo diferentes condiciones sobre la matriz incógnita A. Por ejemplo, se ha resuelto el caso en el que A es hermítica y anti-reflexiva con respecto a una matriz de reflexión generalizada, el caso en que A es una matriz singular reflexiva o anti-reflexiva con respecto a una matriz hermítica tripotente, etc. En este trabajo se resuelve el problema del valor propio inverso para una matriz A hermítica reflexiva con respecto a una matriz $\{k+1\}$-potente y normal. Se presentan condiciones necesarias y suficientes que garantizan la existencia de solución del problema y se da una solución explícita del mismo.

Palabras clave: *valores propios, inversas generalizadas, matrices hermíticas*
2000 AMS Subject Classification: 15A29 - 15A18

1. INTRODUCCIÓN

En este artículo se trata el problema del valor propio inverso. Este problema consiste en encontrar una matriz $A \in \mathbb{C}^{n\times n}$ tal que las m ecuaciones

$$Ax_i = \lambda_i x_i, \qquad i = 1, \ldots, m$$

tengan solución para los vectores dados $x_1, \ldots, x_m$ en $\mathbb{C}^n$ y para los escalares dados $\lambda_1, \ldots, \lambda_m$ en $\mathbb{C}$. Es decir, se resuelve la ecuación matricial $AX = XD$ en A con $X = \begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix}$ donde $x_1, \ldots, x_m$ son sus vectores columna en $\mathbb{C}^n$ y $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ una matriz diagonal siendo los λ_i, $i = 1, \ldots, m$, valores dados. En otras palabras, resolver un problema de valor propio inverso consiste en la reconstrucción de una matriz a partir de sus valores y vectores propios. Este problema tiene numerosas aplicaciones, por ejemplo, en teoría de control, problemas en mecánica, geofísica, física de partículas, física cuántica, etc. [1, 2, 6, 8].

En la literatura, la mayoría de los problemas de este tipo que se estudian introducen alguna restricción sobre la matriz A como, por ejemplo, cuando A es hermítica y anti-reflexiva con respecto a una matriz de reflexión generalizada [7]. Se recuerda que una matriz $A \in \mathbb{C}^{n\times n}$ se llama reflexiva con respecto a la matriz J cuando $A = JAJ$ y anti-reflexiva con respecto a J cuando $A = -JAJ$. Además, una matriz J se llama reflexión generalizada si $J^2 = I$ y $J^* = J$. Por otra parte, en [4], se resuelve un problema de mínimos cuadrados usando diferentes propiedades de una nueva clase de matrices introducidas por los autores: las matrices reflexivas con respecto a un par de matrices que sean ambas reflexiones generalizadas. También, en [5], se analizan las situaciones en que la matriz A sea hermítica bajo las condiciones adicionales de ser reflexiva o anti-reflexiva con respecto a una matriz J tripotente y hermítica.

Se recuerda que para una matriz $M \in \mathbb{C}^{s\times t}$ dada, una $\{1\}$-inversa generalizada de M es una matriz denotada por M^- que satisface $MM^-M = M$. Estas matrices inversas generalizadas siempre existen y están unívocamente determinadas [3]. Se utilizará la siguiente notación: $W^{(l)}(M) = I - M^-M$ y $W^{(r)}(M) = I - MM^-$,

En este trabajo se resuelve el problema del valor propio inverso encontrando las matrices hermíticas $A \in \mathbb{C}^{n\times n}$ reflexivas con respecto a una matriz $J \in \mathbb{C}^{n\times n}$ normal y $\{k+1\}$-potente (es decir, $JJ^* = J^*J$ y $J^{k+1} = J$, $k \in \mathbb{N}$). Se analizan condiciones necesarias y suficientes para la existencia de tales matrices A y se da una solución explícita del problema.

2. Existencia y forma explícita de las soluciones

Sea $X \in \mathbb{C}^{n\times m}$ una matriz dada y $D \in \mathbb{C}^{m\times m}$ una matriz diagonal también conocida. En este trabajo se buscarán soluciones de la ecuación

$$AX = XD \tag{1}$$

satisfaciendo que $A \in \mathbb{C}^{n\times n}$ es hermítica y reflexiva con respecto a la matriz $\{k+1\}$-potente y normal J, es decir $A = JAJ$ donde $J^{k+1} = J$ y $JJ^* = J^*J$. Nótese que la matriz diagonal D tiene elementos reales ya que A es hermítica.

Como J es una matriz normal, es diagonalizable mediante una matriz unitaria y como $J^{k+1} = J$, el espectro de J está incluido en $\{0\} \cup \Omega_k$, siendo Ω_k el conjunto de todas las raíces de la unidad de orden k. Entonces, existe una matriz unitaria $U \in \mathbb{C}^{n\times n}$ tal que

$$J = U \operatorname{diag}(\omega_1 I_{r_1}, \ldots, \omega_k I_{r_k}, O_{r_{k+1}})\, U^* \tag{2}$$

siendo $r_1 + \cdots + r_k + r_{k+1} = \operatorname{rango}(J)$.

Ahora se va a obtener la estructura de la matriz A. Para ello, se particiona la matriz U^*AU en bloques de tamaño adecuado como sigue:

$$U^*AU = \begin{bmatrix} A_{1,1} & \ldots & A_{1,k} & A_{1,k+1} \\ \vdots & \ddots & \vdots & \vdots \\ A_{k,1} & \ldots & A_{k,k} & A_{k,k+1} \\ A_{k+1,1} & \ldots & A_{k+1,k} & A_{k+1,k+1} \end{bmatrix} \tag{3}$$

de acuerdo con los bloques de la partición de J. De (2) y (3), la igualdad $A = JAJ$ lleva a $A_{k+1,j} = O$ para $j \in \{1, \ldots, k+1\}$, $A_{j,k+1} = O$ para $j \in \{1, \ldots, k\}$, y $A_{i,j} = \omega_i\omega_j A_{i,j}$ para $i, j \in \{1, \ldots, k\}$. De aquí resulta que $\omega_i\omega_j = 1$ o bien $A_{i,j} = O$ con $i, j \in \{1, \ldots, k\}$. Esto implica que para cada $i, j \in \{1, \ldots, k\}$ o bien $\omega_i = \overline{\omega}_j$ o bien los bloques $A_{i,j}$ y $A_{j,i}$ son simultáneamente nulos. Ahora se observa que la forma de la matriz U^*AU depende de las raíces de la unidad que figuran en la factorización de la matriz J dando origen a las situaciones que se presentan a continuación:

(I) Si k es par entonces se pueden dar las siguientes posibilidades:

- (a) $\{1, -1\} \subseteq \sigma(J)$,
- (b) $\{1, -1\} \cap \sigma(J) = \emptyset$,
- (c) $1 \in \sigma(J)$, $-1 \notin \sigma(J)$,
- (d) $1 \notin \sigma(J)$, $-1 \in \sigma(J)$.

(II) Si k es impar entonces se pueden dar las siguientes posibilidades:

- (a) $\{1, -1\} \cap \sigma(J) = \emptyset$,
- (b) $1 \in \sigma(J)$, $-1 \notin \sigma(J)$.

Si además se considera que en la factorización de J el orden de los valores propios en la matriz U^*JU es $1, -1, \omega_3, \overline{\omega}_3, \ldots, \omega_p, \overline{\omega}_p, 0$ (en caso que aparezcan 1 y -1), entonces la matriz A tiene la forma

$$A = U \operatorname{diag}(A_{1,1}, A_{2,2}, \tilde{A}_{3,4}, \ldots, \tilde{A}_{p,p+1}, O)\, U^* \tag{4}$$

donde, tomando p de manera adecuada, se tiene que

$$\tilde{A}_{s,s+1} = \begin{bmatrix} O & A_{s,s+1} \\ A_{s+1,s} & O \end{bmatrix} \qquad \text{siendo } s \in \{3, 5, \ldots, p\}. \tag{5}$$

En relación a (4), se asocia el bloque $A_{1,1}$ al valor propio 1 y el bloque $A_{2,2}$ al valor propio -1 (si figuran en J). También, en relación a (5), cada bloque $\tilde{A}_{s,s+1}$ está asociado a los valores propios ω_s y $\overline{\omega}_s$.

Por último, como A es hermítica, se debe cumplir que $A^*_{i,i} = A_{i,i}$ for $i = 1, 2$ y $A^*_{s,s+1} = A_{s+1,s}$ para $s \in \{3, 5, \ldots, p\}$.

La forma explícita de la solución viene dada en el siguiente resultado.

Teorema 1 *Sean $X \in \mathbb{C}^{n\times m}$, $D \in \mathbb{R}^{m\times m}$ una matriz diagonal y $J \in \mathbb{C}^{n\times n}$ una matriz normal $\{k+1\}$-potente como en (2). Sea la partición (de tamaño apropiado)*

$$X = U \left[\begin{array}{ccccccc} X_1^T & X_2^T & \tilde{X}_3^T & \dots & \tilde{X}_p^T & X_{p+1}^T \end{array} \right]^T \text{ con } \tilde{X}_s^T = \left[\begin{array}{cc} \tilde{\tilde{X}}_s^T & \tilde{\tilde{X}}_{s+1}^T \end{array} \right], \quad s \in \{3,5,\dots,p\}. \tag{6}$$

Entonces existe una matriz hermítica $A \in \mathbb{C}^{n\times n}$ reflexiva con respecto a J tal que $AX = XD$ si y sólo si existen matrices $\{1\}$-inversas generalizadas X_i^- de X_i tales que $X_i D W^{(l)}(X_i) = O$ y

$$X_i D X_i^- - (X_i^-)^* D X_i^* = (Y_i W^{(r)}(X_i))^* - Y_i W^{(r)}(X_i), \tag{7}$$

donde Y_i son matrices arbitrarias de tamaños adecuados para $i = 1, 2$; además existen matrices $\{1\}$-inversas generalizadas $\tilde{\tilde{X}}_s^-$ de $\tilde{\tilde{X}}_s$ tales que para $s \in \{3, 4, \dots, p+1\}$ debe ser

$$\tilde{\tilde{X}}_s D W^{(l)}(\tilde{\tilde{X}}_{s+1}) = O, \qquad W^{(r)}(\tilde{\tilde{X}}_s^*) D \tilde{\tilde{X}}_{s+1}^* = O, \qquad \tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s D = D \tilde{\tilde{X}}_{s+1}^* \tilde{\tilde{X}}_{s+1} \tag{8}$$

y también $X_{p+1} D = O$. En este caso, con la misma notación de (5), la solución general viene dada por

$$A = U \operatorname{diag}(X_1 D X_1^- + Y_1 W^{(r)}(X_1), X_2 D X_2^- + Y_2 W^{(r)}(X_2), \tilde{A}_{3,4}, \dots, \tilde{A}_{p,p+1}, O)\, U^* \tag{9}$$

donde

$$A_{s+1,s}^* = A_{s,s+1} = (\tilde{\tilde{X}}_s^*)^- D \tilde{\tilde{X}}_{s+1}^* + \tilde{\tilde{X}}_s D \tilde{\tilde{X}}_{s+1}^- - (\tilde{\tilde{X}}_s^*)^- \tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s D (\tilde{\tilde{X}}_{s+1})^- + W^{(l)}(\tilde{\tilde{X}}_s^*) Y_s W^{(r)}(\tilde{\tilde{X}}_{s+1})$$

para $s \in \{3, 4, \dots, p+1\}$.

Prueba. Primero se supone que existe una matriz hermítica A reflexiva con respecto a J tal que $AX = XD$. Por el razonamiento anterior, la forma de la matriz A está dada en (4). Sustituyendo en $AX = XD$ la partición establecida en (6) se obtiene

$$\left[\begin{array}{cccccc} A_{1,1} & O & O & \dots & O & O \\ O & A_{2,2} & O & \dots & O & O \\ O & O & \tilde{A}_{3,4} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & \tilde{A}_{p,p+1} & O \\ O & O & O & \dots & O & O \end{array} \right] \left[\begin{array}{c} X_1 \\ X_2 \\ \tilde{X}_3 \\ \vdots \\ \tilde{X}_p \\ X_{p+1} \end{array} \right] = \left[\begin{array}{c} X_1 \\ X_2 \\ \tilde{X}_3 \\ \vdots \\ \tilde{X}_p \\ X_{p+1} \end{array} \right] D.$$

Realizando las operaciones entre matrices por bloques se tiene que

$$A_{i,i} X_i = X_i D,\ i = 1, 2 \qquad \tilde{A}_{s,s+1} \tilde{X}_s = \tilde{X}_s D,\ s \in \{3, 5, \dots, p\} \qquad X_{p+1} D = O.$$

Para resolver las dos primeras ecuaciones es necesario utilizar matrices inversas generalizadas. Cada una de las ecuaciones $A_{i,i} X_i = X_i D$ para $i = 1, 2$ tiene solución $A_{i,i}$ si y sólo si existe una matriz $\{1\}$-inversa generalizada X_i^- de X_i tal que $X_i D X_i^- X_i = X_i D$. En estos casos la solución general viene dada por:

$$A_{i,i} = X_i D X_i^- + Y_i (I - X_i X_i^-), \tag{10}$$

siendo Y_i matrices arbitrarias de tamaños adecuados para $i = 1, 2$ [3]. Finalmente, las condiciones $A_{i,i}^* = A_{i,i}$ conducen a las igualdades (7).

Utilizando la notación de (5) y (6) y el hecho de que $A_{s+1,s}^* = A_{s,s+1}$, las ecuaciones dadas por $\tilde{A}_{s,s+1} \tilde{X}_s = \tilde{X}_s D$ conducen al sistema matricial

$$\left\{ \begin{array}{rcl} A_{s,s+1} \tilde{\tilde{X}}_{s+1} & = & \tilde{\tilde{X}}_s D \\ \tilde{\tilde{X}}_s^* A_{s,s+1} & = & D \tilde{\tilde{X}}_{s+1}^* \end{array} \right.$$

para cada $s \in \{3, 4, \ldots, p+1\}$. Las condiciones dadas en (8) garantizan la existencia de solución del sistema matricial anterior [3]. Su solución puede ser expresada como

$$A_{s,s+1} = (\tilde{\tilde{X}}_s^*)^- D\tilde{\tilde{X}}_{s+1}^* + \tilde{\tilde{X}}_s D\tilde{\tilde{X}}_{s+1}^- - (\tilde{\tilde{X}}_s^*)^- \tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s D(\tilde{\tilde{X}}_{s+1})^- + W^{(l)}(\tilde{\tilde{X}}_s^*) Y_s W^{(r)}(\tilde{\tilde{X}}_{s+1}). \quad (11)$$

Finalmente, de (10) y (11), la solución general del problema viene dada por (9). La implicación recíproca es evidente. □

En el siguiente resultado se obtienen condiciones suficientes para obtener la tercera igualdad de (8).

Lema 1 *Bajo la notación del Teorema 1, para cada $i = 1, 2, \ldots, t_s$ se denota por $x_{i,\cdot}^{(s)} = \begin{bmatrix} x_{i1}^{(s)} & \ldots & x_{in}^{(s)} \end{bmatrix}$ a las filas de la matriz $\tilde{\tilde{X}}_s \in \mathbb{C}^{t_s \times n}$. Entonces se tiene que cada igualdad*

$$\tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s D = D\tilde{\tilde{X}}_{s+1}^* \tilde{\tilde{X}}_{s+1} \quad (12)$$

es una condición necesaria para que cualquiera de las dos afirmaciones siguientes (no necesariamente de manera simultánea) se satisfagan:

(a) $\det(D) = 0$ *ó* $\prod_{i=1}^n \sigma_i(\tilde{\tilde{X}}_s) = \pm \prod_{i=1}^n \sigma_i(\tilde{\tilde{X}}_{s+1})$,

(b) $\left\langle \sum_{i=1}^{t_s} |x_{i,\cdot}^{(s)}|^2 - \sum_{j=1}^{t_{s+1}} |x_{j,\cdot}^{(s+1)}|^2, \mathrm{diag}(D) \right\rangle = 0$,

para cada $s = 3, 4, \ldots, p+1$.

Prueba. De la partición dada en (6) es claro que $\tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s \in \mathbb{C}^{n \times n}$. Tomando determinantes a ambos lados de la igualdad (12) se tiene que $\det(D) = 0$ o bien $\det(\tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s) = \det(\tilde{\tilde{X}}_{s+1}^* \tilde{\tilde{X}}_{s+1})$. Teniendo en cuenta que el determinante de una matriz cuadrada coincide con el producto de sus valores propios y que estos no son más que el cuadrado de los valores singulares correspondientes $\sigma_i(.)$ resulta que $\prod_{i=1}^n \left(\sigma_i(\tilde{\tilde{X}}_s)\right)^2 = \prod_{i=1}^n \left(\sigma_i(\tilde{\tilde{X}}_{s+1})\right)^2$ de donde fácilmente se deduce (a).

Por otro lado, por propiedades de la traza se tiene que la hipótesis lleva a $\mathrm{tr}((\tilde{\tilde{X}}_s^* \tilde{\tilde{X}}_s - \tilde{\tilde{X}}_{s+1}^* \tilde{\tilde{X}}_{s+1})D) = 0$. Como la traza del producto de una matriz por otra diagonal se escribe como el producto escalar entre la diagonal de la primera matriz con la de la segunda, después de algunas manipulaciones algebraicas se obtiene la condición (b) del enunciado. □

AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por el Proyecto DGI MTM2010-18228 y por el Proyecto de la Universidad Politécnica de Valencia, PAID-06-09, Ref.: 2659.

REFERENCIAS

[1] M. ARNOLD, *Conditioning and Algorithm for the Eigenvalue Assignment Problem*, Ph.D. thesis, Department of Mathematical Sciences, Northern Illinois University, Dekalb, IL, 1993.

[2] V. BARCILON, *On the multiplicity of solutions of the inverse problems for a vibrating beam*, SIAM J. Appl. Math., 37 (1979), pp. 119-127.

[3] A. BEN-ISRAEL, AND T. GREVILLE, *Generalized inverses: theory and applications*, John Wiley & Sons, Second Edition, Springer-Verlag, New York, 2003.

[4] H. C. CHEN, *Generalized reflexive matrices: special properties and applications*, SIAM J. Matrix Anal., 19, 1, (1998), pp. 140-153.

[5] L. LEBTAHI, AND N. THOME, *The inverse eigenvalue problem for Hermitian reflexive (anti-reflexive) matrices with respect to a tripotent Hermitian matrix*, in Proceedings of the Second ALAMA Meeting, Polytechnical Univ. of Valencia Press, Spain, (2010), 1-6.

[6] R. L. PARKER, AND K. A. WHALER, *Numerical methods for establishing solutions to the inverse problem of electromagnetic induction*, J. Geophys. Res., 86 (1981), pp. 9574-9584.

[7] ZHEN-YUN PENG, *The inverse eigenvalue problem for Hermitian anti-reflexive matrices and its approximation*, Applied Mathematics and Computation, 162, (2005), pp. 1377-1389.

[8] S. J. WANG, AND S. Y. CHU, *An algebraic approach to the inverse eigenvalue problem for a quantum system with a dynamical group*, J. Phys. A, 27 (1994), pp. 5655-5671.

REGULARIZACIÓN ESTADÍSTICA DE PROBLEMAS INVERSOS: MODELOS JERÁRQUICOS.

Gisela L. Mazzieri[♭,♮], Ruben D. Spies[♭,†] y Karina G. Temperini[♭,♯]

[♭]*Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Santa Fe, Argentina.*
[♮]*Depto. de Matemática, Fac. de Bioquímica y Cs. Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[†]*Depto. de Matemática, Fac. de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[♯]*Depto. de Matemática, Fac. de Humanidades y Cs., Universidad Nacional del Litoral, Santa Fe, Argentina.*

Resumen: En este trabajo se presentan algunos resultados estadísticos basados en la perspectiva Bayesiana, que son de gran utilidad en el abordaje de ciertos problemas inversos en los que los datos tienen naturaleza estocástica y/o la información "*a-priori*" sobre la solución exacta del problema es de tipo cualitativo [1], [3], [4]. Mientras que los métodos clásicos de regularización producen una única estimación de la solución del problema, los métodos estadísticos dan como resultado una distribución de probabilidades que puede luego utilizarse para obtener diferentes estimaciones de la solución. Mostraremos que si bien los métodos clásicos y estadísticos parten de enfoques y premisas completamente diferentes, en cierta manera, los mismos se encuentran fuertemente vinculados a través de la teoría generalizada de Tikhonov-Phillips [2]. Finalmente, se presentarán ejemplos numéricos asociados a problemas de procesamiento de señales con el objetivo de mostrar las ventajas que posee este enfoque estadístico de regularización cuando la información "*a-priori*" de la que se dispone es de tipo cualitativa.

Palabras clave: *Problema inverso, Tikhonov-Phillips, Regularización estadística.*
2000 AMS Subject Classification: 47A52 - 65J20

1. INTRODUCCIÓN

El próposito de las técnicas clásicas de regularización es obtener una estimación razonable de la variable de interés basado en los datos de los que se dispone. Por otra parte, desde el punto de vista de la teoría de inversión estadística la solución de un problema inverso es una distribución de probabilidades que puede utilizarse para obtener estimaciones de la incógnita. Dicha distribución, llamada la distribución de probabilidades "*a-posteriori*", describe el grado de conocimiento de la incógnita después que la información contenida en los datos ha sido procesada e incorporada a tal distribución. Por ende, el enfoque estadístico para tratar y resolver un problema inverso es muy diferente, y en principio totalmente disjunto, del enfoque clásico determinístico. Mientras que los métodos clásicos (determinísticos) de regularización producen una única estimación de la solución del problema en estudio, los métodos estadísticos dan como resultado una distribución de probabilidades que puede utilizarse para obtener estimaciones de la incógnita. Sin embargo, como veremos más adelante, los dos enfoques no son disjuntos. En este trabajo abordamos este enfoque para mostrar la relación antes mencionada entre ambos y utilizar las ventajas que este último posee a la hora de incorporar información sobre características o cualidades de la solución exacta. Mostraremos además cómo esta incorporación puede realizarse en distintos niveles de jerarquía dando lugar a los modelos conocidos como "modelos jerárquicos" o "hipermodelos".

2. PRELIMINARES

A lo largo de todo este trabajo, por cuestiones de simplicidad, nos restringiremos al estudio de la teoría en $\mathbb{R}^n$ y supondremos que las variables aleatorias involucradas en los distintos modelos son absolutamente continuas y, por ende, representaremos su distribución de probabilidades por medio de una función de densidad. En general tendremos un modelo de la forma: $Y = f(X, E)$ donde X es la variable incógnita, f es la función asociada al modelo, Y es la variable observable y E el error cometido en las mediciones de Y.

Desde el punto de vista Bayesiano, el problema inverso se expresa de la siguiente manera: *dada una observación de* $Y = y_{obs}$, *hallar la distribución de probabilidad condicional* $\pi(x|y_{obs})$ *de la variable* X. El siguiente teorema, al que nos referiremos como Fórmula de Bayes para problemas inversos, nos provee una manera de determinar dicha distribución.

Teorema 1 *(Fórmula de Bayes para problemas inversos) Sea X una variable aleatoria en $\mathbb{R}^n$ con distribución de densidad de probabilidad "a-priori" dada por π_{pr}. Sea y_{obs} una observación de la variable aleatoria $Y \in \mathbb{R}^k$ tal que $\pi(y_{obs}) > 0$. Entonces la distribución de probabilidad "a-posteriori" de X, dado el dato y_{obs} está dada por $\pi_{post}(x) \doteq \pi(x|y_{obs}) = \frac{\pi_{pr}(x)\pi(y_{obs}|x)}{\pi(y_{obs})} \propto \pi_{pr}(x)\pi(y_{obs}|x)$.*

El símbolo "$\propto$" significa "proporcional a". Teniendo en cuenta el teorema anterior podríamos "resumir" la resolución de un problema inverso en las siguientes tres tareas: 1) basado en la información disponible de la variable incógnita X, hallar una distribución *"a-priori"* que refleje dicha información; 2) hallar la función de verosimilitud, $\pi(y|x)$; 3) desarrollar métodos para explorar la distribución *"a-posteriori"*. Claramente, la definición abstracta de la solución de un problema inverso como la distribución de probabilidad "*a-posteriori*", no es muy útil en la práctica. En general, estaremos interesados en determinar distintos estimadores de dicha distribución. Uno de los estimadores estadísticos más utilizados es el estimador máximo "*a-posteriori*" que definimos a continuación:

Definición 1 *(Estimador MAP) Dada la densidad de probabilidad "a-posteriori" $\pi(x|y)$ de la variable aleatoria incógnita $X \in \mathbb{R}^n$, definimos el estimador máximo "a-posteriori", y lo denotamos con x_{MAP}, como $x_{MAP} \doteq arg \max_{x\in\mathbb{R}^n} \pi(x|y)$.*

3. Resultados principales.

3.1. Tikhonov-Phillips y métodos estadísticos.

Proposición 1 *Sean las variables aleatorias independientes $X \in \mathbb{R}^n$ e $Y, E \in \mathbb{R}^k$ tales que $X \sim \mathcal{N}(0, \gamma^2 I_n)$, $E \sim \mathcal{N}(0, \sigma^2 I_k)$, donde I_n e I_k denotan las matrices identidad de orden n y k, respectivamente. Supongamos además que el modelo que relaciona las variables es un modelo lineal con ruido aditivo, es decir $Y = AX + E$, donde $A \in \mathbb{R}^{k\times n}$ es una matriz conocida. Entonces la distribución de densidad "a-posteriori" de X dada una medición de $Y = y$ está dada por: $\pi(x|y) \propto \exp\left(-\frac{1}{2}(x-\bar{x})^T \Gamma_{post}^{-1}(x-\bar{x})\right)$, donde $\bar{x} = \left(A^TA + \frac{\gamma^2}{\sigma^2} I_k\right)^{-1} A^T y$ y $\Gamma_{post} = \gamma^2 I_n - \gamma^2 I_n A^T(\gamma^2 A I_n A^T + \sigma^2 I_k)^{-1}\gamma^2 A I_n$.*

De la proposición anterior se sigue inmediatamente que encontrar (estimar) la media $\bar{x}$ de $\pi_{post}(x)$ es equivalente a encontrar una solución regularizada del problema inverso $Ax = y$ con el método de Tikhonov-Phillips clásico y parámetro de regularización $\frac{\gamma^2}{\sigma^2}$. Señalamos en este punto que este resultado se obtuvo considerando la formulación funcional de dicho método. Sin embargo observar que $\pi_{post}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|x\|^2\right)$ $\exp\left(-\frac{1}{2\sigma^2}\|y-Ax\|^2\right) = \exp\left(-\frac{1}{2\gamma^2}\|x\|^2 - \frac{1}{2\sigma^2}\|y-Ax\|^2\right)$, de donde $x_{\text{MAP}} = \arg\max_{x\in\mathbb{R}^n} \pi_{post}(x)$ $= \arg\min_{x\in\mathbb{R}^n}\left(\|x\|^2 + \alpha\|y-Ax\|^2\right)$, con $\alpha \doteq \frac{\gamma^2}{\sigma^2}$. De este modo, determinar el estimador x_{MAP} es equivalente a encontrar el mínimo del funcional de Tikhonov-Phillips clásico. Con más generalidad, puede probarse que si $X \sim \mathcal{N}(0, \gamma^2 L^{-1})$, $E \sim \mathcal{N}(0, \sigma^2 I_k)$, donde L es una matriz arbitraria de orden n, entonces $\pi_{post}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2 - \frac{1}{2\sigma^2}\|y-Ax\|^2\right)$. De aquí se sigue entonces que el problema de calcular el estimador x_{MAP} es equivalente al de minimizar un funcional de Tikhonov-Phillips generalizado con penalizante $\|Lx\|^2$ y parámetro de regularización $\alpha \doteq \frac{\gamma^2}{\sigma^2}$. Es importante observar aquí cómo a través de una elección adecuada del potencial asociado a la distribución "*a-priori*", el estimador máximo "*a-posteriori*" de la distribución resultante coincide con una aproximación obtenida con un método de Tikhonov-Phillips en el que el término penalizante es exactamente el potencial de la distribución "*a-priori*".

3.2. Modelos jerárquicos.

La regularización de un problema inverso mal condicionado desde el punto de vista clásico requiere establecer un valor para el parámetro de regularización α. En la subsección anterior se mostró que, en el caso de los métodos estadísticos, el parámetro de regularización α está relacionado con la varianza de la variable incógnita. Es así que el conocimiento del valor de α se traduce, en el enfoque estadístico, en el conocimiento de la varianza mencionada. Sin embargo, la respuesta al interrogante de cómo elegir la

varianza de la distribución "*a-priori*" de la incógnita yace en un axioma propio de este enfoque estadístico Bayesiano según el cual "*si un parámetro es desconocido, entonces debe formar parte del problema de inferencia*". El abordaje de estos problemas da lugar a los llamados "modelos jerárquicos" o "hipermodelos" [1], [4].

El objetivo principal en esta sección es mostrar cómo las herramientas estadísticas permiten la incorporación de información cualitativa y/o estructural y cómo esta incorporación puede realizarse en distintos niveles de jerarquía a través de los "modelos jerárquicos" o "hipermodelos".

Supongamos que deseamos recuperar una señal a partir de un dato y que se sabe que la misma posee discontinuidades cuya ubicación y tamaño son conocidos. En este caso, el uso de una densidad "*a-priori*" del tipo "estructural" (ver [3]) permite incorporar dicha información y así obtener soluciones regularizadas que posean discontinuidades idénticas que la solución exacta. Desde el punto de vista de la teoría clásica este mismo resultado puede obtenerse utilizando el método de regularización de Tikhonov-Phillips generalizado utilizando como penalizante el potencial de la distribución "*a-priori*". Si en cambio la información acerca de las discontinuidades fuese sólo cualitativa, es decir se sabe que la señal posee saltos pero su ubicación y tamaño son desconocidos, entonces podemos formular un modelo jerárquico de nivel 1 que nos permita estimar la ubicación y magnitud de dichas discontinuidades y regularizar el problema simultáneamente, lo cual es imposible de hacer con los métodos clásicos. A continuación mostramos cómo un modelo de Markov autoregresivo de primer orden resulta muy apropiado para detectar este tipo de comportamiento en una señal.

Sea X la variable incógnita (señal). En lugar de representar la señal muestreada como un vector determinístico $x \doteq (x_1, x_2, \ldots, x_n)^T$ la modelamos como un proceso estocástico finito $\{X_j, 1 \leq j \leq n\}$. Un modelo autoregresivo de Markov de primer orden puede escribirse de la siguiente manera:

$$X_j = X_{j-1} + \theta_j^{1/2} V_j, \qquad V_j \sim \mathcal{N}(0,1), \qquad X_0 = 0, \tag{1}$$

donde θ_j es la varianza del "proceso de innovación" y cuantifica el grado de incertidumbre acerca de cuánto cambia la señal al pasar de un punto de la grilla a otro. El modelo en (1) puede reescribirse de la siguiente manera: $LX = D^{1/2}V$, donde X y V son variables aleatorias en $\mathbb{R}^n$ con componentes X_j y V_j respectivamente y $L \doteq \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 1 \end{pmatrix}$ y $D \doteq \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_n \end{pmatrix}$. Del hecho que $LX = D^{1/2}V$ y puesto que $V \sim \mathcal{N}(0, I_n)$ se sigue que $\pi(x)$ está dada por

$$\pi(x) = \left(\frac{det(L^T D L)}{(2\pi)^n}\right)^{1/2} \exp\left(-\frac{1}{2}\left\|D^{-1/2}Lx\right\|^2\right) \propto \exp\left(-\frac{1}{2}\left\|D^{-1/2}Lx\right\|^2\right). \tag{2}$$

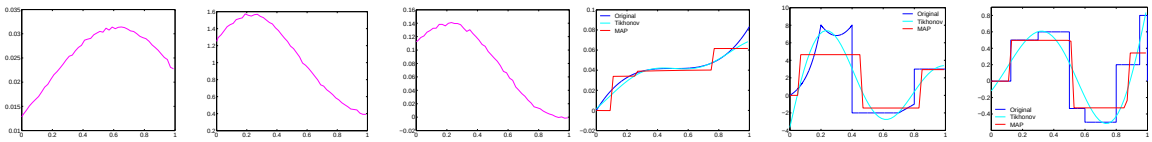
Notar que en este caso se obtiene una densidad "*a-priori*" para X del tipo "densidad estructural", donde la matriz que incorpora la información estructural que se dispone es $D^{-1/2}L$ con D y L definidas como antes. Así es que, si conociéramos la ubicación de un salto y el tamaño del mismo, incorporaríamos esta información a través de la matriz D. En el caso en que sólo dispongamos de información cualitativa, codificamos esta falta de información considerando al vector de varianzas θ como una variable aleatoria en $\mathbb{R}^n$ y la estimación de dicha variable pasa ahora a ser parte del problema. En este caso estamos frente a un modelo jerárquico de orden 1. Más precisamente, sea $\{\Theta_j, 1 \leq j \leq n\}$ un proceso estocástico para las varianzas del proceso de innovación dado en (1). Ahora la densidad de probabilidades "*a-priori*" que debemos proponer es una densidad conjunta para el par (X, Θ). Dicha densidad puede escribirse como $\pi_{pr}(x, \theta) = \pi_{pr}(x|\theta)\pi_h(\theta)$, donde $\pi_h(\theta)$ denota la distribución "*a-priori*" de la variable aleatoria Θ que da lugar al hipermodelo y expresa la falta de información en un nivel jerárquico (en este caso de nivel 1). Observar que en virtud de (2), la distribución $\pi_{pr}(x, \theta)$ está dada por $\pi(x|\theta) = \left(\frac{det(L^T D_\theta L)}{(2\pi)^n}\right)^{1/2} \exp\left(-\frac{1}{2}\left\|D_\theta^{-1/2}Lx\right\|^2\right)$, donde $D_\theta = D$ y el subíndice θ denota que las varianzas se consideran variables aleatorias. Por cuestiones de costo computacional, se introducen nuevas variables aleatorias $Z_1, Z_2, \ldots, Z_n$ que relacionan a las variables $X_1, X_2, \ldots, X_n$ del siguiente modo: $Z = LX = D_\theta^{1/2}V$. Luego, la distribución condicional $\pi(z|\theta)$ está dada por $\pi(z|\theta) = \left(\frac{det(D_\theta^{-1})}{(2\pi)^n}\right)^{1/2} \exp\left(-\frac{1}{2}\left\|D_\theta^{-1/2}z\right\|^2\right)$. Puesto que $Z = LX$ se sigue inmediatamente

que $X = L^{-1}Z$ y puede probarse que $\pi_{post}(z, \theta|y) = \pi(y|z)\pi_{pr}(z|\theta)\pi_h(\theta)$, obteniendo así la distribución "*a-posteriori*" del par (Z, Θ). La exploración de la misma permite obtener los distintos estimadores y, en función de éstos, obtener estimadores de la variable incógnita X. En los ejemplos que se presentan a continuación el estimador x_{MAP} fue calculado utilizando los algoritmos cíclicos presentados en [1].

4. Ejemplos y aplicaciones numéricas.

Consideremos el problema de recuperar una señal $f : [0, 1] \longrightarrow \mathbb{R}$ a partir de un dato g, suponiendo que f y g se relacionan a través del siguiente modelo: $(Kf)(s) \doteq \int_0^1 k(s, t)f(t)\, dt = g(s)$, donde $k \in L^2([0, 1] \times [0, 1])$ se denomina núcleo del operador integral K y está definido como $k(s, t) \doteq \exp(-4\|s - t\|^2)$. Supondremos en este ejemplo que la distribución "*a-priori*" del hiperparámetro θ está dada por una distribución Gamma inversa, es decir: $\pi_{pr}(x, \theta) \propto \prod_{j=1}^{n} \theta_j^{-\alpha-1} \exp\left(-\frac{\theta_0}{\theta_j}\right)$, donde los valores de los parámetros se fijan en $\alpha = 2,001$ y $\theta_0 = 1,7 \times 10^{-8}$.

A continuación se consideran tres señales las cuales presentan distintos grados de regularidad (en azul).



(a) (b) (c) (d) (e) (f)

Figura 1: 1(a), 1(b) y 1(c): señal perturbada para cada una de las señales originales. 1(d), 1(e) y 1(f): señal original, solución regularizada obtenida con el método de Tikhonov-Phillips y estimador x_{MAP} .

Es oportuno mencionar aquí que cuando la curva es regular en todo su dominio y mediante la distribución "*a-priori*" se incorpora información de tipo cualitativa que indica la presencia de discontinuidades, entonces claramente la solución obtenida mediante el estimador x_{MAP} no puede ser mejor que la aproximación obtenida con un método clásico. En contraposición, si la señal presenta marcadas discontinuidades, el estimador x_{MAP} las detecta con gran precisión. Este es un claro ejemplo de que haber incorporado al modelo la información cualitativa acerca de la señal se traduce en mejores soluciones aproximadas.

5. Conclusiones.

En este trabajo se mostró la estrecha relación entre el enfoque estocástico y el determinístico para resolver problemas inversos y, a través de un ejemplo numérico, se puso de manifiesto la flexibilidad de los métodos estadísticos para incorporar información cualitativa de la solución exacta siendo ésta una gran ventaja con respecto a los métodos clásicos.

Agradecimientos.

Este trabajo ha sido subvencionado en parte por el Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, a través del proyecto PIP 2010-2012 Nro. 0219, por la Universidad Nacional del Litoral, U.N.L., a través del proyecto CAI+D 2009-PI-62-315, por la Agencia Nacional de Promoción Científica y Tecnológica, ANPCyT, a través del proyecto PICT-2008-1301 y por la Air Force Office of Scientific Research, AFOSR, USA, a través de la Grant FA9550-10-1-0018.

Referencias

[1] D. Calvetti and E. Somersalo, *Hypermodels in the Bayesian imaging framework*, Inverse Problems, 24(3), 2008.

[2] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of inverse problems*, volume 375 of Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996.

[3] J. Kaipio, V. Kolehmainen, M. Vauhkonen and E. Somersalo, *Inverse problems with structural prior information.*, Inverse Problems, 15(3):713-729, 1999.

[4] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, volume 160 of Applied Mathematical Sciences. Springer-Verlag, New York, 2005.

ON THE CHOICE OF PENALIZERS IN GENERALIZED TIKHONOV-PHILLIPS REGULARIZATION METHODS.

Gisela L. Mazzieri[♭,♮], Ruben D. Spies[♭,†] and Karina G. Temperini[♭,♯]

[♭]*Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Santa Fe, Argentina.*
[♮]*Depto. de Matemática, Fac. de Bioquímica y Cs. Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[†]*Depto. de Matemática, Fac. de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[♯]*Depto. de Matemática, Fac. de Humanidades y Cs., Universidad Nacional del Litoral, Santa Fe, Argentina.*

Abstract: The Tikhonov-Phillips method is widely used for regularizing ill-posed problems due to the simplicity of its formulation as an optimization problem. The use of different penalizers in the functionals associated to the corresponding optimization problems has originated a variety of methods which can be considered as "variants" of the traditional Tikhonov-Phillips method of order zero. Such is the case for instance of the Tikhonov-Phillips method of order one, the total variation regularization method, etc. In this article we study the problem of determining general sufficient conditions on the penalizers in generalized Tikhonov-Phillips functionals which guarantee existence, uniqueness and stability of minimizers. Examples with different penalizers are presented.

Keywords: *Inverse problem, Ill-Posed, Regularization, Tikhonov-Phillips.*
2000 AMS Subject Classification: 47A52, 65J20

1 INTRODUCTION

In a quite general framework an inverse problem can be formulated as the need for determining x in an equation of the form

$$Tx = y, \tag{1}$$

where T is a linear bounded operator between two infinite dimensional Hilbert spaces X and Y (in general these will be function spaces), the range of T, $\mathcal{R}(T)$, is non-closed and y is the data, supposed to be known, perhaps with a certain degree of error. It is well known that under these hypotheses, problem (1) is ill-posed in the sense of Hadamard ([7]). In this case the ill-posedness is a result of the unboundedness of $T^\dagger$, the Moore-Penrose generalized inverse of T. The unboundedness of $T^\dagger$ has as undesired consequence the fact that small errors or noise in the data y can result in arbitrarily large errors in the corresponding approximated solutions, turning unstable all standard numerical approximation methods, making them unsuitable for most applications and inappropriate from any practical point of view. The so called "regularization methods" are mathematical tools designed to restore stability to the inversion process and consist essentially of parametric families of continuous operators approximating $T^\dagger$. Among all regularization methods, probably the best known and most commonly and widely used is the Tikhonov-Phillips method, originally proposed by Tikhonov and Phillips in 1962 and 1963 (see [8], [10], [11]). Although this method can be formalized within a very general framework by means of spectral theory ([5]), the widespread of its use is undoubtedly due to the fact that it can also be formulated in a very simple way as an optimization problem. In fact, the regularized solution of (1) obtained with Tikhonov-Phillips method is the minimizer x_α of the functional

$$J_\alpha(x) \doteq \|Tx - y\|^2 + \alpha \|x\|^2, \tag{2}$$

where α is a positive constant known as the regularization parameter. The penalizing term $\alpha \|x\|^2$ in (2) not only induces stability but it also determines certain regularity properties of the approximating regularized solutions x_α and of the corresponding least squares solution which they approximate as $\alpha \to 0^+$. This method is more precisely known as the Tikhonov-Phillips method of order zero. Choosing other penalizing terms gives rise to different approximations with different properties, approximating different least squares solutions of (1). As a consequence, it is reasonable to assume that an adequate choice of the penalizing term, based on *a-priori* knowledge of certain characteristics of the exact solution of problem (1), will lead to approximated "regularized" solutions which will appropriately reflect those characteristics. With this in mind, we shall consider functionals of the form

$$J_{W,\alpha}(x) \doteq \|Tx - y\|^2 + \alpha W(x) \quad x \in \mathcal{D}, \tag{3}$$

where $W(\cdot)$ is an arbitrary functional with domain $\mathcal{D} \subset \mathcal{X}$ and α is a positive constant. All results presented in this abstract will appear in full extent in a forthcoming article.

2 EXISTENCE AND UNIQUENESS FOR GENERAL PENALIZING TERMS

In this section we shall consider the problem of finding conditions on the penalizer $W(\cdot)$ which guarantee existence and uniqueness of global minimizers of (3). We will previously need a few definitions.

Definition 1 *Let $\mathcal{X}$ be a vector space, W a functional defined over a set $\mathcal{D} \subset \mathcal{X}$ and A a subset of $\mathcal{D}$. We say that A is W-bounded if there exists a constant $k < \infty$ such that $|W(a)| \leq k$ for every $a \in A$.*

Definition 2 *Let $\mathcal{X}$ be a vector space and W, F two functionals defined on a set $\mathcal{D} \subset \mathcal{X}$. We say that F is W-coercive if $\lim_{n\to\infty} F(x_n) = +\infty$ for every sequence $\{x_n\} \subset \mathcal{D}$ for which $\lim_{n\to\infty} W(x_n) = +\infty$.*

Definition 3 *Let $\mathcal{X}$ be a normed vector space, W, F two functionals with $Dom(F) \subset Dom(W) \subset \mathcal{X}$. We say that F is W-subsequentially (weakly) lower semicontinuous if for every W-bounded sequence $\{x_n\} \subset Dom(F)$ such that $x_n \overset{(w)}{\to} x \in Dom(F)$, there exists a subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $F(x) \leq \liminf_{j\to\infty} F(x_{n_j})$. If F is W-subsequentially lower semicontinuous we will simply say that F is W-sls. Similarly, if F is W-subsequentially weakly lower semicontinuous we will say that F is W-swls.*

In the following theorem, sufficient conditions on the functional W guaranteeing the existence and uniqueness of the minimizer of the functional (3) are established.

Theorem 1 *(Existence and uniqueness) Let $\mathcal{X}$, $\mathcal{Y}$ be normed vector spaces, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $y \in \mathcal{Y}$, $\mathcal{D} \subset \mathcal{X}$ a convex set and $W : \mathcal{D} \longrightarrow \mathbb{R}$ a functional bounded from below, W-swls, and such that W-bounded sets are relatively weakly compact in $\mathcal{X}$. More precisely, suppose that W satisfies the following hypotheses: (H1): $\exists\ \gamma \geq 0$ such that $W(x) \geq -\gamma \quad \forall\, x \in \mathcal{D}$; (H2): for every W-bounded sequence $\{x_n\} \subset \mathcal{D}$ such that $x_n \overset{w}{\to} x \in \mathcal{D}$, $\exists$ a subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $W(x) \leq \liminf_{j\to\infty} W(x_{n_j})$; (H3): for every W-bounded sequence $\{x_n\} \subset \mathcal{D}$ $\exists$ a subsequence $\{x_{n_j}\} \subset \{x_n\}$ and $x \in \mathcal{D}$ such that $x_{n_j} \overset{w}{\to} x$. Then the functional $J_{W,\alpha}(\cdot)$ in (3) has a global minimizer. If moreover the operator T is injective or W is strictly convex, then such a minimizer is unique.*

Note 1 *In the previous theorem the convexity of $\mathcal{D}$ is not needed for the existence but only for the uniqueness, in the case that T is not injective. Also, if hypotheses (H2) and (H3) on the functional W are replaced by the assumptions that W be W-sls and that W-bounded sets be relatively compact in $\mathcal{X}$, i.e. by the following hypotheses: (H2'): for every W-bounded sequence $\{x_n\} \subset \mathcal{D}$ such that $x_n \to x \in \mathcal{D}$, there exists a subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $W(x) \leq \liminf_{j\to\infty} W(x_{n_j})$; (H3'): for every W-bounded sequence $\{x_n\} \subset \mathcal{D}$ there exist a subsequence $\{x_{n_j}\} \subset \{x_n\}$ and $x \in \mathcal{D}$ such that $x_{n_j} \to x$, then both existence and uniqueness remain valid.*

Observe that hypothesis *(H1)*, *(H2)* and *(H3)* as well as *(H2')* and *(H3')* impose conditions only on the penalizer $W(\cdot)$ and not on T, so that the corresponding existence and uniqueness results hold for any bounded linear operator T. It is therefore not surprising that those conditions can be relaxed if appropriate information on T in connection to $W(\cdot)$ is provided. The next theorems shows a result in this direction.

Theorem 2 *Let $\mathcal{X}$, $\mathcal{Y}$ be normed spaces, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $D \subset X$ a convex set and W a real functional on $\mathcal{D}$. Consider the following standing hypotheses: (H2"): W is W-T-swls, i.e for every sequence $\{x_n\} \subset \mathcal{D}$ such that $\{\|Tx_n\| + W(x_n)\}$ is bounded in $\mathbb{R}$ (in the sequel we shall refer to such a sequence as a "T-W bounded sequence") and $x_n \overset{w}{\to} x \in \mathcal{D}$, $\exists$ a subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $W(x) \leq \liminf_{j\to\infty} W(x_{n_j})$. (H3"): T-W-bounded sets are relatively weakly compact in $\mathcal{X}$, i.e., for every T-W-bounded sequence $\{x_n\} \subset \mathcal{D}$ $\exists$ a subsequence $\{x_{n_j}\} \subset \{x_n\}$ and $x \in \mathcal{D}$ such that $x_{n_j} \overset{w}{\to} x$. If T and $W(\cdot)$ satisfy hypotheses (H1), (H2") and (H3"), then the functional $J_{W,\alpha}(\cdot)$ in (3) has a global minimizer. If moreover T is injective or W is strictly convex, then such a minimizer is unique.*

Note 2 *Hypotheses (H2") and (H3") are weaker that (H2) and (H3), respectively. Also note that both (H2") and (H3") hold, for instance if $\mathcal{X}$ is reflexive, $W(\cdot)$ is subsequentially weakly lower semicontinuous and T and W are complemented, i.e. $\exists$ a positive constant c such that $\|Tx\| + W(x) \geq c\,\|x\|\ \forall\, x \in \mathcal{D}$.*

3 STABILITY

It is well known that inverse ill-posed problems appear in a wide variety of applications in diverse areas. Solving these problems usually involves several steps starting from modeling, through measurements and data acquisition for the experiment under study, to the discretization of the mathematical model and the derivation of numerical approximations for the regularized solutions. All these steps entail intrinsic errors, many of which are unavoidable. For this reason, in the context of the study of inverse ill-posed problems from the optic of Tikhonov-Phillips methods with general penalizing terms, it is of particular interest the analysis of the stability of the minimizers of the functional (3) under different types of perturbations. To proceed with some results in this direction we shall need the following definitions.

Definition 4 *(Uniform W-coercivity) Let $\mathcal{X}$ be a vector space, W, F_n, $n = 1, 2, \ldots$, functionals defined on a set $\mathcal{D} \subset \mathcal{X}$. We will say that the sequence $\{F_n\}$ is uniformly W-coercive if $\lim_{n\to\infty} F_n(x_n) = +\infty$ for every sequence $\{x_n\} \subset \mathcal{D}$ for which $\lim_{n\to\infty} W(x_n) = +\infty$.*

Definition 5 *(W-consistency) Let $\mathcal{X}$ be a vector space and $W, F, F_n, n = 1, 2, ...$, functionals defined on a set $\mathcal{D} \subset \mathcal{X}$. We will say that the sequence $\{F_n\}$ is W-consistent for F if $F_n \to F$ uniformly on every W-bounded set, that is if for any given $c > 0$ and $\epsilon > 0$, there exists $N = N(c, \epsilon)$ such that $|F_n(x) - F(x)| < \epsilon$ for every $n \geq N$ and every $x \in \mathcal{D}$ such that $|W(x)| \leq c$.*

Next theorem shows a weak stability result for the minimizers of a general functional on a normed space.

Theorem 3 *(Weak stability) Let $\mathcal{X}$ be a normed vector space, $\mathcal{D}$ a subset of $\mathcal{X}$, $W : \mathcal{D} \longrightarrow \mathbb{R}$ a functional satisfying the hypotheses (H1) and (H3) of Theorem 1, J, J_n, $n = 1, 2, \ldots$, functionals on $\mathcal{D}$ such that J is W-swls and $\{J_n\}$ is uniformly W-coercive and W-consistent for J. Assume further that there exists a unique global minimizer $\bar{x} \in \mathcal{D}$ of J and that each functional J_n also possesses on $\mathcal{D}$ a global minimizer x_n (not necessarily unique). Then $x_n \overset{w}{\to} \bar{x}$.*

In the particular case in which J and J_n are of Tikhonov-Phillips type, under certain general conditions on the penalizer W, the previous theorem yields a weak stability result for the minimizers of the functional (3). In fact we have the following corollary.

Corollary 1 *Let $\mathcal{X}$ be a normed vector space, $\mathcal{Y}$ an inner product space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $y \in \mathcal{Y}$, $\alpha > 0$, $\mathcal{D}$ a subset of $\mathcal{X}$, $W : \mathcal{D} \to \mathbb{R}$ a functional satisfying hypotheses (H1), (H2) and (H3) of Theorem 1, $J, J_n, n = 1, 2, ...$, functionals on $\mathcal{D}$ defined as follows: $J(x) \doteq \|Tx - y\|^2 + \alpha W(x)$, $J_n(x) \doteq \|T_n x - y_n\|^2 + \alpha_n W(x)$, such that $\alpha_n \to \alpha$, $y_n \to y$ and $T_n x \to Tx$ as $n \to \infty$ for every $x \in \mathcal{X}$. Suppose further that J has a unique global minimizer x^*. If x_n is a global minimizer of J_n then $x_n \overset{w}{\to} x^*$.*

Slightly changing the hypotheses, analogous strong stability results as in Thm. 1 and Cor. 1 can be obtained.

4 PARTICULAR CASES

We present here two examples of penalizers $W(\cdot)$ for which some of the results obtained in the previous section are valid and therefore, results on existence, uniqueness and/or stability for the minimizers of the corresponding generalized Tikhonov-Phillips functional $J_{W,\alpha}(\cdot)$ in (3) are obtained.

4.1 TOTAL VARIATION PENALIZATION

Bounded variation penalty methods have been studied by Rudin, Osher and Fatemi in 1992 ([9]) and Acar and Vogel in 1994 ([1]), among others. These methods have been proved highly successful in certain image denoising problems where edge preserving is an important issue ([2], [3], [4], [6]). Let $d \geq 2$,

$\Omega \subset \mathbb{R}^d$ a convex, bounded subset with Lipschitz continuous boundary, $1 \leq p \leq \frac{d}{d-1}$, $\mathcal{X} \doteq L^p(\Omega)$, $\mathcal{D} \doteq BV(\Omega)$, where $BV(\Omega)$ is the space of functions of bounded variation on Ω. Recall that the BV norm of u is defined by $\|u\|_{BV(\Omega)} \doteq \|u\|_{L^1(\Omega)} + J_0(u)$, where $J_0(u) \doteq \sup_{v \in \mathcal{V}} \int_\Omega (-u \, div \, v) \; dx$ and $\mathcal{V} \doteq \left\{ v \in C_0^1(\Omega; \mathbb{R}^d) \; : \; |v(x)| \leq 1 \text{ for all } x \in \Omega \right\}$. Let W be the functional defined on $\mathcal{D}$ by $W(u) \doteq \|u\|_{BV(\Omega)}$. It can be shown that W satisfies the hypotheses *(H1), (H2)* y *(H3)* of Theorem 1 and therefore for every $\alpha > 0$, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ ($\mathcal{Y}$ a normed space), the functional $J_{\|\cdot\|_{BV}, \alpha}(u) \doteq \|Tu - v\| + \alpha \|u\|_{BV}$ has a global minimizer on $BV(\Omega)$. If T is injective then such a minimizer is unique. Moreover, it can be shown that the problem of finding such a minimizer is strongly stable under perturbations in T, v and α for $p < \frac{d}{d-1}$, while for $p = \frac{d}{d-1}$ and $d \geq 2$ the problem is only weakly stable.

4.2 Penalization with semi-norms associated to closed operators

Let $\mathcal{X}$, $\mathcal{Z}$ be reflexive Banach spaces, $\mathcal{Y}$ a normed space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $L : \mathcal{D}(L) \subset \mathcal{X} \to \mathcal{Z}$ a closed linear operator such that the range of L, $\mathcal{R}(L)$, is weakly closed (this includes most differential operators). Assume further that T and L are complemented, i.e. $\exists \, \gamma > 0$ such that $\|Tx\| + \|Lx\| \geq \gamma \, \|x\|$ for all $x \in \mathcal{D}(L)$. Let $W_L : \mathcal{D}(L) \longrightarrow \mathbb{R}_0^+$ be defined by $W_L(x) = \|Lx\|^2$. Then it can be shown that W_L satisfies hypotheses *(H1), (H2")* y *(H3")* of Theorem 2. Hence, for any $\alpha > 0$, the functional $J_{L, \alpha}(x) \doteq \|Tx - y\| + \alpha \|Lx\|^2$ has a unique global minimizer on $\mathcal{D}(L)$. We show that this problem is weakly stable, while strong stability can be achieved by imposing slightly stronger hypotheses on T and L.

5 Conclusions

Sufficient conditions on the model operator and the penalizing term in generalized Tikhonov-Phillips functionals, guaranteeing existence and uniqueness of solutions were found. Stability results were given. Two examples were presented.

Acknowledgments

This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2010-2012 Nro. 0219, by Univ. Nac. del Litoral, U.N.L., through project CAI+D 2009-PI-62-315, by Agencia Nacional de Promoción Científica y Tecnológica, ANPCyT, through project PICT-2008-1301 and by the Air Force Office of Scientific Research, AFOSR, through Grant FA9550-10-1-0018.

References

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] A. Chambolle and J. L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[3] T. Chan, A. Marquina, and P. Mullet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516.

[4] T. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.

[5] R. Dautray and J.-L. Lions, *Mathematical analysis and numerical methods for science and technology. Vol. 3: Spectral Theory and Applications*, Springer-Verlag, Berlin, 1990.

[6] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.

[7] J. Hadamard, *Sur les problmes aux drives partielles et leur signification physique*, Princeton University Bulletin, 13 (1902), pp. 49–52.

[8] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comput. Mach., 9 (1962), pp. 84–97.

[9] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.

[10] A. N. Tikhonov, *Regularization of incorrectly posed problems*, Soviet Math. Dokl., 4 (1963), pp. 1624–1627.

[11] ———, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4 (1963), pp. 1035–1038.

APLICACIÓN DEL PROBLEMA DE MOMENTOS PARA RESOLVER UNA ECUACIÓN EN DERIVADAS PARCIALES

María Beatriz Pintarelli† y Fernando Vericat‡

Grupo de Aplicaciones Matemáticas y Estadísticas de la Facultad de Ingeniería (GAMEFI), Universidad Nacional de La Plata

† Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina, mariabea@mate.unlp.edu.ar

‡ Instituto de Física de Líquidos y Sistemas Biológicos(IFLYSIB) CONICET-La Plata, Argentina.

Resumen: Varios investigadores, (por ejemplo [2] cap. 7), han considerado problemas relacionados con la ecuación

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial t^2}-\frac{\partial u}{\partial t}=0 \qquad \forall (x,y,t)\in R^2\times R_+$$

y los han resuelto aplicando técnicas de problema de momentos. Además es conocido que puede aplicarse la transformación de Laplace para resolver ecuaciones en derivadas parciales, como es el caso de la ecuación

$$\frac{\partial^2 u}{\partial t^2}-a^2\frac{\partial^2 u}{\partial x^2}=0$$

con a constante y condiciones iniciales

$$u(0,t)=F(t), \quad u(x,0)=0, \quad \frac{\partial}{\partial t}u(x,0)=0, \quad \lim_{x\to+\infty}u(x,t)=0\,.$$

Consideramos aquí el problema de hallar una función $F(t,x)$ tal que

$$\frac{\partial F(t,x)}{\partial t}=-G(t)\frac{\partial F(t,x)}{\partial x} \qquad (t,x)\in R^+\times R^+ \qquad (1)$$

una ecuación en derivadas parciales de primer orden cuasilineal, con las condiciones de contorno

$$F(t,0)=g(t) \qquad F(0,x)=F_{\text{inicial}}(x)\,.$$

En este trabajo consideramos un enfoque, a nuestro entender novedoso, que combina ambas técnicas para resolver este problema.
Utilizando la transformada bidimensional de Laplace se transformará (1) en un problema de momentos de Hausdorff bidimensional.
Se acotará el error de la solución estimada utilizando las técnicas sobre problema de momentos bidimensional.
También se mostrará que (1) es equivalente a resolver una ecuación integral de Fredholm de primera especie.

Palabras claves: *problema de momentos, densidad bi-dimensional, estabilidad de la solución, ecuaciones integrales.*
2010 AMS Subjects Classification: 44A60 – 49N45

1. INTRODUCCIÓN

El problema de momentos de Hausdorff bidimensional consiste en recobrar una función $f(x,y)$ dados sus momentos

$$\mu_{ij}=\iint_I x^i y^j f(x,y)dxdy \qquad i,j=0,1,2,.... \qquad (2)$$

donde $I=(0,1)\times(0,1)$.
Existe solución cuando $\sum_i\sum_j a_{ij}\mu_{ij}>0$ para todo polinomio $P(x,y)=\sum_i\sum_j a_{ij}x^i y^j$ tomando valores no negativos para todo $(x,y)\in I$ ([2],cap. 1, teorema 1.1).

Si consideramos soluciones para este problema que pertenecen a $L^2(I)$, un teorema de exactitud de la solución aproximada por el método de expansión truncada, asumiendo que los datos $\{\mu_{ij}\}$ de la ecuación (2) podrían contener algún error, sería el siguiente [1], [3].

2. TEOREMA DE EXACTITUD

Teorema 1 *Supongamos que la función $f(x,y)$ en $L^2(I)$ verifica para algún N, ε y E las condiciones*

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\int_0^1\int_0^1 x^i y^j f(x,y)dxdy-\mu_{ij}\right|^2\le\varepsilon^2 \qquad y \qquad \int_0^1\int_0^1\left[f_x^2(x,y)+f_y^2(x,y)\right]dxdy\le E^2\,.$$

Entonces $\int_0^1\int_0^1 |f(x,y)-p_N(x,y)|^2\,dxdy \le \min\left\{\varepsilon^2 tr\left(UU^T\right)\sum_{i=1}^{N^*}\sum_{j=1}^{N^*}\left|\left(C_{ij}\right)\right|^2+\frac{E^2}{2(N^*+1)}, N^*=1,2,...,N\right\}$

donde $p_N(x,y)=\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij}\varphi_{ij}(x,y)$ *siendo* $\{\varphi_{ij}\}$ *base ortonormal de polinomios de Legendre*

$\lambda_{ij}=tr\left(U^T C_{ij}\right)$ $\quad U=\left(\mu_{ij}\right)_{i,j=1,2,...,N}$ $\quad C_{ij}$ *matriz cambio de base*

$p_N(x,y)=\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij}\varphi_{ij}(x,y)$ *sería la solución aproximada por el método de expansión truncada.*

3. RESOLUCIÓN DEL PROBLEMA UTILIZANDO TRANSFORMADA DE LAPLACE

Escribimos la ecuación diferencial (1) como $u_t=-G(t)u_x$. Asumimos que $G(t)$ tiene derivada continua. Si aplicamos la transformada de Laplace con respecto a *x*:

$$L_x(u_t)=-G(t)L_x(u_x) \qquad (3)$$

donde

$$L_x(u)=\int_0^\infty u(t,x)e^{-\xi x}dx=U(t,\xi) \qquad (4).$$

Entonces como

$$L_x(u_x)=\xi\,L_x(u)-u(t,0) \qquad \text{y} \qquad L_x(u_t)=\frac{\partial}{\partial t}L_x(u)$$

reemplazando en (3) tenemos:

$$\frac{\partial}{\partial t}L_x(u)=-G(t)\left[\xi\,L_x(u)-u(t,0)\right].$$

O también

$$\frac{\partial}{\partial t}U(t,\xi)=-G(t)\left[\xi\,U(t,\xi)-u(t,0)\right].$$

Considerando a ξ fijo, escribimos

$$U'(t,\xi)=-G(t)\left[\xi\,U(t,\xi)-u(t,0)\right].$$

Tenemos entonces una ecuación diferencial ordinaria en la variable *t* de la forma

$$U'(t,\xi)+P(t,\xi)\,U(t,\xi)=Q(t)$$

donde

$$P(t,\xi)=\xi G(t) \quad \text{y} \quad Q(t)=G(t)u(t,0).$$

La solución general de esta ecuación diferencial es

$$U(t,\xi)=\frac{1}{I(t,\xi)}\left[\int I(t,\xi)Q(t)dt+C\right] \qquad \text{siendo} \qquad I(t,\xi)=e^{\int P(t,\xi)dt}.$$

Por lo tanto

$$U(t,\xi)=\frac{1}{I(t,\xi)}\left[\int I(t,\xi)G(t)u(t,0)dt+C\right] \qquad I(t,\xi)=e^{\int \xi G(t)dt}.$$

Para determinar C planteamos

$$U(0,\xi)=L_x(u(0,x))=\int_0^\infty u(0,x)e^{-\xi x}dx=\frac{\int I(t,\xi)G(t)u(t,0)dt\Big|_{t=0}+C}{I(0,\xi)} \quad (\text{si } I(0,\xi)\neq 0).$$

Luego de hallar $U(t,\xi)$ multiplicamos ambos miembros de (4) por e^{-st} e integramos con respecto a *t*:

$$\int_0^\infty\int_0^\infty u(t,x)e^{-\xi x}e^{-st}dxdt=\int_0^\infty U(t,\xi)e^{-st}dt=UI(s,\xi) \qquad (5).$$

Por lo tanto $UI(s,\xi)$ es la transformada bidimensional de Laplace de $u(t,x)$.

Consideramos el cambio de variable $v = e^{-t}$ y $w = e^{-x}$. Entonces la transformada (5) puede ser escrita

$$\int_0^1\int_0^1 v^{s-1}w^{\xi-1}u(-\ln v,-\ln w)dvdw = \int_0^1\int_0^1 v^{s-1}w^{\xi-1}f(v,w)dvdw = UI(s,\xi) \qquad (6).$$

Para $s = m+1,\ \xi = n+1$ la ecuación (6) da

$$\int_0^1\int_0^1 v^m w^n f(v,w)dvdw = UI(m+1,n+1) = \mu_{mn} \qquad m,n = 0,1,2,... \qquad (7).$$

Entonces (1) es equivalente a la inversión de la transformada de Laplace (5) , y la inversión de ésta es equivalente al problema de momentos de Hausdorff bidimensional (7).

En el caso de ser $f(v,w) \in L^2(I)$ se resuelve el problema con el método de la expansión truncada donde la solución aproximada será $p_N(v,w) = \sum_{m=1}^{N}\sum_{n=1}^{N}\lambda_{mn}\varphi_{mn}(v,w)$, con $\varphi_{mn}(v,w) = l_m(v)l_n(w)$, $l_m(v)$ polinomio de Legendre en $(0,1)$ de grado *m*.

Entonces $q_N(t,x) = p_N(e^{-t},e^{-x})$ aproxima la solución de (5), por lo tanto también aproxima la solución de (1).

Aplicando el teorema de estabilidad se llega a

$$\int_0^\infty\int_0^\infty \left|u(t,x) - q_N(t,x)\right|^2 dtdx \le \min\left\{\varepsilon^2 tr\left(UU^T\right)\sum_{i=1}^{N^*}\sum_{j=1}^{N^*}\left|\left(C_{ij}\right)\right|^2 + \frac{E^2}{2(N^*+1)}, N^* = 1,2,...,N\right\}$$

donde E^2 es una cota para

$$\int_0^\infty\int_0^\infty \left[(e^t)^2 u_t^2(t,x) + (e^x)^2 u_x^2(t,x)\right]e^{-t}e^{-x}dtdx\ .$$

4. EJEMPLO NUMÉRICO

Partimos de la función $F(t,x) = ce^{\left(-3t^2 + x/12\right)}$, donde *c* es la constante de normalización, la cual satisface (1) con $G(t) = 6t$. De esta forma consideramos las condiciones iniciales $F(t,0) = ce^{\left(-3t^2/12\right)} = g(t)$, $F(0,x) = ce^{\left(x/12\right)} = F_{inicial}(x)$. Entonces el problema sería hallar $F(t,x)$ tal que sea solución de

$$\frac{\partial F(t,x)}{\partial t} = -6t\frac{\partial F(t,x)}{\partial x} \qquad (t,x) \in R^+ \times R^+$$

con las condiciones iniciales

$$F(t,0) = ce^{\left(-3t^2/12\right)} = g(t) \qquad \text{y} \qquad F(0,x) = ce^{\left(x/12\right)} = F_{inicial}(x)\ .$$

Buscamos la solución $U(t,\xi)$. Luego hallamos $UI(s,\xi)$ según (5).

Tomamos $m,n = 0,1,2$ y aplicamos el algoritmo correspondiente al método de expansión truncada. Se encuentra una aproximación $p_N(v,w)$ para $f(v,w)$ en (6).

En la figura 1 superponemos las gráficas de $p_N(v,w)$ y $F(-\ln(v),-\ln(w))$.

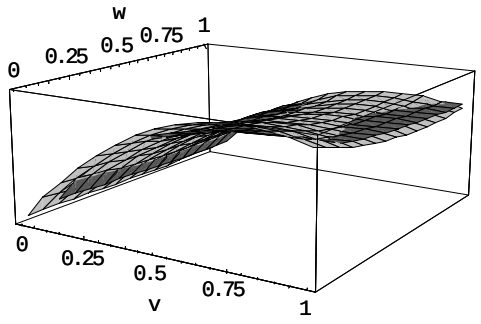


Figura 1: función aproximada ■ y función exacta ■

Otra forma de escribir (1) como una ecuación integral sería la siguiente.

Nuevamente escribimos la ecuación diferencial (1) como $u_t = -G(t)u_x$.

Consideramos la función

$$R(x,\tau,\xi) = e^{-x(\xi+1)-xGp(\tau)} \qquad \text{con} \qquad Gp(\tau) = \int G(\tau)d\tau .$$

Entonces $R_\tau - R_\xi xG(\tau) = 0$.

Y luego de algunos cálculos podemos escribir

$$\nabla^2 R = -RxG'(\tau) + Rx^2(1+G(\tau)^2) .$$

Aplicamos la primera identidad de Green:

$$\iint_D u\nabla^2 R d\xi d\tau = \oint_C u\nabla R n ds - \iint_D \nabla u \nabla R d\xi d\tau$$

donde $D = [0,M]\times[0,M]$ y C es el borde de D.

Desarrollando cada término, tomando límite para $M \to \infty$ y asumiendo que: $u(\tau,\xi)$ acotada ,

$\int_0^M u(M,\xi)R(x,M,\xi)G(M)d\xi \xrightarrow[M\to\infty]{} 0$, $\int_0^M u(\tau,M)R(x,\tau,M)xd\tau \xrightarrow[M\to\infty]{} 0$

se llega a la igualdad

$$\int_0^\infty\int_0^\infty u(\tau,\xi)R(x,\tau,\xi)\left[2G(\tau)^2x^2 - xG'(\tau)\right]d\tau d\xi =$$

$$= xG(0)\int_0^\infty u(0,\xi)R(x,0,\xi)\,d\xi - \int_0^\infty u(\tau,0)R(x,\tau,0)\left[\left(1-G(\tau)^2\right)x+1\right]d\tau \qquad (8).$$

La expresión anterior es una ecuación integral de la forma

$$\int_0^\infty\int_0^\infty u(\tau,\xi)K(x,\tau,\xi)\,d\tau d\xi = \varphi(x) \qquad (9).$$

Si $u(\tau,\xi)$, $K(x,\tau,\xi)$ y $\varphi(x)$ son funciones de cuadrado integrable, entonces una forma de resolver (8) consistiría en tomar una base arbitraria $\psi_m(x)$ de $L^2(0,\infty)$ y resolver el problema de momentos generalizado

$$a'_m = \int_0^\infty\int_0^\infty b'_m(\tau,\xi)u(\tau,\xi)d\tau d\xi \quad m = 1,2,... \qquad (10)$$

con

$$a'_m = \int_0^\infty \varphi(x)\chi_m(x)dx \qquad b'_m(\tau,\xi) = \int_0^\infty K(x,\tau,\xi)\psi_m(x)dx .$$

Se puede probar que resolver (9) es equivalente a resolver (10). Se encuentra una solución aproximada de (10) para $m = 1,2,...,N$ bajo la suposición que las $b'_m(\tau,\xi)$ son linealmente independientes.

CONCLUSIONES

El método clásico de la Transformada de Laplace para resolver ecuaciones en derivadas parciales implica tener que hallar la correspondiente antitransformada. En este trabajo proponemos un método alternativo basado en técnicas de problema de momentos que permiten hallar, bajo ciertas condiciones, una aproximación de la solución que no requiere de la antitransformada y también una cota para el error. Un segundo camino que hemos explorado consiste, en vez de usar Transformada de Laplace, en transformar la ecuación diferencial en una ecuación integral la cual puede intentar resolverse aplicando también técnicas de problema de momentos.

REFERENCIAS

[1] D.D. ANG, R. GORENFLO, V.K. LE and D.D. TRONG, *Moment theory and some inverse problems in potential theory and heat conduction*, Lectures Notes in Mathematics, Springer-Verlag, Berlin Heidelberg, 2002.

[2] J.A. SHOHAT and J.D. TAMARKIN, *The problem of Moments*, Mathematic Surveys, Am. Math. Soc., Providence, RI, 1943

[3] G. TALENTI, *Recovering a function from a finite number of moments*, Inverse Problems 3 (1987), pp.501- 517.

A NOTE ON OPTIMAL DESIGN METHODS FOR PARAMETER ESTIMATION

M.I. Troparevsky[♭], D. Rubio[†] and N. Saintier[‡]

[♭]*Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, Argentina, mitropa@fi.uba.ar,*
[†]*Centro de Matemática Aplicada, Universidad Nacional de San Martín, M. de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina, drubio@unsam.edu.ar*
[‡]*Instituto de Ciencias, Universidad Nacional Gral. Sarmiento, Buenos Aires, Argentina,nsaintie@dm.uba.ar*

Abstract: Mathematical modelling became an important tool for the prediction of the results of processes arising in a variety of disciplines. Once the mathematical system to describe a particular process has been chosen, its parameters must be carefully selected in order to be able to simulate the process. In this context optimal design techniques can help to determine where to measure the outputs of the system in order to accurately estimate the parameter values. In this work we compare two different optimal design criterion and apply them to the Baranyi model for bacterial growth. A numerical example is presented considering a set of data for Salmonellae growth.

Keywords: *Optimal Design, Parameter estimation, Baranyi bacterial growth model.*
2000 AMS Subject Classification: 93A30, 93B35, 92B99

1 INTRODUCTION

The problem of parameter estimation consists in selecting accurate values for the parameters of a mathematical model that describes a process under study, i.e., to fit a parametric mathematical model to experimental data.
Consider a dynamical model that depends on some parameter θ_0. A well known method to estimate the modeling parameter based on real data $y_1, ..., y_N$ measured at time instants $t_1, .., t_N$ respectively, consists in minimizing the functional $J = \sum_{i=1}^{N}(y_i - y_i(\theta))^2$. The estimator of the parameter is then $\hat{\theta} = \arg\min J$. Optimal design techniques consist in seeking N instants, for N fixed, where the output has more information to recover the modeling parameters accurately. These tools are usually based on the sensitivity of the system with respect to the parameter. A good description of these methods can be found in [2],[7],[8]
In this work we compare the performance of a continuous and a discrete D-optimal criterion for the parameter estination of a bacterial growth model. Specifically, we present a numerical example considering a set of data for Salmonellae growth using the Baranyi model [4].

2 MATHEMATICAL FRAMEWORK

Consider a process described by

$$\begin{aligned} x'(t) &= f(t, x(t), \theta) \\ y(t) &= h(t, \theta) + \epsilon \end{aligned} \tag{1}$$

where $\theta \in R^p$ is the unknown parameter that we want to estimate, $x(t) \in R^n$ is the state variable, y are the outputs of the system, $f : R^{1+n+p} \to R^n$ and $h : R^{1+p} \to R$ are smooth funcions (twice continuously differentiable wth respect to θ), and ϵ is a realization of a random variables with media 0 and variance σ^2. We suppose that there exist a real value of θ, that we call θ_0, such that the system (1) describes the process. If we assume that we can obtain N observations $y_1, ..., y_N$ at time instants $t_1, .., t_N$ with N fixed, the modeling parameter can be estimated as $\hat{\theta} = \arg\min J(\theta)$ with $J(\theta) = \sum_{i=1}^{N}(y_i - y_i(\theta))^2$. According to [9] under some assumptions $\hat{\theta}(t_1, .., t_N)$ is approximately and asymptotically normally distributed with mean θ_0 and its covariant matrix is the inverse of the Fisher information matrix $F(t_1, ., t_N, \theta_0) = (F_{ij}(t_1, ., t_N, \theta_0))$ defined by

$$F_{ij}(t_1, ., t_N, \theta_0) = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial h}{\partial \theta^i}(t_n, \theta_0)\frac{\partial h}{\partial \theta^j}(t_n, \theta_0)\,\frac{1}{\sigma^2}.$$

Optimal design techniques consist in seeking the N instants $t_1, .., t_N$ where the output has more information to recover the modeling parameters accurately. A popular optimal criteria, known as D-optimal, consists in looking for $t_1, .., t_N$ such that the determinant of $F(t_1, .., t_N, \theta)^{-1}$ is minimized.

In [3] can be found the following continuous version of the functional J

$$J(\theta, P) = \int_0^T (y(t) - y(t, \theta))^2 dP(t), \tag{2}$$

where P is a probability measure over $[0, T]$. Note that the discrete J can be recovered by taking $P = \frac{1}{N}\sum_{i=1}^{N} \delta(t - t_i)$ where δ denotes the Dirac delta functional.
According to [3], the associated estimator $\hat{\theta}(P)$ is approximately and asymptotically normally distributed with mean θ_0 and its covariant matrix is the inverse of the Fisher information matrix $F(P, \theta_0) = (F_{ij}(P, \theta_0))$ which now reads as

$$F_{ij}(P, \theta_0) = \int_0^T \frac{\partial h}{\partial \theta^i}(t, \theta_0) \frac{\partial h}{\partial \theta^j}(t, \theta_0) \frac{dP(t)}{\sigma^2}.$$

The problem is to choose a probability measure P^* that minimizes the covariance matrix in some sense. In this case the D-optimal criterion consists in looking for a probability measure P^* such that the determinant of $F(P, \theta_0)^{-1}$ is minimized.
In [7] and [8] Molchanov and Zuyev presented a steepest-descent algorithm to calculate P^*. This method considers the space of measure with finite total variation and takes into account the expression of the directional derivative of $P \to -\ln(\det(F(P, \theta_0)))$ at P in the direction η, where η is a signed measure of zero mass. Moreover since this map is convex, the convergence of the algorithm is assured.
In the next section we apply both tecniques to the Baranyi model for bacterial growth (see [4]) and present a numerical example for a set of data for Salmonellae growth.

3 NUMERICAL EXAMPLE

In this section we apply the previous tools to the problem of bacterial growth.
A well-known and widely used mathematical model for this kind of problems is the Baranyi model (see [4]) that combines a logistic model with the Michaelis-Menten one:

$$\begin{aligned} y(t) &= y_0 + \mu_{max} t + \frac{\ln(e^{-\nu t} + e^{-h_0} - e^{-\nu t - h_0})}{\mu_{max}} \\ &\quad -\frac{1}{m}\ln(1 + \frac{e^{m\mu_{max} t + \ln(e^{-\nu t} + e^{-h_0} - e^{-\nu t - h_0})/\mu_{max}} - 1}{e^{m(y_{max} - y_0)}}) \end{aligned} \tag{3}$$

where $x(t)$ is the cell concentration (in CFU/ml), $y(t) = \ln(x(t))$, $y_0 = \ln(x_0)$, $y_{max} = \ln(x_{max})$, being x_0 the initial value and x_{max} the aymptotic value of cell concentration respectively, μ_{max} is the maximun specific growth rate, m is a curvature parameter to characterise the transition from the exponential phase, ν is a curvature parameter to characterise the transition to the exponential phase and h_0 is a dimensionless parameter that indicates the initial physiological state of the cells.
As suggested in [5] we consider $\nu = \mu_{max}$ and $m = 1$ reducing the parameters to y_0, y_{max}, h_0 and μ_{max}.
For the numerical example we consider a set of epxerimental data related to the growth of Salmonellae bacteria extracted from [6] (see table 3). Using least-square method with these data the estimated values for the parameters are

$$(y_0, y_{max}, h_0, \mu_{max}) = (3.32568352835700, 8.92431264170258, 0.33389230974371, 0.28665640784088).$$

The inverse of the Fisher information matrix F is assimptotically related to the covariance matrix of the least-square estimator of the parameters. In particular $\det(F^{-1})$ gives the volume the confidence elipsoide

time	0	1.17	2	2.92	3.92	4.96	5.96	8.08	10.2	13.1	19.8
log(con)	3.39	3.39	3.47	3.46	3.57	3.70	3.98	5.41	4.96	5.74	7.45
time	21.3	22.8	23.8	24.7	26.7	27.7	28.8	29.7	31.3	32.8	49.8
log(con)	7.79	8.10	8.24	8.46	8.69	8.66	8.67	9.16	8.69	8.76	8.78

Table 1: Data for Salmonellae growth from [6]

and the square root of the diagonal part of F is proporcional to the confidence interval of the parameters. With this estimation of the parameters we obtain

$$\det(F^{-1}) = 37.06338129660474 \tag{4}$$

and

$$\sqrt{diag(F^{-1})} = (8.68499074709278, 0.51854847523453, 2.18156793111694, 3.52155853078053). \tag{5}$$

Figure 1 shows the experimental data together with the Baranyi fitting curve obtained with these parameters values.

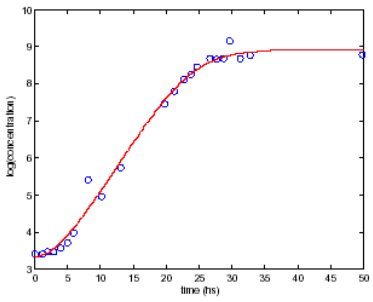


Figure 1: Experimental Data and Baranyi's Fitting Curve.

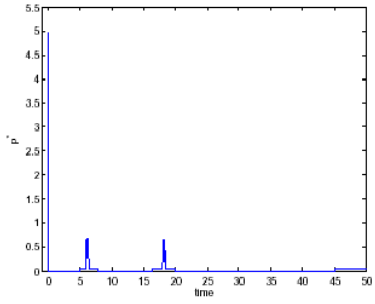


Figure 2: P^* using Molchanov and Zuyev algortihm.

We apply the Molchavnov-Zuyev algorithm to this model The resulting probability function P^* is shown in Figure 2. Note that $\text{supp}(P^*) \subset \{0\} \cup [5, 7.6] \cup [16.5, 19.8] \cup [45, 50]$.
Now, we look for the estimated parameter values obtained by using $N = 4$ observation points. We apply the D-optimal criteria for $N = 4$ considering the previously estimated values for the parameters as initial values. The time instants that we obtained are: $t_1 = 0, t_2 = 6.215, t_3 = 18.1248, t_4 = 50$. Based on the available data, we choose $t_1 = 0, t_2 = 5.96, t_3 = 19.8, t_4 = 49.8$ and estimate the parameter values obtaining:

$$(y_0, y_{max}, h_0, \mu_{max}) = (3.38999999999165, 8.78041597157280, 0.42888760115012, 0.29378328654967).$$

Calculating the corresponding Fisher matrix we obtain

$$\det(F^{-1}) = 0.04563482607606 \tag{6}$$

and

$$\sqrt{diag(F^{-1})} = (2.89389509172062, 0.20744700883892, 1.00064498898728, 1.00000000000000). \tag{7}$$

In Figure 3 we plot the Baranyi curve considering the latter parameter estimated values labeled as D-optimal (dottted curve).

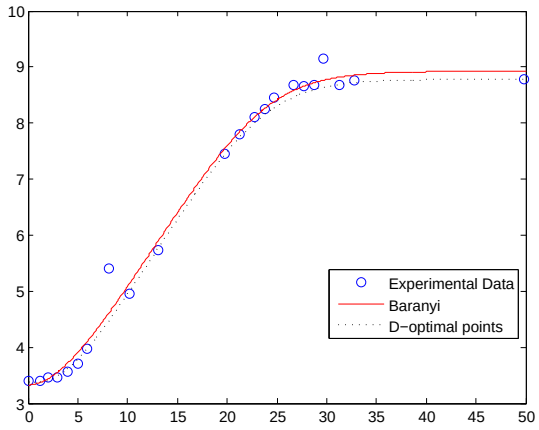


Figura 3: Data and Baranyi curves

4 CONCLUSIONS

We observe that by calculating P^* with Molchavnov-Zuyev algorithm, we obtain a probability function whose support can be expressed as the union of four intervals. Note that each of these intervals contain one of the four *more informative* time instants yield by the D-optimal criteria. Moreover comparing the determinant and the diagonal part of the inverse of the Fisher information matrix (4)-(5) and (6)-(7) obtained using the values of the parameters estimated using the least-square minimization procedure with respectively all the available experimental data and the four most informative data according to the D-optimal criterion, we observe that the latter yields the most reliable estimation. This shows that using all the available data might not be the most efficient way of estimating the parameters of the model since different data may provide redundant information having as a consequence a more degenerate Fisher information matrix.

ACKNOWLEDGMENTS

This work has been supported in part by AFOSR Grant FA9550-10-1-0037.

REFERENCES

[1] H.T. BANKS AND K.L. BIHARI, *Modeling and estimating uncertainty in parameter estimation*, Inverse Problem, 17 (2001), pp.95-111.

[2] H.T. BANKS, S. DEDIU AND S.L. ERNSTBERGER, *Sensitivity functions and their uses in inverse problems*, J. Inverse and Ill-posed Problems, 15 (2007), pp.683-708.

[3] H.T. BANKS, S. DEDIU, S.L. ERNSTBERGER AND F. KAPPEL, *A new optimal approach to optimal design problem*, J. inverse and ill-posed problems, 18 (2010), pp.25-83.

[4] J. BARANYI, T.A ROBERTS AND P MCCLURE, *A non-autonomous differential equation to model bacterial growth*, Food Microbiol., 10 (1993), pp. 4359.

[5] J. BARANYI AND T.A ROBERTS, *Mathematics of predictive food microbiology*, Int. J. Food Microbiol. 26, (1995) pp.199-218.

[6] A.M.GIBSON, N. BRATCHELL AND T.A ROBERTS, *Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by pH, sodium chloride and storage temperature*, Int. J. Food Microbiol., 6 (1988), pp.155-178.

[7] I. MOLCHANOV AND S. ZUYEV, *Steepest descent algorithms in space of measure*, Statistics and computating, 12 (2002), pp.115-123.

[8] I. MOLCHANOV AND S. ZUYEV, *Optimization in space of measures and optimal design*, ESAIM: Probability and statistics, 8 (2004), pp.12-24.

[9] G.A.F. SEBER AND C.J. WILD, *Nonlinear regression*, Wiley Intersciences, Hoboken NJ, 2003.

RESULTS ON THE EXISTENCE OF SATURATION FOR REGULARIZATION METHODS WITH OPTIMAL QUALIFICATION

Gisela Luciana Mazzieri[♭,⋆], Rubén Daniel Spies[♭,◇] and Karina Guadalupe Temperini[♭,∗]

[♭]*Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Santa Fe, Argentina.*
[⋆]*Dpto. de Matemática, Fac. de Bioquímica y Cs. Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[◇]*Dpto. de Matemática, Fac. de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.*
[∗]*Dpto. de Matemática, Fac. de Humanidades y Cs., Universidad Nacional del Litoral, Santa Fe, Argentina.*

Abstract: In this work we provide sufficient conditions for a spectral regularization method with optimal qualification to possess global saturation. Appropriate characterization of both the saturation function and the saturation set are given. Examples and converse results are included.

Keywords: *Inverse problem, qualification, saturation.*
2000 AMS Subject Classification: 47A52, 65J20.

1 INTRODUCTION

In 1994, A. Neubauer [5] showed that certain spectral regularization methods for ill-posed inverse problems "saturate", that is, from the knowledge of a certain degree of regularity of the exact solution of the problem, they become unable to keep on extracting further information about it, independently of the additional hypotheses imposed. The concept of "saturation" is closely associated to the best order of convergence of the total error that the method can achieve independently of the regularity assumptions on the exact solution and on the choice of the regularization parameter. In 2011 Herdman, Spies and Temperini [3] developed a general theory of global saturation for arbitrary regularization methods, formalizing Neubauer's original and intuitive idea.

Related in a somewhat dual way to the concept of saturation is the concept of qualification of a spectral regularization method, which was introduced by Mathé and Pereverzev in 2003 [4]. This concept is strongly related to the optimal order of convergence of the regularization error, under certain "a-priori" assumptions on the exact solution. In 2009 Herdman, Spies and Temperini [2] generalized the concept of qualification and introduced three hierarchical levels of it: weak, strong and optimal qualification. It was shown in [2] that the weak qualification generalizes the definition introduced in [4].

In this work we will shed some light on the existence of saturation for spectral regularization methods with optimal qualification. In particular, we will establish sufficient conditions on the family of functions $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ defining the method and on the optimal qualification ρ, which guarantee the existence of saturation. Moreover, in those cases, we will provide appropriate characterizations of both the saturation function and the saturation set.

2 PRELIMINARIES

Let X, Y be infinite dimensional Hilbert spaces and $T : X \to Y$ a bounded linear operator such that $\mathcal{R}(T)$ is not closed. It is well known that under these conditions, the linear operator equation $Tx = y$ is ill-posed, in the sense that $T^\dagger$, the Moore-Penrose generalized inverse of T, is not bounded [1].

Let $\{E_\lambda\}_{\lambda\in\mathbb{R}}$ be the spectral family associated to the linear selfadjoint operator T^*T and $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ a parametric family of functions, $g_\alpha : [0, \|T\|^2] \to \mathbb{R}$, $\alpha \in (0, \alpha_0)$. Consider also the following standing hypotheses:

H1: For every $\alpha \in (0, \alpha_0)$ the function g_α is piecewise continuous on $[0, \|T\|^2]$.
H2: There exists a constant $C > 0$ (independent of α) such that $|\lambda g_\alpha(\lambda)| \leq C$ for every $\lambda \in [0, \|T\|^2]$.
H3: For every $\lambda \in (0, \|T\|^2]$, there exists $\lim_{\alpha\to 0^+} g_\alpha(\lambda) = \frac{1}{\lambda}$.

It is known that if $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ satisfies hypotheses *H1-H3*, then (see [1], Theorem 4.1) the collection of operators $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$, where $R_\alpha \in \mathcal{L}(Y, X)$ is defined by $R_\alpha y \doteq \int_0^{\|T\|^2+} g_\alpha(\lambda)\, dE_\lambda\, T^*y =$

$g_\alpha(T^*T)T^*y$, is a family of regularization operators for $T^\dagger$. In this case we say that $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$ is a "family of spectral regularization operators" (FSRO) for $Tx = y$ and $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ is a "spectral regularization method" (SRM).

We denote with $\mathcal{O}$ the set of all non-decreasing functions $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\lim_{\alpha\to 0^+} \rho(\alpha) = 0$ and with $\mathcal{S}$ the set of all continuous functions $s : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ satisfying $s(0) = 0$ and such that $s(\lambda) > 0$ for every $\lambda > 0$. Note that if $s \in \mathcal{S}$ is non-decreasing, then s is an *index function* in the sense of Mathé-Pereverzev [4]. The following definitions will be needed to recall the concept of qualification as introduced in [2].

Definition 1 *Let $\rho, \tilde{\rho} \in \mathcal{O}$. We say that "$\rho$ and $\tilde{\rho}$ are equivalent at the origin" and we denote it with $\rho \approx \tilde{\rho}$, if they precede each other at the origin, that is, if there exist constants $\varepsilon > 0$, c_1, c_2, $0 < c_1 < c_2 < \infty$ such that $c_1\, \rho(\alpha) \le \tilde{\rho}(\alpha) \le c_2\, \rho(\alpha)$ for every $\alpha \in (0, \varepsilon)$.*

Clearly "$\approx$" is an equivalence relation and it introduces in $\mathcal{O}$ a partial ordering. Analogous definitions and notation will be used for $s, \tilde{s} \in \mathcal{S}$.

Definition 2 *Let $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ be a SRM, $r_\alpha(\lambda) \doteq 1 - \lambda g_\alpha(\lambda)$, $\rho \in \mathcal{O}$ and $s \in \mathcal{S}$. **i)** We say that (s, ρ) is a "weak source-order pair for $\{g_\alpha\}$" if it satisfies $\frac{s(\lambda)|r_\alpha(\lambda)|}{\rho(\alpha)} = O(1)$ for $\alpha \to 0^+,\ \forall\ \lambda > 0$. **ii)** We say that (s, ρ) is a "strong source-order pair for $\{g_\alpha\}$" if it is a weak source-order pair and there is no $\lambda_0 > 0$ for which $\frac{s(\lambda_0)|r_\alpha(\lambda_0)|}{\rho(\alpha)} = o(1)$ for $\alpha \to 0^+$. That is, if (s, ρ) is a weak source-order pair for $\{g_\alpha\}$ and $\limsup_{\alpha\to 0^+} \frac{s(\lambda)|r_\alpha(\lambda)|}{\rho(\alpha)} > 0, \forall\ \lambda > 0$. **iii)** We say that (ρ, s) is an "order-source pair for $\{g_\alpha\}$" if there exist a constant $\gamma > 0$ and a function $h : (0, \alpha_0) \to \mathbb{R}^+$ with $\lim_{\alpha\to 0^+} h(\alpha) = 0$, such that*

$$\frac{s(\lambda)\,|r_\alpha(\lambda)|}{\rho(\alpha)} \ge \gamma \quad \forall\ \lambda \in [h(\alpha), +\infty). \tag{1}$$

In the context of the previous definitions we refer to the function ρ as the "order of convergence" and to s as the "source function". We are now ready to define the concept of qualification in its three different levels as introduced in [2].

Definition 3 *Let $\{g_\alpha\}$ be a SRM. **i)** We say that ρ is "weak qualification of $\{g_\alpha\}$" if there exists a function s such that (s, ρ) is a weak source-order pair for $\{g_\alpha\}$. **ii)** We say that ρ is "strong qualification of $\{g_\alpha\}$" if there exists a function s such that (s, ρ) is a strong source-order pair for $\{g_\alpha\}$. **iii)** We say that ρ is "optimal qualification of $\{g_\alpha\}$" if there exists a function s such that (s, ρ) is a strong source-order pair for $\{g_\alpha\}$ and (ρ, s) is an order-source pair for $\{g_\alpha\}$.*

Now given the SRM $\{g_\alpha\}$ and $\rho \in \mathcal{O}$, we define $s_\rho(\lambda) \doteq \liminf_{\alpha\to 0^+} \frac{\rho(\alpha)}{|r_\alpha(\lambda)|}$ for $\lambda \ge 0$. Note that $s_\rho(0) = 0$ and if s_ρ is continuous, $s_\rho \in \mathcal{S}$. The next theorem shows the uniqueness of the source function.

Theorem 1 *([2]) If ρ is optimal qualification of $\{g_\alpha\}$ then there exists only one function s (in the sense of the equivalence classes induced by Definition 1) such that (s, ρ) is a strong source-order pair and (ρ, s) is an order-source pair for $\{g_\alpha\}$. Moreover if $s_\rho \in \mathcal{S}$, then s_ρ is such a unique function.*

In order to recall the concept of saturation, a few more definitions are needed.

Definition 4 *Let $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$ be a family of regularization operators for $Tx = y$. The "total error of $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$ at $x \in X$ for a noise level δ" is defined as $\mathcal{E}^{\text{tot}}_{\{R_\alpha\}}(x, \delta) \doteq \inf_{\alpha\in(0,\alpha_0)} \sup_{y^\delta \in \overline{B_\delta(Tx)}} \left\| R_\alpha y^\delta - x \right\|$, where $\overline{B_\delta(Tx)} \doteq \{y \in Y : \|Tx - y\| \le \delta\}$.*

Note that $\mathcal{E}^{\text{tot}}_{\{R_\alpha\}}$ is the error in the sense of the largest possible discrepancy that can be obtained for an observation of y within noise level δ, with an appropriate choice of the regularization parameter α. Next, given $M \subset X$, we shall denote with $\mathcal{F}_M$ the collection of the following functions: we will say that $\psi \in \mathcal{F}_M$ if there exists $a = a(\psi) > 0$ such that ψ is defined in $M \times (0, a)$, with values in $(0, \infty)$ and it satisfies the following conditions: (i) $\lim_{\delta \to 0^+} \psi(x, \delta) = 0$ for all $x \in M$, and (ii) ψ is continuous and increasing as a function of δ in $(0, a)$ for each fixed $x \in M$. Roughly speaking, the collection $\mathcal{F}_M$ contains all possible δ-"*orders of convergence*" on the set M.

Definition 5 *Let $\{R_\alpha\}_{\alpha \in (0,\alpha_0)}$ be a family of regularization operators for the problem $Tx = y$, $M \subset X$ and $\psi \in \mathcal{F}_M$. We say that ψ is an "upper bound of convergence for the total error of $\{R_\alpha\}_{\alpha \in (0,\alpha_0)}$ on M" if there exist a constant $r > 0$ and $p : M \to (0, \infty)$ such that $\psi(x, \delta) \leq p(x)\tilde{\psi}(x, \delta)$ for all $x \in M$ and for every $\delta \in (0, r)$, and we denote it with $\mathcal{E}^{\text{tot}}_{\{R_\alpha\}} \overset{M}{\preceq} \psi$.*

We will denote with $\mathcal{U}_M(\mathcal{E}^{\text{tot}}_{\{R_\alpha\}})$ the set of all functions $\psi \in \mathcal{F}_M$ that are upper bounds of convergence for the total error of $\{R_\alpha\}_{\alpha \in (0,\alpha_0)}$ on M. We are now ready to recall the concept of global saturation as introduced in [3].

Definition 6 *Let $M_S \subset X$ and $\psi_S \in \mathcal{U}_{M_S}(\mathcal{E}^{\text{tot}}_{\{R_\alpha\}})$. It is said that ψ_S is a "global saturation function of $\{R_\alpha\}$ over M_S" if ψ_S satisfies the following three conditions:*

S1. For every $x^* \in X$, $x^* \neq 0$, $x \in M_S$, $\limsup_{\delta \to 0^+} \frac{\mathcal{E}^{\text{tot}}_{\{R_\alpha\}}(x^*, \delta)}{\psi_S(x, \delta)} > 0$.

S2. ψ_S is invariant over M_S (see Definition 2.14 in [3]).

S3. There is no upper bound of convergence for the total error of $\{R_\alpha\}$ that is a proper extension of ψ_S (in the variable x) and satisfies S1 and S2, that is, there exist no $\tilde{M} \supsetneq M_S$ and $\tilde{\psi} \in \mathcal{U}_{\tilde{M}}(\mathcal{E}^{\text{tot}}_{\{R_\alpha\}})$ such that $\tilde{\psi}$ satisfies S1 and S2 with M_S replaced by $\tilde{M}$ and ψ_S replaced by $\tilde{\psi}$.

The function ψ_S and the set M_S are refer to as the saturation function and the saturation set, respectively. This conception of global saturation essentially establishes that no upper bound of convergence for the total error of the regularization method can be, at any non zero point $x^* \in X$, "strictly better" than the saturation function ψ_S at any point of the saturation set M_S.

3 Main Results

In this section a result on existence and uniqueness of global saturation for SRM's with optimal qualification as well as a converse result are presented. The proof of both of these results are quite involved and lengthly and therefore are not presented here. They will appear in a forthcoming paper.

The next theorem provides sufficient conditions on the family of functions $\{g_\alpha\}_{\alpha \in (0,\alpha_0)}$ defining an SRM and on the corresponding optimal qualification ρ, guaranteeing the existence of saturation.

Let $\{g_\alpha\}_{\alpha \in (0,\alpha_0)}$ be a SRM and consider the following hypothesis.

H4: There exists a constant $k > 0$ such that $G_\alpha \doteq \|g_\alpha(\cdot)\|_\infty \leq \frac{k}{\sqrt{\alpha}}$ $\forall\, \alpha \in (0, \alpha_0)$.

Theorem 2 *(Saturation for FSRO with optimal qualification.) Let $\{g_\alpha\}_{\alpha \in (0,\alpha_0)}$ be a SRM satisfying hypothesis H4 and having optimal qualification ρ. Let $r_\alpha(\lambda) \doteq 1 - \lambda g_\alpha(\lambda)$ and suppose that:*

***a)** The function ρ is of local upper type β, for some $\beta \geq 0$ and $s_\rho \in \mathcal{S}$.*

***b)** There exist positive constants $\gamma_1, \gamma_2, \lambda^*, c_1$, with $\lambda^* \leq \|T\|^2$ and $c_1 > 1$ such that*

***i)** $0 \leq r_\alpha(\lambda) \leq 1$, for $\alpha > 0$, $0 \leq \lambda \leq \lambda^*$;*

***ii)** $r_\alpha(\lambda) \geq \gamma_1$, for $0 \leq \lambda < h(\alpha) \leq \lambda^*$, $\alpha \in (0, \alpha_0)$ where h is as in (1). (Note that by virtue of Theorem 1 and the fact that $s_\rho \in \mathcal{S}$, there exists only one function $s \in \mathcal{S}$ satisfying (1), that is, $s = s_\rho$.)*

***iii)** $|r_\alpha(\lambda)|$ is strictly increasing with respect to α for each $\lambda \in (0, \|T\|^2]$;*

***iv)** $g_\alpha(c_1\alpha) \geq \frac{\gamma_2}{\alpha}$ for $0 < c_1\alpha \leq \lambda^*$ and*

***v)** $g_\alpha(\lambda) \geq g_\alpha(\tilde{\lambda})$, for $0 < \alpha \leq \lambda \leq \tilde{\lambda} \leq \lambda^*$.*

c) *There exist* $\{\lambda_n\}_{n=1}^{\infty} \subset \sigma(TT^*)$ *and* $c \geq 1$ *such that* $\lambda_n \downarrow 0$ *and* $\frac{\lambda_n}{\lambda_{n+1}} \leq c$ *for every* $n \in \mathbb{N}$.

Let $\Theta(t) \doteq \sqrt{t}\,\rho(t)$ *for* $t > 0$ *and* $X^{s_\rho} \doteq \mathcal{R}(s_\rho(T^*T)) \setminus \{0\}$. *Then* $\psi(x,\delta) \doteq \rho \circ \Theta^{-1}(\delta)$ *for* $x \in X^{s_\rho}$ *and* $\delta \in (0, \Theta(\alpha_0))$, *is saturation function of* $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$ *over* X^{s_ρ}.

Example 1 *Tikhonov-Phillips regularization method* $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$, *where* $g_\alpha \doteq \frac{1}{\lambda+\alpha}$ *has optimal qualification* $\rho(\alpha) = \alpha$ *and* $s_\rho(\lambda) = \lambda$. *It can be easily checked that the family* $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ *satisfies all hypotheses of Theorem 2. Therefore, the function* $\psi(x,\delta) \doteq \rho \circ \Theta^{-1}(\delta) = \delta^{\frac{2}{3}}$ *defined for* $x \in X^{s_\rho} \doteq \mathcal{R}(T^*T) \setminus \{0\}$ *and* $\delta > 0$ *is a global saturation function of* $\{R_\alpha\}_{\alpha\in(0,\alpha_0)}$ *over* X^{s_ρ}.

The following lemma constitutes a converse result in which regularity properties of the solution are obtained from the rate of convergence of the total error.

Lemma 1 *Under the same hypotheses of Theorem 2, if for some* $x \in X$ *we have that*

$$\sup_{y^\delta \in \overline{B_\delta(Tx)}} \inf_{\alpha\in(0,\alpha_0)} \left\| R_\alpha y^\delta - x \right\| = O(\rho(\Theta^{-1}(\delta))) \quad \textit{when } \delta \to 0^+, \tag{2}$$

then $x \in \mathcal{R}(s_\rho(T^*T))$.

4 Conclusions

In this work we dealt with the existence of saturation for spectral regularization methods with optimal qualification. We established sufficient conditions on the family of functions $\{g_\alpha\}_{\alpha\in(0,\alpha_0)}$ defining the method and on the optimal qualification ρ, guaranteeing the existence of saturation. Moreover, in those cases, appropriate characterizations of both the saturation function and the saturation set were given and an example was shown. Finally a converse result in which regularity properties of the exact solution are obtained from the rate of convergence of the total error was presented.

Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2010-2012 Nro. 0219, by Universidad Nacional del Litoral, U.N.L., through project CAI+D 2009-PI-62-315, by Agencia Nacional de Promoción Científica y Tecnológica, ANPCyT, through project PICT-2008-1301 and by the Air Force Office of Scientific Research, AFOSR, through Grant FA9550-10-1-0018.

References

[1] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of inverse problems*, volume 75 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[2] T. Herdman, R. D. Spies and K. G. Temperini, *Generalized Qualification and Qualification Levels for Spectral Regularization Methods*, J. Optim. Theory Appl., 141 (2009), pp. 547-567.

[3] T. Herdman, R. D. Spies and K. G. Temperini, *Global Saturation of Regularization Methods for Inverse Ill-Posed Problems*, J. Optim. Theory Appl., 148 (2011), pp. 164-196.

[4] P. Mathé and S. V. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789-803.

[5] A. Neubauer, *On converse and saturation results for regularization methods.* In BeitrÄage zur angewandten Analysis und Informatik, Shaker, Aachen, (1994), pp. 262-270.

ANÁLISIS BAYESIANO APLICADO A LA ESTIMACIÓN DEL TAMAÑO DE PARTÍCULAS MEDIANTE MEDICIONES DE DISPERSIÓN DE LUZ

Fernando A. Otero†‡, Gloria L. Frontini†‡ , Guillermo E. Eliçabe‡ y Helcio R. B. Orlande††

†Grupo de Matemática Aplicada, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, ‚Juan B. Justo 4205, 7600 Mar del Plata, Argentina, foterovega@fi.mdp.edu.ar, gfrontin@fi.mdp.edu.ar
‡División Polímeros, Instituto de Investigación en Ciencia y Tecnología de Materiales, Universidad Nacional de Mar del Plata y Consejo Nacional de Investigaciones Científicas y Técnicas Juan B. Justo 4205, 7600 Mar del Plata, Argentina, elicabe@fi.mdp.edu.ar
††Departamento de Ingeniería Mecánica, POLI/COPPE, Universidad Federal de Río de Janeiro, Río de Janeiro, Brasil, helcio@mecanica.ufrj.br

Resumen: Este trabajo presenta un enfoque bayesiano para la solución de un problema inverso de Dispersión de Luz Estática (DLE) empleando información previa obtenida de mediciones de Microscopía Electrónica de Barrido (MEB). Nuestro particular interés se centra en la caracterización de los sistemas de partículas en particular en las estimaciones de la Distribución de Tamaño de Partículas (DTP) que se obtuvieron en trabajos previos empleando un modelo aproximado denominado Aproximación Local Monodispersa (ALM). El objetivo del trabajo es obtener resultados más acordes a los obtenidos por MEB con su correspondiente intervalo de confianza el cual es provisto por el propio método bayesiano. La implementación desarrollada hace uso del algoritmo de Metropolis-Hastings.

Palabras claves: *Problema Inverso, Distribución de Tamaño de Partículas, Métodos Bayesianos*

1. INTRODUCCIÓN

Aplicamos un método bayesiano para resolver un problema inverso en dispersión de luz estática (DLE). Este problema inverso se propone recuperar la Distribución del Tamaño de Partículas (DTP) de los sistemas de partículas analizados. Si bien los métodos usados generalmente en este tipo de problemas corresponden a métodos de Mínimos Cuadrados y técnicas de regularización como Tikhonov ([2], [3] y [4]), estas técnicas pueden llevar a resultados poco satisfactorios. Por esta razón, un esquema bayesiano es desarrollado con el propósito principal de obtener resultados más confiables con un intervalo de confianza.

El enfoque bayesiano considera que todos los parámetros son variables aleatorias, de modo que éstas puedan describirse por su function densidad de probabilidad.

Un enfoque bayesiano se basa en el teorema de Bayes:

$$p_{posterior}(\mathrm{P}) = p(P/Y) = \frac{p_{prior}(P)p(Y/P)}{p(Y)} \qquad (1)$$

donde $p_{posterior}(\mathrm{P})$ es la llamada densidad de probabilidad a posteriori, $p_{prior}(\mathrm{P})$ es la densidad de probabilidad a priori (que se refiere a información previa a las mediciones), $p(Y/P)$ es la función de máxima verosimilitud y $p(Y)$ es la densidad de probabilidad de las mediciones. El enfoque bayesiano es en un sentido estadístico, el conjunto de parámetros P que maximiza $p_{posterior}(\mathrm{P})$. Esta solución es generalmente denominada como Máximo A Posteriori (MAP). Si bien el MAP puede calcularse resolviendo un problema de optimización, la estimación del correspondiente intervalo de confianza requiere la resolución del teorema de Bayes. Dado que para obtener $p(Y)$, es necesario integrar numéricamente, para reducer costo computacional del algoritmo se aplican las llamadas técnicas de Monte Carlo vía Cadenas de Markov (MCMC) como método alternativo. Dentro de este tipo de esquema se optó por el uso del algoritmo Metropolis-Hastings (MH) ([1] y [5]).

2. MODELO DE APROXIMACION LOCAL MONODISPERSA

El modelo de Aproximación Local Monodispersa (ALM) introducido por Pedersen en [6] supone una distribución espacial de las partículas de acuerdo con sus respectivos tamaños. Bajo esta suposición la intensidad de luz dispersada por un sistema de partículas puede expresarse como:

$$I_s(q) = K \int_0^{\infty} f(R)\, S(p,q,R)\, F^2(\mathrm{n}_p, q, R)\, dR \tag{2}$$

donde *F* representa la dispersión de luz debida a cada particular de radio *R*, con índice de refracción n_p para un ángulo de dispersión q definido a través de la magnitud del vector de scattering *q* de la ec. (3), donde n_m es el índice de refracción del solvente, l_0 es la longitud de onda del haz incidente y *p* es un parámetro efectivo del sistema que para concentraciones muy bajas es aproximadamente equivalente a la fracción de volumen de las partículas definida como el cociente entre el volumen de las partículas y el volumen total del sistema .

$$|\mathbf{q}| = \frac{4pn_m}{l_0} \sin\frac{q}{2}\,. \tag{3}$$

La correspondiente formula para F es:

$$F(n_p, q, R) = \frac{1}{q}\int_0^R r \sin qr\, dr = \\ = \frac{1}{q^3}\left[\sin(qR) - qR\cos(qR)\right] \tag{4}$$

donde un término adicional $[n_p - n_m]$ es incluído en la constante global *K* en la ec. (2).

El factor *S*(*p*,*q*,*R*) en la ec. (2) es el llamado factor de estructura y puede calcularse usando la aproximación de Percus-Yevick de acuerdo a:

$$S(p,q,R) = \frac{1}{1 - N_p(2p)^3 C(q)} \tag{5}$$

$$N_p(2p)^3 C(q) = \\ = 24p\left\{\frac{(a+b+d)}{u^2}\cos u - \frac{(a+2b+4d)}{u^3}\sin u - \right. \\ \left. - \frac{2(b+6d)}{u^4}\cos u + \frac{2b}{u^4} + \frac{24d}{u^5}\sin u + \frac{24d}{u^6}(\cos u - 1)\right\} \tag{6}$$

donde *Np* es el número de partículas por unidad de volumen, mientras que el parámetro de fracción efectiva de volumen *p* se define para modelar los efectos de interferencia. Finalmente a, b, d y m se relacionan con *p*, *q* y *R* siguiendo las ecs. (7), (8), (9) y (10).

$$a = \frac{(1+2p)^2}{(1-p)^4} \tag{7}$$

$$b = -6p\frac{(1+\frac{p}{2})^2}{(1-p)^4} \tag{8}$$

$$d = \frac{p(1+2p)^2}{2(1-p)^4} \tag{9}$$

$$u = 2qR \tag{10}$$

3. DESARROLLO

3.1. ANÁLISIS INVERSO

Consideramos el análisis inverso para el modelo con tres parámetros, donde el parámetro K de la ec. (2) es inferido de las estimaciones de MEB y la DTP es modelada mediante la función log-normal de parámetros g y R_0 siguiendo :

$$f(R) = \frac{(g/p)^{1/2}}{R} e^{-g[\log(R/R_0)]^2} \tag{11}$$

de donde el vector de parámetros estimados corresponde a P=[p, g ,R_0].
El algoritmo MH construye una cadena de parámetros [P_0, P_1,... P_m] que sigue la distribución de $p_{posterior}(P)$ de la ec. (1). Para ello se efectúa el cociente $p_{posterior}(P_1^*)/p_{posterior}(P_0)$ donde P_0 y P_1^* son respectivamente el vector de parámetros actuales de la cadena y el de un posible candidato y se le compara con el valor aleatorio de un parámetro a de distribución uniforme. De esta forma el nuevo vector P_1^* sólo es aceptado ($P_1 = P_1^*$) si el cociente es mayor que a. Es importante notar que al realizar el cociente se evita el cálculo de $p(Y)$. Además como las densidades de probabilidades son consideradas normales sólo es necesario calcular los exponentes de las ecs. (12) y (13):

$$p_{prior}(P) = 1/(\sqrt{2pi}|S_{Ap}|)^N e^{-\frac{1}{2}(P-P_{AP})^T (S_{Ap})^{-1}(P-P_{AP})} \tag{12}$$

$$p(Y/P) = 1/(\sqrt{2pi}|S_{med}|)^M e^{-\frac{1}{2}(I_{med} - Is(q,P))^T (S_{med})^{-1}(I_{med} - Is(q,P))} \tag{13}$$

donde P_{Ap} y S_{Ap} son la media y la matriz de covarianza de la distribución a priori, N=3 es la dimensión del vector de parámetros; I_{med} es el vector de M mediciones de DLE, $I_s(q,P)$ son las intensidades generadas por el modelo de la ec.(2) y S_{med} es la matriz de covarianza de las mediciones.

Los posibles candidatos son generados mediante un proceso de actualización de parámetros. Este se realiza de a un parámetro seleccionado aleatoriamente para cada paso de la iteración y los ajustes para el funcionamiento óptimo del algoritmo siguen el desarrollo de [7].

Las simulaciones de MEB fueron realizadas mediante el método de Monte Carlo. En este proceso, primero se generó una población total de partículas y se las ubicó en un espacio tridimensional arbitrario pero fijo, a partir del cual se obtiene una muestra de posición aleatoria sobre el espacio total, repitiendo el proceso de muestreo hasta obtener un número aceptable de muestras para poder analizar estadísticamente el ensamble. Los resultados finales son los parámetros estadísticos de media y varianza de la DTP empleados como información a priori en el método bayesiano.

3.2. EJEMPLOS SIMULADOS

Las mediciones han sido generadas usando el modelo ALM y sumando ruido aditivo normal de media nula y varianza proporcional al valor medio de las mediciones.

Se simularon ejemplos para tres niveles de ruido en las mediciones de DLE (0.1 ,1 y 2% del valor medio de medición de ruido adicionado) y tres tamaños de muestras (50, 100 y 200 partículas) en MEB.
Los valores de los parámetros verdaderos de la simulación son R_0=0.25 g=10, p=0.0216 y K=0.0221

4. DISCUSIÓN Y RESULTADOS

La información estadística permite al método bayesiano emplear simultáneamente información aportada por dos técnicas experimentalmente muy distintas. Así el análisis de los intervalos de confianza de los resultados permite comparar a cuantas micrografías de un tamaño de muestra arbitrario equivale la aplicación de DLE siguiendo el esquema bayesiano.

Se observó que la mejora en el intervalo de confianza tras aplicar MH es equivalente a realizar una micrografía de una muestra de más de 10000 partículas. Asimismo en la tabla 1 se observan resultados ligeramente mejores a los obtenidos aplicando Levenberg-Marquardt (LM). Se vio además que el efecto del ruido en las mediciones de DLE es determinante tanto en los valores estimados como en los intervalos de confianza obtenidos donde no afectó particularmente el tamaño de muestra de MEB.

	R_0	G	P
MEB (a priori)	0.2471 ± 0.0084	14.2520 ± 2.5377	-
LM DLE	0.2497	9.3982	0.0241
Bayesiano (media)	0.2493 ± 0.0004	9.6167 ± 0.2221	0.0247 ± 0.0019
Bayesiano (MAP)	0.2497	9.8385	0.0227
Verdadero	0.25	10	0.0216

Tabla 1. Estimaciones para mediciones de 1% de la media y simulaciones de MEB para una muestra de 50 partículas

5. CONCLUSIONES

Se desarrolló un método bayesiano para combinar el uso de datos obtenidos mediante MEB como información previa con el procesamiento de mediciones de DLE. El método bayesiano se implementó en Matlab® sobre una versión del algoritmo de Metropolis-Hastings. Los resultados obtenidos fueron satisfactorios. Se observó una notable mejora en los intervalos de confianza obtenidos respecto de aquellos hallados mediante las simulaciones de MEB. También se estimaron valores cercanos a los parámetros reales del orden e incluso mejores a los obtenidos mediante el método de Levenberg-Marquardt.

AGRADECIMIENTOS

Agradecemos el apoyo del Departamento de Matemática de la Facultad de Ingeniería de la Universidad Nacional de Mar del Plata, del CONICET y del CNPq.

REFERENCIAS

[1] S. CHIB, AND E. GREENBERG, *Understanding the Metropolis-Hastings Algorithm*, The American Statiscian, 49 (1995), pp. 327-335.

[2] G. FRONTINI AND E. FERNÁNDEZ BERDAGUER, *Inversion of Elastic Light Scattering measurements to determine Refractive Index and Particle Size Distribution of Polymeric Emulsions*, Inverse Problems in Eng., 11 (2003), pp. 329-340

[3] G. FRONTINI ET AL, *Estimation of Size Distribution in Concentrated Particle Systems from Light Scattering Measurements*, Inverse Problems in Eng., 16 (2008), pp. 995-1004

[4] O. GLATTER AND M. HOFER, *Interpretation of Elastic Light Scattering Data. III. Determination of Size Distributions of Polydisperse Systems*, J. Colloid and Inerface Science, 122 (1988), pp. 496-506

[5] N. METROPOLIS ET AL, *Equations of State Calculations by Fast Computing Machines, J. of Chem. Phys., 21 (1953), pp. 1087-1092*

[6] J.S. PEDERSEN, *Determination of Size Distributions from Small-Angle Scattering Data for Systems with Effective Hard-Sphere Interactions*, J. Appl. Cryst, 27 (1994), pp. 595-608

[7] *G. O. ROBERTS ET AL, Weak Convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Prob., 7 (1997), pp. 110-120

DIFUSION – CONSUMO DE OXIGENO EN TEJIDOS VIVOS. UNA FORMULACION GENERAL PARA DISTINTAS GEOMETRIAS.

Angélica Boucíguez[#], Liliana Lazo [##] y Luis T. Villa [###]

 # Facultad de Ciencias Exactas. Universidad Nacional de Salta. Instituto de Investigaciones en Energía No Convencional (INENCO) Av. Bolivia 5150 Salta, Argentina. acbouciguez@gmail.com
[##]*Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente Güiraldes 2160. Ciudad Universitaria. Ciudad Autónoma de Buenos Aires, Argentina,* lazoliliana@yahoo.com.ar
[###]*Facultad de Ingeniería, Universidad Nacional de Salta. Instituto de Investigaciones para la Industria Química (INIQUI) Av. Bolivia 5150 Salta, Argentina,* villal@unsa.edu.ar

Resumen: En este trabajo se presenta una formulación general para el estudio de problemas de difusión - consumo de oxígeno en tejidos vivos para geometrías plana, cilíndrica y esférica. A partir de la ecuación de difusión que determina la concentración de dicho gas en función de la posición y del tiempo, se plantea la resolución general del problema estacionario que permite obtener la distribución inicial de oxígeno en el tejido y su penetración en el mismo. Se realiza un modelo numérico para la resolución del problema en cada una de estas geometrías.

Palabras claves: *difusión – consumo de oxígeno, problema inverso, concentración de oxigeno.*
2000 AMS Subjects Classification: 34A55 – 49N45

1. INTRODUCCIÓN

Se sabe que la presencia de oxígeno en los tejidos tumorales contribuye a una mejor absorción de la radiación, lo que hace que el tratamiento sea más efectivo. A tal efecto, se introduce oxígeno en el interior de ellos, exponiéndolo a altas concentraciones y permitiendo que se absorba hasta que alcance el estado estacionario; a continuación se sella la superficie exterior y comienza el tratamiento con radiación. Paralelamente, el oxígeno presente en el interior del tejido es absorbido y consumido, lo que altera el perfil (estacionario) inicial y gradualmente, el punto de concentración nula, se desplaza hacia la superficie hasta que el oxígeno desaparece por completo del tejido.

El planteo matemático que describe este proceso constituye un problema en que el dominio de trabajo varía en el tiempo y se encuadra en la tipificación de problema inverso de Stefan. El mismo ha sido abordado por diversos autores [1], [2], [4] y [5]; quienes lo plantean en coordenadas cartesianas, utilizando preferentemente modelos adimensionales.

El modelo matemático que lo representa responde a un modelo difusivo en una ecuación no homogénea, debida precisamente a la razón de consumo de oxígeno en el tejido, que en este caso se considera constante y puede resolverse analítica y numéricamente [6].

2. FORMULACIÓN MATEMATICA GENERAL

El problema de difusión - consumo de oxígeno resulta completamente expresado por las ecuaciones (1) a (6):

$$\frac{\partial c(r,t)}{\partial t} = D\left(\frac{\partial^2 c(r,t)}{\partial r^2} + \frac{a}{r}\cdot\frac{\partial c(r,t)}{\partial r}\right) - m \qquad 0 < r < s(t) \qquad t > 0 \tag{1}$$

$$c(r,0) = C_0(r) \qquad 0 \le r \le r_0 \tag{2}$$

$$c(0,t) = 0 \qquad t > 0 \tag{3}$$

$$\frac{\partial c}{\partial r}(0,t)=0 \qquad t>0 \tag{4}$$

$$c(s(t),t)=\frac{\partial c}{\partial r}(s(t),t)=0 \qquad t>0 \tag{5}$$

$$s(0)=r_0 \qquad t>0 \tag{6}$$

Donde el valor de *a*, determina la geometría (*a*=0 corresponde a la plana, *a*=1 a la cilíndrica y *a*=2 a la esférica); *c(r,t)* es la concentración de oxígeno en el tejido, función de la posición *r* y el tiempo *t*; D su difusividad y *m* la razón de su consumo en el mismo; *s(t)* la posición límite, variable en el tiempo, hasta la cual el oxígeno está presente en el tejido. Las constantes C_0 y r_0 denotan, respectivamente, la concentración inicial de oxígeno en el tejido y la profundidad de penetración del gas en el mismo.

Primeramente, es necesario determinar C_0 y r_0, para ello debe resolverse el problema estacionario, el que se resume en las ecuaciones (7) a (9)/(9'):

$$D\left(\frac{d^2C_0(r)}{dr^2}+\frac{a}{r}\cdot\frac{dC_0(r)}{dr}\right)-m=0 \tag{7}$$

$$C_0(r_0)=C_0 \tag{8}$$

$$C_0(0)=0 \tag{9}$$

$$\frac{dC_0}{dr}(0)=0 \tag{9'}$$

Al ser un problema a derivadas segundas, existen dos constantes a determinar, por lo que bastan dos condiciones, expresadas en (8) y (9) o (9'), para evaluarlas; la tercera condición, es necesaria para calcular la penetración inicial de oxígeno en el tejido. Se considera que r=0 representa el interior del tejido y r=r_o la superficie del mismo; esto facilita la determinación de las constantes provenientes de la resolución de la ecuación (7). Con las ecuaciones (8) y (9) se determinan las dos constantes de la ecuación (7), y a continuación con la ecuación (9') la penetración inicial de oxígeno en el tejido, esto es, el valor de r_0.

Resolviendo el problema para cada una de las tres tipos de coordenadas se obtienen las expresiones dadas por las ecuaciones (10) a (12). En ellas se observa que la mayor penetración de oxígeno se presenta en geometría esférica y la menor en la plana.

Coordenadas cartesianas

$$C(r)=\frac{m}{2D}(r-r_0)^2, \qquad r_0=\left(\frac{2DC_0}{m}\right)^{1/2} \tag{10}$$

Coordenadas cilíndricas

$$C(r)=C_0+\frac{m}{4D}(r^2-r_0^2), \qquad r_0=\left(\frac{4DC_0}{m}\right)^{1/2} \tag{11}$$

Coordenadas esféricas

$$C(r)=C_0+\frac{m}{6D}(r^2-r_0^2), \qquad r_0=\left(\frac{6DC_0}{m}\right)^{1/2} \tag{12}$$

3. RESOLUCIÓN NUMERICA Y RESULTADOS OBTENIDOS

Una vez obtenido el perfil inicial de oxígeno en el tejido, se resuelve el problema numéricamente para determinar la concentración del mismo como función de la posición y del tiempo. El valor de C_0, se ha fijado en 100 a fin de observar, rápidamente, cuánto ha disminuido porcentualmente la concentración de oxígeno en el tejido. Los valores de difusividad y razón de consumo de oxígeno se han tomado iguales a D=4 * 10^{-8} m^2/s, m=0,0008 %/s, respectivamente, en concordancia con estudios realizados con anterioridad [3]. Para la resolución numérica del problema se ha empleado el método de diferencias finitas explicitas, mediante un programa realizado en lenguaje Matlab.

En las Figuras 1, 2 y 3 se presenta la concentración de oxígeno en el tejido en función de la posición, para distintos tiempos; en coordenadas cartesianas, cilíndricas y esféricas, respectivamente.

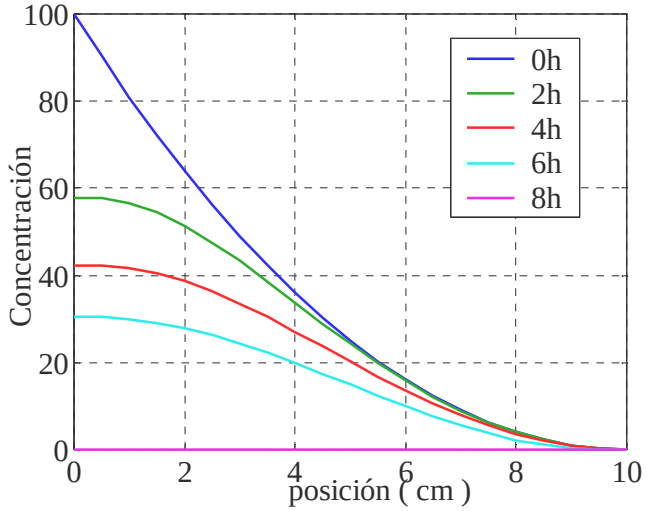


Figura 1: Concentración de oxígeno en el tejido en coordenadas cartesianas.

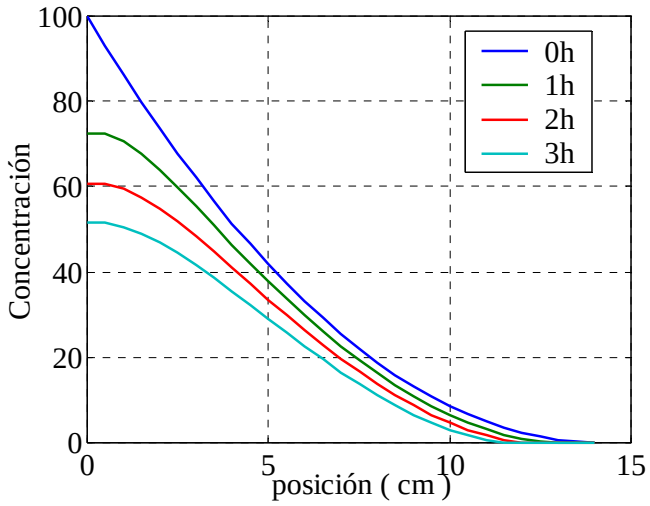


Figura 2: Concentración de oxígeno en el tejido en coordenadas cilíndricas.

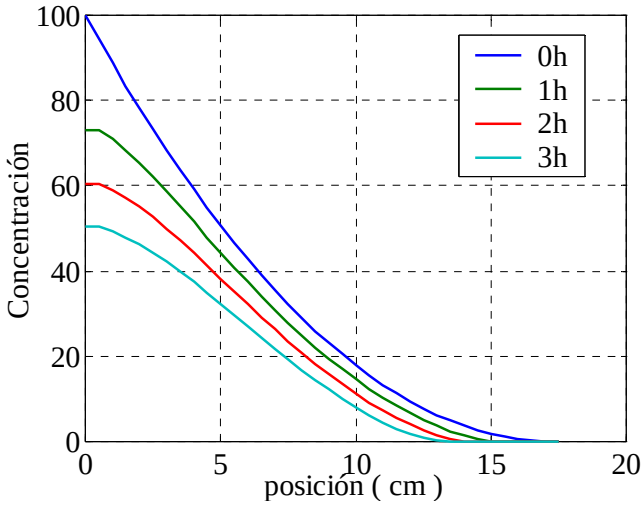


Figura 3: Concentración de oxígeno en el tejido en coordenadas esféricas.

En las Figuras se observa que la penetración inicial de oxígeno en el tejido es más alta en la geometría esférica, siguiéndole en orden decreciente, la cilíndrica y a continuación la plana; con valores de 17,32; 14,14 y 10 cm, respectivamente. Asimismo la concentración de oxígeno en la superficie del tejido disminuye siguiendo el mismo orden de acuerdo a la geometría, sin embargo la permanencia del gas en el tejido se da en orden inverso.

4. CONCLUSIONES

La resolución del problema en forma genérica para las tres geometrías principales, permite abordar el problema en forma general, llegando a una solución analítica básica la que luego puede desglosarse para estudiar cada una de tales geometrías.

La resolución numérica del problema, aún en su formulación más sencilla (todos los parámetros constantes) permite analizar la difusividad - consumo de oxígeno en las tres geometrías básicas y realizar una comparación entre ellas; lo que posibilita establecer las diferencias entre cada una de ellas.

AGRADECIMIENTOS

Los autores agradecen al Consejo de Investigaciones de la UNSa, INIQUI e INENCO.

REFERENCIAS

[1] S. AHAMED, *A Numerical Method for Oxygen Diffusion and Absorption in a Spike Cell.* Applied Mathematics and Computation 173, (2006), pp. 668-682.

[2] M. ASCHIERI Y C. TURNER, *El problema de difusión-consumo de oxígeno en tejidos vivientes*. Mecánica Computacional, XX, (2001), pp. 577-584.

[3] A. BOUCIGUEZ, L. LAZO Y L. VILLA. *Evaluación del Contenido de Oxígeno en Tejidos Tumorales*. Mecánica Computacional, XXVII, (2008), pp. 2705-2714.

[4] S. ÇATAL, *Numerical Approximation for the Oxygen Diffusion Problem*. Applied Mathematics and Computation, 145, 2-3, (2003), pp. 361-369.

[5] A. FRIEDMAN Y F. REITICH, *Analysis of Mathematical Model for the Growth of Tumors.* Journal of Mathematical. Biology. 38, (1999), pp. 262-289.

[6] M. ZERROUKAR Y C. CHATWIN, C. *Computational Moving Boundary Problems.* John Wiley & Sons Inc., 1994.

MODELO MATEMÁTICO PARA DETERMINAR LA HUMEDAD EN LA MADERA USANDO MICROONDAS

Jhon E. Hinestroza R.†, Hernán Estrada B.‡

†Grupo de Investigaciones en Matemática, Universidad Tecnológica, Colombia, Chocó, Quibdó, joe0984@gmail.com
‡ Universidad Nacional de Colombia, Colombia, Bogotá D., hestradab@gmail.com, *www.unal.edu.co*

RESUMEN: En este trabajo se desarrolla la modelación matemática del procedimiento para medir la humedad en la madera usando microondas. Estudiamos con ayuda de la electrodinámica la interacción entre una onda electromagnética (microondas) con la madera, la cual se comporta, por su contenido de agua, como un material dieléctrico. El modelo matemático para la medición de la humedad se desarrolla con ayuda de las ecuaciones de Maxwell en medios materiales. Lo interesante del procedimiento matemático desarrollado está en que el problema no está bien condicionado y es necesario realizar variación del espectro de la señal de microondas para determinar de manera unívoca la humedad del trozo de madera.

Palabras claves: *Ecuaciones de Maxwell, humedad en madera, Ondas microondas, propiedades dieléctricas.*

1. INTRODUCCIÓN

Técnicamente, existe una variedad de métodos para determinar el contenido de humedad en la madera [8]. Entre ellos se ha dado importancia al "*método de las microondas"*. Dicho procedimiento se reconoce por ser no invasivo, y consiste en el uso de un emisor de señal electromagnética (Figura 1) en el rango de las microondas a una frecuencia que permita superar la *skin depth* de la madera, un trozo de madera y un receptor que capta la señal que logra atravesarlo. Mediante las propiedades de la señal recibida como son la amplitud y desfase con la onda incidente se puede determinar el porcentaje de humedad mediante un proceso de inversión.

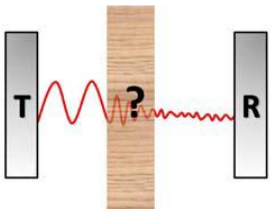


Figura 1: *Prototipo* usado para medir la humedad en la madera, **T** transmite la señal que interactúa con el bloque de madera y **R** recibe la fracción de onda que logra pasar a través del bloque de madera.

El modelo desarrollado tiene importantes aplicaciones en la industria maderera por lo que en la práctica se han diseñado sensores específicos, además, los principios empleados aquí se pueden usar en la medición de humedad en granos como el arroz entre otros.

2. MARCO TEÓRICO

Una característica importante de la madera son sus propiedades dieléctricas originadas por el contenido de humedad que pueden ser descritas con la constante dieléctrica ε' y el factor de pérdida $\varepsilon'' = \varepsilon' tan\delta$, con δ el ángulo de pérdida, ε' y ε'' se escriben como una cantidad compleja única

$$\varepsilon = \varepsilon' + i\varepsilon'' \qquad (1)$$

denominada permitividad dieléctrica [2,3,4].

Debido a que la madera es un material no conductor, las propiedades dieléctricas de la madera húmeda describen la interacción entre esta y el campo eléctrico aplicado [1]. Como consecuencia de las propiedades anisotrópicas de la madera, se distinguen tres direcciones diferentes de fibras al considerar sus propiedades dieléctricas longitudinal L, la radial R y tangencial T [2, 3, 5].

Considerando la propiedad anisotrópica [3] de la madera, sus propiedades físicas son distintas dependiendo de la dirección en que se midan; pero, el valor en la dirección radial y tangencial no difiere mucho, luego pueden ser consideradas iguales, con lo que se escribe la permitividad como $\varepsilon_{\parallel}$ paralela y $\varepsilon_{\perp}$ perpendicular. La permitividad dieléctrica y la humedad se relacionan como muestra la figura 2 [3], y son el eje principal para resolver el problema ya que hacen posible pensar que conocidos los datos referentes a la constante dieléctrica y factor de pérdida, para un determinado tipo de madera es factible, mediante un proceso de inversión determinar su porcentaje de humedad.

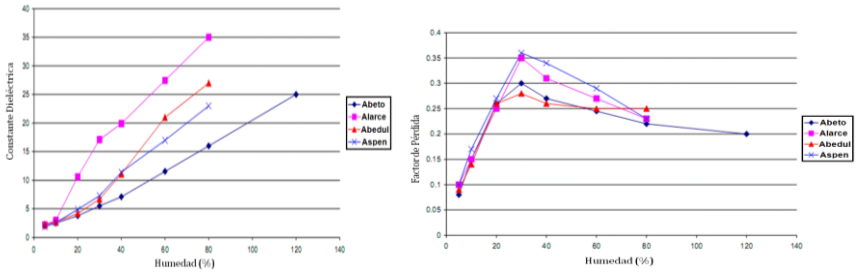


Figura 2: Constante dieléctrica (izquierda) y Factor de pérdida (derecha) como función de la humedad.

2.1. LA MATEMÁTICA

Las ecuaciones de Maxwell describen los fenómenos electromagnéticos y la interacción con medios materiales [1, 7]:

$$\nabla \times H = \frac{4\pi}{\mu_0 \varepsilon_0} A + \mu_0 \varepsilon_0 \frac{\partial D}{\partial t} \qquad (2)$$

$$\nabla \times E = -\mu_0 \varepsilon_0 \frac{\partial B}{\partial t} \qquad (3)$$

$$\nabla \bullet D = \frac{1}{\varepsilon_0} \rho \qquad (4)$$

$$\nabla \bullet B = 0 \qquad (5)$$

Aquí, **E**[volt/m]: es la intensidad del campo eléctrico, **H**: intensidad del campo magnético; **B**[Tesla]: inducción magnética; **A**[Amp/m^2]: densidad de flujo eléctrico o de corriente; **D**: desplazamiento eléctrico; **ρ**[C/m^3]: fuente de carga; ε_0: permitividad eléctrica; μ_0: permeabilidad magnética. Las relaciones constitutivas son:

$$B = \mu H \qquad (6) \qquad D = \varepsilon E \qquad (7) \qquad A = \eta E \qquad (8)$$

Las constantes ε permitividad eléctrica, μ permeabilidad magnética, η conductividad pueden ser reales o complejos, escalares o matrices, constantes o variables. En el caso vacío (aire) ε, μ, η son escalares, pero las propiedades eléctricas en el material se pueden escribir usando matrices diagonales con las correspondientes coordenadas *x, y, z*. A partir de las ecuaciones de Maxwell se obtiene una ecuación de onda para el campo eléctrico dada por:

$$\nabla(\nabla \bullet E) + \mu\eta \frac{\partial}{\partial t} E + \varepsilon\mu \frac{\partial^2 E}{\partial t^2} = \nabla^2 E \qquad (9)$$

Cuya solución puede ser determinada considerando una onda plana de campo eléctrico en la dirección "y":

$$E = \hat{E} e^{-i(\omega t - ky)} \qquad (10)$$

en donde ω es la frecuencia, *k* el número de onda, $\hat{E}$ describe la forma de la onda. Usando las ecuaciones (9-10) conociendo la forma de ε y considerando el campo en direcciones *x*, *y*, *z*, (η=0 en el aire), el número de onda en ambos medios está determinado por:

$$k_A = \frac{\omega}{c_0}, \quad c_0 = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}, \; c_0 \text{ es la velocidad de la luz}$$

$$k_w = \frac{\omega}{c_0}\sqrt{\frac{1}{2}\left(\varepsilon_\perp^{|} + \sqrt{\left(\varepsilon_\perp^{|}\right)^2 + \left(\varepsilon_\perp^{\|}\right)^2}\right)} + \frac{\omega}{c_0}\sqrt{\frac{1}{2}\left(-\varepsilon_\perp^{|} + \sqrt{\left(\varepsilon_\perp^{|}\right)^2 + \left(\varepsilon_\perp^{\|}\right)^2}\right)}i$$

por lo tanto la expresión para el campo eléctrico en la dirección *x* se escribe

$$E_x = \hat{E}e^{-i(\omega t-(\alpha+i\beta)y)} \qquad (11)$$

Considerando las múltiples reflexiones de la señal en las fronteras, como se indica en la figura 3, podemos determinar el coeficiente de transmisión de la señal electromagnética

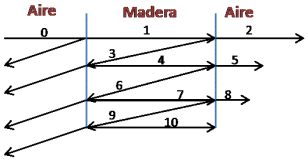


Figura 3: Superposición de ondas en las interfaces onda-madera, madera-aire. Muestra como la onda se refleja en las interfaces, pasa una parte y esta sigue reflejándose de forma infinita.

$$T = t_{Aw}t_{wA}\frac{e^{i(k_w-k_A)d}}{1-r_{wA}^2e^{2ik_wd}} \qquad (12)$$

Que puede ser escrito como:

$$T = Ae^{i\phi} \qquad (13)$$

en donde A corresponde la amplitud y *Φ* la fase del coeficiente, cantidades que son medibles por un receptor apropiado. El resultado dado por (12), describe el coeficiente de transmisión total de la onda inicial que pasa a través del bloque de madera y muestra que para un valor de ω bien determinado y un tamaño de bloque específico, el valor de *A* y *Φ* solo dependen de las propiedades dieléctricas.

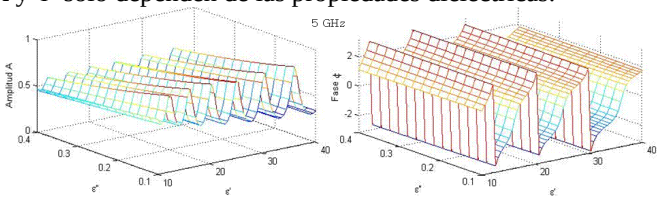


Figura 4: Amplitud y Fase contra la ε', ε'' respectivamente a 5GHZ.

Las figura 4 se obtiene a partir de (12) usando funciones de MATLAB, considerando un ancho fijo del bloque madera y frecuencia y muestra la relación entre ε', ε'' con *A* y Φ. Nótese en la figura 4, que si se escoge un par adecuado de $(\varepsilon', \varepsilon'')$ para *A* y Φ es posible regresar a la figura 2 y determinar el valor de humedad de la madera. Pero, puede verse también, que es posible tener para valores distintos de ε', ε'' un mismo valor de *A* y *Φ*, y se obtendrían distintos valores de humedad para un mismo tipo de madera situación que no se presenta si se toma ω=0.5GHz.

2.2. EL PROBLEMA INVERSO Y EL PROCEDIMIENTO NUMÉRICO

Mediante el modelo anteriormente desarrollado, y con ayuda de las curvas que determinan relación entre la humedad, la constante dieléctrica y la atenuación, es posible determinar a partir de los valores medidos A y *Φ* del coeficiente de transmisión la humedad del trozo estudiado, pero dada la forma de figura 4 su valor no está unívocamente determinado. Esto representa un delicado problema que puede ser resuelto variando la frecuencia de la señal electromagnética hasta obtener un valor de la humedad unívocamente definido.

2.2.1. INVERSIÓN NUMÉRICA DEL MAPEO $\lambda \rightarrow (A,\Phi)$.

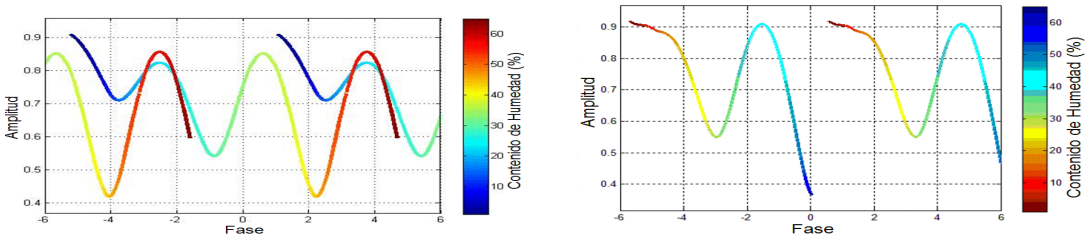


Figura 5: Amplitud vs Fase. La parte izquierda se hizo con una frecuencia de 2GHz y la derecha con una frecuencia de 1.2GHz.

3. CONCLUSIONES

- El método es desarrollado es efectivo y puede verse aplicado en instrumentos desarrollados en [10] y donde se utilizan los principios teóricos mostrados aquí para diseñar un sensor para medir el contenido de humedad en la madera.
- El problema de determinar el porcentaje de humedad mediante ondas microondas es inverso y mal planteado debido a la falta de unicidad en la solución, para cualquier frecuencia de onda visto esto en el hecho de que para algunos pares (Φ, A) la humedad no se puede determinar de forma única (figura 5) sin embargo, esta dificultad puede ser superada al realizar ajustes sobre la frecuencia de la onda emitida. Se halló que para frecuencias menores a 1.35GHz; 1.5GHz; 1.15GHz y 1.85GHz que corresponden el abeto, alerce, abedul, aspen respectivamente, el método numérico funciona bien.
- Para desarrollar los cálculos y el estudio de humedad en la madera se requiere conocer por lo menos las propiedades dieléctricas de la especie de madera. Además, lo desarrollado es muy interesante, ya que ayudaría a la resolución más eficaz de problemas como el estudiado en [4] en el que miden las propiedades eléctricas del arroz, cereal y otros granos Coreanos, y el contenido de humedad mediante un aparato diseñado para tal efecto [4].

REFERENCIAS

[1] M. FREDERICK J. ET. AL, *fundamentos de la teoría electromagnética,* Fondo educativo interamericano 3ª edición. 1984.

[2] W. JAMES. *dielectric properties of Wood and hardboard: variation with temperature, frecuency, moisture content, and grain orientation.* Forest products laboratory. U.S department of Agriculture 1975.

[3] KEAM HOLDEM ASSOCIATES LTD., *summary of the dielectric permittivity of wood.* Aplication note: KHA0420. Auckland, New Zealand July, 1999.

[4] K. KIM, JONG-HEON KIM, L. SEUNG AND S. HA NOH, *Measurement of grain moisture content using microwave at 10.5 GHz and moisture density,* IEEE Transactions on instrumentation and measurement, Vol. 51, Nº 1. 2002.

[5] N. MISATO, Y. TADASHI, *the dielectric properties of wood VI. On the dielectric properties of the chemical constituents of wood and the dielectric anisotropy of wood*. Wood research N052. 1972.

[6] E. PURCELL M, *Electricidad y magnetismo Berkeley phisics course,* Vol 2, Reverté, segunda edición, 2005.

[7] A. SICARD GERMÁN, *electricidad y magnetismo*. Universidad Nacional Colombia, Unibiblos 2008.

[8] M. SILVIA, A. VÍCTOR, V. NORMA. *Calibración de higrómetros para madera, métodos y trazabilidad.* La Guía Metas. Metas & Antropólogos y asociados. Año 4 #12. Simposio de metrología. México 2004.

[9] S. THERESE, B. NILSON Y N. SVEN, *microwave modeling and signal estimation for detection of wood properties. Proceedings of RVK,* Linköping, June 14-16, 2005.

[10] M WILLIAM, H GREGORY R, *microwave wood moisture sensors for dry kilns.* Tennesse forest products center University of Tenessee. 2005.

WAVELET PROJECTION METHODS FOR SOLVING INVERSE PROBLEMS: PSEUDODIFFERENTIAL OPERATOR CASE

María Inés Troparevsky[†] and Eduardo P. Serrano[♭]

[♭] *Centro de Matemática Aplicada, Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Campus Miguelete, CP1650, San Martín, Prov. Bs As, Argentina, Eduardo.Serrano@unsam.edu.ar*
[†] *Departamento de Matemática, Facultad de Ingeniería, Universidad de Buenos Aires, Av. Paseo Colón 850, C1063ACV, Buenos Aires, Argentina, mitropa@fi.uba.ar*

Abstract: Operator equations of the form $Af = g$, where $A : X \to Y$ and X, Y are functional spaces, often appear in mathematical models arising in engeneering and other sciences. In general they cannot be solved exactly and numerical methods must be developed to find approximated solutions $\tilde{f}$. In addition only noisy data $\tilde{g}$ is available. Wavelet Galerkin Projection Methods are sometimes chosen to approximate the solution since the resulting matrix is sparse and computations on increasing $X_N \subset X$ can be done recursively. In [3], the authors develop a general projection method with regularization for a *wavelet-vaguelet* decomposition. In this work we use that scheme to approximate f when A is a pseudodifferential operator of convolution type between the Hilbert space X. We show the errors of the approximations and state some conclusions.

Keywords: *Wavelet-Galerkin, inverse problem, projection methods*
2000 AMS Subject Classification: 65T60, 45Q05, 47AXX, 47A52

1 INTRODUCTION

Operator equations of the form

$$Af = g \tag{1}$$

where $A : X \to Y$, and X, Y are functional spaces, often appear in mathematical models arising in engeneering and other sciences. In general (1) cannot be solved exactly and numerical methods must be developed in order to find approximated solutions $\tilde{f}$. In addition the function g is usually perturbed with noise and only noisy data $\tilde{g}$ that satisfies $\|g - \tilde{g}\| < \varepsilon$ is available. The Galerkin Projection Method searchs for these solutions in a finite dimensional subspace of X, $X_N = span\{\psi_j, \quad j = 1, \cdots N\}$. In this context $\tilde{f}$ is $f_N = \sum_i^N a_j \psi_j$ and the problem of finding f_N becomes the problem of solving an algebraic matrix equation

$$\sum_{j=1}^{N} \langle A\psi_j, v_i \rangle = \langle g, v_i \rangle \qquad i = 1, \cdots N \tag{2}$$

where v_j are the test functions and $Y_N = span\{v_j, \quad j = 1, \cdots N\} \subset Y$.
Considering computational aspects, it is desirable that the resulting matrix is sparse or that it has a small condition number. Orthonormal wavelet basis are sometimes chosen to achieve this goal. Projection methods based on orthogonal wavelet decomposition have also the advantage that computations on increasing X_N can be done recursively (see [3]).
Wavelet Galerkin methods were succesfully applied to solve partial differential equations [6] and to carry on numerical differentiation [5]. An adaptive wavelet Galerkin method for linear inverse problems were developed in [1], where a smaller space adapted to the solution is iteratively constructed. In [2] the analysis of a wavelet based adaptive algorithm for solving elliptic equations is presented. Tikhonov regularization of inverse parabolic problems via adaptive wavelet Galerkin methods is discussed in [4]. In [3], under certain hypothesis, the authors calculate the optimal value N for a general projection method with regularization for a *wavelet-vaguelet* decomposition. In this work we use that scheme to approximate the solution of the problem described below.

We consider inverse problems of the form (1) where $A : X \to X$ is a pseudodifferential operator of convolution type between the Hilbert space X,

$$Af = k * f = g, \qquad \hat{k}(\omega) = \frac{1}{(1+\omega^2)^{\alpha}} \tag{3}$$

α a positive real number , f is the unknow and g is the given, eventually noisy, data. The inverse problem is ill-posed since small perturbations in the data produce large errors in the solution, i.e. $\left\{\overline{f} : \|Af - A\overline{f}\| < \varepsilon\right\}$ is unbounded.
Under these hypothesis, a Galerkin method for solving (3) is defined by an increasing sequence of subspaces $X_m \subset X_{m+1}$ where the solution to (3) is projected. The projection of the solution f into X_m, $f_m \in X_m$, satisfies

$$\langle Af_m, v\rangle = \langle g, v\rangle, \quad \forall v \in X_m.$$

We hope that the resulting sequence f_m approximates the solution f: $\|f - f_m\| \to 0$ when $m \to +\infty$.
We calculate the approximated solutions in two different ways: (1) by deconvolution, working in the frequency domain and (2) using a wavelet-vaguelet decomposition. In the last case, we estimate the error $\|f - f_m\|$.
This paper is organized as follows. In Section 2 we present a summary of the wavelet vaguelet decomposition, build the subspaces V_j and recall some properties. In Section 3 we solve an example and state some conclusions.

2 Wavelet-Vaguelet Decomposition

In this section we present a brief summary of definitions and properties related to wavelets and multiresolution analysis that will be used in the following sections. More details can be found in [7], [8] and [9].
A multi-resolution analysis is a sequence of closed subspaces of $L^2(R)$, $\{V_j, j \in Z\}$ that satisfies:

1. $V_j \subset V_{j+1}$
2. $f(x) \in V_j \iff f(2x) \in V_{j+1}, \quad f(x) \in V_0 \implies f(x-k) \in V_0$
3. $\bigcap_j V_j = \{0\}$
4. $\overline{\bigcup_j V_j} = L^2(R)$
5. $\exists \phi \in V_0$ such that $\{\phi(x-k), k \in Z\}$ is an orthonormal basis for V_0

Associated to V_j we denote W_j as the orthogonal complement of V_j in V_{j+1}, $W_j \oplus V_j = V_{j+1}$.
It results that there is an unique function $\psi \in W_0$ such that the family $\left\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \quad k \in Z\right\}$ is an orthonormal basis of W_j and $L^2(R) = \oplus_{j\in Z} W_j$.
The following decomposition for $f \in L^2(R)$ follows:

$$f = \sum_{j\geq 0}\sum_{k\in Z} d_{jk}\psi_{jk} + \sum_{k\in Z} b_{0k}\phi_{0k}$$

where $d_{jk} = \langle f, \psi_{jk}\rangle$ and $b_{0k} = \langle f, \phi_{0k}\rangle$ are the wavelet and scaling coefficients. The wavelet truncated expansion of order $M > 1$ of f is

$$f_M = \sum_{j\geq 0}^{M-1}\sum_{k\in Z} d_{jk}\psi_{jk} + \sum_{k\in Z} b_{0k}\phi_{0k}$$

Since $X = Y = L^2(R)$, we can choose the subspaces $X_j = W_j \subset X$. The (A, ψ)-vaguelets are the elementary test functions defined by

$$A^* v_{jk} = \lambda_{jk}\psi_{jk}$$

where λ_{jk} are normalization factors.
The vaguelet subspaces are $Y_j = span\{v_{jk},\, j,k \in Z\}$. A summary of the properties of the wavelet-vaguelet decomposition can be found in [3].
In our case, the form of the operator A enables us to evaluate v_{mk} easily. Operating in the frequency domain, we have

$$\hat{v}_{mk}(\omega) = \frac{1}{\hat{k}(\omega)}\lambda_{mk}\hat{\psi}_{mk}(\omega)$$

The coefficients of f are related to g and v_{jk} by:

$$f = \sum_{j\in Z}\sum_{k\in Z} d_{jk}\psi_{jk} = \sum_{j\in Z}\sum_{k\in Z} \langle g, v_{jk}\rangle \, (\lambda_{jk})^{-1}\psi_{jk}$$

We remark that since $\lambda_{jk} \to 0$ the coeffients d_{jk} may be unbounded.

The resulting projection method regularize the inverse problem and the truncation error $\left\| f - \tilde{f}| \right\|$ is bounded. This truncated wavelet-vaguelet decomposition can be combined with adaptive strategies in order to add local refinements (see [3]).

3 EXAMPLE

We consider the convolution operator defined in (3) for $\alpha > 0$, with $g(x) = xe^{-\beta x}(1 + .1\sin(\omega_0 x))$ $\beta > 0$, and some $\omega_0 > 0$. We choose Meyer-type wavelets with Fourier transform of bounded support (see[8]). We assume that the sampling frequency of the data is much greater than ω_0.

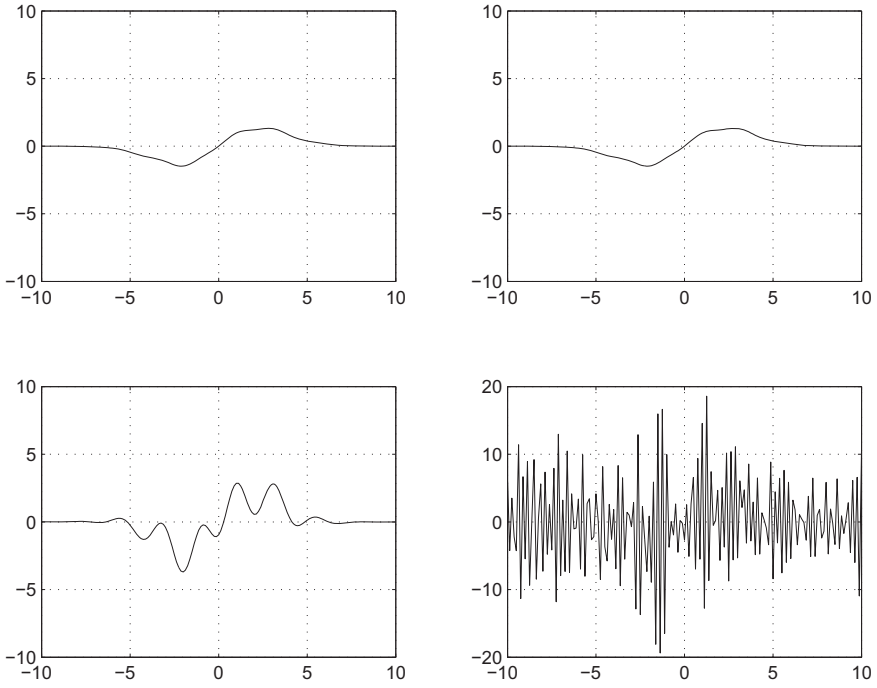


Figure 1: Functions $g(x), \tilde{g}(x)$ (Top left, right), $f(x), \tilde{f}(x)$ (bottom left, right)

In Figure 1 (top) we show the original data g and the perturbed data $\tilde{g}$ considering additive noise with deviation $\sigma \cong 0.005$. The parameters are $\alpha = 1.25$, $\beta = 0.1$, $\omega_0 = 3\pi/4$. The sampling frequency is $8Hz$. The solution f and the approximation $\tilde{f}$ appear in the bottom. The approximated solution $\tilde{f}$ is obtained, working in the frequency domain, by deconvolution. Afterwards, applying the inverse Fourier transform, the function is recovered. We can observe the multiplicative effect of the noise, by the inverse operator $\hat{k}(\omega)$ in high frequencies.

On the other hand, we apply the wavelet-vaguelet projection method described in Section 2.
In Figure 2 several syntesis $\tilde{f}$, corresponding to the level sets $J = [-8, -3]$ $J = [-8, -2]$, $J = [-8, 0]$ and $J = [-8, 1]$ are shown. We observe the improvement of the successive approximations to the solution.

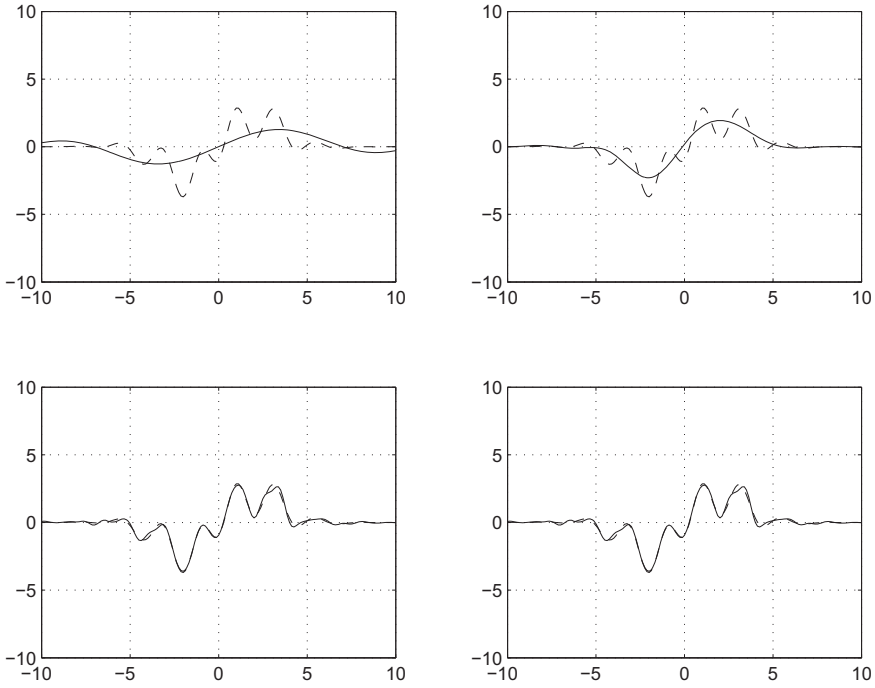


Figure 2: Reconstruction of $\tilde{f}$ for different level sets

4 CONCLUSIONS

The wavelet-vaguelet decompositon method seems to be efficient to calculate approximated solutions for this type of operator equations. In future, numerical efficiency for a more general class of operators will be investigated.

REFERENCES

[1] A. COHEN, M. HOFFMANN AND M. REIβ, *Adaptive Wavelet galerkin Methods for Inverse Problems*,SIAM J. Numer. Anal., 42, 1479-1501, 2004.

[2] A. COHEN, W. DAHMEN AND R. DEVORE, *Adaptive Wavelet Methods for Elliptic Operator Equations - Convergence Rates* , Math. Comp. 70 (2001), 27-75.

[3] V. DICKEN AND P. MAAβ, *Wavelet-Galerkin methods for ill-posed problems* ,J. Inv Ill-Posed Problems, Vol 4, N 3,pp 203-222, 2006.

[4] S. DAHLKE AND P.MAAβ, *An Outline of Adaptive Wavelet galerkin Methods for Tikhonov Regularization of Inverse Parabolic Problems: Recent development in theories and numerics* , International Conference on Inverse Problems, Hong Kong, China, 9-12 January 2002, .

[5] FANG-FANG DOU, CHU-LI FU AND YUN-JIE MA, *A wavelet-Galekin method for high order numerical differentiation* ,Applied Mathematics and Computation, 215, pp 3702-3712, 2010.

[6] J.R. LINHARES DE MATTOS AND E. PRADO LOPES, *A Wavelet Galerkin Method applied to partial differential equations with variable coefficients* Fifth Mississippi State Conference on Differential Equations and Computational Simulations, Electronic Journal of Differential Equations, Conference 10, 2003, pp 211-225.

[7] S. MALLAT, *Multi-Resolution approximation and wavelets* , Trans. Amerc. Math. Soc., Vol 315, pp 69-88, 1988.

[8] Y. MEYER, *Wavelets Algorithms and Applications*, SIAM, Philadelphia, 1993.

[9] Y. MEYER, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations* , AMS, Providence, 2001.

NON-LINEAR NORMAL MODES OF A ROTATING BEAM

Sebastián P. Machado†‡ and C. Martín Saravia†‡

†Grupo Análisis de Sistemas Mecánicos, Universidad Tecnológica Nacional Facultad Regional Bahía Blanca, 11 de Abril 461, B8000LMI, Bahía Blanca, Argentina, smachado@frbb.utn.edu.ar, msaravia@frbb.utn.edu.ar
‡Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET). Argentina

Abstract: Non-linear normal modes of vibration for a cantilever rotating beam are evaluated considering the continuous system in external and internal resonance conditions. The equation of motion is obtained in the form of an integral partial differential equation, taking into account mid-plane stretching, a rotational speed and modal damping. The internal resonance can be activated for a range of the beam´s rotational speed. For a range of the rotational speed, the second natural frequency is approximately three times the first natural frequency. The method of multiple scales is used to derive four first-order ordinary differential equations that govern the evolution of the amplitude and phase of the response. These equations are used to determine the steady state responses and their stability. Comparisons between different models (using linear and non-linear modes) are also provided. Numerical simulations show a complex dynamic scenario and detect chaos and unbounded motions in the instability regions of the periodic solutions.

Key words: Internal resonance dynamic
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCTION

Vibrations of rotating blades or beams have been a subject of constant research interest since they are applied in the design of helicopter blades, turbo-propeller blades, wind-turbine blades and robotic arms. The most simplified representation of a rotating beam is a one-dimensional Euler-Bernoulli model. A uniform rotating beam of doubly symmetric cross-section is a special case (no torsional motion: i.e., out-of-plane (flapping) vibration and in-plane (lead-lag) vibration are uncoupled). Owing to the stiffening effect of the centrifugal tension, one can expect the natural frequencies to increase with an increase in the speed of rotation. In several publications a cantilever beam under rotational speed has been considered and approximate methods such as Rayleigh-Ritz, Galerkin, finite element methods, etc., have been used to estimate the natural frequencies. However, a nonlinear dynamic analysis of rotating beam is rather rare in the literature [1-3]. Systematic procedures have been developed to obtain reduced-order models (ROMs) via nonlinear normal modes (NNMs) that are based on invariant manifolds in the state space of nonlinear systems [4]. These procedures initially used asymptotic series to approximate the geometry of the invariant manifold and have been used to study the nonlinear rotating Euler–Bernoulli beam [1]. Apiwattanalunggarn et al. [2] presented a nonlinear one-dimensional finite-element model representing the axial and transverse motions of a cantilevered rotating beam, which is reduced to a single nonlinear normal mode using invariant manifold techniques. As it can be noted, the interest of most of works about nonlinear dynamic of rotating beams is focused on the reduced-order model such as the invariant manifold solution. Turhan and Bulut [3] investigated the in plane nonlinear vibrations of a rotating beam via single- and two-degree-of-freedom models obtained through Galerkin discretization. In the last four references, the computational cost associated with generating the manifold solution and the efficiency of the resultant model was mainly analyzed. From a literature review, it is found that the study of internal resonance of a cantilever rotating slender beam subjected to a harmonic transverse load has not yet been explored so far. So, this article is focused on the analysis of reduced order models using linear and nonlinear normal modes, for the nonlinear planar vibration of a rotating composite cantilever beam with harmonic transverse load in the presence of internal resonances.

2. PROBLEM FORMULATION

The model is based on one-dimensional Euler-Bernoulli formulation where the geometric cubic nonlinear terms are included in the equation of motion due to midline stretching of the beam. We assume that the motion is planar and the cross sections remains plane during transverse bending. The laminate

stacking sequence of the uniform composite beam is assumed to be symmetric and balanced. A doubly symmetric cross-section box-beam is used and so out-of-plane (flapping) and in-plane (lead-lag) vibration are uncoupled. Neglecting rotary inertia and the transverse shear, the non-linear equations of motion of a rotating beam yields [5]:

$$\rho A\,\ddot{v} + EI\; v^{iv} - \left[\frac{EA}{2L}\int_0^L v'^2 \mathrm{dx} + \frac{\Omega^2 \rho A}{2}\left(\frac{L^2}{3} - x^2\right)\right] v'' + \rho A \Omega^2 v' x = F(x)\cos(\varpi t)\,. \tag{1}$$

Where Ω is the constant angular speed of the beam, ρA is the mass per unit length, EA and EI are the axial and flexural rigidity, ϖ is the excitation frequency, L is the beam length, and $F(x)$ describes the spatial distribution of the applied transverse harmonic load. Overdots indicate differentiation with respect to time and primes with respect to the axial co-ordinate. Introducing a nondimensional quantity for $x^* = x / L$, substituting this relationship in Eq. (1) with the corresponding boundary conditions, adding damping μ, and dropping the asterisk the expressions can be conveniently rewritten as:

$$\ddot{v} + \alpha v^{iv} + 2\mu\dot{v} - \chi v'' - \gamma v''\int_0^1 v'^2 \mathrm{dx} + \lambda v' = f\cos(\varpi t), \qquad BC \quad \begin{cases} v = 0 \quad \text{and} \quad v' = 0 \quad \text{at } x = 0 \\ v'' = 0 \quad \text{and} \quad v''' = 0 \quad \text{at } x = 1 \end{cases}, \tag{2}$$

$$\text{where} \quad \alpha = \frac{EI}{\rho A\, L^4}, \quad \chi = \frac{\Omega^2}{6}, \quad \gamma = \frac{EA}{2\rho A L^4}, \quad \lambda = \frac{\Omega^2}{2}, \quad f = \frac{F(x)}{\rho A}. \tag{3}$$

3. METHOD OF ANALYSIS

3.1. DIRECT METHOD USING NONLINEAR NORMAL MODES

In this section, the asymptotic method of multiple scales is applied directly to the partial differential equation and the associated boundary conditions Eq. (2). We seek an approximate solution to this weakly nonlinear distributed parameter system in the form of a first-order uniform expansion and introduce the time scales $T_n = \varepsilon^n t$, $n = 0,1,2,...$ A small parameter ε is introduced by ordering the linear damping and load amplitude as $\mu = \varepsilon\tilde{\mu}, f = \varepsilon\tilde{f}$. Moreover, the displacement $v(x,t)$ is expanded as:

$$v(x,t) = v_1(T_0, T_1, x) + \varepsilon\; v_2(T_0, T_1, x) + ... \tag{4}$$

Substituting Eq. (4) into Eq. (2) and equating coefficients of like powers of ε on both sides, we obtain

$$\text{Order } \varepsilon^0: \qquad D_0^2 v_1 + \alpha v_1^{iv} - \chi v_1'' + \lambda v_1' = 0\,, \tag{5}$$

$$\text{Order } \varepsilon^1: \qquad D_0^2 v_2 + \alpha v_2^{iv} - \chi v_2'' + \lambda v_2' = -2D_0 D_1 v_1 - 2\mu D_0 v_1 - \gamma v_1''\int_0^1 v_1'^2 \mathrm{dx} + f\cos(\varpi t)\,, \tag{6}$$

where $D_k = \partial / \partial T_k$. In this work, principal parametric resonance of first mode considering internal resonance is analyzed, involving the first two modes. The solution to the first-order perturbation can be expressed by:

$$v_1(T_0, T_1, x) = A_1(T_1)\phi_1(x)e^{i\omega_1 T_0} + A_2(T_1)\phi_2(x)e^{i\omega_2 T_0} + \mathrm{cc}, \tag{7}$$

where $\phi_m(x)$ are the mode shapes of the rotating cantilever beam (see Eq. 8), cc stands for the complex conjugate of the preceding terms and A_i are the unknown complex-valued functions.

$$\begin{aligned} \phi_m(x) = e^{x\beta_{4m}} + \Big\{ & e^{x\beta_{3m}}\left[-e^{\beta_{2m}}\beta_{2m}^2(\beta_{1m}-\beta_{4m}) + e^{\beta_{1m}}\beta_{1m}^2(\beta_{2m}-\beta_{4m}) + e^{\beta_{4m}}\beta_{4m}^2(\beta_{1m}-\beta_{2m})\right] \\ & + e^{x\beta_{2m}}\left[e^{\beta_{3m}}\beta_{3m}^2(\beta_{1m}-\beta_{4m}) - e^{\beta_{1m}}\beta_{1m}^2(\beta_{3m}-\beta_{4m}) - e^{\beta_{4m}}\beta_{4m}^2(\beta_{1m}-\beta_{3m})\right] \\ & + e^{x\beta_{1m}}\left[-e^{\beta_{3m}}\beta_{3m}^2(\beta_{2m}-\beta_{4m}) + e^{\beta_{2m}}\beta_{2m}^2(\beta_{3m}-\beta_{4m}) + e^{\beta_{4m}}\beta_{4m}^2(\beta_{2m}-\beta_{3m})\right]\Big\} \Big/ \Big[-e^{\beta_{2m}}\beta_{2m}^2 \\ & (\beta_{1m}-\beta_{3m}) + e^{\beta_{1m}}\beta_{1m}^2(\beta_{2m}-\beta_{3m}) + e^{\beta_{3m}}\beta_{3m}^2(\beta_{1m}-\beta_{2m})\Big]. \end{aligned} \tag{8}$$

In order to investigate the system response under internal and external resonance conditions, two detuning parameters, σ_1 and σ_2, are introduced: $\omega_2 = 3\omega_1 + \varepsilon\,\sigma_1$, $\varpi = \omega_1 + \varepsilon\,\sigma_2$. Substituting Eq. (7) to find the solution of Eq. (6), we get

$$D_0^2 v_2 + \alpha v_2^{iv} - \chi v_2'' + \lambda v_2' = \Gamma_1(T_1, x)e^{i\omega_1 T_0} + \Gamma_2(T_1, x)e^{i(3\omega_1 T_0 + \sigma_1 T_1)} + \frac{1}{2} f\ e^{i(\omega_1 T_0 + \sigma_1 T_2)} + \text{cc} + \text{NST}, \quad (9)$$

where NST stands for terms that do not produce secular or small divisor terms. By means of the solvability condition and introducing a Cartesian coordinates $A_k = \left[p_k(T_1) - iq_k(T_1)\right]e^{i\nu_k T_1}/2 \left(\text{for } k = 1,2\right)$, we obtain the following complex variable modulation equations for the amplitude and phase.

$$\begin{aligned}
p_1' &= -\mu_1 p_1 - \nu_1 q_1 + \gamma_{11} q_1 \left(p_1^2 + q_1^2\right) + \gamma_{12} q_1 \left(p_2^2 + q_2^2\right) - \delta_1 \left[2 p_1 q_1 p_2 - q_2 \left(p_1^2 + q_1^2\right)\right], \\
q_1' &= -\mu_1 q_1 + \nu_1 p_1 - \gamma_{11} p_1 \left(p_1^2 + q_1^2\right) - \gamma_{12} p_1 \left(p_2^2 + q_2^2\right) - \delta_1 \left[2 p_1 q_1 q_2 + p_2 \left(p_1^2 - q_1^2\right)\right] + \frac{1}{2} f_1, \\
p_2' &= -\mu_2 p_2 - \nu_2 q_2 + \gamma_{21} q_2 \left(p_1^2 + q_1^2\right) + \gamma_{22} q_2 \left(p_2^2 + q_2^2\right) + \delta_2 q_1 \left(3 p_1^2 - q_1^2\right), \\
q_2' &= -\mu_2 q_2 + \nu_2 p_2 - \gamma_{21} p_2 \left(p_1^2 + q_1^2\right) - \gamma_{22} p_2 \left(p_2^2 + q_2^2\right) + \delta_2 p_1 \left(3 q_1^2 - p_1^2\right).
\end{aligned} \quad (10)$$

where, the constants δ_i and γ_{ij} are presented in Table 1.

3.2. Discretized Method Using Linear Normal Modes

To perform the Galerkin discretization, we let

$$v(t, x) = \sum_{n=1}^{\infty} q_n(t)\phi_n(x), \quad (11)$$

where $q_n(t)$ are unknown functions of time to be determined and $\phi_n(x)$ are the eigenfunctions for the bending vibration of the stationary cantilever beam (Eq. 12).

$$\phi_n(x) = \cosh\beta_n x - \cos\beta_n x - \frac{\sinh\beta_n - \sin\beta_n}{\cosh\beta_n + \cosh\beta_n}\left(\sinh\beta_n x - \sin\beta_n x\right). \quad (12)$$

Substituting Eq. (11) into Eq. (2), the discretized equations of motion may then be expressed as:

$$\ddot{q}_n + \omega_n^2 q_n^{iv} + 2\mu\dot{q}_n + \gamma \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} c_{nijk} q_i q_j q_k = f_n \cos(\varpi t), \quad (13)$$

where ω_n is the nth linear frequency, f_n the nth modal force, and c_{nijk} are coefficients depending on eigenfunctions. In this case, applying the method of multiple time scales to study the non-linear equations (13), the displacements q_n are expanded as:

$$q_i(T_0, T_1) = \varepsilon q_i^{(0)}(T_0, T_1) + \varepsilon^2 q_i^{(1)}(T_0, T_1). \quad (14)$$

Following the same methodology as described in the previous section 3.1, we arrive to the same modulation equations (10). The discrepancy between both methods (direct and discretized) is in the eigenfunctions used to obtain the coefficients that govern the modulation equations (see Table 1).

Table 1. Coefficients of the modulation equations, Eq. (10).

Method	γ_{11}	γ_{12}	γ_{21}	γ_{22}	δ_1	δ_2
Direct	1.32	721.38	-95.69	-2256.73	-47.84	1.95
Discretized	9.36	314.71	-46.26	-334.39	-52.60	2.25

4. Results And Discussion

A three-to-one internal resonance is possible for a certain range of rotational speeds. The internal resonance is perfectly tuned when Ω is 326.82 rpm. The beam geometrical characteristics used are: L = 15

m, h = 0.3 m, b = 0.7 m, e = 0.05 m, where h = is the height, b = the width and e = the thickness of the beam cross-section. The analyzed material is graphite-epoxy whose properties are E_1 = 144 GPa, E_2 = 9.65 GPa, G_{12} = 4.14 GPa, G_{13}=4.14 GPa, G_{23} = 3.45 GPa, ν_{12} = 0.3, ν_{13} = 0.3, ν_{23} = 0.5, for a sequence of lamination {*45/-45/-45/45*}. In this case, ρA= 138.9 Kg/m and EI = 9.81 10^7 Nm^2. For the specific value of the rotating speed ω_1 = 5.45 Hz and ω_2 = 16.35 Hz. The equilibrium solutions correspond to periodic motions of the beam. Steady-state solutions are determined by zeroing $p_i' = q_i' = 0$ the right-hand members of the modulation equations, Eq. (10), and solving the non-linear system. The frequency-response curves are shown in Fig. 1, for an internal and external resonance condition. The modal amplitude a_i curves are obtained in function of the external detuning parameter σ_2. In this case, the forcing amplitude is f_1 = 0.005 and 0.05, modal damping d_i = 0.05 and internal detuning parameter σ_1= 0.04. The amplitude of the indirectly excited second mode is smaller in comparison with the first mode. Finally, the discrepancy between both approaches is more remarkable for the larger value of the excitation load amplitude.

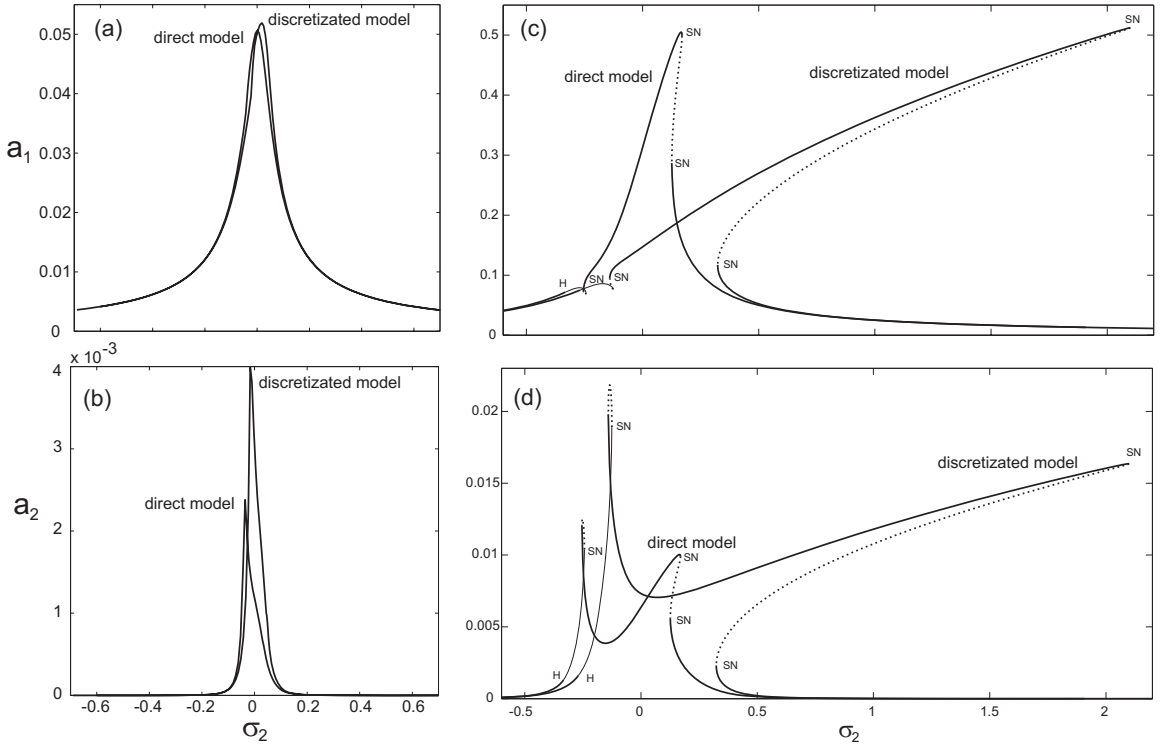


Figure 1: Frequency-response curves for the first and second mode, (a, b) f_1 = 0.005 (c, d) f_1 = 0.05. Solid (dotted) lines denote stable (unstable) equilibrium solutions and thin solid lines denote unstable foci. H = Hopf and SN = saddle-node bifurcation.

REFERENCES

[1] E. PESHECK, C. PIERRE, AND S.W. SHAW, *Accurate reduced order models for a simple rotor blade model using nonlinear normal modes*, Mathematical and computing modeling, 33 (2001), 1085-1097.

[2] P. APIWATTANALUNGGARN, S.W. SHAW, C. PIERRE, AND D. JIANG, *Finite-Element-Based Nonlinear Modal Reduction of a Rotating Beam with Large-Amplitude Motion*, Journal of Vibration and Control, 9 (2003), 235-263.

[3] O. TURHAN, AND G. BULUT, *On nonlinear vibrations of a rotating beam*, Journal of Sound and Vibration, 322 (2009) 314-335.

[4] S.W. SHAW, C. PIERRE, AND E. PESHECK, *Modal analysis-based reduced-order models for nonlinear structures-an invariant manifold approach*, The Shock and Vibration Digest, 31 (1999) 3-16.

[5] S.P. MACHADO, AND C.M. SARAVIA, *Non-linear dynamic response of a rotating thin-walled composite beam*, Mecánica Computacional Vol.29 (2010) 1203-1224.

[6] A.H. NAYFEH, *Nonlinear Interactions*, WILEY, 1996.

EVOLUTION OF AFFINE SHELLS

Salvador D. R. GIGENA†‡, Daniel J. A. ABUD‡ and Moisés BINIA‡

† *Facultad de Ciencias Exactas, Ingeniería y Agrimensura - Universidad Nacional de Rosario Avda. Pellegrini 250 - 2000 Rosario, sgigena@fceia.unr.edu.ar*

‡ *Facultad. de Ciencias Exactas, Físicas y Naturales - Universidad Nacional de Córdoba Avda. Vélez Sarsfield 299 - 5000 Córdoba sgigena@efn.uncor.edu - dabud@efn.uncor.edu - mbinia@hotmail.com*

Abstract: An Affine Shell, already defined in our own recent research papers, composed by a perfectly elastic homogeneous and isotropic material is submitted to certain forces which deform it. So, the shell goes from an original, undeformed state to a final, deformed one. This has been largely studied in our own publications. Now, we analyze the shell in the intermediate states while it is occurring deformation. Our proposal here is to describe these deformations by using PDE methods and their applications, especially Monge-Ampere type equations. The outcome is concerned with the treatment and development of evolution of affine shells in certain and given directions.

Key words: affine shells – evolution – stress-strain relations – basic inequalities
2000 AMS Subjects Classification: 74K25 – 53A15 – 35J60

1. INTRODUCTION

Theory of Shells is a topic of Geometry and Mathematical Physics with a rich history and many, diverse applications to the real world: Engineering, Industry, Avionics, and so on. The usual viewpoint of presentation makes use of the classical, Euclidean Geometry of Surfaces, i.e., invariant under the action of the Euclidean Group ASO(3;$\mathbb{R}$) [10], [11], [12]. The authors have been working on an alternative foundation and development of the theory of shells which is invariant under the action of the unimodular affine group ASL(3; $\mathbb{R}$). This is the so called "*affine geometry of surfaces*". For a given surface in three-dimensional space there are useful concepts such as "affine normal" and "affine distance", corresponding to the ones in Euclidean geometry. For further information, it is recommended to see references [4] to [9]. In the present article, following the definitions of geometrical objects for an initial, original, unstrained state and a final, deformed one, the authors are proposing to study intermediate states, giving rise to the treatment of what is usually called an "Evolution Process". Thus, for a given intermediate state we describe bidimensional compatibility conditions for an affine shell; equations of equilibrium for a solid shell; basic inequalities; estimates for the strain and stress tensors, as well as for their higher order covariant derivatives. Thus, description and validity of the behavior of the physical-geometrical objects of the shell in the intermediate states is the main goal of this article.

2. DESCRIPTION OF AN AFFINE SHELL

The middle surface of a (solid) shell in its original (undeformed) state, is denoted by M_0, parametrized locally by a vector function $X_0 : U \to \mathbb{R}^3$, where $U \subset \mathbb{R}^2$, which is assumed to be enough smooth. We write coordinates in the domain (u^1, u^2), $M_0 = X_0(U)$ and assume that X_0 is a topological immersion (embedding). Particles in the original state have curvilinear Lagrange coordinates (U^1, U^2, U^3) that for our present purposes shall be chosen in a special way: $(U^1, U^2, U^3) = (u^1, u^2, u)$ with $X(u^1, u^2, u) = X_0(u^1, u^2) + u\, N_{ua}$, where $X : U \times (-h, h) \to \mathbb{R}^3$, and N_{ua} is the vector field transversal to M_0 in the ***Unimodular Affine normal direction***. We assume that the shell is in its original, undeformed and unstrained state, until a certain initial time t_0 when certain forces, of any origin, are applied to it producing a deformation, which takes place along a certain finite period, until the final, strained state is reached, at a certain time t_1, with $t_1 > t_0$. At t_1 the shell is in equilibrium. We could

choose, for example, $t_0 = 0$. By looking at this process during deformation it is possible to refer to it as a *generic intermediate state of the shell*, occurring at a time t, where $t_0 < t < t_1$. In the state previous to deformation the border of the shell is made up of two "faces", which are surfaces parallel to M_0 at respective affine distance h, measured along the Affine normal N_{ua}, and of the "border" constituted by segments normal to the faces. Therefore, along the normal to M_0 coordinates U^1, U^2 remain constant while $U^3 := u$ measures the signed distance from M_0. Faces can be represented, then, by equations $U^3 = u = \pm h$ while M_0 is given by $U^3 = u = 0$, [4] to [9]. Consequently, the rest of geometrical objects change from one state to the other and the problem is to determine the nature and extension of such changes for every one of them reducing, under appropriate hypotheses, the obtainable information to all the intermediate middle surfaces. Consider now defined in the ambient space $\mathbb{R}^3$ an exterior 3-form, or non-trivial determinant function, denoted by the symbol $[\ ,\ ,\ ] = \det$. Objects of the geometry are represented, firstly, by $h^t_{\alpha\beta} = \left[\frac{\partial X^t_0}{\partial u^1}, \frac{\partial X^t_0}{\partial u^2}, \frac{\partial^2 X^t_0}{\partial u^\alpha \partial u^\beta}\right]$ and we assume that the surface is non-degenerate, i.e., $H^t = \det(h^t_{\alpha\beta}) \neq 0$. Then, we can write $g^t_{\alpha\beta} = |H|^{-1/4} h^t_{\alpha\beta}$, obtaining the ***Unimodular Affine First Fundamental***: $I^t_{ua} = \sum_{\alpha,\beta} g^t_{\alpha\beta} du^\alpha du^\beta$, a semi-Riemannian structure, [1], [2], [3], and [13]. The ***Unimodular Affine Second Fundamental Form***: $\nabla\left(I^t_{ua}\right) = II^t_{ua}$, in local coordinates written $II^t_{ua} = \sum_{\alpha\beta\gamma} g^t_{\alpha\beta\gamma} du^\alpha du^\beta du^\gamma$, with coefficients $g^t_{\alpha\beta\gamma}$ totally symmetric in their indices. Finally, the ***Affine Third Fundamental Form*** described in the following way: $III^t_{ua} = B^t_{\alpha\beta} du^\alpha du^\beta$ with $B^t_{\alpha\beta} = \sum_\gamma g^t_{\alpha\gamma} \left(B^\gamma_\beta\right)^t$. See [4] through [9].

The shell as a three-dimensional body and the Riemannian structure induced on that object by the ambient space metric is generated in a natural fashion. In Unimodular Affine Geometry that extension is not at all that immediate. However, it can also be realized in a canonical way: starting from the affine invariant pseudometric I_{ua}, defined on M_0: $g_{\alpha\beta} = I_{\text{ua}}\left(\frac{\partial X_0}{\partial u^\alpha}, \frac{\partial X_0}{\partial u^\beta}\right)$. A pseudo-metric in a generic intermediate state t, where $t_0 < t < t_1$, to be denoted by $G^t = \sum G^t_{ij} du^i du^j$. Thus, in the initial original state the shell (as a volume) will be denoted $\mathscr{C}$, the corresponding one in the deformed state is denoted $\mathscr{C}^*$, while in any of the intermediate states will be denoted $\mathscr{C}^t$. The same kind of notation will be used for the rest of geometrical objects, i.e., the metric coefficients of M_0: g_{ij}, in the unstrained state, g^*_{ij}, in the deformed state, and finally g^t_{ij}, in the intermediate state occurring at a t time, where $t_0 < t < t_1$. Let us observe first of all that, since bilinearity must be preserved, in affine normal coordinates of the shell we must necessarily have, on one hand that $G^t_{\alpha\beta} = G^t\left(\frac{\partial X^t_0}{\partial u^\alpha} + u\frac{\partial N_{\text{ua}}}{\partial u^\alpha}, \frac{\partial X^t_0}{\partial u^\beta} + u\frac{\partial N_{\text{ua}}}{\partial u^\beta}\right)$, with, $G^t_{\alpha\beta} := g^t_{\alpha\beta} - 2\ u\ B^t_{\alpha\beta} + u^2 \sum_\lambda \left(B^\lambda_\alpha\right)^t B^t_{\beta\lambda}$, $G^t_{3\alpha} = G^t_{\alpha 3} := 0$ and $G^t_{33} := 1$. See [4] trough [9].

Determination of compatibility conditions are obtained on the behavior of the various difference tensors that can be defined by comparing the different states of the shell. Integrability conditions must be satisfied, in all cases, by the middle surfaces, which are supposed to be enough smooth. Deformation of the shell is guaranteed by these conditions, for any perfectly elastic homogeneous and isotropic material. For these difference tensors $\varepsilon^t_{\alpha\beta} = \frac{1}{2}\left(g^t_{\alpha\beta} - g_{\alpha\beta}\right)$, $\sigma^t_{\alpha\beta\gamma} := g^t_{\alpha\beta\gamma} - g_{\alpha\beta\gamma}$, $w^t_{\alpha\beta} := B^t_{\alpha\beta} - B_{\alpha\beta}$, and the tensor defined by comparison between the corresponding Levi-Civita connections: $\left(\tilde{\Gamma}^{\mu}_{\alpha\beta}\right)^t = C^{\mu}_{\alpha\beta} + \tilde{\Gamma}^{\mu}_{\alpha\beta}$, hold:

1) The Affine Gauss condition:

$$\left(\varepsilon^{\beta,\delta}_{\beta,\delta}\right)^t - \left(\varepsilon^{\delta,\beta}_{\beta,\delta}\right)^t = \frac{1}{2}\left(\left(B^{\beta}_{\beta}\right)^t\left(g^{\delta}_{\delta}\right)^t - \left(B^{\beta}_{\delta}\right)^t\left(g^{\delta}_{\beta}\right)^t + \left(B^{\delta}_{\delta}\right)^t\left(g^{\beta}_{\beta}\right)^t - \left(B^{\delta}_{\beta}\right)^t\left(g^{\beta}_{\delta}\right)^t\right)$$
$$-\frac{1}{2}g^{\beta\alpha}g^{\delta\gamma}\left(\left(A^{\eta}_{\gamma\beta}\right)^t \cdot g^t_{\alpha\eta\delta} - \left(A^{\eta}_{\gamma\delta}\right)^t \cdot g^t_{\alpha\eta\beta}\right)$$
$$+\frac{1}{2}\left(g^{\beta}_{\mu}\right)^t\left(-B^{\delta}_{\delta}\delta^{\mu}_{\beta} + A^{\delta\eta}_{\beta}\cdot A^{\mu}_{\eta\delta}\right)$$
$$-g^t_{\lambda\mu}\,g^{\beta\alpha}g^{\delta\gamma}\left(C^{\lambda}_{\alpha\beta}C^{\mu}_{\gamma\delta} - C^{\lambda}_{\alpha\delta}C^{\mu}_{\beta\gamma}\right)$$

2) The Affine Mainardi-Codazzi condition:

$$\sigma^t_{\alpha\beta\gamma,\delta} - \sigma^t_{\alpha\beta\delta,\gamma} = g^t_{\mu\beta\gamma}C^{\mu}_{\alpha\delta} + g^t_{\mu\alpha\gamma}C^{\mu}_{\beta\delta} - g^t_{\mu\beta\delta}C^{\mu}_{\alpha\gamma} - g^t_{\mu\alpha\delta}C^{\mu}_{\beta\gamma} +$$
$$+B^t_{\alpha\delta}g^t_{\beta\gamma} + B^t_{\beta\delta}g^t_{\alpha\gamma} - B^t_{\alpha\gamma}g^t_{\beta\delta} - B^t_{\beta\gamma}g^t_{\alpha\delta} -$$
$$-B_{\alpha\delta}g_{\beta\gamma} - B_{\beta\delta}g_{\alpha\gamma} + B_{\alpha\gamma}g_{\beta\delta} + B_{\beta\gamma}g_{\alpha\delta}$$

3) The Codazzi condition for the affine shape operators:

$$\left(w^{\alpha}_{\beta,\alpha}\right)^t - \left(w^{\alpha}_{\alpha,\beta}\right)^t = g^{\alpha\rho}\left[B^t_{\beta\mu}\left(C^{\mu}_{\rho\alpha} + \left(A^{\mu}_{\rho\alpha}\right)^t\right) - B^t_{\alpha\mu}\left(C^{\mu}_{\rho\lambda} + \left(A^{\mu}_{\rho\beta}\right)^t\right) + B_{\alpha\mu}A^{\mu}_{\rho\beta}\right]$$

The contravariant components of the stress tensor, $\left(\tau^{mk}\right)^t$, are connected with components of the strain tensor, $\left(\varepsilon_{mk}\right)^t$, by means of the stress-strain relations $\left(\tau^{mk}\right)^t := \sqrt{\frac{G}{G^t}}\frac{\partial W}{\partial\left(\varepsilon_{mk}\right)^t}$ where W is the ***Strain Energy Density*** of the given material. Basic inequalities were previously obtained in [6]. When represented in the form of Monge's M_0 has all of its geometrical properties related to a given function f assumed to be enough differentiable and satisfying a Monge-Ampère PDE type equation: $\det\left(\partial_{\alpha\beta}f\right) = \pm F$ with boundary conditions, whereas there also hold bounds for the function f and its derivatives. Estimates obtained from the *equilibrium equations for the Affine Theory of Shells*, (see [5], [10] for full details) are completely valid. These establish a comparison between the unstrained state and the final, strained one:

a) $$\sum_m \tau_{im\,;\,m} = P_i = F\left(\eta,\tau\right)\left(\tau\tau' + \eta'\tau + \eta\tau'\right)$$

b) $$\sum_r \tau_{hk\,;\,rr} + 2\mu\sum_r \tau_{rr\,;\,hk} = Q_{hk} = F\left(\eta,\tau\right)\begin{pmatrix}\left(\tau'\right)^2 + \eta'\tau\tau' + \left(\eta'\tau\right)^2 + \left(\eta'\right)^2 + \tau'' + \eta''\tau + \\ +\eta'\tau' + \eta\tau + \eta'' + \left(\eta\right)^2 + \eta''\tau^2 + \left(\eta\tau\right)^2 + \\ +\left(\eta'\tau\right)^2 + \eta\eta''\tau + \left(\eta\right)^3\tau + \eta\tau\left(\eta'\right)^2\end{pmatrix}$$

and, consequently,

c) $$\tau_{hk;rr}+\frac{1}{1+\nu}\tau_{rr;hk}=\frac{\nu}{1+\nu}\left(\tau_{1,rr}\delta_h^k-\tau_{1,rk}\delta_r^h-\tau_{1,kr}\delta_k^r\right)+\tau_{hr;kk}+\tau_{rk;hr}+..+\text{higher order terms}$$

d) $$\begin{aligned}\tau_{ij;kl}=\tau_{ij,kl}-(\Gamma_{ik}^h)_{,l}\tau_{hj}-\Gamma_{ik}^h\tau_{hj,l}-(\Gamma_{jk}^m)_{,l}\tau_{im}-\Gamma_{jk}^m\tau_{im,l}-\Gamma_{il}^h(\tau_{hk,j}-\Gamma_{hj}^r\tau_{rk}-\Gamma_{kj}^s\tau_{sh})-\\-\Gamma_{jl}^m(\tau_{mi,k}-\Gamma_{mk}^r\tau_{ri}-\Gamma_{ik}^s\tau_{sm})-\Gamma_{kl}^q(\tau_{qi,j}-\Gamma_{qi}^r\tau_{rj}-\Gamma_{ij}^s\tau_{sq})\end{aligned}$$

Therefore, the following estimates for partial derivatives of higher order

e) $\partial_{k_1,k_2}\ \tau_{i_1,i_2,\ldots,i_n}=O\left(\varepsilon^2\lambda^{1-n}h^{-1}\right)$ and f) $\partial_{k_1,k_2}\ \tau_{\alpha_1,\alpha_2,\ldots,\alpha_n}=O\left(\varepsilon^2\lambda^{-n}\right)$.

While $\Gamma_{kr}^i=F(\eta)(\eta')$, and its successive derivatives, $\eta=O\left(h^2R^{-2}\right)$, $\eta'=O\left(R^{-5/2}\right)$, $\eta''=O\left(R^{-7/2}\right)$, …, the conclusion is that the same kind of estimates are valid for the corresponding covariant derivatives respect to the Levi-Civita connection: $\tau_{k_1,k_2\,;\,i_1,i_2,\ldots,i_n}=O\left(\varepsilon^2\lambda^{1-n}h^{-1}\right)$.

3. CONCLUSION

Validation of general equations in the theory of affine shells, considered for the intermediate states between an initial and a final one, are hereby done. So, for a generic given intermediate state we can describe bidimensional compatibility conditions; equations of equilibrium; basic inequalities; estimates for the strain and stress tensors, as well as for their higher order covariant derivatives. Thus, description and validity of the behavior of the physical-geometrical objects of the shell in the intermediate states are obtained in this article. Particular cases of surface evolution can be seen, for example, in [14].

4. REFERENCES

[1] S. GIGENA *Constant Affine Mean Curvature Hypersurfaces of Decomposable Type*, PROC OF SYMP. IN PURE MATH, AMERICAN MATH SOCIETY, VOL. 54, PART 3, 289-316, 1993.

[2] S. GIGENA *Hypersurface Geometry and Related Invariants in a Real Vector Space*, Editorial Ingreso, Córdoba, Argentina, 1996.

[3] S. GIGENA *Ordinary Differential Equations in Affine Geometry*, Le Matematiche,Vol. LI, Fasc.I, 119-151, 1996.

[4] S. GIGENA, M. BINIA AND D. ABUD *Condiciones de Compatibilidad para Cáscaras Afines*, Mecánica Computacional Vol. XXI, 1862-1881, 2002.

[5] S. GIGENA, M. BINIA AND D. ABUD *Ecuaciones de equilibrio en Cáscaras Afines*, Mecánica Computacional, Vol. XXII, 1953-1963, 2003.

[6] S. GIGENA, D. ABUD AND M. BINIA *Teoría de Cáscaras Afines: Desigualdades Básicas*, Mecánica Computacional, Vol. XXIII, 639-652, 2004.

[7] S. GIGENA, D. ABUD AND M. BINIA *Teoría de Cáscaras Afines: Estimativas de la Tensión y la Deformación*, Mecánica Computacional, Vol. XXIV, 2745-2758, 2005.

[8] S. GIGENA, D. ABUD *Theory Of Affine Shells: Higher Order Estimates of the Strain and Stress Tensors treated by P.D.E. methods*, Actas X Congreso Dr. Antonio Monteiro, 1-17, 2009.

[9] S. GIGENA, D. ABUD AND M. BINIA *Theory of Affine Shells: Higher Order Estimates*, Mecánica Computacional, Vol. XXIX, 969-988, 2010.

[10] F. JOHN *Estimates for the Derivatives of the Stresses in a Thin Shell and Interior Shell Equations*, Comm. Pure Appl. Math. Nº 18, 235-267, 1965.

[11] F. JOHN *Refined Interior Equations for Thin Elastic Shells*, Comm. Pure Appl.Math 24, 583-615, 1971.

[12] W.T. KOITER *On the mathematical foundation of shell theory*, Proc. Int. Congress On Mathematics, Nice Vol. 3, Paris, 1971, 123-130, 1970.

[13] K. NOMIZU AND T. SASAKI *Affine Differential Geometry*, Cambridge U. Press, 1994.

[14] K. LEICHTWEISS *Remarks on Affine Evolutions*, Abh. Math. Sem. Univ. Hamburg, Nº 66, 355-376, 1996.

SERIES DE POTENCIAS CON PARTICIÓN DEL DOMINIO PARA EL ANÁLISIS MODAL DE SISTEMAS CONTINUOS

Ariel E. Matusevich, José A. Inaudi y Julio C. Massa

Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de correo 916, 5000 Córdoba, Argentina, ariel.matusevich@gmail.com, http://www.efn.uncor.edu

Resumen: En este trabajo se analizan dos limitaciones del método convencional de series de potencias en el ámbito del análisis modal de sistemas continuos. La primera limitación está relacionada con el dominio de convergencia de la solución en series; si este dominio no incluye la región bajo análisis, la expansión en series arroja resultados sin sentido. La segunda limitación es de naturaleza computacional; al calcular frecuencias naturales y modos de vibrar de modelos continuos en frecuencias altas ocurren inconvenientes numéricos. Para remediar estas limitaciones, se propone utilizar un método de series de potencias con partición del dominio. Mediante un ejemplo sencillo, el análisis en vibración axial de una barra de sección transversal variable, se demuestra que el método de series de potencias con partición del dominio es más versátil que el método de series de potencias convencional.

Palabras claves: *series de potencias, vibraciones de sistemas continuos, partición del dominio.*

1. INTRODUCCIÓN

El Método de Series de Potencias (MSP) es una técnica ampliamente difundida para resolver ecuaciones diferenciales ordinarias, cuyos orígenes se remontan al siglo diecisiete [1]. Esencialmente, el MSP consiste en proponer como solución de una ecuación diferencial a una serie de Taylor infinita, sustituir la solución propuesta en la ecuación diferencial y mediante manipulaciones algebraicas, obtener relaciones de recurrencia para los coeficientes de la serie. En aplicaciones prácticas, la serie se trunca considerando un número finito de términos que permita aproximar la solución con una precisión determinada. Usualmente, las relaciones de recurrencia se desarrollan a mano o con procesadores simbólicos; esta desventaja se puede subsanar reformulando el método en función de operadores matriciales [2].

El MSP no es un método infalible y no debería usarse en forma indiscriminada. Una solución en series de potencias converge en una región libre de singularidades; cuando existen singularidades, la convergencia de la solución no está garantizada. Adicionalmente, ocurren problemas numéricos cuando se utiliza el MSP para el cálculo de frecuencias naturales y modos de vibrar de modelos continuos a frecuencias altas; no se obtienen resultados satisfactorios aun utilizando cuádruple precisión en los cálculos [3]. Estos conceptos y limitaciones, los cuales se repasan en este trabajo, justifican la utilización de un método de series de potencias con partición del dominio [2].

Este trabajo está organizado de la siguiente manera. En la sección 2 se presenta el problema de valores de frontera de una barra en vibración axial, ejemplo utilizado para ilustrar la aplicación de los métodos analizados. En la sección 2.1 se repasan conceptos importantes sobre la utilización del MSP y en la sección 2.2 se introduce el método de series de potencias con partición del dominio. En la sección 3 se estudia un ejemplo práctico y se discuten los resultados obtenidos. Finalmente, se exponen las conclusiones del trabajo.

2. PROBLEMA DE VALORES DE FRONTERA DE UNA BARRA EN VIBRACIÓN AXIAL

Los modos de vibración ϕ de una barra recta de longitud finita en vibración axial, de sección transversal variable y propiedades del material uniformes (densidad ρ y módulo de Young E), satisfacen la siguiente ecuación diferencial [4]:

$$A\,\frac{d^2\phi}{d\xi^2}+\frac{dA}{d\xi}\,\frac{d\phi}{d\xi}\,L+A\,\bar{\omega}^2\phi=0\,,\qquad \bar{\omega}=\sqrt{\frac{\rho L^2}{E}}\,\omega \tag{1}$$

donde A es el área de la sección transversal de la barra, L es su longitud, $\xi = x/L$ es la coordenada adimensional, $\bar{\omega}$ es la frecuencia adimensional y ω es la frecuencia natural de vibración. Los valores de ω que proveen soluciones no triviales a la ecuación (1) para las condiciones de borde del problema son las frecuencias naturales del modelo ω_i (con $i = 1, 2, \ldots, \infty$).

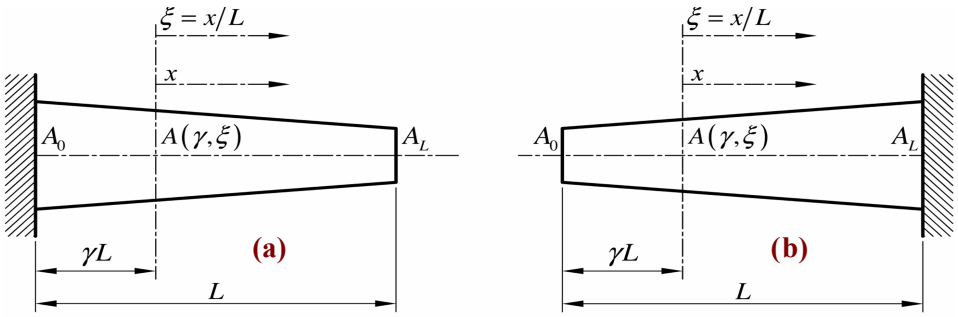


Figura 1: (a) barra fija-libre, (b) barra libre-fija

El análisis en vibraciones libre de una barra fija en un extremo y libre en el otro puede analizarse como "fija-libre" (Fig. 1a) o "libre-fija" (Fig. 1b) dependiendo del sistema de coordenadas adoptado. En las Figs. 1a y 1b se muestra una barra con variación lineal de su sección transversal que se utiliza para ilustrar la aplicación de los métodos numéricos. El origen de las coordenadas x y ξ está ubicado a una distancia γL del extremo izquierdo de la barra, donde γ es un parámetro que puede tomar cualquier valor entre 0 y 1. En el caso de la barra fija-libre, se deben cumplir las siguientes condiciones en los extremos:

$$\phi(-\gamma)=0\,, \qquad E\,A(1-\gamma)\frac{d\phi}{d\xi}(1-\gamma)=0 \quad\rightarrow\quad \frac{d\phi}{d\xi}(1-\gamma)=0 \tag{2}$$

En cambio, en la barra libre-fija se requiere que

$$E\,A(-\gamma)\frac{d\phi}{d\xi}(-\gamma)=0 \quad\rightarrow\quad \frac{d\phi}{d\xi}(-\gamma)=0\,, \qquad \phi(1-\gamma)=0 \tag{3}$$

A continuación se repasan algunos conceptos importantes sobre la aplicación el MSP y se introduce el método de series de potencias con partición del dominio [2].

2.1. SOLUCIÓN MEDIANTE SERIES DE POTENCIAS

Si ξ_0 es un punto regular de la ecuación (1) es posible desarrollar una solución en series de potencias que converja en un dominio definido por

$$|\,\xi-\xi_0\,|<R \tag{4}$$

donde R es la distancia entre ξ_0 y el punto singular más próximo. Dicha distancia, en el plano complejo, representa la cota inferior para el radio de convergencia de la solución en series [5].

Si la serie se centra alrededor del origen del sistema de coordenadas, $\xi_0 = 0$. Moviendo este origen dentro de la región de análisis, se pueden considerar otros puntos de expansión.

Si dividimos ambos términos de (1) por la función $A(\gamma,\xi)$

$$\frac{d^2\phi}{d\xi^2}+\frac{1}{A}\frac{dA}{d\xi}\frac{d\phi}{d\xi}L+\bar{\omega}^2\phi=0 \tag{5}$$

la ecuación diferencial queda expresada en forma estándar [5]. En esta forma se pone en evidencia que las singularidades de la ecuación diferencial son las raíces de $A(\gamma,\xi)$. Si consideramos que A varía en forma lineal a lo largo de la barra

$$A(\gamma,\xi)=a\,\xi+b, \qquad a=A_L-A_0, \qquad b=A_0+(A_L-A_0)\gamma \tag{6}$$

existe una sola singularidad ubicada en $\xi=-b/a$; por lo tanto, el límite inferior para el radio de convergencia resulta

$$R=\left|\frac{-b}{a}\right|=\left|\frac{A_0}{A_0-A_L}-\gamma\right| \tag{7}$$

Como indica la expresión (7), R depende de dos aspectos: (i) la geometría de la barra y (ii) la ubicación del sistema de coordenadas alrededor del cual se centra la serie.

2.2. SOLUCIÓN EN SERIES DE POTENCIAS CON PARTICIÓN DEL DOMINIO

La Figura 2 muestra una barra libre-fija dividida en s elementos de longitudes L_1, L_2, … , L_s. La partición del elemento en regiones cuyas longitudes L_i satisfacen: $L_i < R$, garantiza la existencia de soluciones en series de potencias en cada región. La aproximación del modo ϕ se transforma en una función definida en subdominios 1, 2, … , s.

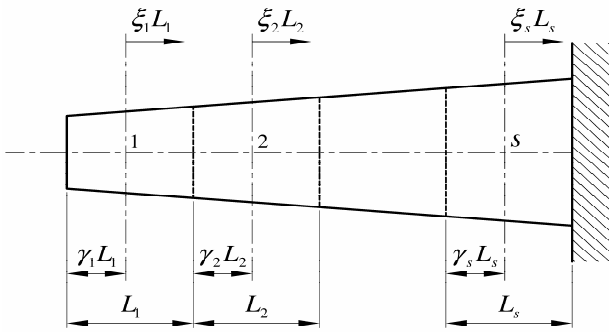


Figura 2: Barra con partición del dominio

Para el modelo libre-fijo de la Fig. 2, se deben satisfacer condiciones en los bordes $\xi_1 = -\gamma_1$, $\xi_s = 1 - \gamma_s$ y deben imponerse $2s$–2 condiciones de continuidad en las uniones entre particiones. Más detalles de este método y su implementación matricial pueden consultarse en [2].

3. EJEMPLOS DE ANÁLISIS

Se calcula la frecuencia fundamental de una barra libre en un extremo y fija en el otro, utilizando el MSP en cuatro casos equivalentes que se detallan en la Tabla 1. Luego se resuelve el mismo ejemplo mediante el método de series de potencias con partición del dominio.

Tabla 1: Casos analizados

Caso	Borde izquierdo	Borde derecho	Parámetro γ	Función $A(\gamma,\xi)$	R [Ec. (7)]
1	fijo	libre	0	$A(0,\xi) = -\xi + 2$	2
2	libre	fijo	0	$A(0,\xi) = \xi + 1$	1
3	fijo	libre	$1/2$	$A(1/2,\xi) = -\xi + 3/2$	$3/2$
4	libre	fijo	$1/2$	$A(1/2,\xi) = \xi + 3/2$	$3/2$

Notar que en los en los casos 1, 3 y 4 la convergencia del MSP está asegurada ya que $R > 1$. En el caso 2, $R = 1$ y por lo tanto, la existencia de una solución en series debe analizarse. En la Tabla 2 se muestran los resultados para estos cuatro casos en función del grado (n) del polinomio que aproxima la solución.

Como se observa en la Tabla 2, la convergencia en el caso 2 es extremadamente lenta; la solución nunca se estabiliza. Ese ejemplo representa un caso límite entre el éxito y el fracaso de una solución en series centrada en el borde izquierdo de la barra. Si se analizara una barra libre-fija cuya función de área tuviera una pendiente $a > 1$, resultaría $R < 1$ y el MSP arrojaría resultados sin sentido.

La velocidad de convergencia es, en general, muy sensible a la distancia al centro de la expansión. Si se hace coincidir el centro de la expansión con un borde del elemento, los puntos cercanos al otro borde pueden experimentar una convergencia lenta. En cambio, al ubicar el centro de la expansión a la mitad del elemento ($\gamma = 1/2$) se minimiza la distancia máxima entre los puntos considerados y el centro de la expansión. Esta elección puede resultar en una convergencia más rápida, hecho que se verifica en los casos 3 y 4. Notar que en esos casos, los resultados de los modelos fijo-libre y libre-fijo son idénticos.

Los resultados de la Tabla 3 corresponden a barras libres-fijas divididas en regiones uniformes y fueron obtenidos mediante polinomios centrados a la izquierda de igual grado en cada subdominio. Estos resultados demuestran que mientras mayor es la razón R/L_i, más rápido converge la serie. Cuando se utilizan polinomios centrados a la mitad de cada partición, la convergencia se acelera notablemente [2].

Tabla 2: Soluciones en series de potencias para la frecuencia fundamental $\bar{\omega}_1$

Grado	Modelo: fijo-libre		Modelo: libre-fijo	
n	Caso 1 ($\gamma = 0$)	Caso 3 ($\gamma = 1/2$)	Caso 2 ($\gamma = 0$)	Caso 4 ($\gamma = 1/2$)
5	1,787376881424038	1,796864135421148	1,948343381090857	1,796864135421148
7	1,799262256950063	1,794220613434518	1,877980005594379	1,794220613434518
13	1,794067262363255	1,794011212724370	1,842212513691217	1,794011212724370
14	1,794039244169542	1,794010998518423	1,751667896486866	1,794010998518423
15	1,794025121850327	1,794010938952015	1,835965764956050	1,794010938952015
20	1,794011353980687	1,794010904890254	1,763650488053592	1,794010904890254
30	1,794010905201035	1,794010904758691	1,773410102254346	1,794010904758691
40	1,794010904759122	1,794010904758688	1,778428453437813	1,794010904758688
60	1,794010904759122		1,783536220889923	
80	1,794010904758689		1,786122928034675	
100	1,794010904758689		1,787685291203742	

Tabla 3: Cálculo de $\bar{\omega}_1$ mediante series de potencias con partición del dominio

Grado	Número de subdivisiones del dominio, modelo libre-fijo ($\gamma = 0$)							
n	1	2	3	4	5	6	7	8
5	1,9483	1,8072	1,7963	1,7947	1,7943	1,7941	1,7941	1,7940
6	1,7037	1,7884	1,7934	1,7939	1,7940	1,7940	1,7940	1,7940
7	1,8780	1,7967	1,7942	1,7940	1,7940	1,7940	1,7940	
8	1,7248	1,7927	1,7939	1,7940				
9	1,8624	1,7947	1,7940					
10	1,7367	1,7937	1,7940					
11	1,8506	1,7942						
12	1,7453	1,7939						
13	1,8422	1,7941						
14	1,7517	1,7940						
15	1,8360	1,7940						

3.1. PROBLEMAS NUMÉRICOS EN ALTA FRECUENCIA

El cálculo de frecuencias naturales altas requiere órdenes elevados en los polinomios de aproximación utilizados. En el caso del ejemplo analizado, utilizando el MSP sobre el dominio completo del elemento y $\gamma = 1/2$, se pueden calcular con precisión las primeras 21 frecuencias naturales [2]. Más allá de este límite, la combinación de alta frecuencia y grados elevados en los polinomios de aproximación ocasiona inconvenientes numéricos que impiden proseguir con los cálculos. Esta situación puede mejorarse utilizando la partición del dominio propuesta en la sección 2.2. Dividiendo la barra en 5 dominios iguales y utilizando polinomios centrados ($\gamma = 1/2$) de grado $n = 35$ en cada partición, se pueden calcular las primeras 35 frecuencias naturales de este ejemplo con excelente precisión.

4. CONCLUSIONES

Se han analizado dos limitaciones del método de series de potencias convencional: (*i*) la existencia de singularidades dentro de la región de análisis puede ocasionar que el método convencional falle y (*ii*) dificultades numéricas en la estimación de frecuencias naturales y modos de vibrar en modelos continuos en alta frecuencia. Ambas limitaciones pueden remediarse mediante la partición del dominio, buscando soluciones en series de potencias en cada intervalo o subdominio. Esta simple modificación del método de series de potencias aumenta su rango de aplicación en problemas de vibraciones de sistemas continuos en alta frecuencia.

REFERENCIAS

[1] E. HAIRER, S. NORSET, G. WANNER AND M. CULLEN, *Solving ordinary differential equations I: Nonstiff problems*, Springer Verlag, 2008.

[2] J. INAUDI AND A. MATUSEVICH, *Domain-partition power series in vibration analysis of variable-cross-section rods*, Journal of Sound & Vibration, 321 (2010), pp. 4539-4549.

[3] S. NAGULESWARAN, *A direct solution for the transverse vibration of Euler Bernoulli wedge and cone beam*, Journal of Sound & Vibration, 172 (1994), pp. 289-304.

[4] R. CLOUGH AND J. PENZIEN, *Dynamics of structures*, McGraw-Hill, 1993.

[5] G. ZILL AND M. CULLEN, *Differential equations with boundary-value problems*, Brooks/Cole, 2001.

SOLUCIÓN NUMÉRICA DE LA ECUACIÓN DNLS NO DIFUSIVA CON UNA ONDA COMO CONDICIÓN INICIAL

Gustavo Krause† y Sergio Elaskar‡

Departamento de Aeronáutica, Universidad Nacional de Universidad Nacional de Córdoba y CONICET
Av. Vélez Sársfield 1611, 5000 Córdoba, Argentina
† *gustavojavierkrause@gmail.com*
‡ *sergio.elaskar@gmail.com*

Resumen: La ecuación Derivada No Lineal de Schrödinger (DNLS) posee la capacidad de describir la propagación de ondas de Alfvén de amplitud finita circularmente polarizadas tanto para plasmas calientes como fríos. Considerando esta ecuación sin efectos difusivos manteniendo los términos no lineal y dispersivo, es posible establecer condiciones analíticas de estabilidad modular, pero esto no describe en qué forma evoluciona el sistema ni determina el tiempo en el que se produce la inestabilidad. En el presente trabajo se soluciona numéricamente la ecuación DNLS mediante técnicas espectrales para las derivadas espaciales y un esquema de Runge-Kutta de cuarto orden para el avance en el tiempo. La investigación consiste en verificar numéricamente las condiciones analíticas de estabilidad modular y obtener además el tiempo de inestabilidad y la forma de evolución, encontrándose que para este caso de una onda inicial la misma depende de la amplitud inicial.

Palabras claves: *ondas de Alfvén, DNLS, métodos espectrales*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

La estructura de Alas de Alfvén o del campo electromagnético que se genera por el movimiento de una amarra espacial sumergida en un plasma puede abordarse diferenciando las regiones cercanas al cuerpo (campo próximo) de las alejadas (campo lejano). Para el segundo caso es posible un estudio mediante análisis lineal [1], mientras que en las cercanías del elemento existe una serie de fenómenos que ameritan un análisis más detallado. En este sentido, una posibilidad para el estudio de la evolución de los campos es por medio de la ecuación "Derivative Non-Linear Schrödinger Equation (DNLS)" la cual describe la propagación paralela o casi paralela de ondas de Alfvén circularmente polarizadas [2]. La característica que hace atractiva a la DNLS es la gran cantidad de soluciones exactas que se conocen [3], además esta ecuación ha sido estudiada por tres técnicas alternativas: búsqueda de soluciones exactas [4], integración numérica [5] y [6], y reducción a un sistema de ecuaciones diferenciales ordinarias suponiendo ondas viajeras estacionarias [7] y mediante un modo finito de modos [8] y [9].

2. GENERALIDADES

La ecuación DNLS puede obtenerse a partir de las ecuaciones de la Magnetogasdinámica considerando plasma neutro, compuesto por dos fluidos, despreciando la inercia de los electrones y la corriente de desplazamiento [10].

Si el campo magnético B_0 posee la dirección del eje z y el sistema de referencia se mueve con velocidad de Alfvén, la DNLS puede ser expresada adimensionalmente de la siguiente manera [2] - [4]

$$\frac{\partial \phi}{\partial t} \pm \mathrm{i} \frac{\partial^2 \phi}{\partial z^2} \pm \frac{\partial}{\partial z}\left(\phi |\phi|^2\right) = 0 \,, \tag{1}$$

donde $\phi = \dfrac{B_x \pm \mathrm{i} B_y}{B_0}$, $\omega_{ci}\, t \to t$, $\dfrac{\omega_{ci}}{V_A} z \to z$, siendo ω_{ci} la frecuencia de ciclotrón iónica y V_A la velocidad de Alfvén. El signo superior en el término dispersivo corresponde a ondas polarizadas a izquierda (LH), mientras que el inferior a derecha (RH) propagándose en la dirección del eje z. El signo del término no lineal puede ser negativo solamente para el caso de plasmas calientes [4].

3. SOLUCIÓN MEDIANTE TÉCNICAS ESPECTRALES

La simulación numérica de la DNLS considera métodos espectrales para evaluar las derivadas espaciales y un esquema de Runge-Kutta de cuarto orden para avanzar en el tiempo. Los métodos espectrales exigen que las soluciones satisfagan condiciones de contorno periódicas, es decir $\phi(z,t) = \phi(z+L,t)$, con L la longitud del dominio. En muchas situaciones para la DNLS el fenómeno de interés no está influenciado por lo que sucedo en el contorno por lo que simulaciones numéricas con condiciones de borde periódicas son una buena solución [3]. Considerando la ecuación (1), siendo $\mathbf{F}\left[\phi(z)\right] = \hat{\phi}(k)$ la Transformada de Fourier y $\mathbf{F}^{-1}\left[\hat{\phi}(k)\right] = \phi(z)$ su inversa, se tiene que

$$\frac{\partial \phi}{\partial t} = \mathbf{F}^{-1}\left\{\mp(\mathrm{i}k)\mathbf{F}\left[\phi(z)\left|\phi(z)\right|^2\right] \pm (\mathrm{i}k^2)\mathbf{F}\left[\phi(z)\right]\right\}. \tag{2}$$

Para la discretización espacial de la ecuación (2) se utiliza la Transformada de Fourier Discreta (DFT), la cual se calcula a través de la Transformada Rápida de Fourier. La DFT y su inversa están dadas por

$$\bar{\phi}(k) = \sum_{j=1}^{N} \phi(z_j)\,\mathrm{e}^{\mathrm{i}kz_j} \quad ; \quad k = -\frac{N}{2}+1,\ldots,\frac{N}{2} \qquad\qquad \phi(z_j) = \sum_{k=-\frac{N}{2}+4}^{\frac{N}{2}} \bar{\phi}(z)\,\mathrm{e}^{\mathrm{i}kz_j} \quad ; \quad j = 1,\ldots,N \tag{3}$$

con lo que la ecuación (2) resulta

$$\mathrm{Re}\left(\frac{\partial \phi_j}{\partial t}\right) = \mathbf{F}^{-1}\left\{\pm k_j\,\mathbf{F}\left[\mathrm{Im}\left(\phi_j\right)\left|\phi_j\right|^2\right] \mp k^2\,\mathrm{Im}\left(\bar{\phi}_j\right)\right\};\ \mathrm{Im}\left(\frac{\partial \phi_j}{\partial t}\right) = \mathbf{F}^{-1}\left\{\mp k_j\,\mathbf{F}\left[\mathrm{Re}\left(\phi_j\right)\left|\phi_j\right|^2\right] \pm k^2\,\mathrm{Re}\left(\bar{\phi}_j\right)\right\}. \tag{4}$$

Para la simulación el espacio físico es discreto y cerrado y por lo tanto también lo será el espacio de Fourier, entonces:

$$z \in \left(-\frac{L}{2}+h, -\frac{L}{2}+2h, \ldots, \frac{L}{2}-h, \frac{L}{2}\right) \rightarrow\ k \in \left[\left(\frac{2\pi}{L}\right)\left(-\frac{N}{2}+1, -\frac{N}{2}+2, \ldots, \frac{N}{2}-1, \frac{N}{2}\right)\right],$$

siendo N el número de divisiones del espacio físico, con h la distancia entre dos puntos consecutivos y L la longitud del dominio de integración (en este trabajo se utiliza $L = 64$ y $N = 256$):

4. RESULTADOS NUMÉRICOS

Para evaluar el esquema numérico se presentan en esta sección resultados obtenidos para la ecuación (1) con condición inicial consistente en una onda polarizada a izquierda la cual satisface las condiciones de periodicidad y tiene la siguiente forma

$$\phi(z,0) = \phi_0 = A_0\,\mathrm{e}^{\mathrm{i}k_0 z} = A_0\left[\cos(k_0 z) + \mathrm{i}\sin(k_0 z)\right]. \tag{5}$$

En este trabajo se considera que el vector de onda es constante en todas las simulaciones con $k = 1.08$ y se utilizan distintos valores de amplitud inicial tal que $0.04 \le A_0 \le 2$.

El criterio de estabilidad modular de la DNLS cuando existen condiciones de contorno periódicas. Se establece siguiendo el desarrollo de Fla [11] que consiste en suponer que la solución base $A_0\,\mathrm{e}^{\mathrm{i}(k_0 z - \omega t)}$ es perturbada levemente por una función $\varepsilon(z,t)$ que cumple las condiciones de periodicidad y cuya transformada de Fourier es

$$\varepsilon(z,t) = \sum_{j=-\infty}^{\infty} \alpha_j(t)\,\mathrm{e}^{\mathrm{i}\mu_j z}; \qquad\qquad \alpha_j(t) = \alpha_j(0)\,\mathrm{e}^{\lambda_j t}, \tag{6}$$

donde $\mu_j = 2\pi j / L$ son los números de onda, entonces considerando la ecuación (1)

$$\lambda_j = \mathrm{i}\,2\left(A^2 - k_0\right)\mu_j \pm \left|\mu_j\right|\sqrt{\left(2k_0 - A^2\right)A^2 - \mu_j^{\,2}}\,. \tag{7}$$

La ecuación (7) muestra que la onda de amplitud constante es inestable para $2k_0 > A^2$ y marginalmente estable para $2k_0 < A^2$, por lo tanto para $k_0 = 1.08$ existirá inestabilidad modular para valores de amplitud que satisfagan $A < 1.470$.

Para evaluar el tiempo en que se produce la inestabilidad, se define define el parámetro E_{k0}, el cual representa la relación entre la energía transportada por la onda inicial y la energía total del sistema. Cuando la onda inicial pierde el 0.1% de su energía inicial se asume que comienza la inestabilidad, con ese criterio se ha confeccionado la Tabla 1, donde se especifica el tiempo de inestabilidad (T.I.) en función de la amplitud de la onda inicial. Se observa que el tiempo en el que se produce la inestabilidad no es constante y depende fuertemente de la amplitud de la onda inicial. También se deduce de este análisis que para valores de amplitud inicial cada vez más pequeños el tiempo necesario para la inestabilidad crece cada vez más.

A	0.08	0.1	0.13	0.14	0.15	0.2	0.4	0.6	0.8	1	1.2	1.4	1.45	1.46	1.7	2.0
T.I.	4217	2534	1695	1519	1378	663	73	40	24	19	22	52	410	631	–	–

Tabla 1: Tiempo de Inestabilidad para diferentes amplitudes de onda A

El otro aspecto que desea estudiarse en este trabajo es la evolución de la onda luego de alcanzada la inestabilidad. En la Figura 1 se grafica la evolución del parámetro E_{k0} para una onda inicial de amplitud $A = 0.1$. Se observa que se produce una evolución cuasi periódica de transferencia de energía, donde el intervalo de tiempo entre pico y pico decrece lentamente al avanzar el tiempo. En la Figura 2 se muestra la energía de la transformada de Fourier para tres tiempos distintos, donde se destaca que claramente que la energía es transferida fundamentalmente entre la onda original o madre y dos ondas hijas, verificando la relación $2k_0 = k_1 + k_2$. Para valores de $A = 0.08, 0.13, 0.14$ los resultados fueron similares.

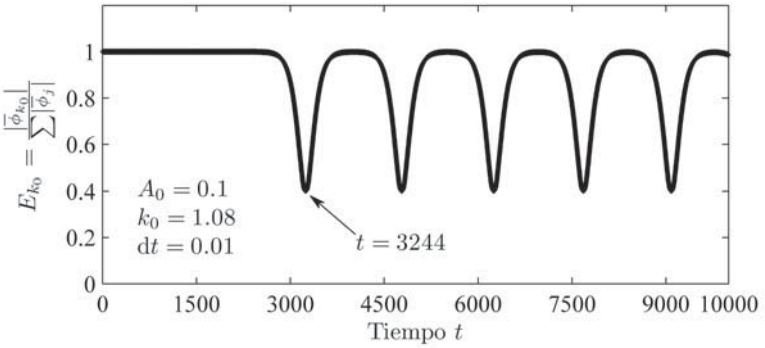


Figura 1: Evolución de la energía de la onda inicial en función del tiempo

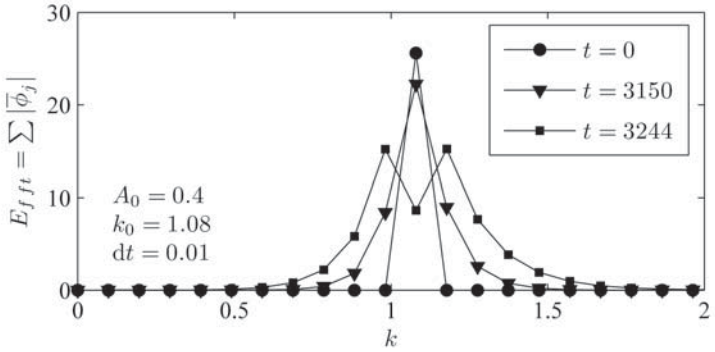


Figura 2: Energía de la Transformada de Fourier

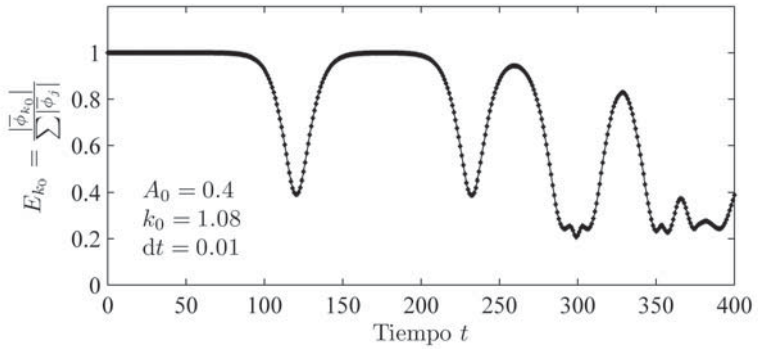


Figura 3: Evolución de la energía de la onda inicial en función del tiempo

5. CONCLUSIONES

Se ha solucionado numéricamente la ecuación DNLS sin efectos difusivos con condición inicial de una onda polarizada a izquierda, utilizando métodos espectrales para las derivadas espaciales y un esquema de Runge-Kutta de cuarto orden para la integración en el tiempo. El objetivo del trabajo fue la determinación del tiempo de inestabilidad modular de la solución y la forma de la evolución una vez que la inestabilidad ocurre. El método desarrollado ha demostrado una buena correlación entre los resultados obtenidos y los evaluados analíticamente por otros autores, satisfaciendo las condiciones analíticas de estabilidad modular, encontrándose además que el tiempo para el cual sucede la inestabilidad y la evolución posterior del sistema dependen de forma no lineal de la amplitud de la onda inicial, siendo ésta una característica no predicha por el análisis analítico. Por otro lado también se observó que la evolución del sistema luego de la inestabilidad da origen fundamentalmente a un intercambio de energía entre tres ondas: la onda inicial y dos ondas hijas que satisfacen la relación $2k_0 = k_1 + k_2$.

REFERENCIAS

[1] J. SANMARTÍN, R. ESTES, *Alfvén Wave Far Field from Steady-Current Tethers,* Journal of Geophysics Research, 102-A7(14): 625, 1997.

[2] A. ROGISTER, *Parallel Propagation of Nonlinear Low-Frequency Waves in High-β Plasma*, Physics of Fluids, 1971

[3] V. BELASHOV, S. VLADIMIROV, *Solitary Waves in Dispersive Complex Media*, Springer, Berling, 2005

[4] E. MJOLHUS, T. HADA, *Nonlinear Waves and Chaos in Space Plasmas*, edited by T. Hada and H. Matsumoto (Terrapub, Tokio), 121-169, 1997

[5] S. SPLANGER, J. SHEERIN, G. PAYNE, *A Numeric Study of Nonlinear Alfvén Waves and Solitons*, Physics of Fluids, 28: 104-109, 1985

[6] S. DAWSON. C. FONTA, *Soliton Decay of Nonlinear Alfvén Waves: Numerical Studies*, Physics of Fluids, 31(1): 83-89, 1988

[7] T. HADA, C. KENNEL, B. BUTI, E. MJOLHUS, *Chaos in Driven Alfvén Systems*, Physics of Fluids B2(11): 2581-2590, 1990

[8] J. SANMARTÍN, O. LÓPEZ-REBOLLAL, E. DEL RÍO, S. ELASKAR, *Hard Transition to Chaotic Dynamics in Alfvén Wave-Fronts*, Physics of Plasmas, 11(5): 2026-2035, 2004.

[9] S. ELASKAR, G. SÁNCHEZ-ARRIAGA, J. SANMARTÍN, *Chaos in Nonlinear Alfvén Waves Using the DNLS Equation,* International Symposioum on Electrohydrodynamics – 2006 ISEHD. Buenos Aires, 4-6 iciembre, Publicado en los Proceeding del Congreso, pp. 131-134 (ISBN 950-29-0964-X), 2006

[10] F. BACCELLI, G. COHEN, G.J. OLSDER, AND J-P. QUADRAT, *Synchronization and linearity. An algebra for discrete event systems*, Wiley and Sons, 1992.

[11] T. FLA, *A Numerical Conserving Method for the DNLS Equation*, Journal of Computacional Physics, 101: 71-79, 1992

[12] L. TREFETHEN, *Spectral Methods in MATLAB,* SIAM, Philadelphia, 2005

NEW RPD FUNCTION FOR TYPE-I INTERMITTENCY

Sergio Elaskar[a], Ezequiel del Rio[b] and José Donoso[b]

[a] *Departamento de Aeronáutica, Universidad Nacional de Córdoba and CONICET, Av. Vélez Sarfield 1611, Córdoba, Argentina, selaskar@efn.uncor.edu*

[b] *Escuela Técnica Superior de Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, Plaza Cardenal Cisneros 3, Madrid, España, ezequiel.delrio@upm.es*

Abstract: We apply a new methodology to evaluate the reinjection probability function for type-I intermittency. In previous papers we introduced this methodology for type-II and type-III intermittencies. For type-II intermittency we presented a new one-parameter family of functions describing the reinjection probability, being the usual type-II uniform reinjection probability a particular case of our RPD [4]. For the type-III case, a new two-parameter family of RPD has been found from which one can derive the lower bound of reinjection (LBR) [5] and [8]. By extending the preceding analysis of type-II and type-III intermittencies, we give here a new RPD for the type-I case, from which we also derive the densities of the laminar phase lengths and the new characteristic relations.

Keywords: intermittency, chaos, reinjection

1 INTRODUCTION

Intermittency is a particular form of deterministic chaos, in which transition between laminar and chaotic phases occurs. A system is in regular behavior until, with a small change in a parameter, it begins to show chaotic burst at irregular intervals. Pomeau and Maneville introduced the intermittency concept in relation to the Lorenz system [1,2,3]. It is classified into three types: I, II and III, according to the Floquet multipliers or eigenvalue in the local Poincaré map. For continuous-time systems, the type-I intermittency arises in a cyclic-fold bifurcation, for which a stable and an unstable orbits collapse, therefore, the system loses the stable orbits in the vicinity of the vanished periodic orbits. For some maps, type-I intermittency occurs by means of an inverse tangent bifurcation, in this case an eigenvalue leaves the unit circle through +1. Intermittency type-II begins in a subcritical Hopf bifurcation, so that, two complex-conjugate Floquet multipliers or two complex-conjugate eigenvalues of the local Poincaré map exit the unit circle. Intermittency of type-III is related to a subcritical period-doubling or flip bifurcation and one Floquet multiplier leaves the unit circle through -1.

In some previous papers, we have presented a new methodology to evaluate the main defining properties for type-II and type-III intermittencies, such as the reinjection probability density function (RPD), the probability density of the laminar phase, the average laminar length and the characteristic relation [4,5,6,7,8,9,10]. In this work we extend the new methodology to type-I intermittency. The local Poincaré maps for type-I intermittency is written as $x_{n+1} = \varepsilon + x_n + a\,x_n^2$, with $a > 0$ for which the intermittency phenomenon exists only for $\varepsilon > 0$ [11,12,13,14].

It is clear that the reinjection probability density $\phi(x)$, accounting with the transition from chaotic burst into the laminar zone, depends on each particular system or map making ϕ to be governed by the chaotic behavior of the system itself. The local Poincaré map of the intermittency does not give the necessary information to determine the reinjection probability density. In general, it is very difficult to obtain $\phi(x)$ analytically and it is also very complicated to set experimentally or numerically, because the large number of data needed to cover each interval of length δx in the reinjection region due to the noise introduced in numerical evaluations or in experimental measurements. Because of this, different approaches have been used in the literature to study the intermittent systems. The most usual and simple approximation considers $\phi(x)$ as a uniform function, not depending on the reinjection point [12]. Due to the disparity observed in modeling ϕ, we can conclude that it is very important to provide a method to obtain a correct form for the RPD for each different map.

2 REINJECTION PROBABILITY DISTRIBUTION

We do not directly measure the reinjection probability density $\phi(x)$ from the numerical data, instead of this, we numerically compute the function $M(x)$, defined as

$$M(x) = \frac{\int_{x_i}^{x} \tau\, \phi(\tau)\, d\tau}{\int_{x_i}^{x} \phi(\tau)\, d\tau} \quad if \quad \phi(\tau) \neq 0; \qquad M(x) = 0 \quad if \quad \phi(\tau) = 0\,, \tag{1}$$

where x_i is the closed point to the unstable fixed point where the reinjection takes place, *i.e.* it is the lower bound of the reinjection. The integration interval $[x_i\ ,\ x]$ defines the laminar region. $M(x)$ has been calculated for a broad class of maps numerically, and it has been stated that if exhibits the linear form $M(x) = m\,x + x_h$ as a good approximation [4,9]. From Eq.(1) it possible to determine that $M(x_i) = x_i$, then

$$M(x) = m\left(x - x_i\right) + x_i\,, \tag{2}$$

where the slope m plays an important role in the intermittency dynamics. Therefore, the function $M(x)$ has been proved to be a useful tool to study type-II and type-III intermittencies. From Eqs.(1) and (2) the reinjection probability density can be deduced, giving

$$\phi(x) = \Lambda\left(x - x_i\right)^{\alpha} \quad with \quad \alpha = \frac{1-2m}{m-1}, \qquad \Lambda = \frac{\alpha+1}{\left(c - x_i\right)^{\alpha+1}} = \frac{m}{1-m}\left(c - x_i\right)^{m/(m-1)}\,, \tag{3}$$

where Λ is the normalization constant and c is the upper limit of the laminar interval. Note that the slope m must satisfy the condition $0 < m < 1$ which has been met in all our numerical tests. The usual uniform probability reinjection is recovered for $m = ½$ with $x_i = 0$, leading to $M(x) = 0.5\,x$.

3 APPLICATION TO INTEMITTENCY TYPE-I

The new technique is now applied to study the type-I intermittency using the illustrating map

$$x_{n+1} = \varepsilon + x_n + a\,x_n^2 \quad if \quad x_n \le x_l \qquad x_{n+1} = \left(\frac{x - x_l}{1 - x_l}\right)^{s} \quad if \quad x_n \le x_l\,, \tag{4}$$

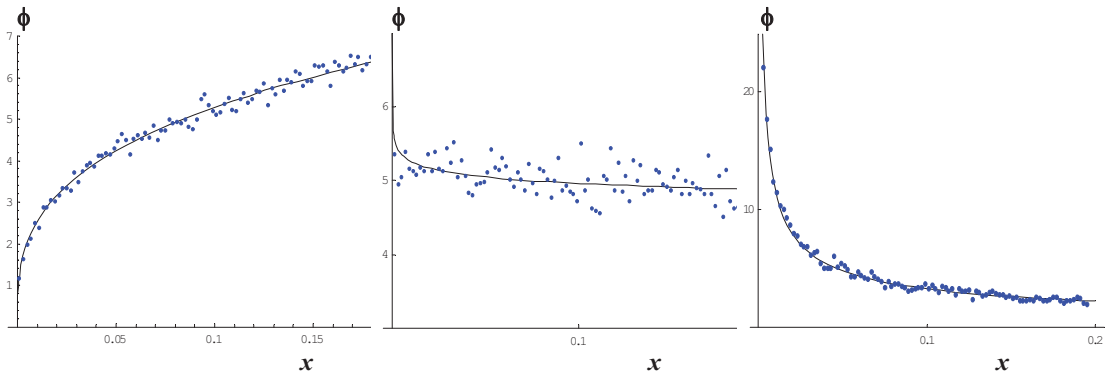
where x_l is such that $\varepsilon + x_l + a\,x_l^2 = 1$. For $\varepsilon = 0$ the origin is a fixed point, however, for $\varepsilon > 0$ all points x close to the origin move away in a process driven by the parameters ε and a. When the n-th iterated value x_n approaches x_l the reinjection mechanism starts, governed by exponent s.

3.1 REINJECTION PROBABILITY DENSITY FUNCTION

In this section we compute the RPD by using $M(x)$ computed after having carried out several numerical tests. The results are straight lines, crossing the origin, $x_i = 0$ in Eq.(2). After applying the least square method, we have obtained the corresponding m values of the $M(x)$ slope and each exponent α appearing in Eq.(3):

$s = 0.75$,	$m = 0.5686$	$\alpha = 0.318$
$s = 1.0$,	$m = 0.4936 \approx 0.5$	$\alpha = -0.02516 \approx 0$
$s = 2.0$,	$m = 0.3104$	$\alpha = -0.55$

The comparison between the RPD obtained numerically with the analytical RPD calculated by means of Eq.(3) is depicted in Figures 1a, b and c, where dots stand for the numerical results and the solid lines correspond to analytical expressions. In these figures it can be checked out how the theoretical RDP properly assembles the numerical RPD for the three test cases, each one having a characteristic distinguishable non-linear (global) behavior, in particular, the RPD is approximately constant in Fig 1b since α is close to zero.



Figures 1a,b,c: The RPD as a function of x for $s = 0.75$, $s = 1$ and s = 2. With ε = 0.00001, $c = 0.2$. Dots stand for numerical results and the solid line plots the function given by Eq.(3)

3.2 PROBABILITY OF THE LAMINAR LENGTH

Following the usual method based on interpretation transposing the map local difference equation into a continuous differential equation inside the laminar region (Shuster and Just, 2005), for type-I intermittency, Eq.(1), we have $dx/dl = \varepsilon + a\,x^2$, where l counts the number of iterations in the laminar region. After integration we have $l(x,c) = \left[\arctan\left(c\sqrt{a/\varepsilon}\right) - \arctan\left(x\sqrt{a/\varepsilon}\right)\right]/\sqrt{a\varepsilon}$, which clearly evidences that the laminar iteration number (length of the laminar region) only depends on the local map but not on the global one. The probability of finding a laminar phase length inside the interval (l; l+dl), $\phi_l(l)$, is given by $\phi_l(l) = \phi(X(l,c))\left|dX(l,c)/dl\right|$, where $X(l,c)$ is the inverse of the $l(x,c)$. The required probability is

$$\phi_l(l) = \frac{\varepsilon^{1+\alpha/2}}{a^{\alpha/2}}\Lambda\ \sec^2(z)\tan^{\alpha}(z)\,; \qquad z = \arctan\left(c\frac{\sqrt{a}}{\sqrt{\varepsilon}}\right) - l\sqrt{a\varepsilon}\,, \tag{5}$$

which for when $l \to 0$, behaves as $\lim_{l\to 0}\phi_l(l) \to \Lambda c^{\alpha}\left(\varepsilon + a c^2\right)$. This last equation indicates that for a very small ε, $\phi_l(0)$ is approximately constant and independent of ε: $\phi_l(0) \approx a c^{2+\alpha}\Lambda$. For any positive α, the function $\phi_l(l)$ is a decreasing function of l, being $\phi_l(l) = 0$ when l equals the value $l_m = \arctan\left(c\sqrt{a/\varepsilon}\right)/\sqrt{a\varepsilon}$ and for $\alpha < 0$ the probability of the laminar length satisfies $\lim_{l\to l_m}\phi_l(l) \to \infty$, meaning that for negative α values, the laminar length $l = l_m$ is a cut-off. Having in mind the previous relations, we can conclude that there always exists a limit value l_m for l, meanwhile the behavior of $\phi_l(l)$ depends on the sign of α since for $\alpha \le 0$, $\lim_{l\to l_m}\phi_l(l) \to \infty$ and for $\alpha > 0$ $\lim_{l\to l_m}\phi_l(l) \to 0$.

In general, the behavior of $\phi_l(l)$ has two relevant cases. For $\alpha \neq 0$ two factors govern Eq.(15), $\sec^2(z)$ and $\tan^{\alpha}(z)$, the former is always positive whereas $\tan^{\alpha}(0) = 0$, furthermore, for $z = 0$, a limit value $l = l_m$, exists for $\alpha < 0$ giving $\lim_{l\to l_m}\phi_l(l) \to \infty$ and for $\alpha > 0$, $\lim_{l\to l_m}\phi_l(l) \to 0$. For $\alpha = 0$, the factor $\tan(z)$ disappears and $\phi_l(l)$ is only depending on the factor $\sec^2(z)$, *i.e.* $\lim_{l\to l_m}\phi_l(l) \to \infty$, however in this case $l = l_m$ when z = π/2: $l_m = \left[\arctan\left(c\sqrt{a/\varepsilon}\right) - 0.5\pi\right]/\sqrt{a\varepsilon}$, which satisfies $\lim_{\varepsilon\to 0} l_m \to \infty$.

4 CONCLUSIONS

In this paper we have extended to the type-I intermittency phenomenon the analysis procedure we developed in a previous work in studying type-II and type-III intermittencies. We have found that our function $M(x)$ is also a key tool to analyze the type-I intermittency, especially when numerical or experimental data are required in the investigation, since it is easily obtained. Therefore, $M(x)$ is more useful and simpler than the reinjection probability density ϕ which can be derived from the former. As a matter of fact, the reinjection probability density function and the probability of the laminar length have been obtained finding a good agreement with theoretical predictions. In all numerical tests we have obtained that $M(x)$ is linear, $M(x) = m\,x$, and we have found a power law for the RPD as $\phi(x) = \lambda\, x^{\alpha}$ which extended the usual uniform RPD, a particular case of ours for $\alpha = 0$ or $m = ½$.

ACKNOWLEDGEMENTS

This paper was supported by grants PIP - No 11220090100809 of CONICET, Universidad Politécnica de Madrid, Universidad Nacional de Córdoba, Ministerio de Ciencia y Tecnología de Córdoba and Ministerio de Ciencia de España under projects ENE2007-67406-C02-01, AYA2008-04769.

5 REFERENCES

[1] P. MANNEVILLE, Y. POMEAU, Intermittency and the Lorenz model, *Physical Letters A*, 75: 1-2, 1979.

[2] Y. POMEAU, P. MANNEVILLE., *Intermittent transition to turbulence in dissipative dynamical systems*, Communications in Mathematical Physics, 74: 189-197, 1980.

[3] P. MANNEVILLE, *Intermittency, self-similarity and 1/f spectrum in dissipative dynamical systems,* Le Journal de Physique, 41 (11): 1235-1243, 1980.

[4] E. DEL RIO, S. ELASKAR, *New characteristic relation for intermittency type II*, International Journal of Bifurcation and Chaos, 20 (4): 1185–1191, 2010.

[5] E. DEL RIO, S. ELASKAR, *Characteristic Relations and Reinjection Probability Densities (RPD) of Type-II and III Intermittencies. Effect of the noise on RPD. 8th AIMS Conference on Dynamical Systems, Differential Equations and Applications*. University of Technology Dresden , Dresden, Germany, May 25 - 28, 2010a.

[6] E. DEL RIO, S. ELASKAR, J. DONOSO, L. CONDE, *Noise influence on the Characteristic Relations and Reinjection Probability Densities of type-II and type-III Intermittencies*. *3rd Chaos 2010 International Conference*, Chania Crete Greece, 1 - 4 June 2010.

[8] S. ELASKAR, E. DEL RIO, *Reinjection Probability Function with Lower Bound of the Reinjection for Intermittency Type III*. Mecánica Computacional, 28: 1463-1476, 2009.

[9] S. ELASKAR, E. DEL RIO, J. DONOSO, *Reinjection probability density in type-III intermittency*, Physica A. Submitted, 2010.

[10] S. ELASKAR, E. DEL RIO, J. DONOSO, *Studies for type-I, type-II and Type-III intermittencies*, CILAMCE 2010, XXXI Congreso Ibero-Latino-Americano de Métodos Computacionales en la Ingeniería, Bs. As., November. Published in Mecánica Computacional, 29: 3389-3406, 2010.

[11] N. RUSBAND, *Chaotic Dynamics of Nonlinear Systems*, John Wiley & Sons, New York, USA, 1990.

[12] H. SCHUSTER, W. JUST, *Deterministic Chaos. An Introduction*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2005.

[13] CH. KIM, G. KIM, G. KIM, H. LEE, *Experimental evidence of characteristic relations of type-I intermittency in an electronic circuit*, Physical Review E, 56 (3): 2573-2577, 1997.

[14] M. KIM, H. LEE, K. CHIL-MIN; P. HYUN-SOO; L. EOK-KYUN, O. KNOWN, *New characteristic relations in Type II and III intermittency,* International Journal of Bifurcation and Chaos, 7 (4): 831-836, 1997a.

[15] I. PRIGOGINE, *Las leyes del caos*. Ed. Drakonitos, Barcelona, Spain, 2009.

NONLINEARIZED FOURIER APPROACH AND COHERENCE. APPLICATIONS TO SHOCK WAVE - TURBULENCE INTERACTION

Liviu Florin Dinu[♭] and Marina Ileana Dinu[†]

[♭]*Romanian Academy, Institute of Mathematics, P.O.Box 1-764, Bucharest, RO-014700, Romania,, liviu.dinu@imar.ro, lfdinu2@gmail.com*
[†]*Polytechnical University, Bucharest, Romania, marinadinu@gmail.com*

Abstract: In a *minimally nonlinear* context [Lax−Majda] we *construct* a solution [analytic, closed, optimal, admissible (entropy), highly nontrivial] associated to the shock-turbulence interaction. ● In order to complete such a construction we have to [naturally] identify two hierarchies [of *partitions* and, respectively, of *factorizations*] and to essentially notice a *gasdynamic coherence* between them. ● The constructed solution is finally used into a *classifying* approach making evidence of a *critical* ["pseudo relativistic"] separation inside the interaction solution. ● The details of this "pseudo relativistic" separation are finally compared with the criticity arguments considered in recent fundamental numerical studies on the shock-turbulence interaction.

Keywords: *Hyperbolic systems of conservation laws, multidimensional Fourier-Snell analysis, gasdynamic interaction.*
2000 AMS Subject Classification: 35L65, 35L99, 35Q35, 76N10, 76N15.

1 INTRODUCTION

The present paper considers, in a linearized context, the interaction between two gasdynamic objects: a *turbulence model* and, respectively, a *planar shock discontinuity* ([2], [3]). ● The *linearized* (with shock) context assumes a *minimal* nonlinearity in the form of a *nonlinear subconscious* (in the sense of P.D. Lax and A. Majda; see [7]). It considers a *linearized* problem: a *linear* problem with a *nonlinear* subconscious. A nonlinear subconscious results when the nonlinearity is allowed only at the zeroth order of a perturbation expansion: a piecewise constant admissible solution [with shock; zeroth order] is perturbed; one linearizes and proves that the zeroth order requirement of admissibility is still active at the first order and essentially structures the linearized description. ● Consequently, and essentially, the interaction solution is constructed as an *admissible (entropy)* solution. ● The resultant perturbation is regarded as a solution ("interaction solution") of such a *linearized* problem.

In the present paper the incident turbulence is considered to have a *vorticity* nature and is modelled, using the *linearized* context, by a nonstatistical/noncorrelative *superposition* of some finite (or point core) planar vortices. ● The turbulence − planar shock interaction is associated with a class of interaction elements. An interaction element formally models the interaction between a planar shock and a *single* incident vortex corresponding to a certain inclination of the vortex axis with respect to the shock. ● Modelling the incident turbulence by a superposition of compressible planar vortices appears to correspond to a *first level* of decomposition. In J.M. Lighthill's fundamental paper [5] the turbulence is acoustically modelled by a distribution of quadrupoles − which is equivalent with a "weighted" distribution of point vortices. ● Next, in order to proceed, each incident vortex is Fourier decomposed into planar monochromatic waves − a *second level* of decomposition. ● Finally, each incident planar monochromatic wave is Snell passed [optically refracted] through the shock discontinuity (see [1], [2] for the details of the admissibility). ● The composition of the mentioned levels leads to a Fourier−Snell representation of the interaction solution. ● The main point of the analysis is that the result of the passage through the shock can again be presented by two levels of *recombination* so that each incident level of decomposition has a correspondent in the emergent solution.

A Fourier−Snell representation of the linearized interaction between a planar shock discontinuity and a planar compressible finite-core vortex whose axis is *parallel* to the shock, we call it a *parallel interaction*, has been considered and exploited numerically by H.S. Ribner (see [8]).

2 THE MAIN RESULTS

The present analysis has essentially two objectives: (*a*) finding an *analytic, closed, and optimal* form for the interaction solution associated to an incident superposition of vortices (eventually *oblique* with respect to the shock), and (*b*) offering an *exhaustively classifying characterization* of this mentioned solution (Figure 1).

Realizing the objective (*a*) is connected with: • considering a *singular limit* of the interaction solution, • considering a *hierarchy of (natural) partitions* of the singular limit, • identifying a sequence of (natural) *gasdynamic factorizations* and • noticing a *compatibility* between partitions and factorizations (indicating a gasdynamic *inner coherence*), • *predicting some exact details* of the interaction solution, • indicating some parasite singularities [= strictly depending on the method] to be compensated [= pseudosingularities], and • *re-weighting* the singular limit of the interaction solution into an analytic, closed, and optimal form.

Realizing the objective (*b*) is connected with finding some *Lorentz arguments of criticity*. The interaction solution appears essentially to include a *subcritical* and respectively a *supercritical* contribution distinguished by differences of a *"pseudo relativistic"* nature (Figure1). Precisely: in the singular limit of the interaction solution the emergent sound is *singular* in the subcritical contribution and it is *regular* in the supercritical contribution. This "pseudo relativistic" discontinuity in the nature of the emergent sound, corresponding to the singular limit of the interaction solution, appears to be dissembled (hidden) in the re-weighted interaction solution.

Remark 1. The present paper begins with the construction of the *emergent sound* in the parallel interaction solution. The other modal emergent contributions of this solution could be expressed in terms of the emergent sound via the equations in the adjacent regions of the shock and the Rankine−Hugoniot jump relations ([3]). □

The analytic, closed, optimal form of the parallel interaction solution appears to be *extensible* to the oblique cases [where the axis of the incident vortex is oblique with respect to the shock]. • For a subcritical extension see [3]. • The details of a supercritical extension are in final progress.

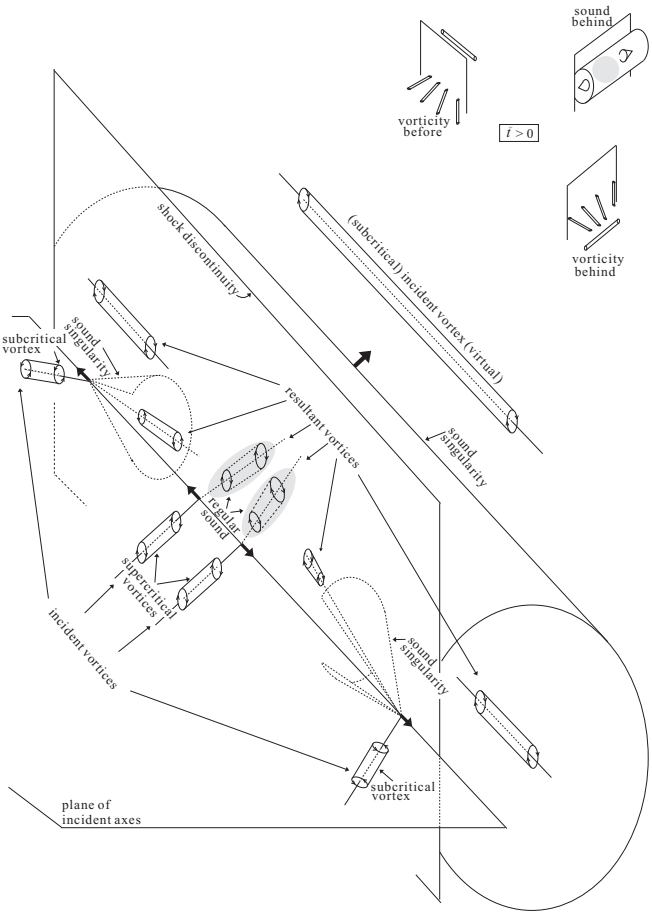


Figure 1 The interaction solution: a classifying structure

3 SOME CONSTRUCTIVE DETAILS

3.1 THE CASE OF A PARALLEL INTERACTION

At its second level of decomposition the parallel interaction solution is constructively made of polymodal elementary structures (p.e.s.). A p.e.s. puts together, in an optimal manner, some monochromatic waves [here planar] related by an optical [Snell] refraction [their phases] and by the Rankine−Hugoniot jump relations [their amplitudes]. It can be shown ([1], [2]) that only four p.e.s. are *admissible (entropy)*

$$\mathcal{V}_{li}\mathcal{S}^{+}_{rd}\mathcal{V}_{rd},\ \ \mathcal{S}^{+}_{li}\mathcal{S}^{+}_{rd}\mathcal{V}_{rd},\ \ \mathcal{S}^{-}_{li}\mathcal{S}^{-}_{rd}\mathcal{V}_{rd},\ \ \mathcal{S}^{-}_{ri}\mathcal{S}^{+}_{rd}\mathcal{V}_{rd}. \tag{1}$$

Since the incident turbulence has a vorticity character, the construction uses the p.e.s. $(1)_1$ only, because of their incident vorticity [$(1)_1$ is the first element in the sequence (1); $\mathcal{V}/\mathcal{S}$ mean vorticity/sound nature; *l,r* mean left/right to the shock; *i,d* mean incident to / divergent from the shock; $-/+$ mean backward/forward]. • There exist two types of p.e.s. $(1)_1$: which respectively include a *real* or a *strictly complex* phase in their emergent sound contributions. • We notice that a real / strictly complex phase of the emergent sound is

associated with a subcritical / supercritical inclination of the vorticity incidence. The two mentioned types of p.e.s. $(1)_1$ will be therefore said to be respectively subcritical or supercritical. Their incident vorticity inclinations are separated by a *critical* inclination denoted $\mathfrak{z}_c$.

We only consider the details of the *emergent sound* in the interaction solution [cf. Remark 1].

The subcritical / supercritical contributions [unions of subcritical / supercritical p.e.s. $(1)_1$] in the emergent sound will be said concurrently to be pseudo hyperbolic / pseudo elliptic [label *h/e*]. • A first representation of the emergent sound in the interaction solution has the form (2): associated to a first partition. • This partition persists in the singular limit of the emergent sound [the limit $r_*\to 0$ with r_* the radius of the incident vortex core]:

$$[\widetilde{p},\widetilde{u},\widetilde{v}]=[\widetilde{p}_h,\widetilde{u}_h,\widetilde{v}_h]+[\widetilde{p}_e,\widetilde{u}_e,\widetilde{v}_e]. \tag{2}$$

The representation (2) of the emergent sound concurrently includes some other *significant* implicit details: associable for example with the formal interaction shock-vorticity [label *int*] or with the shape of the incident vortex [label *vs*]. We approach the optimal form of the emergent sound representation by following a significant sequence of partitions. One of them moves (2) into

$$[\widetilde{p}_h,\widetilde{u}_h,\widetilde{v}_h]+[\widetilde{p}_e,\widetilde{u}_e,\widetilde{v}_e]=[\widetilde{p}_{vs},\widetilde{u}_{vs},\widetilde{v}_{vs}]+[\widetilde{p}_{int},\widetilde{u}_{int},\widetilde{v}_{int}] \tag{3}$$

making evidence of the details just mentioned. The passage from (2) to (3) appears to be "stored" by a *factorizable* "memory" (4)

$$(\xi^2+\eta^2+\zeta_i)^2-4\xi^2\zeta_i=\frac{1}{(x^2+y^2)^2}\left[\left(\mathfrak{z}_c t-x\sqrt{\mathfrak{z}_c^2-\zeta_i}\right)^2-\zeta_i y^2\right]\left[\left(\mathfrak{z}_c t+x\sqrt{\mathfrak{z}_c^2-\zeta_i}\right)^2-\zeta_i y^2\right] \tag{4}$$

where we use the *Lorentz type coordinates* x,y,t

$$x=\frac{\widetilde{x}+M\tilde{t}}{\sqrt{1-M^2}}=\frac{X}{\sqrt{1-M^2}},\quad y=\widetilde{y},\quad t=\frac{\tilde{t}+M\widetilde{x}}{\sqrt{1-M^2}}; \tag{5}$$

with $\widetilde{x}=X-MT,\ \widetilde{y}=Y,\ \tilde{t}=T$; X,Y is a frame fixed on the shock, $\widetilde{x},\widetilde{y}$ is a Lagrangian frame fixed on the zeroth order flow after the shock, M is the Mach number of this flow and ζ_i are some ratios associated with the passage through the shock [ζ_1 has an entropy-vorticity nature, and $\zeta_{2,3}$ have a sound nature]; see [2], [3].

We also use the *Lorentz type entities*

$$\xi=\frac{\mathfrak{z}_c t y}{x^2+y^2}\ ,\ \eta=\frac{\mathfrak{z}_c x\sqrt{t^2-x^2-y^2}}{x^2+y^2};\qquad t^2-x^2+y^2=\tilde{t}^2-\widetilde{x}^2+\widetilde{y}^2 \tag{6}$$

and notice the presence, in the integrals corresponding to the *vs*-contribution, of a part which is *singular*, concurently with η^{-1}. This circumstance naturally offers a last element in the sequence (2), (3), (7) (the labels r/s mean *regular/singular*)

$$[\widetilde{p}_{vs},\widetilde{u}_{vs},\widetilde{v}_{vs}]+[\widetilde{p}_{int},\widetilde{u}_{int},\widetilde{v}_{int}]=[\widetilde{p}_r,\widetilde{u}_r,\widetilde{v}_r]+[\widetilde{p}_s,\widetilde{u}_s,\widetilde{v}_s]. \tag{7}$$

On carrying and re-arranging the calculations (3) into the last term of (7), we are able to [naturally] identify a set of *gasdynamic factorizations*, mutually compatible and compatible with (4). This indicates a *gasdynamic coherence*. In order to exemplify this coherence we present the *prefinal* details of the terms p_r,p_s

$$\widetilde{p}_r(x,y,t)=\frac{\pi y}{2(x^2+y^2)^2}\sum_{i=1}^{4}\frac{(2-i)(3-i)\sqrt{|\zeta_i|}\widetilde{k}_i(\zeta)}{(\xi^2+\eta^2+\zeta_i)^2-4\xi^2\zeta_i}\left[\sqrt{\mathfrak{z}_c^2-\zeta_i}\mathcal{E}_1^p(\zeta_i)-\mathcal{E}_2^p(\zeta_i)\right]\left[\left(\mathfrak{z}_c t+x\sqrt{\mathfrak{z}_c^2-\zeta_i}\right)^2-\zeta_i y^2\right] \tag{8}_p$$

$$\begin{aligned}\widetilde{p}_s(x,y,t)=&-\frac{y}{\sqrt{t^2-x^2-y^2}}\cdot\frac{\pi}{(x^2+y^2)^2}\sum_{i=1}^{4}\frac{\widetilde{k}_i(\zeta)}{(\xi^2+\eta^2+\zeta_i)^2-4\xi^2\zeta_i}\\&\cdot\left\{\left[\sqrt{\mathfrak{z}_c^2-\zeta_i}\mathcal{E}_1^p(\zeta_i)-\mathcal{E}_2^p(\zeta_i)\right]\left(t\sqrt{\mathfrak{z}_c^2-\zeta_i}-\mathfrak{z}_c x\right)\left[\left(\mathfrak{z}_c t+x\sqrt{\mathfrak{z}_c^2-\zeta_i}\right)^2-\zeta_i y^2\right]\right.\\&\left.+\left[\sqrt{\mathfrak{z}_c^2-\zeta_i}\mathcal{E}_1^p(\zeta_i)+\mathcal{E}_2^p(\zeta_i)\right]\left(t\sqrt{\mathfrak{z}_c^2-\zeta_i}+\mathfrak{z}_c x\right)\left[\left(\mathfrak{z}_c t-x\sqrt{\mathfrak{z}_c^2-\zeta_i}\right)^2-\zeta_i y^2\right]\right\}.\end{aligned} \tag{9}_p$$

Similar expressions result for u_r,u_s and v_r,v_s in (7).

A *final* form of the emergent sound results when the mentioned gasdynamic coherence [a factorial compatibility between the denominator (4) and numerators in $(8)_{p,u,v}$, $(9)_{p,u,v}$] is taken into account.

We have to notice that the singularities of this final form, distinct from those laid in $X>0$ along $\widetilde{t}^2-\widetilde{x}^2+\widetilde{y}^2=0$ or at $(\widetilde{x}=0,\widetilde{y}=0)$ are mutually compensated in the sum $[\widetilde{p}_r,\widetilde{u}_r,\widetilde{v}_r]+[\widetilde{p}_s,\widetilde{u}_s,\widetilde{v}_s]$: they are pseudo singularities [dependent on the method]. • The geometric elements corresponding to $\widetilde{t}^2-\widetilde{x}^2+\widetilde{y}^2=0$ or $(\widetilde{x}=0,\widetilde{y}=0)$ appear to be enveloped by the subcritical p.e.s. $(1)_1$.

The final form of the emergent sound in the parallel interaction solution can be put into an "encoded" form

$$\begin{cases} \widetilde{p}_r+\widetilde{p}_s \equiv \widetilde{p}_{\parallel}(x,y,t;\ \zeta_1,\zeta_2,\zeta_3,\zeta_4;\ \mathfrak{z}_c;\ Q_1,Q_2,Q_3) \\ \widetilde{u}_r+\widetilde{u}_s \equiv \widetilde{u}_{\parallel}(x,y,t;\ \zeta_1,\zeta_2,\zeta_3,\zeta_4;\ \mathfrak{z}_c;\ Q_1,Q_2,Q_3) \\ \widetilde{v}_r+\widetilde{v}_s \equiv \widetilde{v}_{\parallel}(x,y,t;\ \zeta_1,\zeta_2,\zeta_3,\zeta_4;\ \mathfrak{z}_c;\ Q_1,Q_2,Q_3) \end{cases} \tag{10}$$

where x,y,t are the Lorentz type coordinates (5) and Q_1,Q_2,Q_3 are coefficients associated to the shock-vorticity interaction. All the other modal contributions in the interaction solution can be then expressed directly in an analytic, closed, optimal form in terms of the emergent sound [Remark 1].

Finally, we have to *re-weight* the singular interaction solution associated to (10) into an analytic, closed and optimal form.

3.2 The case of an oblique interaction

The "encoded" form (10), corresponding at the *first level* of decomposition (cf. Introduction) to a parallel interaction element, appears to be *extensible* to the case of an oblique interaction [for which the axis of the incident vortex is oblique (subcritical/supercritical) with respect to the shock]. • Such an extension results when the *parallel* Lorentz type coordinates (5) are replaced in (10) by some *generalized* (subcritical, supercritical) Lorentz type coordinates. • For a subcritical extension see [3]. • The details of a supercritical extension are in final progress. • There is a "pseudo relativistic" contrast between the two mentioned types of interaction elements (subcritical/supercritical); cf. Figure 1, associated to the first level of decomposition. • The subcritical or, respectively, the supercritical nature associated with the *first level* of decompositon do not appear to be in a strict correspondence with the pseudo hyperbolic / pseudo elliptic nature associated with the *second level* of decompositon [see for example the partition (2) in the parallel (subcritical) case].

4 Final details

• The present interaction solution parallels and extends, from an *analytical* and *classifying* prospect, allowing *oblique* vortices, the Ribner *parallel* representation and computational approach ([8]).

• The structure of the present interaction solution is naturally associated, from a classifying prospect, to the Lighthill fundamental representation of the incident turbulence ([5]).

• The details of the "pseudo relativistic" separation mentioned above are compared with the criticity arguments considered in the recent fundamental numerical studies on the shock-turbulence interaction due to S.K. Lele ([4]) or K. Mahesh, S.K. Lele and P. Moin ([6]).

Acknowledgments. Support from Romanian Grant PN2, No.573, 2009.

References

[1] A.M. BLOKHIN, AND Y. TRAKHININ, *Stability of strong discontinuities in fluids and MHD,* Handbook of Fluid Dynamics, Vol. 1 (2002), pp.1−100.

[2] L.F. DINU, *Mathematical concepts in nonlinear gas dynamics,* Monographs and Surveys in Pure and Applied Mathematics, CRC Press, Boca Raton - London [forthcoming monograph].

[3] L.F. DINU, AND M.I. DINU, *Nonlinearized Fourier approach and gasdynamic coherence,* Communications in Mathematical Analysis [Washington DC], Vol.8 (2010) (Special Volume in Honor of Professor Peter D. Lax), No.3, pp.66-91.

[4] S.K. LELE, *Compressibility effects on turbulence,* Annual Rev. Fluid Mech., Vol. 26 (1994), pp.211−254.

[5] J.M. LIGHTHILL, *On sound generated aerodynamically 1. General theory,* Proc. Roy. Soc. London, Vol. A 211 (1952), pp.564−587.

[6] K. MAHESH, S.K. LELE, AND P. MOIN, *The influence of entropy fluctuations on the interaction of turbulence with a shock wave,* J. Fluid Mechanics, Vol.334 (1997) pp.353−379.

[7] A. MAJDA, *The stability of multidimensional shock fronts,* Memoirs AMS, Vol. 275 (1983); *The existence of multidimensional shock fronts,* Memoirs AMS, Vol. 281 (1983).

[8] H.S. RIBNER, *Cylindrical sound wave generated by shock−vortex parallel interaction,* A.I.A.A. Journal, Vol. 23 (1985) pp.1708−1715.

WAVELETS DEFINIDAS SOBRE GRILLAS TETRAÉDRICAS IRREGULARES. CÁLCULO DE LAS MATRICES DE ANÁLISIS Y SÍNTESIS

Liliana Boscardín[♭], Liliana Castro[†] y Silvia Castro[♭,†]

[♭]*Departamento de Matemática, Universidad Nacional del Sur, Avda Alem 1253, Bahía Blanca, Argentina, lboscar@uns.edu.ar*

[†]*Departamento de Matemática, Universidad Nacional del Sur, Avda Alem 1253, Bahía Blanca, Argentina, lcastro@uns.edu.ar*

[♭,†]*Departamento de Ciencias de la Computación, Universidad Nacional del Sur, Avda Alem 1253, Bahía Blanca, Argentina, smc@cs.uns.edu.ar*

Resumen:

Las wavelets definidas sobre grillas tetraédricas no anidadas permiten representar funciones definidas sobre una tetraedrización irregular dada y esto tiene como aplicación inmediata la representación de distintos atributos definidos sobre un objeto como pueden ser su color, su densidad, etc. Esta representación consiste en un conjunto de coeficientes correspondientes a una aproximación gruesa seguida por una sucesión de coeficientes de detalle que miden el error entre dos aproximaciones sucesivas. El análisis, es decir el paso de una resolución fina $k+1$ a una más gruesa k, proceso llamado *análisis*, es realizado a través de la matriz de *análisis* N^k. Para la *síntesis*, es decir para pasar de una resolución gruesa k a una más fina $k+1$, se utiliza la matriz de *síntesis* S^k. En este trabajo se definen wavelets sobre grillas tetraédricas no anidadas que permiten representar funciones constantes por tramos sobre una tetraedrización irregular dada. Por otro lado, para el caso de un operador de aproximación suryectivo cualquiera, se deduce una forma de hallar la matriz de análisis conociendo la de síntesis y viceversa.

Palabras clave: *wavelets, tetraedrizaciones irregulares, representación multirresolución, subdivisión anidada y no anidada.*

2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

El motivo básico por el cual es aplicable la teoría de wavelets en [1], [2], [3], [5], es que el método de subdivisión elegido, ya sea en el caso de grillas triangulares o tetraédricas, regulares o semiregulares, permite obtener mallas anidadas y esto a su vez permite definir los espacios anidados que son característicos en la teoría de wavelets. Ahora bien, cuando los datos están distribuidos de forma irregular, es conveniente tetraedrizar el dominio con una malla tetraédrica irregular y aplicar alguna de las técnicas de simplificación conocidas para esta clase de mallas. Sin embargo con estas técnicas no se obtienen mallas anidadas y por lo tanto no puede aplicarse la teoría clásica de wavelets. Se requiere entonces extender dicha teoría de modo tal que ya no sea necesario tener una sucesión de espacios anidados. En este trabajo presentamos un análisis multirresolución para datos definidos sobre redes tetraédricas arbitrarias. Este análisis multirresolución está definido sobre tetraedrizaciones jerárquicas que no poseen la propiedad de conectividad de subdivisión.

1.1. ANÁLISIS DE MULTIRRESOLUCIÓN CON ESPACIOS NO ANIDADOS.

Indicaremos con E a un espacio de Hilbert, que en la mayoría de las aplicaciones será el espacio funcional $L_2(\Omega)$, siendo Ω un dominio de $\mathbb{R}^3$.

Un análisis multirresolución de $L^2(\Omega)$ consiste en una sucesión de espacios V^k no anidados de dimensión finita, $dimV^k \leq dimV^{k+1}$, e isomorfos a un subespacio $\widetilde{V}^k \subset V^{k+1}$. Los espacios V^k son el dominio de los operadores de aproximación: $P^k : V^{k+1} \to V^k$; a partir de ellos se definen los espacios $W^k := NuP^k$.

1.2. BASES Y MATRICES DE ANÁLISIS Y DE SÍNTESIS.

Introduciremos ahora las bases y notaciones para los diferentes espacios.

$i)$ n_k la dimensión de V^k y (ϕ_i^k), $i = 1, ..., n_k$, una base de V^k;

ii) m_k la dimensión de $\widetilde{V}^k$ y $\widetilde{\phi}_i^k$, $i = 1, ..., m_k$, una base de $\widetilde{V}^k$;

iii) r_k la dimensión de W^k y (ψ_i^k), $i = 1, ..., r_k$, una base de W^k;

iv) $(a_\cdot^k)$, $(\widetilde{a}_\cdot^k)$ y $(b_\cdot^k)$ representarán, respectivamente, los coeficientes de las funciones $f_k \in V^k$, $\widetilde{f}_k \in \widetilde{V}^k$ y $g_k \in V^k$ en las bases indicadas en los incisos anteriores. Los coeficientes $(b_\cdot^k)$ también son llamados los coeficientes wavelets de la función f_k.

*Análisis: la matriz que realiza esta operación es por definición la *matriz de análisis* N^k. Podemos escribir entonces:

$$\begin{bmatrix} a_\cdot^k \\ b_\cdot^k \end{bmatrix} = N^k[a_\cdot^{k+1}] = \begin{bmatrix} P^k_{\phi_\cdot^{k+1},\phi_\cdot^k} \\ Q^k_{\phi_\cdot^{k+1},\phi_\cdot^k} \end{bmatrix} [a_\cdot^{k+1}], \tag{1}$$

donde N^k es una matriz $(n_k + r_k) \times n_{k+1}$ que resulta cuadrada si P^k es suryectivo.

*Reconstrucción: la matriz que realiza esta operación es la *matriz de síntesis* S^k que aparece en la siguiente ecuación:

$$[a_\cdot^{k+1}] = S^k \begin{bmatrix} a_\cdot^k \\ b_\cdot^k \end{bmatrix} = \begin{bmatrix} \Phi^k & \Psi^k \end{bmatrix} \begin{bmatrix} a_\cdot^k \\ b_\cdot^k \end{bmatrix}, \tag{2}$$

donde Ψ^k es la matriz cuyos vectores columna son las funciones $\psi_\cdot^k$ expresadas en la base $\phi_\cdot^{k+1}$ y Φ^k es la matriz cuyos vectores columna son los transformados por S^k de la base de V^k expresados en la base de V^{k+1}.

Cuando el operador de aproximación es suryectivo, las matrices de análisis y de síntesis son cuadradas e inversas una de la otra.

Notación: indicaremos con $N = \begin{bmatrix} P \\ Q \end{bmatrix}$ a la matriz de análisis y con $S = [\Phi \ \ \Psi]$ a la de síntesis. Consideraremos ahora el caso en que el operador de aproximación es suryectivo. Bajo esta hipótesis, el siguiente lema provee una forma de hallar las matrices P y Q conociendo Φ y Ψ y recíprocamente.

Lema 1 *.Sea P un operador de aproximación suryectivo y notemos con P^* a la matriz traspuesta de P. Entonces:*

1) $PP^* = [(\Phi^*\Phi) - (\Phi^*\Psi)(\Psi^*\Psi)^{-1}(\Psi^*\Phi)]^{-1}$.

2) $PQ^* = -[(\Phi^*\Phi) - (\Phi^*\Psi)(\Psi^*\Psi)^{-1}(\Psi^*\Phi)]^{-1}(\Phi^*\Psi)(\Psi^*\Psi)^{-1}$.

3) $QQ^* = [(\Psi^*\Psi) - (\Psi^*\Phi)(\Phi^*\Phi)^{-1}(\Phi^*\Psi)]^{-1}$.

4) $P = [(\Phi^*\Phi) - (\Phi^*\Psi)(\Psi^*\Psi)^{-1}(\Psi^*\Phi)]^{-1}[\Phi^* - (\Phi^*\Psi)(\Psi^*\Psi)^{-1}\Psi^*]$.

5) $Q = [(\Psi^*\Psi) - (\Psi^*\Phi)(\Phi^*\Phi)^{-1}(\Phi^*\Psi)]^{-1}[\Psi^* - (\Psi^*\Phi)(\Phi^*\Phi)^{-1}\Phi^*]$.

6) $\Phi = P^*X_{11} + Q^*X_{21}$.

7) $\Psi = P^*X_{12} + Q^*X_{22}$,

siendo:

$$X_{11} = (PP^*)^{-1} + (PP^*)^{-1}(PQ^*)[(QQ^*) - (QP^*)(PP^*)^{-1}(PQ^*)]^{-1}(QP^*)(PP^*)^{-1}$$

$$X_{12} = -(PP^*)^{-1}(PQ^*)[(QQ^*) - (QP^*)(PP^*)^{-1}(PQ^*)]^{-1}$$

$$X_{21} = -[(QQ^*) - (QP^*)(PP^*)^{-1}(PQ^*)]^{-1}(QP^*)(PP^*)^{-1}$$

$$X_{22} = [(QQ^*) - (QP^*)(PP^*)^{-1}(PQ^*)]^{-1}.$$

1.3. EJEMPLO DE UN OPERADOR DE APROXIMACIÓN.

En esta sección daremos un ejemplo de un operador de aproximación suryectivo para ejemplificar la teoría descrita en las secciones anteriores. Para esto indicaremos previamente la notación a usar.
Sea τ una tetraedrización arbitraria de un dominio $\Omega \subset \mathbb{R}^3$. Indicaremos con τ^i, $i = 0, ..N$, las tetraedrizaciones obtenidas por la aplicación del colapsado de medio lado a la tetraedrización $\tau^N := \tau$. Notaremos C^i, $i = 0, ..., N$ a los espacios de aproximación formados por las funciones constantes por tramos sobre la tetraedrización τ^i. Si notamos con T^i_j un tetraedro de la tetraedrización τ^i y con $X^i_{T_j}$ su función característica, es claro que el conjunto de funciones $\left\{X^i_{T_j}\right\}_j$ forma una base ortogonal de C^i.

Definiremos a continuación un operador de aproximación suryectivo, que llamaremos operador *promedio* y notaremos P^i. Sean T^{i+1}_j, $j = 1, .., J$, los triángulos de la tetraedrización τ^{i+1} y T^i_k, $k = 1, ..., K$, los tetraedros de la tetredrización τ^i. Definimos:

$$u_j(k) = \begin{cases} 1, & T^{i+1}_j \bigcap T^i_k \neq \emptyset \\ 0, & T^{i+1}_j \bigcap T^i_k = \emptyset \end{cases}.$$

Por otro lado indicaremos con l_j al siguiente número natural:

$$l_j\text{= cantidad de } T^{i+1}_j \text{ tales que } T^{i+1}_j \bigcap T^i_k \neq \emptyset.$$

Luego, el operador promedio P^i es tal que:

$$\begin{array}{rcl} P^i : C^{i+1} & \rightarrow & C^i \\ f^{i+1} & \rightarrow & P^i(f^{i+1}), \end{array}$$

y tiene la siguiente representación matricial:

$$\begin{pmatrix} \frac{u_1(1)}{l_1} & \frac{u_1(2)}{l_1} & \cdots & \frac{u_1(J)}{l_1} \\ \frac{u_2(1)}{l_2} & \frac{u_2(2)}{l_2} & \cdots & \frac{u_2(J)}{l_2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{u_K(1)}{l_K} & \frac{u_K(2)}{l_K} & \cdots & \frac{u_K(J)}{l_K} \end{pmatrix}.$$

Se puede probar que este operador es suryectivo y por lo tanto para calcular las matrices Q, Φ y Ψ se puede aplicar el lema anterior.

REFERENCIAS

[1] L.BOSCARDÍN, *Wavelets definidas sobre volúmenes*, Tesis de Magister, (2001), Universidad Nacional del Sur, Bahía Blanca, Argentina.

[2] M. LOUNSBERY, T. DE ROSE, J. WARREN, *Multiresolution Analysis for Surfaces of Arbitrary Topological Type*, ACM Transactions on Graphics, 16, (1997), pp. 34-73.

[3] P. SCHRÖEDER, W. SWELDENS, *Spherical Wavelets: Efficiently representing functions on the Sphere*, ACM Proceedings of SIGGRAPH'95, (1995) , pp.161-172.

[4] J. BEY, *Tetrahedral Grid Refinement*, Computing, 55, (1995), pp. 355-378.

[5] L. BOSCARDÍN, L. CASTRO, S. CASTRO, A. DE GIUSTI, *Wavelets basis defined over tetrahedra*, Journal of Computer Science Technology, 6, (2006), pp. 46-52.

[6] A. GERUSSI, *Analyse multirésolution non emboîtée. Applications à la visualisation scientifique*, Thèse de Docteur de L'Université J. Fourier, (2000), Grenoble, Francia.

TRANSFORMADA DE FOURIER NO UNIFORME EN EL PROCESAMIENTO DE IMÁGENES: APLICACIONES Y PERSPECTIVAS

Julieth Manta†, Cristyan Manta‡ y Octavio Salcedo†‡

†Proyecto Curricular de Matemáticas, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia, cjmantac@correo.udistrital.edu.co, www.udistrital.edu.co
‡Maestría en Ciencias de la Información y las Comunicaciones, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia, hcmantac@udistrital.edu.co, www.udistrital.edu.co
†‡Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia, osalcedo@udistrital.edu.co, www.udistrital.edu.co

Resumen: Novedosos métodos de obtención de imágenes han promovido el desarrollo de investigaciones de carácter científico, médico y de ingeniería. El procesamiento de señales e imágenes, que representan diversos fenómenos físicos, se consolida en un campo clave para la resolución de preguntas de investigación en cada una de estas áreas. Desde las aplicaciones de tomografía computarizada, imágenes por resonancia magnética, magneto-encefalografía hasta desarrollos en imágenes foto-acústicas, geo-estacionarias y sensores remotos, requieren de herramientas y métodos matemáticos que apoyen las tareas y actividades de investigación. Una de las herramientas matemáticas vitales de mayor impacto en el procesamiento de imágenes es la transformada de Fourier, su extensión a dominios discretos y problemas de tipo no-lineal. En este artículo, se presenta una aplicación y nuevas perspectivas de reconstrucción no-lineal de imágenes mediante la pseudo-inversión de la transformada rápida no uniforme de Fourier.

Palabras claves: *problemas inversos, procesamiento de imágenes, transformada de Fourier*.

Abstract: Novel imaging methods have promoted the development of scientific, medical and engineering research. The processing of signals and images that represent different physical phenomena are consolidated in a key component for the resolution of research questions in each of these areas. From applications of computerized tomography, magnetic resonance imaging, magneto-encephalography to developments in photo-acoustic imaging, geo-stationary and remote sensors require mathematical tools and methods in order to support tasks and research activities. One of the vital mathematical tools with greater impact on image processing is the Fourier transform, its extension to discrete domains and non-linear problems. In this article, we present an application and new prospects for non-linear reconstruction of images using the pseudo-inversion of the non-uniform fast Fourier transform.

Keywords: *Fourier transform, image processing, inverse problems.*

1. INTRODUCCIÓN

En 1822, con la publicación del libro la teoría del calor [6] escrita por J. B. Fourier, nace una de las bifurcaciones más importantes en el estudio matemático y físico, que conllevarían al estudio de las Series de Fourier, la transformada integral de Fourier, y en la actualidad al desarrollo de métodos de computación rápida de la transformada con la extensión al dominio discreto y su combinación con las ciencias computacionales para el estudio de su complejidad e implementación en lenguajes de programación, con la finalidad de solucionar problemas de tipo lineal y no-lineal en todas las áreas del saber.

En el procesamiento de señales e imágenes existen numerosos problemas multi-dimensionales y de orden no-lineal, cuya complejidad requiere el uso de herramientas matemáticas y computacionales para su resolución, ya sea exacta o aproximada. Los intentos por solucionar estos problemas específicamente en el área de la reconstrucción de imágenes médicas convergen a la búsqueda de la llamada transformada rápida no uniforme de Fourier [1], [11], la cual podría tener la complejidad de los algoritmos de *Cooley-Tukey* [4] y ser exacta. No obstante, los resultados hacia esta búsqueda de la transformada rápida no uniforme, son soluciones de carácter aproximado [5]. Con relación a las aplicaciones y perspectivas de la transformada de Fourier en el campo del procesamiento de imágenes, se pueden citar diversos acercamientos de investigación. Algunos, en la adquisición de imágenes por resonancia magnética, en donde las trayectorias de muestreo de tipo espiral [2], *Lissajous*, estocásticas (arbitrarias) [10] y en general de tipo no-lineal, más sofisticadas que las convencionales cartesianas, imponen restricciones y requerimientos para la

computación de la transformada de Fourier y su inversa en el dominio espacio-frecuencia. En consecuencia, la imagen no puede ser reconstruida por simple aplicación de la transformada inversa rápida de Fourier y deben implementarse nuevas estrategias de reconstrucción. Se presentan dos estrategias de reconstrucción de imágenes por pseudo-inversión del operador de Fourier, a. iterativa y b. por reducción de complejidad del operador.

2. HACIA LA EVOLUCIÓN DE LA TRANSFORMADA INTEGRAL DE FOURIER

El corazón de la reconstrucción de imágenes de muestras en espacios no cartesianos y trayectorias no-lineales en el dominio de la frecuencia, ya sea de tipo iterativa y/o por reducción del operador inverso de Fourier, es la *Non Uniform Fast Fourier Transform* NU-FFT hacia delante [3], [5], [7], [9], la cual puede ser definida de la siguiente forma:

2.1. LA TRANSFORMADA DE FOURIER RÁPIDA NO-UNIFORME

La transformada discreta de Fourier de un vector $\mathcal{g} = (\mathrm{g_n})_{\mathrm{n=0}}^{\mathrm{N-1}} \in \mathbb{C}^{\mathrm{N}}$ con respecto a los nodos $\mathcal{w} = (\mathcal{w}_{\mathrm{k}})_{\mathrm{k=-N/2}}^{\mathrm{N/2-1}}$ (con N par) se define por:

$$\mathrm{T}[\mathcal{g}](\mathcal{w}_{\mathrm{k}}) := \sum_{\mathrm{n=0}}^{\mathrm{N-1}} e^{-\mathrm{i}\mathcal{w}_{\mathrm{k}}\mathrm{n}2\pi/\mathrm{N}}\mathrm{g_n}\ , \qquad \mathrm{k} = -\mathrm{N}/2\,, \ldots, \mathrm{N}/2-1. \tag{1}$$

Una evaluación directa de las N sumas en (1) requiere $\mathcal{O}(\mathrm{N}^2)$ operaciones. Usando la clásica transformada rápida de Fourier (FFT) este esfuerzo se puede reducir a $\mathcal{O}(\mathrm{N}\log\mathrm{N})$ operaciones. Sin embargo, la aplicación de la clásica FFT se restringe al caso de nodos equidistantes $\mathcal{w}_{\mathrm{k}} = \mathrm{k}$, $\mathrm{k} = -\mathrm{N}/2\,, \ldots, \mathrm{N}/2-1$.

La FFT no-uniforme unidimensional [5], [8], [9] es una aproximación de la transformada continua no uniforme de Fourier, no obstante es un método altamente exacto para evaluar (1) en nodos arbitrarios $\mathcal{w}_{\mathrm{k}}, \mathrm{k} = -\mathrm{N}/2, \ldots, \mathrm{N}/2-1$, en $\mathcal{O}(\mathrm{N}\log\mathrm{N})$ operaciones.

2.2. DERIVACIÓN DE LA FFT NO-UNIFORME

Para obtener la FFT no-uniforme se sigue el acercamiento de la presentación de *Karsten-Fourmont* [7], el cual se basa en el siguiente lema:

Lema 1 [7, Proposición 1]. Si $c > 1$ y $\propto < \pi(2c-1)$. Supongamos que $\psi\colon \mathbb{R} \to \mathbb{R}$ es continua en $[-\propto, \propto]$, y positiva en $[-\pi, \pi]$. Luego

$$e^{-\mathrm{i}\mathcal{w}\theta} = \frac{c}{2\pi\psi(\theta)}\sum_{\mathrm{j}\epsilon\mathbb{Z}} \Psi(\mathcal{w} - \mathrm{j}/c)e^{-\mathrm{ij}\theta/c}, \qquad \mathcal{w} \in \mathbb{R}\,, |\theta| \le \pi. \tag{2}$$

Donde $\Psi(\mathcal{w}) := \int_{\mathbb{R}} e^{-\mathrm{i}\mathcal{w}\theta}\psi(\theta)\,\mathrm{d}\theta$ denota la transformada de Fourier unidimensional de ψ.

Proposición 2. Si $c, \propto, \psi$ y Ψ sean como en el Lema 1. Luego, para todo $\mathcal{g} = (\mathrm{g_n})_{\mathrm{n=0}}^{\mathrm{N-1}} \in \mathbb{C}^{\mathrm{N}}$ y $\mathcal{w} \in \mathbb{R}$ tenemos

$$\sum_{\mathrm{n=0}}^{\mathrm{N-1}} e^{-\frac{\mathrm{i}\mathcal{w}2\pi}{\mathrm{N}}}\mathrm{g_n} = \sum_{\mathrm{j}\in\mathbb{Z}} e^{-\mathrm{i}\pi(\mathcal{w}-\mathrm{j}/c)}\Psi(\mathcal{w} - \mathrm{j}/c)\,\widehat{\mathrm{G}}_{\mathrm{j}} \tag{3}$$

con

$$\widehat{\mathrm{G}}_{\mathrm{j}} := \frac{c}{2\pi}\left(\sum_{\mathrm{n=0}}^{\mathrm{N-1}} \frac{\mathrm{g_n}e^{-\mathrm{ijn}2\pi/(\mathrm{N}c)}}{\psi(\mathrm{n}2\pi/\mathrm{N}-\pi)}\right), \qquad \mathrm{j} \in \mathbb{Z} \tag{4}$$

Prueba: Tomando $\theta = \mathrm{n}2\pi/\mathrm{N} - \pi \in [-\pi, \pi]$ en (2), tenemos

$$e^{-\mathrm{i}\mathcal{w}\mathrm{n}2\pi/\mathrm{N}} = \frac{c}{2\pi\psi(\mathrm{n}2\pi/\mathrm{N}-\pi)}\sum_{\mathrm{j}\in\mathbb{Z}} \Psi(\mathcal{w} - \mathrm{j}/c)\, e^{-\mathrm{ijn}2\pi/(c\mathrm{N})}e^{-\mathrm{i}\pi(\mathcal{w}-\mathrm{j}/c)}$$

Por tanto,

$$\sum_{\mathrm{n=0}}^{\mathrm{N-1}} e^{-\mathrm{i}\mathcal{w}\mathrm{n}2\pi/\mathrm{N}}\mathrm{g_n} = \frac{c}{2\pi}\sum_{\mathrm{n=0}}^{\mathrm{N-1}}\sum_{\mathrm{j}\in\mathbb{Z}} e^{-\mathrm{i}\pi(\mathcal{w}-\mathrm{j}/c)}\Psi\,(\mathcal{w} - \mathrm{j}/c)\,\frac{\mathrm{g_n}e^{-\mathrm{ijn}2\pi/(c\mathrm{N})}}{\psi(\mathrm{n}2\pi/\mathrm{N}-\pi)}$$

Intercambiando el orden de la suma en el lado derecho de la ecuación anterior se obtiene (3), (4) lo que concluye la prueba.

2.3. Pseudo Inversión del Operador rápido no-uniforme de Fourier

Una solución sencilla del problema inverso de (1), la que puede ser notada en forma matricial como $\mathrm{T} = \Psi \mathrm{g}$ es una operación computacionalmente extensa, la cual está dada por la pseudo-inversa *Moore-Penrose* [8]:

$$\mathrm{g} = \Psi^{+}\mathrm{T} = (\Psi^{\mathrm{H}}\Psi)^{-1}\Psi^{\mathrm{H}}T \tag{5}$$

El superíndice *H* denota transpuesta Hermitiana. Sin embargo, tal solución es prácticamente imposible cuando el número total de muestras N es grande, puesto que necesita de la inversión de una matriz de $N^2 x N^2$ en el caso bidimensional 2D. Alternativamente, (5) puede ser reformulada como un problema de minimización del error de la señal reconstruida:

$$\min_{\mathrm{g}} \|\Psi \mathrm{g} - \mathrm{T}\|_2^2 \tag{6}$$

El problema puede ser resuelto iterativamente con varias técnicas de resolución de ecuaciones integro-diferenciales no lineales, dentro de tales métodos se propone en la presente investigación el uso del método no-lineal del Gradiente Conjugado [8] con la fórmula de actualización de *Fletcher-Reeves*.

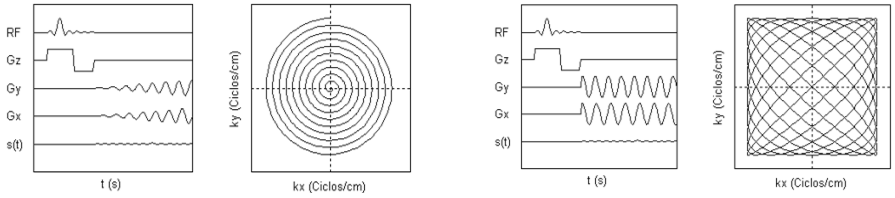


Figura 1. Trayectorias de Adquisición. (a) Lissajous y (b) Espiral.

3. Matemática Aplicada al Procesamiento de Imágenes

3.1. Adquisición de Imágenes: Perspectivas en Resonancia Magnética

El método de adquisición de Imágenes con base en Resonancia Magnética IRM consiste en producir imágenes de una delgada sección del cuerpo humano. En la adquisición, el objeto se coloca en un campo magnético constante, al irradiar el objeto de estudio mediante un impulso de radiofrecuencia de banda angosta, sólo los espines en la frecuencia de resonancia serán excitados. Los gradientes de campo son controlables en los tres ejes espaciales ortogonales entre sí, debido a estos la frecuencia de resonancia del espín es función de su posición.

La codificación espacial con el plano seleccionado se realiza mediante modificación de los gradientes en función del tiempo en los ejes x, y, z concepto este, que recibe el nombre de trayectorias del Espacio k. Estas trayectorias poseen un alto valor informativo, ya que muestran el momento en que se mide cada frecuencia espacial, y para el caso particular de muestreo en frecuencia de tipo no uniforme, corresponden a formas espirales, *Lissajous* (ver Figura 1) y en general no-uniformes. El muestreo de tipo no-uniforme como el citado, es implementado con el fin de reducir la complejidad y limitaciones derivadas para la construcción física de los gradientes de campo. La señal recibida *Free Induction Decay* FID, resultante de modificar los gradientes controlables de campo magnético es igual a la transformada continua de Fourier de la imagen de la sección, a lo largo de la trayectoria $K(t)$. De tal forma, la imagen puede ser obtenida por inversión del operador no-uniforme de Fourier [3], [5], [10].

3.2. Adquisición de Imágenes: Perspectivas en Foto-Acústica

Las imágenes foto-acústicas PAI son una herramienta novedosa para la visualización de la luz que absorben las estructuras en un medio óptico mediante dispersión, las cuales poseen valiosa información para el diagnóstico médico. Tienen como base la generación de ondas acústicas al iluminar un objeto con pulsos de radiación electromagnética no-ionizante, y combina el alto contraste de la óptica pura con la alta resolución de la proyección de imágenes ultrasónicas. Cuando un objeto se ilumina con pulsos cortos de radiación electromagnética no-ionizante, absorbe una fracción de la energía y se calienta. Esto a su vez induce ondas acústicas que se registran con detectores de acústica fuera del objeto.

El ancho de banda en frecuencia de las señales registradas es, por tanto, amplio y depende del tamaño y la forma de las estructuras iluminadas. Las señales acústicas registradas se utilizan para reconstruir la presión acústica generada inicialmente, que representa estructuras del objeto investigado. Para la grabación de la geometría plana, dos tipos de reconstrucción teórica exacta se han reportado: proyección hacia atrás temporal y fórmulas en el dominio de Fourier. En este sentido, un algoritmo de reconstrucción eficiente que utiliza la FFT No-uniforme aumenta la calidad de la reconstrucción mediante interpolación en datos irregularmente espaciados [9].

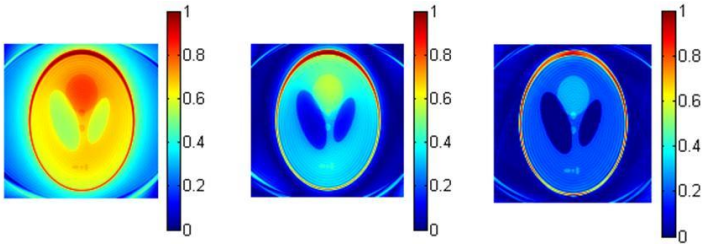


Figura 2. Experimento y Perspectiva de Reconstrucción Iterativa.

4. EXPERIMENTOS Y CONCLUSIONES EN RECONSTRUCCIÓN DE IMÁGENES

La reconstrucción de imágenes se realiza sobre experimentos realizados con el *phantom* de *Shepp-Logan*, ver Figura 2. La expresión analítica de la transformada continua de Fourier del *phantom*, es la base para la adquisición no-uniforme de los datos, mediante muestreo en trayectoria espiral. Se presenta una aplicación y nuevas perspectivas de reconstrucción no-lineal de imágenes mediante la pseudo-inversión de la transformada rápida no uniforme de Fourier, por método iterativo aplicando el gradiente conjugado; y se presentan nuevas perspectivas de reconstrucción de imágenes foto-acústicas. Las perspectivas propuestas se basan en la FFT no-uniforme. Se presentan experimentos sobre el *phantom* de *Shepp-Logan*.

REFERENCIAS

[1] S. BAGCHI, S. K. MITRA, *Non-uniform Discrete Fourier Transform and its Signal Processing Applications*, Norwell, M.A.: Kugler, 1999.

[2] M. BRONSTEIN, A. BRONSTEIN, M. ZIBULEVSKY, H. AZHari, *Reconstruction in ultrasound diffraction tomography using non-uniform FFT*, IEEE Transaction on Medical Imaging, Vol. 21 No. 11, pp. 1395-1401, November 2002.

[3] M. BRONSTEIN, A. BRONSTEIN, M. ZIBULEVSKY, *The non-uniform FFT and some of its applications*, pp. 1-51, November 2002.

[4] J. W. COOLEY, J. W. TUKEY, *An algorithm for the machine calculation of complex Fourier Series*. Mathematical Computation, Vol. 19, pp 297-301, April 1965.

[5] J.A. FESSLER, B.P SUTTON, *Non-uniform Fast Fourier Transforms Using Min-Max Interpolation*. IEEE Transaction on Signal Processing, Vol. 51, pp 560-574, February 2003.

[6] J. B. FOURIER, *Théorie analytique de la chaleur*, Libraires pour les mathématiques, l´architecture hydraulique et la Marine, Rue Jacoob, No. 24, 1822.

[7] K. FOURMONT, *Non-equispaced fast Fourier transforms with applications to tomography*, Journal of Fourier Analysis Applications, vol. 9, no. 5, pp. 431–450, 2003.

[8] J. C. GILBERT, J. NOCEDAL, *Global Convergence Properties of Conjugate Gradient Methods for Optimization*. SIAM Journal on Optimization, Vol. 2, No. 1, pp. 21-42, 1992.

[9] M. HALTMEIER, O. SCHERZER, G. ZANGERL, *A Reconstruction Algorithm for Photoacoustic Imaging based on the Non-uniform FFT*, IEEE Transactions on Medical Imaging, Vol. 28:11, pp. 1727-1735 2009.

[10] C. LIU, J. ZHANG, ME MOSELEY, *Auto-calibrated parallel imaging reconstruction for arbitrary trajectories using k-space sparse matrices (kSPA)*, IEEE Transactions on Medical Imaging, No. 29(3), pp. 950-959, 2010.

[11] A. OPPENHEIM, D. JOHNSON, *Computation of spectra with unequal resolution using the fast Fourier transform*, Proceedings. IEEE, Vol. 59, pp. 299–301, 1971.

UN ESTUDIO ACERCA DE MÉTODOS DE SELECCIÓN DE UMBRAL

Cintia Copa[♭], Zulema Guaymás[♭], María Elena Buemi[†] y Cristian Martínez[♭,†]

[♭]*Departamento de Informática, Facultad de Ciencias Exactas, Universidad Nacional de Salta, Argentina, cintia.dlac@gmail.com, zuleguaymas@gmail.com, martinezdro@yahoo.com.ar*
[†]*Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, mebuemi@gmail.com*

Resumen: El presente trabajo aborda diferentes métodos de selección de umbral para la segmentación de imágenes. La segmentación de imágenes es un área del procesamiento de imágenes que propone técnicas para la detección de bordes sobre imágenes, permitiendo así un reconocimiento automático de objetos en las mismas. Las pruebas computacionales realizadas sobre diferentes imágenes han resultado competitivas, en tiempo computacional y calidad de los resultados.

Palabras clave: *agrupamiento, histograma, segmentación, umbral.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

La segmentación de imágenes es una de las etapas más importantes en el análisis de imágenes. El objetivo es identificar propiedades comunes o detectar zonas que se encuentran muy correlacionadas entre sí, estableciendo contornos que las separan. La transformación más simple para segmentar una imagen es la aplicación de un umbral. La trasformación de una imagen aplicando umbral es uno de los métodos más utilizados en la segmentación de imágenes. Consiste en, dado un umbral, separar los objetos de interés respecto del fondo de la imagen, provocando distinción entre los objetos y el fondo de la escena. Para ello, se debe asignar cada píxel (de la imagen) a un determinado grupo o segmento. La pertenencia de un píxel a un segmento se define mediante la comparación de su nivel de gris y un valor de umbral. El nivel de gris de un píxel equivale a su nivel de luminosidad.

Una técnica muy utilizada es la segmentación por análisis del histograma. Cuando éste presenta dos picos y entre ambos existe un valle, el umbral corresponde a un valor de intensidad perteneciente al valle. Otras técnicas utilizan agrupamiento (clustering) de píxeles según algunos parámetros estadísticos. Algunos casos reales donde se utilizan las técnicas de segmentación son el ánalisis de imágenes provenientes de documentos en el cual se busca extraer información de caracteres impresos, logotipos o gráficos y el análisis digital de diferentes estudios médicos como ser tomografías, resonancias magnéticas, radiografías, entre otros, desde los cuales se intenta auxiliar a los profesionales en el diagnóstico y tratamiento de enfermedades.

En este trabajo se han estudiado diferentes técnicas de selección de umbral para segmentación. La evaluación de la calidad de los resultados obtenidos por las mismas se realizó usando diferentes mediciones del error propuestos en la literatura. De los resultados obtenidos durante las pruebas, podemos afirmar que es posible una detección de umbral de imágenes de manera rápida y segura, libre de apreciaciones subjetivas.

El resto del trabajo se organiza de la siguiente manera: en la sección 2, se realiza una breve introducción a los métodos de detección de umbral. En la sección 3, se explican las pruebas realizadas y los resultados alcanzados. Finalmente, el trabajo futuro y las conclusiones finales se muestran en la sección 4.

2. TÉCNICAS DE SELECCIÓN DE UMBRAL

Dada una imagen $f(i)$ de tamaño N, con L niveles de intensidad, tal que $0 \leq f(i) < L-1$, con $i = [1, \ldots, N]$, T es un umbral si la transformación $g(i)$ cumple,

$$g(i) = \begin{cases} 1 & \text{si } f(i) > T \\ 0 & \text{si } f(i) \leq T \end{cases}$$

con $0 \leq T < L-1$.

Esta transformación es sencilla, pero la decisión de un valor para T ha sido motivo de mucho estudio y análisis, dado que es importante establecer un umbral antes de aplicar otras técnicas como, por ejemplo,

morfología binaria, donde es necesario establecer un umbral antes de dilatar, erosionar, etc. Por esto, es que se buscan métodos automáticos para determinar un valor de umbral óptimo a cada imagen. Mehmet y Bülent [2] categorizaron a los métodos de selección de umbral de la siguiente manera:

- Métodos basados en histograma: estos proponen la obtención del umbral mediante el análisis de picos, valles y curvaturas del histograma.
- Métodos basados en agrupamiento: proponen agrupar píxels mediante características comunes entre ellos.
- Métodos basados en entropía: usan las entropías de las regiones del fondo y del primer plano, y la entropía cruzada entre la imagen original y la binarizada.
- Métodos basados en atributos del objeto: la obtención del umbral se obtiene mediante el análisis de similitud entre la imagen (en escala de grises) y la binarizada.
- Métodos espaciales de distribución de probabilidad de orden superior y / o correlación entre píxeles.
- Métodos locales, adaptados a los valores locales de umbral.

A continuación, describiremos los métodos estudiados.

2.1. MÉTODOS BASADOS EN HISTOGRAMA

2.1.1. Método del Triángulo

Este método obtiene el umbral a través de una distancia. Para ello, Sadeghian y Seman [6] trazan una línea entre el máximo y el mínimo valor del histograma. Luego, calculan la distancia entre la línea y la función de histograma $h(r_k) = n_k$, donde $0 \leq r_k < L$ y n_k es el número de píxeles de la imagen con el nivel de intensidad r_k. El umbral corresponde al valor de distancia obtenido.

2.1.2. Método del Valle

Este método trata de encontrar un buen valle entre todos los presentes. Su dificultad yace en la sensibilidad a pequeñas fluctuaciones o protuberancias. Para salvar esto, Qui y Bo-Li [4] propusieron un método de detección de valle del histograma mediante el siguiente procedimiento:

1. Suavizado del histograma, según un método que comprende a cinco vecinos.
2. Reconocimento de valles en el histograma, basándose en comparaciones respecto de simetría, ancho y profundidad de los mismos.

2.2. MÉTODOS BASADOS EN AGRUPAMIENTO

2.2.1. Método de Otsu

Este método calcula el valor de umbral mediante técnicas estadísticas. El umbral óptimo propuesto por Otsu [3] se obtiene maximizando la varianza entre clases mediante una búsqueda exhaustiva.

Descripción del Método de Otsu para un umbral óptimo: Sea una imagen de tamaño N con valores de gris entre 1 y L. La cantidad de píxeles con nivel de gris i se denota n_i, y la probabilidad de ocurrencia del nivel de gris i en la imagen está dada por: $p_i = \frac{n_i}{N}$.

Para la segmentación en dos niveles de una imagen, los píxels son divididos en dos clases: C_1 para aquellos que poseen nivel de gris comprendido en $\{1, ..., t\}$ y C_2 para aquellos que poseen nivel de gris comprendido en $\{t+1, ..., L\}$, las densidades de probabilidad para C_1 y C_2 son:

$$\frac{p_1}{P(C_1)}, \quad \frac{p_2}{P(C_1)}, \quad ..., \quad \frac{p_t}{P(C_1)} \quad \text{y} \quad \frac{p_{t+1}}{P(C_2)}, \quad \frac{p_{t+2}}{P(C_2)}, \quad ..., \quad \frac{p_L}{P(C_2)}$$

con $P(C_1) = \sum_1^t p_i \quad \text{y} \quad P(C_2) = \sum_{t+1}^L p_i$

El método calcula la media y la varianza de cada clase y la media de toda la imagen, estableciendo una medida de varianza entre clases para todos los pares de clases posibles denominada σ_B^2. El umbral óptimo t^* es: σ_B^2 tal que: $t^* = max\{t : \sigma_B^2(t)\}$ con $1 \leq t \leq L$.

2.2.2. Método de González y Woods

Para la elección automática de umbral, González y Woods [1] describen el siguiente procedimiento:

1. Seleccionar un umbral T inicial (por ejemplo, el punto medio entre el mínimo y máximo valor de intensidad de la imagen).
2. Segmentar la imagen utilizando T, obteniendo dos grupos de píxels: C_1 que contiene todos los píxeles con valores de intensidad $\leq T$ y C_2 que contiene los píxeles con valores $> T$.
3. Calcular la intensidad media entre μ_1 y μ_2 para los píxeles pertenecientes a los grupos C_1 y C_2 respectivamente.
4. Calcular el nuevo valor de umbral: $T = \frac{1}{2}(\mu_1 + \mu_2)$

Los pasos 2 a 4 deben repetirse hasta que la diferencia entre dos umbrales consecutivos sea menor a un parámetro predefinido T_0. Puede notarse que este método separa de forma automática el histograma en dos secciones equilibradas relativas al umbral. Se ha demostrado que este algoritmo converge al mismo umbral obtenido por el Método de Otsu.

3. EXPERIMENTOS

Para medir la calidad de los resultados alcanzados por los métodos de selección de umbral analizados, se han usado los siguientes errores:

- Error en la clasificación (Misclassification Error, ME) [2]: mide el porcentaje de píxeles del fondo asignados incorrectamente al primer plano y viceversa. Para el problema de segmentación en dos clases, ME puede ser expresado como:

$$ME = 1 - \frac{|B_O \bigcap B_T| + |F_O \bigcap F_T|}{|B_O| + |F_O|} \tag{1}$$

  donde B_O y F_O representan el fondo y primer plano de la imagen original (*ground-truth*), B_T y F_T el área de píxeles del fondo y primer plano de la imagen de prueba y $|\,.\,|$, la cardinalidad del conjunto. ME varía desde 0 para una imagen perfectamente clasificada hasta 1 para una imagen binarizada totalmente errónea.

- Error relativo de primer plano (Relative Area Error, RAE): en [2] se propuso la modificación de la medida de precisión relativa de medición final (Relative Ultimate Measurement Error, ***RUMA***) propuesta por [5] que compara características del objeto como área y forma. RAE se obtiene de la siguiente manera:

$$RAE = \begin{cases} \frac{A_O - A_T}{A_O} & \text{si } A_T < A_O \\ \frac{A_T - A_O}{A_T} & \text{si } A_T \geq A_O \end{cases}$$

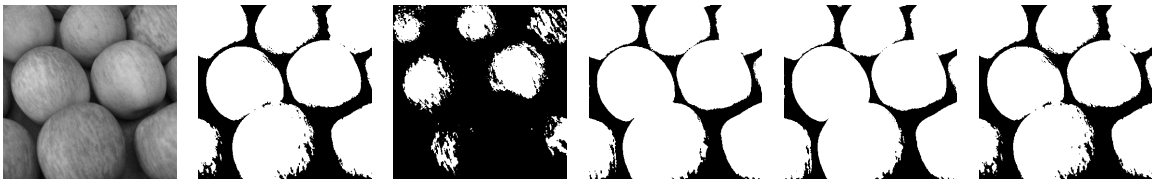
  donde A_0 representa el área de la imagen de referencia y A_T el área de la imagen binarizada. Para una combinación perfecta de las regiones segmentadas, RAE vale cero; mientras que si existen superposiciones de las áreas de objetos, puede valer hasta 1.

La Figura 1 muestra una imagen utilizada en la literatura (Pimientos) y las imágenes obtenidas de la aplicación de los métodos de selección de umbral Otsu, Valle, Triángulo y González y Woods.

La Figura 2 muestra un nuevo experimento de segmentación con otra imagen (Manzanas) y en la Tabla 1 se muestran los errores ME y RAE alcanzados por los métodos mencionados. Para poder realizar esta evaluación cuantitativa se construyó manualmente un *phantom* de la imagen a analizar. De los valores obtenidos y de la evaluación visual surge que, para esta imagen en particular, el método del valle con una operación de suavizado dió resultados satisfactorios, mientras que con 5 operaciones arrojó el mejor resultado. La complejidad de la imagen Pimientos no permitió la creación de su correspondiente *phantom*, por lo tanto no pudo hacerse una evaluación numérica con la misma. Sin embargo, de la apreciación visual surge que las segmentaciones obtenidas con los diferentes métodos presentan diferencias apreciables ya que se puede observar una clara diferencia de los objetos del primer plano respecto del fondo.

(a) Imagen Original (b) Método de Otsu, T=119 (c) Método del Valle, imagen suavizada 5 veces, T=68 (d) Método del Valle, imagen suavizada 10 veces, T=128 (e) Método de Triángulo, T=19 (f) Método de González y Woods, T=119

Figura 1: Imagen Pimientos y la segmentación obtenida por técnicas de selección de umbral



(a) Imagen Original (b) Método de Otsu, T=101 (c) Método del Valle, imagen suavizada 1 vez, T=155 (d) Método del Valle, imagen suavizada 5 veces, T=87 (e) Método de Triángulo, T=83 (f) Método de González y Woods, T=102

Figura 2: Imagen Manzanas y la segmentación obtenida por técnicas de selección de umbral

Errores / Métodos	ME	RAE
Otsu	0.0584	0.0700
Valle (suavizado 1 vez)	0.5831	0.7229
Valle (suavizado 5 veces)	0.0221	0.0015
González y Woods	0.0631	0.0762
Triángulo	0.0239	0.0129

Tabla 1: Errores entre el phantom de la imagen original Manzanas y las binarizadas

4. CONCLUSIONES

En este trabajo se implementaron cuatro métodos de selección de umbral: Otsu, Valle, González y Woods y Triángulo, los cuales fueron evaluados según dos criterios: ME y RAE. Los resultados obtenidos fueron competitivos, dado que los métodos separan correctamente los objetos del primer plano de la escena como así también del fondo de la misma. Esto fue corroborado al analizar los errores de clasificación alcanzados por todos los métodos.

Como trabajo futuro, se preve un estudio más pormenorizado de los métodos vistos, el cual estará incluido en un software de uso académico.

REFERENCIAS

[1] R. GONZALEZ AND R. WOODS, *Digital Image Processing (3rd Edition)*, Prentice-Hall Inc., 2006.

[2] S. MEHMET AND S. BÜLENT, *Survey over image thresholding techniques and quantitative performance evaluation*, Journal of Electronic Imaging, 13-1 (2004), pp.146-165.

[3] N. OTSU, *A Threshold Selection Method from Gray-Level Histograms*, International Conference on Computer Science and Software Engineering, 9-1 (1979), pp.62-66.

[4] C. QUI, X. BO-LI AND K. GANG-YAO, *An Approach on Analyzing Histogram and Selecting Threshold*, International Conference on Computer Science and Software Engineering, 6 (2008), pp.185-188.

[5] Y. ZHANG, *A survey on evaluation methods for image segmentation*, Pattern Recognition, 29 (1996), pp.1335-1346.

[6] F. SADEGHIAN, Z. SEMAN AND A. RAMLI, *A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing*, Biological Procedures Online, 11-1 (2009), pp.196-206.

AUTOESPACIOS DEL GRAFO DE HAMMING $H(2n, 2)$. APLICACIONES EN COMPRESIÓN DE IMÁGENES

F. Levstein[♭], J. Lezama[♭], C. Maldonado[†] y D. Penazzi[♭]

[♭]*Facultad de Matemática, Astronomía y Física. Universidad Nacional de Córdoba. Córdoba, Argentina, Haya de la Torre y Medina Allende, +54-351-4334051/363, www.famaf.unc.edu.ar*

[†]*Facultad de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de Córdoba. Córdoba, Argentina, Av. Velez Sarsfield 1611, +54-351-4334141/4152, www.efn.uncor.edu*

[†,♭]*Universidad Nacional de Córdoba. CIEM CONICET Córdoba.*

Resumen:

Una imagen en escala de grises se representa con una matriz en donde cada entrada corresponde a un pixel representado por un entero entre 0 y 255.

Dada una matriz correspondiente a una imagen, la subdividimos en submatrices $2^n \times 2^n$, que miraremos como vectores en $\mathbb{R}^{4^n}$ (el espacio de las matrices de orden 2^n está naturalmente identificado con $\mathbb{R}^{4^n}$).

A cada $v \in \mathbb{R}^{4^n}$, le aplicamos una matriz H tal que Hv son las coordenadas de v respecto de una base de autovectores de una matriz A. Descartamos la información del vector Hv correspondientes a los autoespacios asociados a los autovalores de menor valor, obteniendo un vector mas chico w. Guardamos w para luego reconstruir un vector $\tilde{v}$ que será una aproximación de v. H debe ser bien elegida para que el ojo humano no detecte mayores diferencias entre v y $\tilde{v}$. Elegimos H utilizando técnicas de combinatoria algebraica.

Palabras clave: *Esquemas de asociación, grafos de Hamming, marcos ajustados finitos, compresión de imágenes*
2000 AMS Subject Classification: 05E30 - 94C15

1. INTRODUCCIÓN

Una de las técnicas utilizadas en compresión de imágenes es descomponer la matriz asociada a una imagen y guardar algunas de las proyecciones en lugar de la imagen total (ver [3], [4], [8], [9]).

En este trabajo consideramos una descomposición de $\mathbb{R}^{64}$ asociada a autoespacios de la matriz de adyacencia A del grafo de Hamming. Utilizamos resultados conocidos de la combinatoria algebraica desarrollados en [1], [2]. Exploramos aplicaciones a la compresión de imágenes utilizando marcos ajustados finitos, (ver [5], [13]), también usados en [3], [4], [8], [9]. Algunos artículos publicados por nuestro grupo relacionados con este trabajo son [10], [11].

A $\mathbb{R}^{64}$ lo descomponemos como suma directa de autoespacios de una matriz A.

$$\mathbb{R}^{64} = V_{\lambda_0} \oplus V_{\lambda_1} \oplus V_{\lambda_2} \oplus V_{\lambda_3} \oplus V_{\lambda_4} \oplus V_{\lambda_5} \oplus V_{\lambda_6}$$

donde λ_j son autovalores distintos de A y V_{λ_j} los autovectores asociados a cada λ_j. Asumimos que los autovalores están ordenados de manera decreciente:

$$\lambda_0 > \lambda_1 > ... > \lambda_6.$$

Dada una imagen y F la matriz asociada, consideramos submatrices de orden 8 (bloques) y lo miramos como vectores $v \in \mathbb{R}^{64}$. Queremos comprimir la imagen F recorriendola por bloques.

Tomamos una matriz H_{64} ortogonal (es decir $H_{64}H_{64}^t = I$) cuyas columnas son autovectores de A. Dado un bloque $v \in \mathbb{R}^{64}$, el vector $H_{64}v$ contiene las coordenadas de v en la base de autovectores. Uno de los métodos para comprimir consiste en guardar sólo las proyecciones a los primeros autoespacios V_0, V_1, V_2 *y* V_3. La compresión viene dada por:

$$\tilde{H_{64}}v$$

$\tilde{H_{64}}$ proviene de sacarle a H_{64} algunas filas. La reconstrucción de v, (con perdida de información), se realiza mediante :

$$\tilde{H_{64}}^t \tilde{H_{64}}v$$

Con este método logramos obtener una compresión del 80 % de la imagen original. A continuación desarrollamos la teoría y desarrollo del trabajo.

2. ESQUEMA DE HAMMING

El hipercubo o esquema de Hamming (ver [1], [2]) $H(n,2) = (X,E)$ es un grafo que tiene como vértices a $X = \mathbb{Z}_2^n$, donde dos vértices $\mathbf{x} = (x_1, x_2, ..., x_n),\ \mathbf{y} = (y_1, y_2, ..., y_n) \in X$ son adyacentes si $|\{i\ :\ x_i \neq y_i\}| = 1$.

Definición 1 *Dado un grafo $G = (X,E)$ se define la matriz de adyacencia A indexada por los vértices del grafo x, y por:*

$$(A)_{xy} = \begin{cases} 1 & \textit{si } x, y \textit{ son adyacentes} \\ 0 & \textit{si no} \end{cases}$$

$\mathbb{R}^{64}$ se descompone como:

$$\mathbb{R}^{64} = V_0 \oplus V_1 \oplus ... \oplus V_6$$

donde $\{V_j\}_{j=0}^6$ son autoespacios de A correspondientes a autovalores distintos.

Ordenamos la descomposición teniendo en cuenta $\lambda_0 > \lambda_1 > ... > \lambda_6$, donde λ_j es el autovalor de A asociado al autoespacio V_{λ_j}.

Sean $\bigwedge_0 = V_0,\ \bigwedge_1 = V_0 \oplus V_1,\ \bigwedge_2 = V_0 \oplus V_1 \oplus V_2,\ \ldots, \bigwedge_6 = V_0 \oplus V_1 \oplus V_2 \oplus \ldots \oplus V_6$

Obteniendose así la filtración:

$$\bigwedge_0 \subset \bigwedge_1 \subset \bigwedge_2 \cdots \subset \bigwedge_6$$

En [11] se prueba que para algunos grafos existe un subconjunto de V_1, $\{v_\tau, \tau \in \Omega\}$ y una constante c tal que si $f \in V_1$ entonces

$$\sum_{\tau \in \Omega} < f, v_\tau > v_\tau = cf$$

i.e. $\{v_\tau, \tau \in \Omega\}$ es un marco ajustado finito (ver [5]) para V_1, además se tiene que el conjunto de índices Ω es a su vez un esquema de asociación. Más aún, este resultado se puede extender a otros autoespacios, generando así una familia de marcos ajustados.

En el caso del Hamming se observa que un subconjunto de los marcos ajustados obtenidos en [11] forman una base ortogonal de autovectores de A, con propiedades que permiten predecir que la compresión será buena en engañar al ojo humano. Esta forma una matriz ortogonal de orden 2^n que denotamos H_{2n}.

2.1. APLICACIONES EN COMPRESIÓN DE IMÁGENES

Por otro lado una imagen en escala de grises se representa con una matriz en donde cada entrada de la matriz corresponde a un pixel dado en una escala de grises, (intensidad), representada por un entero entre 0 y 255. El color negro corresponde al 0 y el blanco al 255. Trabajamos con imágenes 2^9 x 2^9 que son las más comunes. Para profundizar más en el tema ver [7], [12].

Dada una matriz F, 512 x 512, correspondiente a una imagen en $\bigwedge_{2n}$, subdividimos F en submatrices (bloques) de orden 2^n. A cada submatriz la miraremos como vectores en $\mathbb{R}^{4^n}$. Llamemos $v \in \mathbb{R}^{4^n}$ a una de ellas. El vector $H_{2n}v$ contiene las coordenadas de v en la base de autovectores. Guardamos con mayor precisión las coordenadas correspondientes a autoespacios asociados a los autovalores de mayor valor. Siendo $\bigwedge_{2n}$ el nivel de mayor resolución y $\bigwedge_0$ el de menor resolución.

2.2. TRUNCAMIENTO

Se consideraron diversos métodos para comprimir la imagen. Uno de ellos consiste en guardar sólo las proyecciones a los primeros autoespacios $V_0, V_1, \ldots, V_j$, es decir, quedandonos en el nivel de resolución $\bigwedge_j$ y reconstruir la imagen con dicha información de la siguiente manera:

Sea k = $\sum_{j=i+1}^{2n} \binom{2n}{j}$ y H^* = H_{2n} sin las ultimas k filas. La compresión viene dada por:

$$H^* v$$

La reconstrucción de v, (con perdida de información), se realiza mediante la formula:

$$H^{*T} H^* v$$

Este procedimiento lo aplicamos dos veces aumentando así el nivel de compresión.

2.3. OPTIMIZACIÓN DEL TRUNCAMIENTO

Otro método utilizado consiste en guardar las proyecciones de Hv a los primeros V_j con mayor nivel de intensidad en la escala de grises, y la información contenida en los restantes autoespacios con menor resolución.

Para optimizar nuestro trabajo medimos la energía contenida en todos los vectores Hv de la imagen original con respecto a un porcentaje dado por una constante optima y reconstruimos cada vector Hv con la proyección a los autoespacios asociados, en donde se registre mayor concentración de la energía medida, optimizando de esta forma el nivel de compresión.

Para analizar nuestros resultados consideramos distintas imágenes y calculamos el error cuadrático medio de la imagen reconstruida con su original. El histograma de la diferencia de estas imágenes muestra una gran concentración alrededor del 0. También analizamos nuestros resultados con la métrica signal to noise ratio (SNR) (ver [6]):

$$Métrica\ signal - to - noise\ ratio\ (SRN)$$

$$SNR(f,g) = (\frac{\sum_{x\in X} g(x)^2}{\sum_{x\in X}(f(x)-g(x))^2})^{\frac{1}{2}}$$

donde f es la imagen original y g es la imagen obtenida. Esta métrica es muy conocida en el área de compresión de imágenes, mide la cantidad de ruido obtenido entre la imagen original y la imagen obtenida. Esta métrica tiene sentido siempre que sea mayor a 1, obteniendo mejores resultados mientras mayor sea la misma.

El software utilizado en este trabajo es ENVI 4.3 con una base de datos de imgenes con distintos contrastes de grises.

3. CONCLUSIONES

Con este método obtenemos una compresión del 80 %, con grandes expectativas de aumentar y mejorar el nivel de compresión utilizando otros esquemas de asociación y sus respectivos marcos ajustados finitos y combinarlos con los procedimientos aplicados por los métodos estándar, con el objetivo de alcanzar a los métodos estándar, por ejemplo JPEG, que tienen una compresión mayor al 90 %.

Figura 1: Imagen original

Figura 2: Imagen obtenida con compresión del 80 %

REFERENCIAS

[1] R. A .BAILEY. *Association Schemes: Designed Experiments, Algebra and Combinatorics*, Cambridge University Press.

[2] E. BANNAI, T. ITO, *Algebraic Combinatorics I: Association Schemes*, Benjamin Cummings,Lecture Notes Series, Menlo Park, Cal, 1984.

[3] A. CHEBIRA AND J. KOVACEVIC, *Life Beyond Bases: The Advent of Frames (Part I)*, IEEE SP Mag., vol. 24, no. 4, Jul. 2007, pp. 86-104.

[4] A. CHEBIRA AND J. KOVACEVIC, *Life Beyond Bases: The Advent of Frames (Part II)*, IEEE SP Mag., vol. 24, no. 5, Sep. 2007, pp. 115-125.

[5] I. DAUBECHIES, *Painless Nonorthogonal Expansions*, J. Math. Phys. vol. 27, pp. 1271-1283 (1986).

[6] E. DEZA AND M. DEZA, *Dictionary of distances*, ELSEVIER, Chapter 21.

[7] J. M. DE LA CRUZ GARCA Y G. PAJARES MARTINSANZ, *Visin por Computador*, 2 ed. Alfaomega (2002).

[8] R. FOOTE, G. MIRCHANDANI, D. ROCKMORE, D. HEALY AND T. OLSON, *A Wreath Product Group Approach to Signal and Image Processing: Part II - Convolution, Correlation, and Applications*, IEEE transactions on signal processing, 2000, vol. 48, no1, pp. 102-132.

[9] R. FOOTE, G. MIRCHANDANI, D. ROCKMORE, D. HEALY AND T. OLSON, *A Wreath Product Group Approach to Signal and Image Processing: Part II - Convolution, Correlation, and Applications*, IEEE transactions on signal processing, 2000, vol. 48, no1, pp. 102-132.

[10] F. LEVSTEIN, C. MALDONADO AND D. PENAZZI, *The Terwilliger algebra of a Hamming scheme H(d,q)*, European Jornual of Combinatories 27 (2006) 1-10.

[11] F. LEVSTEIN, C. MALDONADO AND D. PENAZZI, *Lattices, frames and Norton algebras of dual polar graphs*, Special Volume of the Contemporary Mathematics. Series:Proceedings of the VII Workshop in Lie Theory and its Applications. Edited by: Carina Boyallian and Linda Saal.

[12] V. MADISETTI AND D. WILLIAMS, *The digital signal processing handbook*, (CRC PRESS, IEEE PRESS, 1998).

[13] R. VALE AND S. WALDRON, *Tight frames and their symmetries*, Const. Approx., 21:83-112, 2005.

EVALUACIÓN DE CALIDAD DE IMÁGENES DE RADAR DE APERTURA SINTÉTICA

Gustavo Lazarte y Elizabeth Vera de Payer
Facultad de Ciencias Exactas, Físicas y Naturales – Universidad Nacional de Córdoba – Av. Vélez Sarsfield 1611
Córdoba – Argentina, g_lazarte@hotmail.com, epayer@efn.uncor.edu

Resumen: Los sistemas SAR permiten obtener imágenes de reflectividad del terreno utilizando el concepto de apertura sintética, basado en la síntesis de una antena ficticia de gran longitud para obtener de este modo una alta resolución en azimut. En un control de calidad rutinario, realizado con el despliegue de la imagen, es posible detectar anomalías fuertemente evidentes aunque las sutiles no pueden ser observadas. El análisis de calidad riguroso de las imágenes SAR debe ser entonces introducido en una producción masiva de productos, para lo cual se ha desarrollado un software que permite observar el resultado del producto, complementando el control de calidad visual rutinario. La información aportada permite conocer la calidad de la imagen desarrollada y alertar sobre la necesidad de reprocesamiento con, por ejemplo, datos orbitales definitivos. Este control de calidad es a posterior del control rutinario operativo.

Palabras Claves: *Procesado SAR, parámetros estadísticos*
2000 AMS Subjects Classification: 68U10 - 46N30

1. INTRODUCCIÓN

La detección a distancia (teledetección o percepción remota) es una técnica que permite obtener información sin que el sensor se encuentre en contacto con el objeto, superficie o fenómeno bajo estudio. Al margen de algunas aplicaciones ya consolidadas como la meteorología radar, sondeo ionosférico y del subsuelo, etc., los trabajos de I+D actuales se centran en tres tipos de sensores embarcados en satélite: altímetros, radares de apertura sintética (SAR) y dispersómetros.

Los Radares de Apertura Sintética permiten formar, mediante un elaborado procesado de la señal radar, imágenes de la superficie planetaria con resoluciones del orden de algunos metros. Las aplicaciones potenciales de estos sistemas son innumerables: cartografía de zonas de alta nubosidad (inaccesibles mediante sensores ópticos), obtención de modelos topográficos a escala mundial de alta precisión, exploración de otros planetas o satélites con atmósfera, determinación de recursos hídricos, vegetación, clasificación de cultivos, etc. [2]

Tanto la generación como la interpretación de una imagen radar son complejas, resultando la sola observación visual no suficiente si se la quiere caracterizar desde el punto de vista de la calidad. Por ello, se hace indispensable analizar la imagen desde su proceso de formación estudiando los parámetros más importantes que la caracterizan. Este estudio habilita estimar la calidad tanto del sensor SAR como del software de procesamiento de la imagen resultante.

El Sistema Radar de Apertura Sintética permite obtener imágenes de alta resolución del terreno. Esto es posible mediante una antena virtual sintetizada en forma matemática que representa a una antena de considerable dimensiones, y que es ubicada sobre una plataforma móvil que puede ser un avión o un satélite. Básicamente, el sistema SAR utiliza un transmisor de radiofrecuencias ubicado en la plataforma móvil. Para su funcionamiento, el transmisor emite un haz de radiación electromagnética de corta duración que impacta contra el terreno, luego las ondas reflejadas son interceptadas y procesadas por el receptor de la plataforma móvil y, finalmente, estos datos se envían a la Estación Terrena para el procesamiento definitivo de la información. [4]

Un píxel de la imagen final es el resultado de la composición matemática de varios ecos recibidos en el satélite de la zona correspondiente durante el viaje del transceptor. Es como si la antena tuviera una gran extensión durante el sobrevuelo (de ahí el nombre de apertura sintética). Una vez conocida la velocidad orbital del satélite, el desplazamiento en frecuencia relativo de los movimientos (efecto Doppler-Fizeau) y la frecuencia patrón de referencia, es posible lograr una matriz bidimensional que representa el terreno. Esta matriz es el resultado de una conversión de un sistema radial de referencia, con centro en el sensor y el

sistema de referencia plano cartesiano que representa el terreno, operación que se denomina focalización dentro del procesamiento SAR.

En este trabajo se desarrolla una herramienta que permite analizar, desde el punto de vista estadístico, tanto la respuesta del sistema SAR a un reflector puntual [1] como a una zona de características homogéneas, estudiando en forma sectorizada el comportamiento de los pixeles procesados. Los datos obtenidos pueden ser utilizados para un análisis de variación temporal de la zona bajo estudio, como por ejemplo, desplazamientos de grietas, deformaciones y/o alteraciones importantes del terreno bajo estudio como así también para complementar el análisis visual de un control operativo rutinario permitiendo la observación detallada a nivel del pixel. Se elige una zona homogénea en Antártida donde se espera y posteriormente se verifica que los niveles de señal procesados son idénticos, en tanto que en zonas con rugosidades (terreno difuso) se aprecian variaciones en los niveles de pixeles procesados

2. ESTIMACIÓN DE LA CALIDAD DE LA IMAGEN SAR

Una característica importante de las imágenes generadas con el sistema de radar de apertura sintética es la posibilidad que brinda de detectar visualmente ciertas anomalías, algunas de las cuales pueden ser solucionadas. Sin embargo, hay otras que no son visibles o que son difícilmente detectables en un ambiente operativo de generación de imágenes, como por ejemplo, el ajuste del pulso de compresión en azimut, el cálculo del centroide Doppler, la migración en rango, la ambigüedad, el ruido speckle, la reflexión direccional, los brillos y sombras en la imagen, el movimiento de la plataforma, etc.

El desarrollo consiste en el diseño y la realización de un software en lenguaje gráfico, Labview (www.ni.com), que integra rutinas producidas con Matlab (www.mathworks.com) para la obtención de gráficos y cómputos estadísticos, tanto para el análisis de la respuesta del sistema SAR frente a una fuente puntual como para el análisis radiométrico frente a una fuente homogénea, como así también el acceso a Google Earth para la posibilidad del análisis de georeferencia de la imagen radar obtenida. Los valores estadísticos obtenidos permiten complementar el control de calidad de la imagen radar procesada.

3. PARÁMETROS DE CALIDAD

La respuesta SAR a un objeto puntual (por ejemplo, reflector, antena de la Estación Terrena) es caracterizada por la función respuesta al impulso. Su análisis permite determinar parámetros relacionados con la resolución espacial, como así también la presencia de lóbulos laterales indeseados y por lo tanto, el nivel de ambigüedad. La resolución radiométrica generalmente se estudia en una región homogénea (donde el coeficiente de retrodispersión es constante) observando las variaciones de los valores de los pixels vecinos.[5]

En cuanto a la calidad radiométrica de los datos SAR, se puede decir que ésta es afectada por factores inherentes al instrumento y por la geometría de la iluminación. Las dos causas principales de distorsiones radiométricas que perjudican la interpretación de las imágenes de radar son: el ruido speckle y el efecto del patrón de la antena.[4]

3.1 METODOLOGÍA PARA EL ANÁLISIS DE LAS IMÁGENES

El método que se propone en el trabajo parte de obtener una porción de imagen centrada en un punto brillante, preferentemente libre de otros puntos de características semejantes que pudieran afectar la base del estudio y/o una porción de imagen de brillo uniforme [3]. Se pretende trabajar en un formato TIFF y emplear un software realizado en Labview que integra algoritmos generados en Matlab y la herramienta Google Earth para obtener los parámetros de calidad de la imagen.

Para el estudio de la respuesta al impulso del sistema se propone generar a partir de la imagen CEOS (Committee on Earth Observation Satellites**)** original una imagen en formato TIFF no comprimido de, por ejemplo, 300 x 300 puntos, alrededor del reflector puntual, empleando para esta transformación, el software ENVI. Google Earth permite el estudio de georeferencia de esta zona de observación. Para el estudio de la resolución radiométrica, también se propone generar una imagen TIFF de las mismas características, aunque

aquí ya no es posible el análisis de georeferencia. Se genera para este caso una imagen TIFF a partir de una imagen suministrada por Alaska Sar Facility (The Ends of the Earth).

El Software desarrollado, SAR_QC, realizado en Labview, permite obtener parámetros estadísticos necesarios para complementar el control de calidad rutinario y visual sobre las imágenes de un radar de apertura sintética tanto para el estudio sobre una fuente puntual como sobre un terreno uniforme obteniéndose datos estadísticos imprescindibles para la detección de cualquier anomalía no visible presente en la imagen. El software requiere como entrada de datos, además de los archivos de texto de encabezamiento propios del producto CEOS, las imágenes en formato TIFF no comprimido de 8 bits a las que se desea realizar el control de calidad. Esta transformación de CEOS a TIFF tiene como ventajas el independizar el sistema de control de calidad del procesador con el que se realiza el producto, el cual varía de acuerdo al proveedor o de acuerdo a mejoras implementadas a lo largo del tiempo sobre el sistema de procesamiento.

4. APLICACIÓN

La zona seleccionada para estudio de radiometría es Antártida, Figura 1-a, correspondiente a Latitud 77° 52' Sur y Longitud 34° 37' Oeste, Figura 1-b, que es la correspondiente a Base Belgrano II donde existen zonas de hielo uniforme en grandes extensiones.

Se trata de una zona altamente reflectiva la cual produce valores radiométricos que pueden ser comparados con los esperados en calibración. Cualquier desviación significativa debiera ser reportada en el control de calidad. Puede usarse la Base Belgrano para estudiar la georeferencia de la imagen adquirida aunque este proceso puede resultar inapropiado por la alta presencia de nieve durante todo el año. Aquí es necesario instalar un reflector puntual para la correcta georeferenciación y el estudio de la respuesta impulsiva del sistema.

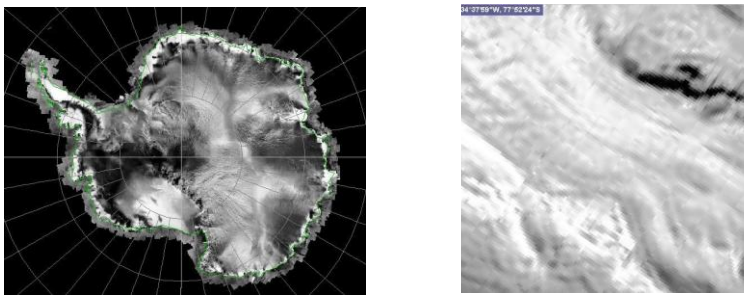


Figura 1: a) Antártida b) Región de estudio

Realizado el análisis estadístico para el estudio radiométrico se obtuvieron los resultados indicados en la Figura 2:

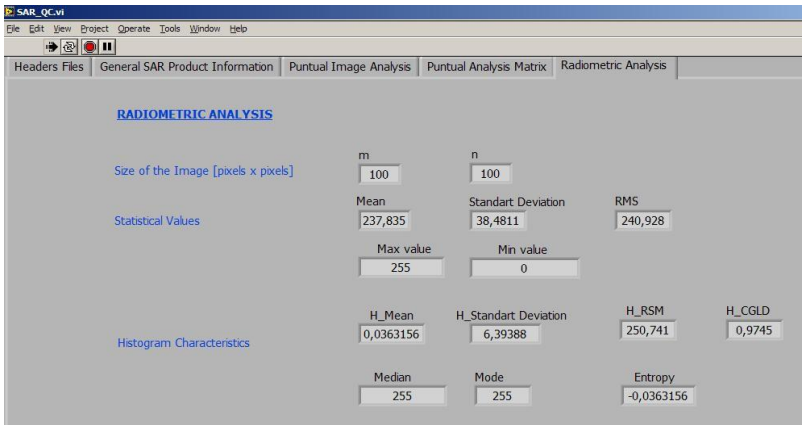


Figura 2: Análisis Radiométrico de una porción de la región en estudio

En la figura 2 pueden apreciarse los siguientes resultados:

- Tamaño de la imagen en píxeles: 100 x 100
- Valores medio, desviación estándar, valor eficaz, máximo y mínimo.
- Valores característicos obtenidos sobre el histograma: valor medio, desviación estándar, valor eficaz, distribución de nivel de gris acumulada, mediana, modo y entropía.

La imagen bajo estudio presenta valores límites de 255 y 0, en tanto que su entropía es -0,036 indicando la textura no uniforme del terreno. El modo de la imagen, que representa el nivel de gris asociado a la mayor frecuencia, es 255.

Teniendo en cuenta que todas las distorsiones geométricas afectan la calidad radiométrica de la imagen es que es imprescindible que el terreno bajo estudio sea totalmente plano libre de obstáculos. Las distorsiones radiométricas existen en conexión con el relieve del terreno y no pueden ser completamente corregidas. Cualquier variación en los valores de pixeles vecinos puede indicar desniveles en el terreno o grietas considerables.

5. CONCLUSIONES

Los sistemas SAR permiten obtener imágenes de reflectividad del terreno utilizando el concepto de apertura sintética, basado en la síntesis de una antena ficticia de gran longitud para obtener de este modo una alta resolución en azimut.

La principal ventaja de estos sistemas en la generación de imágenes de la superficie terrestre consiste en que, al tratarse de sistemas activos que suministran la fuente de iluminación, permiten adquirir datos en ausencia de la luz solar y en condiciones meteorológicas adversas.

Básicamente, los sistemas SAR emiten pulsos a una frecuencia PRF, reciben los ecos y los almacenan en una matriz de datos RAW. Existen varios algoritmos que realizan el procesado de estos datos para obtener la imagen resultante.

En un control de calidad rutinario, realizado con el despliegue de la imagen es posible detectar anomalías fuertemente evidentes aunque las sutiles no pueden ser observadas. El análisis de calidad riguroso de las imágenes SAR debe ser entonces introducido en una producción masiva de productos, para lo cual, el uso de herramientas tal como la desarrollada, permite observar el resultado del producto, complementando el control de calidad visual rutinario. La información aportada permite a priori conocer la calidad de la imagen desarrollada y alertar sobre la necesidad de reprocesamiento con, por ejemplo, datos orbitales definitivos. Este control de calidad es a posterior del control rutinario operativo.

6. REFERENCIAS

[1] G. BENITO, PROSAR: PROCESADOR DE IMÁGENES SAR, INVAP S.E., 2004

[2] CURLANDER, MCDONOUGH, *Synthetic Aperture Radar: Systems and Signal Processing*, John Wiley & Sons, New York, 1991

[3] A. MARTINEZ A., J.L. MARCHAND, *SAR Image Quality Assessment, Revista de Teledetección, 1993.*

[4] C. OLIVER, S.QUEGAN, *Understanding Synthetic Aperture Radar Images*, SciTech Publishing, Inc., 2004. ISBN: 1-891121-31-6

[5] S. SRIVASTAVA, *Radarsat-1 Image Quality: The continuing success* Operations Planning Manager, Satellite Operations, Canadian Space Agency. Argentina on-site visit, 2001.

SOBRE EL TAMAÑO DE LA TDF EN MÉTODOS DE CONVOLUCIÓN POR BLOQUES

Eduardo E. Paolini

Departamento de Ingeniería Eléctrica y de Computadoras - Universidad Nacional del Sur
Instituto de Investigaciones en Ingeniería Eléctrica "Alfredo C. Desages"
Av. Alem 1253 - (B8000CPB) Bahía Blanca, Argentina. epaolini@uns.edu.ar

Resumen: Los métodos de convolución o filtrado por bloques se basan en la aplicación de la transformada discreta de Fourier (TDF) y en particular de la transformada rápida de Fourier (FFT) para procesar eficientemente grandes volúmenes de datos. En este trabajo se discute cómo elegir el tamaño de la FFT a partir del conocimiento de la longitud de las sucesiones a convolucionar para minimizar de manera subóptima el número de operaciones requeridas por unidad de tiempo, y el número total de operaciones.

Palabras clave: *convolucion por bloques - metodos rápidos - FFT*
2000 AMS Subject Classification: 21A54-55P54

1. INTRODUCCIÓN

Los métodos de convolución o filtrado por bloques permiten procesar eficientemente grandes volúmenes de datos, sobre todo cuando una de las sucesiones, que se denominará *entrada* $x[n]$, tiene una longitud N_x mucho mayor que el tamaño P de la otra sucesión, el *filtro* $h[n]$. Estos métodos emplean la transformada rápida de Fourier (FFT por sus siglas en inglés) para calcular un conjunto de convoluciones parciales de tamaño menor. Las dos técnicas más conocidas son las de *overlap-add* y *overlap-save* [4], [6], y se diferencian en la manera en que se toman los bloques de la entrada $x[n]$ y en la forma de combinar los resultados de cada una de las convoluciones parciales. Estas se pueden implementar de manera muy eficiente aplicando la convolución circular [3] entre los bloques. El procedimiento consiste en tomar bloques de longitud L de la sucesión de entrada $x[n]$, calcular la TDF de orden N, multiplicarla por la TDF de orden N de la respuesta impulsiva del filtro $h[n]$ de largo P, antitransformar el producto y concatenar adecuadamente los bloques obtenidos. Cuando las convoluciones circulares se calculan utilizando la transformada rápida de Fourier (FFT), el tamaño N de la TDF debe ser potencia de 2, $N = 2^\gamma$.

El problema que se trata en este trabajo es el de determinar el tamaño óptimo N de cada una de las convoluciones parciales (el tamaño de la FFT) de manera de (a) minimizar el número de operaciones por unidad de tiempo que es necesario realizar para calcular cada muestra de salida, y (b) minimizar el número total de operaciones necesario para convolucionar la sucesión $x[n]$ con el filtro $h[n]$. Posiblemente, el primer objetivo es más importante que el segundo, ya que una menor cantidad de operaciones por unidad de tiempo permite implementar los algoritmos en procesadores menos veloces, y por lo tanto, de menor costo.

2. PRELIMINARES

La convolución lineal entre dos sucesiones reales discretas $x[n]$ y $h[n]$ que no son idénticamente nulas para $0 \le n \le N_x - 1$, y $0 \le n \le P - 1$, respectivamente, donde se supone que $N_x \gg P$, da como resultado una sucesión, denominada *salida* $y[n]$ de longitud $N = N_x + P - 1$. La convolución linear discreta entre $x[n]$ y $h[n]$ se expresa como

$$y[n] = x[n] * h[n] = \sum_\ell x[n-\ell]h[\ell] = \sum_\ell x[\ell]h[n-\ell], \tag{1}$$

donde los extremos de las sumatorias dependen del instante n en que se evalúa la salida, y de la expresión utilizada (tercer o cuarto miembro) para calcularla; por razones de espacio no se incluyen en este trabajo, pero pueden derivarse muy fácilmente [5]. El número de operaciones necesarias para calcular cada muestra de la convolución (1) es de P productos reales y de $P-1$ sumas reales. Al principio y al final del conjunto

de datos el número de operaciones es menor porque cierto número de muestras de $x[n]$ son nulas, pero esta diferencia puede despreciarse si $N_x \gg P$.

En la convolución por bloques, la sucesión $x[n]$ de longitud N_x se particiona en bloques $x_i[n]$ de longitud $L > P$ que se convolucionan con la respuesta impulsiva $h[n]$ del filtro. Aunque estas convoluciones parciales pueden calcularse usando (1), desde el punto de vista de la eficiencia computacional es más conveniente utilizar una convolución circular basada en la FFT de orden N. La relación entre las dimensiones L, P, y N depende del método de filtrado por bloque se utilice: para *overlap-add*, $L = N-(P-1)$, mientras que para *overlap-save*, $L = N$. Aunque la forma de combinar los resultados parciales es diferente en cada caso, el número de muestras "útiles" N_u que se obtiene al procesar cada bloque es el mismo para los dos métodos: $N_u = N-(P-1)$. Este número de "muestras útiles" tiene en cuenta que el resultado de cada convolución circular no coincide necesariamente con el de la convolución lineal entre $x_i[n]$ y $h[n]$ (overlap-save) o no corresponde exactamente con las muestras correspondientes del resultado final (overlap-save): sólo N_u muestras coinciden exactamente con un intervalo de las muestras de la convolución completa.

Para evaluar la eficiencia computacional, se considerará el número de operaciones (sumas y productos) que deben realizarse para obtener la salida $y[n]$. Si bien en una implementación real otros factores pueden ser tanto o más relevantes para el desempeño del algoritmo (velocidad de los accesos a memoria, cantidad de registros auxiliares, etc.) el número de operaciones es un indicador confiable para procesadores de propósito general y también para procesadores dedicados (DSPs). En general, si $N \gg 1$ y $P \gg 1$, el número de sumas y productos es del mismo orden del magnitud cuando las FFT se calculan por los métodos de decimación en tiempo o en frecuencia [4]. Por lo tanto, de aquí en adelante la eficiencia computacional se relacionará con el número de productos necesarios para implementar un dado algoritmo.

Para computar la FFT de orden $N = 2^\gamma$ se requieren $(N/2)\log_2 N$ productos complejos [2]. Como cada multiplicación compleja demanda 4 operaciones reales, el número de productos reales es $2N\log_2 N$. Si se descartan las multiplicaciones por 0, 1, etc., el número se reduce a $2N\log_2 N - 5N$, y si los datos $x[n]$ son reales, el número de multiplicaciones es $N_{FFT} = N\log_2(N/2) - N/2$ [1].

Para procesar cada bloque es necesario: (*i*) calcular la TDF del bloque de entrada (N_{FFT} operaciones), (*ii*) multiplicar esta TDF por la TDF de la respuesta impulsiva del filtro ($4N$ multiplicaciones reales), y (*iii*) calcular la TDF inversa del producto (N_{FFT} operaciones). La TDF del filtro $h[n]$ sólo debe calcularse una vez, y puede estar pre-calculada. La TDF inversa se calcula con el mismo algoritmo que la TDF, conjugando los datos de entrada y salida [4]. La "conjugación" no se computa como una operación ya que sólo implica cambiar el signo de un registro. En consecuencia, cada bloque demanda N_{PR} multiplicaciones reales , donde

$$N_{PR} = 2 \times N_{FFT} + 4N = 2N\log_2(N/2) + 3N. \tag{2}$$

3. El número de operaciones por unidad de tiempo

El número de operaciones por unidad de tiempo se define como la cantidad de operaciones que se deben realizar para calcular cada muestra de salida. Para un funcionamiento adecuado del algoritmo deben concluirse en un lapso de tiempo menor que el período de muestreo, y frecuentemente es deseable que este tiempo sea aún menor para permitir que el procesador realice otras tareas (lectura de teclado, comunicaciones, etc.)

En la convolución lineal (1) se necesitan a lo sumo P productos reales y $P-1$ sumas para calcular cada muestra de salida, y por lo tanto, el número de operaciones por unidad de tiempo es $N^C_{O/S} = P$.

En el procesado por bloques, en cada iteración se obtienen $N_u = N-(P-1)$ muestras "útiles", para las cuales es necesario efectuar N_{PR} operaciones (ec. 2), y el número de operaciones por unidad de tiempo es $N^B_{O/S} = N_{PR}/(N-P+1)$. Como N debe ser potencia de 2, y además $N > P$, se notará $N = 2^\gamma 2^{\lceil\alpha\rceil}$, con $\gamma \in \mathbb{N}$, $\alpha = \log_2 P$, y donde $\lceil\cdot\rceil$ es la función "techo": $\lceil x\rceil$ es el menor entero mayor a x. En consecuencia,

$$N^B_{O/S}(P,\gamma) = \frac{2^{\gamma+\lceil\log_2 P\rceil}\,(2\gamma + 2\lceil\log_2 P\rceil + 1)}{2^{\gamma+\lceil\log_2 P\rceil} - P + 1}. \tag{3}$$

Para minimizar $N^B_{O/S}$ interesa encontrar el valor de $\gamma = \hat{\gamma} \in \mathbb{N}$ que hace mínimo (3) para cada valor de P. Es evidente que $\hat{\gamma}$ no es una función continua de P, como se muestra en la Figura $1(a)$, para P variando

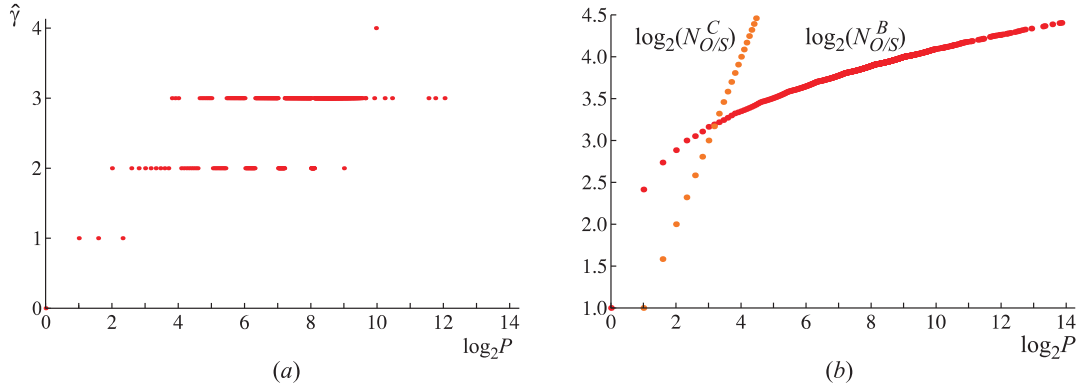


Figura 1: (a) Valor óptimo del exponente γ y (b) cantidad de operaciones por segundo, en función de la longitud P.

entre 1 y $16384 = 2^{14}$, un rango más que adecuado de longitudes de filtros. El número de operaciones por segundo $N^B_{O/S}$ para la convolución por bloques se grafica en la Figura $1(b)$, junto con $N^C_{O/S}$ para la convolución lineal. El cruce de las curvas se produce para $P = 10$, lo que revela que aún con filtros muy sencillos la implementación de convolución o filtrado por bloques es más eficiente (aunque más compleja).

No es sencillo encontrar una expresión para $\hat{\gamma}$ en función de P. Sin embargo, este esfuerzo no se justifica, ya que la diferencia entre el valor óptimo $N^B_{O/S}(P, \hat{\gamma})$ y el valor de $N^B_{O/S}$ para un valor determinado de γ es pequeño, como muestra la Figura 2, que representa la diferencia $\Delta = \log_2 N^B_{O/S}(P, 3) - \log_2 N^B_{O/S}(P, \hat{\gamma})$ entre el $N^B_{O/S}$ que corresponde a $\gamma = 3$ y el $N^B_{O/S}$ óptimo, para cada valor de P entre 0 y 16384. Esto indica que un número casi óptimo de operaciones por segundo se obtiene eligiendo el tamaño N de la TDF como $2^\gamma = 8$ veces la longitud del filtro FIR redondeado a la potencia de 2 más próxima.

4. El número total de operaciones

Para la convolución lineal, el número total de operaciones necesarias para realizar el filtrado de la señal es $N_P = PN_x$ productos reales, y $N_S = (P-1)(N_x-1)$ sumas reales. Como en general $P \gg 1$ y $N_x \gg 1$, se puede considerar que el número de productos y sumas reales es aproximadamente PN_x.

Para la convolución por bloques, como la longitud de la convolución lineal entre $x[n]$ y $h[n]$ es N_x+P–1, el número de bloques a procesar es $N_b = \lceil (N_x+P-1)/N_u \rceil = \lceil (N_x+P-1)/(N-P+1) \rceil$. Si N_x es suficientemente grande, este resultado es aproximadamente el mismo tanto para overlap-add como para overlap-save. Para filtrar toda la señal se deben calcular N_b convoluciones circulares, y a esto deben sumarse las N_{FFT} operaciones necesarias para calcular la TDF de la respuesta impulsiva del filtro $h[n]$. Por lo tanto,

$$N_P = N_b N_{PR} + N_{FFT} = \left\lceil \frac{N_x + P - 1}{N - (P-1)} \right\rceil (2N \log_2(N/2) + 3N) + N \log_2(N/2) - N/2 \qquad (4)$$

Como en la sección anterior, $N = 2^\gamma 2^{\lceil \alpha \rceil}$, donde $\alpha = \log_2 P$, y $N_x = 2^\beta P$. De acuerdo a estas hipótesis, el

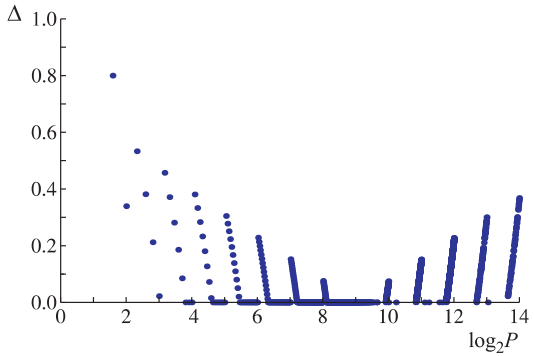


Figura 2: Diferencia entre el número de operaciones por unidad de tiempo $N^B_{O/S}$ óptimo y el correspondiente a $\gamma = 3$.

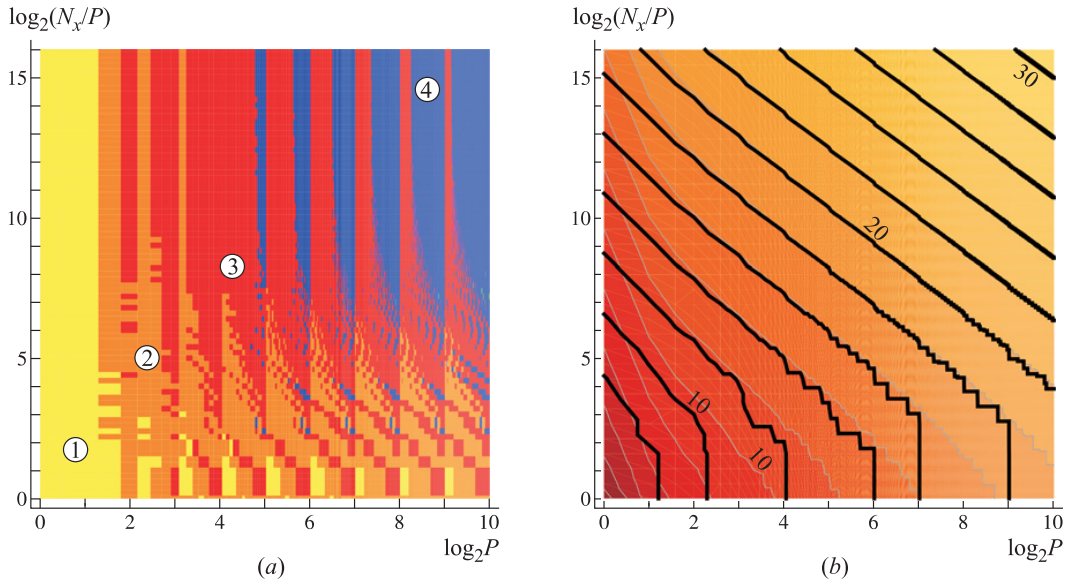


Figura 3: (a) Valor óptimo del exponente γ, y (b) curvas de nivel del total de operaciones para $\gamma = \hat{\gamma}$ (óptimo) (superficie de color) y para $\gamma = 3$ (trazo negro), en función de la longitud P del filtro y de $\beta = \log_2(N_x/P)$.

número de productos es

$$N_P(P,\beta,\gamma) = 2^{\gamma+\lceil\alpha\rceil}\left[-\frac{3}{2}+\gamma+\lceil\alpha\rceil+(1+2\gamma+2\lceil\alpha\rceil)\left[\frac{2^\beta P+P-1}{2^{\gamma+\lceil\alpha\rceil}-(P-1)}\right]\right]. \tag{5}$$

En la Figura 3(a) se grafica el valor de $\gamma = \hat{\gamma}$ que minimiza (5) en función de P y de $\beta = \log_2(N_x/P)$, para P variando entre 1 y 1024, y β entre 0 y 16, que corresponde a valores de N_x entre 1 y $2^{10} \times 2^{16} \approx 67 \times 10^6$ muestras (aproximadamente, el tamaño de un archivo de audio de una canción). El valor de $\hat{\gamma}$ varía entre 1 y 4, como se indica en la figura. Nuevamente, $\hat{\gamma}$ no es una función continua de P y β, y por lo tanto es difícil encontrar una aproximación razonable, pero al igual que en la sección anterior, para este rango de valores de P un valor subóptimo es $\gamma = 3$. En la Figura 3(b) se comparan las curvas de nivel de $\log_2 N_P(P,\beta,\gamma)$ para $\gamma = \hat{\gamma}$ (superficie de color) y las de $\log_2 N_P(P,B,3)$, para $\gamma = 3$ (trazo negro). Se observa que hay algunas diferencias para valores bajos de P y β, que disminuyen para rangos de P y β donde se justifica el empleo de la convolución por bloques. Como es natural, la diferencia en el número total de operaciones necesarias para calcular la convolución lineal y la convolución por bloques disminuye a medida que crece N_x.

5. Conclusiones

En este trabajo se determinó el tamaño subóptimo de la FFT para minimizar el número de operaciones por unidad de tiempo y el número de operaciones totales en algoritmos de filtrado por bloques. Se encontró que para filtros de longitud L hasta $2^{14} \sim 2^{16}$ el tamaño N subóptimo de la FFT es aproximadamente 8 veces la longitud L redondeada a la potencia de 2 más próxima.

Agradecimientos

El autor agradece el apoyo de la SGCyT de la UNS (PGI 24ZK21).

Referencias

[1] C. S. Burrus, *Block realization of digital filters*, IEEE Trans. Aud. Electroac., AU-20 (1972), pp. 230-235.

[2] C. W. Cooley, J. W. Tukey, *An Algorithm for the Machine Computation of Complex Fourier Series*, Mathematics of Computations, 19 (1965), pp. 297-301.

[3] P. S. R. Diniz, E. A. B. da Silva, S. L.Ñetto, *Digital Signal Processing: System Analysis and Design*, CUP, 2010.

[4] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, 2009.

[5] A. V. Oppenheim, A. S. Willsky, I. T. Young, *Signals and Systems*, Prentice-Hall, 1983.

[6] J. G. Proakis, D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, Prentice-Hall, 2006.

ALGORITMO CONJUNTO KALMAN–WAVELETS PARA EL FILTRADO DEL RUIDO EN SEÑALES

Guillermo La Mura†, Ricardo O. Sirne‡ y Eduardo P. Serrano†‡

†Centro de Matemática Aplicada, Universidad Nacional de San Martín, 25 de Mayo y Francia, 1650 San Martín, Buenos Aires, Argentina, guillermo.lamura@gmail.com
‡Departamento de Matemática, Facultad de Ingeniería, Universidad de Buenos Aires, Av. Paseo Colón 850, 1063 Ciudad Autónoma de Buenos Aires, Argentina, rsirne@fi.uba.ar
†‡Escuela Superior Técnica del Ejército "General Manuel N. Savio", Instituto de Enseñanza Superior del Ejército, Av. Cabildo 15, 1426 Ciudad Autónoma de Buenos Aires, Argentina – Centro de Matemática Aplicada, Universidad Nacional de San Martín, 25 de Mayo y Francia, 1650 San Martín, Buenos Aires, Argentina, eduardo.eduser@gmail.com

Resumen: En este trabajo proponemos un algoritmo para el filtrado de ruido en señales que, combinando el método de Kalman con el procesamiento con wavelets, aprovecha las ventajas relativas de ambos métodos.

Palabras claves: *filtrado de ruido, filtro Kalman, wavelets*
2000 AMS Subjects Classification: 42C40

1. INTRODUCCIÓN

Considerando el problema del filtrado del ruido en señales, interesa aprovechar la estimación óptima dada por el filtrado de Kalman y la posibilidad de filtrar usando onditas (wavelets) en el marco del análisis de multirresolución (AMR). Existen varias propuestas para resolver problemas relacionados con el análisis de señales usando onditas y/o filtros de Kalman; esquemas recursivos de filtrado y predicción usando onditas [6], de estimación de movimiento Browniano fraccionario con filtros de Kalman usando onditas [9], con esquema combinado sub-óptimo Kalman–Haar [7]. En este último la metodología de procesamiento propuesta es sub-óptima desde el punto de vista de Kalman, pero aprovecha ambas herramientas en forma conjunta, con bajo costo computacional. En la presentación actual enfocamos la supresión del ruido de medición con un nuevo esquema basado en el diseño de un modelo sistémico, que permite filtrar los coeficientes del desarrollo en serie de onditas.

2. MÉTODOS DE FILTRADO

En esta sección especificamos los dos tipos de filtrado que proponemos combinar, definimos la nomenclatura y sintetizamos los métodos de trabajo requeridos para su aplicación en forma individual.

2.1. MULTIRRESOLUCIÓN Y FILTRADO CON WAVELETS

Consideremos la representación en serie de onditas (wavelets) ortogonales de una señal s_0 en el marco de un análisis de multirresolución (AMR) con función de escala ϕ [4, 8]. Denotamos $V_j \subset V_{j-1}$ con $j \in Z$ a los subespacios encajados del AMR; siendo la ondita ψ ortogonal, resulta $V_{j-1} = W_j \oplus V_j$, donde W_j es el complemento ortogonal de V_j respecto de V_{j-1}. Con este esquema, suponiendo $s_0 \in V_0$:

$$s_0 = r_1 + \cdots + r_J + s_J \quad \text{con} \quad r_j \in W_j \qquad \text{y} \qquad s_J \in V_J \quad \text{para todo } J \in \mathbb{N}, \tag{1}$$

$$r_j(t) = \sum_{k=-\infty}^{\infty} d_{j,k}\, 2^{-j/2}\, \psi(2^{-j}t - k) \quad \text{y} \quad s_J(t) = \sum_{k=-\infty}^{\infty} c_{J,k} 2^{-j/2} \phi(2^{-j}t - k)\,; \tag{2}$$

r_j es la señal residual correspondiente al nivel de resolución j (proyección de s_0 sobre W_j) y s_j la proyección de s_0 sobre V_j. En particular, cuando se usa la ondita de Haar resulta:

$$\phi(t) = 1 \text{ si } t \in [0,1) \quad \text{y} \quad \phi(t) = 0 \text{ si } t \notin [0,1), \text{ con } \psi(t) = \phi(2t) - \phi(2t-1)\,; \tag{3}$$

siendo $\phi(n) = \delta_n$ para $n \in \mathbb{N}$, con $\delta_0 = 1$ y $\delta_n = 0$ si $n \neq 0$.

En este caso los coeficientes se calculan recursivamente mediante:

$$c_{j,k} = 2^{-1/2}(c_{j-1,2k} + c_{j-1,2k+1}) \quad \text{y} \quad d_{j,k} = 2^{-1/2}(c_{j-1,2k} - c_{j-1,2k+1}) \tag{4}$$

donde, si la señal se analiza en bloques de $2^J N$ valores, corresponden $2^{J-j} N$ coeficientes de cada tipo en el nivel j de resolución (j = 1, ..., J). Desde (2) resulta $s_0(k) = \sum_n c_{0,n}\phi(k-n) = \sum_n c_{0,n}\delta_{k-n} = c_{0,k}$; es decir, el proceso inicia tomando los valores de la señal muestreada como coeficientes de nivel 0. En este contexto si la señal $s_0(k) = s_{0,k}$ está contaminada con ruido aditivo v_k , de media nula y desvío σ , con sólo suponer que v_k y v_{k+i} son incorrelacionados para $i \neq 0$ los coeficientes del AMR –según (4)– también resultan con ruido aditivo de igual media y desvío σ . En particular, si el ruido es gaussiano los coeficientes también lo son.

Teniendo en cuenta que la señal tiene un contenido frecuencial en un intervalo [$f_{mín}$, $f_{máx}$], el análisis de multirresolución se realiza para los niveles j = 1, ..., J de modo que s_J sea despreciable desde el punto de vista energético. Entonces, para la reconstrucción, en (1) puede considerarse:

$$s_0 \cong r_1 + \cdots + r_J \ ; \tag{5}$$

la estrategia del filtrado con wavelets consiste en realizar esta reconstrucción calculando las señales residuales aplicando (2) con coeficientes $\bar{d}_{j,k}$ que se obtienen desde los $d_{j,k}$ mediante:

$$\bar{d}_{j,k} = \begin{cases} sg(d_{j,k})L(|d_{j,k}|) & \text{si} \quad |d_{j,k}| \geq \alpha_j \\ 0 & \text{si} \quad |d_{j,k}| < \alpha_j \end{cases} \tag{6}$$

donde L generalmente es una función lineal [5, 8] y α_j es el umbral para el nivel de resolución j . Este filtrado es eficiente cuando la energía de la señal está concentrada en determinados niveles de resolución.

2.2. FILTRADO DE KALMAN

Mediante el método recursivo de Kalman [2, 3] se logra, minimizando el error cuadrático medio, el filtrado óptimo del ruido de un proceso estocástico cuyo modelo es un sistema lineal discreto del tipo:

$$\begin{cases} x_{k+1} &= A_k x_k + B_k u_k + w_k \\ y_k &= H_k x_k + v_k \end{cases}, \tag{7}$$

donde x_k , u_k e y_k representan respectivamente el *estado*, el *control* y la *observación* (*medición*) del sistema en el tiempo discreto $k \in \mathbb{Z}$. En general $A_k \in \Re^{n\times n}$, $B_k \in \Re^{n\times p}$ y $H_k \in \Re^{m\times n}$ son funciones no aleatorias del tiempo; $x, w \in \Re^{n\times 1}$, $y, v \in \Re^{m\times 1}$, $u \in \Re^{p\times 1}$, mientras que w_k y v_k son ruidos blancos gaussianos independientes, con media nula y matrices de covariancia Q_k y R_k respectivamente. El esquema de Kalman es un proceso iterativo de estimación que contempla los siguientes cálculos para $k \in \mathbb{N}$:

$$\text{predicción} \begin{cases} \hat{x}_k^- &= A_{k-1}\hat{x}_{k-1} + B_{k-1}u_{k-1} \\ P_k^- &= A_{k-1}P_{k-1}A_{k-1}^t + Q_{k-1} \end{cases} \text{ y corrección} \begin{cases} K_k &= P_k^- H_k^t [H_k P_k^- H_k^t + R_k]^{-1} \\ \hat{x}_k &= \hat{x}_k^- + K_k(y_k - H_k\hat{x}_k^-) \\ P_k &= (I_n - K_k H_k)P_k^- \end{cases} \tag{8}$$

Siendo $\hat{x}_k^-$ la estimación *a priori* de x_k obtenida en base a las mediciones y_i para tiempo discreto $i < k$, denotaremos $\hat{x}_k$ a la estimación *a posteriori* de x_k cuando se conoce y_k e $I_n \in \Re^{n\times n}$ a la matriz identidad. Para comenzar el procedimiento deben definirse los valores iniciales $\hat{x}_0$ y P_0, para ello es posible elegir $\hat{x}_0 = E[x_0] = 0$ y $P_0 = X_0$; si el sistema dinámico es uniformemente completamente observable y controlable, una condición inicial $P_0 \geq 0$ pierde influencia a medida que crece la cantidad de mediciones procesadas [1].

3. ALGORITMO CONJUNTO KALMAN–WAVELETS

Dado que el filtrado de Kalman y el de wavelets son radicalmente distintos, resulta atractivo combinar ambos métodos con la intención de potenciar sus ventajas para cierto tipo de señales, correspondientes a sistemas lineales modelados según se indica en la sección 2.2.

Trabajando ambos filtrados individualmente, si primero se aplica Kalman y luego se filtra con wavelets cada componente estimada del estado del sistema, el primer filtrado lograría la relación señal/ruido óptima para cada componente, pero esto no asegura dicha optimización para cada residual en las diferentes escalas del AMR. Por otra parte, filtrando primero con wavelets cada componente observada –de y_k según (7)– no es posible asegurar que la reconstrucción obtenida mediante (5) y (6) cumpla con las hipótesis necesarias para la posterior aplicación del procesamiento de Kalman (modelo y tipo de ruido).

Nuestra propuesta consiste en generar –a partir de (7)– un modelo sistémico para los coeficientes del AMR. Esto permite aplicar Kalman en cada nivel de resolución para luego implementar la reconstrucción de cada componente del estado filtrando con wavelet sobre los coeficientes ondita. Para ello, en el sistema lineal (7), cada componente del estado x es una señal del tipo s_0 que se procesa en el marco de un AMR.

Ilustremos el primer paso del modelo sistémico propuesto. Según 2.1, para la ondita de Haar los coeficientes de nivel 0 coinciden con los valores de la señal; por lo tanto dichos coeficientes para cada componente de x se puede agrupar en un vector $C_0 = x$ que cumple con (7). Si las matrices *A,B,H* son constantes y el ruido del sistema (*w*) es despreciable, denotando $U_0^C = Bu$, $V_{0,k}^C = v_k$ y $A_{[0]} = A$ resulta:

$$C_{0,k+1} = A_{[0]}\, C_{0,k} + U_{0,k}^C \;, \quad y_{0,k} = HC_{0,k} + V_{0,k}^C \;; \tag{9}$$

entonces $C_{0,2k+3} = A_{[0]}(\underbrace{A_{[0]}\, C_{0,2k+1} + U_{0,2k+1}^C}_{C_{0,2k+2}}) + U_{0,2k+2}^C = A_{[0]}^2\, C_{0,2k+1} + \underbrace{A_{[0]}\, U_{0,2k+1}^C + U_{0,2k+2}^C}_{a_0}$. Análogamente resulta $C_{0,2k+2} = A_{[0]}^2\, C_{0,2k} + \underbrace{A_{[0]}\, U_{0,2k}^C + U_{0,2k+1}^C}_{b_0}$. Luego aplicando (4) a los vectores, se obtiene:

$$\begin{cases} C_{1,k+1} = A_{[1]}\; C_{1,k} + U_{1,k}^C \\ D_{1,k+1} = A_{[1]}\; D_{1,k} + U_{1,k}^D \end{cases} \quad \text{con} \quad \begin{cases} y_{1,k}^C = H\; C_{1,k} + V_{1,k}^C \\ y_{1,k}^D = H\; D_{1,k} + V_{1,k}^D \end{cases} \tag{10}$$

donde $A_{[1]} = A_{[0]}^2$, $U_{1,k}^C = (a_0 + b_0)/\sqrt{2}$, $U_{1,k}^D = (a_0 - b_0)/\sqrt{2}$ y D_1 es el vector de los coeficientes detalle de las componentes del estado del sistema en el nivel de resolución 1; el ruido se comenta en nota final. El modelo sistémico general para niveles de resolución $j \geq 1$ resulta:

$$\begin{cases} C_{j+1,k+1} = A_{[j+1]}\; C_{j+1,k} + U_{j+1,k}^C \\ D_{j+1,k+1} = A_{[j+1]}\; D_{j+1,k} + U_{j+1,k}^D \end{cases} \quad \text{con} \quad \begin{cases} y_{j+1,k}^C = H\; C_{j+1,k} + V_{j+1,k}^C \\ y_{j+1,k}^D = H\; D_{j+1,k} + V_{j+1,k}^D \end{cases} \tag{11}$$

donde $A_{[j+1]} = A_{[j]}^2$, $U_{j,k}^C = (a_{j-1} + b_{j-1})/\sqrt{2}$, $U_{j,k}^D = (a_{j-1} - b_{j-1})/\sqrt{2}$, $a_j = A_{[j]} U_{j,2k+1}^C + U_{j,2k+2}^C$, $b_j = A_{[j]} U_{j,2k}^C + U_{j,2k+1}^C$, siendo C_{j+1}, D_{j+1} los vectores de los coeficientes del AMR de las componentes del estado del sistema para el nivel de resolución $j+1$. Respecto a los vectores de ruido, para $j \geq 1$ son $V_{j,k}^C = 2^{-1/2}(V_{j-1,2k}^C + V_{j-1,2k+1}^C)$ y $V_{j,k}^D = 2^{-1/2}(V_{j-1,2k}^C - V_{j-1,2k+1}^C)$ que, con las hipótesis dadas en la sección 2.2, son gaussianos con media nula y la misma matriz de covarianza *R* que perturba a las mediciones del sistema.

4. EJEMPLO DE FILTRADO CONJUNTO

Sea un sistema lineal que en estado puro (sin ruido) puede modelarse mediante la ecuación diferencial $z'' + z' + 2z = \text{sen}(2\pi \frac{1}{64} t) + \text{sen}(2\pi \frac{1}{18} t)$ con $z(0) = 0$, $z'(0) = 0$. Si definimos el vector de estado x, tal

que $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ con $x_1 = z$, $x_2 = z'$, para el caso en que la salida se observa perturbada con ruido aditivo gaussiano, el modelado correspondiente resulta ser del tipo:

$$x' = Ax + Bu\ ,\ y = Hx + v\ ,\ A = \begin{pmatrix} 0 & 1 \\ -2 & -1 \end{pmatrix}\ ,\ B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\ ,\ H = (1\ 1),\ u = \begin{pmatrix} 0 \\ f(t) \end{pmatrix}\ ,$$

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{con} \quad f(t) = \text{sen}(2\pi \tfrac{1}{64} t) + \text{sen}(2\pi \tfrac{1}{18} t)\ ,\ R = 0.25\ ,$$

donde se respeta la nomenclatura definida y se supone que –este caso– el ruido tiene desvío de 0.25. El muestreo se realizó a intervalos regulares de tiempo $\Delta t = 1/32$, simulándose la medición de 2^{15} datos.

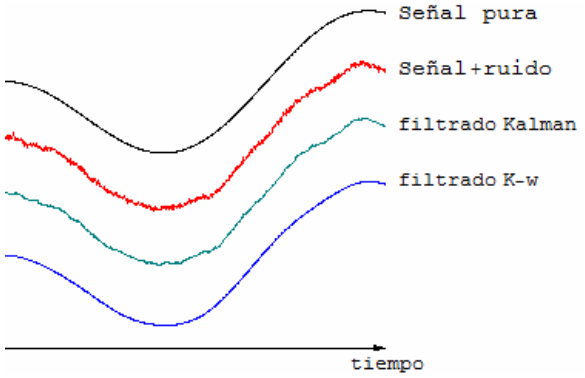


Figura 1: Comparación de las señales (vista local)

En la Fig. 1 se representa una vista local (corto intervalo de tiempo) de la señal pura, la simulada con ruido aditivo, la filtrada con método de Kalman y la estimada mediante el filtrado Kalman-wavelet propuesto.

5. CONCLUSIONES

El modelo propuesto en este trabajo, dado que el filtrado se aplica por niveles, combina las ventajas de ambos métodos resultando eficiente para señales con composición frecuencial concentrada en algunos niveles de resolución. Tal es el caso de la señal expuesta en el ejemplo precedente que concentra el 90% de su energía en tres niveles de resolución, cumpliendo con las suposiciones bajo las cuales se resaltan las ventajas del filtrado combinando Kalman con wavelets.

REFERENCIAS

[1] C.E. D'Attellis, *Estimadores Óptimos y sus Aplicaciones*, CONICET, Argentina, 1981.

[2] R.E. Kalman, *A New Approach to Linear Filtering and Prediction Problems*, Trans. ASME-Journal of Basic Engineering, 82 (1960), pp. 35-45.

[3] R.E. Kalman, R.S. Bucy, *New Results in Linear Filtering and Prediction Theory*, Trans. ASME-Journal of Basic Engineering (1961), pp. 95-108.

[4] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[5] S. Postalcioglu, K. Erkan, E.D. Bolat, *Comparison of Kalman Filter and Wavelet Filter for Denoising*, International Conference on Neural Network and Brain, Beijing, China, Vol. 2 (2005), pp. 951-954.

[6] O. Renaud, J. Starck, F. Murtagh, *Wavelet-Based Combined Signal Filtering and Prediction*, IEEE Trans. on Sytems, Man, and Cybernetics, Part B: Cybernetics, Vol. 35, Issue 6 (2005), pp. 1241-1251.

[7] A. Viegener, R.O. Sirne, E.P. Serrano, M. Fabio, C.E. D'Attellis, *Algoritmo conjunto Kalman–Haar aplicado al procesamiento de señales*, trabajo presentado en el XVII International Symposium on Mathematical Methods Applied to the Sciences (XVII SIMMAC), San José de Costa Rica, 16-19 de febrero de 2010; en revisión para su publicación en la Revista Matemática: Teoría y Aplicaciones.

[8] D.F. Walmut, *An Introduction to Wavelet Analysis*, Birkhaüser, Boston, 2002.

[9] J. Zhao, H. Ma, Z. You, M. Umeda, *Multiscale Kalman Filtering of Fractal Signals Using Wavelet Transform*, Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg (2001), pp. 305-313.

UN MODELO PARA LA ESTIMACIÓN DE LA FUNCIÓN DE ESCALA MULTIFRACTAL UTILIZANDO CASCADAS MULTIPLICATIVAS MULTINOMIALES

Eduardo Serrano[†] and Alejandra Figliola[♭]

[†]*Centro de Matemática Aplicada, ECyT, Universidad Nacional de San Martín, Irigoyen 3100, San Martín, Pcia. de Buenos Aires, Argentina, eduardo.eduser@gmail.com*

[♭]*(1) Instituto del Desarrollo Humano, Universidad Nacional de General Sarmiento, J. María Gutiérrez 1150, Los Polvorines, Pcia. de Buenos Aires, Argentina, afigliol@ungs.edu.ar*

Abstract: Los datos empíricos provenientes de los mercados financieros, tales como índices burstiles, tasas de interés o la variación de los precios de los productos básicos, presentan características autosimilares y multifractales, las que, a su vez, se relacionan con el grado de ineficiencia de los mercados financieros. En consecuencia, el análisis del espectro multifractal o de singularidades de estos sistemas provee de interesante información para conocer el estado de cada mercado, desde un novedoso enfoque. Para estos sistemas se propone un modelo del comportamiento multifractal basado en procesos del tipo Cascadas Multinomiales Multiplicativas. El modelo se aplica en el caso de la serie temporal del índice de precios de las operaciones de la bolsa diaria de Buenos Aires (Merval).

Keywords: *Sistemas dinámicos complejos, Fractales, Análisis de series temporales económicas*
2000 AMS Subject Classification: 37Fxx, 28A80, 91B84

1 INTRODUCCIÓN

Las características multifractales de un sistema quedan totalmente determinadas por su espectro multifractal, $d_f(H)$, el cual es función de los exponentes de Hölder, H, del sistema. Estos exponentes determinan el grado de irregularidad de una función o en otra palabras, cuán rugosa es su gráfica. Tanto H como $d_f(H)$ se relacionan entre sí a través de la función de escala $\eta(q)$, utilizando la Transformada de Legendre y en el caso que ésta sea una función cóncava, [7], [3]:

$$H(q) \;=\; \eta'(q), \tag{1}$$

y

$$d_f(H) \;=\; q\,H \;-\; \eta(q), \tag{2}$$

donde q es el índice y $q \in \mathbb{R}$.
Existen diversos métodos para estimar $\eta(q)$ en el caso de señales provenientes de sistemas naturales, económicos o sociales. Se pueden mencionar los que provienen del análisis wavelet, tal como *Wavelet Transform Modulo Maxima*, [1] o *Wavelet Leaders*, [5] o el *Multifractal Detrended Fluctuation Analysis* (MFDFA), que resulta una generalización del método DFA para calcular el exponente de Hurst, [6]. Es destacable que, utilizando cualquiera de estas metodologías, la función de escala puede construirse a partir de un conjunto de datos discretos. Sin embargo, la estimación numérica presenta detalles delicados, lo que compromete el cálculo estable de $d_f(H)$ y de H via las ecuaciones (1) y (2).
Por otra parte, dado que es conocida la forma analítica de la función de escala para las *cascadas multinomiales multiplicativas*, este trabajo plantea una alternativa para diseñar un modelo de $\eta(q)$ utilizando estas funciones. Una breve síntesis de sus características será presentada en la próxima sección.

2 CASCADAS MULTINOMIALES MULTIPLICATIVAS

Se considerarán una clase de procesos multifractales conocidas como *cascadas multinomiales multiplicativas* (CMM), las cuales resultan análogas a las medidas de Cantor.
Dado un entero, $m \geq 2$, se consideran para cada uno de los $j \geq 0$, $I_{jk} = [k, k+1)/m^j$ con $0 \leq k < m^j$, intervalos disjuntos que cubren $I = [0, 1)$ y verifican que $|I_{jk}| = m^{-j}$.

Sea $\mu([0,1)) = 1$ y $(p_0, p_1, \ldots, p_{m-1})$, conjunto de pesos no negativos, que verifican: $\sum_k p_k = 1$. Entonces, una cascada de medida multinomial μ se define de la siguiente manera recursiva:

$$\mu(I_{j+1,r(k)}) \quad = \quad p_r \mu(I_{j,k}), \tag{3}$$

donde $0 \leq r(k) < m$ y $k \equiv r(k) \mod (m)$.
Para cada paso j, la cascada multinomial puede ser representada en los puntos $x_{jk} = km^{-j}$ a través de la serie finita y_{jk} de longitud m^j, donde:

$$\sum_{k=0}^{j} y_{jk} \quad = \quad \sum_{k=0}^{j} p_0^{\alpha_{j,k,0}} p_1^{\alpha_{j,k,1}} \ldots p_{m-1}^{\alpha_{j,k,m-1}} = (p_0 + p_1 + \cdots + p_{m-1})^j \tag{4}$$

La *función de partición* del sistema $S(q, j)$ es:

$$S(q,j) \quad = \quad m^{-j} \sum_{0 \leq k \leq m^j} y_{jk}^q \quad = \quad m^{-j} \left(p_0^q + \cdots + p_{m-1}^q\right)^j, \tag{5}$$

y el formalismo multifractal asume la relación:

$$S(q,j)(H) \sim m_j^{(1-d_f(H)+Hq)}, \tag{6}$$

donde $\eta = 1 - d_f(H) + Hq$ (para más detalles, ver [4] y [8]). A partir de las Eq. (5) y (6) es fácil demostrar

$$\eta(q) \quad = \quad 1 - \log_m \left(p_0^q + \cdots + p_{m-1}^q\right). \tag{7}$$

La función $\eta(q)$ puede estimarse a partir de los pesos p_k. Una importante propiedad de esta relación es que las asíntotas de la gráfica, para $q \to \pm\infty$, permiten estimar de manera eficiente el menor y el mayor de los pesos. Además, se puede obtener la expresión analitica para los exponentes Hölder, $H(q)$, y para el espectro multifractal, $d_f(H)$, de manera analítica:

$$H(q) \quad = \quad \frac{d\eta(q)}{dq} \quad = \quad - \frac{p_0^q \log_m(p_0) + \cdots + p_{m-1}^q \log_m(p_{m-1})}{\left(p_0^q + \cdots + p_{m-1}^q\right)} \tag{8}$$

y

$$d_f(H) \quad = \quad \log_m \left(p_0^q + \cdots + p_{m-1}^q\right) - q \, \frac{p_0^q \log_m(p_0) + \cdots + p_{m-1}^q \log_m(p_{m-1})}{\left(p_0^q + \cdots + p_{m-1}^q\right)}. \tag{9}$$

Algunos autores han propuesto un modelo binomial para el estudio de la multifractalidad en las series del índice de la variación diaria del flujo de ríos, [2].

3 El modelo. Aplicación a un caso de la economía

Se muestra como ejemplo la serie del índice promedio de las operaciones de la Bolsa de Buenos Aires diarias, Merval ($\{x_i\}, i = 1...N$, con $N = 1304$ datos), correspondiente al período comprendido entre el 5 de marzo de 2004 y el 5 de marzo de 2009. Se trabaja con el índice del retorno de precios $rt_i = log(x_i/x_{i+1})$. La Figura 1 muestra la serie $\{rt_i\}$ en el período indicado, mientras que la Figura 2 grafica la la correspondiente función de escala $\eta(q)$. Los datos fueron obtenidos en el sito web http://www.mscibarra.com.

Se ajustó la función $\eta(q)$, según la ecuación (7) con un modelo de cinco pesos p_i no decrecientes: $a = p_1$, $c = p_5$, definidos por las asíntotas $\eta(q)$ y los tres valores "centrales" e iguales, $p_2 = p_3 = p_4 = b$:

$$\eta(q) = 1 - \log_6 \left(a^q + b^q + b^q + b^q + c^q\right) = 1 - \log_6 \left(a^q + 3b^q + c^q\right). \tag{10}$$

En particular, podemos observar que

$$\lim_{q \to -\infty} \frac{\eta(q)}{q} = -log_m(P_0) = -log_6(a) \tag{11}$$

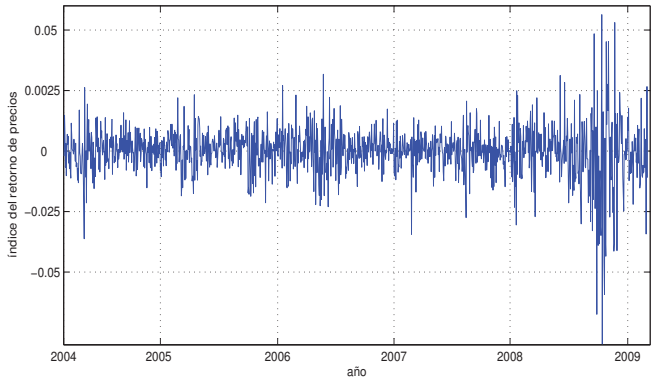


Figure 1: Serie del índice de retorno de precios para el MERVAL, entre marzo de 2004 a marzo de 2009

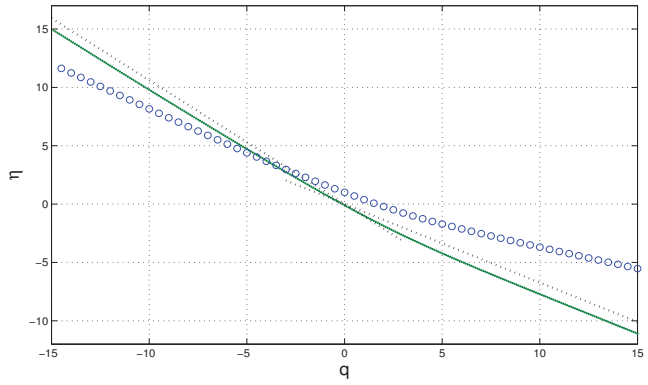


Figure 2: Función de escala de los datos de la Figura 1. La curva con puntos corresponde a la estimación hecha con el modelo CMM y en círculos vacíos a la del MFDFA. Las rectas en líneas punteadas constituyen las asíntotas, cuya pendiente se muestra en las ecuaciones (11) y (12).

y

$$\lim_{q\to\infty} \frac{\eta(q)}{q} = -log_m(P_{m-1}) = -log_6(c). \tag{12}$$

El rango de los exponentes Hölder resulta, escrito en términos de las asíntotas,

$$\Delta H_{CMM} =\mid log_6(a) - log_6(c) \mid=\mid log_6(\frac{a}{c}) \mid . \tag{13}$$

a	b	c	ΔH_{MFDFA}	ΔH_{CMM}
0.1493	0.1838	0.2993	0.4090	0.3889

Tabla I: Valores de los parámetros de ajuste y cálculo del grado de multifractalidad del sistema, utilizando MFDFA directamente y utilizando el modelo CMM.

4 CONCLUSIONES

En este trabajo se propone tomar como modelo de la función de escala, la cual es representativa del comportamiento multifractal de un sistema, a la cascada multinomial multiplicativa. Tanto la expresión de los exponentes de Hölder como el espectro multifractal, para la CMM tienen expresiones analíticas conocidas, lo que equivale a calcularlos directamente y evitar de ese modo el correspondiente cálculo de la derivada de la función de escala.

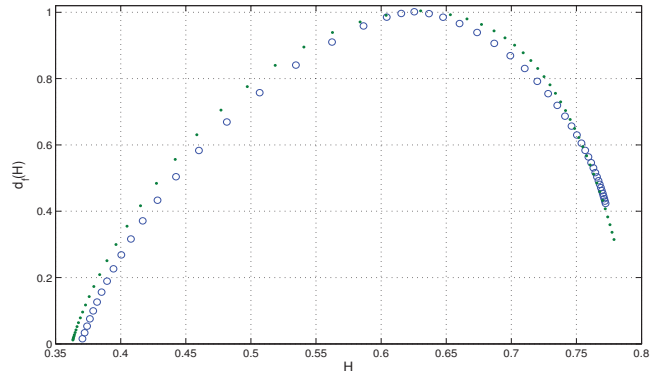


Figure 3: Espectro multifractal para la serie retornos de precios del Merval. Con círculos vacíos la estimación del MFDFA y con puntos la estimación del modelo de CMM

Asimismo, la medida de la multifractalidad ΔH(como medida de la ineficiencia del mercado) puede ser calculada directamente a partir de los valores del ajuste, tal como se mostró en la ecuación (13) y de esta manera, evitar el uso de la Transformación de Legendre para encontrar el rango de los exponentes Hölder. En el caso en que $a/c \approx 1$, $\Delta H \approx 0$ que se corresponde con una función $\eta(q)$ cuya gráfica es una recta y el sistema tiene características monofractales.

AGRADECIMIENTOS

E. Serrano agradece a la Universidad de San Martín y A. Figliola a la Universidad de General Sarmiento y al CONICET.

REFERENCES

[1] A. ARNEODO, A. ARGOUL, J.F. MUZY, M. TABARD AND E. BACRY, *Beyond classical multifractal analysis using wavelets: uncovering a multiplicative process hidden in the geometrical complexity of diffusion limited aggregates*, Fractals 1 (1995) pp. 629-646.

[2] E. KOSCIELNY-BUNDE, J. W. KANTELHARDT, P. BRAUN, A. BUNDE, S. HAVLIN, *Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies*. Journal of Hydrology, 322 (2006) 120–137.

[3] K. FALCONER,*Techniques in fractal geometry*, John Wiley and Sons Ltd, New York, (1997).

[4] A. FIGLIOLA AND E. SERRANO *A Model of the Multifractality for the Stock Market Indices*, Procceding of V Meeting on dynamics of social and economic systems, (Dyses 2010) September 20-25, Benevento, Italy (2010)

[5] B. LASHERMES, S. JAFFARD AND P. ABRY, Wavelet Leaders in multifractal Analysis in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal*, IEEE Acoustics, Speech and Signal 4, (2005) pp. 161–164

[6] J. W. KANTELHARDT, S. A. ZSCHIEGNER, E. KOSCIELNY-BUNDE, S. HAVLIN, A. BUNDE AND H. E. STANLEY, *Multifractal detrended fluctuation analysis of nonstationary time series*, Physica A 316 (2002) pp. 87–114.

[7] S. MALLAT, *A Wavelet Tour of Signal Processing* , Academic Press, San Diego, (2009).

[8] E. SERRANO AND A. FIGLIOLA, *Wavelet leaders: a new method to estimate the multifractal singularity spectra*, Phys. A388 (2009) pp. 2793–2805.

[9] L. ZUNINO, A. FIGLIOLA, B. M. TABAK, D. G. PÉREZ, M. GARAVAGLIA AND O. A. ROSSO, *Multifractal structure in Latin-American market indices*, Chaos, Solitons & Fractals 6443, 41 (5) (2009) pp. 2330–2339.

CARACTERIZACIÓN DE LA FRECUENCIA INSTANTÁNEA EN SEÑALES TIPO PASA-BANDA

M. Fabio†, A. Aragón† y E. Serrano†

†*Centro de Matemática Aplicada, Universidad Nacional de San Martín, Martín de Irigoyen N° 3100 (1650), San Martín, Buenos Aires, Argentina ,*
mfabio@unsam.edu.ar

Resumen: las denominadas *señales tipo pasa-banda* poseen una particular estructura que brinda la eficiente estimación de su amplitud y su frecuencia instantáneas, en el sentido dado por la Transformada de Hilbert, a partir de datos de muestreo. Esta estructura sugiere descomponer una señal dada, de potencia o energía finita, en un Análisis de Multirresolución, utilizando especiales wavelets tipo pasa-banda. Entonces, las proyecciones en cada espacio wavelet son señales tipo pasa-banda y pueden analizarse en consecuencia.

En este trabajo deducimos las fórmulas de estimación y detallamos la metodología de análisis para descomponer la señal en ondas con espectros instantáneos bien definidos. Exponemos resultados numéricos y sugerimos ulteriores refinamientos.

Palabras clave: *transformada de Hilbert, frecuencia y amplitud instantánea, señales tipo pasa-banda, wavelets.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Existen diversos criterios para representar una señal de energía finita $y \in L^2(\mathbb{R})$ en la forma:

$$y(t) \;=\; g(t)\cos(\phi(t)) \tag{1}$$

de modo que la función $g(t)$ y derivada de la fase $\phi'(t)$ se interpreten como la *amplitud* y la *frecuencia instantánea*, respectivamente, [1], [2], [3].

El criterio formal, unívoco, se fundamenta en la *función analítica* [2] asociada a la señal:

$$y_a(t) \;=\; y(t) + \imath\tilde{y}(t) \;=\; A(t)\,\exp(\imath\phi(t)), \tag{2}$$

donde $\tilde{y} = \mathcal{H}y$ denota la transformada de Hilbert de la señal, [8]. Se definen la amplitud instantánea $A(t) = (y^2(t) + \tilde{y}^2(t))^{1/2}$ y la frecuencia instantánea $\omega(t) = \phi'(t)$, donde la derivada existe.

Esta definición normaliza la representación (1) y es válida para cualquier señal de energía finita. En general, el cálculo de la función $\phi(t)$ no es trivial: hay singularidades y cambios de fase donde el módulo se anula y, por sobre todo, $\omega(t)$ varía en el rango del espectro de la señal.

Por estas razones, es aconsejable descomponer previamente la señal $y(t)$ en apropiadas funciones y_λ simples o elementales:

$$y(t) = \sum_\lambda y_\lambda(t) \;= \sum_\lambda A_\lambda(t)\cos(\phi_\lambda(t)). \tag{3}$$

y calcular las amplitudes y frecuencias, $A_\lambda(t)$ y $\phi'_\lambda(t)$ en cada componente. La descomposición en *funciones modales intrínsecas*, asociadas a la transformada de Hilbert-Huang [1] o en *chirplets* [4], son algunas alternativas propuestas en la literatura. Remarcamos que el espectro instantáneo no es lineal y que cada esquema organiza la información tiempo-frecuencia con estructuras distintas.

En este trabajo proponemos la descomposición de la señal como superposición de funciones elementales del tipo pasa-banda, ortogonales y asociadas a un esquema de Análisis de Multirresolución, [2], [3]:

$$y(t) = \sum_j q_j(t) \;= \sum_j A_j(t)\cos(\phi_j(t)). \tag{4}$$

donde q_j es la proyección ortogonal de la señal sobre el espacio wavelet W_j. Aplicamos una especial clase de wavelets, de modo que es posible estimar las amplitudes y la frecuencias instantáneas para cada j, utilizando los valores de muestreo de las proyecciones.

2. SEÑALES TIPO PASA-BANDA

Una señal y se dice *tipo pasa-banda* si $\widehat{y}(\omega) \in C^1(\mathbb{R})$ y $|\widehat{y}(\omega)|$ está soportada en una banda bilátera $\omega_0 - \Omega/2 < |\omega| < \omega_0 + \Omega/2$, con $\omega_0 > \Omega$.

Estas señales son suaves y decaen rápidamente, lo que posibilita el eficiente cálculo de su transformada de Hilbert. En tal sentido, se demuestra que si y es una señal del tipo pasa-banda [8], entonces:

$$y(t) = \alpha(t)\cos(\omega_0 t) + \beta(t)\sin(\omega_0 t), \tag{5}$$

$$\mathcal{H}(y)(t) = \alpha(t)\sin(\omega_0 t) - \beta(t)\cos(\omega_0 t) \tag{6}$$

donde $\alpha(t),\ \beta(t) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, son diferenciables y de banda limitada, con $|\widehat{\alpha}(\omega)|$ y $|\widehat{\beta}(\omega)|$ soportadas en el intervalo $[-\Omega/2,\ \Omega/2]$.

3. CÁLCULO DE LA FRECUENCIA INSTANTÁNEA

Para deducir la fórmula de estimación de la frecuencia instantánea demostramos, en primer lugar, que si $y \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ es una función diferenciable, entonces, donde $\tilde{y}^2(t) + y^2(t) > 0$:

$$\phi'(t) = \frac{\tilde{y}'(t)y(t) - \tilde{y}(t)y'(t)}{\tilde{y}^2(t) + y^2(t)} \tag{7}$$

Este resultado se aplica en particular si y es una señal pasa-banda. En consecuencia, combinando los resultados anteriores (5), (6) y (7), demostramos que:

Teorema 1 *Bajo las hipótesis precedentes, resulta:*

$$\phi'(t) = \omega_0 + \frac{\beta(t)\alpha'(t) - \beta'(t)\alpha(t)}{\alpha^2(t) + \beta^2(t)}, \tag{8}$$

donde $\alpha^2(t) + \beta^2(t) > 0$.

Observamos que la frecuencia instantánea varía respecto de la frecuencia de referencia ω_0, dependiendo del segundo término de la expresión (8).

La aplicación de esta fórmula requiere conocer las funciones α y β y sus derivadas, pero dado que son de banda limitada, $|\omega| \leq \Omega/2 < \omega_0$, pueden ser exactamente determinadas a partir de los valores de muestreo de la señal y. Para ello definimos los puntos de muestreo $t_n = \frac{\pi}{2\omega_0}\, n$, con $n \in \mathbb{Z}$. Entonces:

$$y(t_{2k}) = \alpha(t_{2k})\,\cos(\pi k) = (-1)^k\,\alpha(t_{2k}), \tag{9}$$

$$y(t_{2k+1}) = \beta(t_{2k+1})\,\sin\left(\frac{\pi}{2}(2k+1)\right) = (-1)^k\,\beta(t_{2k+1}) \tag{10}$$

e interpolamos ambas funciones en la red $\{t_n\}$.

De manera análoga pueden estimarse los valores de las derivadas. Sin embargo, la suavidad de las funciones α y β permiten una eficiente aproximación de las mismas mediante diferencias:

$$\phi'(t_n) \cong \omega_0 \left[1 + \frac{\beta(t_n)(\alpha(t_{n+1}) - \alpha(t_{n-1})) - \alpha(t_n)(\beta(t_{n+1}) - \beta(t_{n-1}))}{\pi\,(\alpha^2(t_n) + \beta^2(t_n))}\right], \tag{11}$$

fórmula discreta basada exclusivamente en los valores de muestreo $\{y(t_n)\}$.

4. WAVELETS TIPO PASA-BANDA

Los autores desarrollan en [7] una clase de wavelets tipo pasa-banda $\psi^{(m)}$, con $\widehat{\psi}^{(m)}$ soportada en la banda bilátera $(\pi - \pi/m) < |\omega| < 2(\pi + \pi/m)$, con $m \geq 3$. Su diseño está basado en las funciones de Malvar, [3]. Cada familia $\{\psi_{jk}^{(m)}(t) = 2^{j/2}\,\psi^{(m)}(2^j t - k),\ j,k \in \mathbb{Z}\}$ constituye una base ortonormal de $L^2(\mathbb{R})$. Para cada j la frecuencia central es $\omega_{0j} = \pi(3 + 1/m)2^{j-1}$.

El algoritmo de análisis y síntesis se realiza en forma eficiente utilizando la transformada rápida de Fourier (FFT), [7].

Seleccionado el parámetro m, y dada la señal $y \in L^2(\mathbb{R})$, se calculan sus proyecciones q_j sobre los subespacios W_j. Dado que éstas son funciones pasa-banda se aplica, para cada j, el método propuesto en la sección anterior para calcular el correspondiente espectro instantáneo $(A_j(t), \phi_j'(t))$. Finalmente, se obtiene la representación en ondas definida en (3).

En la práctica, el análisis se reduce a los espacios W_j y a los intervalos temporales donde se concentra la energía de la señal.

5. EJEMPLO

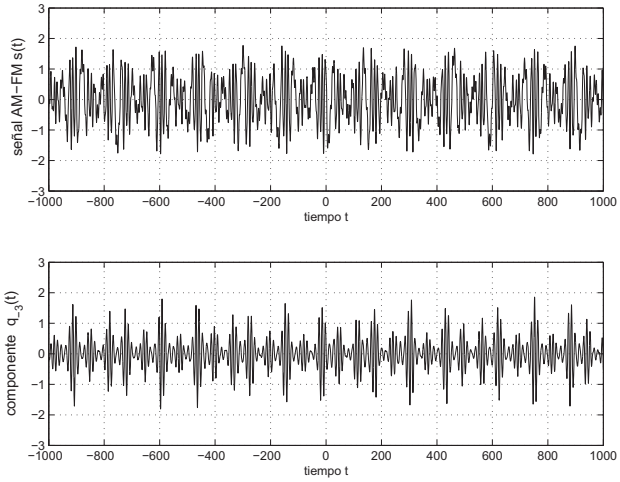


Figura 1: Señal AM-FM $s(t)$ y componente $q_{-3}(t) \in W_{-3}$

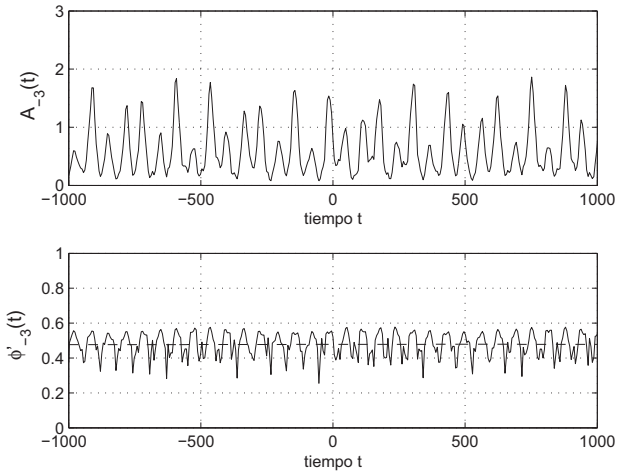


Figura 2: Amplitud $A_{-3}(t)$ y frecuencia instantánea $\phi'_{-3}(t)$

Para ilustrar el método utilizamos una clásica señal AM-FM de prueba, [9]:

$$s(t) \quad = \quad (1 + 0.5\cos(2\pi t/150))\sin\left(2\pi/17 + 2.5\cos(2\pi/64)\right) + 0.3\sin(2\pi/7)$$

Las propiedades de esta señal de prueba pueden verse en la citada bibliografía de referencia.

Seleccionamos la familia de wavelets pasa-banda generada por $\psi^{(4)}$. El análisis de la señal s nos brinda las cuatro componentes principales q_j, proyecciones sobre los subespacios W_j para $j = -5, -4, -3, -2$, respectivamente. En este ejemplo, tomamos la componente de mayor energía, q_3. Ésta y la señal de prueba

se exhiben en la Figura 1. Los datos se muestrean con paso $\Delta t = 1/4seg$. El análisis se limita al intervalo temporal $|t| \leq 1000seg$.

Aplicando el método expuesto en la sección 3., estimamos la amplitud $A_{-3}(t)$ y la frecuencia instantáneas $\omega_{-3}(t) = \phi'_{-3}(t)$, expuestas en la Figura 2. Para este nivel, $j = -3$ y para $m = 4$ la banda de localización es $[0.094\pi, 0.31\pi]$ y la frecuencia de referencia $\omega_{-3,0} = 0.2\pi$.

En el intervalo de análisis, el valor medio de la frecuencia instantánea calculada es $\bar{\phi}' = 0.152\pi$.

6. Conclusiones

En este trabajo se proponemos un eficiente método numérico para determinar la amplitud y la frecuencia instantáneas de funciones del tipo pasa-banda.

Se aplica en el contexto de un Análsis de Multirresolución generado por wavelets ortogonales pasa-banda. Primeramente se realiza el análisis de la señal y luego se calculan las amplitudes y las frecuencias de cada proyección, tipo pasa-banda, sobre los subespacios wavelet.

Las correspondientes fórmulas de cálculo se aplican utilizando los datos de muestreo de las proyecciones.

La síntesis final representa la señal como superposición de ondas ortogonales con un espectro instantáneo bien definido.

Las experiencias numéricas realizadas y, en particular en el ejemplo expuesto, muestran que las frecuencias instantáneas oscilan respecto de su media, próxima a la frecuencia de referencia. Este fenómeno se refleja en la fórmula (8).

Si la componente pasa-banda fuese casi-monocromática, la frecuencia instantánea debería ser casi constante, y las variaciones en el segundo término de la citada fórmula, poco significativas.

Esta observación sugiere la posibilidad de refinar la localización en frecuencia de las funciones elementales. En particular, diseñar y aplicar paquetes de wavelets del tipo pasa-banda. Este objetivo motiva una futura línea de desarrollo.

Referencias

[1] N.E. Huang et al., *The empirical mode decomposition and the Hilber spectrum for non-stationary tyme series analysis* , Proc. R. Soc. Lond. A 454, 1998.

[2] S. Mallat, *A Wavelet Tour of Signal Processing, The Sparse Way*, Academic Press,Elsevier, 2009.

[3] Y. Meyer, *Wavelets, Algorithms and Applications*, SIAM, Philadelphia, 1993.

[4] Y. Meyer, *Oscillating Pattern in Image Processing and Nonlinear Evolution Equations* , American Mathematical Society, Providence, 2001.

[5] E. Serrano, A. Figliola, *Littlewood-Paley spline wavelets: a simple and efficient tools for signal and image processing in industrial applications*, Proceedings in Applied Mathematics and Mechanics, Wiley InterScience, PAMM, Vol. 7, Issue 1 (2008) pp. 1040313-1040314.

[6] E. Serrano, M. Fabio, A. Aragón, *Paquetes de Wavelets Analíticas*, II Congreso de Matemática Aplicada, Computacional e Industrial, Argentina, 2009.

[7] E. Serrano, M. Fabio, *Diseño de funciones elementales combinando la transformada wavelet y la transformada de Hilbert* , UMA 2010, Tandil, Argentina, 2010.

[8] A. D. Poularikas, *Transforms and Applications*, CRC Press, 2010.

[9] Yi et al., *Research on iterated Hilbert Transformand its applications in mechanical fault diagnostic* , Mechanical System and Signal Processing, Vol. 22 (2008), pp. 1967-1980.

UNA ENTROPÍA BASADA EN WAVELET LEADERS Y SU APLICACIÓN A SERIES DE DATOS FINANCIEROS

M. Rosenblatt[†], E. Serrano[♭] y A. Figliola[†]

[†]*Instituto de Desarrollo Humano, Universidad Nacional de General Sarmiento, Los Polvorines, Argentina, mrosen@ungs.edu.ar, afigliola@ungs.edu.ar, www.ungs.edu.ar*
[♭]*Centro de Matemática Aplicada, Universidad Nacional de San Martín, San Martín, Argentina, eserrano@unsam.edu.ar, www.unsam.edu.ar*

Resumen: En este trabajo presentamos un nuevo cuantificador de la regularidad local de una señal: la *entropía wavelet leaders puntual*. Definimos esta nueva medida de la regularidad combinando el concepto de entropía, proveniente de la teoría de la información y de la mecánica estadística, con los coeficientes wavelet leaders. Además establecemos su relación con uno de los exponentes de regularidad más conocido, el exponente Hölder puntual. Por último, aplicamos esta metodología a una serie de datos financieros de la bolsa de USA, registrados en el período 2007-2010, a fin de comparar la evolución temporal de la *entropía wavelet leaders puntual* y del exponente Hölder puntual.

Palabras clave: *regularidad local, exponente Hölder puntual, entropía, wavelet leaders*
2000 AMS Subject Classification: 65T60 - 26A16 - 28D20 - 37M10

1. INTRODUCCIÓN

En este trabajo proponemos una novedosa aproximación para el análisis de la dinámica de las series temporales: el estudio de la evolución de la regularidad local de estas señales por medio de una nueva medida, la *entropía wavelet leaders puntual*.

Diversos cuantificadores han sido propuestos para medir la regularidad local de una función. El más conocido es el exponente Hölder puntual, que se define en cada $x_0 \in Dom(f) \subseteq \mathbb{R}$, donde f es una función localmente acotada, como

$$H_f(x_0) = \sup_{0\leq\alpha<+\infty} \{\alpha : f \in \mathcal{C}^\alpha(x_0)\} \tag{1}$$

La función f está en la clase $\mathcal{C}^\alpha(x_0)$ si existen $C > 0$ y un polinomio $P_{x_0}(x)$ de grado menor que α tales que, en un entorno de x_0: $|f(x) - P_{x_0}(x)| < C\,|x - x_0|^\alpha$.

Este exponente detecta las singularidades de una función. Cuanto más cerca de cero está el valor del exponente, más irregular es el gráfico de la función en ese punto, mientras que las porciones suaves están asociadas a exponentes altos. En [4], S. Jaffard formula una nueva caracterización del exponente Hölder puntual a través del estudio del decaimiento de los denominados coeficientes *wavelet leaders*, que se calculan a partir de los supremos locales de los coeficientes wavelet de una señal $f \in L^2(\mathbb{R})$, concentrando su información y reorganizando su estructura.

A partir del cálculo de los coeficientes wavelet leaders de una señal dada por 2^m datos, construimos, en cada punto x_0, una distribución de probabilidades discreta $\mathcal{P}_{x_0} = \{\rho_j : j = 1, .., m\}$ y definimos la *entropía wavelet leaders puntual* usando la entropía de Shannon, [7]. Un antecedente de combinar el concepto de entropía con el de coeficientes wavelet puede encontrarse en [1], [8]. El nuevo cuantificador que definimos tiene características diferentes de los definidos con coeficientes wavelet y además refina lo formulado en [1] y [8] donde las entropías se definen en intervalos no solapados de determinado tamaño, limitado por la cantidad de datos de la señal y su contenido frecuencial.

En este trabajo mostramos que la *entropía wavelet leaders puntual* alcanza valores muy cercanos al máximo cuando el exponente Hölder puntual toma valores próximos a cero, lo que indica que este cuantificador también detecta las singularidades de una función.

Como ejemplo, presentamos una aplicación de la metodología propuesta para analizar la dinámica de una serie de datos del índice bursátil de USA en 2007-2010.

2. WAVELET LEADERS Y REGULARIDAD LOCAL

El estudio de la regularidad de una función f en un punto x_0 está relacionado con el comportamiento de la amplitud de la transformada wavelet en un entorno de x_0.

Suponiendo que la wavelet madre ψ ortogonal es una función $\mathcal{C}^r$, con r momentos nulos y derivadas de rápido decaimiento, para $r \in \mathbb{N}$ suficientemente grande, la familia $\mathcal{F} = \left\{2^{j/2}\psi(2^j x - k)\right\}_{j,k\in\mathbb{Z}}$ resulta una base ortonormal de $L^2(\mathbb{R})$ y $f \in L^2(\mathbb{R})$ puede representarse como

$$f(x) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} c_{j,k}\psi(2^j x - k)$$

donde $c_{j,k} = \left\langle f, 2^j\psi(2^j x - k)\right\rangle$ son los *coeficientes wavelet* de f. Si ψ está esencialmente localizada en el intervalo $[0, 1]$ entonces $c_{j,k}$ posee información relevante de f en el intervalo diádico $I_{j,k} = \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right)$. Para ampliar estos tópicos pueden consultarse [5] y [6].

En [3], S. Jaffard encuentra que la pertenencia de f a la clase $\mathcal{C}^\alpha(x_0)$ está correlacionada con el decaimiento de los coeficientes wavelet $c_{j,k}$ y prueba que si $f \in \mathcal{C}^\alpha(x_0)$ entonces existe $C > 0$ tal que, para todo $j \geq 0$, los coeficientes wavelet decaen siguiendo la desigualdad: $|c_{j,k}| \leq C\ 2^{-j\alpha}(1 + \left|2^j x_0 - k\right|)^\alpha$. Posteriormente, propone una nueva formulación de esta desigualdad en términos de los coeficientes *wavelet leaders* [4], que se definen para una función f acotada, en cada nivel j y para cada x_0, como:

$$d_j(x_0) = \sup_{I_{l,h}\subset 3I_j(x_0)} |c_{l,h}| \tag{2}$$

donde $I_j(x_0)$ es el único intervalo diádico que contiene a $x_0 \in \mathbb{R}$ en el nivel j y $3I_j(x_0)$ es el intervalo dilatado. Entonces, si f es una función acotada en la clase $\mathcal{C}^\alpha(x_0)$, $\alpha > 0$, se tiene que existe $C > 0$ tal que $\forall j > 0$:

$$d_j(x_0) \leq C\ 2^{-j\alpha} \tag{3}$$

y además, si f es uniformente Hölder, el exponente Hölder puntual de f verifica

$$H_f(x_0) = \liminf_{j\to+\infty} \frac{\log(d_j(x_0))}{\log(2^{-j})} \tag{4}$$

3. ENTROPÍA WAVELET LEADERS

La entropía es un concepto que surge de la termodinámica y es reformulado por la teoría de la información para cuantificar la información promedio contenida en un mensaje, [2]. Si X es una fuente de información con símbolos $\{x_1, \ldots, x_m\}$ que ocurren con probabilidad $\{\rho_1, \ldots, \rho_m\}$ se define la entropía de la distribución de probabilidades $\mathcal{P} = \{\rho_1, \ldots, \rho_m\}$ como:

$$S(\mathcal{P}) = -\sum_{i=1}^m \rho_i log_2(\rho_i) = \sum_{i=1}^m \rho_i log_2(1/\rho_i)$$

y se define $\rho_i log_2(\rho_i) = 0$ si $\rho_i = 0$, [7].

$S(\rho_1, \ldots, \rho_m)$ es una función continua que verifica que $0 \leq S(\rho_1, \ldots, \rho_m) \leq log_2(m)$ y alcanza su máximo cuando la distribución es equiprobable.

Para definir la *entropía wavelet leaders puntual* construimos una distribución de probabilidades $\mathcal{P}_{x_0}$ asociada a cada x_0 en el dominio de $f \in L^2(\mathbb{R})$, una función acotada: $\mathcal{P}_{x_0} = \{\rho_1, \ldots, \rho_m\}$ donde

$$\rho_i = \frac{d_i^2(x_0)}{\sum_{j=1}^m d_j^2(x_0)} \quad si \quad d_i(x_0) \neq 0 \quad y \quad \rho_i = 0 \quad en\ otro\ caso \qquad (i = 1, \ldots, m)$$

con $d_i(x_0)$ los coeficientes wavelet leaders asociados a x_0 en los niveles $j = 1, \ldots, m$, definidos en (2).

A partir de estos conceptos:

Definición 1 *Sea $f \in L^2(\mathbb{R})$ una función acotada. La entropía wavelet leaders de f en x_0 es*

$$S_f(x_0) = S(\mathcal{P}_{x_0}) = -\sum_{i=1}^m \rho_i log_2(\rho_i)$$

y se define $\rho_i log_2(\rho_i) = 0$ si $\rho_i = 0$

Si los coeficientes wavelet de mayor amplitud, en un entorno de x_0, están concentrados en el nivel j más alto entonces los coeficientes wavelet leaders en x_0 son todos iguales y en consecuencia $S_f(x_0)$ alcanza su valor máximo $log_2(m)$. Por el contrario, $S_f(x_0) = 0$ si los coeficientes wavelet, en un entorno de x_0, son nulos o bien si los coeficientes wavelet no nulos están concentrados en el nivel j más bajo.

Usando (3) y (4) y que $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_m$ (pues $(d_j(x_0))_j$ es una sucesión decreciente) probamos el siguiente resultado que establece una relación inversa entre el exponente Hölder puntual y la *entropía wavelet leaders puntual*.

Proposición 1 *Sea $H = H_f(x_0)$ el exponente Hölder de f en x_0 y $f \in \mathcal{C}^H(x_0)$ una función uniformemente Hölder, entonces*

$$4^{-(m-1)H} log_2\left(m\, 4^{-(m-1)H}\right) \leq S_f(x_0)$$

si m es suficientemente grande.

Como corolario se deduce que si $H_f(x_0)$ toma valores cercanos a cero entonces la *entropía wavelet leaders puntual* $S_f(x_0)$ alcanza valores próximos al máximo $log_2(m)$. También se puede probar que si existe $C > 0$ tal que $d_i(x_0) \cong C\, 2^{-iH}$ para todo $i = 1, \ldots, m$, entonces

$$S_f(x_0) \leq -log_2(1 - 4^{-H}) + 6H4^{-H}.$$

En este caso la entropía $S_f(x_0)$ está cerca de cero cuando el exponente $H_f(x_0)$ es muy grande.

4. Aplicación y conclusiones

Analizamos una serie de datos del índice bursátil de USA (SPX), que registra diariamente el Morgan Stanley Capital Index (http://www.mscibarra.com), con un total de 1045 datos del período 2007-2010.

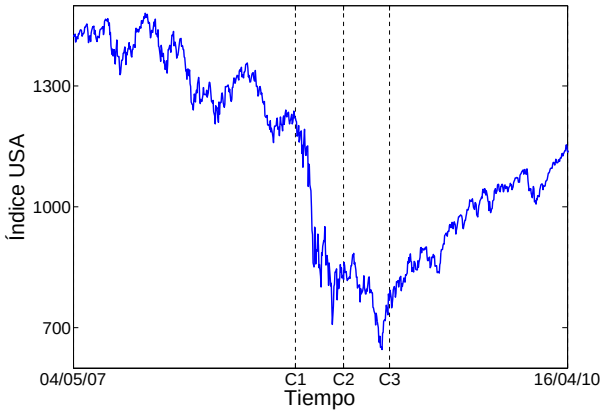


Figura 1: Evolución temporal del índice bursátil de USA. C1=*01/09/08* C2=*15/12/08* C3=*25/03/09*

Calculamos la serie de los retornos logarítmicos del índice bursátil $r(t) = log(x(t+1)/x(t))$, con $x(t)$ el valor del índice bursátil en el tiempo t. A través de métodos numéricos estimamos el exponente Hölder puntual y la entropía wavelet leaders puntual de $r(t)$, utilizando una wavelet spline cúbica para calcular los coeficientes wavelet hasta el nivel $m = 10$, vía análisis de multirresolución [5]. Para estimar el exponente Hölder puntual usamos una regresión lineal suponiendo que $log(d_j(x_0)) \approx log(C) + H_f(x_0) log(2^{-j})$.

Las figuras 2 y 3 revelan que la variación temporal de estos cuantificadores refleja la evolución de la crisis financiera. En el período más crítico de la crisis bursátil -entre C1 y C2- el exponente Hölder toma valores muy cercanos a cero mientras que la entropía toma valores próximos al máximo $log_2(10)$, lo que indica que estos cuantificadores distinguen el fenómeno.

La serie original alcanza un mínimo en el período comprendido entre C2 y C3. Sin embargo, durante este período, ambos cuantificadores muestran una recuperación de la regularidad de la señal, lo que puede interpretarse como un precursor de un cambio en la tendencia de la crisis.

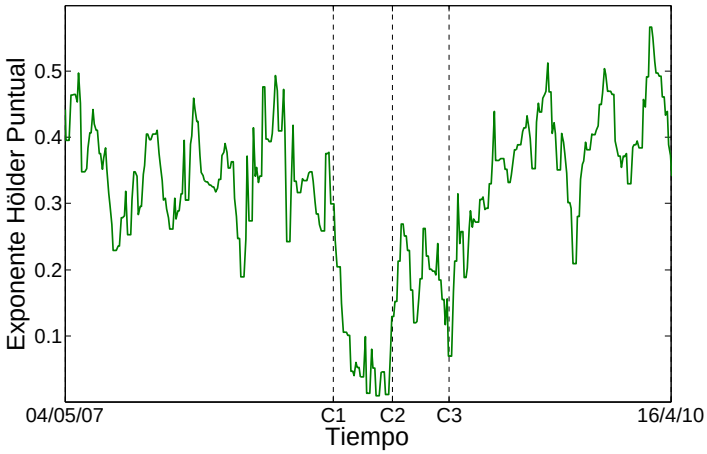


Figura 2: Evolución temporal del exponente Hölder puntual. C1=*01/09/08* C2=*15/12/08* C3=*25/03/09*

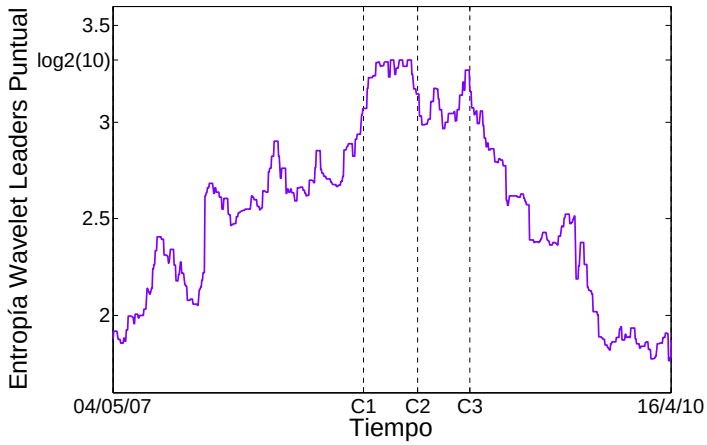


Figura 3: Evolución temporal de la entropía wavelet leaders puntual. C1=*01/09/08* C2=*15/12/08* C3=*25/03/09*

En este trabajo proponemos una nueva forma de entropía, que se correlaciona con la regularidad local de la señal, procurando iniciar una nueva línea de desarrollo. Además, aplicamos esta metodología a una serie de datos del índice bursátil de USA, intentando hacer un aporte a la comprensión de la dinámica de los mercados financieros.

REFERENCIAS

[1] S. BLANCO, A. FIGLIOLA, R. QUIAN QUIROGA, O.A. ROSSO, E. SERRANO, *Time-frequency analysis of electroencephalogram series (III): information transfer function and wavelets packets*, Phys. Rev. E, 57 (1998), pp.932-940.

[2] T.M. COVER, J.A. THOMAS, *Elements of Information Theory, 2nd Edition*, J. Wiley; New York, 2006.

[3] S. JAFFARD, *Exposants de Hölder en des points donnés et coefficients d'ondelettes*, C.R.A.S. Série I, 308 (1989), pp.79-81.

[4] S. JAFFARD, *Wavelet techniques in multifractal analysis*, Proc. Sympos. Pure Math., AMS, 72 (2004), pp.91-151.

[5] S. MALLAT, *A Wavelet Tour of Signal Processing, The Sparse Way, 3rd Edition*, Academic Press: Burlington, 2009.

[6] Y. MEYER, *Ondelettes et opérateurs*, Hermann, 1990.

[7] C.E. SHANNON, *A mathematical theory of communication*, Bell Syst. Technol. Journal, 27 (1948), pp.379-423 y 623-656.

[8] M.E. TORRES, L. GAMERO, C.E. D'ATTELLIS, *A multirresolution entropy approach to detect epileptic form activity in the EEG*, IEEE Workshop on Nonlinear Signal and Image Processing, II (1995), pp.791-794.

SUB-WAVELETS: UNA NUEVA FAMILIA DE FUNCIONES ELEMENTALES EN EL CONTEXTO DE UN ANÁLISIS DE MULTIRRESOLUCIÓN

M. Fabio† y E. Serrano†

†*Centro de Matemática Aplicada, Universidad Nacional de San Martín, Martín de Irigoyen N 3100 (1650), San Martín, Buenos Aires, Argentina ,*
mfabio@unsam.edu.ar

Resumen: En este trabajo presentamos una novedosa familia de funciones elementales casi monocromáticas, suaves e infinitamente oscilantes que denominamos sub-wavelets. Se incluye el diseño, basado en la aplicación en cada subespacio wavelet, de una descomposición del tipo Littlewood-Paley y la metodología de aplicación que conduce a la descomposición de señales en ondas de frecuencia instantánea bien definida.

Palabras clave: *transformada de Hilbert, frecuencia y amplitud instantánea, wavelet de Malvar, sub-wavelets*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Consideramos una *señal* en $L^2(\mathbb{R})$ en la que conviven estructuras oscilantes, fenómenos transitorios y parásitos sin estructura definida. Suponemos que el objetivo del procesamiento es caracterizar patrones oscilantes significativos, presuntamente asociados a fenómenos físicos.

Para ello, se pretende una descomposición de la señal en la forma:

$$f(t) = \sum_{\lambda\in\Lambda} A_\lambda(t)\cos(\phi_\lambda(t)) + r(t),$$

donde cada componente $f_\lambda(t) = A_\lambda(t)\cos(\phi_\lambda(t))$, representa una *onda* con un patrón oscilante característico, con amplitud $A_\lambda(t)$ y fase $\phi_\lambda(t)$, variando en el tiempo con cierta regularidad. La componente residual $r(t)$ resume la tendencia y las componentes de muy baja frecuencia, los transitorios y, en general, la información sin estructura.

Este esquema de descomposición general plantea las siguientes cuestiones:

- ¿Cómo deben ser las componentes $f_\lambda(t)$ para ser consideradas *ondas* desde un punto de vista físico?
- En consecuencia, ¿cómo debe realizarse la descomposición de la señal en forma eficiente y en un apropiado marco analítico?

Se pretende además:

- Las componentes $f_\lambda(t)$ sean suaves, oscilantes, bien localizadas en intervalos T_λ y con *frecuencias instantáneas* $\omega_\lambda(t)$ bien definidas.
- La interferencia entre distinas ondas $f_{\lambda_1}(t)$, $f_{\lambda_2}(t)$ sea mínima (ondas casi-ortogonales).

Revisamos previamente varias soluciones propuestas en la literatura:

- Sintetizar las ondas a partir de una descomposición primaria, usando wavelets, wavelet packets, funciones de Gabor o de Malvar-Wilson, [2], [3].
- Aplicar un esquema del tipo Littlewood-Paley, [4], [5].
- Descomponer la señal en término de funciones modales intrínsecas, mediante la transformada de Hilbert-Huang, [1], [6].

A partir de estos desarrollos, en este trabajo, diseñamos una nueva alterntiva.

2. METODOLOGÍA PROPUESTA

Partimos de la función analítica:

$$F_\lambda(t) \;=\; f_\lambda(t) + \imath\mathcal{H}f_\lambda(t) = A_\lambda(t)(\exp(\imath\phi_\lambda(t)))$$

donde $\mathcal{H}$ representa la transformada de Hilbert.

Entonces, $A_\lambda(t) \geq 0$ es la *amplitud instantánea* y $\omega_\lambda(t) = \phi'_\lambda(t)$ la *frecuencia instantánea*.

Utilizando apropiadas wavelets ortonormales podemos descomponer la señal

$$f(t) \;=\; \sum_{j\in J}\sum_{k\in K} c_{jk}\,\psi_{jk}(t) \;+\; R(t) \;=\; \sum_{j\in J} f_j(t) \;+\; R(t)$$

donde J y K son finitos y resumen la energía de la misma. Una alternativa para refinar la localización en frecuencia es descomponer, en cada subespacio wavelet W_j, la señales $f_j(t)$, en la forma:

$$f_j(t) \;=\; \sum_m f_{jm}(t)$$

donde $f_\lambda(t) = f_{jm}(t)$ son funciones casi-monocromáticas, con frecuencia instantánea precisa.

Para ello, en este trabajo diseñamos novedosas funciones elementales, *sub-wavelets*, que realizan una descomposición del tipo Littlewood-Paley en cada subespacio W_j, [3], [5].

Esta operación asegura que las componentes $f_{jm}(t)$ se asemejen a funciones modales intrínsecas, con fase $\phi(t)$ casi-lineal.

3. WAVELETS DE MALVAR

El diseño propuesto se realiza en el contexto de un *Análisis de Multirresolución* generado por wavelets del tipo de Malvar-Wilson-Meyer, [3]. Para las señales dadas por un conjunto finito de datos las operaciones de análisis y de síntesis se realizan en el dominio de la frecuencia mediante la aplicación de un eficiente algoritmo basado en la transformada rápida de Fourier, (FFT).

Dada la partición de la recta real en intervalos

$$[a_j, a_{j+1}],\;\; \text{con}\;\; \cdots < a_{-1} < a_0 < a_1 < \cdots,$$

donde, $j \in \mathbb{Z}$, $l_j = a_{j+1} - a_j$ y $\alpha_j > 0$ suficientemente pequeños, se diseñan las ventanas $w_j(t) = w(2^{-j}t)$, que verifican las siguientes condiciones:

- $w_j(t) = 1$ si $a_j + \alpha_j \leq t \leq a_{j+1} + \alpha_{j+1}$
- $w_j(t) = 0$ si $t \leq a_j - \alpha_j$ ó $t \geq a_{j+1} - \alpha_{j+1}$
- $w_j^2(a_j + \nu) + w_j^2(a_j - \nu) = 1$ si $|\nu| \leq \alpha_j$
- $w_{j-1}(a_j + \nu) = w_j(a_j - \nu)$ si $|\nu| \leq \alpha_j$

La familia de funciones

$$u_{jk}(t) = \sqrt{\frac{2}{l_j}}\, w_j(t) \cos\left(\frac{\pi}{l_j}(k + 1/2)(t - a_j)\right),\; j \in \mathbb{Z}, k \in \mathbb{N}_0,$$

constituye una base ortonormal de $L^2(\mathbb{R}_{\geq 0})$. Obtenemos una segunda clase, análoga, reemplazando los cosenos por senos.

En consecuencia, tomando $\omega > 0$ como variable y $a_j = 2^j\pi,\; \alpha_j = 2^j r,\; 0 < r \leq \pi/3$, contamos con dos bases ortonormales de $L^2(\mathbb{R}_{\geq 0})$, en el dominio frecuencial:

$$u_{jk}(\omega) = \sqrt{\frac{2^{1-j}}{\pi}}\, w_j(\omega) \cos\left(2^j(k+1/2)(\omega - 2^j\pi)\right),\ j \in \mathbb{Z}, k \in \mathbb{N}_0,$$

$$v_{jk}(\omega) = \sqrt{\frac{2^{1-j}}{\pi}}\, w_j(\omega) \sin\left(2^j(k+1/2)(\omega - 2^j\pi)\right),\ j \in \mathbb{Z}, k \in \mathbb{N}_0,$$

y extendiendo por paridad e imparidad, respectivamente, para $\omega < 0$:

$$\left\{\frac{1}{2}(u_{jk} + v_{jk}),\ \frac{1}{2}(u_{jk} - v_{jk})\ j \in \mathbb{Z}, k \in \mathbb{N}_0\right\}$$

resulta una base ortonormal de $L^2(\mathbb{R})$ y concluimos que:

$$\left\{\sqrt{\frac{1}{2\pi}}\, w_j(\omega)\, e^{-i(k+1/2)\omega},\ j, k \in \mathbb{Z}\right\}$$

es también una base ortonormal de $L^2(\mathbb{R})$, [3].

Tomando $\alpha = \pi/m,\ m \geq 3$, y la correspondiente ventana para $j = 0$, $w_\alpha(\omega)$, definimos la *wavelet madre*:

$$\widehat{\psi}(\omega) \ = \ w_\alpha(\omega)\, e^{-i\omega/2} \ = \ |\widehat{\psi}(\omega)|\ e^{-i\omega/2},$$

nula si $|\omega| \geq [\pi - \alpha, 2\pi + 2\alpha]$. En el dominio del tiempo, la familia

$$\psi_{jk}(t) \ = \ 2^{j/2}\, \psi(2^j t - k),\ j, k \in \mathbb{Z}$$

es una base ortonormal de wavelets de $L^2(\mathbb{R})$. Entonces, si $f_j(t) \in W_j$,

$$\widehat{f_j}(\omega) \ = \ 2^{-j/2} w_\alpha(\omega/2^j) \sum_k c_{jk}\, e^{-i\omega(k+1/2)/2^j},$$

siendo $c_{jk} \in l^2(\mathbb{Z})$ los coeficientes wavelets.

4. SUB-WAVELETS

Proponemos en cada subespacio W_j un refinamiento utilizando *sub-ventanas*, que realizan la partición:

$$w_\alpha(\omega) \ = \ w_\alpha^{(1)}(\omega) \ + \ \cdots \ + \ w_\alpha^{(m)}(\omega),$$

una descomposición del tipo Littlewood-Paley, [4], [5].

Para $\alpha = \frac{\pi}{m+2}$ y $m \in \mathbb{N}_{>2}$, las sub-ventanas $w_\alpha^{(p)}(\omega)$ están localizadas en torno a la frecuencia

$$\omega_{mp} = \pi(1 + \frac{p}{m+2}),\ 1 \leq p \leq m$$

La Figura 1 muestra el gráfico de estas sub-ventanas, en el nivel $j = -1$, $m = 7$ y $1 \leq p \leq 7$.

Las *sub-wavelets* $\psi_\alpha^{(p)}(t)$ se definen como:

$$\widehat{\psi}_\alpha^{(p)}(\omega) \ = \ w_\alpha^{(p)}(\omega)\, e^{-i\omega/2}$$

El gráfico de algunas de ellas, para el nivel $j = -1$, $m = 7$ y $1 \leq p \leq 3$, se exhibe en la Figura 2. La colección de sub-wavelets resulta sobrecompleta en cada W_j.

Entonces, descomponemos:

$$\widehat{f_j}(\omega) \ = \ \sum_{p=1}^{m} \sum_{k \in Z} c_{jk}\, \widehat{\psi}_\alpha^{(p)}(\omega) \ = \ \sum_{p=1}^{m} f_j^{(p)}(\omega)$$

y las funciones $f_j^{(p)}(t)$ se asemejan a funciones modales intrínsecas y son casi-ortogonales.

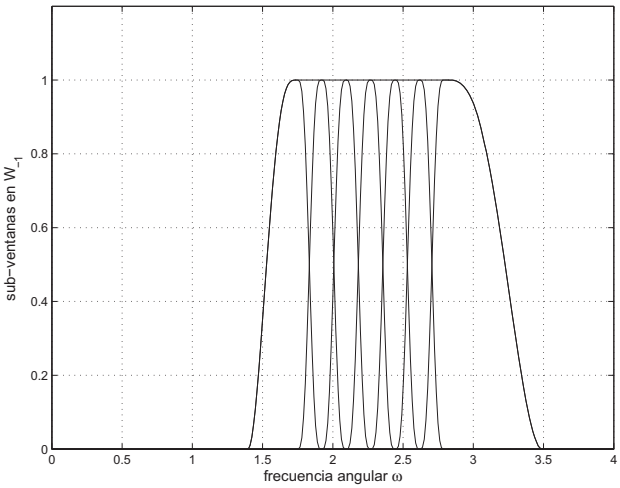


Figura 1: sub-ventanas, $j = -1$, $m = 7$ y $1 \leq p \leq 7$

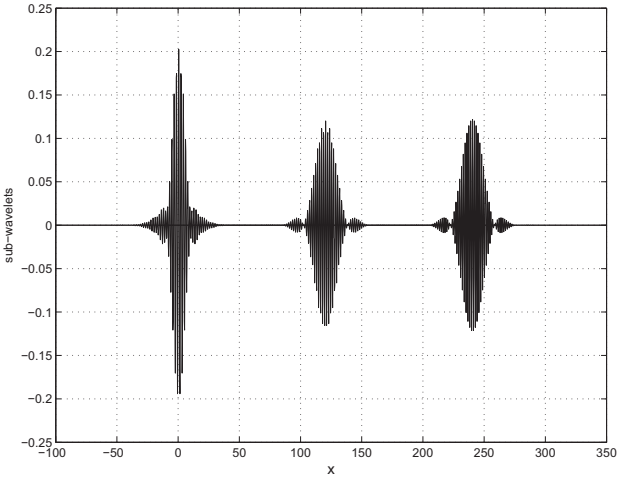


Figura 2: sub-wavelets, $j = -1$, $m = 7$ y $p = 1, 2, 3$

5. CONCLUSIONES

En este trabajo presentamos una nueva familia de funciones elementales en el contexto de un Análsis de Multirresolución. Las mismas descomponen las proyecciones sobre los subespacios wavelet en ondas casi-monocromáticas, mejorando la precisión en frecuencia.

REFERENCIAS

[1] N.E. HUANG ET AL., *The empirical mode decomposition and the Hilber spectrum for non-stationary tyme series analysis* , Proc. R. Soc. Lond. A 454, 1998.

[2] S. MALLAT, *A Wavelet Tour of Signal Processing, The Sparse Way*, Academic Press,Elsevier, 2009.

[3] Y. MEYER, *Wavelets, Algorithms and Applications*, SIAM, Philadelphia, 1993.

[4] Y. MEYER, *Oscillating Pattern in Image Processing and Nonlinear Evolution Equations* , American Mathematical Society, Providence, 2001.

[5] E. SERRANO, A. FIGLIOLA, *Littlewood-Paley spline wavelets: a simple and efficient tools for signal and image processing in industrial applications*, Proceedings in Applied Mathematics and Mechanics, Wiley InterScience, PAMM, Vol. 7, Issue 1 (2008,) pp. 1040313-1040314

[6] E. SERRANO, M. FABIO, A. ARAGÓN, *Paquetes de Wavelets Analíticas*, II Congreso de Matemtica Aplicada, Computacional e Industrial, Argentina, 2009.

VOLTAGE ENVELOPE, NOISE AND HILBERT TRANSFORM

Federico Muiño[♭], Maximiliano Carabajal[♭], Marcela Morvidone[♭, †] and Carlos D'Attellis[♭, †]

[♭]*Facultad Regional Buenos Aires, Universidad Tecnológica Nacional, Mozart 2300, C1407IVT Buenos Aires, Argentina*

[†]*Centro de Matemática Aplicada, Universidad Nacional de San Martín. Martín de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina, cdattellis@yahoo.com.ar*

Abstract: The fluctuation of voltage is one of the important power quality events. Different methods have been proposed for estimating the voltage envelope, but the presence of noise is, in general, not considered. A method for estimating the envelope in presence of noise, based on the Hilbert transform and a low-pass filter, is presented. The results obtained from a real signal measured from an arc furnace are shown.

Keywords: *Hilbert transform, flicker, voltage fluctuation*

1 INTRODUCTION

Voltage fluctuations can be described as systematic variations or random variations in the voltage envelope. The fluctuation of voltage is one of the important power quality events due to the effects of electronic and control systems, and in the light flicker. There are several sources of voltage flickers as arc furnaces, fans, pumps, lifts, switching of powers factor capacitors, large motors [1], [6].

Different methods have been proposed for estimating the magnitude and frequency of flicker. The IEC [2] and IEEE [3] standards recommend the square demodulation, a method used in demodulation of AM signals, which consists in tracking the flicker envelope by squaring the input voltage signal. Other methods proposed are Fast Fourier Transform [16], Least Absolute Value [17], Kalman filters [11], Wavelet transform [8], Teager Energy Operator [4]. Hilbert transform is also used [5], [18], [14], [15], and, in particular, using Prony analysis and Hilbert Transform [10].

Recently, the performance of several flicker detecting methods were compared [9]. The core of flicker analysis is to track the envelope of voltage signal, that is, the instantaneous amplitude. Then, an important characteristic of the algorithms proposed is their on-line behavior; the faster is the estimation of the voltage envelope values, the better is the on-line behavior.

Another aspect of the problem is the presence of noise. In [19] this problem is pointed out and solved using the Hilbert transform for estimating the flicker envelope and the wavelet transform for extracting other noises contained in the voltage flicker. A signal obtained from an arc furnace shows that the high frequency noise can not be ignored. However, the influence of noise in the methods previously cited is not considered.

2 ENVELOPE ESTIMATION USING THE HILBERT TRANSFORM

2.1 ESTIMATING THE ENVELOPE

In this section we review some results concerning the estimation of the envelope of a discrete signal.

The Hilbert transform is used in signal processing to derive the analytic representation of a signal $x[n]$. The analytic representation of a signal is well known for continuous-time signals [7] and it is also defined for discrete signals as

$$z[n] = x[n] + i\mathcal{H}x[n],$$

where $\mathcal{H}x[n]$ denotes the discrete Hilbert transform of the sequence $x[n]$ (see [13]). This representation allows a straightforward identification of the envelope of an amplitude modulated signal. An amplitude modulated signal is modeled by:

$$x[n] = a[n]\cos(\omega n), \tag{1}$$

where the frequency content of $a[n]$ has an upperbound less than ω. In this conditions, Bedrosian theorem for discrete signals [13] states that:

$$\mathcal{H}x[n] = a[n]\mathcal{H}\cos(\omega n),$$

which turns into $\mathcal{H}x[n] = a[n]\sin(\omega n)$. Now, the analytic representation of the signal takes the simple form

$$z[n] = a[n]e^{i\omega n},$$

and the amplitude (or the envelope) $a[n]$ is easily obtained from $a[n] = |z[n]|$.

For the sake of completeness, we include Bedrosian theorem as stated in [13]:

Theorem 1 *Suppose that $z_1[n]$ and $z_2[n]$ are complex sequences with discrete-time Fourier transforms $Z_1(e^{i\omega})$and $Z_2(e^{i\omega})$. Then*

$$\mathcal{H}(z_1 z_2)[n] = z_1[n]\mathcal{H}(z_2)[n]$$

if there exists a nonnegative number $\sigma < \pi$ such that

$$Z_1(e^{i\omega}) = 0, \ \textit{for } 0 < \sigma < |\omega| < \pi, \ \textit{and } \ Z_2(e^{i\mu}) = 0, \ \textit{for } 0 < |\mu| \leq \sigma < \pi.$$

2.2 HILBERT FILTER

In this section we describe the Hilbert filter in more detail. The discrete Hilbert transform $\mathcal{H}x[n]$ of the sequence $x[n]$ is defined in the frequency domain as [12]

$$(\mathcal{H}x)(\omega) = -i\,\mathrm{sgn}(\omega)X(\omega), \tag{2}$$

where $X(\omega)$ is the discrete Fourier transform of x:

$$X(\omega) = (\mathcal{F}x)(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n}.$$

From equation (2), the transfer function of the Hilbert transform for discrete signals is

$$H(\omega) = \begin{cases} -i, & 0 < \omega < \pi \\ 0, & |\omega| = \pi \\ i, & -\pi < \omega < 0. \end{cases} \quad \Rightarrow \quad H(\omega) = -i\,\mathrm{sgn}[\sin(\omega)] = G(\omega)\,e^{i\frac{\pi}{2}},$$

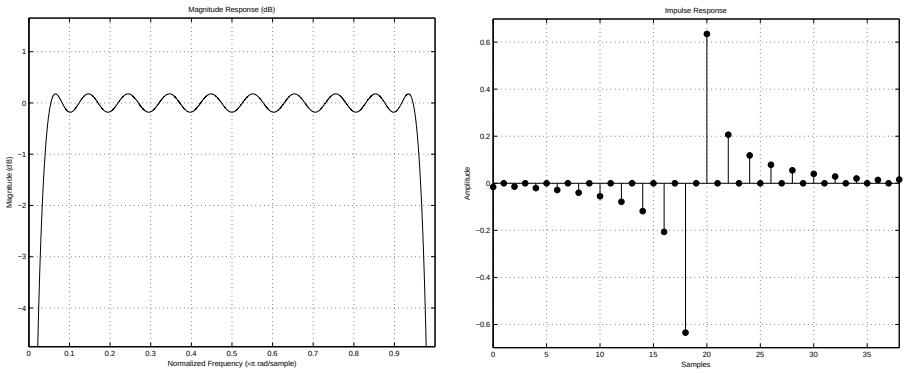
with $G(\omega) = -\mathrm{sgn}[\sin(\omega)]$. The discrete time representation of the Hilbert filter is easily obtained from this expression. In fact, $G(\omega)$ is an odd function whose Fourier transform reads

$$G(\omega) = \frac{4}{\pi}\sum_{m=0}^{\infty}\frac{1}{2m+1}\sin[(2m+1)\,\omega].$$

Denoting $h(k) = \mathcal{F}^{-1}[H(\omega)]$ the inverse Fourier transform of $H(\omega)$, we have

$$h(k) = \frac{2}{\pi k}\sin k\frac{\pi}{2}, \ \ k \geq 0, \ \text{ and } \ h(-k) = -h(k).$$

Figure 1 shows the $|H(\omega)|$ and the impulse responses $h[k]$ of the Hilbert filter used in this work, which is described in detail in Section 4.



(a) Magnitude response (b) Impulse response

Figure 1: Hilbert Filter

3 A SIGNAL FROM AN ARC FURNACE

In this paper, the performance of the proposed model on tracking the voltage flicker signal envelope is examined with a signal coming from real measurements. It is a typical AC arc furnace application in a steel plant. This arc furnace is served from a 13.8 kV bus. The measured signal is one of the phase voltages and it was sampled at a sampling frequency of 1000 Hz.

Since this signal is distorted by noise that comes from making physical measurements, a serious issue is the robustness of the method for estimating flicker. Previous works hardly consider this problem.

As an example, we make some comments on Prony algorithm which has been used in this problematic [10]. Prony algorithm is good at system identification provided that the available samples come from a signal completely predictable and free of any randomness. Under these conditions, it is a good alternative for obtaining mathematical models in the form of damped complex exponentials from a small number of samples. This type of representation allows a straightforward calculation of the Hilbert transform of the signal [10]. However, when the signals have some degree of randomness like signals immersed in noise, Prony algorithm is very unstable and it requires a large amount of samples to achieve an acceptable approximation. This method has some other drawbacks: it is very difficult to estimate the optimal number of exponentials to use in the approximation, the calculation involves two pseudo-inverse matrices whose systems are poorly conditioned, which increases the instability, and finally it has a high computational cost.

4 NUMERICAL RESULTS

We present numerical results on the flicker estimation of the arc furnace signal described in the previous section. The method is implemented in Simulink from Matlab (see Figure 2). The Hilbert filter is a FIR, all zeros filter of order 39, linear phase, whose magnitude and impulse response are shown in Figure 1. As a first step, the envelope is estimated, then a low-pass filter is applied to minimize the influence of noise. This is a FIR equiripple filter of order 42, with cutoff frequency 60 Hz.

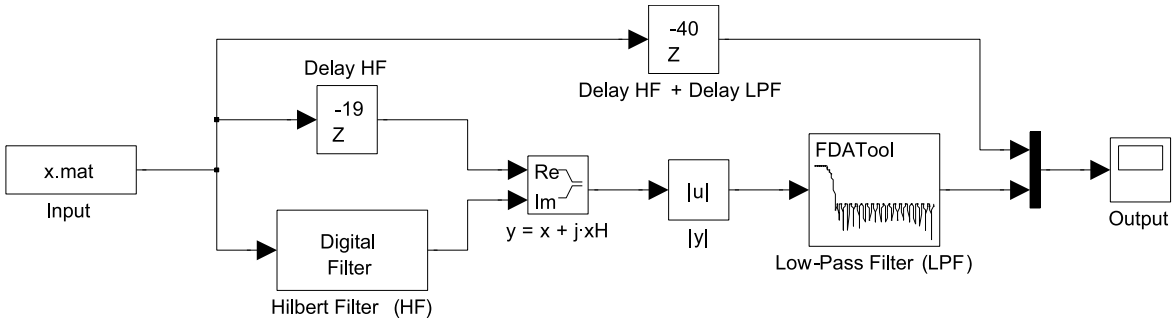


Figure 2: Block diagram of the flicker estimator

Because of the fact that the signal passes through two FIR filters, there is a delay time in the tracking processes. However, as the two filters have a linear phase response, these have a constant group delay response. Therefore, this time delay is constant for all frequencies and it can be calculated. In this case, it resulted in a total delay of 40 samples, and it represents 0.04 sec at a sampling frequency of 1000 Hz.

In Figure 3 the estimations of the envelope corresponding to the diagram of Figure 2 are shown. There are two intervals of 0.5 sec: [16.5,17] and [17,17.5].

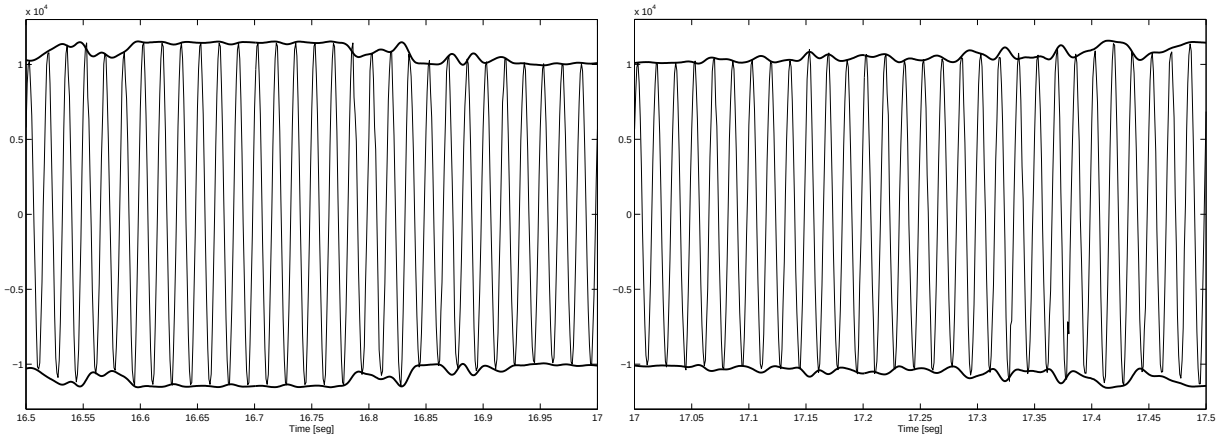


Figure 3: The signal and its estimated envelope.

5 CONCLUSIONS

We have presented a method for estimating signal flicker in presence of noise using the Hilbert transform. We test this technique on a real world signal produced by an arc-furnace, showing the method robustness.

REFERENCES

[1] *IEC 38*, 1983. IEC Standard Voltages.

[2] *IEC Standard 61000-4-15*, 1999. Flickermeter: Functional and Design Specifications.

[3] *IEEE Standard 1453*, 2004. IEEE Recommended Practice for Measurement and Limits of Voltage Fluctuations and Associated Light Flicker on AC Power Systems.

[4] T. K. Abdel-Galil, E. F. El-Saadany, and M. M. Salama. Energy operator for on-line tracking of voltage flicker levels. *Proc. 2002 PES Winter Meeting*, 3:1153–1157, 2002.

[5] T. K. Abdel-Galil, E. F. El-Saadany, and M. M. Salama. On-line tracking of voltage flicker utilizing energy operator and Hilbert transform. *IEEE Trans. Power Delivery*, 19:861–867, 2004.

[6] J. Arrillaga, N. R. Watson, and S. Chen. *Power System Quality Assessment*. Wiley, N.Y., 2000.

[7] R. Carmona, W-L. Hwang, and B. Torrésani. *Practical Time-Frequency Analysis, Gabor and Wavelet Transforms with an Implementation in S*. Academic Press, 1998.

[8] M. T. Chen and A. P. S. Meliopoulos. Wavelet-based algorithm for voltage flicker analysis. *Proc. 9th Int. Conf. Harmonics and Quality of Power*, 2:732–738, 2000.

[9] Q. Chen, X. Jia, and C. Zhao. *Analysis on Measuring Performance of Three Flicker Detecting Methods*, pages 1–7. Power & Energy Society General Meeting, 2009. PES '09. IEEE. July 2009.

[10] E. A. Feilat. Detection of voltage envelope using Prony analysis-Hilbert transform method. *IEEE Transactions on Power Delivery*, 21(4):2091–2093, Oct. 2006.

[11] A. A. Girgis, J. W. Stephens, and E. B. Makram. Measurement and prediction of voltage flicker magnitud and frequency. *IEEE Trans. Power Delivery*, 10(3):1600–1605, 1995.

[12] S. L. Hahn. *Hilbert Transform in Signal Processing*. Artech House, 1996.

[13] H. Li, L. Li, and T. Qian. Discrete-time analytic signals and bedrosian product theorems. *Digital Signal Processing*, 20:982–990, 2010.

[14] Tianyun Li, Yan Zhao, and Yongquiang Han. Application of Hilbert-Huang transform method in detection of harmonic and voltage flicker. *Power System Technology*, 29:74–77, 2005.

[15] M.I. Marei, T.K. Abdel-Galil, and E.F. El-Saadany. Hilbert transform based control algorithm of the DG interface for voltage flicker mitigation. *IEEE Transactions on Power Delivery*, 20(2):1129–1133, April 2005.

[16] C. Schauder. STATCOM for compensation of large electric arc furnace instalations. *Proc. 1999 IEEE-PES Summer Meeting*, 2:1109–1112, 1999.

[17] S. A. Soliman and M. E. El-Haway. Measurements of power system voltage and flicker levels for power quality analysis: a static LAV state estimation based algorithm. *International Journal of Electrical Power and Energy Systems*, 22:447–450, 2000.

[18] Hong Su and Yi Wang. Voltage flicker detection method based on mathematical morphology filter and Hilbert transform. *Proc of CSEE*, 28:111–114, 2008.

[19] W. Tong, S. Yuan, Z. Li, and X. Song. *Detection of Voltage Flicker Based on Hilbert Transform and Wavelet Denoising*, pages 2286–2289. Proc. of the 2008-DRPT, Nanjing-China. 2008.

MÉTODOS NUMÉRICOS PARA PROCESAMIENTO DE SEÑALES EN TIEMPO DISCRETO APLICADOS AL SENSADO REMOTO POR ONDAS DE RADIO

María G. Molina[1,2], Miguel A. Cabrera[2,3] , Patricia M. Fernández de Campra[1] y Rodolfo G. Ezquer [2,4,5]

1. Dpto. de Ciencias de la Computación, Facultad de Ciencias Exactas y Tecnología (FACET), Universidad Nacional de Tucumán (UNT), Argentina. pfernandez@herrera.unt.edu.ar
2. Laboratorio de Ionósfera, Dpto. de Física, FACET, UNT, Argentina. gmolina@herrera.unt.edu.ar
3. Laboratorio de Telecomunicaciones, DEEC, FACET, UNT. mcabrera@herrera.unt.edu.ar
4. CONICET, Argentina
5. CIASUR, Fac. Reg. Tucumán, Universidad Tecnológica Nacional, Argentina. rezquer@herrera.unt.edu.ar

Palabras claves: *Compresión de Pulso, Correlación, Ionósfera*

RESUMEN

Un sondador o ionosonda es un tipo particular de radar utilizado en estudios geofísicos de la alta atmósfera terrestre. Este instrumento determina alturas virtuales de la ionosfera y frecuencias críticas de los diferentes estratos ionosféricos mediante una serie de procesos numéricos. Los métodos modernos de procesamiento de señales para el sensado remoto por ondas de radio, como la denominada compresión de pulso, se apoyan fuertemente en algoritmos numéricos. En particular, los modernos sondadores utilizan esta técnica para la detección de ecos débiles, la cual permite alcanzar altas resoluciones en la determinación del rango, emitiendo bajas potencias de radio. En este trabajo se presentan resultados de la aplicación del método de compresión de pulso sobre datos del sondador AIS-INGV, ubicado en la estación ionosférica de Roma. Se concluye que es posible la determinación de alturas virtuales incluso en el caso de ecos inmersos en grandes niveles de ruido.

1. INTRODUCCIÓN

Mediante el sondaje vertical realizado por una ionosonda, es posible obtener alturas virtuales y frecuencias críticas de los distintos estratos ionosféricos para una dada frecuencia (Risbeth and Garriot, 1969).

Los sondadores tradicionales utilizaban la detección por envolvente para determinar la presencia del blanco, mientras que los equipos modernos detectan fase de portadora y código. Así aprovechando la gran capacidad de cálculo computacional actual basan su funcionamiento en el procesamiento digital de señales. Actualmente, la técnica más utilizada es la compresión de pulso, cuyo propósito es el de mejorar la resolución de rango de los radares emitiendo menor potencia de radio. Esta técnica requiere de procesos numéricos para la detección de los ecos con el fin de extraer información de las características del blanco detectado (Skolnik, 1980).

En este trabajo, se implementaran los métodos numéricos involucrados en la estimación de la altura virtual para una frecuencia fija usando datos reales generados en la estación ionosférica Roma con el equipo AIS-INGV (Ariokiasamy et al, 2003). Se utilizará la correlación de funciones discretas en el dominio del tiempo para implementar la compresión de pulso. A través de simulaciones se estimará la distancia a la cual se encuentra el blanco, que en este caso, representa la altura virtual de la ionósfera.

2. METODO DE COMPRESION DE PULSO

La compresión de pulso consiste en la transmisión de una larga ráfaga de portadora de radio modulada que permite lograr ancho de banda similar a la emisión de un pulso de RF no modulado (Skolnik, 1980).

Así se aprovechan las características espectrales de los diferentes tipos de códigos utilizados en la modulación, para detectar la presencia del objetivo dentro de un eco embebido en ruido, en la etapa receptora del radar. La detección implica la ejecución de procesamiento digital, y en particular la correlación numérica de las muestras con el código modulador. La señal transmitida es una portadora codificada (modulada) convenientemente. En el caso de las simulaciones realizadas, el esquema de codificación seleccionado es el de secuencias binarias de código complementario (Golay, 1961; Molina et al, 2010; entre otros). Para realizar la detección es necesario comparar o correlacionar este código, llamado código local, con la señal recepcionada. Una vez que la señal analógica del eco es recepcionada, ésta es muestreada para ser almacenada digitalmente. A partir de este momento las muestras discretas son "comparadas" con el código local. Si es posible determinar la marca temporal donde la señal recepcionada coincide con el código local, entonces será posible determinar la distancia en la que ocurre la reflexión de la señal. Esta distancia se denomina rango y se puede calcular con la ecuación,

$$h_v = \frac{c\,\tau}{2} \quad (1)$$

Donde h_v es la distancia, τ es la marca temporal y $c = 3x10^8\ m/s$ la velocidad de la luz en el vacío.

La manera de establecer esta marca temporal es utilizar un algoritmo numérico que compare las muestras recepcionadas con el código local, denominada correlación de señales discretas en el dominio del tiempo.

Dadas dos señales discretas de energía, x e y, su función de correlación es,

$$R_{xy}[m] = \sum_{n=-\infty}^{\infty} x[n]y[n+m] = \sum_{n=-\infty}^{\infty} x[n-m]y[n] \quad (2)$$

Donde m representa el desplazamiento en el tiempo que sufre una de las dos señales (Roberts, 2004).

Si estas señales discretas son finitas se las puede expresar vectorialmente de la forma $x = (x_0, x_1, ..., x_i, ...x_{N-1})$ e $y = (y_0, y_1, ..., y_i, ...y_{N-1})$, y los índices de la sumatoria se pueden reescribir como,

$$R_{xy}[m] = \sum_{n=0}^{N-1} x[n]y[n+m] = \sum_{n=0}^{N-1} x[n-m]y[n] \quad (3)$$

Las operaciones vectoriales que deben realizarse, no implican grandes costos computacionales o de almacenamiento debido a que no se utiliza un gran número de muestras en esta aplicación particular (512 muestras para un pulso de 480 μs).

3. DATOS Y RESULTADOS

Mediante una serie de sondajes realizados durante el año 2008 en la estación ionosférica Roma se probaron los métodos antes mencionados. Las señales utilizadas son las resultantes de la adquisición en la etapa de recepción del sondador, están compuestas por señal útil más ruido. El objetivo es determinar la presencia de la señal útil dentro de la recepcionada, lo que implica la determinación del blanco. En todos los casos para establecer un criterio que determine una detección positiva, se utilizaron como referencia diferentes umbrales de ruido. Se estableció que si un determinado pico supera en un 20% el umbral de ruido, éste representa la presencia del blanco (Sultzer y Woodman, 1984).

El primer caso de estudio corresponde al sondaje realizado el día 16 de octubre de 2008 a las 12:05 hs. Se realizó la correlación entre la señal recepcionada y el código local almacenado. La Figura 1 muestra el resultado final donde se observa claramente un valor máximo, el cual representa una correlación máxima.

Para estimar si realmente corresponde a la presencia del blanco, se debe tener en cuenta el umbral de ruido. En este primer caso, el valor máximo supera ampliamente los niveles de ruido, incluso en más del 100%.

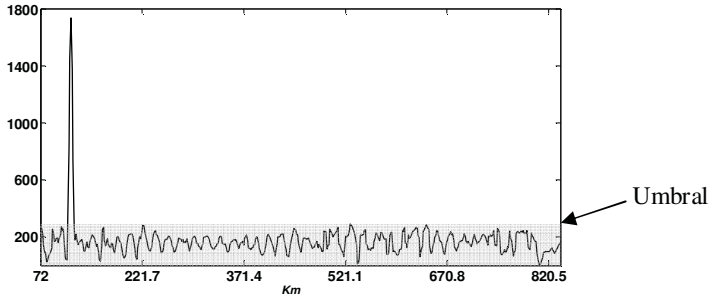


Figura 1. Estimación altura virtual del 16/10/2008 horas 12:05

El segundo caso corresponde a una señal inmersa en grandes niveles de ruido que corresponde al sondaje realizado el día 16 de octubre de 2008 a las 13:20 hs. La Figura 2 muestra que a pesar que hay un umbral de ruido más alto, aún es posible obtener un máximo distinguible. La evaluación de este máximo respecto a la base de ruido, muestra que este valor máximo es 5 veces mayor al valor base de ruido y con ello que este pico representa, efectivamente, la presencia del blanco.

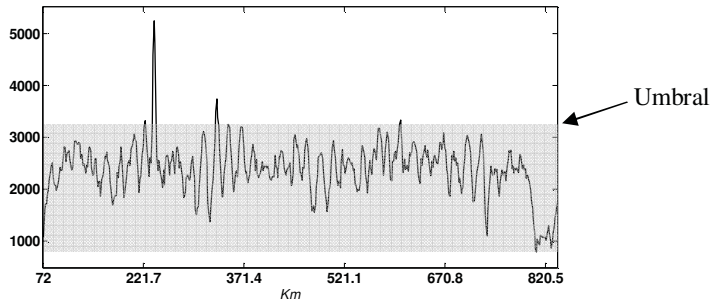


Figura 2. Estimación altura virtual del 16/10/2008 horas 13:20

Finalmente se analizó la posibilidad de que existan dos ecos en la misma señal recepcionada. En este caso el sondaje es el correspondiente a las 15:30 hs del mismo día que los casos anteriores. En la Figura 3 se pueden observar no uno, sino dos picos distinguibles inmersos en grandes niveles de ruido. El eco máximo claramente supera el 20% de la base de ruido, sin embargo el eco débil se encuentra apenas por debajo de este valor.

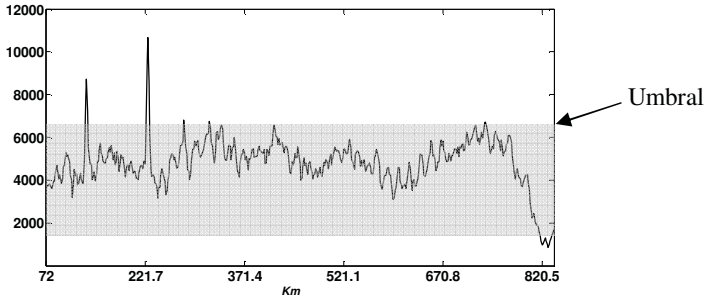


Figura 3. Estimación altura virtual del 16/10/2008 horas 15:30

Es claro que, para establecer que existe detección, es fundamental decidir cual va a ser el umbral de ruido. En todos los casos se utilizó un umbral variable que depende del valor promedio de las muestras que representan el ruido, previa eliminación de aquellas que presuponen la presencia del blanco. Existen numerosas formas de determinar el umbral de ruido dinámico, que serán evaluadas en trabajos posteriores. Sin embargo, para los casos de señales ionosféricas, se mostró empíricamente que es posible tener una detección positiva aún cuando se utiliza un método tan simple como el desarrollado en este trabajo. Una vez obtenida la marca temporal, y usando la ecuación 1, se puede estimar fácilmente la distancia recorrida por la señal y con ello, estimar la altura virtual de la ionósfera.

4. CONCLUSIONES

El procesamiento digital de señales ionosféricas utiliza numerosas técnicas numéricas y estadísticas. El objetivo de este trabajo es poder estimar la altura virtual de la ionósfera mediante la transmisión y recepción de ondas de radio HF. Nuestros resultados muestran que, aún en circunstancias donde la señal recepcionada se encuentra inmersa en grandes niveles de ruido, es posible detectar la señal útil dentro de ésta. Este eco representa la marca temporal que permitirá, posteriormente, determinar la altura virtual.

El proceso descripto es factible de implementarse en el dominio de la frecuencia donde nuevas herramientas numéricas deben ser utilizadas. El costo computacional para trabajar en el dominio del tiempo es relativamente bajo y sencillo, y no es preciso recurrir al dominio de la frecuencia.

Las simulaciones numéricas son un método muy útil al momento del diseño del equipo sondador, de tal manera que permiten con bajos costos preveer resultados y evaluar los criterios de diseño.

5. Agradecimientos

Este trabajo se realizó dentro del Proyecto de Investigación 26/E408 CIUNT-UNT. Los autores agradecen al Istituto Nazionale de Geofisica e Vulcanologia de Italia por los datos del sondador AIS de Roma.

6. Referencias

[1] JAMES B. ARIOKIASAMY, C. BIANCHI, U. SCIACCA, G. TUTONE AND E. ZUCCHERETTI, *The new AIS-INGV Digital Ionosonde*,Annals of Geophysics, pp 647-659, 2003.
[2] M.J.E GOLAY, *Complementary series,* IRE Trans. Inf.Theory, 7, 82-87, 1961.
[3] M. G. MOLINA, M. A CABRERA., J. LOPEZ, R. G. EZQUER, C. IVAN, *Análisis de Esquemas de Codificación de Señales para Aplicaciones en Radares Ionosféricos,* Libro de resúmenes de las XVIII Jornadas de Jóvenes Investigadores UGM, Santa Fé, Argentina (2010), 135.
[4] H. RISHBETH AND O. K. GARRIOT, *Introduction to Ionospheric Physics*, Academic Press, England,1969.
[5] M. J. ROBERTS, *Señales y Sistemas,* Mc Graw Hills, 2004.
[6] M. I SKOLNIK, *Introduction to radar systems,* Mc Graw Hills, 1980.
[7] M. P. SULTZER, AND R. F. WOODMAN, *Quasi-complementary codes: A new technique for radar sounding,* Radio Science., 19, pp 337-344 (1980).

RECONOCEDOR DE NÚMEROS TELEFÓNICOS BASADO EN MODELOS DE MARKOV OCULTOS

Patricio Perez Preiti, Claudio Estienne, Damian Simkin, Sebastian Perez y Patricia Pelle

Facultad de Ingeniería, Universidad de Buenos Aires, Av. Paseo Colón 850, Ciudad de Buenos Aires, Argentina, www.fi.uba.ar
ppreiti@fi.uba.ar, cestien@fi.uba.ar, damiansimkin04@hotmail.com, sebaperez31@gmail.com, ppelle@fi.uba.ar

Resumen: En este trabajo se implementa un sistema de reconocimiento de números telefónicos emitidos en forma oral, con adaptación a un usuario determinado. La adaptación da como resultado la disminución de los errores cuando es usado siempre por el mismo hablante, mejorando la satisfacción del usuario. La realización de este trabajo consta de tres partes. La primera tiene como fin la generación de una "línea de base", es decir, un sistema de reconocimiento de habla sin ninguna aplicación específica, que permita testear el desempeño inicial contra otros sistemas similares considerados estado-del-arte. La segunda parte es la aplicación de la gramática particular de los números telefónicos limitando el vocabulario a las posibles formas de mencionar los números en sus variantes más usuales, y disminuyendo la tasa de errores comparado con el sistema de línea de base. Por último se agrega la etapa de adaptación que produce una disminución adicional de los errores.

Palabras clave: *Reconocimiento del habla, Reconocimiento de patrones*
2000 AMS Subject Classification: 68T10

1. INTRODUCCIÓN

La tarea de un reconocedor de habla es encontrar la transcripción escrita de una frase emitida en forma oral. Varias aproximaciones son posibles, siendo el enfoque estadístico del problema el más comunmente utilizado por ser el que da mejor desempeño al presente. En este trabajo se describe la implementación de un sistema de reconocimiento y la evaluación de su desempeño. El sistema es uno de los más simples, que utiliza un vocabulario restringido a números de teléfono. La tarea es sencilla pero describe adecuadamente la totalidad de los pasos involucrados en tareas más complejas, pudiéndose apreciar mediante los resultados experimentales obtenidos las diferentes alternativas que hay que sortear en su diseño.

2. DESCRIPCIÓN GENERAL DEL SISTEMA

2.1. EXTRACCIÓN DE CARACTERÍSTICAS DE LA SEÑAL

La señal de habla es una secuencia de sonidos concatenados, de duración variable y cuyas características pueden ser distinguidas desde un punto de vista espectral. La primer etapa en un sistema de reconocimiento es codificar la señal de habla (de forma computacionalmente manejable), de modo tal de poder captar las diferencias entre estos sonidos. Para esto el primer paso es convertir la señal de habla analógica en una señal digital mediante un proceso de muestreo y luego dividirla en ventanas de una longitud que permita apreciar las características acústicas, pero suficientemente cortas como para que la señal pueda ser considerada estacionaria dentro de la ventana. Cada ventana de señal es luego procesada para capturar las características espectrales de la señal en una forma compacta (parametrización). Existen diversos métodos de parametrización de las ventanas, en este trabajo se utilizará Mel Frequency Cepstral Coefficients (MFCC,[1]). Los coeficientes MFCC son una representación basada en características perceptuales e intentan emular las características de resolución espectral en el sistema auditivo humano.

2.2. MODELOS ESTADÍSTICOS ACÚSTICOS

Para generar los modelos estadísticos se debe definir primeramente cuál será la una unidad básica que los componen. Una posibilidad es modelar cada palabra del vocabulario. Pero la cantidad de variantes de palabras y la complejidad de éstas hace que se prefiera utilizar los fonemas como la unidad elemental. De este modo, los modelos de palabras serán obtenidos concatenando los modelos de fonemas entre sí. Una vez que se define a los fonemas como la unidad básica, podemos utilizar cadenas ocultas de Markov para

modelar de manera estadística las emisiones de la señal de habla [3]. Las cadenas de Markov son secuencias de estados que se van sucediendo de acuerdo a las probabilidades de transición entre estados. Las emisiones de la señal están compuestas por la secuencia de ventanas parametrizadas correspondientes a una frase. Las frases se componen de palabras que a su vez se componen de fonemas, de modo que la secuencia de ventanas parametrizadas correspondientes a una frase se corresponde a una secuencia de fonemas. La aplicación de las cadenas de Markov a este caso se da asociando cada estado con un dado fonema. La secuencia de ventanas parametrizadas será entonces asociado a una secuencia de estados recorridos. En las cadenas *ocultas* de Markov, el nexo entre el estado-fonema y la evidencia acústica (ventanas de señal parametrizada) se da asumiendo que cada estado tiene una función de distribución de probabilidad asociada a la emisión de un dado fonema. De este modo es posible determinar de forma genérica la probabilidad de emitir una secuencia de ventanas parametrizadas dada por un cierto recorrido sobre la cadena. En este contexto, se puede construir una cadena de Markov oculta para cada palabra del vocabulario, y por ende para cada frase posible, mediante la concatenación de los estados correspondientes a los fonemas de esa palabra, y determinando luego las funciones de densidad y las probabilidades de transición de la cadena. Por otra parte, aplicando el teorema de Bayes será posible también calcular la probabilidad a posteriori de una dada secuencia de ventanas parametrizadas dado un cierto recorrido por la cadena. En la etapa de reconocimiento consideramos que la secuencia de fonemas emitidos está oculta y sólo conocemos las probabilidades de cada secuencia de estados dada la evidencia acústica. Entonces, la etapa de reconocimiento consiste en determinar las probabilidades a posteriori de cada secuencia de estados posible dada la evidencia acústica, es decir la secuencia de ventanas parametrizadas. Y se asume como resultado del reconocedor a la más probable de todas esas secuencias posibles.

El desempeño del reconocedor está influido principalmente por dos factores: primero, el ajuste de las funciones de distribución de cada fonema con respecto a los fonemas observados. Segundo, la cantidad de secuencias de estados posibles a recorrer. En el primer caso, un desajuste de las funciones que caracterizan a los fonemas puede producir que secuencias equivocadas den mayores probabilidades que las verdaderas secuencias. En el segundo caso, si las secuencias de estados son muchas habrá muchas secuencias entre las cuales decidir, y mayores posibilidades de encontrar secuencias con probabilidades altas semejantes. Una mejora para el primer problema será explicado en esta sección. La solución al segundo problema se resuelve básicamente restringiendo la conexión entre palabras posibles, es decir, definiendo la gramática específica para la tarea, que será explicada en la siguiente sección.

Con respecto a la definición de las probabilidades de emitir parámetros de cada estado, es posible hacer una discriminación más fina entre fonemas. Cuando tomamos a los fonemas como unidad básica el modelo de un fonema siempre es el mismo sin importar entre qué sonidos aparezca. La realidad es que los sonidos correspondientes a un mismo fonema están muy afectados por los fonemas anterior y posterior (efectos co-articulatorios). Para mejorar el desempeño del reconocedor se emplean modelos que tienen en cuenta dichos efectos, que se denominan modelos contexto dependientes. En este caso habrá varios modelos distintos del mismo fonema de acuerdo al contexto en el que aparezca. El problema que esto trae aparejado es el aumento del número de modelos y por ende, el número de parámetros a estimar. Para lograr que todos los parámetros sean bien estimados se emplean técnicas de agrupación de modelos similares, como es el caso de los Arboles de decisión [6]. Este método heurístico permite decidir de acuerdo a reglas lingüísticas cuáles contextos producen el mismo efecto sobre el fonema, y por lo tanto cuáles modelos pueden ser agrupados, reduciendo así la posibilidad de que haya modelos sin datos de entrenamiento asociados. El uso de árboles de decisión permite tener mayor discriminación que en el caso contexto independiente (un sólo modelo por fonema) pero permitiendo obtener modelos más robustos al ser entrenados con una cantidad suficiente de datos.

2.3. Modelos de conexión entre palabras

En la construcción de las cadenas de Markov que representen todas las posibles secuencias de palabras de nuestro vocabulario se procede primero construyendo cadenas individuales para cada palabra. Estas cadenas son lineales, avanzando por cada uno de los fonemas que componen cada palabra. La conexión entre palabras en cambio no está tan claramente determinada. Existen gran cantidad de aproximaciones al problema,

ya que gran parte del desempeño del reconocedor radicará en la adecuada determinación de las conexiones y de sus probabilidades de transición. Una posibilidad, la más simple pero la de peor desempeño, es considerar que las N palabras que componen el vocabulario pueden estar conectadas aleatoriamente entre sí. Eso equivale a conectar cada palabra con todas las demás, incluso consigo misma, definiendo una probabilidad de transición $1/N$ uniforme. Otra posibilidad sería considerar que es posible que cada palabra esté conectada a todas las demás, pero la probabilidad de transición derivarla experimentalmente de la ocurrencia de dos palabras seguidas en las bases de entrenamiento. Esta determinación de las transiciones se denomina bigramas, aunque también es posible usar información de "historias" más largas para ello (trigramas, cuatrigramas). Otra posibilidad es restringir las conexiones si la tarea específica tiene alguna estructura clara. Por ejemplo se pueden utilizar reglas gramaticales, o como en otros casos de tareas muy específicas las conexiones se pueden determinar por el uso habitual. Este tipo de restricción es el que aplicamos en nuestro sistema de reconocimiento de dígitos correspondientes a números de teléfono.

2.4. Adaptación al hablante

Otra técnica que suele dar una mejora del desempeño para la determinación de las probabilidades de emisión acústica de cada fonema es la adaptación a las características de un dado hablante. Mediante la recopilación de datos del usuario que utilizará el sistema y el re-entrenamiento de los modelos para éste, se modelarán las características del orador con mayor exactitud. Alguna estrategia debe usarse para que la adaptación se produzca con una cantidad de datos reducida y no con las cantidades usuales con las que se entrenan en general los sistemas (típicamente horas de grabación). El método que se utilizará en nuestro caso es entrenar un mapeo entre los datos ya entrenados y las características particulares del hablante, conocido como Regresión Lineal de Máxima Verosimilitud (Maximum Likelihood Linear Regresion, MLLR,[2]), que computa un conjunto de transformaciones que reducirán las diferencias que existen entre el modelo inicial (modelo independiente del hablante) y los datos de adaptación.

3. Experimentos y resultados

La primera etapa es desarrollar un reconocedor básico no aplicado a ninguna tarea específica de modo que los resultados de reconocimiento sean comparables a otros considerados estado-del-arte sobre una base de datos prefijada. La base de datos sobre la que trabajamos es Latino 40 (distribuida por LDC, Linguistic Data Consortium, *http://www.ldc.upenn.edu*), una base de datos de pequeño vocabulario, compuesta de frases dichas en español por 40 hablantes diferentes latinoamericanos. La base de datos no está etiquetada en fonemas, sino que tiene únicamente la transcripción de palabras, sin alineación temporal. Una primera etapa entonces consistió en utilizar una base de datos más pequeña pero segmentada fonéticamente (TIMIT, distribuida por LDC, Linguistic Data Consortium, *http://www.ldc.upenn.edu*) para plantear modelos fonéticos básicos, y con esos modelos realizar una segmentación fonética automática de Latino 40. Esta segmentación se realiza planteando las cadenas de Markov correspondientes a las frases que aparecen en Latino 40 que son conocidas, y efectuando un reconocimiento. En este caso el reconocimiento no tiene el sentido de determinar la secuencia correcta (porque ésta es única), sino solo alinear los estados a las secuencias de ventanas de señal. De este modo se logra una segmentación inicial. Con esa segmentación inicial se pueden estimar los parámetros de la cadena de Markov oculta del sistema completo.

El entrenamiento consiste en determinar las unidades básicas a utilizar (determinación de los trifonos a utilizar), las probabilidad de distribución para cada uno de ellos, y las probabilidades de transición entre estados de una misma palabra y entre palabras. Para la determinación de la probabilidad de distribución se plantea un modelo basado en mezcla de gaussianas [3]. El número de gaussianas elegido es determinado de forma experimental, agregando gaussianas hasta que el desempeño empieza a disminuir. También es posible aumentar o disminuir el número de trifonos distintos a utilizar modificando parámetros de los árboles de decisión. El entrenamiento se realiza utilizando el algoritmo de Baum-Welch conocido también como Expectación-Maximización (EM,[3]) sobre la base de datos de entrenamiento. El reconocimiento se realiza utilizando el algoritmo de Viterbi [3] sobre la base de datos de reconocimiento. El sistema es desarrollado utilizando el sistema Hidden Markov Model Toolkit (HTK,[5]) desarrollado en la universidad de Cambridge,

que es de dominio público. Para evaluar el rendimiento del reconocedor a lo largo de este trabajo se utiliza como medida el Word Error Rate (WER), definido como el porcentaje de palabras erradas con respecto al total. El mínimo WER para los modelos iniciales utilizando TIMIT se obtuvo con 80 gaussianas, utilizando modelos de fonemas sin contexto. El mínimo WER para el sistema entrenado sobre Latino 40 con modelos de trifonos se obtuvo para 7 gaussianas. En lugar de reducir el número de trifonos se utilizó el árbol de decisión para determinar qué subconjuntos de trifonos pueden compartir las mismas gaussianas entre sus modelos. Para esta primera etapa se utilizaron conexiones entre palabras correspondientes a bigramas y trigramas, generados a partir de un vocabulario de 3000 palabras (estas palabras son las que aparecen en la base Latino 40, en la parte de entrenamiento). La herramienta utilizada para construir los modelos de lenguaje es el SRI Language Modeling Toolkit (SRILM,[4]). Los resultados de reconocimiento obtenidos (con y sin adaptación al hablante) se muestran en la tabla 1a.

	Sin adaptación		**Con adaptación**	
Hablante	**%WER (Bigrama)**	**%WER (Trigrama)**	**%WER (Bigrama)**	**%WER (Trigrama)**
cm05	17,4	8,0	17,2	5,8
gf18	10,3	4,7	11,6	3,0
hf39	8,6	9,3	3,3	7,3
nm15	10,2	10,8	9,0	7,1
pf40	5,7	7,8	3,2	5,1
vm25	15,8	14,9	13,5	11,8
Total	***11,3***	***9,2***	***9,6***	***6,7***

(a) Resultados de reconocimiento para el sistema de línea de base, con y sin adaptación

Hablante	**Adaptación**	**%WER**
Pa	No	14,78
Pa	Si	6,46

(b) Resultados de reconocimiento para el sistema de reconocimiento de números telefónicos, con y sin adaptación

Tabla 1: Resultados de reconocimiento obtenidos

Con el sistema base construido se pasa a la aplicación particular, en este caso, el reconocimiento de números telefónicos. Para ello, se debe armar la gramática del dictado de los números telefónicos (la gramática deberá incluir todas las formas posibles usuales de dictado). A partir de dicha gramática se grabaron 75 oraciones, todas del mismo hablante, de las cuales se emplearon 50 para la evaluación del sistema y 25 para realizar la posterior adaptación al hablante. Los resultados de reconocimiento obtenidos (con y sin adaptación al hablante) se muestran en la tabla 1b.

4. CONCLUSIONES

A lo largo de este trabajo hemos mostrado las distintas etapas en el diseño de un sistema de reconocimiento de habla basado en Modelos de Markov Ocultos. La experiencia en el campo del habla nos ha demostrado la gran influencia de los parámetros de diseño en los resultados finales (gramática de la aplicación, parametrización de las señales, tipos de modelos, formas de agrupamiento, modelos de lenguaje, etc.). Por último vimos como mejoran los valores de reconocimiento cuando un sistema es utilizado por un único hablante cuando se realiza adaptación al mismo.

REFERENCIAS

[1] FRANCO.H., *Modelación de Habla con HMM*, Tesis Doctoral, Universidad de Buenos Aires, 1990

[2] LEGGETTER C.J., WOODLAND P.C. , *Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression*, Cambridge University, Engineering Dept., 1994

[3] RABINER L.R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, Vol.77 No. 2, February, 1989

[4] STOLCKE A., *SRILM-An Extensible Language Modeling Toolkit*, Speech Technology and Research Laboratory, 2000

[5] YOUNG S.J., EVERMANN G., ODELL J.J., *The HTK Book. Version 3.3*, Cambridge University, April, 2005

[6] YOUNG S.J., ODELL J.J., WOODLAND P.C., *Tree-based state tying for high accuracy acoustic modelling*, HLT '94 Proceedings of the workshop on Human Language Technology, Cambridge University, April, 1994

ESTIMATING THE QUEUE LENGTH TO OPTIMIZE THE GREEN TIME FOR AN URBAN TRAFFIC CONTROL SYSTEM

Juan D'Amato[1], Pablo Negri[1] and Pablo Lotito[1]

[1]*PLADEMA, Universidad del Centro de la Provincia de Buenos Aires, Tandil, Argentina , {jdamato,pnegri,plotito} (a) exa.unicen.edu.ar*

Abstract: We describe a queue length estimation module that is part of an Urban Traffic Control System (UTCS). The main objective of this module is to estimate the traffic queue length at an intersection using image processing algorithms. This measurement feeds a traffic green time optimizer that chooses adequate green periods, based on the traffic load, and reduces the global time spent by drivers.

We compare the results of three algorithms that estimate the queue length based on Gaussian Mixture Models and Level Lines. A hybrid method, which combines both approaches, is better at minimizing the system errors. In addition, this framework makes it possible to test different camera positions to improve results.

Keywords: *Video processing, traffic flow, estimating*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

This paper presents and analyzes a methodology used in the congestion reduction module of a UTCS. This methodology incorporates recent advances in image processing, simulation models and control and optimization methods. The main objective of this module is to adapt the computation of adequate green times to demand variations (traffic load), and to reduce the global time spent by drivers accordingly. In order to achieve this goal, the full state of the traffic network should be known in each step [4].

Historically, inductive loops have been the most widely used sensors to measure the queue length. The vehicles entering and exiting the road section are detected, and the number of vehicles on the local road section is calculated by difference [3]. Many algorithms, which depend on the data provided by the loop detectors (number of vehicles, individual speed, etc.), still have some problems to be solved. As the algorithms need accurate data, the loss of one loop dramatically disturbs the system, and the length of the queue is under-estimated. Additionally,the number of loop detectors increase rapidly according to the number of roads at the inspected crossing and the number of lanes on each road. In contrast, the cost of one camera covering all the lanes of a road is equal to the cost of one loop in a lane.

Our application uses a computer vision system to estimate traffic metrics, such as the queue length. Several tests used to design the queue measurement system through virtual video captures simulating strategically placed cameras are presented. The advantage of 3D virtual reconstructions is that they yield the complete state of the system and make it possible to analyze and compare different configurations in order to reduce measurement errors. An efficient algorithm to estimate the queue length, which is advantageous as it does not increase the overall computation time, is also shown.

Figure 1: Screen-shot of the UTCS at PLADEMA showing three views of the same intersection generated by the 3D visualization module.

2 MEASURING SYSTEM USING COMPUTER VISION

The 3D virtual reconstructions are generated by a UTCS module developed by this group. The visualization module recreates the vehicle movement in a realistic 3D scenario and generates synthetic videos subsequently used in our study. This virtual tool makes it possible to analyze scenarios with structural variations (e.g., changing the camera position on a corner) and provides instantaneous information, such as vehicle speed and position used for automatic validation. Figure 1 shows different points of view of the same state of the simulator, located in three places where cameras could be potentially installed. This methodology, compared to video campaigns, leads to a significant reduction in the time and cost of video acquisition, at the same time the analysis task is significantly reduced.

To start our study, we must first define a polygonal section for each view. This polygon is separated into two lines and then into four cells each, representing on the image the place where the vehicles are queuing up. The first cell of each line is the head of the queue. As we are modeling a traffic light cycle, the vehicle movement discriminates both cycles (green and red). We assume that a vehicle queue exists if there is detection in a cell without motion (red cycle). The queue length is defined as the amount of occupied cells without motion if the first one (the head) is stationary.

We analize three algorithms to estimate occupancy and motion on the images. These strategies model the background to detect differences between consecutive frames in a video. The first one uses Gaussian Mixture Models (GMMs) [6], and the second one uses Level Lines [1] but as both have certain limitations, we propose a new strategy that keeps their advantages and enhances the global performance.

2.1 GAUSSIANS MIXTURE MODELS

The detection method proposed for Stauffer [6] considers each pixel as an independent statistical process where the n values over time of each pixel can be modeled by K Gaussians. The background is made up of those Gaussians which are stables in a sequence. One pixel is considered as a foreground (movement) if their probability of belonging to one of the K Gaussians background models is lower than a threshold. A pixel $\mathbf{p}$ belonging to the foreground switches on the binary detection matrix: $D(\mathbf{p}) = 1$.

To determine if a cell is occupied in our model, we consider all pixels $\mathbf{p}$ of the foreground ($D(\mathbf{p}) = 1$) contained in it. Using a similar approach to [7], the number of detected points must be divided by the cell area to get a standarized measure. If this value is greater than a threshold, the cell is assumed to be occupied. Formally, for each cell C_i we obtain ratio d_i:

$$d_i = \frac{N(D_i)}{A(C_i)} \geq T_i^d$$

where $N(D_i)$ is the number of detections in cell i and $A(C_i)$ is the area of cell i.

The motion detection method computes the difference of two successive frames in the sequence. This operation is performed only on the foreground pixels $\mathbf{p}$ with $D(\mathbf{p}) = 1$. We get motion matrix M:

$$M(\mathbf{p}) = \{\mathbf{p} | D(\mathbf{p}) = 1 \wedge |F_t(\mathbf{p}) - F_{t+1}(\mathbf{p})|\} \quad (1)$$

To estimate the movement inside a cell, we accumulate the values of motion pixels $\mathbf{m}$ and divide them by the number of detected pixels, obtaining a new ratio m_i for cell i:

$$m_i = \frac{S(M_i)}{N(D_i)} \geq T_i^m$$

If m_i is greater than threshold T_i^m, we consider that there are a moving vehicles in cell i.

2.2 LEVELS LINES

Other effective detection method, using objects contours as primitives, is the Level Lines representation [1]. The Level Lines are defined as follows. Let be $I(x, y)$ be the intensity value on image I at coordinates (x, y), the level set of I is the set of pixels $\mathbf{p}(x, y)$ whose intensity is greater than or equal to λ:

$$\Xi_\lambda = \{\mathbf{p} / I(\mathbf{p}) \geq \lambda$$

The boundaries of a level set Ξ_λ are called Level Lines λ ($LL\lambda$) and, for the same image, we can obtain a set of $LL\lambda$ choosing different thresholds $\lambda = \{0, 1, 2, ..., 255\}$. Each $LL\lambda$ also contains information about the direction of the level line in the pixel. The direction of the pixel $\mathbf{p}(x, y) \in LL\lambda$ is calculated as the slope of the straight line created by the neighbors of $\mathbf{p}$, using the least square method. Then, this value is quantified in η values (for our application $\eta = 12$). A binary function $f_t(\mathbf{p}, \theta_k)$ signals the presence of a direction θ_k on the pixel $\mathbf{p}$ at the time t, and $1 \leq k \leq \eta$. Pixel $\mathbf{p}$ belonging to the background model have a stable $f_t(\mathbf{p}, \theta_k)$ for all the directions, during an interval of time T.

Figure 2: Detection matrix D for the GMM and the Level Lines algorithms.(left) Capture (center) GMM detection (right) LL detection

One pixel $\mathbf{p}(x, y) \in LL\lambda$ will be considered as a detection if its directions θ_k do not match the direction of the reference at the (x,y) location. Then, the detection matrix is switched on at this pixel: $D(\mathbf{p}) = 1$ (see Fig. 2).The vehicle detection follows the same procedure as that of the GMM algorithm, counting the number of detected lines in the cell.

In the motion detection, we match the detections in the previous frame $D_{t-1}(\mathbf{p})$ and the present frame $D_t(\mathbf{p})$. We define a new matrix $H(\mathbf{p})$ where each point $\mathbf{p}$ has a distance to the nearest detection $D(\mathbf{p}')$. For each $\mathbf{p}$ where $D_{t-1}(\mathbf{p}) = 1$ (a detection), we look for the most similar pixel $\mathbf{p}'$ in D_t in a NxN centered window at $\mathbf{p}$, using an intensity difference criterion $I_{t-1}(\mathbf{p}) - I_t(\mathbf{p}')$, the number of differences in the directions between $f_{t-1}(\mathbf{p}, \theta_k)$ and $f_t(\mathbf{p}', \theta_k)$, and comparing $H_t(\mathbf{p})$ and $H_{t-1}(\mathbf{p}')$.

Each pair of matched points $\mathbf{p}$ and $\mathbf{p}'$ generates a velocity vector $\vec{\mathbf{v}}_i = \mathbf{p}_i\vec{\mathbf{p}'}_i$. Inside the cell, we add up all the generated velocity vectors and the modulus is divided by the number of detected pixels to obtain a motion ratio. If this ratio is greater than a threshold, we conclude that a moving vehicle is inside the cell.

2.3 Hybrid Algorithm

We propose a new hybrid algorithm that improves the system performance while preserving the advantages of the previous methods. On the one hand, the GMM method detects vehicles more slowly but more accurately than the LL method (as we will see in the results). On the other hand, the LL method provides a velocity vector, which can represent the speed of a vehicle in the cell. It is impossible to obtain these vectors through the motion detection algorithm of the GMM method.

To compute vehicle detection, the LL algorithm computes the movement in the frame very quickly. The resulting curves D_t^{LL} are projected on a horizontal profile. The GMM algorithm is applied only to the areas of the image where the profile has a value that is greater than a threshold. Then, we obtain the detection matrix D_t^{GMM}, speeding up the vehicle detection.

As it is very expensive to compute the motion vectors over all the pixels of the level lines, we use motion matrix M (see eq. 1) where $D_t^{GMM}(\mathbf{p}) = 1$ instead. Those pixels where intensity difference is greater than a threshold generate a binary mask of the level lines in the frame D_{t-1}^{LL}. The velocity vectors are then generated by matching the pixels in the mask with the corresponding ones in I_t.

3 Results

The test sequence is made up of a virtual scene with more than 50 queues. Among other advantages already mentioned, the virtual simulator provides detailed information, such as vehicle position and speed at any time. In this regard, the user does not need to label the images used in learning and testing, nor even estimate the vehicle speed, which is a rather complex task. The motion T_i^m and detection T_i^d thresholds

Table 1: Table of results for the detection algorithms.

Views	Level Lines			Gaussians Mixture Models			Hybrid Method		
	FP (%)	FN (%)	Mean er.	FP (%)	FN (%)	Mean er.	FP (%)	FN (%)	Mean er.
1	0.78	20.43	0.65	1.31	5.20	0.92	4.34	2.64	0.85
2	1.09	23.08	0.55	0.50	9.87	0.59	3.44	4.43	0.55
3	0.67	13.05	0.62	1.22	7.71	0.57	2.59	2.61	0.61

of each detection method were adjusted to minimizing the errors. To analyze the results of the algorithms, like in [2], we define the false detections rate (false positives (FPs) over the total number of queues), the non-detection rate (false negatives (FNs) over the total number of queues), and the queue length mean error (in cells) with respect to detected queues. Table 3 shows the results obtained on assessed sequences.

The LL method yields a very large number of FNs, especially in the presence of dark vehicles, as detected lines are not enough to trigger the detector cell occupancy. The GMM and the Hybrid methods have fewer FNs and the latter has the best results for this evaluation metric. In fact, this is the most important metric because an FN error means that the UTCS required the queue length and the detection system estimated that the vehicles were not queuing up. The worst queue length error is observed when there is a truck in the line. Because of its height, the truck detection fills the cell behind it, increasing the error.

The proposed Hybrid algorithm shows an intermediate behavior between the GMMs and Level Lines, reducing the FN percentage and, at the same time, providing reliable information about the vehicle speed. As regards the view configuration, the best one can be found between views 2 and 3. The problem of view 1 is that vehicles in the same row are concealed, as can be seen in Figure 1. This effect is reduced in view 2 and 3, concluding that the installation position should be somewhere between them. For view 3, it is also possible to use algorithms like those proposed in [5] to reduce the occlusions.

4 Conclusions

A framework for the evaluation of video analysis algorithms to detect the length of a queue in front of a traffic light has been presented. We used three algorithms to count the number of vehicles waiting at the red light. Two of them were based on classical methods: Gaussian Mixture Models and Level Lines. The third one is an efficient combination of the other two, and it obtains better results minimizing the system errors. This framework also allows the user to test different camera positions to improve results.

As an extension of our method, we plan to incorporate a line length estimator, which tracks incoming vehicles in the sequence video before they queue up. Thus, we would be able to handle cases of undetected vehicles hidden by others.

References

[1] D. Aubert, F. Guichard, S. Bouchafa, *Time-scale change detection applied to real-time abnormal stationarity monitoring*, Real-Time Imaging, Vol. 10, 2004, pp 9-14.

[2] D. Aubert, F. Boillot,*Automatic measurement of traffic variables by image processing application to urban traffic control*, Recherche - Transports - Securite, Vol. 62, pp. 7-21, 1999.

[3] L.A. Klein, M.K. Mills and D.R.P. Gipson, "Traffic Detector Handbook: Third Edition", Technical Report, Vol. I & II, 2006.

[4] Mayorano, F., Rubiales, A., Lotito, P. A., *Optimal Control Based Heuristics for Congestion Reduction in Traffic Networks*, In: Proceedings of EngOpt, Brazil, 2008.

[5] C.C. Pang, W.W. Lam, N.H. Yung, *A method for vehicle count in the presence of multiple-vehicle occlusions in traffic images*, Intelligent Transportation Systems, Vo. 8, No. 3, pp. 441-459, 2007.

[6] C. Stauffer, W. Grimson,*Adaptive background mixture models for real-time tracking*, In: Proceedings of CVPR Vol. 2, 1999.

[7] M. Zanin, S. Messelodi, CM. Modena, *An Efficient Vehicle Queue Detection System Based on Image Processing*, In: Proceedings of ICIAP, pp. 232-237, 2003.

CONTROL ADAPTATIVO DE SISTEMAS NO LINEALES QUE ADMITEN LINEALIZACIÓN EXACTA. APLICACIÓN AL SISTEMA GLUCOREGULATORIO HUMANO.

Guillermo R. Cocha†, Carlos E. D'Attellis‡

†Facultad Regional La Plata, Universidad Tecnológica Nacional, La Plata, Buenos Aires, Argentina, gcocha@frlp.utn.edu.ar

‡Grupo de Ing. Clínica U.T.N.; Dto. de Matemática, Univ. Favaloro; Centro de Matemática Aplicada, UNSAM, Buenos Aires, Argentina, cdattellis@yahoo.com.ar

Resumen: Si un sistema tiene su grado relativo igual al número de estados en el entorno de un punto de equilibrio es posible efectuar una transformación de coordenadas y una realimentación de estados que convierte al sistema no lineal en uno lineal y controlable.
Esta técnica es aplicada en la regulación de glucosa en el organismo humano pero, dado que la misma se basa en la cancelación exacta de los términos no lineales, cuando aparecen parámetros que son variantes en el tiempo o incertidumbre en los términos no lineales del modelo, la cancelación ya no es exacta y el control puede dejar de ser efectivo. Este trabajo presenta una ley de control adaptativo no lineal basado en técnicas de linealización exacta aplicada al problema de regulación de glucosa en pacientes diabéticos por medio de infusión de insulina intravenosa y monitoreo continuo de glucosa.

Palabras claves: *control no lineal, control adaptativo, diabetes mellitus.*

2000 AMS Subjects Classification: 15 Procesamiento de Señales e Imágenes.

1. INTRODUCCIÓN

Consideremos un sistema de una entrada y de una salida

$$\left.\begin{aligned}\dot{x} &= f(x) + g(x)u\\ y &= h(x)\end{aligned}\right\} \tag{1}$$

donde $x \in \mathbb{R}^n$; $f,\ g,\ h$ son funciones suaves. Diferenciando a y respecto del tiempo, obtenemos

$$\dot{y} = L_f h + L_g h u$$

donde $L_f h,\ L_g h$ son las derivadas de Lie sobre f, g respectivamente. Si $L_g h(x) \neq 0\ \forall\ x \in \mathbb{R}^n$, entonces la ley de control tiene la forma $\alpha(x) + \beta(x)v$, es decir

$$u = \frac{1}{L_g h}\left(-L_f h + v\right)$$

obteniendo el sistema lineal $\dot{y} = v$. Para el caso en que $L_g h(x) = 0$ se diferencia nuevamente la (2) y se obtiene

$$\ddot{y} = L_f^2 h(x) + L_g L_f h(x) u \tag{2}$$

De una manera más general, si llamamos δ al entero más pequeño tal que $L_g L_f^i \equiv 0$ para $i = 0, \dots, \delta - 2$, y $L_g L_f^{\delta-1} h(x) \neq 0\ \forall\ x\ \in \mathbb{R}^n$; entonces la ley de control será

$$u = \frac{1}{L_g L_f^{\delta-1} h(x)}\left(-L_f^{\delta} h(x) + v\right) \tag{3}$$

$$y^{\delta} = v \tag{4}$$

Esta ley de control, llamada control por realimentación por linealización exacta presupone el conocimiento en tiempo continuo de los estados [1] y de los parámetros del proceso. Su aplicación al problema de regulación de glucosa puede verse en [2]. Los estados del proceso se obtienen a partir del uso de observadores de estado [3], [4] pero subsiste el problema de la variación temporal de los parámetros. Esta variación puede llevar a que los términos no lineales no se cancelen totalmente que es la base de la teoría de linealización exacta.
A fin de considerar este problema se presenta aquí una ley de control adaptativa basada en el trabajo de Sastry e Isidori [5] en el que sugieren el uso de un control con parámetros adaptativos, que hace asintóticamente exacta la cancelación de los términos no lineales.

2. CONTROL ADAPTATIVO DE SISTEMAS QUE ADMITEN LINEALIZACIÓN EXACTA

La base de la estimación "on line" es la comparación de la respuesta $y(t)$ del sistema observado con la salida del sistema parametrizado $\hat{y}(\theta, t)$ cuya estructura es la misma que el modelo de la planta. Bajo ciertas condiciones de la entrada, decir que $\hat{y}(\theta, t)$ aproxima a $y(t)$ implica que $\theta(t)$ se aproxima a θ^*que es el vector de parámetros del modelo de la planta.

2.1. SISTEMAS CON GRADO RELATIVO IGUAL A UNO.

Sea un sistema de la forma (1) con $L_g h(x) \neq 0$ lo cual implica que el grado relativo es igual a 1. Si podemos escribir a las funciones f y g en la forma

$$f(x) = \sum_{i=1}^{n} \theta_i^1 f_i(x) \tag{5}$$

$$g(x) = \sum_{j=1}^{n_2} \theta_j^2 g_j(x) \tag{6}$$

donde $\theta_i^1,\ i = 1, \dots, n_1$; $\theta_j^2,\ i = 1, \dots, n_2$; son parámetros desconocidos de las funciones conocidas $f_i(x)$ y $g_j(x)$. En un instante de tiempo t el valor estimado de las funciones f y g son respectivamente

$$\hat{f}(x) = \sum_{i=1}^{n_1} \hat{\theta}_i^1(t) f_i(x) \tag{7}$$

$$\hat{g}(x) = \sum_{j=1}^{n_2} \hat{\theta}_j^2(t) g_j(x) \tag{8}$$

donde $\hat{\theta}_i^1$ y $\hat{\theta}_j^2$ son los valores estimados de θ_i^1 y de θ_j^2 respectivamente en el instante t. En consecuencia, la ley de control está dada por

$$u = \frac{1}{\widehat{L_g h(x)}} \left(-\widehat{L_f h(x)} + v \right)$$

donde $\widehat{L_f h(x)}$, $\widehat{L_f h(x)}$ son los valores observados de $L_g h(x)$ y de $L_f h(x)$ basados en

$$\widehat{L_f h} = \sum_{i=1}^{n_1} \hat{\theta}_i^1(t) L_{fi} h$$

$$\widehat{L_g h} = \sum_{j=1}^{n_2} \hat{\theta}_j^2(t) L_{gj} h$$

donde $\theta \in \mathbb{R}^{n_1+n_2}$ es el vector de parámetros "reales" $(\theta^{1T}, \theta^{2T})^T$, $\hat{\theta} \in \mathbb{R}^{n_1+n_2}$ es el vector estimado de parámetros, y $\phi = \theta - \hat{\theta}$ es el error del parámetro. Entonces, ahora podemos expresar a

$$\dot{y} = v + \phi^{1T} w_1 + \phi^{2T} w_2$$

donde

$$w_1 \in \mathbb{R}^{n_1} \coloneqq \begin{bmatrix} L_{f_1} h \\ \vdots \\ L_{f_{n1}} h \end{bmatrix} \quad \text{y} \quad w_2 \in \mathbb{R}^{n_2} \coloneqq \begin{bmatrix} L_{g_1} h \\ \vdots \\ L_{g_{n1}} h \end{bmatrix} \frac{-\widehat{L_f h} + v}{\widehat{L_g h}}$$

La ley de control usada para el seguimiento es

$$v = \dot{y}_M + \alpha(y_M - y)$$

y la ley de control no lineal resulta ahora

$$u = \frac{1}{\widehat{L_g h}}(-\widehat{L_f h} + \dot{y}_M + \alpha(y_M - y)$$

2.2. SISTEMAS CON GRADO RELATIVO MAYOR A UNO.

Extendiendo los resultados a sistemas que tienen grado relativo δ donde $L_g h = L_g L_f h = L_g L_f^{\delta-2} \equiv 0$ y con $L_g L_f^{\delta-1} \neq 0$, la ley de control no lineal es

$$u = \frac{1}{L_g \widehat{L_f^{(\delta-1)}} h} \left(\widehat{L_f^{(\delta-1)}} h + \hat{v} \right)$$

En ausencia de información precisa sobre $L_f h$, $L_f^2 h$, ..., etc. la componente del control que puede ser sintonizada con las herramientas del control lineal está dada por

$$\hat{v} = y_M^{(\delta)} + \alpha_1 \left(y_M^{(\delta-1)} - \widehat{L_f^{(\delta-1)}} h \right) + \ ... + \alpha_i (y_M - y)$$

2.3. APLICACIÓN AL SISTEMA GLUCOREGULATORIO.

El modelo utilizado para el control es el modelo de mínima de Bergman [6], cuya expresión vectorial es

$$\begin{bmatrix} \dot{G} \\ \dot{X} \\ \dot{I} \end{bmatrix} = \begin{bmatrix} -P_1 G - X(G + G_b) + D(t) \\ -P_2 X + P_3 I \\ -n(I + I_b) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/V \end{bmatrix} u(t), \tag{9}$$

$$y = h(x) = (G + G_b).$$

cuyo grado relativo δ es igual a tres. Para poder expresar a $f(x)$ y a $g(x)$ en la forma (7) y (8) se puede expresar a (9) como

$$f(x) = \sum_{i=1}^{n} \theta_i^1 f_i(x) = \begin{bmatrix} -G & 0 & 0 & 0 & -X(G + G_b) \\ 0 & -X & I & 0 & 0 \\ 0 & 0 & 0 & -(I + I_b) & 0 \end{bmatrix} * \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ n \\ 1 \end{bmatrix} \tag{10}$$

$$g(x) = [0 \quad 0 \quad 1] * \begin{bmatrix} 0 \\ 0 \\ 1/V \end{bmatrix} = [0 \quad 0 \quad 1] * \begin{bmatrix} 0 \\ 0 \\ \beta_3 \end{bmatrix} = g_3\beta_3 \tag{11}$$

la expresión (10) queda como

$$f_1(x) = \begin{bmatrix} -x_1 \\ 0 \\ 0 \end{bmatrix}, f_2(x) = \begin{bmatrix} 0 \\ x_2 \\ 0 \end{bmatrix}, f_3(x) = \begin{bmatrix} 0 \\ x_3 \\ 0 \end{bmatrix}, f_4(x) = \begin{bmatrix} 0 \\ 0 \\ -(x_3 + I_b) \end{bmatrix}, f_5(x) = \begin{bmatrix} -x_2(x_1 + G_b) \\ 0 \\ 0 \end{bmatrix},$$

con $x_1 = G, x_2 = X, x_3 = I,\ \theta_1 = P_1,\ \theta_2 = P_2,\ \theta_3 = P_3,\ \theta_4 = n,\ \theta_5 = 1,$

y realizando las operaciones

$$L_f^3 h = \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \quad \sum_{k=1}^{n_1} \frac{\partial}{\partial x}\left(\frac{\partial L_f}{\partial x} f_i\right) f_i \theta_i^1 \theta_j^1 \theta_k^1 \quad \text{y} \quad L_g L_f h = \sum_{i=1}^{n_2} \quad \sum_{j=1}^{n_1} \frac{\partial}{\partial x}\left(\frac{\partial L_f}{\partial x} f_j\right) g_i \theta_i^2 \theta_j^1$$

se obtiene la ley de control adaptativa del sistema glucoregulatorio

$$u = \frac{1}{L_g \widehat{L_f^{(\delta-1)}} h}\left(\widehat{L_f^{(\delta-1)}} h + \hat{v}\right)$$
$$= \frac{1}{\beta_3\theta_5(x_1 + G_b)}\big((-x_1(\theta_1 - \theta_5 x_2) - x_2\theta_5 + x_3\theta_5)\theta_1\big) - x_2(-\theta_5\theta_1 x_1$$
$$- \theta_5(x_1 + G_b)\theta_2 + x_3\theta_3\big(-\theta_5\theta_1 x_1 - \theta_5(x_1 + G_b)\big) + (x_3 + I_b)\theta_3\big(\theta_5(x_1 + G_b)\big)\theta_4$$
$$- x_2(x_1 + G_b)(\theta_1(\theta_1 - \theta_5 x_2) - x_2\theta_5 + x_3\theta_5)\theta_5 + \hat{v})$$

Por último, se realiza la sintonización de los componentes lineales.

3. CONCLUSIONES

Se presenta aquí una primera aproximación al problema del control adaptativo de procesos biológicos, se destaca como resultado interesante que la componente $1/L_g \widehat{L_f^{(\delta-1)}} h$ no involucra al parámetro θ_3 como si ocurre con la versión invariante en el tiempo lo cual simplifica notablemente la sintonización del control. En una próxima etapa se compararán las simulaciones realizadas con datos experimentales de mediciones continuas de glucosa, las cuales son extremadamente difíciles de obtener.

REFERENCIAS

[1] C. D´ATTELLIS, *"Introducción a los sistemas no lineales de control y sus aplicaciones."*, AADECA, 1992.

[2] G. COCHA, V. COSTANZA, C. D´ATTELLIS,. "*Control No Lineal de la Diabetes Mellitus*". Anales de la RPIC, Rosario, 2009.

[3] G. COCHA, V. COSTANZA, C. D´ATTELLIS, "*Observadores No Lineales en el Control de la Diabetes Mellitus*" Anales 1º Congreso de BioIngeniería, *Costa Rica 2009*. Vol 1, 325-331.

[4] G. COCHA, M. PODESTÁ, C. D´ATTELLIS, "*Regulación automática de glucosa en pacientes insulinodependientes*". Anales del Congreso de Ingeniería Clínica y Bioingeniería, Paraná, Octubre de 2010.

[5] S. SARTRY, A. ISIDORY, *"Adaptive control of linearizable systems."*, IEEE Transactions on Automatic Control, Vol. 34, NO. 11**,** pp 1123-1131, Noviembre 1989.

[6] E. H.C: Morris, B.O´Reilly, D Streja *"A New Bifasic Minimal Model"*, Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco USA, Septiembre 1-5, 2004.

SISTEMA ROBUSTO AL RUIDO PARA LA DETECCIÓN DE FRECUENCIA GLÓTICA BASADO EN LA REPRESENTACIÓN DE SINTONÍA

Matías L. Capeletto y Patricia A. Pelle

Facultad de Ingeniería de la Universidad de Buenos Aires - Av. Paseo Colón 850 - Buenos Aires - Argentina, www.fi.uba.ar - mcape@fi.uba.ar, ppelle@fi.uba.ar

Resumen: En este trabajo se presenta un nuevo sistema para la detección de frecuencia glótica basado en la Representación de Sintonía. Esta representación, fundamentada en evidencia biológica del funcionamiento del sistema auditivo humano, se implementa mediante un banco de filtros cocleares seguidos de lazos de enganche de fase (PLLs) y permite analizar la estructura armónica de la señal de habla en forma robusta al ruido. El sistema de detección de frecuencia glótica propuesto utiliza la Representación de Sintonía de la señal de habla para construir una nueva señal que solo posee la información correspondiente a la estructura armónica de la señal original. Esta señal es luego usada como entrada de un algoritmo de detección tradicional, logrando un sistema robusto al ruido. Para probar la validez de la propuesta se utilizaron bases de datos estándar para realizar mediciones de desempeño. Estos experimentos presentan resultados prometedores para el diseño propuesto.

Palabras clave: *frecuencia glótica, PLL, Representación de Sintonía*
2000 AMS Subject Classification: 60G35, 93E10, 94A12

1. INTRODUCCIÓN

La señal de habla es una señal no estacionaria que se divide en porciones sonoras y no sonoras. Las porciones sonoras pueden considerarse señales cuasi-periódicas de frecuencia fundamental lentamente variable, que es determinada por los órganos fonadores, en particular la glotis, y es llamada frecuencia glótica, o f_o. Se han desarrollado gran cantidad de algoritmos de detección de f_o [4] [10] [2] [11]. Sin embargo es en la actualidad un problema abierto, especialmente debido al pobre desempeño que estos métodos poseen en condiciones de señal a ruido (Signal to Noise Ratio, SNR) adversas.

En trabajos anteriores [6] [5] [7] se exploró una nueva herramienta de análisis de la señal de habla llamada *Representación de Sintonía* basada en el sistema auditivo humano. La misma utiliza lazos de enganche de fase (PLL, Phase Locked Loop), dispositivos de control que generan una salida sinusoidal cuya fase intenta copiar la fase de la entrada en forma robusta al ruido. En este trabajo se presenta un nuevo diseño para la detección de f_o basado en esta representación, que reconstruye una nueva señal en la que se concentra toda la información armónica de los segmentos sonoros. Esta nueva señal puede ser analizada luego por un detector básico de f_o, obteniéndose desempeños muy estables cuando la SNR se vuelve más adversa.

El resto del trabajo se organiza del siguiente modo: en la sección 2 se describe la Representación de Sintonía. En la sección 3 se presenta el nuevo sistema de detección de f_o. En la sección 4 se describen los experimentos y las bases de datos utilizadas para la medición del desempeño. Por último, en la sección 5 se presentan las conclusiones del trabajo.

2. LA REPRESENTACIÓN DE SINTONÍA

En la cóclea, parte principal del sistema auditivo periférico humano, se observa una descomposición tonotópica del sonido que puede ser vista en primera aproximación como la aplicación de un banco de filtros sobre la señal de habla [3]. Estos filtros poseen una forma asimétrica con respecto a su frecuencia pico como se muestra en la Figura 1. A partir de experimentos biológicos sobre las fibras del nervio auditivo se mostró que el sistema no puede ser explicado utilizando directamente un modelo lineal, observándose fenómenos de sincronía con la fase de los armónicos de la señal de habla [9].

Basada en esta evidencia, la Representación de Sintonía comienza con una descomposición espectral mediante un banco de N filtros cocleares. La respuesta en frecuencia de estos filtros esta basada en el trabajo de [12]. Las frecuencias características se elijen siguiendo la escala Mel, con valores que van desde

los 100 Hz hasta los 2000 Hz. La salida de cada filtro es luego aplicada a un PLL, que se engancha en fase al armónico con más energía dentro de la banda de paso. Cada PLL entrega tres señales como salida: $v_{si}(n)$, una señal sinusoidal cuya fase intenta copiar la fase del armónico que este canal este siguiendo; $frec_i(n)$, una estimación de la frecuencia instantánea de $v_{si}(n)$; $lock_i(n)$, una indicación de grado de enganche del PLL. En la Figura 2 se puede ver un diagrama del sistema utilizado para generar la Representación de Sintonía, que es el conjunto de las señales $v_{si}(n)$, $frec_i(n)$ y $lock_i(n)$.

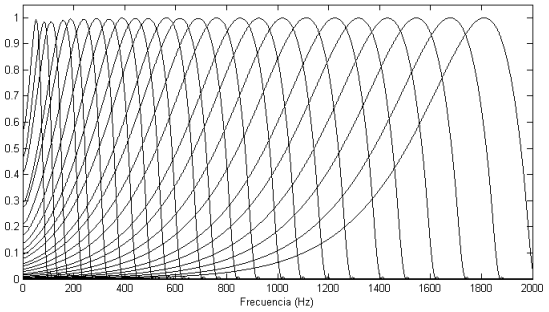


Figura 1: Banco de filtros cocleares

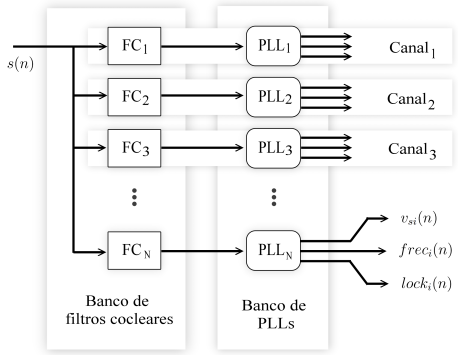


Figura 2: Sistema de Representación de Sintonía

En la Figura 3 se muestra el espectrograma de una señal de habla, sobre el cual se grafican las señales $frec_i(n)$. A partir de los 300 Hz, los filtros cocleares permiten el paso de algún armónico de la señal de habla y los PLLs se enganchan con la fase del mismo, y por lo tanto siguen las variaciones de frecuencia del armónico al que se encuentra sincronizado.

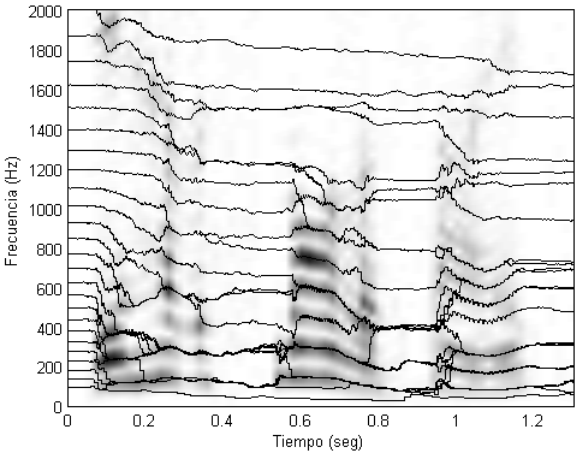


Figura 3: $frec_i(n)$ sobre espectrograma

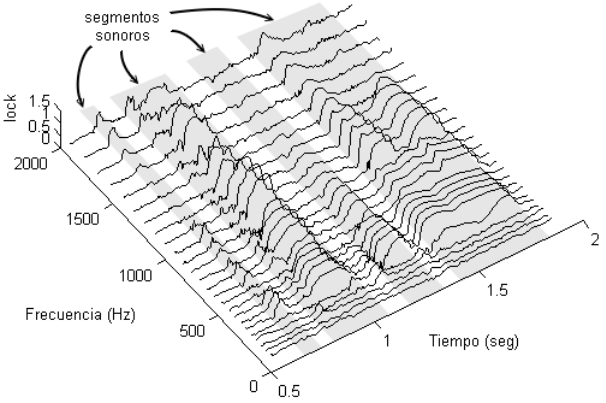


Figura 4: Indicación de enganche, $lock_i(n)$

Para la misma señal de habla, en la Figura 4 se muestran las señales de $lock_i(n)$, que han sido dispuestas en el eje de frecuencia según la frecuencia pico del filtro coclear correspondiente. Los primeros canales presentan una señal $lock_i(n)$ cercana a cero indicando que no se encuentran enganchados. En los segmentos sonoros puede verse un aumento de las señales $lock_i(n)$ con valores cercanos a uno para los canales que están enganchados con algún armónico.

3. Descripción del nuevo sistema de detección de f_o

En este diseño toda la información de la Representación de Sintonía que es de interés para la detección de frecuencia glótica es concentrada en esta señal $x_{rs}(n)$, que luego puede ser analizada por un detector de f_o básico obteniendo mejores resultados que al trabajar sobre la señal original.

En la Figura 5 se puede ver un diagrama de bloques del sistema propuesto. El mismo consta de cuatro etapas: en la primera etapa, la señal de habla $s(n)$ es descompuesta en los N canales de la Representación de Sintonía como fue explicado en la sección 2; en la segunda etapa las señales $lock_i(n)$ y $frec_i(n)$ son utilizadas para definir para cada canal i las señales de peso $p_i(n)$ que miden el aporte de cada uno de los canales en la representación final; en una tercera etapa se construye la señal $x_{rs}(n)$ sumando las señales $v_{si}(n)$ multiplicadas por los pesos $p_i(n)$ calculados en la etapa anterior; por último, se utiliza un detector básico de f_o sobre $x_{rs}(n)$ para generar las estimaciones de f_o.

De acuerdo a esto la señal $x_{rs}(n)$ será entonces una suma pesada con pesos $p_i(n)$ de las señales $v_{si}(n)$:

$$x_{rs}(n) = \sum_{i=1}^{N} p_i(n) v_{si}(n) \tag{1}$$

Dos propiedades son deseables en la señal $x_{rs}(n)$: eliminar la mayor cantidad de ruido posible, dejando solamente la información armónica; y obtener un espectro de deltas lo más blanco posible para facilitar la tarea del detector de f_o. Con estos objetivos presentes, sólo serán incluidos los aportes de los PLLs claramente enganchados. Los pesos $p_i(n)$ correspondientes a cada canal se calcularán a partir de la señal $lock_i(n)$, definiendo un límite mínimo a partir del cual se considera que un PLL esta enganchado y un límite tope para ayudar a lograr el blanqueo del espectro. Debido a que varios canales pueden estar siguiendo al mismo armónico, el aporte del PLL con mayor $lock_i(n)$ en los canales enganchados en el mismo armónico será incluido, descartando los demás definiendo $p_i(n)$ igual a cero en estos canales.

En la Figura 6 se muestran los espectrogramas de una señal de habla con SNR 0dB y de la señal $x_{rs}(n)$ obtenida, en donde se puede apreciar la reducción del ruido, mientras que simultáneamente se observa que la estructura armónica se mantiene y con el blanqueo espectral buscado.

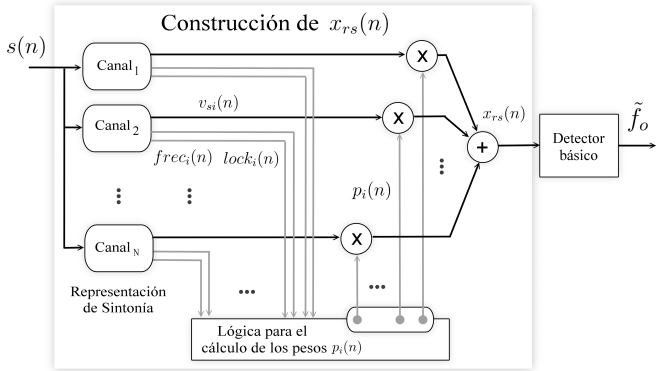


Figura 5: Esquema nuevo sistema

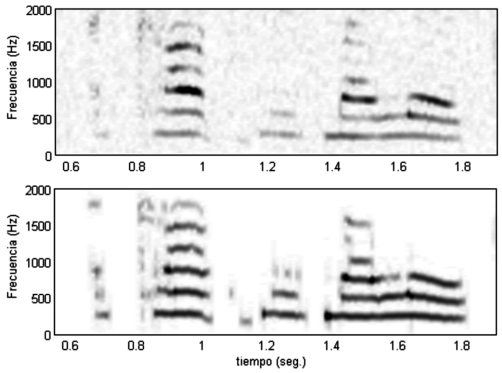


Figura 6: Señal original y $x_{rs}(n)$, espectrogramas

La elección del detector básico que estimará el valor de f_o a partir de la señal $x_{rs}(n)$ no es crucial en nuestro planteo. En el presente trabajo se utilizó un detector de corto tiempo basado en transformada doble con operación alineal de modulo (ver [4]).

4. Experimentos y resultados

Para las mediciones de desempeño se utilizaron dos bases de datos: la base *Bagshaw* [1] y la base *Keele* [8], totalizando aproximadamente 15 minutos de grabación digitalizada a 20KHz con 16 bits. Estas bases son acompañadas por una referencia de frecuencia glótica. A partir de éstas se construyeron nuevas señales de prueba, adicionando ruido blanco de la base Noisex de la Signal Processing Information Base (SPIB), de la Universidad de Rice, USA, en distintos niveles de relación de señal a ruido (Signal to Noise Ratio, SNR).

El primer indicador de desempeño es la tasa de error grosero (*GER*, Gross Error Rate), que mide el porcentaje de estimaciones en las que se ha cometido un error relativo mayor al 20 % con respecto a la referencia. Para medir la precisión se realiza una estadística de errores finos: el promedio (MAE) y el desvío (STD) del error absoluto de muestras que no posean error grosero.

La Tabla 1 muestra el resultado del algoritmo propuesto que llamaremos RS-PDA, en comparación con el algoritmo PLL-PDA (también basado en la Representación de Sintonía [7]) y el algoritmo *RAPT* [10], uno de los algoritmos más populares basado en la función de autocorrelación. Analizando el GER, la medida de mayor importancia en nuestro caso, se puede ver que para condiciones favorables de SNR, el algoritmo PLL-PDA logra un mejor desempeño que esta propuesta. Sin embargo, el desempeño del algoritmo RS-PDA es más estable a medida que la SNR se vuelve más adversa, llegando a superar en más de un uno por ciento al algoritmo PLL-PDA para SNR de 0 dB. Además, la complejidad computacional es mucho más baja en este nuevo diseño, aumentando el atractivo de este enfoque. Contrariamente a estos algoritmos basados en la Representación de Sintonía, el desempeño del algoritmo *RAPT* se degrada rápidamente cuando la SNR se vuelve más adversa.

	RS-PDA			PLL-PDA			*RAPT*		
SNR(dB)	GER %	MAE(Hz)	STD(Hz)	GER %	MAE(Hz)	STD(Hz)	GER %	MAE(Hz)	STD(Hz)
30	**3.30**	4.44	7.12	**1.92**	2.80	4.14	**6.30**	2.91	4.12
20	**3.32**	4.43	7.10	**2.02**	2.80	4.11	**8.10**	2.83	3.93
10	**3.22**	4.46	7.14	**2.35**	2.89	4.19	**18.35**	2.63	3.36
0	**3.83**	4.63	7.38	**4.87**	3.20	4.41	**61.19**	2.56	2.73

Tabla 1: Evaluación sobre bases *Bagshaw* y *Keele*, con ruido blanco adicionado en distintas SNR

5. CONCLUSIONES

El desempeño obtenido por el diseño nos permite concluir que esta transformación de la señal de habla basada en la Representación de Sintonía es claramente ventajosa en los segmentos sonoros comparada con el mejor algoritmo del estado-del-arte, RAPT, y es parcialmente superadora frente al otro algoritmo basado en la Representación de Sintonía, PLL-PDA. El análisis propuesto de la señal parece corresponderse adecuadamente al objetivo propuesto, y promete ser un punto de partida de mejoras adicionales, como por ejemplo evaluar otros detectores básicos, o agregar una etapa de postprocesamiento para llevar el GER aún a valores más bajos.

REFERENCIAS

[1] P. C. Bagshaw, S. M. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of F0 contours for computer and intonation teaching. In *Eurospeech 1993*, pages 1003–1006.

[2] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Institute of Phonetic Sciences, Amsterdam*, 1993.

[3] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(1):115–132, Jan. 1994.

[4] Wolfgang J. Hess. Pitch and voicing determination of speech with an extension toward music signals. In Jacob Benesty, M. Mohan Sondhi, and Yiteng (Arden) Huang, editors, *Springer Handbook of Speech Processing*, pages 181–212. Springer Berlin Heidelberg, 2008.

[5] P.A. Pelle. A robust pitch extraction system based on phase locked loops. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I, Toulouse, France, May. 2006.

[6] Patricia Alejandra Pelle and Matias Capeletto. Pitch estimation using phase locked loops. In *8th European Conference on Speech communication and technology (EUROSPEECH 2003)*, Geneva, Switzerland, Sep. 1-4 2003.

[7] Patricia Alejandra Pelle and Claudio Francisco Estienne. A pitch extraction system based on phase locked loops and consensus decision. In *International Conference on Speech communication and technology (INTERSPEECH 2007)*, Antwerp, Belgica, Ago. 27-31 2007. ISSN 1990-9772.

[8] F. Plante, G. Meyer, and W. A. Ainsworth. A pitch extraction reference database. In *Eurospeech 1995*, pages 837–840.

[9] J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology*, 30(4):769–793, 1967.

[10] D. Talkin. *Speech Coding and Synthesis*, chapter A robust algorithm for pitch tracking (RAPT), page 495–518. Elsevier, 1995.

[11] Chao Wang and S. Seneff. Robust pitch tracking for prosodic modeling in telephone speech. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1343–1346 vol.3, 2000.

[12] Kuansan Wang and S. Shamma. Zero-crossing and noise suppression in auditory wavelet transformations. Technical Report ISR; TR 1992-94, Systems Research Center and Department of Electrical Engineering, University of Maryland, 1992.

BIFURCACIONES EN LA ECUACIÓN DE VAN DER POL REALIMENTADA CON RETARDO

Andrea Bel[†‡] y Walter Reartes[†]

[†]*Departamento de Matemática, Universidad Nacional del Sur, Alem 1253, 8000 Bahía Blanca, Buenos Aires, Argentina,* {*andrea.bel,reartes*}*@uns.edu.ar, www.uns.edu.ar*
[‡]*CONICET, Argentina*

Resumen: En este trabajo se estudia la ecuación de van der Pol realimentada con retardo usando el método de análisis homotópico (HAM). El trabajo está centrado en el cálculo de las soluciones periódicas y las bifurcaciones asociadas a estas, Hopf, Hopf doble, Neimark-Sacker, etc.. Se analiza en particular una bifurcación de Hopf doble y el comportamiento del sistema en un entorno de la misma.

Palabras clave: *bifurcaciones, Hopf, Hopf doble, van der Pol, HAM*
2000 AMS Subject Classification: 34K07 - 34K13 - 34K18

1. INTRODUCCIÓN

Consideramos la ecuación de van der Pol realimentada con retardo

$$x''(t) + \epsilon(x^2(t) - 1)x'(t) + x(t) = d\epsilon x(t - \tau), \tag{1}$$

donde ϵ, d son constantes reales positivas, y $\tau > 0$ representa el retardo. Utilizaremos el método de análisis homotópico [3, 4] para construir expresiones analíticas de las soluciones periódicas de la misma. Al estudiar estas soluciones y su dinámica nos encontramos con distinto tipo de bifurcaciones [2], algunas se relacionan con los equilibrios, la bifurcación de Hopf, bifurcación de Hopf doble, etc., y otras son bifurcaciones que sufre el propio ciclo, *fold, flip,* y Neimark-Sacker. El análisis de estas bifurcaciones es esencial para determinar el comportamiento del sistema considerado al variar los parámetros.

Este trabajo se organiza de la siguiente manera. En la Sec. 2 se describe el método de análisis homotópico enfocado en la búsqueda de soluciones periódicas. En la Sec. 3 se considera la estabilidad de equilibrios y ciclos, y se estudia en particular el entorno de un punto de Hopf doble no resonante. Finalmente, se detallan las conclusiones en la Sec. 4.

2. MÉTODO DE ANÁLISIS HOMOTÓPICO

Supongamos que existe una solución periódica con amplitud a y frecuencia ω de la ecuación (1). Para hallar una expresión analítica de la misma utilizando el método de análisis homotópico es conveniente normalizar la ecuación. Esto se logra reescalando las variables t y x de manera que la solución periódica correspondiente resulte con amplitud y frecuencia 1, seguiremos llamando t y x a las variables. La ecuación que resulta de la normalización es

$$\omega^2 x''(t) + \epsilon\omega(a^2x^2(t) - 1)x'(t) + x(t) = d\epsilon x(t - \omega\tau), \tag{2}$$

donde $\epsilon, d, \tau \in \mathbb{R}^+$. Buscaremos una solución de (2) que verifique: $x(0) = 1$ y $x'(0) = 0$.

Consideremos la homotopía definida por

$$\mathcal{H}[\phi, \Omega, A, \epsilon, d, \tau, h, q] = (1 - q)\mathcal{L}[\phi - x_0] - qh\mathcal{N}[\phi, \Omega, A, \epsilon, d, \tau], \tag{3}$$

donde $h \neq 0$ es un parámetro de control de convergencia, x_0 es una aproximación inicial que verifica $x_0(0) = 1$ y $x_0'(0) = 0$, $\mathcal{L}$ es el operador lineal

$$\mathcal{L}[\psi(t, q)] = \frac{\partial^2\psi(t, q)}{\partial t^2} + \psi(t, q),$$

y $\mathcal{N}$ es un operador que corresponde a la ecuación (2), y está definido por

$$\mathcal{N}[\phi(t,q),\Omega(q),A(q),\epsilon,d,\tau] = \Omega^2(q)\frac{\partial^2\phi(t,q)}{\partial t^2} + \phi(t,q) + \\ + \epsilon\,\Omega(q)\Big(A^2(q)\phi^2(t,q)-1\Big)\frac{\partial\phi(t,q)}{\partial t} - d\epsilon\phi(t-\Omega\tau,q).$$

Sea $\mathcal{H}\equiv 0$ para $q\in[0,1]$. Ya que

$$\mathcal{H}[\phi,\Omega,A,\epsilon,d,\tau,h,0] = \mathcal{L}[\phi(t,0)-x_0], \quad y \quad \mathcal{H}[\phi,\Omega,A,\epsilon,d,\tau,h,1] = -h\mathcal{N}[\phi(t,1),\Omega(1),A(1),\epsilon,d,\tau],$$

observamos que $\phi(t,0)-x_0(t)$ es solución de $\mathcal{L}[\psi]=0$, luego $\phi(t,0)-x_0(t) = c_1\cos(t)+c_2\,\mathrm{sen}(t)$ con $c_1,c_2\in\mathbb{R}$. Además, $\phi(t,1)$, $\Omega(1)$ y $A(1)$ determinan una solución de la ecuación $\mathcal{N}[\phi,\Omega,A,\epsilon,d,\tau]=0$. De acuerdo a la definición de $\mathcal{N}$, $\phi(t,1)$ corresponde a la solución periódica que buscamos, siendo $\Omega(1)$ la frecuencia y $A(1)$ la amplitud de la misma.

Las condiciones que verifica la solución de la ecuación (2) implican que $\phi(0,1)=1$ y $\frac{\partial\phi(t,1)}{\partial t}\big|_{t=0}=0$, consideramos

$$\phi(0,q)=1, \qquad \frac{\partial\phi(t,q)}{\partial t}\Big|_{t=0}=0, \tag{4}$$

para todo $q\in[0,1]$. Estas condiciones junto con las impuestas a x_0 implican que $\phi(t,0)=x_0(t)$.

Si las funciones ϕ, Ω y A son analíticas para $q\in[0,1]$ existirán los desarrollos

$$\phi(t,q)=\sum_{k=0}^{+\infty}x_k(t)q^k, \quad \Omega(q)=\sum_{k=0}^{+\infty}\omega_k q^k, \quad A(q)=\sum_{k=0}^{+\infty}a_k q^k. \tag{5}$$

Luego, considerando $q=1$, se obtiene $x(t)=\phi(t,1)$, $\omega=\Omega(1)$ y $a=A(1)$. Las series anteriores dependerán del parámetro h, y serán convergentes si es posible fijar un valor adecuado para el mismo. Calculamos los términos k-ésimos, para $k\geq 1$ utilizando las ecuaciones

$$\mathcal{L}[x_k(t)-(1-\delta_{1k})x_{k-1}(t)] = h\frac{1}{(k-1)!}\,\frac{\partial^{k-1}\mathcal{N}_q[\phi,\Omega,A,\epsilon,d]}{\partial q^{k-1}}\Big|_{q=0}, \tag{6}$$

y las condiciones iniciales $x_k(0)=0$, $x_k'(0)=0$. Estas ecuaciones se obtienen reemplazando (5) en (3), derivando m veces respecto de q y considerando $q=0$ (recordemos que $\mathcal{H}\equiv 0$ para $q\in[0,1]$). De manera análoga, a partir de (4) se obtienen las condiciones iniciales mencionadas. Además, se puede probar que (6) define un problema de valores iniciales para cada término x_k en función de los k términos anteriores.

Como el objetivo es encontrar soluciones periódicas, tanto la aproximación inicial $x_0(t)$ como cada términos $x_k(t)$ en la serie de $x(t)$ deberán ser periódicos. Luego, al resolver las ecuaciones (6) debemos asegurar que no aparezcan términos no periódicos de la forma $t\cos(t)$ o $t\,\mathrm{sen}(t)$. Esto se logrará fijando valores adecuados de ω_{k-1} y a_{k-1} en cada paso k, obteniendo así los desarrollos en serie de ω y a.

En nuestro caso, teniendo en cuenta las condiciones para las soluciones de (2) consideramos $x_0(t)=\cos(t)$, el lado derecho de la ecuación (6) para $k=1$ resulta

$$(1-\omega_0^2-d\epsilon\cos(\omega_0\tau))\cos(t)+\epsilon\omega_0\Big(1-\frac{a_0^2}{4}-\frac{d}{\omega_0}\,\mathrm{sen}(\omega_0\tau)\Big)\,\mathrm{sen}(t)-\epsilon\omega_0\frac{a_0^2}{4}\,\mathrm{sen}(3t),$$

por lo tanto ω_0 y a_0 deben ser soluciones del sistema

$$1-\omega_0^2-d\epsilon\cos(\omega_0\tau)=0, \quad 1-\frac{a_0^2}{4}-\frac{d}{\omega_0}\,\mathrm{sen}(\omega_0\tau)=0. \tag{7}$$

Para $k\geq 1$ las condiciones que definen ω_k y a_k son lineales y pueden resolverse con facilidad.

Una vez determinados los valores de ω_0 y a_0, podemos calcular $x_1(t)$, y plantear la ecuación para $x_2(t)$, a partir de la cual obtendremos un sistema, en este caso lineal, que nos permitirá determinar ω_1 y a_1.

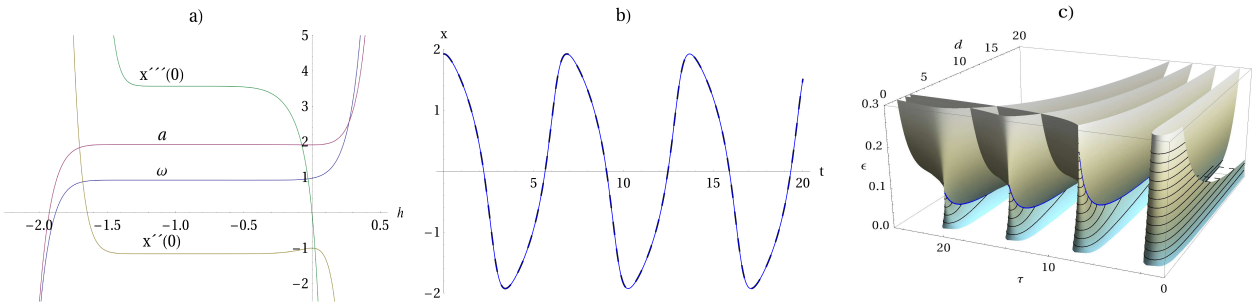


Figura 1: a) Polinomios en función de h. b) Ciclo: (–) expresión de orden 15, (- -) aproximación numérica. c) Superficie S en el espacio (d, τ, ϵ), (– azul) curvas de autointersecciones, (- -) segmentos en que $d\epsilon = 1$.

Continuando de esta manera podemos obtener expresiones analíticas de la solución periódica, y desarrollos de la frecuencia y la amplitud hasta un orden de aproximación arbitrario en función de h.

Sólo resta determinar un valor adecuado para h. Las aproximaciones de ω y a serán polinomios en h, y lo mismo sucederá con $x(t)$ y sus derivadas para t fijo. Como se muestra en el libro de Liao [3] la observación del comportamiento de estos polinomios permite seleccionar un valor adecuado para h. En los valores de h para los cuales la serie converge, dichos polinomios tenderán en el límite, cuando el orden tiende a infinito, a un valor independiente de h. Así, graficando los polinomios podremos tener una idea aproximada del lugar en que se encuentran estas regiones y seleccionar un valor apropiado para h. En la Figura 1 a) se muestran varios de estos polinomios para una expresión de orden 15 de un ciclo, se distinguen claramente las zonas mencionadas. En b) se grafica el ciclo correspondiente en función de t para $h = -0{,}8$, junto con la aproximación numérica del mismo calculada con Mathematica, se observa la similitud entre ambos. Esto ocurre en general cuando comparamos los ciclos con los obtenidos numéricamente, incluso, en ciertos valores de los parámetros, para expresiones de orden menor al considerado antes.

3. Análisis de estabilidad y bifurcaciones

Primero, notemos que $x \equiv 0$ es el único equilibrio de la ecuación (2) para todos los valores de los parámetros que verifiquen $d\epsilon \neq 1$. Si $d\epsilon = 1$ la ecuación tiene infinitos equilibrios no aislados.

La condición $a_0 = 0$ en el sistema (7) define una superficie S en el espacio de parámetros (d, τ, ϵ). En los puntos de la superficie se cumple la condición necesaria para que exista bifurcación de Hopf, esto es, que exista un par de autovalores característicos de la forma $\lambda = \pm i\omega$, $\omega > 0$. Además, en las curvas determinadas por las autointersecciones de la superficie existirán dos pares de autovalores complejos conjugados $\lambda = \pm i\omega_1$ y $\lambda = \pm i\omega_2$, con $\omega_{1,2} > 0$, condición necesaria para la existencia de la bifurcación de Hopf doble.

De acuerdo a las restricciones impuestas a los parámetros, la superficie S se encuentra en el primer octante del espacio (d, τ, ϵ), y se puede probar que presenta autointersecciones sólo si $\epsilon < \sqrt{2}$. Estudiaremos en particular aquellas zonas de la superficie S en las que el equilibrio trivial presenta un cambio en su estabilidad. En la Figura 1 c) se muestra la superficie S, y se recubren con curvas las zonas consideradas, para valores de los parámetros en el interior de las mismas el equilibrio $x \equiv 0$ es estable. Además, en la misma figura se pueden observar las curvas de autointersecciones y algunos de los segmentos en que $d\epsilon = 1$.

En las zonas en que existan soluciones periódicas su estabilidad será calculada utilizando la expresión analítica obtenida. Considerando una perturbación de la misma, reemplazando en (2) y linealizando se obtiene una ecuación con retardo con coeficientes periódicos. A partir de ésta es posible definir un operador de monodromía, similar al que se obtiene en el caso de ODEs (ver [1]). Sin embargo, este operador actúa sobre el espacio infinito dimensional en que se encuentran las condiciones iniciales de (2), y esto dificulta el estudio de sus autovalores. Se verifica que siempre existe un autovalor trivial 1, y que el ciclo será asintóticamente estable si los autovalores distintos del trivial tienen módulo menor a 1, e inestable si alguno de ellos tiene módulo mayor a 1. Para conocer estos autovalores construímos un operador finito dimensional que aproxima al operador de monodromía utilizando polinomios de Chebyshev.

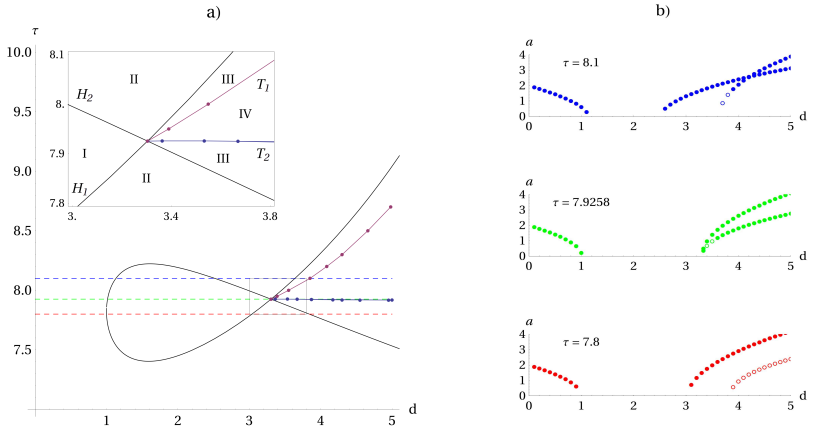


Figura 2: a) Cercanías del punto de Hopf doble. b) Amplitudes para valores de τ constantes indicados en a).

3.1. Cerca del punto de Hopf doble

Conociendo los ciclos y su estabilidad podemos detectar y analizar bifurcaciones en las que ellos intervienen. En este caso nos centramos en el análisis de una bifurcación de Hopf doble.

Fijando $\epsilon = 0{,}1$, una bifurcación de Hopf doble no resonante ocurre cuando $d = 3{,}30471$ y $\tau = 7{,}92541$. Los pares de autovalores correspondientes son $\pm i1{,}14456$ y $\pm i0{,}82461$. Analizando el comportamiento de los ciclos cercanos al punto de Hopf doble se puede observar que éstos sufren bifurcaciones de Neimark-Sacker (NS), lo que determina la aparición de toros 2D. Los puntos en los que se observa la bifurcación NS pertenecen a dos curvas T_1 y T_2 en el espacio (d, τ) que se muestran en la Figura 2 a), en la misma se pueden observar además las curvas de Hopf H_1 y H_2. La dinámica observada corresponde a uno de los conocidos como casos simples [2]. Resumimos lo que sucede al variar los parámetros d y τ: en la región I el equilibrio es estable, pasando a la región II en las curvas H_1 y H_2 el sistema sufre una bifurcación de Hopf supercrítica, y así el ciclo que nace es estable, en III el ciclo estable se mantiene y surge un nuevo ciclo esta vez inestable, que se estabiliza por medio de una bifurcación NS, en la región IV ambos ciclos son estables y el toro 2D que surge es inestable. En la Figura 2 b) muestra las amplitudes de los ciclos para varios valores constantes de τ, los resultados coinciden con la dinámica explicada antes.

4. Conclusiones

En este trabajo se utilizó el método de análisis homotópico para hallar expresiones analíticas de las soluciones periódicas de una ecuación diferencial con retardo. El método permite calcular estas soluciones con gran precisión. Este hecho se observó en todos los cálculos que se hicieron para la realización de este trabajo. La extensión del resumen no nos permite abundar en más ejemplos que esperamos publicar próximamente en otro trabajo. Además, a partir de las expresiones de las soluciones se determina la estabilidad de las mismas. Así, para valores arbitrarios de los parámetros es posible calcular las soluciones periódicas existentes y su estabilidad. Esto nos brinda la posibilidad de obtener un panorama general del comportamiento de los ciclos de la ecuación al variar los parámetros y facilita el estudio de bifurcaciones en las que ellos intervienen. En particular, se analizan las soluciones periódicas en un entorno de una bifurcación de Hopf doble no resonante.

Referencias

[1] J. Hale, S. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, 1993.

[2] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, 1995.

[3] S.J. Liao, *Beyond perturbation: introduction to homotopy analysis method*, CHAPMAN & HALL/CRC, 2004.

[4] S.J. Liao, *Notes on the homotopy analisys method: Some definitions and theorems*, Commun Nonlinear Sci Numer Simulat 14 (2009), pp.983-997.

[5] S. Ma, Q. Lu, Z. Feng, *Double Hopf bifurcation for van der Pol-Duffing oscillator with parametric delay feedback control*, Journal of Mathematical Analysis and Applications 338 (2008), pp. 993-1007.

CARACTERIZACIÓN DE FORMAS NORMALES DE BIFURCACIONES DE HOPF EN EL DOMINIO FRECUENCIA

A. TORRESI[♭], G. CALANDRINI[♭,†], P. BONFILI[♭♭] y J. MOIOLA[†]

[♭]*Departamento de Matemática, Universidad Nacional del Sur, Bahía Blanca, (B8000CPB), Argentina*
[†]*Instituto de Investigaciones de Ingeniería Eléctrica, IIIE (UNS-CONICET) , Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Bahía Blanca, (B8000CPB), Argentina*
[♭♭]*Departamento de Matemática, Universidad Nacional San Juan Bosco, Trelew, (B9100CPB), Argentina*

Resumen: Utilizando métodos frecuenciales y teoría de singularidades se determinan condiciones sobre los coeficientes de la ecuación de bifurcación de alto orden para determinar las formas normales de ciertos tipos de bifurcaciones dinámicas. Se aplican los resultados obtenidos en un circuito electrónico con diodo túnel.

Palabras clave: *bifurcación de Hopf degenerada, dominio frecuencia, formas normales*
2000 AMS Subject Classification: 34F10 - 37G05

1. INTRODUCCIÓN

La existencia y estabilidad de oscilaciones de un sistema autónomo no lineal planteado como un sistema realimentado $\mathcal{S}$ en el dominio frecuencia están formuladas en el Teorema gráfico de Hopf [4, 5]. En éste se determina la existencia de un único ciclo límite, donde la forma normal que vincula la amplitud de la oscilación con el parámetro presenta el clásico diagrama de bifurcación de forma cuadrática. En los casos en que alguna de las hipótesis del teorema falle aún es posible determinar la existencia de uno o más ciclos límites. En este trabajo se determinan condiciones de definición y no degeneración sobre los coeficientes de la ecuación de bifurcación de alto orden [6] para caracterizar las formas normales de bifurcaciones dinámicas en el dominio frecuencia, cuando alguna de las hipótesis clásicas no se verifica. Se utiliza la teoría de singularidades para la clasificación de los diagramas de bifurcación Z_2-equivalentes [1, 2]. En las aplicaciones, se determinan todos los diagramas no equivalentes de bifurcaciones de órbitas periódicas que se obtienen perturbando la forma normal singular. Los resultados se aplican en un circuito electrónico con diodo túnel.

1.1. PRELIMINARES

Sea el sistema dinámico no lineal planteado como un sistema realimentado $\mathcal{S}$

$$\dot{x} = A(\mu)x + Bu, \quad y = Cx, \quad u = -f(y,\mu),$$

$A(\mu) \in \mathbb{R}^{n\times n}$ y $det\,(sI - A(\mu)) \neq 0$, $\mu \in \mathbb{R}$ es el parámetro de bifurcación, $f : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^l$, es C^k con $k > 3$ y $f(0,\mu) = 0$, u es el vector de entradas e y el vector de las salidas, $C \in \mathbb{R}^{m\times n}$ y $B \in \mathbb{R}^{n\times l}$.

Sea $\hat{y}$ la solución de equilibrio de $G(0,\mu)f(y,\mu) + y = 0$. Linealizando en el punto de equilibrio $\hat{y}$ se obtiene la matriz $J = (Df)_{\hat{y}}$, luego $G(s,\mu)J$ es la función de transferencia de lazo abierto.

$\mathcal{S}$ satisface, **(H1)**$_{df}$: Existe una única función característica de $G(i\omega,\mu)J$, que se nota $\hat{\lambda}(\omega,\mu) \in \mathbb{C}$, tal que para una única frecuencia ω_0 y un valor crítico del parámetro μ_0 se verifica que $\hat{\lambda}(\omega_0,\mu_0) = -1$.

A continuación se presenta una versión extendida del teorema de bifurcación de Hopf gráfico, donde se obtiene una ecuación de bifurcación de alto orden [6].

Teorema 1 *Sea $\mathcal{S}$ un sistema que verifica **(H1)**$_{df}$, la ecuación de bifurcación de órbitas periódicas de alto orden en el dominio frecuencia es*

$$\theta\,\Phi(z,\omega,\mu) = \theta\left(\hat{\lambda}(\omega,\mu) + 1 + \sum_{k=1}^{q} z^k \xi_k(\omega,\mu) + \mathcal{O}(z^{q+1})\right) = 0, \tag{1}$$

donde $z = \theta^2$ y $\xi_k = \xi_k(\omega,\mu) \in \mathbb{C}$. Las soluciones no nulas de (1) están en correspondencia uno a uno con las soluciones periódicas de pequeña amplitud θ del sistema $\mathcal{S}$ con período cercano a $2\pi/\omega_0$.

Observación: Las definiciones de $\xi_k(\omega,\mu) \in \mathbb{C}$ con $k = 1, \cdots$ se encuentran en el desarrollo de la demostración del teorema y sus expresiones $\forall k$ se obtienen de forma algorítmica, ver [6].

Si además de **(H1)**$_{df}$, $\mathcal{S}$ verifica las condiciones:

$$\textbf{(H2)}_{df}: \quad \delta =< \vec{\lambda}_{\partial\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0 \quad \text{y} \quad \textbf{(H3)}_{df}: \quad \epsilon =< \vec{\xi}_1, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0,$$

donde $\vec{\lambda}_{\partial\mu} = (\frac{\partial \Re\hat{\lambda}}{\partial\mu}, \frac{\partial \Im\hat{\lambda}}{\partial\mu})|_{(\omega_0,\mu_0)}$, $\vec{\lambda}^{\perp}_{\partial\omega} = (\frac{\partial \Im\hat{\lambda}}{\partial\omega}, -\frac{\partial \Re\hat{\lambda}}{\partial\omega})|_{(\omega_0,\mu_0)}$, $\vec{\xi}_1 = (\Re\xi_1, \Im\xi_1)|_{(\omega_0,\mu_0)}$, donde $\Re$ y $\Im$ notan la parte real e imaginaria y $< \cdot,\cdot >$ es el producto escalar real; a partir de las ecuaciones reales $\theta\,\Re\Phi(z,\omega,\mu) = 0$ y $\theta\,\Im\Phi(z,\omega,\mu) = 0$ para $\theta \neq 0$ suficientemente pequeño se obtiene la forma normal clásica de la ecuación de bifurcación

$$\delta(\mu - \mu_0) + \epsilon\,\theta^2 = 0,$$

la cual relaciona el parámetro de bifurcación con la amplitud de la órbita y donde el signo de ϵ/δ determina la estabilidad de la órbita bifurcada. Para obtener esta ecuación sólo es necesario la ecuación (1) para $q = 1$; si **(H2)**$_{df}$ o **(H3)**$_{df}$ no se verifican en algunos casos es necesario desarrollar (1) para valores mayores de q.

2. RESULTADOS

En esta sección se hace uso de la teoría de singularidades [1, 2] para determinar condiciones de definición y no degeneración sobre los coeficientes de la ecuación de bifurcación (1) cuando no se verifica alguna de las hipótesis **(H2)**$_{df}$ o **(H3)**$_{df}$. Se logran caracterizar las formas normales de ciertos tipos de bifurcaciones dinámicas en el dominio frecuencia. Esto permitirá ampliar la información obtenida en los comportamientos locales.

A partir de aquí sólo se supone que el sistema satisface **(H1)**$_{df}$, se agrega la notación $\vec{\lambda} = (\Re\hat{\lambda}, \Im\hat{\lambda})|_{(\omega_0,\mu_0)}$, $\vec{\xi}_k = (\Re\xi_k, \Im\xi_k)|_{(\omega_0,\mu_0)}$, $\vec{\lambda}_{\partial^k\mu} = (\frac{\partial^k \Re\hat{\lambda}}{\partial\mu^k}, \frac{\partial^k \Im\hat{\lambda}}{\partial\mu^k})|_{(\omega_0,\mu_0)}$ y $\vec{\lambda}^{\perp}_{\partial^k\mu} = (\frac{\partial^k \Im\hat{\lambda}}{\partial\mu^k}, -\frac{\partial^k \Re\hat{\lambda}}{\partial\mu^k})|_{(\omega_0,\mu_0)}$ para $k = 1, \cdots$.

Aplicando diferenciación implícita a las ecuaciones $\theta\,\Re\Phi(z,\omega,\mu) = 0$ y $\theta\,\Im\Phi(z,\omega,\mu) = 0$ alrededor de $(0,\omega_0,\mu_0)$, se obtienen bajo ciertas condiciones, $g(\theta,\mu) = 0$ y $h(\theta,\omega) = 0$. En particular la ecuación escalar que relaciona el parámetro de bifurcación con la amplitud de la órbita periódica, puede escribirse como $g(\theta,\mu) = \theta\, r(\theta^2,\mu) = 0$, ya que g es impar en θ ($\mathbb{Z}_2$-simétrica). Las soluciones no nulas de esta ecuación determinan la existencia de soluciones periódicas no triviales del sistema original.

Dentro del contexto de la teoría de singularidades se dice que la ecuación $g(\theta,\mu) = 0$ es $\mathbb{Z}_2$-equivalente a otra ecuación $g_1(\theta,\mu) = 0$, si existe un difeomorfismo local, que respeta la simetría, que mapea una ecuación en otra. Ambos conjuntos solución tendrán el mismo comportamiento cualitativo.

Lema 1 *En un entorno de (ω_0,μ_0) y $z = 0$, la ecuación $\theta\, r(z,\mu) = 0$ es $\mathbb{Z}_2$-equivalente a la forma normal:*

$a)$
$$\theta\,(\delta\,(\mu - \mu_0) + \epsilon\, z^q) = 0, \tag{2}$$

si se verifican las condiciones:

i) si $q = 1$, de no degeneración, $\epsilon =< \vec{\xi}_1, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$ y $\delta =< \vec{\lambda}_{\partial\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$, (Hopf clásico)

ii$_1$) de definición, $< \vec{\xi}_k, \vec{\lambda}^{\perp}_{\partial\omega} >= 0$ y $< \vec{\xi}_k, \vec{\lambda}^{\perp}_{\partial\mu} >= 0$ para $k = 1, \cdots, q-1$; y de no degeneración $\epsilon =< \vec{\xi}_q, \vec{\lambda}^{\perp}_{\partial\omega} > /q! \neq 0$ y $\delta =< \vec{\lambda}_{\partial\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$, ó

ii$_2$) de definición, $< \vec{\xi}_k, \vec{\lambda}^{\perp}_{\partial\omega} >= 0$ para $k = 1, \cdots, q-1$; y de no degeneración $\dfrac{\partial^q r}{\partial z^q}(z,\mu)|_{(0,\mu_0)} \neq 0$ y $\delta =< \vec{\lambda}_{\partial\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$.

$b)$
$$(\delta\,(\mu - \mu_0)^m + \epsilon\, z)\,\theta = 0, \tag{3}$$

si se verifican las condiciones

i) de definición, $< \vec{\lambda}_{\partial^k\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >= 0$ y $< \vec{\xi}_1, \vec{\lambda}^{\perp}_{\partial^k\mu} >= 0$ para $k = 1, \cdots, m-1$; y de no degeneración $\epsilon =< \vec{\xi}_1, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$ y $\delta =< \vec{\lambda}_{\partial^m\mu}, \vec{\lambda}^{\perp}_{\partial\omega} > /m! \neq 0$, ó

ii) de definición, $< \vec{\lambda}_{\partial^k\mu}, \vec{\lambda}^{\perp}_{\partial\omega} >= 0$ para $k = 1, \cdots, m-1$; y de no degeneración $\epsilon =< \vec{\xi}_1, \vec{\lambda}^{\perp}_{\partial\omega} >\neq 0$ y $\dfrac{\partial^m r}{\partial\mu^m}(z,\mu)|_{(0,\mu_0)} \neq 0$.

3. APLICACIONES

Se considera $\mathcal{S}$ un sistema dinámico con un parámetro de bifurcación μ y $\mathcal{P} \in \mathbb{R}^l$ parámetros auxiliares que aparecen en $\hat{\lambda}(\omega, \mu, \mathcal{P})$ y $\xi_k(\omega, \mu, \mathcal{P})$ de la ecuación (1).

La codimensión de la forma normal de tipo (2) $\delta(\mathcal{P}_0)\,(\mu - \mu_0(\mathcal{P}_0)) + \epsilon(\mathcal{P}_0)\,\theta^{2q} = 0$ es $q-1$, alrededor de $\mu = \mu_0(\mathcal{P}_0)$ y $z = 0$, donde $\mathcal{P}_0$ representa a valores específicos de $\mathcal{P}$ que verifican las condiciones de definición Lema 1 a). El desarrollo universal, es decir la expresión que contiene todas las posibles perturbaciones, que se obtiene de (1) tiene la forma:

$$\delta(\mathcal{P})\,(\mu - \mu_0(\mathcal{P})) + \mu_2(\mathcal{P})\,\theta^2 + \mu_4(\mathcal{P})\,\theta^4 + \cdots + \epsilon(\mathcal{P})\,\theta^{2q} = 0, \tag{4}$$

donde se verifica que $\mu_2(\mathcal{P}_0) = 0, \cdots, \mu_{2q-1}(\mathcal{P}_0) = 0$.

La codimensión de la forma normal de tipo (3) $\epsilon(\mathcal{P}_0)z + \delta(\mathcal{P}_0)(\mu - \mu_0(\mathcal{P}_0))^m = 0$ es $m-1$, alrededor de $\mu = \mu_0(\mathcal{P})$ y $z = 0$, donde $\mathcal{P}_0$ representa a valores específicos de $\mathcal{P}$ que verifican las condiciones de definición Lema 1 b). El desarrollo universal que se obtiene de (1) tiene la forma

$$\epsilon(\mathcal{P})z + z_0(\mathcal{P}) + z_1(\mathcal{P})\,(\mu - \mu_0(\mathcal{P})) + \cdots + z_{m-2}(\mathcal{P})\,(\mu - \mu_0(\mathcal{P}))^{m-2} + \delta(\mathcal{P})(\mu - \mu_0(\mathcal{P}))^m = 0. \tag{5}$$

donde se verifica que $z_0(\mathcal{P}_0) = 0, \cdots, z_{m-2}(\mathcal{P}_0) = 0$.

Las variedades de transición son subconjuntos en el espacio de parámetros del desarrollo universal definidos por relaciones entre sus coeficientes ($\mu_k(\mathcal{P})$ en (4) y $z_k(\mathcal{P})$ en (5)). Estas son de mucha importancia, ya que separan regiones donde los diagramas que se obtienen perturbando la forma normal singular son equivalentes. La elección de los parámetros $\mathcal{P}$ dará diagramas que pertenecerán a alguna de las distintas regiones lo que permitirá controlar las posibles dinámicas del problema.

3.1. CIRCUITO CON DIODO TÚNEL

Se considera el circuito con un diodo túnel de la Fig. 1, que se modela como un sistema planar y donde la función característica que representa la corriente versus el voltaje en dicho elemento no lineal se muestra en la Fig. 2. Este es uno de los circuitos osciladores más simples y estudiados en la literatura de la ingeniería electrónica, pero sigue estando vigente (ver [3]). Sea $x_1 = i_L$, $x_2 = v_C$, $\mu = v_B$, $\gamma_1 = 1/L$ (L inductancia) y $\gamma_2 = 1/C$ (C capacidad).

$$\begin{cases} \dot{x}_1 = \gamma_1 x_2, \\ \dot{x}_2 = \gamma_2(-x_1 - x_2) - f(x_2, \mu); \end{cases}$$
$$f(x_2, \mu) = -\gamma_2(h(\mu - x_2) + x_2)$$
$$h(y) = \kappa - (\kappa(y-\rho)^2(2y+\rho))/\rho^3.$$

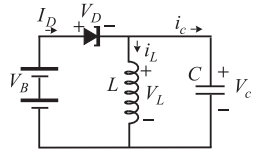


Fig. 1. Circuito con diodo túnel.

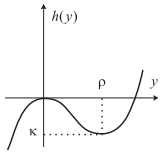


Fig. 2. Función $h(y)$.

Se considera la siguiente representación en frecuencia

$$A(\mu) = \begin{bmatrix} 0 & \gamma_1 \\ -\gamma_2 & -\gamma_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & -1 \end{bmatrix}, \quad G(i\omega, \mu) = -\frac{\gamma_2(6\kappa\mu(\mu-\rho) + \rho^3)\omega}{\rho^3(-i\gamma_1\gamma_2 + (\gamma_2 + i\omega)\omega)},$$

μ es el parámetro de bifurcación, ω la frecuencia y $\mathcal{P} = (\gamma_1, \gamma_2, \rho, \kappa)$ parámetros auxiliares. $\hat{y} = 0$ el equilibrio ($G(0)\,f(\hat{y}, \mu) + \hat{y} = 0$) y $J = \gamma_2\left(1 + \frac{2\kappa(\mu-\rho)^2}{\rho^3} + \frac{2\kappa(\mu-\rho)(2\mu+\rho)}{\rho^3}\right)$.

La parte real e imaginaria de la función característica $\hat{\lambda}(\omega, \mu, \mathcal{P})$ de $G(i\omega, \mu)J$,

$$\Re\hat{\lambda}(\omega, \mu, \mathcal{P}) = -\frac{\gamma_2^2\left(6\kappa\mu(\mu-\rho)+\rho^3\right)\omega^2}{\rho^3\left(\gamma_1^2\gamma_2^2 - 2\gamma_1\gamma_2\omega^2 + \gamma_2^2\omega^2 + \omega^4\right)}, \quad \Im\hat{\lambda}(\omega, \mu, \mathcal{P}) = -\frac{\gamma_2\left(6\kappa\mu(\mu-\rho)+\rho^3\right)\omega\left(\gamma_1\gamma_2 - \omega^2\right)}{\rho^3\left(\gamma_1^2\gamma_2^2 - 2\gamma_1\gamma_2\omega^2 + \gamma_2^2\omega^2 + \omega^4\right)}.$$

La hipótesis **(H1)**$_{df}$, se verifica en $\mu = 0$, $\omega = \sqrt{\gamma_1\gamma_2}$, y en $\mu = \rho$, $\omega = \sqrt{\gamma_1\gamma_2}$.

Se tiene la siguiente ecuación de bifurcación en el dominio frecuencia

$$\begin{cases} \Re\hat{\lambda}(\omega, \mu, \mathcal{P}) + 1 + \theta^2\,\Re\xi_1(\omega, \mu, \mathcal{P}) & = 0 \\ \Im\hat{\lambda}(\omega, \mu, \mathcal{P}) + \theta^2\,\Im\xi_1(\omega, \mu, \mathcal{P}) & = 0, \end{cases} \tag{6}$$

donde $\xi_1(\omega, \mu, \mathcal{P}) = \dfrac{6\kappa\omega\left(-i\gamma_1\gamma_2\rho^3 + 2\omega\left(3\kappa\left(-2(-2+\gamma_2)\mu^2 + 2(-2+\gamma_2)\mu\rho + \rho^2\right) + 2i\rho^3\omega\right)\right)}{\rho^3(\gamma_1\gamma_2 + i(\gamma_2 + i\omega)\omega)\left(\gamma_1\gamma_2\rho^3 - 4\omega\left(3i\gamma_2\kappa\mu(\mu-\rho) + \rho^3\omega\right)\right)}$.

En $z=0$, $\mu_0=0$ y $\omega_0=\sqrt{\gamma_1\gamma_2}$ se verifican las condiciones de no-degeneración $<\vec{\xi}_1,\vec{\lambda}^{\perp}_{\partial\omega}>=-\frac{12\kappa}{\gamma_2^2\rho^3}$ y $<\vec{\lambda}_{\partial\mu},\vec{\lambda}^{\perp}_{\partial\omega}>=\frac{12\kappa}{\gamma_2\rho^2}$, luego usando el Lema 1 a) i) se obtiene la forma normal $\mu-\frac{1}{\gamma_2\rho}\theta^2=0$.

En $z=0$, $\mu_0=\rho$ y $\omega_0=\sqrt{\gamma_1\gamma_2}$, se verifican las condiciones de no-degeneración $<\vec{\xi}_1,\vec{\lambda}^{\perp}_{\partial\omega}>=-\frac{12\kappa}{\gamma_2^2\rho^3}$ y $<\vec{\lambda}_{\partial\mu},\vec{\lambda}^{\perp}_{\partial\omega}>=-\frac{12\kappa}{\gamma_2\rho^2}$, luego usando el Lema 1 a) i) se obtiene la forma normal $\mu+\frac{1}{\gamma_2\rho}\theta^2=0$.

Alrededor de $\mu=0$ y $\mu=\rho$ se verifican **(H1)**$_{df}$, **(H2)**$_{df}$ y **(H3)**$_{df}$, luego las ecuaciones de bifurcación obtenidas son de codimensión cero y capturan localmente las caracteristicas principales del diagrama, que son válidos en entornos disjuntos de μ, (ver en la Fig. 4 las curvas (a) y (b)).

Para determinar si las dos ramas se unen, se realiza un desarrollo de la ecuación de bifurcación alrededor del punto $\mu=\rho/2$ donde $<\vec{\lambda}_{\partial\mu}(\omega,\mu,\mathcal{P}),\vec{\lambda}^{\perp}_{\partial\omega}(\omega,\mu,\mathcal{P})>=0$.

En $z=\frac{\gamma_2\rho^2}{4}$, $\mu_0=\frac{\rho}{2}$ y $\omega_0=\sqrt{\gamma_1\gamma_2}$ se verifican las condiciones de definición $<\vec{\lambda}_{\partial\mu},\vec{\lambda}^{\perp}_{\partial\omega}>=0$ y $<\vec{\lambda}_{\partial\mu},\vec{\xi}_1^{\perp}>=0$ y las de no degeneración $\epsilon=<\vec{\xi}_1,\vec{\lambda}^{\perp}_{\partial\omega}>=\frac{6\kappa(3\kappa-2\rho)}{\gamma_2^2\rho^4}\neq 0$ y $\delta=<\vec{\lambda}_{\partial^2\mu},\vec{\lambda}^{\perp}_{\partial\omega}>/2=\frac{6\kappa(3\kappa-2\rho)}{\gamma_2\rho^4}\neq 0$, luego usando el Lema 1 b) i) se obtiene la forma normal singular $\delta\,(\mu-\rho/2)^2+\epsilon z=0$, de codimensión uno. De (6) se obtiene el desarrollo universal $\delta\,(\mu-\rho/2)^2+\epsilon(z-\gamma_2\rho^2/4)=0$ de la forma normal. Luego obtenemos la expresión del coeficiente del desarrollo universal en función de los parámetros originales del problema $z_0(\mathcal{P})=-\frac{6\kappa(3\kappa-2\rho)}{4\gamma_2\rho^2}$.

En la Fig. 3 se muestran los distintos diagramas de bifurcación para $\epsilon>0$ y $\delta>0$ separados por la variedad de transición $z_0(\mathcal{P})=0$.

En la Fig. 4 se muestran los diagramas de bifurcación locales para $\gamma_1=10$, $\gamma_2=10^6$, $\rho=1$, $\kappa=-1$. Se observa que la validez de la información local proporcionada por los primeros diagramas (ver en la Fig. 4 las curvas (a) y (b)), se completa con la información dada en el diagrama (c) de la Fig. 4. A partir de $\mu=0$ nace una órbita periódica cuya amplitud crece, hasta $\mu=1/2$ y a partir de aquí comienza a decrecer hasta desaparecer en $\mu=1$, características que coinciden con los resultados experimentales.

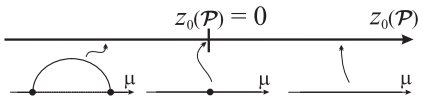


Fig. 3 Diagramas para $\epsilon>0$ y $\delta>0$.

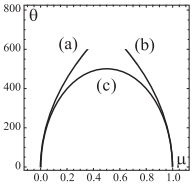


Fig. 4 (a) $\mu-10^{-6}\theta^2=0$, (b) $\mu+10^{-6}\theta^2=0$; (c) $-\frac{1}{4}+10^{-6}\,\theta^2+\left(\mu-\frac{1}{2}\right)^2=0$.

4. Conclusión

Utilizando la ecuación de bifurcación de alto orden se han podido determinar los diagramas de bifurcaciones locales cuando fallan algunos postulados del clásico teorema de bifurcación de Hopf. Se han verificado los mismos con resultados analíticos y simulaciones en un circuito electrónico con diodo túnel.

5. Agradecimientos

Se agradece la financiación de PICT-2006-00828, PIP 112-200801-01112, PGI 24/K041 y 24/L075.

Referencias

[1] M. Golubitsky and W. F. Langford, *Classification and unfoldings of degenerate Hopf bifurcation*, Journal of Differential Equations, 41 (1981), pp. 375-415.

[2] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Volume I Springer-Verlag, Berlin, New-York, 1985.

[3] M. Heinrich, and T. Dahms, V. Flunkert, S. Teitsworth and E. Schöll, *Symmetry-breaking transitions in networks of nonlinear circuits elements*, New Journal Of Physics, 12 (2010), pp. 1-26.

[4] A. Mees, *Dynamics of Feedback Systems*, Wiley, New-York, 1981.

[5] J. Moiola and G. Chen, *Hopf Bifurcation Analysis: A Frequency Domain Approach*, Nonlinear Science, World Scientific Co., Singapore, 1996

[6] A. Torresi, G. Calandrini and J. Moiola , *Oscilaciones múltiples y su control en el sistema de Bautin* , XXII Congreso Argentino de Control Automtico, 31 de agosto al 1 de septiembre 2010, Buenos Aires, 8 pags.

HOPF BIFURCATION IN AN INTERNET CONGESTION CONTROL MODEL: A FREQUENCY-DOMAIN APPROACH

Franco S. Gentile and Jorge L. Moiola

Instituto de Investigaciones en Ingeniería Eléctrica IIIE (UNS-CONICET), Departamento de Ingeniería Eléctrica y de Computadoras, Universidad Nacional del Sur, Av. Alem 1253 (B8000CPB), Bahía Blanca, Argentina, franco.gentile@uns.edu.ar, jmoiola@criba.edu.ar

Abstract: In this article, a class of internet congestion control is studied. The effect of delays in the transmission lines is considered, leading to a nonlinear delay-differential equation model. By using frequency-domain techniques, we study the occurrence of Hopf bifurcations, which causes the instability of the system and the appearance of smooth oscillations.

Keywords: *Time-delayed systems, internet congestion control, Hopf bifurcation, frequency-domain.*

1 INTRODUCTION

Internet congestion is actually a serious and challenging problem, which arises when a resource within the network becomes overloaded, causing an overflow on the communication channels. Then, some pieces (packets) of information are lost, and even more, the system may collapse. This trouble has driven to the active research of congestion control algorithms. Transmission Control Protocol (TCP) and Active Queue Management (AQM) are among the mainstays of these control strategies. Roughly speaking, TCP consists of the setting of the transmission rate through a window flow-mechanism depending on the state of the network. On the other hand, AQM consists of controlling the congestion level at each router through different algorithms.

Essentially, what an internet congestion control algorithm needs is to be stable, *i.e.*, the flow of information through its links should tend towards a stationary value, preferably close to the link capacity. However, under the variation of network parameters, some schemes exhibit poor performance, and even more, may become unstable. Concerning this topic, the stability analysis generally involves the study of a trascendental equation, because the control schemes are modeled by delay-differential equations (DDEs). Many authors dedicated their efforts to gain insight about the dynamics of these kinds of models, for example [1], [5], [6] to mention a few. The utilization of a DDE model comes from the necessity to considering the inherent communication delay, which is comprised of propagation and queuing delays.

Through this paper, we shall consider a TCP/AQM control model presented in [1], where a bifurcation analysis was made via the normal form theory. In [6], the same model was studied with the method of multiple scales. In both papers, the authors characterized the appearance of limit cycles in the system caused by Hopf bifurcations, a mechanism which provokes the onset of smooth oscillations via the variation of a distinguished parameter. Particularly, we study the variation of the time delay.

In this article, we use the frequency-domain (FD) approach [4] to study the dynamical behavior of the mentioned model. This method is based on the Graphical Hopf Bifurcation Theorem (GHBT) [3], and provides an alternative tool for the analysis of stability and the appearance of oscillations via the Hopf mechanism. The main advantage of the utilization of this tool is the avoidance of dealing with a trascendental equation. Instead, we study a curve in the complex plane (called Nyquist plot) determined by a characteristic function [2].

2 THE FREQUENCY-DOMAIN APPROACH

Let us consider a system of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bg(y(t), y(t-\tau); \mu), \\ y(t) = -Cx(t), \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^{2n} \times \mathbb{R} \to \mathbb{R}^n$ is a smooth nonlinear function, $\mu \in \mathbb{R}$ is the main bifurcation parameter and $\tau > 0$ is a time-delay. System (1) can be represented by the

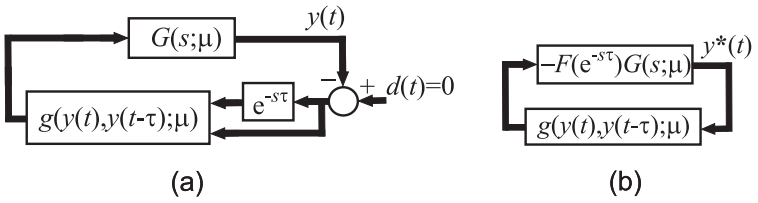


Figure 1: (a): Block representation of system (1).(b): Equivalent block representation with the delay absorbed on the linear part.

feedback scheme shown in Fig. 1(a), by taking $G(s) = C(sI_n - A)^{-1}B$, where s is the complex variable of the Laplace transform and I_n is the $n \times n$ identity matrix. The input $d(t)$ is assumed to be zero, because the system is autonomous. The delay can be absorbed in the linear block through a simple manipulation. Let $F(e^{-s\tau}) := \begin{pmatrix} I_m \\ I_m e^{-s\tau} \end{pmatrix}$, then we can represent the system as shown in Fig. 1(b), where the linear part becomes $G^*(s;\mu) := F(e^{-s\tau})G(s;\mu)$ and the output turns in $y^*(t) := (y(t)^T \quad y(t-\tau)^T)^T$. The equilibrium points $\widehat{y}$ can be found by solving $G^*(0)g(\widehat{y},\widehat{y};\mu) = -\widehat{y}$. Finally, we can linearize the nonlinear feedback at equilibrium by taking

$$J(\mu) = \left(\frac{\partial g(\cdot)}{\partial y(t)} \quad \frac{\partial g(\cdot)}{\partial y(t-\tau)} \right) |_{(y(t),y(t-\tau))=(\widehat{y},\widehat{y})}.$$

At this point, it is convenient to consider the following result given in [4].

Lemma 1 *If an eigenvalue of the corresponding Jacobian of the nonlinear system (1), in the time domain, assumes a purely imaginary value $i\omega_0$ at a particular value $\mu = \mu_0$, then the corresponding eigenvalue of the constant matrix $G^*(i\omega_0;\mu_0)J(\mu_0)$ in the FD must assume the value $-1 + i0$ at $\mu = \mu_0$.*

The eigenvalues of FGJ are given by the equation $|\lambda I_{2m} - G^*J| = 0$. Suppose that for $\mu = \mu_0$, there exists a simple root $\widehat{\lambda}(s;\mu)$ which takes the value $-1 + i0$ for a given $s = i\omega_0$. If we consider the geometrical locus of $\widehat{\lambda}(i\omega;\mu)$, it describes a curve (Nyquist locus) parameterized on the frequency ω. Particularly, for $\mu = \mu_0$ this curve crosses the point $-1 + i0$ at $\omega = \omega_0$. Now, let μ vary slightly from μ_0 and compute the auxiliary vector

$$\xi(\omega;\mu) = -\frac{u^T G^*(i\omega;\mu)p(\omega;\mu)}{u^T v}, \tag{2}$$

where u and v are the left and right eigenvectors of FGJ associated to $\widehat{\lambda}(s;\mu)$, and $p(i\omega) = D_2 v \otimes V_{02} + \frac{1}{2}D_2\bar{v} \otimes V_{22} + \frac{1}{8}D_2 v \otimes v \otimes \bar{v}$. Here, $\otimes$ denotes the tensor product operator, D_2 and D_3 are the second and third derivatives of $g(\cdot)$, and

$$V_{02}=-\tfrac{1}{4}(I_{2m}+G^*(0;\mu)J(\mu))^{-1}G^*(0;\mu)D_2 v\otimes\bar{v},\ V_{22}=-\tfrac{1}{4}(I_{2m}+G^*(i2\omega;\mu)J(\mu))^{-1}G^*(i2\omega;\mu)D_2 v\otimes v. \tag{3}$$

Let us consider the following result given in [4], which for convenience to the reader is stated as follows

Theorem 1 *(Graphical Hopf Bifurcation Theorem) Suppose that when ω varies, the vector $\xi(\omega;\mu) \neq 0$, and that the half line starting from $-1 + i0$ and pointing to the direction parallel to that of $\xi(\omega;\mu)$, first intersects the locus of $\widehat{\lambda}(i\omega;\mu)$ at the point $\widehat{P} = \widehat{\lambda}(i\widehat{\omega};\widehat{\mu}) = -1+\xi(\widehat{\omega};\widehat{\mu})\theta^2$, where constant $\theta = \theta(\widehat{\omega}) \geq 0$. Suppose, furthermore, that the above intersection is transversal,* i.e., *$\widehat{\lambda}(i\widehat{\omega};\widehat{\mu})$ and $\xi(\widehat{\omega};\widehat{\mu})$ are not parallel. Then:*

1. *The nonlinear system (1) has a limit cycle which is unique in a ball of radius $\mathcal{O}(1)$ centered at $\widehat{y}$.*

2. *If the total number of anticlockwise encirclements of the point $\widehat{P} + \delta\xi(\widehat{\omega};\widehat{\mu})$, for a small enough δ, is equal to the number of poles of $\widehat{\lambda}$ with positive real parts, then the limit cycle is stable.*

3 ANALYSIS OF THE TCP/AQM NETWORK

In this paper, we investigate the model

$$\begin{cases} \dot{w}(t) = \frac{1}{\tau} - \frac{1}{2\tau} w(t)^2 K q(t-\tau), \\ \dot{q}(t) = \frac{N}{\tau} w(t) - Q, \end{cases} \tag{4}$$

where $w(t)$ is the average of TCP windows size, $q(t)$ is the average queue length, τ is the round-trip time which consists of the propagation and queuing delay, Q is the queue capacity, N is the number of TCP sessions and K is a positive constant. Considering that τ is fixed is equivalent to take into account that the propagation delay is much greater than the queuing delay. The unique equilibrium point of (4) is $(\widehat{w}, \widehat{q}) = \left(\tau Q/N, 2N^2/KQ^2\tau^2\right)$. Then, defining $x_1(t) := w(t) - \widehat{w}$, $x_2(t) := q(t) - \widehat{q}$, system (4) can be written as

$$\begin{cases} \dot{x}_1(t) = -\alpha_1 x_1(t) - \alpha_2 x_2(t-\tau) - \alpha_3 x_1(t)^2 - \alpha_4 x_1(t) x_2(t-\tau) - \alpha_5 x_1(t)^2 x_2(t-\tau), \\ \dot{x}_2(t) = \alpha_6 x_1(t), \end{cases} \tag{5}$$

where for simplicity we call $\alpha_1 := \frac{2N}{Q\tau^2}$, $\alpha_2 := \frac{KQ^2\tau}{2N^2}$, $\alpha_3 := \frac{N^2}{Q^2\tau^3}$, $\alpha_4 := \frac{KQ}{N}$, $\alpha_5 := \frac{K}{2\tau}$, and $\alpha_6 := \frac{N}{\tau}$. Now, the equilibrium is $(\widehat{x}_1, \widehat{x}_2) = (0, 0)$. System above can be represented in the form (1) by choosing

$$A = \begin{pmatrix} -\alpha_1 & 0 \\ \alpha_6 & -1 \end{pmatrix},\ B = C = I_2,\ g = \begin{pmatrix} \alpha_2 y_2(t-\tau) - \alpha_3 y_1(t)^2 - \alpha_4 y_1(t) y_2(t-\tau) + \alpha_5 y_1(t)^2 y_2(t-\tau) \\ -y_2(t) \end{pmatrix}.$$

Then, we have matrices

$$G^*(s) = \begin{pmatrix} \frac{1}{s+\alpha_1} & 0 \\ \frac{\alpha_6}{(s+1)(s+\alpha_1)} & \frac{1}{s+1} \\ \frac{e^{-s\tau}}{s+\alpha_1} & 0 \\ \frac{\alpha_6 e^{-s\tau}}{(s+1)(s+\alpha_1)} & \frac{e^{-s\tau}}{s+1} \end{pmatrix}, \quad J = \begin{pmatrix} 0 & 0 & 0 & \alpha_2 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

It can be seen that matrix G^*J has a unique nonzero eigenvalue (characteristic function) given by $\widehat{\lambda}(s) = (\alpha_2\alpha_6 e^{-s\tau} - s - \alpha_1)/(s+1)(s+\alpha_6)$. The Hopf bifurcation condition $\widehat{\lambda}(i\omega) = -1 + i0$ leads to

$$\alpha_2\alpha_6 \cos(\omega\tau) = \omega^2, \qquad \alpha_2\alpha_6 \sin(\omega\tau) = \alpha_1\omega, \tag{6}$$

from which we obtain $\omega^4 + \alpha_1^2\omega^2 - \alpha_2^2\alpha_6^2 = 0$, and the critical frequency results

$$\omega_0(\tau) = \frac{\alpha_1}{\sqrt{2}} \left(\sqrt{1 + 4\frac{\alpha_2^2\alpha_6^2}{\alpha_1^4}} - 1 \right)^{1/2} = \frac{\sqrt{2}N}{Q\tau^2} \left(\sqrt{1 + \frac{K^2Q^8\tau^8}{16N^6}} - 1 \right)^{1/2}. \tag{7}$$

In addition, from (6) we have $\tau = \frac{1}{\omega_0} \arctan(2N/\omega_0 Q\tau^2)$. Thus by using (7), we can find the critical value of delay. This can be achieved by solving the last equation graphically, intersecting $f_1(\tau) = \tau$ and $f_2(\tau) = \arctan(2N/\omega_0(\tau)Q\tau^2)$ as in [1]. The left and right eigenvectors of G^*J associated to $\widehat{\lambda}(\cdot)$ are obtained as $u = (0\ u_2\ 0\ 1)^T$ and $v = (1\ v_2\ \ e^{-s\tau}\ v_2 e^{-s\tau})^T$, where $u_2 = -(s+\alpha_1)/\alpha_2\alpha_6$ and $v_2 = (\alpha_2\alpha_6 e^{-s\tau} - s - \alpha_1)/\alpha_2(s+1)$. Vectors V_{02} and V_{22} are obtained from (3) as

$$V_{02} = -\frac{\gamma_1(\omega)}{4\alpha_2} \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}^T, \quad V_{22} = -\frac{\gamma_2(\omega)}{4\eta(i2\omega)} \begin{pmatrix} i2\omega & \alpha_6 & i2\omega e^{-i2\omega\tau} & \alpha_6 e^{-i2\omega\tau} \end{pmatrix}^T,$$

where $\gamma_1(\omega) := -2\alpha_3 - 2\alpha_4 \Re\left\{v_2 e^{-i\omega\tau}\right\}$, $\gamma_2(\omega) := -2\alpha_3 - 2\alpha_4 v_2 e^{-i\omega\tau}$ and $\eta(s) = s^2 + s\alpha_1 + \alpha_2\alpha_6 e^{-s\tau}$ ($\Re\{\cdot\}$ stands for the real part). Thus, we can obtain $p(i\omega) = [0 \quad p_1(\omega)]^T$, where

$$p_1(\omega) = \frac{\alpha_4}{4\alpha_2}\gamma_1 - \frac{\gamma_2}{8\eta(i2\omega)}(i2\omega\bar{\gamma_2} - \alpha_4\alpha_6 e^{-i2\omega\tau}) + \frac{\alpha_5}{4}(2v_2 e^{-i\omega\tau} + \bar{v_2} e^{i\omega\tau}),$$

where $\bar{(\cdot)}$ means complex conjugate. Finally, from (2) we have $\xi(\omega) = -\frac{\alpha_2\alpha_6}{(\alpha_1+i\omega)(\alpha_2\alpha_6 - (\alpha_1+i\omega)e^{i\omega\tau})} p_1(\omega)$. Figures 2 and 3 illustrate the numerical results for $K = 0.001$, $Q = 1000$ and $N = 50$. For these parameter

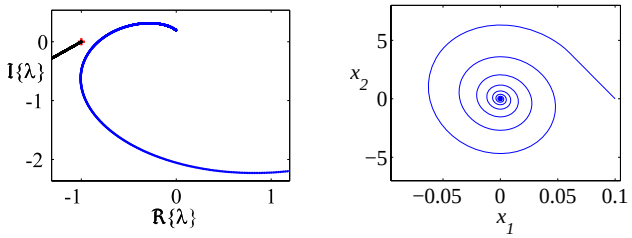


Figure 2: Nyquist diagram (left) and phase plot (right) for $K = 0.001$, $Q = 1000$, $N = 50$ and $\tau = 0.2$.

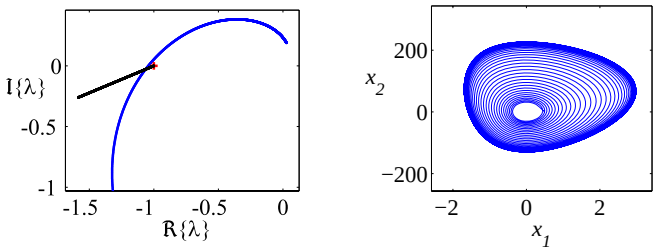


Figure 3: Nyquist diagram (left) and phase plot (right) for $K = 0.001$, $Q = 1000$, $N = 50$ and $\tau = 0.225$.

values, the critical delay value for which the Hopf bifurcation occurs is $\tau_0 \simeq 0.2203$. Figure 2 corresponds to $\tau = 0.2$. The Nyquist curve does not intersect vector $\xi(\cdot)$, thus the equilibrium is stable, as shown in the phase plot. Figure 3 shows the results obtained for $\tau = 0.225$. In this case, there exists an intersection between the Nyquist diagram and $\xi(\cdot)$. Thus, according to Theorem 1, the equilibrium is unstable, and a stable limit cycle has emerged. It shows that the control algorithm looses its stability for large enough delay values.

4 Conclusions

The GHBT provides a valuable tool for detecting the emergence of oscillations in nonlinear delayed systems. It allows to determine the stability of the equilibrium as well as emergent periodic solutions without dealing with a transcendental equation. In this work, we have illustrated the usefulness of this tool in a TCP/AQM model, were the system stability is lost via Hopf bifurcation. In future works, we will consider a varying delay, in order to deal with a more evolved model.

Acknowledgements

The financial support of the following grants is greatly appreciated: PICT 2006-00828 (ANPCyP), PIP 112-200801-01112 (CONICET) and PGI 24/K041 (UNS).

References

[1] D. Ding, J. Zhu, and X. Luo, *Hopf bifurcation analysis in a fluid flow model of internet congestion control algorithm*, Nonlinear Anal.: Real World Appl., 10 (2009), pp. 824–839.

[2] A. G. J. MacFarlane and I. Postlethwaite, *The generalized Nyquist stability criterion and multivariable root loci*, Int. J. Control, 25 (1977), pp. 81–127.

[3] A. I. Mees and L. O. Chua, *The Hopf bifurcation theorem and its applications to nonlinear oscillations in circuits and systems*, IEEE Trans. Circuits Syst. I, 26 (1979), pp. 235–254.

[4] J. L. Moiola and G. Chen, *Hopf Bifurcation Analysis - A Frequency Domain Approach*, vol. 21, World Scientific, Singapore, 1996.

[5] M. Xiao and J. Cao, *Delayed feedback-based bifurcation control in an internet congestion model*, J. Math. Anal. Appl., 332 (2007), pp. 1010–1027.

[6] Y. G. Zheng and Z. H. Wang, *Stability and Hopf bifurcation of a class of TCP/AQM networks*, Nonlinear Anal.: Real World Appl., 11 (2010), pp. 1552–1559.

CONVERGENCIA DEL MÉTODO EN FRECUENCIA AL APROXIMAR BIFURCACIONES DE DOBLE PERÍODO EN MAPAS CUADRÁTICOS

Guillermo Calandrini[♭,†] y María Belén D'Amico[♭]

[♭]*Instituto de Investigaciones en Ingeniería Eléctrica "Alfredo Desages", Universidad Nacional del Sur – CONICET, Avda Alem 1253, B8000CPB Bahía Blanca, Argentina*
[†]*Departamento de Matemática , Departamento de Ingeniería Eléctrica y de Computadoras , UNS*
calandri@criba.edu.ar, mbdamico@uns.edu.ar

Resumen: Se estudia en este trabajo la convergencia del método en frecuencia al aproximar las órbitas emergentes de una bifurcación de doble período en mapas cuadráticos. El análisis se realiza mediante el desarrollo en series de potencias de la solución exacta. La aplicación de los resultados obtenidos se ilustra a través de un mapa sencillo.

Palabras clave: *mapas, bifurcación, método en frecuencia, convergencia*
2000 AMS Subject Classification: 37C05, 37G15, 65P30

1. INTRODUCCIÓN

La bifurcación de doble período se presenta en mapas no lineales cuando la variación de un parámetro provoca el cambio de estabilidad de un punto fijo y la aparición de una órbita de período dos en torno al mismo. En una bifurcación *supercrítica* coexisten un punto fijo inestable con una órbita estable y, en una *subcrítica*, un punto fijo estable con una órbita inestable. En la literatura es frecuente encontrar que el análisis de este fenómeno se realice de dos formas diferentes. Una manera, que surge naturalmente, es hallar el segundo mapa iterado y ver los puntos de período dos como puntos fijos del nuevo mapa [1]. La otra opción consiste en utilizar los teoremas de la variedad centro y de las formas normales [2]. Cualquiera de estos métodos puede involucrar cálculos extensos y complicados.

Se presentó en [3] una técnica alternativa para el tratamiento de bifurcaciones de doble período formulada en el dominio frecuencia. Este método se basa en la representación del tipo entrada-salida del sistema y en la realización de un balance de armónicos para capturar el comportamiento dinámico en torno del punto fijo. Como un número significativo de mapas de más de una dimensión pueden llevarse a representaciones realimentadas de una única entrada y una única salida, se logra que dichos sistemas se analicen directamente en el dominio frecuencia sin tener que recurrir a su reducción a través del teorema de la variedad centro. No obstante, el alcance del método es local pues las expresiones de las órbitas son aproximaciones, con el grado de exactitud que se desee [4], de las soluciones de período dos reales.

El propósito de este trabajo es realizar un análisis de convergencia del método en frecuencia al aproximar las órbitas de período dos que emergen de sistemas con nolinealidades del tipo cuadráticas. Para ello, se comparan las soluciones exactas y aproximadas a través de descomposiciones en series de potencias. Las condiciones obtenidas pueden utilizarse también para tener una primera noción de la región de convergencia del método al emplearse en mapas con no linealidades de orden superior.

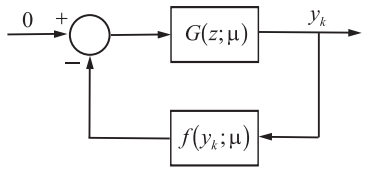
El trabajo se organiza de la siguiente manera. En la sección 2 se describen las fórmulas necesarias en el dominio frecuencia para la aproximación de órbitas de período dos. El estudio de convergencia del método para el caso de funciones cuadráticas se presenta en la sección 3. En la sección 4 se ilustra con un ejemplo sencillo la aplicación de los resultados obtenidos. Finalmente, las conclusiones se resumen en la sección 5.

2. MÉTODO EN FRECUENCIA

Sea la familia de mapas nolineales dependientes del parámetro $\mu \in \mathbf{R}$ dada por

$$x_{k+1} = Ax_k + Bh(y_k) \tag{1}$$
$$y_k = Cx_k \tag{2}$$

con $x_k \in \mathbf{R}^n$, $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times 1}$, $C \in \mathbf{R}^{1\times n}$ y $h(\cdot) : \mathbf{R} \to \mathbf{R}$. Dicha familia puede llevarse a una representación del tipo entrada-salida como la de la Fig. 1 y que consiste en un lazo cerrado formado por un



$$G(z;\mu) = C[zI - A]^{-1}B,$$
$$f(y_k;\mu) = -h(y_k).$$

Figura 1: Representación entrada-salida de un mapa nolineal.

bloque lineal y escalar $G(\cdot)$ y un bloque no lineal $f(\cdot) : \mathbf{R} \to \mathbf{R}$. De hecho, aplicando la transformada z y realizando una serie de operaciones algebraicas, resulta $G(z;\mu) = C[zI - A]^{-1}B$ y $f(y_k;\mu) = -h(y_k)$. Bajo este enfoque, los puntos fijos $\widehat{y}$ del sistema se obtienen resolviendo $G(1;\mu)f(\widehat{y};\mu) + \widehat{y} = 0$ y el comportamiento dinámico en torno de los mismos se analiza mediante la función de lazo abierto $\lambda\,(z;\mu) = G(z;\mu)J(\mu)$ donde $J(\cdot) = D_y f(\widehat{y};\mu)$. Así, la condición necesaria para la aparición de una bifurcación de doble período dada por el cruce de uno de los autovalores de (1) por el punto -1 del círculo unitario para $\mu = \mu_o$ es equivalente a que la curva que representa a $\lambda(z;\mu)$ en el plano complejo para $z = e^{i\omega}$ ($\omega \in [0,\pi]$) pase por el punto crítico $-1 + i0$ para $\omega = \pi$ y $\mu = \mu_o$ [3].

Una vez detectado el punto de bifurcación, se aplica el método del balance de armónicos para capturar la órbita emergente. Para ello, se propone la solución de período dos de la forma $y_k = \widehat{y} + Y_0 + Y_1 e^{i\pi k}$ donde Y_0 es una corrección del centro o valor medio e Y_1 representa la distancia desde el centro hasta aquellos de período dos. La respuesta de $f(\cdot)$ ante dicha entrada posee su misma característica, es decir, $f(y_k;\mu) = f(\widehat{y};\mu) + F_0 + F_1 e^{i\pi k}$ donde F_0 y F_1 dependen de la composición de $f(\cdot)$ y de Y_0 e Y_1. Teniendo en cuenta que $G(z;\mu)$ es lineal, las relaciones que deben cumplir los coeficientes de cada armónico de la oscilación de período dos al recorrer el lazo de realimentación están dada por

$$Y_0 = -G\,(1;\mu)\,F_0, \qquad Y_1 = -G(-1;\mu)F_1, \tag{3}$$

donde $G(1;\mu)$ y $G(-1;\mu)$ son las respuestas de $G(\cdot)$ ante señales de frecuencia $\omega = 0$ (corrección) y $\omega = \pi$ (período dos), respectivamente. El grado de dificultad en la resolución de las ecuaciones de balance así planteadas depende directamente de la complejidad de la función no lineal $f(\cdot)$. Se propuso en [3] una manera de obtener fórmulas explicitas que permiten calcular Y_0 e Y_1 con suficiente exactitud.

Asumiendo que $f(\cdot)$ es una función suave (C^r con $r \geq 4$), la misma puede expandirse en una serie de Taylor en torno del punto $\widehat{y}$ como

$$f(y_k;\mu) = f(\widehat{y};\mu) + D_y f(\widehat{y};\mu)(y_k - \widehat{y}) + \tfrac{1}{2!}D_y^2 f(\widehat{y};\mu)(y_k - \widehat{y})^2 + \tfrac{1}{3!}D_y^3 f(\widehat{y};\mu)(y_k - \widehat{y})^3 + O(|y_k - \widehat{y}|^4).$$

Así, los coeficientes F_0 y F_1 pueden expresarse como

$$\begin{aligned} F_0 &= J(\mu)Y_0 + \tfrac{1}{2}D_y^2 f(\widehat{y};\mu)(Y_0^2 + Y_1^2) + \tfrac{1}{6}D_y^3 f(\widehat{y};\mu)Y_0(Y_0^2 + 3Y_1^2) + O(|y_k - \widehat{y}|^4), \\ F_1 &= J(\mu)Y_1 + D_y^2 f(\widehat{y};\mu)Y_0 Y_1 + \tfrac{1}{6}D_y^3 f(\widehat{y};\mu)(3Y_0^2 Y_1 + Y_1^3) + O(|y_k - \widehat{y}|^4). \end{aligned}$$

Considerando además que $|Y_0| = O(|Y_1|^2)$, pues esta condición se verifica al inicio de la bifurcación, las ecuaciones de balance (3) se reducen a

$$\begin{aligned} Y_0 &= -G\,(1;\mu)\,[J(\mu)Y_0 + \tfrac{1}{2}D_y^2 f(\widehat{y};\mu)Y_1^2] + O(|Y_1|^4), \\ Y_1 &= -G\,(-1;\mu)\,[J(\mu)Y_1 + D_y^2 f(\widehat{y};\mu)Y_0 Y_1 + \tfrac{1}{6}D_y^3 f(\widehat{y};\mu)Y_1^3] + O(|Y_1|^4). \end{aligned}$$

Resulta sencillo verificar entonces que una expresión aproximada para calcular Y_0 es

$$\widetilde{Y}_0 = -\tfrac{1}{2}H(\mu)D_y^2 f(\widehat{y};\mu)\widetilde{Y}_1^2, \tag{4}$$

con $H(\cdot) = [1 + G(1;\mu)J(\mu)]^{-1}G(1;\mu)$, y que la expresión para el cálculo del coeficiente $\widetilde{Y}_1$ es

$$[1 + G(-1;\mu)J(\mu)]\widetilde{Y}_1 = -G(-1;\mu)p(\mu)\widetilde{Y}_1^3$$

con $p(\cdot) = D_y^3 f(\widehat{y};\mu)/6 - H(\mu)[D_y^2 f(\widehat{y};\mu)]^2/2$. Finalmente, suponiendo que $\widetilde{Y}_1 \neq 0$ y definiendo $\xi(\mu) = -G\,(-1;\mu)\,p(\mu)$, la solución existirá siempre que

$$\lambda\,(-1;\mu) = -1 + \xi(\mu)\widetilde{Y}_1^2. \tag{5}$$

3. SISTEMAS CON NO LINEALIDAD CUADRÁTICA

Se estudiará a continuación la convergencia de las soluciones de doble período planteadas en [3] para el conjunto de sistemas donde el bloque no lineal $f(\cdot)$ es un polinomio de segundo grado, es decir,

$$f(y_k;\mu) = f(\widehat{y};\mu) + J(\mu)(y_k - \widehat{y}) + K(\mu)(y_k - \widehat{y})^2,$$

donde $K(\mu) = D_y^2 f(\widehat{y};\mu)/2!$ y $D_y^n f(\widehat{y};\mu) = 0, \forall n \geq 3$. En este caso, (4)-(5) se reducen a

$$\begin{aligned} \widetilde{Y}_0 &= -H(1;\mu)K(\mu)\widetilde{Y}_1^2, \\ \lambda(-1;\mu)+1 &= -2\,G(-1;\mu)K(\mu)\widetilde{Y}_0 = 2\,G(-1;\mu)K(\mu)^2H(1;\mu)\widetilde{Y}_1^2 \end{aligned}$$

con $\xi(\mu) = 2G(-1;\mu)H(\mu)K(\mu)^2$. Considerando $H_{-1}(\mu) = G(-1;\mu)^{-1}[1 + G(-1;\mu)J(\mu)]$, se tiene

$$\widetilde{Y}_0 = \frac{-H_{-1}(\mu)}{2\,K(\mu)}, \qquad \widetilde{Y}_1^2 = \frac{H_{-1}(\mu)}{2\,K(\mu)^2H(\mu)}.$$

Por otro lado, dada la estructura de $f(\cdot)$ resulta posible resolver en forma exacta las ecuaciones del balance armónico (3). Esto es,

$$Y_0 = -G\,(1;\mu)\,[J(\mu)Y_0 + K(\mu)Y_0^2 + K(\mu)Y_1^2], \tag{6}$$

$$Y_1 = -G\,(-1;\mu)\,[J(\mu)Y_1 + 2\,K(\mu)Y_0Y_1]. \tag{7}$$

Suponiendo que $Y_1 \neq 0$, se obtiene

$$Y_0 = -\frac{1 + G\,(-1;\mu)\,J(\mu)}{2\,K(\mu)G\,(-1;\mu)} = \frac{-H_{-1}(\mu)}{2\,K(\mu)}, \qquad Y_1^2 = \frac{H_{-1}(\mu)\,(2 - H(\mu)H_{-1}(\mu))}{4\,K(\mu)^2H(\mu)}.$$

Las ecuaciones de balance exacto y aproximado se diferencian en el término de Y_0^2 que aparece en (6), pues el mismo fue despreciado anteriormente por ser $O(|Y_1|^4)$. Aún así, las expresiones para calcular Y_0 e $\widetilde{Y}_0$ son idénticas. Se realiza entonces un desarrollo en serie de potencias de Y_0 para poder determinar su influencia en los cálculos aproximados. Operando (6), se obtiene $Y_0 = \left(-H(\mu)K(\mu)Y_1^2\right)(1 + H(\mu)K(\mu)Y_0)^{-1}$. Reemplazando en (7) y considerando la serie geométrica $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$ para $|x| < 1$, resulta

$$\lambda(-1;\mu)+1 = 2\,G(-1;\mu)H(1;\mu)K(\mu)^2\left(1 + \sum_{n=1}^{\infty}(-H(\mu)K(\mu)Y_0)^n\right)Y_1^2.$$

La serie involucrada es convergente cuando $|H\,(\mu)\,K(\mu)Y_0| = \left|\frac{1}{2}H(\mu)H_{-1}\,(\mu)\right| < 1$. Se observa además que el primer término coincide con la expresión del método aproximado para el cálculo de $\xi(\mu)$. El resto de los términos actúan como correcciones del valor de $\xi(\mu)$.

Un análisis similar puede realizarse considerando una expansión en potencias de Y_1. Resolviendo Y_0 en función de Y_1 en (6) y eligiendo la solución que corresponde a la bifurcación de doble período (que debe verificar $Y_0 = 0$ para $Y_1 = 0$), se tiene $Y_0 = \left(-1 + \sqrt{1 - (2\,K(\mu)H(\mu)Y_1)^2}\right)(2\,K(\mu)H(\mu))^{-1}$.

En base al desarrollo en serie de $\sqrt{1-x} = 1 - \frac{1}{2}x - \sum_{n=2}^{\infty}\frac{(2n-3)!}{2^{2n-2}n!(n-2)!}x^n$ para $|x| < 1$, resulta

$$Y_0 = -K(\mu)H(\mu)Y_1^2 - \frac{1}{2\,K(\mu)H(\mu)}\sum_{n=2}^{\infty}\frac{(2n-3)!}{2^{2n-2}n!(n-2)!}(2\,K(\mu)H(\mu)Y_1)^{2n}$$

con $|2\,K(\mu)H(\mu)Y_1|^2 < 1$. Puede apreciarse que nuevamente el primer término corresponde al aproximado por el método. Finalmente, reemplazando en (7),

$$\lambda\,(-1;\mu)+1 = 2\,G\,(-1;\mu)\,K(\mu)^2H(\mu)Y_1^2 + \sum_{n=2}^{\infty}\frac{4\,(2n-3)!}{n!(n-2)!}G\,(-1;\mu)\,K(\mu)^{2n}H(\mu)^{2n-1}Y_1^{2n},$$

y que corresponde al desarrollo de alto orden $\lambda\,(-1;\mu)+1 = \sum_{n=1}^{\infty}\xi_nY_1^{2n}$ propuesto en [4]. Esta serie es convergente para $|2\,K(\mu)H(\mu)Y_1|^2 = |H(\mu)H_{-1}(\mu)\,(2 - H(\mu)H_{-1}(\mu))| < 1$.

$$Y_0 = -1 + \frac{1}{2}\sqrt{1-4\mu}, \qquad Y_1^2 = -\frac{3}{4} - \mu,$$
$$y_k = -\frac{1}{2} + e^{i\pi k}\sqrt{-\frac{3}{4} - \mu},$$

$$\widetilde{Y}_0 = -1 + \frac{1}{2}\sqrt{1-4\mu}, \qquad \widetilde{Y}_1^2 = \frac{1-4\mu}{2} - \sqrt{1-4\mu},$$
$$\widetilde{y}_k = -\frac{1}{2} + e^{i\pi k}\sqrt{\frac{1-4\mu}{2} - \sqrt{1-4\mu}}.$$

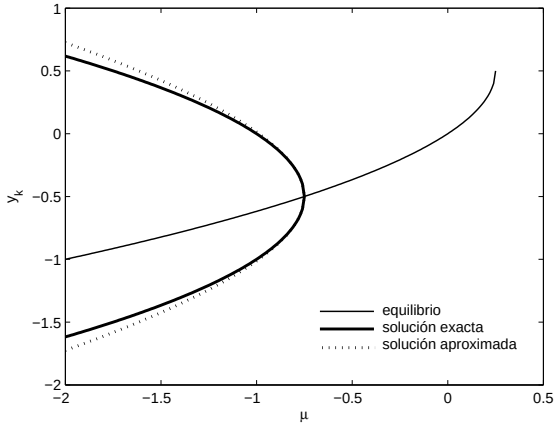


Figura 2: Expresiones y gráfico de las soluciones exacta y aproximada del doble período.

4. EJEMPLO

A continuación se aplican los resultados obtenidos al mapa $x_{k+1} = x_k^2 + \mu$. Para ello, se elige la realización $G(z;\mu) = 1/z$, $f(y_k;\mu) = -(y_k^2+\mu)$ con $y_k = x_k$. Luego, $J(\mu) = -2\widehat{y}$, $K(\mu) = -1$ y resolviendo la ecuación de equilibrio se observa que existen dos puntos fijos para $\mu < 1/4$, pero sólo $\widehat{y} = \frac{1}{2} - \frac{1}{2}\sqrt{1-4\mu}$ verifica la condición $\lambda(-1;\mu_0) = -1$ para $\mu_0 = -3/4$.

Siguiendo el procedimiento desarrollado anteriormente se obtienen las soluciones exacta y aproximada que se describen y grafican en la Fig. 2. En principio, se verifica que ambas existen si $\mu \leq -\frac{3}{4}$. Para validar la aproximación se analizan las condiciones de convergencia de las expansiones en serie. El desarrollo en potencias de Y_0 resulta convergente para $\left|\frac{1}{2} - \frac{1}{\sqrt{1-4\mu}}\right| < 1$, se deduce entonces que la aproximación se cumple para $\mu < 5/36$. La serie en potencias de Y_1 resulta convergente para $\left|1 + \frac{4}{4\mu-1}\right| < 1$ que se verifica siempre que $\mu < -1/4$. Se observa en ambos desarrollos que la aproximación resulta convergente en todo el rango de valores del parámetro $\mu \leq -3/4$ donde existen las soluciones de período dos.

5. CONCLUSIONES

Se estudió la convergencia del método en frecuencia al aproximar las órbitas emergentes de una bifurcación de doble período en mapas cuadráticos. El análisis se realizó mediante el desarrollo en series de potencias de la solución exacta. Para el caso en estudio, la condición de convergencia validó la aproximación. En otras situaciones podría resultar un limitante de la región de convergencia. Las condiciones obtenidas podrían emplearse además como una primera referencia sobre la región de convergencia del método al abordar mapas con no linealidades de orden mayor a dos.

AGRADECIMIENTOS

Los autores agradecen a la SGCyT de la UNS (PGI 24/K041), al CONICET (PIP 112-200801-01112) y a la ANPCyT (PICT 2006-00828) por el apoyo brindado para la realización de este trabajo.

REFERENCIAS

[1] R. L. DEVANEY, *A First Course in Chaotic Dynamical Systems*, Addison-Wesley Publishing Inc., Massachusetts, 1992.

[2] Y. A. KUZNETSOV, *Elements of Applied Bifurcation Theory*, Springer-Verlag, Nueva York, 3era edición, 2004.

[3] M. B. D´AMICO, J. L. MOIOLA Y E. E. PAOLINI, *A frequency domain method for analyzing period doubling bifurcations in discrete-time systems*, Circuits, Systems and Signal Processing, 23 (2004), pp. 516-535.

[4] M. B. D´AMICO, J. L. MOIOLA Y E. E. PAOLINI, Study of degenerate bifurcations in maps: A feedback system approach, *Int. J. of Bifurcation and Chaos*, 14 (2004), pp. 1625-1641.

INTERACCIONES ENTRE BIFURCACIONES DE CODIMENSIÓN 2 DE CICLOS LÍMITES

Gustavo Revel, Diego M. Alonso y Jorge L. Moiola

Instituto de Investigaciones en Ingeniería Eléctrica IIIE (UNS-CONICET), Departamento de Ingeniería Eléctrica y de Computadoras, Universidad Nacional del Sur, Av. Alem 1253 (B8000CPB), Bahía Blanca, Argentina
grevel@uns.edu.ar

Resumen: En este trabajo se estudia la interacción entre bifurcaciones de ciclos límites halladas en el análisis de la dinámica de un oscilador eléctrico. Realizando continuaciones numéricas, se describen interacciones entre bifurcaciones de doble período (flip), silla nodo de ciclos (fold) y Neimark-Sacker (NS). Se detecta una estructura particular que vincula curvas de bifurcaciones de codimensión 2 tales como fold-NS, fold-flip y resonancias fuertes 1:2, organizadas en torno a dos singularidades de codimensión 3.

Palabras clave: *bifurcaciones de ciclos límites, interacción de oscilaciones, resonancias*
2000 AMS Subject Classification: 37G15 - 34C15

1. INTRODUCCIÓN

Para comprender el comportamiento de un sistema dinámico sobre una región amplia del espacio de parámetros, resulta útil identificar los centros organizadores donde confluyen diferentes curvas de bifurcaciones. Estos centros pueden ser simplemente puntos en el espacio de los parámetros o estructuras más complejas como curvas cerradas o islas que pueden atribuirse a singularidades de mayor codimensión. Las interacciones entre bifurcaciones de ciclos límites en sistemas autónomos representados por ODEs pueden dar lugar a este tipo de estructuras. En este caso particular el estudio de sus bifurcaciones se puede realizar a partir del análisis de las singularidades de los puntos fijos del mapa de Poincaré asociado al ciclo. Las bifurcaciones locales más simples (codimensión 1) se dan cuando la linealización del mapa en el punto de interés presenta multiplicadores sobre el círculo unitario. En forma genérica existen tres casos de acuerdo al valor que toma el multiplicador [6]: $\mu = 1$, origina una bifurcación silla-nodo o *fold*, $\mu = -1$, conduce a una bifurcación de doble período o *flip*, y $\mu_{1,2} = e^{\pm i\theta}$ con $0 < \theta < \pi$ da lugar a una bifurcación de Neimark-Sacker (en adelante NS). Si se consideran variaciones de un segundo parámetro independiente, se pueden estudiar bifurcaciones de codimensión 2. Estos casos, de mayor complejidad que los anteriores, combinan curvas de bifurcaciones de codimensión 1. De esta manera, pueden ocurrir singularidades *fold-flip* ($\mu_1 = 1$ y $\mu_2 = -1$), *fold*-NS ($\mu_1 = 1$ y $\mu_{2,3} = e^{\pm i\theta}$), *flip*-NS ($\mu_1 = -1$ y $\mu_{2,3} = e^{\pm i\theta}$) y doble NS ($\mu_{1,2} = e^{\pm i\theta}$ y $\mu_{3,4} = e^{\pm i\delta}$). Cabe mencionar que cada caso puede darse sólo si el sistema ODE y el mapa de Poincaré asociado al ciclo tienen las dimensiones apropiadas. Además varias de ellas están vinculadas al nacimiento de dinámica compleja. El estudio analítico de estas singularidades es reciente, por ejemplo la interacción *fold-flip* se estudia en [8], la *fold*-NS en [1, 12] y la *flip*-NS y doble NS en [7]. Las bifurcaciones de codimensión 2 también incluyen los casos en los cuales se tiene un par de multiplicadores en $\mu_{1,2} = e^{\pm i2\pi/q}$ con $q = 1, 2, 3, 4$. Estas bifurcaciones se denominan resonancias fuertes $1 : q$ y han sido ampliamente estudiadas (véase por ejemplo [4, 5]). Completan las bifurcaciones de codimensión 2 de puntos fijos o ciclos, los casos degenerados del fold, flip y NS, que se denominan cusp, flip generalizada y Chenciner, respectivamente.

En este trabajo se describe la interacción entre curvas de bifurcaciones de codimensión 2 de ciclos límites, tales como fold-flip, fold-NS y resonancias 1:2, halladas en la investigación del comportamiento dinámico de un oscilador electrónico. Se detectan bifurcaciones de codimensión 3 que actúan como centros organizadores de la dinámica. El análisis se realiza en forma numérica utilizando paquetes para la continuación de soluciones y sus bifurcaciones [2, 3]. El trabajo se encuentra organizado de la siguiente manera. En la Sec. 2, se presenta el circuito del oscilador y se introducen algunas de las bifurcaciones encontradas. En la Sec. 3, se describe localmente una singularidad de codimensión 3. En la Sec. 4 se muestra una estructura de bifurcaciones distintiva en tres parámetros. Por último, en la Sec. 5 se exponen las conclusiones.

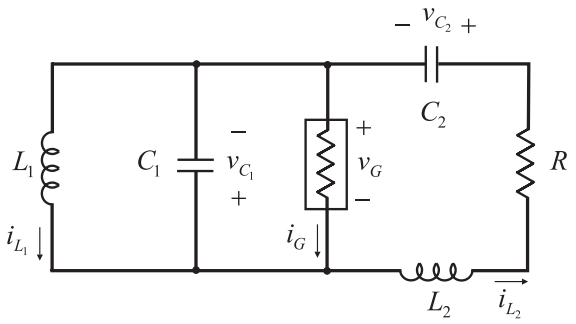


Figura 1: Circuito del oscilador eléctrico.

2. CIRCUITO DEL OSCILADOR ELÉCTRICO Y CARACTERÍSTICAS DINÁMICAS

El diagrama del oscilador utilizado en el estudio se muestra en la Fig. 1 y su modelo matemático resulta

$$\begin{aligned} \dot{x}_1 &= \eta_1 \left(\tfrac{1}{2}x_1 + x_2 - x_4 - \tfrac{3}{5}x_1^2 - x_1^3\right), \\ \dot{x}_2 &= -\eta_3 x_1, \\ \dot{x}_3 &= \left(1+\sqrt{2}\right) x_4, \\ \dot{x}_4 &= \left(2-\sqrt{2}\right)(x_1 - x_3 - \eta_2 x_4), \end{aligned} \tag{1}$$

donde $x_1 = v_{C_1}$, $x_2 = i_{L_1}$, $x_3 = v_{C_2}$, y $x_4 = i_{L_2}$ son los estados, η_1, η_2 y η_3 son los parámetros de bifurcación que se relacionan con los parámetros físicos a través de $\eta_1 = 1/C_1$, $\eta_2 = R$, $\eta_3 = 1/L_1$, $C_2 = 1/(1+\sqrt{2})$, $L_2 = 1/(2-\sqrt{2})$, e $i_G = -\frac{1}{2}v_G - \frac{3}{5}v_G^2 + v_G^3$ es la característica del elemento no lineal.

Este circuito, y algunas de sus versiones modificadas, se ha utilizado como ejemplo de estudio para el análisis de la bifurcación de Hopf doble o Hopf-Hopf. En [9] se han estudiado cuatro despliegues de su forma normal en el espacio de parámetros $\eta_1 - \eta_2$ para diferentes valores de η_3. Tres de ellos corresponden a los casos simples que contienen dos curvas de bifurcaciones de NS, y un caso complejo que además posee curvas que originan un toro 3D. Para $\eta_3 = 0.60355$ la singularidad Hopf-Hopf experimenta una resonancia 1:1, mientras que para $\eta_3 = 0.7740$ se tiene una resonancia 2:3. Esta última tiene como característica principal la generación de una estructura de bifurcaciones compuesta por dos resonancias fuertes 1:2, una isla o burbuja de doble período sobre una de las curvas de Neimark-Sacker y una resonancia 1:3 sobre la otra [10, 11].

3. DESCRIPCIÓN DE LAS INTERACCIONES ENTRE FOLD, FLIP Y NEIMARK-SACKER

Analizando la evolución de la isla de doble período originada en la resonancia 2:3, se ha observado que aproximadamente para $\eta_3 = 0.6648$ ésta interacciona con una curva de bifurcaciones de silla-nodo o fold de ciclos dando lugar a dos singularidades de codimensión 2 de tipo fold-flip [9]. Una parte de esta interacción se manifiesta en el escenario dinámico hallado para $\eta_3 = 0.6642$. El correspondiente diagrama de bifurcaciones en el plano de parámetros $\eta_1 - \eta_2$ se muestra en la Fig. 2a. Dado que los fenómenos ocurren en una región muy estrecha del espacio de parámetros, para lograr una mejor interpretación gráfica, se ha optado por incluir un diagrama esquemático donde se han deformado las escalas en forma conveniente. En el diagrama se observa la curva cerrada o isla (identificada como PD, en rojo) correspondiente a una bifurcación de doble período (flip). Sobre esta curva se detectan dos puntos de singularidades fold-flip indicados como FF_1 y FF_2, donde una curva de silla-nodo de ciclos (CF, en azul) se hace tangente a la isla (PD). Sobre la curva de PD ocurren además cuatro fallas del coeficiente de curvatura (GPD) y dos resonancias fuertes 1:2 denominadas $R^a_{1:2}$ y $R^b_{1:2}$ donde confluyen curvas de Neimark-Sacker (NS, en verde). Sobre la curva NS inferior se produce otra tangencia con la curva de fold CF dando lugar a una singularidad fold-NS (FNS). Cabe mencionar que el ciclo que experimenta estos fenómenos dinámicos nace a partir de la bifurcación de Hopf indicada como H en la Fig. 2.

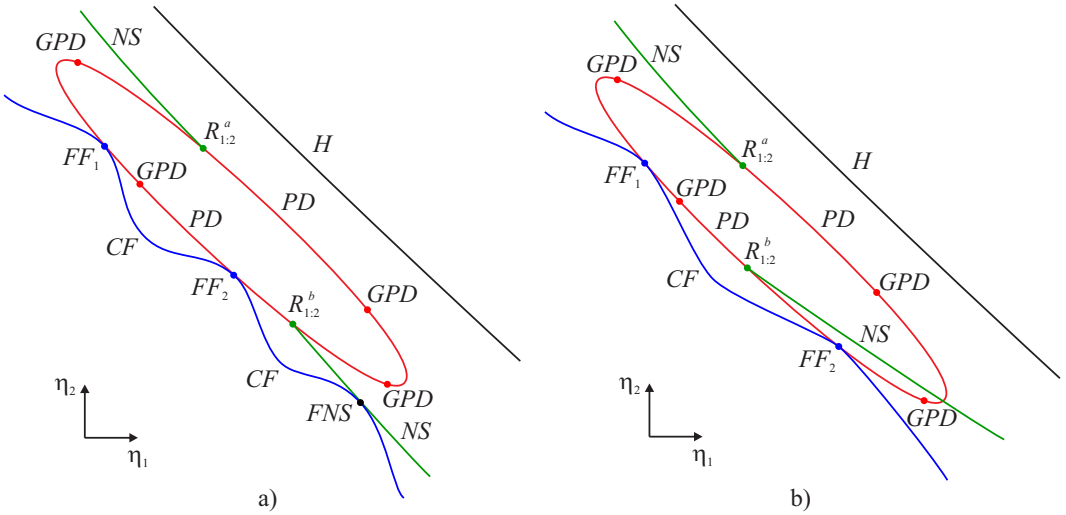


Figura 2: Diagrama de bifurcaciones esquemático en el plano de parámetros $\eta_1 - \eta_2$ para: (a) $\eta_3 = 0.6642$ y (b) $\eta_3 = 0.6640$.

De acuerdo al diagrama de bifurcaciones de la Fig. 2a, para $\eta_3 = 0.6642$ existen tres puntos de tangencia de la curva de fold CF, dos veces con la isla de PD y una con la curva NS inferior. Esta situación se ve modificada al disminuir η_3. Por ejemplo, para $\eta_3 = 0.6640$ la estructura de bifurcaciones se grafica en la Fig. 2b. Se observa un cambio cualitativo en la parte inferior del diagrama, ya que ha variado la ubicación relativa de los puntos R^b_{12} y FF_2. Además ha desaparecido el punto FNS, por lo que existen sólo los dos puntos de tangencia sobre la isla de PD. Esto se debe a una interacción entre las singularidades de codimensión 2, es decir entre FF_2 (fold-flip), R^b_{12} (resonancia 1:2) y FNS (fold-NS). Esta situación se da para $(\eta_1, \eta_2, \eta_3) = (2.172127, 1.730513, 0.664077)$ y corresponde a una singularidad de codimensión 3 que aún no ha sido reportada en la literatura especializada.

Para valores decrecientes del parámetro η_3 se tiene una situación similar entre el fold-flip FF_1 y la resonancia R^a_{12}. Para $(\eta_1, \eta_2, \eta_3) = (2.1761, 1.7456, 0.66138)$ el fold-flip FF_1 interactúa, en otra singularidad de codimensión 3, con la resonancia R^a_{12} y da lugar a una bifurcación fold-NS sobre la curva NS superior. Ambos puntos singulares de codimensión 3 se encuentran vinculados por medio de una estructura de bifurcaciones de codimensión 2 que se analiza a continuación.

4. Estudio de la dinámica variando tres parámetros

Para analizar la forma en que se conectan las singularidades descriptas anteriormente es necesario variar 3 parámetros simultáneamente. Los paquetes de continuación estándares, en sus implementaciones actuales, tienen limitaciones para continuar los puntos de codimensión 2. En particular, la continuación del fold-flip solamente está implementada en LocBif [3], mientras que las restantes singularidades no lo están. En la Fig. 3 se muestran las proyecciones del gráfico de bifurcaciones en el plano de parámetros $\eta_1 - \eta_2$ (Fig. 3a) y en el plano $\eta_3 - \eta_2$ (Fig. 3b). Se distingue una curva cerrada correspondiente a la bifurcación fold-flip (FF) realizada con LocBif. Sobre esta curva los multiplicadores críticos asociados al ciclo límite resultan $(1, -1)$. Además se han detectado dos puntos en los cuales se da la condición de multiplicadores $(1, -1, -1)$ que indican la intersección de la isla con dos curvas de resonancias fuertes 1:2 (R^a_{12} y R^b_{12}). Ambos puntos marcan además el nacimiento de curvas fold-NS con multiplicadores en $(1, e^{\pm i\theta})$. Estos puntos corresponden a las singularidades de codimensión 3 mencionadas anteriormente. Exceptuando el caso de la curva FF, las restantes se han realizado para valores puntuales del parámetro η_3 dada la imposibilidad de continuarlas con los paquetes estándares. Es de esperar que en cercanías de esta estructura se desarrollen comportamientos de dinámica compleja dado que, por ejemplo, se ha mostrado que la bifurcación fold-NS [12] genera atractores extraños a partir de tangencias homóclinas. Este aspecto no es analizado en el presente trabajo y en los diagramas no se han incluido curvas de bifurcaciones globales como homóclinas y heteróclinas.

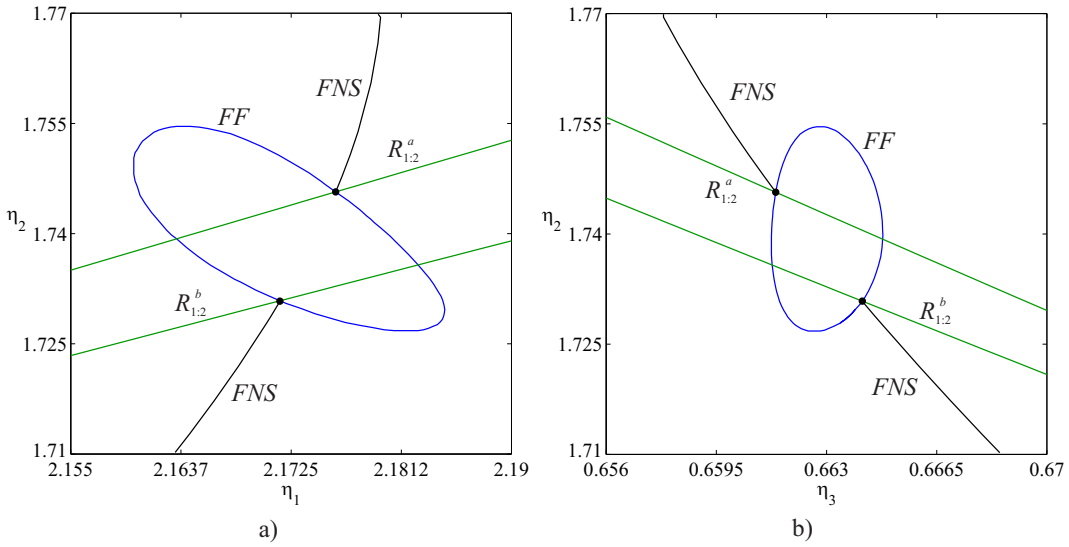


Figura 3: Centro organizador de la dinámica en el espacio de parámetros (η_1, η_2, η_3).

5. Conclusiones

En este trabajo se ha descripto una estructura de bifurcaciones que actúa como centro organizador de la dinámica en tres parámetros, reuniendo curvas de bifurcaciones de codimensión 2 en torno a dos puntos de codimensión 3. Este escenario no se encuentra estudiado en la literatura y los resultados numéricos presentados pueden servir como punto de partida para realizar su estudio analítico. Dado que algunas de las curvas nacen en puntos singulares de la bifurcación de Hopf-Hopf se continuará investigando si la estructura identificada se genera a partir de una bifurcación de más alta codimensión sobre dicha singularidad.

Agradecimientos

La investigación fue financiada por los siguientes subsidios: PICT-2006-00828 (ANPCyT), PGI 24/K041 (UNS) y PIP 112-200801-01112 (CONICET).

Referencias

[1] H. Broer, C. Simó, and R. Vitolo, *Hopf saddle-node bifurcation for fixed points of 3d-diffeomorphisms: Analysis of a resonance 'bubble'*, Physica D, 237 (2008), pp. 1773–1799.

[2] A. Dhooge, W. Govaerts, Y. A. Kuznetsov, W. Maestrom, and B. Sautois, *MATCONT and CL-MATCONT continuation toolboxes in MATLAB*, manual de usuario, Gent University and Utrech University, 2006.

[3] A. I. Khibnik, Y. A. Kuznetsov, V. V. Levitin, and E. V. Nikolaev, *LOCBIF, interactive LOCal BIFurcation analyzer*, manual de usuario, 1992.

[4] B. Krauskopf, *Bifurcation sequences at 1:4 resonance: an inventory*, Nonlinearity, 7 (1994), pp. 1073–1091.

[5] ———, *Strong resonances and Takens' Utrecht preprint*, in Global Analysis of Dynamical Systems, Festschrift dedicated to Floris Takens for his 60th birthday, H. W. Broer, B. Krauskopf, and G. Vegter, eds., Institute of Physics, 2001, pp. 89–111.

[6] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, third ed., 2004.

[7] Y. A. Kuznetsov and H. G. E. Meijer, *Remarks on interacting Neimark-Sacker bifurcations*, Journal of Difference Equations and Applications, 12 (2006), pp. 1009–1035.

[8] Y. A. Kuznetsov, H. G. E. Meijer, and L. V. Veen, *The fold-flip bifurcation*, Int. J. of Bif. and Chaos, 14 (2004), pp. 2253–2282.

[9] G. Revel, D. M. Alonso, and J. L. Moiola, *A gallery of oscillations in a resonant electric circuit: Hopf-Hopf and fold-flip interactions*, Int. J. of Bif. and Chaos, 18 (2008), pp. 481–494.

[10] ———, *Resonancia 2:3 en la bifurcación de Hopf doble*, in Anales del II Congreso de Matemática Aplicada, Computacional e Industrial (II MACI 2009), Rosario, Argentina, 14-16 nov. 2009, pp. 431–434.

[11] ———, *Interactions between oscillatory modes near a 2:3 resonant Hopf-Hopf bifurcation*, Chaos, 20 (2010), pp. 431061–431068.

[12] R. Vitolo, H. Broer, and C. Simó, *Routes to chaos in the Hopf-saddle-node bifurcation for fixed points of 3d-diffeomorphisms*, Nonlinearity, 23 (2010), pp. 1919–1947.

CONTROLLING CHAOS IN THE LOGISTIC MAP BY MODULATION*

Graciela A. González† and Roberta Hansen

Dpto. de Matemática, FIUBA, Universidad de Buenos Aires, Av. Paseo Colón 850, 1063 Buenos Aires, Argentina, ggonzal@fi.uba.ar, rhansen@fi.uba.ar
†*Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.*

Abstract: An approach to stabilize the chaotic dynamics of the logistic map by modulating a control parameter is presented. The work concentrates on proportional and delayed feedback control methods. Some proposals to overcome certain requirements of the OGY method of control chaos are proposed.

Keywords: *chaos control, logistic map, feedback control, parameter modulation.*
2000 AMS Subject Classification: 37N35 - 93D15

1 INTRODUCTION

Chaotic behavior is a very interesting nonlinear phenomena, but in many practical situations it is desirable to be avoided, for example, because it restricts the operating range of electronic or mechanics devices. Moreover, this goal must be achieved with the only help of tiny perturbations properly chosen.

The seminal idea of Ott, Grebogy and Yorke [7] turned the presence of chaos into an advantage. Indeed, the system may be stabilized in a particular unstable periodic orbit (UPO) embedded within a strange attractor. When the trajectory is close to a desired UPO, a small time-dependent feedback perturbation is applied to some accessible parameter or variable system. The periodic orbit is preserved, but its stability is modified keeping the trajectory to stay close to the UPO. This control strategy is known as the OGY method.

A simple proportional feedback (SPF) control method basically consists in perturbing a system parameter by an amount proportional to the difference between the current system value and an unstable fixed point (or an UPO's component). It is well known that it can control chaos for some 1-dimensional maps [2],[7],[9]. The OGY method is a particular case of a SPF. With the aim to overcome some limitations of the OGY method (such as sensitivity to fluctuation of external noises), Pyragas [8] proposed an alternative one, based on a self-controlling delayed feedback, which, as the OGY, does not modify the original UPO, but does not depend explicitely on it. Its discrete-time version results in a perturbation which is proportional to the difference between the current system value and a previous one. This delayed feedback control (DFC) successfully controls chaotic behavior in a variety of experiments [3] (and its references).

The logistic map is the prototypical discrete-time dynamical system with simplest algebraic equation and exhibiting all ingredients of chaos, and so extensively studied [1],[6],[9] (and references wherein):

$$x_{k+1} = f(x_k, r) = r\, x_k\, (1 - x_k)\,, \tag{1}$$

where $x \in [0, 1]$ and r is a control parameter. This map develops a cascade of period–doubling bifurcation that accumulates at $r = r_\infty \approx 3.569$, where the chaotic regime starts with the presence of chaotic attractors, until $r = 4$, where the motion becomes unbounded. We will consider the case $r_0 = 3.8$ for which there is one chaotic attractor, with the interval $[0, 1]$ as its basin of attraction.

The OGY method requires the exact knowledge of both, the UPO to be stabilized, and the linearized dynamics about it. This is not always the case in the real-world implementations. Moreover, as said above, the method also results very sensitive to nonlinearities and noise, mainly for large orbit oscillations. In order to explore the possibility of relaxing these requirements, we investigate mathematically and numerically the approach to stabilize the chaotic dynamics of the map (1), proposing different types of feedback controls by modulating the nonlinear parameter r.

*This work was partially supported by Programación UBACyT 2010–2012.

2 CONTROLLING CHAOS BY FEEDBACK CONTROL

2.1 THE OGY METHOD

Suppose we want trajectories resulting from a randomly chosen initial condition x_0 to be as close as possible to a m–period UPO, $\{p_1, p_2, \ldots, p_m\}$. Suppose also, that the parameter r can be finely tuned in a small range around r_0, namely we allow r to vary in the range $(r_0-\delta, r_0+\delta)$, where $0<\delta\ll 1$. We briefly give the ideas of this method outlined in [7]. Assume that at time k, the trajectory falls into the neighborhood of the component p_i. The linearized dynamics about p_i and r_0 is:

$$x_{k+1}-p_{i+1} \;=\; f_x(p_i, r_0)\,(x_k-p_i) \;+\; f_r(p_i, r_0)\,(r_k-r_0)\,, \tag{2}$$

where $f_x(p_i, r_0) = (1-2p_i)$ and $f_r(p_i, r_0) = p_i(1-p_i)$. To force the system towards the UPO, we set $x_{k+1}-p_{i+1}=0$, and so we have, from (2):

$$u_k^i = (\Delta r)_k = r_k-r_0 = r_0\,\frac{(2p_i-1)}{p_i(1-p_i)}\,(x_k-p_i) = \alpha_i\,(x_k-p_i)\,. \tag{3}$$

Eq. (3) holds only when $|x_k-p_i|\leqslant\varepsilon\ll 1$, hence the required parameter perturbation $(\Delta r)_k$ is small, and the maximum parameter perturbation δ is proportional to ε (with factor α_i). When the trajectory is outside the ε–neighborhood of p_i, we do not apply any perturbation, and the system evolves at its nominal chaotic parameter r_0. Then, for given $\varepsilon, \delta>0$, we want to to stabilize

$$x_{k+1} = [r_0 + u_k^i]\,x_k\,(1-x_k)\,, \quad \text{with} \quad u_k^i = \begin{cases} \alpha_i\,(x_k-p_i) & \text{if} \quad |x_k-p_i|\leqslant\varepsilon\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{4}$$

at the m–period orbit $\{p_i\}_{1\leqslant i\leqslant m}$, with the control bounded by δ, i.e., $|u_k^i|\leqslant\delta$, for all k. The parameter perturbation is of SPF type, i.e., $u_k^i\propto|x_k-p_i|$, and it is time-dependent.

2.2 USING THE OGY METHOD

Note that the control in Eq. (4) depends also on which p_i was selected to stabilize the dynamics on it, namely it is turned on only at the end of each oscillation. This fact makes this control very sensitive to nonlinearities and to fluctuation of external noises that are common in real implementations, mainly for large values of the orbit's period, m. Besides, the control gains, α_i, may differ a lot according to i, which yields to very different performances as regarding the waiting time, even for the same UPO.

In order to improve this drawback, we first propose a simple "switching control", that works as the OGY, but forcing the trajectory to keep close to the m–period UPO, by applying the perturbation u_k^i for all $1\leqslant i\leqslant m$, i.e., each time the trajectory is close to any component p_i. As a first and natural consequence, a net reduction on waiting time will be obtained. The control algorithm is

$$x_{k+1} = [r_0 + u_k]\,x_k\,(1-x_k)\,, \quad \text{with} \quad u_k = \begin{cases} \alpha_i\,(x_k-p_i) & \text{if} \quad |x_k-p_i|\leqslant\varepsilon, 1\leqslant i\leqslant m\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{5}$$

We must take into account that the modulation of r should keep the dynamics to remain globally bounded, so $r_0+|u_k|\leqslant 4$. In our case, for $\varepsilon>0$, $|\alpha_i\,\varepsilon|\leqslant\delta\leqslant 4-r_0=0.2$, or $\varepsilon\leqslant\delta/|\alpha_i|$, for all $1\leqslant i\leqslant m$. As a second consequence, this strategy of control displays a notably better performance than control (4) in presence of external noise. This is seen in Figure 1 where the response of applying the control (5) about a 4-period UPO of map (1), and the controls u_k^i of (4) for each $i=1, 2, 3, 4$, in presence of noise, are compared.

2.3 USING SPF MODULATION

Here we propose to perturb the parameter r_0 with a SPF modulation selected from a set of control laws, similar to (5), but replacing the fixed gain α_i by coefficients β_i adequately chosen,

$$x_{k+1} = [r_0 + u_k]\,x_k\,(1-x_k)\,, \quad \text{with} \quad u_k = \begin{cases} \beta_i\,(x_k-p_i) & \text{if} \quad |x_k-p_i|\leqslant\varepsilon, 1\leqslant i\leqslant m\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{6}$$

This dynamics also preserves the m-UPO. The linear stability criterion for p_i to be an asymptotically stable (a.s.) point, states the condition for β_i. We want $\left|\frac{\partial x_{k+1}}{\partial x_k}(p_i, r_0)\right| < 1$, so this implies a range $(\beta_i^{\inf}, \beta_i^{\sup})$

of possible values for each β_i, with $\beta_i^{\inf} = -\frac{1+r_0(1-2p_i)}{p_i(1-p_i)}$ and $\beta_i^{\sup} = \frac{1-r_0(1-2p_i)}{p_i(1-p_i)}$. The β_i must verify $|\beta_i\,\varepsilon| \leqslant \delta \leqslant 0.2$, $\forall i$, to ensure the desired bound on the control effort and a globally bounded dynamics. We claim that under the conditions stated on the control coefficients, convergence of the algorithm is formally proven. Note that $\alpha_i = \frac{\beta_i^{\inf}+\beta_i^{\sup}}{2}$, and it corresponds to set $\frac{\partial x_{k+1}}{\partial x_k}(p_i, r_0) = 0$, in the system (4). Because $\beta^{\inf} < \alpha$, a benefit is obtained, since smaller values for the control gain are allowed, or else, for the same control effort, δ, a greater ε is allowed, improving the waiting time to achieve the control. As an example, we take the unstable fixed point $p = 1 - \frac{1}{r_0}$, requiring $\delta = 0.2$ as bound on control effort. In Figure 2(a), it is appreciated both: the neat reduction of control effort by changing the control gain α by β, and an increase in the convergence time once the control is turned on. Figure 2(b) shows, for the similar control effort, the reduction in the waiting time explained above.

2.4 Using DFC modulation

In [8] Pyragas proposed to stabilize the logistic map (1) to its unstable fixed point, $p = 1 - \frac{1}{r_0}$, by adding a linear perturbation in the form of one-time delay, $\gamma(x_k - x_{k-1})$. Here we propose to use the same DFC but modulating the parameter r, and to explore the likelihood of its extension to greater than one period orbits,

$$x_{k+1} = [r_0 + u_k]\, x_k\,(1-x_k)\,, \quad \text{with} \quad u_k = \begin{cases} \gamma\,(x_k - x_{k-1}) & \text{if} \quad |x_k - p| + |x_{k-1} - p| \leqslant \varepsilon\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{7}$$

This perturbation can be useful in applications were the fixed point may be unknown or may drift. It vanishes when the system (1) state attains p, and preserves the fixed point. The system (7) yields to the two dimensional one: $x_{k+1} = [r_0 + \gamma(x_k - y_k)]x_k(1-x_k)$; $y_{k+1} = x_k$, which has $P = (p, p)$ as a fixed point. The Jacobian matrix at P is a *companion matrix*, and the conditions for P to be a.s. [5], yields to $\gamma^{\inf} = \frac{r_0^2(r_0-3)}{2(r_0-1)} < \gamma < \gamma^{\sup} = \frac{r_0^2}{r_0-1}$. Fixing a γ in this range, and an adequate $\varepsilon = \varepsilon(\delta)$, the convergence of the algorithm (7) to p is obtained. The resulting control performance may be comparable, or even better, than the one obtained by applying (6) if adequate coefficients are chosen (see Figures 2(c) and (d)).

Our proposal of an extension of (7) to an m-UPO, is to set a "switching" control, as follows

$$x_{k+1} = [r_0 + u_k]x_k(1-x_k),\; u_k = \begin{cases} \gamma_i\,(x_k - x_{k-m}), \text{ if } \sum_{j=0}^{m} |x_{k-j} - p_{(i+j) \bmod(m)}| \leqslant \varepsilon,\; 1 \leqslant i \leqslant m, \\ 0 \qquad\qquad \text{otherwise}\,. \end{cases} \tag{8}$$

This yields to an $(m+1)$-dimensional system with $\{P_i = (p_i, \ldots, p_m, p_1 \ldots, p_i)\}_{1 \leqslant i \leqslant m}$ as m-UPO. The conditions for the orbit to be a.s. involve a product of the m Jacobian matrices at each P_i, which are companions [4]. It is not possible to obtain explicitely, if it exists, a range of γ_i for orbit stability, in any case. However, we can look for a solution, by simply proposing that only a particular $\gamma_i \neq 0$, and the rest of them to be zero. Indeed, this is the case, e. g., for the 2-UPO, $\{p_1, p_2\}$ of (1) shown in Figure 2(e), stabilized by the control (8), where $\gamma_1 = 0$ and $\gamma_2 \in (\gamma_2^{\inf}, \gamma_2^{\sup})$. The response of the control u_k is shown in Figure 2(f). The performance of control (8) to the presence of noise, even for higher period orbits, just as other possible perturbations of the map (1), are under study and they will be communicated elsewhere.

References

[1] Z. Elhadj and J.C. Sprott, *The effect of modulating a parameter in the logistic map*, Chaos 18, 023119 (2008).

[2] D.J. Gauthier, *Resource Letter: CC-1: Controlling chaos*, Am. J. Phys., 71 (2003), pp.750-759.

[3] E. Hellen and K. Thomas, *Prediction and measurement of transient responses of first difference based chaos control for 1-dimensional maps*, arXiv:0807.2637v3 (2010).

[4] E.S. Key and H. Volkmer, *Eigenvalue multiplicities of products of companion matrices*, Elect. J. Linear. Alg. (ELA), 11 (2004), pp. 103-114.

[5] B.C. Kuo, *Sistemas de Control Automático*, Prentice Hall, 1995.

[6] A. Nandi, D. Dutta, J.K. Bhattacharjee and R. Ramaswamy, *The phase-modulated logistic map*, Chaos 15, 023107 (2005).

[7] E. Ott, C. Grebogi and J.A. Yorke, *Controlling Chaos*, Phys. Rev. Lett., 64 (1990), pp.1196-1199.

[8] K. Pyragas, *Continuous control of chaos by self-controlling feedback*, Phys. Lett. A, 170 (1992), pp.421-428.

[9] R.J. Wiener, E. Callan, S.C. Hall and T. Olsen, *Proportional feedback control of chaos in a simple electronic oscillator*, Am. J. Phys., 74 (2006), pp.200-206.

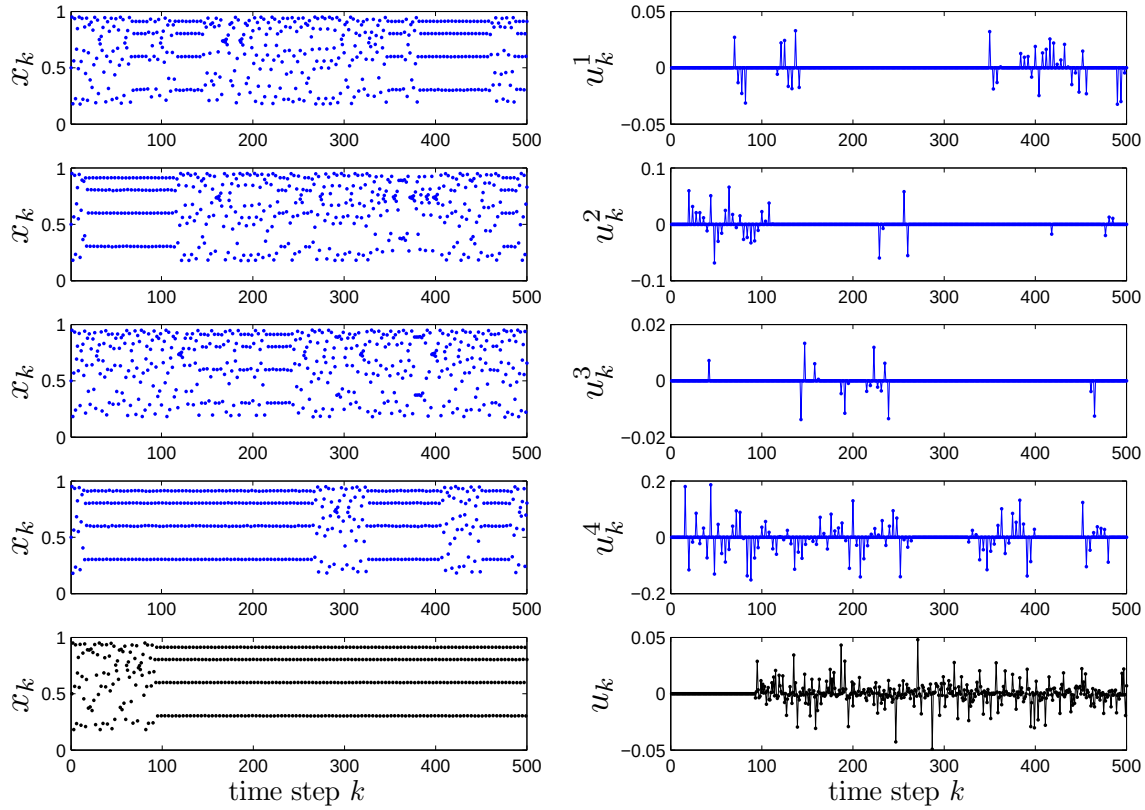


Figure 1: (blue) Time series of x_k, and controls $u_k^i, i=1\text{-}4$ (4), applied to (1) at each i-th of the 4-UPO, $\{p_1 \approx 0.3, p_2 \approx 0.8, p_3 \approx 0.6, p_4 \approx 0.91\}$, with the effect of additive noise modeled by $5\times10^{-4}\sigma_k$, $\sigma_k \sim N(0,1)$, i.c. $x_0 = 0.5, \varepsilon = 0.005$. (black) The same, but when applying the control (5).

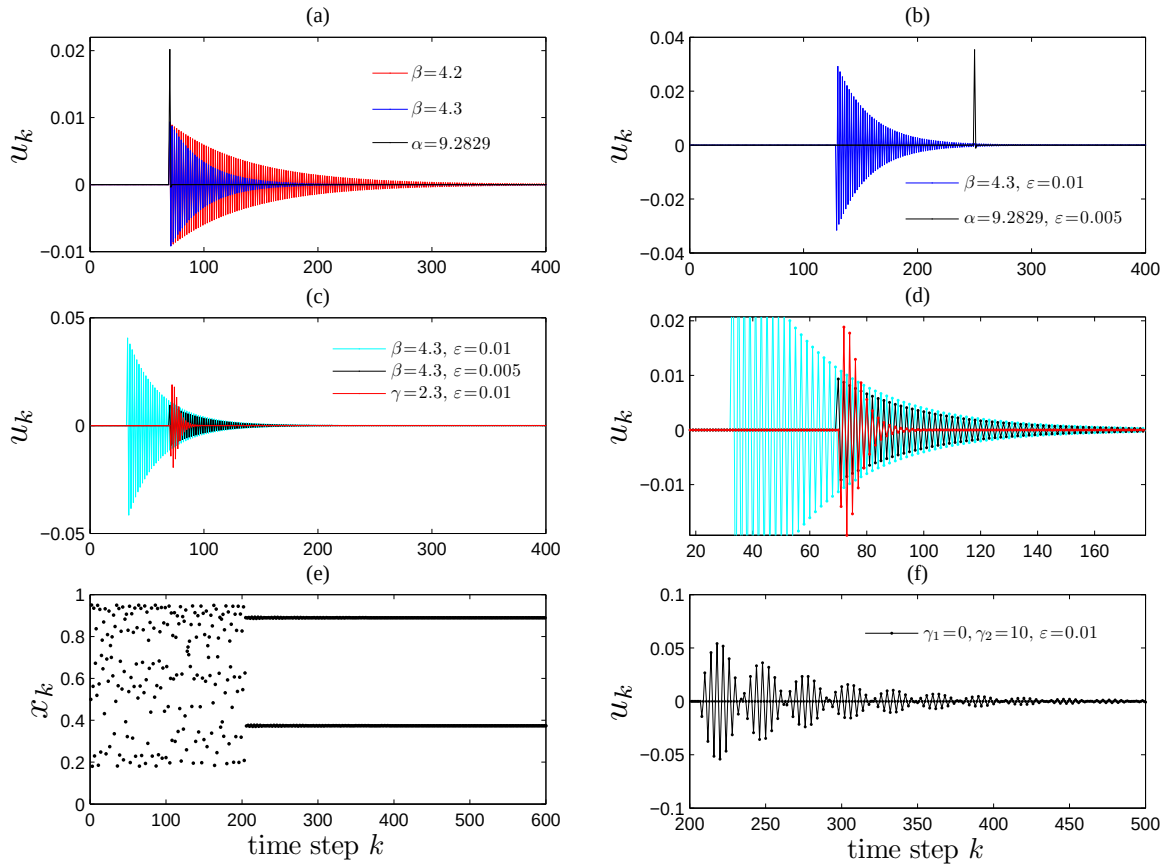


Figure 2: (a) Performance of the controls (5) and (6) to stabilize (1) about the fixed point $p \approx 0.736$, $x_0 = 0.94, \varepsilon = 0.005$, ($\beta^{\inf} \approx 4.126, \beta^{\sup} \approx 14.44$). (b) Idem (a), but keeping the control effort, and varying the ε values, $x_0 = 0.5$. (c) Idem (a) and performance of the control (7), $x_0 = 0.94, y_0 = 0.45$. (d) Expansion of (c). (e)-(f) Time series of x_k, and control (8) applied to stabilize the 2-UPO, $\{p_1 \approx 0.374, p_2 \approx 0.89\}$, i.c. $x_0 = y_0 = z_0 = 0.5, \varepsilon = 0.01$ ($\gamma_2^{\inf} \approx 9.12, \gamma_2^{\sup} \approx 763.97$).

ANÁLISIS DE ESTABILIDAD DE SOLUCIONES PERIÓDICAS

Griselda R. Itovich[♭] y Jorge L. Moiola[†,‡]

♭ Escuela de Tecnología, Producción y Medio Ambiente - Sede Alto Valle - Universidad Nacional de Río Negro, Tacuarí 669, (8336) Villa Regina, ARGENTINA, gitovich@arnet.com.ar

[†]Instituto de Investigaciones en Ingeniería Eléctrica - IIIE (UNS-CONICET) y [‡]Dpto. de Ing. Eléctrica y de Computadoras - Universidad Nacional del Sur, Avda. Alem 1253, (B8000CPB) Bahía Blanca, ARGENTINA

Resumen: En este trabajo se analiza la aplicación de polinomios de Tchebyshev para aproximar la matriz fundamental de soluciones de un sistema lineal periódico. Esta metodología permite analizar la estabilidad de una órbita cuya expresión aproximada se obtuvo empleando métodos frecuenciales, a través de un balance de armónicos.

Palabras claves: Soluciones periódicas. Dominio frecuencia. Estabilidad, Polinomios deTchebyshev
2000 AMS Subject Classification: 34C25 - 41A50

1. INTRODUCCIÓN

La aparición de una rama de soluciones periódicas y el análisis de su estabilidad puede advertir sobre otros posibles escenarios dinámicos que pueden presentarse, incluyendo el caos. Un enfoque para analizar bifurcación de soluciones en sistemas de ecuaciones diferenciales ordinarias consiste en llevar el problema al dominio frecuencia, técnica que resulta muy familiar para ingenieros y especialistas en control [4], [5]. Una vez probada la existencia de una órbita, la evaluación de su estabilidad se consigue por medio de la resolución de un sistema lineal con coeficientes periódicos [2]. Este tipo de sistemas han sido profundamente estudiados debido a su aparición reiterada en una variedad de modelos y se conocen diferentes técnicas para resolverlos con métodos de promediación, perturbación, teoría de Floquet e integración numérica y también los que consideran, un sistema distinto del original, donde la matriz periódica se aproxima por medio de funciones constantes o lineales a trozos o bien por polinomios ortogonales, como los de Tchebyshev [6],[7].

A continuación, en la Sección 2 se presentan los sistemas lineales con coeficientes periódicos. En la Sección 3, se introducen los polinomios de Tchebyshev y se establecen propiedades y aplicaciones sencillas. Posteriormente, en la Sección 4, se describe la metodología en frecuencia como marco para el análisis dinámico de ecuaciones diferenciales, en particular para el tratamiento de la bifurcación de Hopf y sus ciclos emergentes. En la Sección 5, se combinan los resultados de la Secciones 2, 3 y 4 para estudiar la estabilidad en soluciones periódicas. Por último, en la Sección 6, se presentan algunas conclusiones y objetos de trabajo futuro.

2. SISTEMAS LINEALES CON COEFICIENTES PERIÓDICOS

Se considera un sistema de ecuaciones diferenciales del tipo

$$\dot{x} = A(t)x, \tag{1}$$

donde $x = x(t) \in R^n, A(t+T) = A(t)$ donde T es el periodo minimal de A, esto es, T es el número positivo más pequeño para el cual se cumple la última igualdad. Hay que notar que las soluciones de la ecuación (1) no son necesariamente periódicas. Sin embargo, un sistema tal tiene por lo menos una solución no trivial que cumple la condición $x(t+T) = \mu\, x(t)$ donde μ es una constante [1]. Si $\mu = 1$, dicha solución resulta periódica. Se puede demostrar que μ es un autovalor de $\Phi(T)$ donde Φ es una matriz fundamental de soluciones del sistema (1) que satisface $\Phi(0) = I_n$ (la matriz identidad de orden n). Los autovalores de $\Phi(T)$ se conocen como multiplicadores característicos o de Floquet. En cualquier caso, para construir Φ, se requiere hallar una expresión aproximada de la solución periódica y esto puede lograrse por medio de polinomios de Tchebyshev [7]. El planteo de una tal aproximación lleva a la resolución de un sistema de ecuaciones lineales cuyo orden es n veces la cantidad de polinomios de Tchebyshev que se propone emplear. Se ha visto que esta metodología mejora los resultados que se obtienen empleando teoría de perturbaciones, aunque debería recomendarse sólo para sistemas de baja dimensión.

3. POLINOMIOS DE TCHEBYSHEV

3.1. GENERALIDADES

Los polinonios de Tchebyshev de primer tipo se definen en el intervalo $[-1,1]$ a partir de la relación de recurrencia:

$$T_0(t)=1,\ \ T_1(t)=t,\ \ T_n(t)=2tT_{n-1}(t)-T_{n-2}(t),\quad \text{para}\ \ n\geq 2.$$

De esta manera, allí resultan polinomios ortogonales con respecto a la función de peso $p(t)=(1-t^2)^{-\frac{1}{2}}$, esto es, $\int_{-1}^{1} T_n(t)T_m(t)p(t)dt=0$, si $n\neq m$. Usando un cambio de variables como $t^*=\frac{t+1}{2}$, se obtienen los polinomios de Tchebyshev de primer tipo T_n^*, trasladados al intervalo $[0,1]$, que se definen como $T_n^*(t)=T_n(2t-1),\ \ 0\leq t\leq 1$, y son los que utilizaremos en este trabajo. Las relaciones de ortogonalidad que estos polinomios satisfacen son: $\int_0^1 T_n^*(t)T_m^*(t)p^*(t)dt=0$ si $n\neq m$, $\frac{\pi}{2}$ si $n=m\neq 0$ y π si $n=m=0$, donde $p^*(t)=(t-t^2)^{-\frac{1}{2}}$. Generalmente una función continua $f(t)$ en el intervalo $[0,1]$ puede ser representada por medio de una serie de Tchebyshev como $f(t)=\sum_{n=0}^{\infty} a_nT_n^*(t),\ \ 0\leq t\leq 1$, donde los coeficientes a_n se pueden obtener como $a_n=\frac{1}{\delta}\int_0^1 f(t)T_n^*(t)p^*(t)dt$ donde $\delta=\frac{\pi}{2}$ si $n\neq 0$ y π si $n=0$. Una representación finita de una función continua en polinomios de Tchebyshev, como se describe arriba, puede ser obtenida en forma "automática" mediante paquetes especiales incluidos en los programas de cálculo simbólico más conocidos. Por ejemplo, una aproximación de octavo orden para la función $f(t)=e^t,\ \ 0\leq t\leq 1$ resulta $P_8(t)=\sum_{n=0}^{8} a_nT_n^*(t)$ cuyos coeficientes de Tchebyshev son: $a_0=1{,}75338, a_1=0{,}85039, a_2=0{,}10520, a_3=0{,}00872, a_4=0{,}54343*10^{-3}, a_5=0{,}27115*10^{-4}$, $a_6=0{,}11281*10^{-5}, a_7=0{,}40245*10^{-7}$ y $a_8=0{,}12565*10^{-8}$.

3.2. APLICACIONES

Ejemplo 1. Se considera la ecuación diferencial lineal con coeficiente periódico $\dot{x}(t)=\cos(3t)x(t)$. La función $x(t)=\exp(\frac{1}{3}\sin(3t))$ es una solución exacta y periódica de la ecuación dada que satisface $x(0)=1$ y tiene periodo minimal $T=\frac{2\pi}{3}$. Para hallar una aproximación de la solución de la ecuación anterior mediante polinomios de Tchebyshev, se efectúa un cambio de variables, a los efectos de trabajar en el intervalo $[0,1]$. La ecuación normalizada resulta $\dot{y}(t)=\frac{2\pi}{3}\cos(2\pi t)y(t)$ y su solución exacta es $y(t)=\exp(\frac{1}{3}\sin(2\pi t))$.
Una solución aproximada que emplea 11 polinomios de Tchebyshev es $Y_{10}(t)=\sum_{n=0}^{10} a_nT_n^*(t)$, $Y_{10}(0)=1$, por lo que deberán determinarse los 11 coeficientes a_n. Para esto, se sustituye esta expresión en la ecuación diferencial normalizada de donde resulta

$$Y_{10}'(t)=\sum_{n=1}^{10} a_nT_n^{*\prime}(t)=\sum_{n=0}^{9} b_nT_n^*(t)=\tilde{A}(t)\sum_{n=0}^{10} a_nT_n^*(t)=\sum_{n=0}^{10} c_nT_n^*(t),$$

donde $\tilde{A}(t)$ es una representación de la función $A(t)=\frac{2\pi}{3}\cos(2\pi t)$ en polinomios de Tchebyshev, cuya precisión es del orden de 10^{-10}. De aquí, se pueden extraer 10 ecuaciones y añadiendo la condición inicial se completa el sistema de ecuaciones lineales a resolver. Así resulta que la solución aproximada es $Y_{10}(t)=\sum_{n=0}^{10} a_nT_n^*(t)$ donde $a_0=1{,}02177, a_1=-0{,}19183, a_2=-0{,}01611, a_3=0{,}22565, a_4=-0{,}01767, a_5=-0{,}03570, a_6=0{,}01561, a_7=0{,}00139, a_8=-0{,}00413, a_9=0{,}00044$ y $a_{10}=0{,}00050$. Se comparó la solución numérica obtenida con Matlab con la exacta y esto arrojó que el error cometido es del orden de 10^{-4}.

Para llevar adelante la aplicación de la técnica elegida, se han empleado algoritmos escritos ad-hoc en un programa de cálculo simbólico. Este procedimiento puede extenderse al caso de sistemas de ecuaciones diferenciales lineales con coeficientes periódicos, como se ha efectuado en los ejemplos que siguen.

Ejemplo 2. Se analizará la ecuación de Mathieu $\ddot{x}(t)+(v+u\cos(\Omega t))x(t)=0,\ t>0$, donde v,u son parámetros del modelo y $T=\frac{2\pi}{\Omega}$ es su periodo. Reescalando como $y(t)=x(Tt)$ resulta la ecuación de Mathieu normalizada $\ddot{y}(t)+T^2(v+u\cos(2\pi t))y(t)=0$, que se puede escribir como

$$\begin{bmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ -T^2(v+u\cos(2\pi t)) & 0 \end{bmatrix}\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}=A(t)\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}. \tag{2}$$

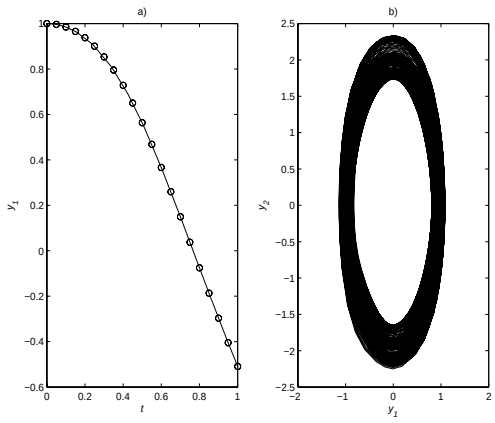


Figura 1: a) Soluciones de la ecuación de Mathieu normalizada con $v = 1, u = -0{,}32$ y $\Omega = 3$. ('$\circ$' Polinomios de Tchebyshev, '–' Matlab) b) Plano de fases de la solución.

Se consideran los valores $v = 1, u = -0{,}32$ y $\Omega = 3$, para cotejar resultados con [7], donde se aproxima la solución partiendo de las condiciones iniciales $y_1(0) = 1, y_2(0) = 0$. En (2), la matriz fundamental $\Phi(1)$ tiene como autovalores $-0{,}5091 \pm 0{,}8600i$, que significa que sus soluciones no son periódicas. Para hallar una aproximación de la solución se establece $\tilde{y}_i(t) = \sum_{j=0}^{m-1} a_{ij} T_i^*(t),\ i = 1, 2$ con $m = 11$. Estas expresiones se sustituyen en (2) y añadiendo las condiciones iniciales queda determinado un sistema de 22 ecuaciones lineales cuyas 22 incógnitas son los coeficientes a_{ij} de las soluciones $\tilde{y}_1$ e $\tilde{y}_2$. De esta forma, para la primera incógnita se obtiene que $a_{10} = 0{,}39523, a_{11} = -0{,}78843, a_{12} = -0{,}15816, a_{13} = 0{,}36103 * 10^{-1}, a_{14} = 0{,}90929 * 10^{-2}, a_{15} = -0{,}26324 * 10^{-2}, a_{16} = -0{,}88083 * 10^{-3}, a_{17} = 0{,}34871 * 10^{-3}, a_{18} = 0{,}80185 * 10^{-4}, a_{19} = -0{,}24802 * 10^{-4}, a_{1(10)} = -0{,}47637 * 10^{-5}$. En la Figura (1) a), se muestra la representación de la solución aproximada $\tilde{y}_1$, junto con la que se obtiene con el programa Matlab. Por otra parte, en la Figura (1) b), aparece el plano de fases de la solución de (2), cuando $y_1(0) = 1, y_2(0) = 0$.

4. Sistemas autónomos

4.1. La metodología en el dominio frecuencia

Se considera un sistema de ecuaciones diferenciales ordinarias del tipo

$$\dot{x} = f(x, \mu), \tag{3}$$

donde $x = x(t) \in R^n$ y μ es un parámetro de bifurcación. Para analizar la dinámica de un sistema como (3) se reescribe el mismo introduciendo, en primer lugar, variables de salida $y \in R^m$ y de entrada $u \in R^s$ para interpretarlo como un sistema realimentado, esto es $\dot{x} = Ax + Bu, \quad y = -Cx, \quad u = g(y, \mu)$, donde A, B y C son matrices de órdenes apropiados que pueden depender del parámetro μ. Luego, si se adiciona la condición inicial $x(0) = 0$ y se aplica transformada de Laplace, la ecuación de equilibrios resultante es $\hat{y} = -G(0, \mu) g(\hat{y}, \mu)$, donde $G(s, \mu) = C(sI_n - A)^{-1}B$, s es la variable de la transformada y para el caso de órbitas su estabilidad se deduce a partir de los autovalores de la matriz $G(., \mu)\frac{\partial g}{\partial y}$. Así se establece el teorema gráfico de Hopf que da condiciones suficientes para la aparición de una rama de soluciones periódicas [3],[4]. Al mismo tiempo, esta herramienta permite hallar una expresión cuasianalítica aproximada con distintos niveles de precisión, conforme el número de armónicos involucrados [5].

4.2. Estabilidad de soluciones periódicas

Para estudiar la estabilidad de una solución $T-$periódica $\gamma = \gamma(t, \mu^*)$ de (3) que resulta de una bifurcación de Hopf, se considera el siguiente sistema $\dot{\Phi} = \frac{\partial f}{\partial x}\Big|_{\Gamma} \Phi, \quad \Phi(0) = I_n$, donde $\Phi = \Phi(t) \in R^{n \times n}$ y $\Gamma = (\gamma(t, \mu^*), \mu^*)$, que es un sistema de tipo (1). La dinámica del ciclo considerado depende de los autovalores de $\Phi(T)$, que, en este caso, tiene siempre el autovalor trivial 1. Si todos los restantes tienen módulo menor que 1, se trata de una órbita estable. Cuando exista un multiplicador de Floquet situado por afuera del círculo unitario, el ciclo será inestable.

5. APLICACIÓN

Se considera el modelo de van der Pol dado por la ecuación $\ddot{x} + \epsilon\dot{x}\left(x^2 - 1\right) + x = 0, \epsilon \geq 0$, que resulta equivalente al sistema: $\dot{x}_1 = x_2, \dot{x}_2 = -x_1 + \epsilon x_2 - x_1^2 x_2$. En $\epsilon = 0$, aparece una bifurcación de Hopf supercrítica con frecuencia inicial $\omega = 1$. Por medio de la metodología en el dominio frecuencia, se consigue una expresión aproximada de segundo orden para la solución periódica como $(x_1^*(t),\ x_2^*(t)) = (-2\sqrt{\epsilon}\sin t, -2\sqrt{\epsilon}\cos t)$. De acuerdo con la última sección, para analizar la estabilidad de esta órbita, se debe resolver el sistema

$$\begin{bmatrix} \dot{\phi}_1 & \dot{\phi}_2 \\ \dot{\phi}_3 & \dot{\phi}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 - 4\epsilon\sin 2t & -4\epsilon\sin^2 t + \epsilon \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_3 & \phi_4 \end{bmatrix}, \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_3 & \phi_4 \end{bmatrix}_{t=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

para luego hallar los autovalores de la matriz solución para $t = 2\pi$. Estos valores se pueden obtener numéricamente, por ejemplo usando Matlab y así analizar para distintos valores de ϵ. Por otra parte, si se normaliza este sistema de modo que la matriz de coeficientes tenga periodo 1, como

$$\begin{bmatrix} \dot{\phi}_1 & \dot{\phi}_2 \\ \dot{\phi}_3 & \dot{\phi}_4 \end{bmatrix} = 2\pi \begin{bmatrix} 0 & 1 \\ -1 - 4\epsilon\sin 4\pi t & -4\epsilon\sin^2 2\pi t + \epsilon \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_3 & \phi_4 \end{bmatrix},$$

se puede aproximar sus soluciones empleando polinomios de Tchebyshev en el intervalo $[0, 1]$ y luego evaluar para $t = T = 1$, de donde sigue la estabilidad de la órbita en cuestión. En este sentido, empleando desarrollos que incluyen estos polinomios hasta el grado 10, se hallaron los multiplicadores de Floquet y se compararon con los valores que resultan de Matlab. Estos resultados, que concuerdan, se muestran en la Tabla 1.

6. CONCLUSIONES

En este trabajo se empleó una técnica de resolución de sistemas lineales periódicos para analizar la estabilidad de ciclos límites, que aparecen a través de una bifurcación de Hopf. La determinación de las expresiones aproximadas para las órbitas, necesarias para el procedimiento anterior, se consiguieron por medio de la metodología en frecuencia. Se espera poder extender la aplicación de estos recursos para determinar la pérdida de la estabilidad de soluciones periódicas.

AGRADECIMIENTOS

G. R. I. agradece la ayuda brindada por la UNRN (40/A-008) y J. L. M. hace un similar reconocimiento al CONICET. Este trabajo ha sido subvencionado por aportes de la SGCyT de la UNS (24/K041) y por el PICT-2006-00828 (ANPCyT) y PIP 112-200801-01112 (CONICET).

REFERENCIAS

[1] D. JORDAN AND P. SMITH, Nonlinear Differential Equations, Oxford University Press, Avon, Gran Bretaña, tercera edición, 1999.

[2] Y. A KUZNETSOV, Elements of Applied Bifurcation Theory, Applied Mathematics Sciences 112. Springer-Verlag, Nueva York, segunda edición, 1998.

[3] A.I. MEES AND L. CHUA, The Hopf bifurcation theorem and its applications to nonlinear oscillations in circuits and systems, IEEE Transactions on Circuits and Systems, 26(4) (1979), 235- 254.

[4] A.I. MEES, Dynamics of Feedback Systems, John Wiley and Sons, Nueva York, 1981.

[5] J.L. MOIOLA AND G. CHEN, Hopf Bifurcation Analysis: A Frequency Domain Approach, World Scientific, Singapur, 1996.

[6] S. C. SINHA, C. C. CHOU AND H.H. DENMAN, Stability analysis of systems with periodic coefficients: an approximate approach, Journal of Sound and Vibration, 64 (1979), 515-527.

[7] S. C. SINHA AND D.-H. WU, An efficient computational scheme for the analysis of periodic systems, Journal of Sound and Vibration, 151(1) (1991), 91-117.

ϵ	0,01	0,05	0,1
Tcheb.	1,00018	1,00092	1,00197
Matlab	0,99991	0,99995	1,00021

ϵ	0,01	0,05	0,1
Tcheb.	0,93961	0,73024	0,53643
Matlab	0,93900	0,73029	0,53331

Tabla 1: Multiplicadores de Floquet. Izquierda: Multiplicador trivial, Derecha: Multiplicador restante.

FURTHER INVARIANCE RESULTS FOR SWITCHED SYSTEMS

J. L. Mancilla-Aguilar[♭] and R.A. García[†]

Department of Mathematics, Instituto Tecnológico de Buenos Aires,
jmancill@itba.edu.ar[♭], ragarcia@itba.edu.ar[†]

Abstract: In this paper we present invariance principles for switched nonlinear systems with otherwise arbitrary compact index sets and subjected to restrictions originated by the timing of the switchings, state-dependent constrained switching and switching whose logic has memory, i.e., the active subsystem only can switch to a prescribed subset of subsystems.

Keywords: *switched systems, invariance principles, robust stability.*
2000 AMS Subject Classification: 34A60 - 93D05

1 INTRODUCTION

In the last decade various extensions to switched systems of LaSalle's invariance principle for differential equations (see [8]) were derived in order to establish the convergence of the solutions of the switching system to an equilibrium point, and consequently the asymptotic stability. Thus, in [6] Hespanha introduced an invariance principle for switched linear systems under *persistently dwell-time* switching signals and in [7] Hespanha *et al.* extended some of those results to a family of nonlinear systems. Bacciotti and Mazzi presented in [1] an invariance principle for switched systems with *dwell-time* signals. An invariance principle for switched nonlinear systems with *average dwell-time* signals that satisfy *state-dependent* constraints was derived by Mancilla-Aguilar and García in [10] from the sequential compactness of particular classes of trajectories of switched systems. Based on invariance results for hybrid systems, Goebel *et al.* in [5] obtained recently invariance results for switched systems assuming different dwell-time conditions. Lee and Jiang in [9] gave a generalized version of Krasovskii-LaSalle Theorem for time-varying switched systems. Under certain ergodicity and dwell-time conditions on the switching signal, some stability results were also obtained in [3, 11, 12].

Most of the invariance results for switched systems already published only consider restrictions originated by the timing of the switchings or by the state dependence of it. Nevertheless there is also an important restriction to take into account: the fact that not all the subsystems may be accessible from a particular one, i.e. the case in which the switching logic has memory. This restriction is clearly exhibited, for example, in switched systems which are the continuous portion of a hybrid automaton (see [4]).

In this paper we present invariance results that hold for trajectories of switched systems with a non necessarily finite number of subsystems and whose switching signals verify certain property (which we call **L**) with respect to a certain subclass $\mathcal{S}^*$ of switching signals. These results enable us to obtain in an unified way invariance theorems for switched systems whose switchings satisfy some of the restrictions mentioned above. It must be pointed out that property **L** does not involve any dwell-time condition, as the above mentioned results do, and that it is robust w.r.t. a certain type of perturbations.

2 BASIC DEFINITIONS

In this work we consider switched systems described by

$$\dot{x} = f(x, \sigma) \tag{1}$$

where x takes values in $\mathbb{R}^n$, $\sigma : \mathbb{R} \to \Gamma$, with (Γ, d_Γ) a compact metric space, is a *switching signal, i.e.,* σ is piecewise constant (it has at most a finite number of jumps in each compact interval) and is continuous from the right and $f : \mathrm{Dom}(f) \to \mathbb{R}^n$, with $\mathrm{Dom}(f)$ a closed subset of $\mathbb{R}^n \times \Gamma$, is continuous.

For each $\gamma \in \Gamma$, the dynamical system defined by $\dot{x} = f_\gamma(x)$, with $f_\gamma(\xi) = f(\xi, \gamma)$ for all $\xi \in \chi_\gamma := \{\xi \in \mathbb{R}^n : \ (\xi, \gamma) \in \mathrm{Dom}(f)\}$, is the subsystem or mode γ of the switched system and χ_γ is its corresponding state space.

Given $\sigma \in \mathcal{S}$, where $\mathcal{S}$ denotes the set of all the switching signals, a solution of (1) corresponding to σ is a locally absolutely continuous function $x : I_x \to \mathbb{R}^n$, with $I_x \subset \mathbb{R}$ a nonempty interval, such that $(x(t), \sigma(t)) \in \mathrm{Dom}(f)$ for all $t \in I_x$ and $\dot{x}(t) = f(x(t), \sigma(t))$ for almost all $t \in I_x$. The solution x is complete (forward complete) if $I_x = \mathbb{R}$ ($\mathbb{R}_{\geq 0} \subset I_x$). A pair (x, σ) is a trajectory of (1) if $\sigma \in \mathcal{S}$ and x is a solution of (1) corresponding to σ. The trajectory is (forward) complete if x is (forward) complete.

Given a subset $\mathcal{O}$ of $\mathbb{R}^n$, we say that the trajectory (x, σ) is precompact relative to $\mathcal{O}$ if there exists a compact set $B \subset \mathcal{O}$ such that $x(t) \in B$ for all $t \in I_x$.

Finally, we denote by $\pi_1 : \mathbb{R}^n \times \Gamma \to \mathbb{R}^n$ the projection onto the first component

Remark 1 *By assuming that* $\mathrm{Dom}(f)$ *is a subset of* $\mathbb{R}^n \times \Gamma$ *instead of the whole of* $\mathbb{R}^n \times \Gamma$ *one can take into account, in the analysis of the asymptotic behavior of a given trajectory* (x, σ) *of (1), some kind of state-dependent constraints which the trajectory under study must satisfy.*

In this paper we consider forward complete solutions of (1) corresponding to switching signals σ which belong to particular subclasses of $\mathcal{S}$. Let $\Lambda(\sigma)$ be the set of times (switching times) where σ has a jump. Following [6] we say that $\sigma \in \mathcal{S}$ has *a)* a dwell-time $\tau_D > 0$ if $|t - t'| \geq \tau_D$ for any pair $t, t' \in \Lambda(\sigma)$ such that $t \neq t'$, *b)* an average dwell-time $\tau_D > 0$ and a chatter bound $N_0 \in \mathbb{N}$ if for any open finite interval $(\tau_1, \tau_2) \subset \mathbb{R}$, $\mathrm{card}(\Lambda(\sigma) \cap (\tau_1, \tau_2)) \leq N_0 + (\tau_2 - \tau_1)/\tau_D$.

We denote by $\mathcal{S}_a[\tau_D, N_0]$ the set of all the switching signals which have an average dwell-time $\tau_D > 0$ and a chatter bound $N_0 \in \mathbb{N}$ and by $\mathcal{S}_d[\tau_D]$ the set of switching signals σ which have a dwell-time $\tau_D > 0$.

For Γ finite and $T > 0$, $\mathcal{S}_e[T]$ denotes the family of all the switching signals σ which verify the following "ergodicity" condition ([3]): for every $t_0 \in \mathbb{R}$ and every $\gamma \in \Gamma$, $\sigma^{-1}(\gamma) \cap [t_0, t_0 + T] \neq \emptyset$ and $\mathcal{S}_e = \bigcup_{T>0} \mathcal{S}_e[T]$.

The families of switching signals already introduced have no restrictions on the accessibility from any subsystem to another. The family of switching signals —and their corresponding trajectories— that we introduce next, models systems in which the switching logic has memory, i.e. when a subsystem corresponding to an index $\gamma \in \Gamma$ can only switch to subsystems corresponding to modes γ' that belong to a certain subset $\Gamma_\gamma \subset \Gamma$. Given a set-valued map $H : \Gamma \rightsquigarrow \Gamma$, $\mathcal{S}^H$ is the set of all the switching signals σ which verify the condition $\sigma(t) \in H(\sigma(t^-))$ for every time $t \in \Lambda(\sigma)$. Here $\sigma(t^-) = \lim_{s \to t^-} \sigma(s)$. This class of switching signals enable us, for example, to model the restrictions imposed by the discrete process of a hybrid system whose continuous portion is as in (1) (see [4]).

3 Invariance results

The invariance results that we present below enable us to characterize the asymptotic behavior of a precompact forward complete trajectory (x, σ) of (1) when σ verifies property **L** with respect to a certain subclass $\mathcal{S}^*$ of switching signals. By considering such subclass we can to obtain in an unified way invariance results for systems whose switching signals undergo different restrictions. Next we introduce the concept of invariance considered along this work and property **L**.

Definition 1 *Given a family* $\mathcal{T}^*$ *of complete trajectories of (1), we say that a nonempty subset* $M \subset \mathbb{R}^n \times \Gamma$ *is* weakly-invariant w.r.t $\mathcal{T}^*$ *if for each* $(\xi, \gamma) \in M$ *there is a trajectory* $(x, \sigma) \in \mathcal{T}^*$ *such that* $x(0) = \xi$, $\sigma(0) = \gamma$ *and* $(x(t), \sigma(t)) \in M$ *for all* $t \in \mathbb{R}$.

Definition 2 *Let* $\mathcal{S}^*$ *be a family of translation-invariant switching signals (i. e. for any* $s \in \mathbb{R}$ *and any* $\overline{\sigma} \in \mathcal{S}^*$, $\overline{\sigma}(\cdot + s) \in \mathcal{S}^*$*). We say that a switching signal* $\sigma \in \mathcal{S}$ *has property* **L** *with respect to* $\mathcal{S}^*$ *if for any sequence of times* $\{s_k\}$ *with* $s_k \nearrow \infty$, *and* $\sigma_k(\cdot) = \sigma(\cdot + s_k)$ *there exist* $\sigma^* \in \mathcal{S}^*$ *and a subsequence* $\{\sigma_{k_l}\}$ *such that* $\lim_{l \to \infty} \sigma_{k_l}(t) = \sigma^*(t)$ *for almost all* $t \in \mathbb{R}$.

The following result gives conditions under which a switching signal σ has property **L** w.r.t. some of the translation-invariant families introduced above.

Lemma 1 *Let* $\sigma \in \mathcal{S}$ *then*

1. *σ has property* **L** *w.r.t. $\mathcal{S}_a[\tau_D, N_0]$ if for some $s \in \mathbb{R}$, $\mathrm{card}(\Lambda(\sigma) \cap (\tau_1, \tau_2)) \leq N_0 + (\tau_2 - \tau_1)/\tau_D$ for all τ_1, τ_2 such that $s \leq \tau_1 < \tau_2$;*

2. *σ has property* **L** *w.r.t. $\mathcal{S}_d[\tau_D] \cap \mathcal{S}^H$, with $H : \Gamma \rightsquigarrow \Gamma$ such that $\mathrm{Graph}(H) = \{(\gamma, \gamma') \in \Gamma \times \Gamma : \gamma' \in H(\gamma)\}$ is closed, if for some $s \in \mathbb{R}$, $\sigma(t) \in H(\sigma(t^-))$ for all $t \in \Lambda(\sigma) \cap (s, \infty)$ and $|t - t'| \geq \tau_D$ for all $t, t' \in \Lambda(\sigma) \cap (s, \infty)$, $t \neq t'$;*

3. *σ has property* **L** *w.r.t. $\mathcal{S}_d[\tau_D] \cap \mathcal{S}_e[T]$ if for some $s \in \mathbb{R}$, $|t - t'| \geq \tau_D$ for all $t, t' \in \Lambda(\sigma) \cap (s, \infty)$, $t \neq t'$, and $\sigma^{-1}([t, t + T]) = \Gamma$ for all $t > s$.*

The next result shows that property **L** is robust with respect to some kind of perturbations.

Lemma 2 *Suppose that $\sigma \in \mathcal{S}$ has property* **L** *w.r.t. $\mathcal{S}^*$. Let $\tilde{\sigma} \in \mathcal{S}$ such that*

$$\lim_{t \to +\infty} \int_t^{t+T} d_\Gamma(\tilde{\sigma}(s), \sigma(s)) ds = 0 \quad \forall T > 0, \tag{2}$$

Then, $\tilde{\sigma}$ has property **L** *w.r.t. $\mathcal{S}^*$.*

In what follows, given a Lebesgue measurable subset $A \subset \mathbb{R}$, $\mu(A)$ stands for its Lebesgue measure.

Remark 2 *It follows readily that condition (2) is equivalent to the following:*

$$\lim_{n \to +\infty} \mu(I_\varepsilon \cap [n, n+1]) = 0 \quad \forall \varepsilon > 0, \tag{3}$$

where $I_\varepsilon = \{s \in \mathbb{R} : \ d_\Gamma(\tilde{\sigma}(s), \sigma(s)) \geq \varepsilon\}$.

Next, we present two invariance results that involve the existence of a function V which is nonincreasing along a trajectory of (1).

Definition 3 *We say that a function $V : \mathrm{Dom}(V) \to \mathbb{R}$ belongs to class $\mathcal{V}$, if it verifies:* (a) $\mathrm{Dom}(V) = \mathcal{O} \times \Gamma$, *with $\mathcal{O} \subset \mathbb{R}^n$ open;* (b) *for all $\gamma \in \Gamma$, $V_\gamma(\cdot) := V(\cdot, \gamma)$ is differentiable on $\mathcal{O}_\gamma := \mathcal{O} \cap \chi_\gamma$.*

In our first invariance result we consider the following assumption.

Assumption 1 *The forward complete trajectory (x, σ) of (1) verifies the following: there exists a function $V \in \mathcal{V}$ whose restriction to $\mathrm{Dom}(f) \cap \mathrm{Dom}(V)$ is continuous, (x, σ) is precompact relative to $\mathcal{O}$ and $v(t) = V(x(t), \sigma(t))$ is nonincreasing on $[0, +\infty)$.*

Theorem 1 *Suppose that (x, σ), with $\sigma \in \mathcal{S}$, is a forward complete trajectory of (1) for which Assumption 1 holds. Suppose in addition that σ has property* **L** *w.r.t. a family of translation-invariant switching signals $\mathcal{S}^*$ and let $\mathcal{T}^*$ be the set of all the complete trajectories $(\overline{x}, \overline{\sigma})$ of (1) with $\overline{\sigma} \in \mathcal{S}^*$.Then there exists $c \in \mathbb{R}$ such that x converges to a connected component of $\pi_1(M(c))$, where $M(c)$ is the maximal weakly-invariant set w.r.t. $\mathcal{T}^*$ contained in $V^{-1}(c) \cap \{(\xi, \gamma) \in \mathrm{Dom}(f) \cap \mathrm{Dom}(V) : \ \nabla V_\gamma(\xi) f_\gamma(\xi) = 0\}$.*

When Γ is a finite set, we can relax the nonincreasing condition in Assumption 1 as follows.

Assumption 2 *The forward complete trajectory (x, σ) of (1) verifies the following: there exists a function $V \in \mathcal{V}$ such that (x, σ) is precompact relative to $\mathcal{O}$ and $v(t) = V(x(t), \sigma(t))$ is nonincreasing on $\mathcal{I}_{\sigma,\gamma} = \sigma^{-1}(\gamma) \cap [0, +\infty)$, for all $\gamma \in \Gamma$.*

In what follows, when Γ is a finite set, we identify it with the set $\{1, \ldots, N\} \subset \mathbb{N}$, where $N = \mathrm{card}(\Gamma)$.

Theorem 2 *Suppose that Γ is finite and let $\mathcal{S}^*$ and $\mathcal{T}^*$ be as in Theorem 1. Suppose that (x, σ), with $\sigma \in \mathcal{S}$, is a forward complete trajectory of (1) which has property* **L** *with respect to $\mathcal{S}^*$ and for which Assumption 2 holds. Then there exists $\vec{c} = (c_1, \ldots, c_N) \in \mathbb{R}^N$ such that x converges to a connected component of $\pi_1(M(\vec{c}))$, where $M(\vec{c})$ is the maximal weakly-invariant set w.r.t. $\mathcal{T}^*$ contained in $\cup_{\gamma \in \Gamma} \{(\xi, \gamma) \in \mathrm{Dom}(f) \cap \mathrm{Dom}(V) : \ V_\gamma(\xi) = c_\gamma \ \wedge \ \nabla V_\gamma(\xi) f_\gamma(\xi) = 0\}$.*

The following result is an integral-invariance principle which is an extension to switched systems of one of Byrnes and Martin for differential equations ([2]).

Theorem 3 *Let $\mathcal{S}^*$ and $\mathcal{T}^*$ be as in Theorem 1. Suppose that (x, σ), with $\sigma \in \mathcal{S}$, is a forward complete trajectory of (1) such that σ has property* **L** *with respect to $\mathcal{S}^*$. Suppose in addition that there exists a lower semi-continuous function $h : \mathrm{Dom}(h) \to [0, +\infty]$, with $\mathrm{Dom}(h) \subset \mathbb{R}^n \times \Gamma$, such that $(x(t), \sigma(t))$ evolves in a compact subset K of $\mathrm{Dom}(h)$, and that for some $\tau > 0$ $\int_t^{t+\tau} h(x(s), \sigma(s))ds$ converges to 0 as $t \to +\infty$. Then x converges to a connected component of $\pi_1(M^*)$, where M^* is the maximal weakly-invariant set w.r.t. $\mathcal{T}^*$ contained in $h^{-1}(0) \cap \mathrm{dom}(f)$.*

In the case in which σ of Theorem 3 is as in 1. of Lemma 1, the hypotheses can be weakened by considering *weakly meagre functions*. We recall that a function $y : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} = [-\infty, +\infty]$, is weakly meagre if $\lim_{k\to\infty}(\inf\{|y(t)| : t \in I_k\}) = 0$ for every family $\{I_k : \ k \in \mathbb{N}\}$ of nonempty and pairwise disjoint intervals in $\mathbb{R}_{\geq 0}$ with $\inf\{\mu(I_k) : \ k \in \mathbb{N}\} > 0$.

Theorem 4 *Let $\mathcal{S}^*$ and $\mathcal{T}^*$ be as in Theorem 1. Suppose that (x, σ), with $\sigma \in \mathcal{S}$ as in 1. of Lemma 1, is a forward complete trajectory of (1) and that σ has property* **L** *with respect to $\mathcal{S}^*$. Suppose in addition that there exists a lower semi-continuous function $h : \mathrm{Dom}(h) \to [0, +\infty]$, with $\mathrm{Dom}(h) \subset \mathbb{R}^n \times \Gamma$, such that $(x(t), \sigma(t))$ evolves in a compact subset K of $\mathrm{Dom}(h)$ for all $t \geq 0$ and that $y(\cdot) = h(x(\cdot), \sigma(\cdot))$ is weakly meagre. Then x converges to a connected component of $\pi_1(M^*)$, where M^* is the maximal weakly-invariant set w.r.t. $\mathcal{T}^*$ contained in $h^{-1}(0) \cap \mathrm{dom}(f)$.*

4 Conclusions

In this paper we have obtained some invariance results for bounded trajectories of switched systems whose switchings verify a certain property with respect to a certain subclass of switching signals. These results enable us to study, in an unified way, properties of bounded trajectories of switched systems subjected to perturbations and/or restrictions in the switching originated not only by its timing, but also by state-dependent constraints, and by accessibility constraints from each subsystem to other ones.

References

[1] A. Bacciotti and L. Mazzi, *An invariance principle for nonlinear switched systems*, Systems Contr. Lett., 54 (2005), pp.1109–1119.

[2] C. I. Byrnes and C. F. Martin, *An integral-invariance principle for nonlinear systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 983-994.

[3] D. Cheng, J. Wang and X. Hu, *An extension of LaSalle's invariance principle and its applications to multi-agent consensus*, IEEE Trans. Automat. Control, 53 (2008), pp. 1765–1770.

[4] R.A. Decarlo, M.S. Branicky, S. Pettersson and B. Lennartson, *Perspectives and results on the stability and stabilizability of hybrid systems*, Proc. IEEE 88 (2000), pp.1069–1082.

[5] R. Goebel, R.G. Sanfelice and A. Teel, *Invariance principles for switching systems via hybrid systems techniques*, Systems Contr. Lett., 57 (2008), pp.980–986.

[6] J.P Hespanha, *Uniform stability properties of switched linear systems: extensions of LaSalle's invariance principle*, IEEE Trans. Automat.Control, 49 (2004), pp.470–482.

[7] J.P. Hespanha, D. Liberzon, D. Angeli and E. Sontag, *Nonlinear observability notions and stability of switched systems*, IEEE Trans. Automat. Control, 50 (2005), pp.154–168.

[8] J.P. LaSalle, *Some extensions of Lyapunov's second method*, IRE Trans. Circuit Theory, 7 (1960), pp. 520–527.

[9] T.C. Lee and Z.P. Jiang, *Uniform asymptotic stability of nonlinear switched systems with an application to mobile robots*, IEEE Trans. Automat. Contr., 53 (2008), pp.1235–1252.

[10] J.L. Mancilla Aguilar and R.A. García, *An extension of LaSalle's invariance principle for switched systems*, Systems Contr. Lett., 55 (2006), pp.376–384.

[11] J. Wang and D. Cheng, *Stability of switched nonlinear systems via extensions of LaSalle's invariance principle*, Sci. China, Ser. F: Information Sciences, 52 (2009), pp. 84–90

[12] J. Wang, D. Cheng and X. Hu, *An extension of LaSalle's Invariance Principle for a class of switched linear systems*, Systems Control Lett., 58 (2009), pp. 754–758.

EVALUACIÓN DE UN PROCEDIMIENTO PARA LA IDENTIFICACIÓN DE PARÁMETROS ESTRUCTURALES DE SISTEMAS DINÁMICOS

Juan F. Giro[1,2], José E. Stuardi[1] y Ariel E. Matusevich[1]

[1]*Departamento de Estructuras, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de Correo 916, 5000 Córdoba, Argentina, juanfgiro@gmail.com, http://www.efn.uncor.edu*

[2]*Departamento de Ingeniería de Sistemas de Información, Facultad Regional Córdoba, Universidad Tecnológica Nacional, Maestro M. López esq. Cruz Roja Argentina, C.P. (X5016ZAA) Córdoba, Argentina*

Resumen*:* En este trabajo se continúa la evaluación un procedimiento, propuesto por los mismos autores, para determinar parámetros de sistemas estructurales lineales a partir de registros de respuesta en el dominio del tiempo. El objetivo final es desarrollar una herramienta efectiva para perfeccionar modelos estructurales y también para identificar daños y seguir su posterior evolución. Los parámetros identificados son los elementos de las matrices de rigidez y de amortiguamiento. Para ello es necesario conocer la matriz de masas y se utilizan registros de desplazamientos de la estructura bajo ciertas condiciones de excitacion. Este procedimiento está formulado de manera sistemática a través del álgebra matricial, lo que lo independiza de la complejidad o dimensión del sistema estructural estudiado. Se analizan las condiciones que deben cumplir las señales de excitación de la estructura a fin de que el procedimiento no presente problemas numericos y el sistema sea identificable. Se presentan y comentan resultados obtenidos.

Palabras claves: *respuesta dinámica de estructuras, identificación de parámetros*

1. INTRODUCCIÓN

El objeto de este trabajo es continuar la evaluación de un procedimiento destinado a la identificación de parámetros estructurales de sistemas dinámicos lineales presentado con anterioridad [1] [2]. Los parámetros que se identifican a través de este procedimiento son los elementos de las matrices de rigidez y de amortiguamiento, que se determinan a partir de registros de respuesta a la acción de cargas conocidas que varían en el dominio del tiempo. Para ello es necesario conocer previamente los elementos de la matriz de masas.

Los recientes desarrollos en tecnología computacional, en áreas tales como adquisición de datos, tratamiento de señales y procesamiento masivamente paralelo, han contribuido a estimular el interés por la identificación de parámetros a partir de la respuesta dinámica de sistemas mecánicos, conduciendo al desarrollo de nuevos procedimientos y al perfeccionamiento de otros existentes [3].

Sin embargo, la identificación directa de parámetros estructurales a partir de mediciones experimentales ofrece inconvenientes, tales como: *i*) La dificultad, muchas veces imposibilidad, de excitar la estructura en forma apropiada para la medición de los valores buscados, *ii*) Los errores inherentes a las mediciones, que se propagan a través de los procesos numéricos con impacto incierto en los resultados y *iii*) Las condiciones requeridas por el proceso de identificación para asegurar la no singularidad y buen condicionamiento de los sistemas de ecuaciones involucradas. En muchos casos estos problemas restringen la potencialidad de las técnicas de identificación, desalientan su utilización y/o los limitan al tratamiento de casos simples, de menor interés práctico.

Para superar la primera de las dificultades señaladas, es decir poder excitar la estructura adecuadamente, en [1] se propuso construir un modelo de la estructura en estudio utilizando redes neuronales artificiales, de tipo multicapa de perceptrones. Luego, en una segunda etapa se utiliza el modelo neuronal para evaluar numéricamente la respuesta del sistema a partir de condiciones iniciales y en ausencia de cargas exteriores.

El segundo inconveniente, el impacto de los errores de las mediciones en la evaluación de los parámetros estructurales, fue objeto del trabajo [2]. También a partir de condiciones iniciales se estudió la relación entre los niveles de error en los datos disponibles y la calidad de los parámetros estructurales obtenidos, con diversos modelos neuronales y sistemas mecánicos progresivamente más complejos. Así se analizó la robustez del procedimiento y su aptitud para operar correctamente, aun a partir de registros de entradas ruidosos.

El análisis de la sensibilidad del proceso de identificación a las condiciones con que son obtenidos los registros de datos, que es la última de las dificultades señaladas, dio lugar al estudio que es presentado en este trabajo. La condición que deben cumplir las cargas a fin de que un sistema sea identificable está relacionado con el concepto de excitación persistente, que establece que deben tener un espectro de frecuencias suficientemente amplio, de manera de perturbar al sistema en todos sus modos naturales de vibración.

2. PLANTEO DEL PROBLEMA

Considerando el caso de un sistema elástico lineal de n grados de libertad, su equilibrio dinámico queda representado por un sistema de ecuaciones diferenciales que tienen la siguiente forma general:

$$M\,\ddot{y} + C\,\dot{y} + K\,y = u \tag{1}$$

donde M, C y K representan las matrices de inercia, amortiguamiento y rigidez respectivamente.

Utilizando condiciones de excitación que aseguren la presencia de sus principales modos de vibración en la respuesta, se obtiene el vector respuesta (y) en sucesivos intervalos de tiempo Δt. A partir de estas respuesta, y utilizando formulas de diferencias finitas de orden elevado, se calculan los vectores de velocidad ($\dot{y}$) y de aceleración ($\ddot{y}$), todos ellos de dimensión n. Estos desplazamientos y velocidades son agrupados en un vector de estado y su derivada, ambos de dimensión $2n$, denominados z y $\dot{z}$:

$$z = \begin{Bmatrix} \dot{y} \\ y \end{Bmatrix} \quad , \quad \dot{z} = \begin{Bmatrix} \ddot{y} \\ \dot{y} \end{Bmatrix} \tag{2}$$

Con los m vectores z y $\dot{z}$ correspondientes a sucesivos instantes de la respuesta del sistema se definen las columnas de las matrices Z y $\dot{Z}$, ambas de dimensión $2n \times m$. Esto es:

$$Z = \left[z_1\, z_2\, z_3 \ldots z_m \right] \quad , \quad \dot{Z} = \left[\dot{z}_1\, \dot{z}_2\, \dot{z}_3 \ldots \dot{z}_m \right] \tag{3}$$

Ordenando en forma similar a los sucesivos m vectores de cargas exteriores en una matriz U, de dimensión $n \times m$, se obtiene:

$$U = \left[u_1\, u_2\, u_3 \ldots u_m \right] \tag{4}$$

Reordenando la ecuación (1) en función del vector de estado z y su derivada, y considerando la respuesta del sistema en los m instantes de tiempo, se tiene:

$$\dot{Z} = F\,Z + G\,U \qquad \text{donde} \qquad F = \begin{bmatrix} -M^{-1}C & -M^{-1}K \\ I & 0 \end{bmatrix} \quad \text{y} \quad G = \begin{bmatrix} -M^{-1} \\ 0 \end{bmatrix} \tag{5}$$

Despejando ahora la matriz F en función de la matriz Z, su derivada $\dot{Z}$ y la matriz cargas U:

$$F^T = \begin{bmatrix} F_{11}^T & I \\ F_{12}^T & 0 \end{bmatrix} = (Z^T)^+ \left[\dot{Z}^T - U^T G^T \right] \qquad \text{donde} \quad F_{11} = -M^{-1}C \quad \text{y} \quad F_{12} = -M^{-1}K \tag{6}$$

donde "+" denota la matriz seudo inversa de Moore-Penrose. Nótese que resulta conveniente transponer la expresión anterior ya que siempre m >> n y el cálculo de la seudo inversa incluye así la inversión de una matriz de orden n. Luego, considerando que es conocida la matriz de masa M, habitualmente diagonal, se determinan las matrices de rigidez $\overline{K}$ y amortiguamiento $\overline{C}$ a partir de las ecuaciones (6):

$$\overline{K} = -M\,F_{12} \qquad \overline{C} = -M\,F_{11} \tag{7}$$

La utilización de la seudo inversa o inversa de Moore-Penrose de una matriz implica la optimización implícita de un problema sobredeterminado, es decir que se dispone de más ecuaciones que incógnitas y cuya solución corresponde a un planteo de mínimos cuadrados. Nótese que este procedimiento está supeditado a que la seudo inversa de la matriz Z exista, lo que implica que debe ser de rango completo.

La determinación de la submatriz F_{12} abre las puertas al cálculo de las frecuencias y modos normales de vibración del sistema dinámico. En efecto, volviendo a la ecuación (1), omitiendo las fuerzas de amortiguamiento y suponiendo una respuesta armónica, queda planteado el clásico problema de autovalores, donde $\lambda = \omega^2$:

$$(M^{-1}K - \lambda I)\,\overline{y} = 0 \qquad \Rightarrow \qquad (-F_{12} - \omega^2 I)\,\overline{y} = 0 \tag{8}$$

Todo el planteo anterior esta supeditado al calculo de la matriz F, que a su vez depende no solo de que Z sea de rango completo sino también de que este bien condicionada. Para responder a esta exigencia habría que evaluar anticipadamente indicadores de consistencia temporal y espacial, tales como los propuestos por Enecio [4], pero de la forma en que el problema esta planteado se carece de información suficiente para hacerlo. Por otra parte, mas que un diagnostico sobre el buen condicionamiento de Z, lo que se busca aquí es asegurar esta condición a través de una apropiada excitación de la estructura. Como ya fue anticipado, en este trabajo se buscaron recomendaciones que conduzcan a la perturbación del sistema en forma apropiada.

3. CASO DE ESTUDIO Y RESULTADOS OBTENIDOS

Para alcanzar el objetivo propuesto se considera un caso simple que ya fue utilizado y descrito en detalle en [2]. Se trata de un sistema elástico lineal de tres grados de libertad, con masas concentradas y amortiguamiento viscoso proporcional (o de Rayleigh). Para el amortiguamiento se adopta la forma particular $C = \beta K$ con valores de β en el rango $0{,}08 \le \beta \le 0{,}40$. Como referencia, para el valor $\beta = 0{,}1592$ los factores de amortiguamiento ξ_i correspondientes a los tres modos de vibración del sistema son $\xi_1 = 0{,}0153$; $\xi_2 = 0{,}0282$ y $\xi_3 = 0{,}0369$.

Las respuestas del sistema a las diferentes condiciones de excitación fueron obtenidas a través de integración numérica. Esto debido a la necesidad de: *i*) Disponer de respuestas precisas que permitieran cuantificar los errores cometidos y *ii*) Poder imponer variadas condiciones de excitación, aun aquellas de difícil implementación en sistemas reales. Las respuestas fueron obtenidas sobre 1000 puntos en un intervalo de 20 seg, es decir con un incremento $\Delta t = 0{,}02$ seg. Posteriormente, los vectores velocidad y aceleración fueron determinados con formulas regresivas de derivación numérica.

Para determinar los errores cometidos en la evaluación de las matrices de rigidez ($\bar{K}$) y amortiguamiento ($\bar{C}$) se calcularon las diferencias medias cuadráticas de los elementos de cada matriz divididas por el valor máximo en cada una. Se obtuvieron así los indicadores de error $E_K^{\%}$ y $E_C^{\%}$, donde K y C representan los valores conocidos de la rigidez y amortiguamiento del sistema estudiado.

$$E_K^{\%} = \frac{100\, E_K}{\text{mayor}\left\{ \left| K_{ik} \right| \right\}} \qquad \text{donde} \quad E_K = \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} (K_{ik} - \bar{K}_{ik})^2} \tag{9}$$

$$E_C^{\%} = \frac{100 E_C}{\text{mayor}\left\{ \left| C_{ik} \right| \right\}} \qquad \text{donde} \quad E_C = \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} (C_{ik} - \bar{C}_{ik})^2} \tag{10}$$

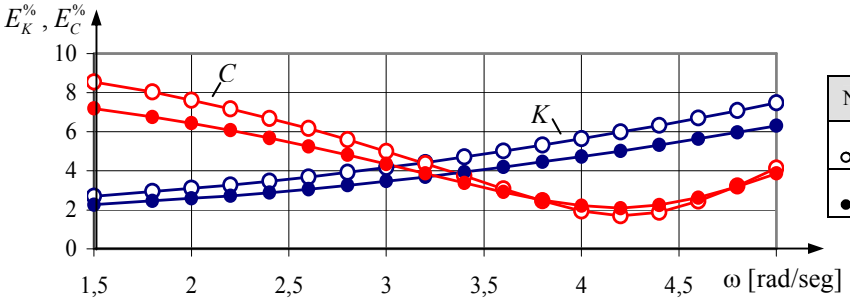
Para evaluar la respuesta del sistema a ciertas condiciones externas de excitación se estudiaron tres casos que contemplaron: *i*) carga periódica de frecuencia fija, *ii*) carga periódica de frecuencia variable en el intervalo estudiado, barriendo un rango de frecuencias que incluye las frecuencias de sus modos naturales de vibración y *iii*) carga constantes aplicada súbitamente y en forma gradual (escalón y rampas). En cada caso se obtuvieron resultados con una carga sobre uno de los grados de libertad y con dos cargas aplicadas en fase, donde la segunda carga tiene una magnitud del 20% de la primera.

Con las cargas periódicas de frecuencia fija se obtuvieron resultados que son representados en la Figura 1, donde el eje de las abscisas corresponde a la frecuencia de excitación. Se muestra que los mejores resultados para K se obtienen con la frecuencia de excitación más baja, bajo la acción de dos cargas, y los mejores resultados para C corresponden a la acción de una sola carga con una frecuencia próxima a la del tercer modo de vibración. Los mejores resultados obtenidos tuvieron errores $E_K^{\%} = 2{,}62$ % y $E_C^{\%} = 1{,}68$ %.

En la Figura 2 se presentan resultados obtenidos con cargas periódicas de frecuencia variable, donde la abscisa representa el incremento de la variación de frecuencia. Aquí puede observarse que los valores de K y C muestran tendencias opuestas; los mejores resultados de una corresponden a condiciones en que se obtienen los peores de la otra. Los mejores valores de K se obtienen con un incremento $\Delta\omega = 0{,}7$ rad/seg^2 y los de C con un incremento $\Delta\omega = 2{,}1$ rad/seg^2. También es conveniente usar dos cargas para determinar K y una sola carga para determinar C, y los mejores resultados obtenidos tuvieron errores $E_K^{\%} = 1{,}29$ % y $E_C^{\%} = 2{,}17$ %.

Por último, en la Figura 3 se presentan los errores de la rigidez y amortiguamiento obtenidos a partir de cargas constantes que son aplicadas gradualmente, donde el eje de las abscisas muestra la duración "t_e" de la rampa de carga creciente. Con la carga en escalón ($t_e = 0$) se obtuvieron los mejores resultados, tanto para K como para C. Las rampas brindan también valores muy buenos de K con duraciones de 1,6 y 3,2 seg, es decir la mitad del periodo del modo fundamental de vibración, mientras el error en el amortiguamiento también es oscilante pero con valores muy superiores al de la carga escalón. Los mejores resultados obtenidos tuvieron errores $E_K^{\%} = 0{,}05$ % y $E_C^{\%} = 0{,}11$ %, mientras que aquí la diferencia entre usar una o dos cargas es mínima.

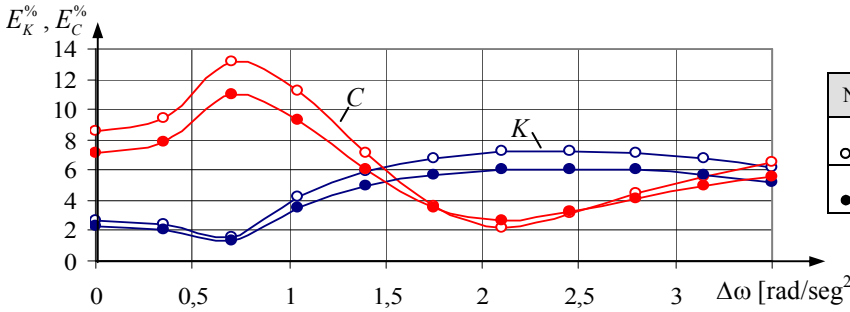
En resumen, al asegurar la carga en escalón la necesaria presencia de todos los modos de vibración en la respuesta del sistema, se descarta toda posibilidad de problema numérico y se explica la obtención de los mejores valores de K y C. En el caso de utilizarse cargas periódicas habrá que evaluar K y C con frecuencias iguales a las de los modos de vibración extremos, sin descartarse el riesgo de mal condicionamiento en el sistema de ecuaciones resultante. Por ultimo, las cargas periódicas de frecuencia variable no demostraron ventajas significativas.



Errores mínimos

NC	$E_K^{\%}$	$E_C^{\%}$
1	2,70	1,68
2	2,26	2,09

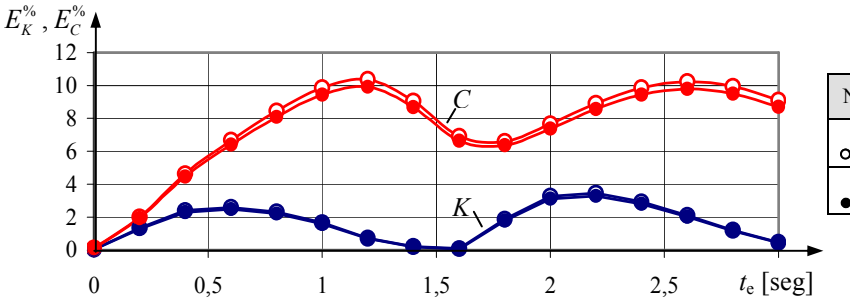
Figura 1: Errores cometidos con cargas periódicas de frecuencia constante y sus valores mínimos



Errores mínimos

NC	$E_K^{\%}$	$E_C^{\%}$
1	1,55	2,17
2	1,29	2,68

Figura 2: Errores cometidos con cargas periódicas de frecuencia variable y sus valores mínimos



Errores mínimos

NC	$E_K^{\%}$	$E_C^{\%}$
1	0,05	0,12
2	0,05	0,11

Figura 3: Errores cometidos con cargas constante aplicada gradualmente (rampa) y sus valores mínimos

4. CONCLUSIONES

En este trabajo se continuó evaluando un procedimiento para obtener los parámetros característicos de sistemas mecánicos lineales a partir su matriz de masa y los registros de su respuesta en el dominio del tiempo. Los parámetros evaluados son los elementos de las matrices de rigidez y amortiguamiento. Las experiencias realizadas confirmaron la posibilidad de obtener resultados de muy buena calidad y que las cargas en escalón son la condición de excitación más conveniente. Se continuará trabajando con casos de mayor dimensión y complejidad, para luego reemplazar los registros de respuesta obtenidos de simulación numérica por valores experimentales. El objetivo final es la predicción de daños estructurales en sistemas mecánicos reales.

REFERENCIAS

[1] J. GIRÓ, A. GARCÍA Y J. STUARDI, *Identificación de parámetros de sistemas dinámicos a través de redes neuronales artificiales*, Mecánica Computacional, 26 (2007), pp. 2585-2599.

[2] J. GIRÓ, A. GARCÍA Y J. STUARDI, *Sensibilidad de modelos neuronales usados para evaluar propiedades dinámicas de estructuras a partir de mediciones de su respuesta en el tiempo*, *Mecánica Computacional*, 27 (2008), pp. 1983-1997.

[3] K. ALVIN, A. ROBERSTON, G. REICH AND K. PARK, *Structural system identification: from reality to models*, Computers and Structures, 81 (2003), pp. 1149-1176.

[4] G. ENECIO AND M. ABE, *Structural Damage Detection for bolted connection between two steel plates using laser doppler vibrometry*, Research Report, University of San Carlos (2002), pp. 162-193.

MODELLING OF DYNAMICAL SYSTEMS WITH PERIODIC ORBITS USING CONTINUOUS PIECEWISE LINEAR APPROXIMATIONS

Andrés G. García and Osvaldo E. Agamennoni

Universidad Nacional del Sur, Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca, Argentina,
agarcia@uns.edu.ar,oagamen@uns.edu.ar

Abstract: This paper presents formal conditions for the existence of a continuous piecewise linear (CPWL) approximation of nonlinear ODE's possessing periodic orbits. The main theorem ensures the existence of a grid size of a CPWL ODE which captures the same periodic pattern as the nonlinear model under analysis.

An example of application is considered using a passive walking robot dynamic model. This kind of examples exhibiting periodic motion are the key to understand deeper concepts for a more advanced biped walking robots. Finally, some conclusions are depicted as well.

Keywords: *first, second, third*
2000 AMS Subject Classification: 34C05, 34C07, 37C27

1 INTRODUCTION

The analysis and simulation of dynamical systems modelled with ordinary differential equations (ODE's) is the heart of any scientific research with real applications. However, while the study of stable or attractive attractors could be approximated with satisfactory results using numerical simulators, the presence of oscillations like limit cycles, homoclinic, heteroclinic, chaos, etc (see [3] and [1] for further details), can not always be approximated with enough accuracy by numeric methods.

Many physical systems of interest in real applications exhibits oscillations which, moreover, are the essence of the dynamic of the system. For instance the problem of stable walking of walking robots (see [7] and [13]) is in fact a very rich dynamics which still have to be distilled in order to achieve satisfactory numerical models, also the modelling of internal combustion motors requires more and more efficient and precise numerical predictions in order to optimize the performance as much as possible (see [16]).

As the problem in $\Re^n$ is difficult in general, the study of planar dynamical systems ($\Re^2$) has always been a source of inspiration for system of bigger order and also because they model many physical phenomena (see for instance [8], [2] and [1]).

In this paper a generalization of previous works [4], [14] and [15] applicable to ODE's with periodic orbits will be developed. The approach introduces continuous piecewise linear (CPWL) approximating vector fields which captures the periodic orbits of the given nonlinear ODE. This kind of methodologies where pioneered applied in [9] and [10], moreover, using this approach Tonnelier derived a condition for the number of limit cycles in CPWL planar ODE's of the Liénard type (see [11]).

A clear lack in the existent literature, is an existence result: given an ODE with periodic orbits, there exists a CPWL approximating vector field with the same number and location of periodic orbits?. This paper answer this question in the affirmative: if it is known that a certain ODE posses periodic dynamic (like walking robots, combustion motors, etc), then the existence of a CPWL approximating ODE with periodic orbits is proved. This result allow to ensure that for certain degree of approximation, the periodic behavior is captured by the CPWL ODE, it turns out that systems like walking robots where the intermittent contact with the ground produces a piecewise holonomic model it is better formulated using CPWL vector fields.

This paper is organized as follows: Section 2 presents a formal introduction to the problem considered, Section 2.1 recalls some definitions in CPWL theory, Section 3 present the main theorem in this paper, where the Section 4 shows an example using the model of a walking robot. Finally, some conclusions are depicted in Section 5.

2 THE PROBLEM CONSIDERED

Given an ODE: $\dot{x}(t) = f(x)$ with a given initial condition: $x(0) = x_0$, to provide a CPWL vector field, so once the number of simplices (subdivision of the state space) and the size of each simplex (grid size) is settled, using any of the available tools to perform this approximation (for instance [12]), then a CPWL vector field:

$A^{(i)} \cdot x(t) + B^{(i)}$ approximating a given nonlinear one: $f(x)$ in the i^{th} simplex, it is obtained:

$$\left| f(x) - A^{(i)} \cdot x(t) + B^{(i)} \right| \leq \lambda$$

for some $\lambda \in \Re^{+}$. Notice that it is immediate to suggest that the study of the CPWL approximating ODE: $\dot{x}(t) = A^{(i)} \cdot x(t) + B^{(i)}$ possesses the same qualitative properties of the given nonlinear one for some small enough λ, however to prove this assertion it is not so immediate.

These error bounds indicates that once the grid size tend to zero, then the trajectories of the CPWL approximating ODE and the real non-linear one converges each other, however this is not telling anything about the existence of a finite grid size to achieve this same behavior.

2.1 PRELIMINARY DEFINITIONS

The following definitions are in the core of piecewise linear functions (see [12]):

Definition 1 (Continuous Piecewise Linear Functions (CPWL)) *Let a domain $\mathbb{D}$ be partitioned into a set of convex polyhedrons called regions $R^{(i)}, \quad i = 1, \ldots, r$, such that $\mathbb{D}$ is the union of the compact closures of each region $R^{(i)}$: $\bar{R}^{(i)}$ by a set:*

$$H = \{H_i \subset \mathbb{D}, \quad i = 1, \ldots, h\}$$

of a finite number of $n-1$-dimensional hyperplanes, also called boundaries or borders:

$$H_i = \{x : \pi_i(x) = \alpha_i' \cdot x + \beta_i = 0\}$$

where $i = 1, \ldots, h$, $\alpha_i \in \Re^n, \quad \beta_i \in \Re$. Then, a piecewise linear function (PWL) is defined by the local (linear) functions:

$$f^{(i)}(x) = J^{(i)} \cdot x + \omega^{(i)}$$

where $J^{(i)} \in \Re^{1 \times n}, \quad \omega^{(i)} \in \Re$. Moreover if $J^{(p)} \cdot z + \omega^{(p)} = J^{(q)} \cdot z + \omega^{(q)}, \quad \forall z \in \bar{R}^{(p)} \cap \bar{R}^{(q)}$, where $\bar{R}^{(p)} \cap \bar{R}^{(q)} \neq \emptyset$ for every possible pair of contiguous regions $R^{(p)}, R^{(q)} \subset \mathbb{D}$, then the PWL function is continuous (CPWL).

Definition 2 (Simplex) *Let $\{x_0, x_1, \ldots, x_n\}$ be $n+1$ points in a n-dimensional space. A simplex $\Delta(x_0, x_1, \ldots, x_n)$ is defined by:*

$$\Delta(x_0, x_1, \ldots, x_n) = \{x : x = \sum_{i=0}^{n} \mu_i \cdot x_i\}$$

where $\mu_i \in [0, 1], \quad i = 1, \ldots, n$ and $\sum_{i=0}^{n} \mu_i = 1$.

3 MAIN RESULT

In this section, the main result of the paper is presented:

Theorem 1 *Given a nonlinear ODE: $\dot{x}(t) = f(x)$ with periodic orbits and with $f(x)$ Lipschitz, there exists a CPWL approximating vector field of $f(x)$ with error bounds $\lambda \in \Re^{+n}$ such that the CPWL ODE: $\dot{x^{(k)}} = A^{(k)} \cdot x^{(k)} + B^{(k)}$ also possesses periodic orbits.*

Proof. Let's construct a CPWL vector field: $(A^{(i)} \cdot x(t) + B^{(i)})$ as an approximation of the given ODE: $\dot{x}(t) = f(x)$ with static error λ and define a new ODE: $\dot{x}^{(k)}(t) = A^{(k)} \cdot x(t) + B^{(k)}$:

$$\begin{cases} \left| f(x) - (A^{(i)} \cdot x(t) + B^{(i)}) \right| \leq \lambda \\ \dot{x}(t) = f(x) \\ \dot{x}^{(k)}(t) = A^{(k)} \cdot x(t) + B^{(k)} \end{cases}$$

where $A^{(i)} \cdot x(t) + B^{(i)}$ and $A^{(k)} \cdot x(t) + B^{(k)}$ indicates the CPWL vector field in the simplex i^{th} and k^{th} respectively. This takes into account the possibility for the ODE $\dot{x}(t) = f(x)$ (dictating the location of the simplex for $A^{(i)} \cdot x(t) + B^{(i)}$) with their states in the simplex i^{th} and the ODE $\dot{x}^{(k)}(t) = A^{(k)} \cdot x(t) + B^{(k)}$ running in the simplex k^{th}.

Then, defining the error as: $E^{(k)} = x(t) - x^{(k)}$, the procedure depicted in [14], [15] and the thesis [5] yields the following bounds:

$$\begin{cases} min\{E^{1*}, E^{2*}\} \leq E^{(k)} \leq max\{E^{1*}, E^{2*}\} \\ \dot{E^{1*}}(t) = A^{(i)} \cdot E^{1*}(t) - \left[\left(A^{(k)} - A^{(i)} \right) \cdot x^{(k)} + \left(B^{(k)} - B^{(i)} \right) \right] + \lambda \\ \dot{E^{2*}}(t) = A^{(i)} \cdot E^{2*}(t) - \left[\left(A^{(k)} - A^{(i)} \right) \cdot x^{(k)} + \left(B^{(k)} - B^{(i)} \right) \right] - \lambda \end{cases} \tag{1}$$

Defining two new variables $z_1(t) = E^{1*}(t) + x^{(k)}$ and $z_2(t) = E^{2*}(t) + x^{(k)}$, their dynamics are:

$$\dot{z_1}(t) = A^{(i)} \cdot z_1(t) + B^{(i)} + \lambda, \quad \dot{z_2}(t) = A^{(i)} \cdot z_1(t) + B^{(i)} - \lambda$$

Finally, if the trajectories $x(t)$ are periodic, then both: $\{i^{th}, A^{(i)}\}$ are periodic sequences. Also the difference: $\dot{z_1}(t) - \dot{z_2}(t) = A^{(i)} \cdot (z_1(t) - z_2(t))$ is a time-varying linear ODE with periodic coefficients, so their solutions are periodic. Then, the difference: $E^{1*}(t) - E^{2*}(t)$ is periodic, thus showing that both $\{E^{1*}, E^{2*}\}$ are periodic. To complete the proof, let's note that $\{A^{(i)}, B^{(i)}, z_1, z_2, E^{1*}, E^{2*}\}$ are periodic, then using the definition (1): $\dot{x^{(k)}} = -\dot{E^{1,2*}}(t) + A^{(i)} \cdot E^{1,2*}(t) + B^{(i)} \pm \lambda$ is periodic and this completes the proof. □

4 Example of Application

The following system corresponds to a simple walking robot (see Figure 1a), taken from [6]:

$$\begin{cases} \dot{x_1}(t) = x_2(t), \quad \dot{x_2}(t) = sin(x_1(t) - \gamma) \\ \dot{x_3}(t) = x_4(t), \quad \dot{x_4}(t) = sin(x_1(t) - \gamma) + x_2(t)^2 \cdot sin(x_3(t)) - cos(x_1(t) - \gamma) \cdot sin(x_3(t)) \end{cases}$$

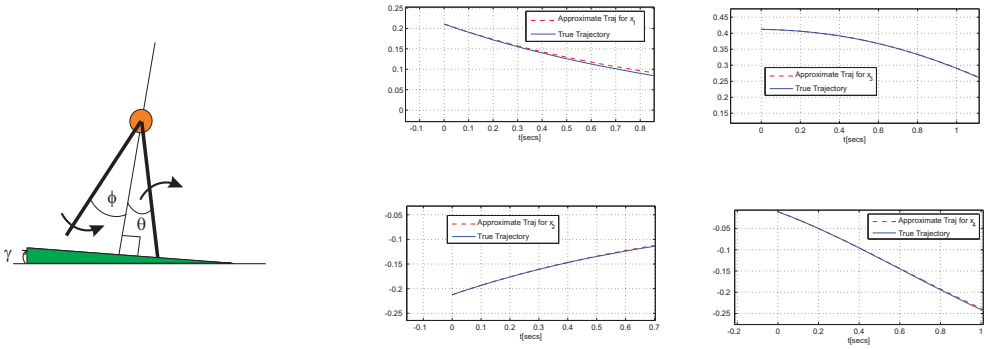
where γ is the ramp slope since this a walking robot without motors, just moved by the gravity force. For the simulation $\gamma = 0.009$ was used along with 20 simplices for every coordinate and only 200 points for the CPWL vector field.

The simulation results are shown in Figure 1b, it is clear that the accuracy is high enough (less than 0.12 percent) to capture correctly the dynamics before the contact change of the legs of the robot. It is worth to mention that a toolbox in Matlab was created to simulate nonlinear ODE's using CPWL approximations.

5 Conclusion

An existence result of periodic orbits of ODE's using CPWL approximations, was proved. The result is forward exploitable to those systems with known periodic behavior. This is of interest in real applications where periodic trajectories are the core of the motion (walking robots, combustion motors, etc).

Numerical results for a simplified model of a walking robot was considered, the high precision of the results shows the potentiality of using CPWL approximations as a numerical and efficient tool in the view of the linearity of each simplex. In this sense it is worth to mention that a Matlab code was prepared to simulate nonlinear ODE's allowing the user to specify the number of simplices, number of points to get the CPWL approximations, initial conditions and even the possibility to enforce the equilibrium of the CPWL vector field so as to coincide with the equilibriums of the nonlinear ODE under analysis.



(a) (b)

Figure 1: Simplified model of a walking robot and simulation results.

ACKNOWLEDGMENTS

This work was completed with the complete support of Universidad Nacional del Sur under the PROMEI program, CIC and ANPCyT.

6 REFERENCES

REFERENCES

[1] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear oscillations in dynamical systems and bifurcations of vector fields (Applied mathematical sciences)*, Springer-Verlag, 1993.

[2] E. M. IZHIKEVICH, *Dynamical Systems in Neuroscience. The Geometry of Excitability and Bursting*, The MIT Press. Cambridge, Massachusetts. London, England, 2007.

[3] Y. A. KUZNETSOV, *Elements of Applied Bifurcation Theory*, Springer-Verlag, 1995.

[4] A. GARCÍA, S. BIAGIOLA, J. FIGUEROA, L. CASTRO AND O. AGAMENNONI, *Approximate Solutions for Non-Linear Autonomous ODE's on the Basis of PWL Approximation Theory*, World Scientific, 2006, pp. 113-126.

[5] A. GARCÍA, *Resolución de ecuaciones diferenciales ordinarias no lineales mediante funciones lineales a tramos : aplicaciones al control de sistemas*, Phd thesis, Universidad Nacional del Sur. Departamento de Ingeniera Eléctrica y Computadoras, 2009.

[6] M GARCIA, A. CHATTERJEE, A, RUINA AND M. COLEMAN, *TheSimplest Walkign Model: Stability, Complexity , and Scaling*, ASME Journal of Biomechanical, (1998), pp.1-15.

[7] J RUMMEL AND Y BLUM AND A SEYFARTH, *Robust and efficient walking with spring-like legs*, Bioinspiration and Biomimetics, Vol. 5 (2010), pp.1-13.

[8] R. FITZHUGH, *I*mpulses and physiological states in theoretical models of nerve membrane, Biophysical J, Vol. 1 (1961), pp. 445-466.

[9] J. LLIBRE AND E. PONCE, *P*iecewise-linear feedback systems with arbitrary number of limit cycles, International Journal of Bifurcation and Chaos, Vol. 2 (2003), pp. 2073-2097.

[10] G. TIGAN AND A. ASTOLFI, *A* note on a piecewise-linear Duffing-type system, International Journal of Bifurcation and Chaos, Vol. 12 (2007), pp. 4425-4429.

[11] A. TONNELIER, On the number of limit cycles in piecewise linear Liénard systems, Int. J. Bifurcation Chaos, Vol. 15 (2005), pp. 1417-1422.

[12] P. JULIÁN, A. DESAGES AND O. AGAMENNONI, High Level Canonical Piecewise Linear Representation Using a Simplicial Partition, IEEE Trans. Circuit and Systems-I: Fundamental Theory and Apllications, Vol. 46 (1999), pp. 463-480.

[13] JAE-SUNG MOON AND M. W. SPONG, *Bifurcation and Chaos in Passive Walking of a Compass-Gait Bid with Assymetries*, Proc. 2010 IEEE International Conference on Robotics and Automation Anchorage Convention District, Anchorage, Alaska, USA (2010).

[14] A. GARCÍA AND O. AGAMENNONI, *Continuous piecewise linear control for nonlinear systems: The parallel model technique*, Proc. Applied Computing Conference (2008).

[15] A. GARCÍA, O. AGAMENNONI AND J. FIGUEROA, *Applying continuous piecewise linear approximations to affine non-linear control systems*, Proc. 6th IFAC Symposium on Robust Control Design (2009).

[16] P. ORTNER, P. LANGTHALER, J. VICENTE GARCÍA ORTIZ AND L. DEL RE, *MPC for a Diesel Engine Air Path using an Explicit Approach for Constraint Systems*, Proc. IEEE International Conference on Control Applications, Munich, Germany (2006).

CONVERGENCIA DE CONTROLES ÓPTIMOS FRONTERA PARA INECUACIONES VARIACIONALES ELÍPTICAS

Mahdi Boukrouche[a], Claudia M. Gariboldi[b] y Domingo A. Tarzia[c]

[a]*PRES Lyon University, University of Saint-Etienne, Laboratory of Mathematics, LaMUSE EA-3989, 23 rue Paul Michelon, 42023 Saint-Etienne, France. E-mail: Mahdi.Boukrouche@univ-st-etienne.fr*
[b]*Departamento de Matemática, FCEFQyN, Uni. Nac. de Río Cuarto, Ruta 36 Km 601, 5800 Río Cuarto, Argentina. E-mail:cgariboldi@exa.unrc.edu.ar*
[c]*Departamento de Matemática-CONICET, FCE, Univ. Austral, Paraguay 1950, S2000FZF Rosario, Argentina. E-mail: DTarzia@austral.edu.ar*

Resumen: Se considera un sistema complementario S en un dominio acotado n-dimensional con condiciones de frontera mixtas y para cada $\alpha > 0$ se tiene otro sistema complementario S_α con diferente condición (tipo Robin) sobre una porción de la frontera del dominio. Se establece una estimación de la distancia entre la combinación convexa de las soluciones del sistema S, $u_3(t) = tu_{q_1} + (1-t)u_{q_2}$ para cada q_1, q_2 y la solución de la combinación convexa de los datos $u_4(t) = u_{tq_1+(1-t)q_2}$. Similarmente se prueba una análoga estimación vinculada al sistema S_α, para cada $\alpha > 0$. Para un funcional de costo cuadrático, se considera un problema de control óptimo frontera gobernado por una inecuación variacional elíptica en relación al sistema S y usando la propiedad de monotonía $u_4(t) \le u_3(t)\ \forall t \in [0,1]$ se obtiene la unicidad del control óptimo. La misma propiedad es probada para la familia de problemas de control óptimo frontera vinculados al sistema S_α, para cada $\alpha > 0$. Se demuestra además la convergencia, cuando $\alpha \to +\infty$, de los controles óptimos y de los estados asociados a esta familia de problemas de control frontera. Este resultado se obtiene sin el uso del estado adjunto del sistema lo cual es una ventaja respecto a la prueba dada en Gariboldi-Tarzia, Adv. in Diff. Eq. and Control Processes, 1 (2008), pp. 113-132, para sistemas gobernados por ecuaciones variacionales elípticas.

Palabras clave: *inecuaciones variacionales elípticas, control óptimo frontera, combinaciones convexas, convergencia de controles óptimos, problema de complementariedad.*
2000 AMS Subject Classification: 35R35, 35B37, 35J85, 49J20

1. INTRODUCCIÓN

Sea Ω un dominio abierto en $\mathbb{R}^n$ cuya frontera regular Γ consiste de la unión de dos porciones disjuntas Γ_1 y Γ_2. con $med(\Gamma_i) > 0$ para $i = 1, 2$. Se considera el siguiente problema de complementariedad:

$$u \ge 0, \quad u(-\Delta u - g) = 0, \quad -\Delta u - g \ge 0 \quad \text{en} \quad \Omega \tag{1}$$

$$u = b \quad \text{sobre} \quad \Gamma_1, \qquad -\frac{\partial u}{\partial n} = q \quad \text{sobre} \quad \Gamma_2 \tag{2}$$

y para el parámetro $\alpha > 0$, se considera el problema de complementariedad (1) con condiciones mixtas:

$$-\frac{\partial u}{\partial n} = \alpha(u - b) \quad \text{sobre} \quad \Gamma_1 \qquad -\frac{\partial u}{\partial n} = q \quad \text{sobre} \quad \Gamma_2 \tag{3}$$

donde α es el coeficiente de transferencia de calor sobre Γ_1, g es la energía interna, b es la temperatura sobre Γ_1 y q es el flujo de calor sobre Γ_2. Se definen los espacios y los conjuntos convexos siguientes:

$$V = H^1(\Omega), \qquad V_0 = \{v \in V : v_{|\Gamma_1} = 0\}$$

$$K = \{v \in V : v|_{\Gamma_1} = b,\ v \ge 0 \text{ en } \Omega\}, \qquad K_+ = \{v \in V : v \ge 0 \text{ en } \Omega\}.$$

Es conocido que, para una temperatura positiva $b \in H^{\frac{1}{2}}(\Gamma_1)$, $q \in Q = L^2(\Gamma_2)$ y $g \in H = L^2(\Omega)$, los problemas de frontera libre (1)-(2) y (1)-(3) son respectivamente equivalentes a los siguientes problemas variacionales elípticos:

$$\text{Hallar } u \in K \text{ tal que} \qquad a(u, v-u) \ge (g, v-u) - \int_{\Gamma_2} q(v-u)ds, \qquad \forall v \in K \tag{4}$$

Hallar $u \in K_+$ tal que $\quad a_\alpha(u\,,\,v-u) \geq (g\,,\,v-u) - \displaystyle\int_{\Gamma_2} q(v-u)ds + \alpha \int_{\Gamma_1} b(v-u)ds \quad \forall v \in K_+$ (5)

donde

$$a(u,v) = \int_\Omega \nabla u \nabla v dx, \qquad (g,v) = \int_\Omega g v dx, \qquad a_\alpha(u,v) = a(u,v) + \alpha \int_{\Gamma_1} uv ds.$$

Se conoce que a y a_α son formas bilineales, continuas, simétricas y coercivas sobre V_0 y V [16, 17].

Sean u_1 y u_2 dos soluciones de la inecuación variacional (4) con flujo q_1 y q_2 respectivamente, se define $\forall t \in [0,1]$

$$u_3(t) = (1-t)u_1 + tu_2 \quad ; \quad u_4(t) = u_{(1-t)q_1 + tq_2}$$

.

En [4], se estableció la condición necesaria y suficiente para obtener que la combinación convexa $u_3(t)$ sea la única solución de la inecuación variacional elíptica (4) con flujo $q_3(t) = tq_1 + (1-t)q_2$, esto es:

$$u_3(t) = u_4(t) \quad \forall t \in [0,1] \quad \text{ si y sólo si } \quad A = B = 0$$

con

$$A = a(u_1, u_2 - u_1) - (g, u_2 - u_1) + \int_{\Gamma_2} q_1(u_2 - u_1)\, d\gamma \geq 0,$$
$$B = a(u_2, u_1 - u_2) - (g, u_1 - u_2) + \int_{\Gamma_2} q_2(u_1 - u_2)\, d\gamma \geq 0.$$

En la Seccion 2 se establece una estimación de la distancia entre $u_3(t)$ y $u_4(t)$ en el caso en que A y B no sean iguales a cero y se da una propiedad de monotonía, de la cual se desprende la unicidad de los controles óptimos. De manera análoga se definen $u_{\alpha 3}(t)$ y $u_{\alpha 4}(t)$ (para cada $\alpha > 0$) y se obtienen resultados similares vinculados a la inecuación variacional elíptica (5). En [5] se consideraron problemas de control óptimo distribuido, donde la variable de control era la fuente de energía g.

En este trabajo se definen los funcionales de costo $J : Q \to \mathbb{R}$ y $J_\alpha : Q \to \mathbb{R}, \forall \alpha > 0$ [10, 12]:

$$J(q) = \frac{1}{2}\|u_q\|_H^2 + \frac{M}{2}\|q\|_Q^2, \tag{6}$$

$$J_\alpha(q) = \frac{1}{2}\|u_{\alpha q}\|_H^2 + \frac{M}{2}\|q\|_Q^2, \tag{7}$$

y se consideran los problemas de control óptimo frontera siguientes:

$$\text{Hallar} \quad q_{op} \in Q \quad \text{tal que} \quad J(q_{op}) = \min_{q \in U_{ad}} J(q), \tag{8}$$

$$\text{Hallar} \quad q_{op_\alpha} \in Q \quad \text{tal que} \quad J(q_{op_\alpha}) = \min_{q \in U_{ad}} J_\alpha(q), \tag{9}$$

donde

$$U_{ad} = \{q \in Q : q \geq 0 \text{ en } \Gamma_2\}\,,$$

.

En la Sección 3 se demuestra la existencia y unicidad de los controles óptimos de los problemas (8) y (9) con una prueba diferente a la dada en [14].

En la Sección 4 se demuestra que el control óptimo q_{op_α} y su correspondiente estado $u_{\alpha q_{op_\alpha}}$ convergen fuertemente hacia q_{op} y $u_{q_{op}}$ respectivamente, cuando $\alpha \to +\infty$, en adecuados espacios funcionales. Se resalta que no se necesita considerar el estado adjunto para los problemas (4) y (5) como en [6, 7, 13] para probar la convergencia cuando $\alpha \to +\infty$. Esta es una muy importante ventaja del presente resultado con respecto a la demostración previa dada para sistemas gobernados por ecuaciones variacionales elípticas en [7].

En los libros [2, 12] se pueden encontrar diferentes problemas de control óptimo gobernados por ecuaciones diferenciales a derivadas parciales. Algunos trabajos sobre control óptimo de sistemas gobernados por inecuaciones variacionales elípticas son [1, 3, 8, 9, 15].

2. ESTIMACIONES

Lema 1 *Las funciones $u_3(t)$ y $u_4(t)$ satisfacen la siguiente estimación:*

$$m\|u_4(t)-u_3(t)\|_V^2 + tI_{14}(t) + (1-t)I_{24}(t) \leq t(1-t)(A+B)$$

donde

$$I_{14}(t) = a(u_1, u_4(t)-u_1) - (g, u_4(t)-u_1) + \int_{\Gamma_2} q_1(u_4(t)-u_1)\, d\gamma \geq 0$$

$$I_{24}(t) = a(u_2, u_4(t)-u_2) - (g, u_4(t)-u_2) + \int_{\Gamma_2} q_2(u_4(t)-u_2)\, d\gamma \geq 0.$$

Más aún, se tiene la siguiente propiedad de monotonía [14]

$$u_4(t) \leq u_3(t), \quad \forall t \in [0,1], \quad \forall q_1, q_2 \in Q.$$

De manera similar, para cada $\alpha > 0$, se consideran las únicas soluciones $u_{\alpha 1}$ y $u_{\alpha 2}$ de la inecuación variacional elíptica (5) para flujos q_1 y q_2 respectivamente y se definen $\forall t \in [0,1]$:

$$u_{\alpha 3}(t) = (1-t)u_{\alpha 1} + tu_{\alpha 2} \quad ; \quad u_{\alpha 4}(t) = u_{\alpha[(1-t)q_1+tq_2]}$$

De forma análoga al Lema 1 se definen A_α, B_α, $I_{\alpha 14}(t)$, $I_{\alpha 24}(t)$ y se prueba el siguiente resultado:

Lema 2 *Las funciones $u_{\alpha 3}(t)$ and $u_{\alpha 4}(t)$ satisfacen la estimación:*

$$m_\alpha\|u_{\alpha 4}(t)-u_{\alpha 3}(t)\|_V^2 + tI_{\alpha 14}(t) + (1-t)I_{\alpha 24}(t) \leq t(1-t)(A_\alpha + B_\alpha)$$

y la propiedad de monotonía

$$u_{\alpha 4}(t) \leq u_{\alpha 3}(t), \quad \forall t \in [0,1], \quad \forall q_1, q_2 \in Q.$$

3. EXISTENCIA Y UNICIDAD DE LOS CONTROLES ÓPTIMOS

Teorema 1 *Los funcionales de costo J y $J_\alpha (\forall \alpha > 0)$ satisfacen las siguientes propiedades:*

a) son semicontinuos inferiormente en Q débil,

b) $\lim\limits_{\|q\|_Q \to +\infty} J(q) = +\infty \quad ; \quad \lim\limits_{\|q\|_Q \to +\infty} J_\alpha(q) = +\infty,$

c) son estrictamente convexos sobre Q,

y por lo tanto, existen únicas soluciones q_{op} y q_{op_α} en Q de los problemas de optimización (8) *y* (9) *respectivamente.*

Prueba. a) y b) Resultan usando [11],[12].

c) Sean $u = u_{q_i}$ y $u_{\alpha q_i}$ las soluciones de las inecuacionales variacionales (4) y (5) respectivamente con $q = q_i$ para $i = 1, 2$. Se deducen las siguientes igualdades:

$$\|u_3(t)\|_H^2 = t\|u_{q_1}\|_H^2 + (1-t)\|u_{q_2}\|_H^2 - t(1-t)\|u_{q_2} - u_{q_1}\|_H^2, \tag{10}$$

$$\|u_{\alpha 3}(t)\|_H^2 = t\|u_{\alpha q_1}\|_H^2 + (1-t)\|u_{\alpha q_2}\|_H^2 - t(1-t)\|u_{\alpha q_2} - u_{\alpha q_1}\|_H^2. \tag{11}$$

Luego, de la propiedad de monotonía

$$u_4(t) \leq u_3(t) \quad en \quad \Omega, \quad \forall t \in [0,1], \tag{12}$$

se deduce que

$$tJ(q_1) + (1-t)J(q_2) - J(q_3) \geq \frac{t(1-t)}{2}\left\{\|u_{q_1} - u_{q_2}\|_V^2 + M\|q_1 - q_2\|_Q^2\right\} > 0 \tag{13}$$

para todo $t \in]0,1[$ y para todo q_1, q_2 en Q. Por lo tanto J es un funcional estrictamente convexo con lo cual se tiene la unicidad del problema de control óptimo (8). La unicidad del problema de control óptimo (9) se deduce de una manera similar para cualquier $\alpha > 0$. □

4. CONVERGENCIA DE PROBLEMAS DE CONTROL ÓPTIMO CUANDO $\alpha \to +\infty$

En esta sección se estudia la convergencia del estado $u_{\alpha q_{op_\alpha}}$, y del control óptimo q_{op_α}, cuando el coeficiente de transferencia de calor α sobre Γ_1, tiende a infinito. Para un dado $q \in Q$ fijo, se tiene primero la siguiente propiedad que generaliza la obtenida para ecuaciones variacionales en [16, 17].

Lema 3 *Sean $u_{\alpha q}$ la única solución de la inecuación variacional* (5) *y u_q la única solución de la inecuación variacional* (4)*, entonces se tiene que*

$$u_{\alpha q} \to u_q \quad en \quad V \text{ fuertemente cuando } \alpha \to +\infty \quad \forall q \in Q.$$

Ahora se da el resultado más importante del trabajo que generaliza el resultado de convergencia obtenido en [7], para inecuaciones variacionales elípticas, y sin necesidad de utilizar los estados adjuntos. Se resalta la doble dependencia del parámetro α en la expresión del estado del sistema $u_{\alpha q_{op_\alpha}}$ correspondiente al control óptimo q_{op_α}.

Teorema 2 *Sean $u_{\alpha q_{op_\alpha}}$, g_{op_α} y $u_{q_{op}}$, q_{op} los estados y los controles óptimos definidos en los problemas* (9) *y* (8) *respectivamente. Entonces, se obtienen los comportamientos asintóticos siguientes:*

$$\lim_{\alpha \to +\infty} \|u_{\alpha q_{op_\alpha}} - u_{q_{op}}\|_V = 0, \quad \lim_{\alpha \to +\infty} \|q_{op_\alpha} - q_{op}\|_Q = 0. \tag{14}$$

AGRADECIMIENTOS

Trabajo subsidiado por PIP No. 0460 de CONICET-UA y Grant FA9550-10-1-0023, Rosario, Argentina.

REFERENCIAS

[1] K. AIT HADI, *Optimal control of the obstacle problem: optimality conditions*, IMA J. Math. Control Inform., 23 (2006), pp. 325-334.

[2] V. BARBU, *Optimal control of variational inequalities*. Research Notes in Mathematics, 100. Pitman (Advanced Publishing Program), Boston, 1984.

[3] M. BERGOUNIOUX, *Optimal Control of problems governed by abstract elliptic variational inequalities with state constraints*, SIAM J. Control and Optimization, 36 (1998), pp. 273-289.

[4] M. BOUKROUCHE AND D. A. TARZIA, *On a convex combination of solutions to elliptic variational inequalities*, Electro. J. Dif. Eq., 2007, No. 31 (2007), pp. 1-10.

[5] M. BOUKROUCHE AND D. A. TARZIA, *Convergencia de controles óptimos distribuidos para inecuaciones variacionales elípticas*, in Congreso II MACI 2009, MACI, 2 (2009), pp. 459-462.

[6] C.M. GARIBOLDI AND D.A. TARZIA, *Convergence of distributed optimal controls on the internal energy in mixed elliptic problems when the heat transfer coefficient goes to infinity*, Appl. Math. Optim., 47 (2003), pp. 213-230. Ver también, *A new proof of the convergence of distributed optimal controls on the internal energy in mixed elliptic problems*, MAT - Serie A, 7 (2004), pp. 31-42.

[7] C.M. GARIBOLDI AND D.A. TARZIA, *Convergence of boundary optimal control problems with restrictions in mixed elliptic Stefan-like problems*, Adv. in Diff. Eq. and Control Processes, 1 (2008), pp. 113-132.

[8] J. HASLINGER AND T. ROUBICEK, *Optimal control of variational inequalities. Approximation Theory and Numerical Realization*, Appl. Math. Optim., 14 (1987), pp. 187-201.

[9] K. ITO AND K. KUNISCH, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343-364.

[10] S. KESAVAN AND T. MUTHUKUMAR, *Low-cost control problems on perforated and non-perforated domains*, Proc. Indian Acad. Sci. (Math. Sci.), 118 (2008), pp. 133-157.

[11] D. KINDERLEHRER AND G. STAMPACCHIA, *An introduction to variational inequalities and their applications*. Academic Press, New York, 1980.

[12] J.L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[13] J.L. MENALDI AND D. A. TARZIA, *A distributed parabolic control with mixed boundary conditions.* Asymptotic Analysis, 52 (2007), pp. 227-241.

[14] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques*, J. Functional Analysis, 22 (1976), pp. 130–185.

[15] F. MIGNOT AND J.P. PUEL, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466-476.

[16] E.D. TABACMAN AND D. A. TARZIA, *Sufficient and/or necessary condition for the heat transfer coefficient on Γ_1 and the heat flux on Γ_2 to obtain a steady-state two-phase Stefan problem*, J. Dif. Eq., 77 (1989), pp. 16-37.

[17] D. A. TARZIA, *Una familia de problemas que converge hacia el caso estacionario del problema de Stefan a dos fases*, Math. Notae, 27 (1979), pp. 157-165.

UNA GENERALIZACIÓN SOBRE LAS RESTRICCIONES DE ESTADO PARA UN SISTEMA DINÁMICO CON SALTOS

Eduardo A. Philipp[♭,†] y Elina M. Mancinelli[♭,†]

[♭]*FCEIA, Universidad Nacional de Rosario, Pellegrini 250, 2000 Rosario, Argentina*
[†]*CONICET, Argentina*
eduardo@fceia.unr.edu.ar, elina@fceia.unr.edu.ar, www.fceia.unr.edu.ar/optimiz_control/

Resumen: Consideramos un problema de control con saltos en la dinámica y restricciones de estado. Fijado el control, asumiendo el conjunto de restricciones de estado convexo, se tiene una noción satisfactoria de solución para la dinámica y, a partir de la técnica de completación del gráfico, se obtiene la existencia y unicidad de solución en este sentido. En el presente trabajo generalizamos el conjunto de restricciones de estado a un conjunto estrella y damos una condición suficiente sobre la dinámica para asegurar la existencia y unicidad de solución.

Palabras clave: *medida de Radon, completación del gráfico, restricciones de estado*
2000 AMS Subject Classification: 93C30 - 49J55 - 34A37

1. PLANTEO DEL PROBLEMA

Consideramos el siguiente sistema dinámico controlado. Dados los campos vectoriales $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ y $g_i : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n$, $i = 1 \dots M$, consideramos el problema de Cauchy:

$$(E) \begin{cases} \dot{y}(t) = f(t, y(t), \alpha(t)) + \sum_{i=1}^{M} g_i(t, \alpha(t))\, \dot{u}_i(t), & \text{para } t \in (\tau, T], \\ y(\tau^-) = x, \end{cases}$$

donde:

- El control $\alpha \in \mathcal{A}$.
- $u = (u_1, .., u_M)$ es una función dada fija perteneciente a $VA^-([0,T], \mathbb{R}^M)$ y por lo tanto $\dot{u}$ es una medida de Radon.
- f, g están acotadas globalmente.

Asumiremos lo siguiente:

1. El conjunto de controles es $\mathcal{A} := \{\alpha : (0,T) \to A \text{ medible}\}$, con $A \subseteq \mathbb{R}^m$ compacto.
2. Las funciones $f, g_i,\ i = 1 \dots M$ son medibles en t, y continuas en (Y, a) y en a respectivamente. Más aun, para cada $Y \in \mathbb{R}^n, a \in A$ tenemos $f(\cdot, Y, a) \in L^1(\mathbb{R}^+)$ y $g_i(\cdot, a) \in L^1_{u_i}(\mathbb{R}^+),\ i = 1 \dots M$.
3. Existe una función $k_0 \in L^\infty(\mathbb{R}^+; \mathbb{R}^+)$ tal que
$$|f(t, Y, a) - f(t, Z, a)| \le k_0(t)|Y - Z| \qquad \forall Y, Z \in \mathbb{R}^n, a \in A \text{ y p.c.t. } t > 0$$
$$|g_i(t, Y) - g_i(t, Z)| \le k_0(t)|Y - Z| \qquad \forall Y, Z \in \mathbb{R}^n, \text{ p.c.t. } t > 0, i = 1..M$$
4. Existe $K > 0$ tal que
$$|g_i(t, a)| \le K, \quad |f(t, Y, a)| \le K \qquad \forall Y \in \mathbb{R}^n, a \in A \text{ y p.c.t. } t > 0, i = 1..M$$

2. COMPLETACIÓN DEL GRÁFICO

Consideraremos el caso donde tenemos sólo un salto. Utilizamos la técnica de completación del gráfico introducida por Dalmaso-Rampazzo en [5]. Sea $\eta \in (0,T)$ el único punto de discontinuidad de u. Llamaremos $\mathcal{T} := \{0, \eta\}$. Sean ψ_0 y ψ_η funciones Lipschitzianas y de VA que mapean $[0,1]$ en $\mathbb{R}^M$ tales que

$$\psi_\eta(0) = u(\eta^-)\ ,\ \psi_\eta(1) = u(\eta^+) \text{ y } \psi_0(1) = u(0^+). \tag{1}$$

Para $t \in \mathcal{T}$ denotamos ξ la solución de

$$\begin{cases} \dfrac{d\xi}{d\sigma} = g(\sigma, \xi(\sigma), \alpha(\sigma)) \dfrac{d\psi_t}{d\sigma}, & \sigma \in (0,1], \\ \xi(0) = \bar{\xi}, \end{cases}$$

y escribimos $\xi(\bar{\xi}, \psi_t) := \psi(1) - \bar{\xi}$. Definiremos ahora lo que consideraremos solución de (E).

Definición 1 *Fijamos un punto inicial (x,τ) y un control $\alpha \in \mathcal{A}$. La función $y_x \in VA([\tau,T],\mathbb{R}^n)$ es una solución de (E) si para cada conjunto de Borel B de (τ,T) tenemos*

$$\int_B \dot{y}_x(t) = \int_B f(t,y_x(t),\alpha(t))dt + \sum_{i=1}^{M} \int_{B\cap\mathcal{T}^c} g_i(t,y_x(t),\alpha(t))d\dot{u}_i + \sum_{t\in\mathcal{T}\cap B} \xi(y_x(t^-),\psi_t)$$

y $y_x(\tau^-) = x$. Más aún , si $\tau \in \mathcal{T}$ tenemos $y_x(\tau^+) = \xi(x,\psi_\tau)$.

Ahora fijamos $\quad a_t := V_0^1(\psi_t), \qquad a := a_0 + a_\eta, \qquad w(t) := \frac{t+V_0^t(u)}{T+V_0^T(u)},$

y definimos $W : [0,T] \to [0,1]$ como sigue:

$$W(t) := \frac{1}{1+a}\,(w(t) + \sum_{s\in\mathcal{T},s<t} a_s\,). \tag{2}$$

La completación del gráfico de u correspondiente a (ψ_0,ψ_η) queda definida por:

$$\Phi(s) = (\phi^0,\phi^1,..,\phi^M)(s) = \begin{cases} (t,u(t)), & \text{si } s = W(t), \qquad t \in [0,T]\backslash\mathcal{T}, \\ (t,\psi_t(\frac{s-W(t)}{W(t^+)-W(t)})), & \text{si } s \in [W(t),W(t^+)], \quad t \in \mathcal{T}. \end{cases}$$

El caso particular en donde en la familia $(\phi_t)_{t\in\mathcal{T}}$, cada miembro es una función lineal, la correspondiente completación del gráfico Φ es llamada completación canónica del gráfico. Tenemos los elementos para construir la reparametrización del sistema. Sea $\sigma := W(\tau)$, para cada control $\alpha \in \mathcal{A}$ y posición inicial x denotamos $z_{x,\sigma} : [\sigma,1] \to \mathbb{R}^n$ la solución de

$$(E_R)\begin{cases} \dfrac{dz}{ds}(s) &= \displaystyle\sum_{i=1}^{M} g_i(\phi^0(s),z(s))\left[\dot{u}_i^a(\phi^0(s))\frac{d\phi^0}{ds}(s) + \frac{d\phi^i}{ds}(s)\right] \\ &\quad + f(\phi^0(s),z(s),\alpha(\phi^0(s))\dfrac{d\phi^0}{ds}(s), & s \in (\sigma,1] \\ z(\sigma) &= x, \end{cases}$$

donde $\dot{u}^a$ es la parte absolutamente continua de la medida $\dot{u}$ con respecto a la medida de Lebesgue, es decir $\dot{u}(t) = \dot{u}^a(t)dt + \dot{u}^s$. Observemos que las derivadas de ϕ^0,ϕ^i son funciones medibles, por lo tanto las hipótesis 2 y 3 aseguran la aplicabilidad del teorema de Caratheodory para obtener la existencia y unicidad de solución en $AC([\sigma,1];\mathbb{R}^n)$.

Nuestro objetivo es conectar los problemas (E) y (E_R). En [4], en el caso sin restricciones de estado se demuestra una equivalencia entre ellos:

Teorema 1 *Supongamos válidos (1)-(4). Sea $\dot{u}$ una medida de Radon y $(\phi_t)_{t\in\mathcal{T}}$ como antes. Entonces $y_{x,\tau} \in VA^-([\tau,T],\mathbb{R}^n)$ es una solución de (E) si y sólo si existe una solución $z_{x,\sigma} \in AC([\sigma,1],\mathbb{R}^n)$ del problema reparametrizado correspondiente a la completación de gráfico Φ tal que:*

$$z_{x,\sigma}(W(t)) = y(t) \qquad \forall t \in [\tau,T]$$

donde W esta dado en (2)*. Más aún, para cada medida de Radon $\dot{u}$ y cada familia (ϕ_0,ϕ_η), la ecuación (E) tiene una única solución (hasta un conjunto de medida de Lebesgue cero).*

3. La dinámica con restricciones de estado

Sea $\mathcal{K} \subseteq \mathbb{R}^n$ un conjunto cerrado de restricción de estado. Para un $(x,\tau) \in \mathbb{R}^n \times [0,+\infty)$ fijo definimos el conjunto de trayectorias admisibles:

$$S^{\mathcal{K}}_{[\tau,T]} := \{y_x \text{ sol. de } (E), y_x(s) \in \mathcal{K}\ \forall s \in [\tau,T]\}.$$

Similarmente definimos el conjunto de trayectorias admisibles para el problema reparametrizado (E_R):

$$SR^{\mathcal{K}}_{[\sigma,1]} := \{z_{x,\sigma} \text{ sol. de } (E_R), z_{x,\sigma}(s) \in \mathcal{K}\ \forall s \in [\sigma,1]\}.$$

En [6] hemos probado que si $M = 1$ y $\mathcal{K}$ es convexo, tenemos la siguiente correspondencia entre las soluciones de (E) y (E_R)

Proposición 1 *Sea $\alpha \in \mathcal{A}$ tal que la correspondiente solución de (E_R) es admisible para el problema restringido, esto es: $z_{x,\sigma} \in SR^{\mathcal{K}}_{[\sigma,1]}$. Sea $(\psi_t)_{t\in\mathcal{T}}$ que verifica* (1) *y Φ la correspondiente completación del gráfico de u. Entonces la función definida como $y_x(t) := z_{x,\sigma}(W(t))$ para todo $t \in [\tau,T]$ pertenece a $S^{\mathcal{K}}_{[\tau,T]}$.*

La prueba de la Proposición 1 no depende del valor de M, por lo tanto el mismo resultado es válido para cualquier M. Más aun, este resultado no depende del conjunto $\mathcal{K}$ por lo tanto es válido para cualquier hipótesis impuesta al mismo.

Proposición 2 *Consideramos $\alpha \in \mathcal{A}$ tal que la correspondiente solución de (E) es admisible para el problema restringido, esto es: $y_x \in S^{\mathcal{K}}_{[\tau,T]}$. Sea Φ la completación canónica del gráfico de u. Entonces existe una única $z_{x,\sigma} \in SR^{\mathcal{K}}_{[\sigma,1]}$ tal que $y_x(t) := z_{x,\sigma}(W(t))$ para todo $t \in [\tau, T]$.*

Daremos la demostración para un M general en base a la prueba dada en [6] para $M = 1$.

Prueba. Sabemos que $y_x(s) \in \mathcal{K}\ \forall s \in [\tau, T]$. Naturalmente asumimos que $\sigma < \eta$ en otro caso caeríamos en un caso trivial. Por Teorema 1 sabemos que existe una única $z_{x,\sigma} \in AC([\sigma, 1]; \mathbb{R}^n)$ solución de (E_R) tal que $z_{x,\sigma}(W(t)) = y(t) \qquad \forall t \in [\tau, T]$.

Luego debemos chequear solamente que $z_{x,\sigma}(s) \in \mathcal{K}$ para todo $s \in [\sigma, 1]$. Sea $s \in [\sigma, 1]$.

• Si $s = W(t)$ para algún $t \in [\tau, T]$ entonces sabemos que $z_{x,\sigma}(s) = z_{x,\sigma}(W(t)) = y_x(t) \in \mathcal{K}$.

• Si $s \neq W(t)$ para todo $t \in [\tau, T]$ entonces $s \in [W(\eta), W(\eta^+)]$. Notemos que para todo $v \in [W(\eta), W(\eta^+)]$, dado que $z_{x,\sigma}$ resuelve (E_R), tenemos

$$\frac{dz}{dv}(v) = \underbrace{\sum_{i=1}^{M} g_i(\eta, \alpha(\eta)) \frac{d\phi^i}{dv}\left(\frac{v - W(\eta)}{W(\eta^+) - W(\eta)}\right)}_{=\text{constante}} + f(\eta, z(v), \alpha(\eta)) \underbrace{\frac{d}{dv}\eta}_{=0},$$

puesto que $(\phi^1, .., \phi^M) = \psi_\eta$ en $[W(\eta), W(\eta^+)]$. Luego $\frac{dz}{dv}(v)$ es constante para todo $v \in [W(\eta), W(\eta^+)]$. Por lo tanto la trayectoria $z_{x,\sigma}$ es una función lineal en $[W(\eta), W(\eta^+)]$ que une los puntos $z(W(\eta)) = y(\eta)$ y $z(W(\eta^+)) = y(\eta^+)$ los cuales pertenecen a $\mathcal{K}$. Ahora, teniendo en cuenta que $\mathcal{K}$ es convexo concluimos que $z_{x,\sigma}(s) \in \mathcal{K}$. Por lo tanto $z_{x,\sigma} \in SR^{\mathcal{K}}_{(\sigma,1]}$. □

4. El conjunto de restricciones de estado en forma de estrella

Nuestro propósito es generalizar el resultado para un $\mathcal{K}$ más general. Podemos pensar en primer lugar en utilizar una completación del gráfico de u más general que la canónica. En el caso en que $M = 1$, observando la dinámica (E_R) si $v \in [W(\eta), W(\eta^+)]$:

$$\frac{dz}{dv}(v) = g_1(\eta, \alpha(\eta)) \frac{d\phi^1}{dv}\left(\frac{v - W(\eta)}{W(\eta^+) - W(\eta)}\right)$$

notamos que cualquiera sea la completación del gráfico de u, la imágen de $z_{x,\sigma}$ siempre será el segmento que une $y_x(t)$ y $y_x(t^+)$, dado que $\frac{dz}{dv}(v)$ es siempre un múltiplo de $g_1(\eta, \alpha(\eta))$ que es constante, con lo cual la generalización de nuestro previo resultado para un $\mathcal{K}$ más general no es posible si $M = 1$ para ninguna completación del gráfico de u alternativa (es decir que $\mathcal{K}$ convexo es lo más restrictivo que podemos pedir). Esto nos induce a considerar el caso en el cual $M > 1$.

Definición 2 *Decimos que $C \subseteq \mathbb{R}^n$ es un conjunto estrella si existe $P \in C$ tal que para todo $Q \in C$, el segmento $\overline{PQ}$ está contenido en C.*

Consideramos $\mathcal{K}$ un conjunto estrella y $P \in \mathcal{K}$ como en la definición. Sin fijar ψ_η, observamos que para $v \in [W(\eta), W(\eta^+)]$, la dinámica (E_R) se reduce a:

$$\frac{dz}{dv}(v) = \sum_{i=1}^{M} g_i(\eta, \alpha(\eta)) \frac{d\phi^i}{dv}\left(\frac{v - W(\eta)}{W(\eta^+) - W(\eta)}\right). \tag{3}$$

Teniendo en cuenta la definición de conjunto estrella, deseamos encontrar una ψ_η que determine una $z_{x,\sigma}$ admisible. Consideremos una completación del gráfico seccionalmente lineal (de hecho con un solo punto de no derivabilidad).

Teorema 2 *Consideramos $\alpha \in \mathcal{A}$ tal que la correspondiente solución de (E) es admisible para el problema restringido, esto es: $y_x \in S^{\mathcal{K}}_{[\tau,T]}$. Suponemos válida la condición*

$$P - y(\eta) \in gen\{g_i(\eta, \alpha(\eta))\}_{i=1}^{M}. \tag{4}$$

Luego existe Φ, una completación del gráfico de u, seccionalmente lineal y tal que existe una única $z_{x,\sigma} \in SR^{\mathcal{K}}_{[\sigma,1]}$ tal que $y_x(t) := z_{x,\sigma}(W(t))$ para todo $t \in [\tau, T]$.

Prueba. En el caso en que $P = y(\eta)$ o $P = y(\eta^+)$, el segmento $y(\eta), y(\eta^+)$ está contenido en $\mathcal{K}$ y por lo tanto podemos considerar la completación canónica del gráfico como en el caso de $\mathcal{K}$ convexo. Suponemos lo contrario. Construiremos una completación del gráfico de u tal que $z_{x,\sigma}(s) \in \mathcal{K}$ para todo $s \in [\sigma, 1]$. Por (4) sabemos que existen $A_1, .., A_M$ no todos nulos tales que $P - y(\eta) = \sum_{i=1}^{M} A_i g_i(\eta, \alpha(\eta))$. Como $\mathcal{K}$ es un conjunto estrella, sabemos que los segmentos $\overline{y(\eta)P}$ y $\overline{Py(\eta^+)}$ están en $\mathcal{K}$. Deseamos que la completación del gráfico sea tal que la imagen de $z_{x,\sigma}(v)$ si $v \in [W(\eta), W(\eta^+)]$ sea precisamente estos dos segmentos. Sea

$$\psi_\eta(s) = \begin{cases} u(\eta) + \frac{2}{W(\eta^+) - W(\eta)}(A_1, .., A_M)s, & \text{si } s \in [0, \frac{1}{2}], \\ u(\eta^+) + 2[\psi_\eta(\frac{1}{2}) - u(\eta^+)](1 - s), & \text{si } s \in (\frac{1}{2}, 1], \end{cases}$$

Notamos que ψ_η es seccionalmente lineal y que $\psi_\eta(0) = u(\eta)$, $\psi_\eta(1) = u(\eta^+)$. Además como $\psi_\eta(\frac{1}{2}^+) = \psi_\eta(\frac{1}{2})$ resulta que ψ_η es continua, por lo tanto se trata de una completación del gráfico válida. Sea $v \in [W(\eta), \frac{W(\eta)+W(\eta^+)}{2}]$, entonces por (3) se tiene:

$$\frac{dz}{dv}(v) = \frac{2}{W(\eta^+) - W(\eta)} \sum_{i=1}^{M} A_i g_i(\eta, \alpha(\eta)) = \frac{2}{W(\eta^+) - W(\eta)}(P - y(\eta)).$$

Luego $z_{x,\sigma}(v) = y(\eta) + \frac{2}{W(\eta^+) - W(\eta)}(P - y(\eta))(v - W(\eta))$ si $v \in [W(\eta), \frac{W(\eta)+W(\eta^+)}{2}]$, que es una función lineal y además $z_{x,\sigma}(\frac{W(\eta) - W(\eta^+)}{2}) = P$, luego esta porción de imagen de $z_{x,\sigma}$ permanece en $\mathcal{K}$.
Ahora, si $v \in (\frac{W(\eta)+W(\eta^+)}{2}, W(\eta^+)]$, sabemos que siendo ψ_η una completación del gráfico válida, deberá valer que $z_{x,\sigma}(W(\eta^+)) = y(\eta^+)$ y además es claro que $z_{x,\sigma}$ también es una función lineal en este intervalo, dado que $\frac{dz}{dv}(v)$ es constante, en efecto:

$$\frac{dz}{dv}(v) = -2 \sum_{i=1}^{M} g_i(\eta, \alpha(\eta))(\psi_\eta(\tfrac{1}{2}) - u(\eta^+))_i.$$

Luego la imagen de $z_{x,\sigma}$ si $v \in (\frac{W(\eta)+W(\eta^+)}{2}, W(\eta^+)]$ es el segmento $\overline{Py(\eta^+)}$ que está contenido en $\mathcal{K}$.
Queda probado entonces que $z_{x,\sigma} \in SR^{\mathcal{K}}_{[\sigma,1]}$. □

Hemos probado entonces la equivalencia de soluciones entre (E) y (E_R) bajo la suposición (4) en el caso en que $\mathcal{K}$ es un conjunto estrella.

AGRADECIMIENTOS

El trabajo ha sido parcialmente subsidiado por los proyectos PIP 112-200801-00460 CONICET e ING298 UNR.

REFERENCIAS

[1] O. BOKANOWSKI, N. FORCADEL, H. ZIDANI, *Deterministic state constrained optimal control problems without controllability assumptions*. ESAIM CONTROL, OPTIMISATION AND CALCULUS OF VARIATIONS. DOI: 10.1051/COCB/2010030, (2010)

[2] O. BOKANOWSKI, N. FORCADEL, H. ZIDANI, *Reachability and minimal times for state constrained nonlinear problems without any controllability assumption*. SIAM JOURNAL OF CONTROL AND OPTIMIZATION. VOL. 48(7) (2010). PP. 4292-4316.

[3] A. BRIANI, *A Hamilton-Jacobi equation with measures arising in* Γ*-convergence of optimal control problems*. DIFFERENTIAL AND INTEGRAL EQUATIONS, 12(6) (1999), PP. 849-886.

[4] A. BRIANI, H. ZIDANI, *Characterisation of the value function of final state constrained control problems with BV trajectories*. PREPRINT.

[5] G. DAL MASO Y F. RAMPAZZO, *On systems of ordinary differential equations with measures as controls*, DIFFERENTIAL AND INTEGRAL EQUATIONS, VOL. 4. NUMBER 4, (1991) PP. 739-765.

[6] E. PHILIPP, *Problemas de control óptimo con restricciones de estado para sistemas que involucran medidas de Radon*. TESINA DE GRADO DE LICENCIATURA EN MATEMÁTICA. FCEIA, UNR, ARGENTINA. 2010.

NUMERICAL SOLUTION OF A MIN-MAX PROBLEM USING A SPECIALLY DESIGNED NECESSARY OPTIMAL CONDITION

Laura S. Aragone[♭] and Pablo A. Lotito[†]

[♭]*OPTyCON, Universidad Nacional de Rosario y CONICET, laura@fceia.unr.edu.ar*
[†]*PLADEMA-UNCPBA, OPTyCON-UNR y CONICET, plotito@exa.unicen.edu.ar*

Abstract: We consider a min-max problem where the objective function is evaluated over a trajectory given by an ordinary differential equation parametrized by the control. We derive a necessary condition that allows the design of a numerical method that performs very well in the examples given.

Keywords: *min-max optimization, necessary optimal condition, adjoint operator*
2000 AMS Subject Classification: 21A54 - 55P54

lo deje en spanish porque no se donde va la opcion english

1 DESCRIPTION OF THE PROBLEM

We consider in the interval $[0, T]$ a dynamic system which evolves according to the ordinary differential equation

$$\begin{cases} \dfrac{dy}{ds}(s) = g(y(s), \alpha(s)) \quad 0 \leq s \leq T, \\ \\ y(0) = x \in \Omega \subseteq \mathbb{R}^r, \qquad \Omega \text{ an open domain.} \end{cases}$$

The optimal control problem consists in minimizing the functional $J : \Omega \times \mathcal{U} \mapsto \mathbb{R}$

$$J(x, \alpha(\cdot)) = \operatorname{ess\,sup} \{f(y(s), \alpha(s)) : s \in [0, T)\},$$

over the set of controls

$$\mathcal{U} = \{\alpha : [0, T] \to A \subset \mathbb{R}^m : \alpha(\cdot) \text{ measurable}\}$$

where A is compact.

This problem arises when the minimization of the maximum deviation of the controlled trajectories with respect to a given special trajectory is sought. This differs from those problems usually considered in the optimal control literature, where an accumulated cost is minimized. As considering an accumulated cost is not always the best method to qualify a controlled system with an unique real parameter, minimax optimal control problems seem to be more naturally in many applications.

2 HYPOTHESES

We make the following hypothesis:

- f and g are bounded and uniformly continuous functions on $\Omega \times A$ and f is independent of α,
- $g(y, \alpha) = g_1(y) + g_2(y)\alpha$ and $g \in C^1$
- A convex,
- we also suppose that the trajectory $y(\cdot)$ remains in Ω, for any control in $\mathcal{U}$.

In order to obtain a necessary condition for α to be a minimizer, we would like compute the gradient or, at least, a directional derivative of J for α along an admissible direction v. It is easy to see that because of the involved definition of J, it could not exist. Nevertheless, if J has some degree of convexity we can guarantee the directional differentiability. We can prove the following lemma:

Lemma 1 *Let f be convex, then the functional J is convex in α.*

Proof. It is easy to see that the application $\alpha \mapsto y_\alpha$ is linear from $\mathcal{U}$ to $C[0,T]$. Then, the composition of that operator with the convex function f is convex. Taking the supremum of functions convex in α we get a function convex in α. For a detailed analysis of this operators see [3, 4]. □

3 Optimality conditions

Using classical convex analysis (see [5]) we can write the directional derivative as

$$J'(\alpha; v) = \sup_{t \in C_\alpha} \langle \nabla f(y_\alpha(t)), z_v(t) \rangle \tag{1}$$

where z_v solves the differential equation

$$\begin{cases} \dot{z_v} = & h_1(t) z_v + h_2(t) v, \\ z_v(0) = & 0, \end{cases} \tag{2}$$

the functions h_1 and h_2 are given by

$$h_1(t) = \nabla g_1(y_\alpha(t)) + \nabla g_2(y_\alpha(t)) \alpha(t), \tag{3}$$

$$h_2(t) = g_2(y_\alpha(t)) \tag{4}$$

and C_α is the set of critical times

$$C_\alpha = \arg\max_t f(y_\alpha(t)). \tag{5}$$

If α is an optimal point it is necessary that every directional derivative be positive, or that there is no direction with negative directional derivative. The last assertion is equivalent to

$$\inf_v \sup_{t \in C_\alpha} \langle \nabla f(y_\alpha(t)), z_v(t) \rangle \geq 0. \tag{6}$$

Let us explicit the linear operator $v \mapsto z_v$, the image is the solution of the linear differential equation

$$\begin{cases} \dot{z} = & h_1(t) z + h_2(t) v, \\ y(0) = & 0, \end{cases} \tag{7}$$

and the solution is given by

$$z_v(t) = \int_0^t S_{ts} v(s) ds \tag{8}$$

where the matrix S_{ts} is a matrix solution of the system

$$\begin{cases} \frac{d}{ds} S_{ts} = & h_1(s) S_{ts} + h_2(s), \\ S_{tt} = & I. \end{cases} \tag{9}$$

Now the directional derivative can be written as

$$J'(\alpha; v) = \sup_{t \in C_\alpha} \left\langle \nabla f(y_\alpha(t)), \int_0^t S_{ts} v(s) ds \right\rangle \tag{10}$$

using the linearity of the scalar product and the function $I_t(s) = 1$ if $s \leq t$ and 0 otherwise we can write the directional derivative as

$$J'(\alpha; v) = \sup_{t \in C_\alpha} \int_0^1 \langle I_t(s) S_{ts}^* \nabla f(y_\alpha(t)), v(s) \rangle \, ds \tag{11}$$

and if we define $q_{\alpha,t}(s) = I_t(s) S_{ts}^* \nabla f(y_\alpha(t))$ for $t \in C_\alpha$ and $Q_\alpha = \{q_{\alpha,t} : t \in C_\alpha\}$, we can write the directional derivative as

$$J'(\alpha; v) = \sup_{q \in Q_\alpha} \langle q, v \rangle \tag{12}$$

where the last scalar product is in L^2.

Using the expression in (12) the following theorem easily follows.

Theorem 1 *Let α be a minimum of J then*

$$\inf_{v} \sup_{q \in Q_\alpha} \langle q, v \rangle = 0. \tag{13}$$

In order to facilitate the computation of the condition before, we can prove the following proposition.

Proposition 1 *The necessary condition (13) implies*

$$\inf_{v} \sup_{t \in [0,T]} f(y_\alpha(t)) - \varphi_\alpha + \langle q_t, v \rangle = 0, \tag{14}$$

where $\varphi_\alpha = \max_{t \in [0,T]} f(y_\alpha(t))$.

Note 1 *If we suppose* $\max_{t \in [0,T]} f(y_\alpha(t)) = 0$*, then the condition (14) is equivalent to*

$$\inf_{v} \sup_{t \in [0,T]} \langle q_t, v \rangle = 0, \tag{15}$$

and because of the previous proposition this is also a (weaker) necessary condition.

4 Numerical results

Basically, the proposed algorithm at each step computes a descent direction solving the (discretized) necessary condition (15) and uses it to reduce the objective function. In order to evaluate the function and its derivatives along the trajectory we follow a similar discretization to the used in [1].

Example

We consider the following data $\Omega = R^2 \times [0,6]$, $T = 6$, $h = 2$, $f = x_3\sqrt{x_1^2 + x_2^2}$, $A = [-1,1] \times [-1,1]$ and

$$x_0 = \begin{pmatrix} 0.7778 \\ 1 \\ 0 \end{pmatrix} \qquad g_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad g_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

In the figure 1 we can see the value of $f(y_\alpha(t))$ at the left and the value of the optimal control (both components) at the right.

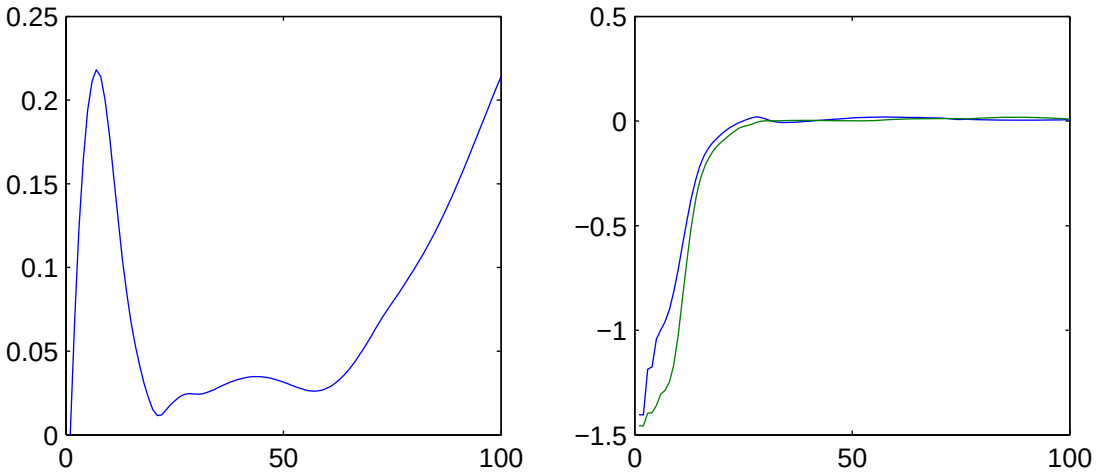


Figure 1: The value of the function ovwer the optimal trajectory (left) and both components of the optimal control α (right)

In this case the discretized version of the necessary condition (15) is given by

$$\min_{\delta v \in K} \left(\max_{n=0,\ldots,\mu} \left(\langle q_n, \delta v \rangle + f(y^h(t_n)) - \bar{f} \right) \right)$$

which is a linear programming problem and can be solved efficiently by usual methods. In the figure 2 we show the value of f at the successive trajectories, note that the max is always decreasing.

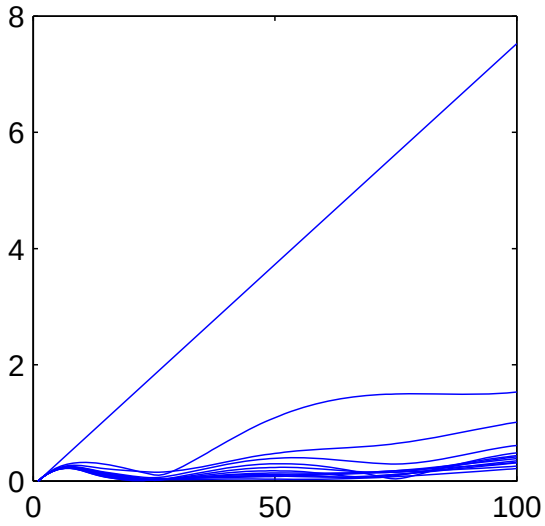


Figure 2: Iterations f

REFERENCES

[1] L.S. ARAGONE, E. M. MANCINELLI, G. F. REYERO, *El principio del mximo de Pontryagin para problemas de control optimo de tipo minimax*, Anales MACI 2009, pp. 467-470.

[2] E. POLAK , *Computational methods in optimization. A unified approach*, Academic Press, New York.

[3] H. CARTAN, *Cours de calcul differentiel*, Hermann, Paris, 1977.

[4] B. DACOROGNA, *Direct Methods in the Clculus of VaritionsPertubations analysis of optimi- zations problems. Springer Verlag,* Berlin, 1989.

[5] A. SHAPIRO, J.F. BONNANS, *Pertubations analysis of optimizations problems*. Springer Verlag, New York, 2000.

NON-LINEAR OPTIMAL CONTROL APPLIED TO ENERGY MANAGEMENT IN HYBRID ELECTRIC VEHICLES

Laura V. Pérez[♭], Cristian H. de Angelo[♭] and Víctor L. Pereyra[†, †1]

[♭]*Grupo de Electrónica Aplicada, Facultad de Ingeniería Universidad Nacional de Río Cuarto, Ruta 36 Km 601, Río Cuarto, Córdoba, Argentina, lperez@ing.unrc.edu.ar, http://www.ing.unrc.edu.ar/grupos/gea/*
[†]*Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, U.S.A*
[†1]*Energy Resources Engineering Department, Stanford University, California, U.S.A, victor@ca.wai.com*

Abstract: To minimize fuel consumption in hybrid electric vehicles it is necessary to define which one of the energy sources must be used at each time. Under the assumption that the velocity required of the vehicle is known a priori, this problem may posed as a non linear optimal control problem with control and state constraints. We find the solution to this problem using the optimality conditions given by Pontryagin Maximum Principle. This approach leads to boundary value problems that we solve using a software tool named PASVA4. On real time operation, the velocity to be required of the vehicle is not known in advance. Then, we show how the adjoint state obtained from the former problem may be used as a weighting factor, called ”equivalent consumption”. This weighting factor may be used to design power management suboptimal real time algorithms.

Keywords: *non linear constrained optimal control, Pontryagin Maximum Principle, boundary value problems solvers, hybrid electric vehicles, equivalent consumption minimization algorithms*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

Hybrid electric vehicles are those whose powertrain is formed by an electric motor fed by a bank of batteries and by a conventional internal combustion engine. This engine not only contributes to the traction each time the driver power requirements are greater than those able to be supplied by the electric system, but is also used to recharge the batteries. These vehicles are more efficient from the energetic point of view because of several reasons: the engine may be smaller than those in conventional vehicles; because of the existence of an additional energy source, it may be set constantly in its most efficient operating point; they have the facility of regenerative braking, which means that during braking the driving electric motor may change its operating mode to generator, using the kinetic energy that the vehicle has been accumulating to generate electrical energy that is sent back to the batteries. Because of these features, the powertrain in these vehicles shows energy flows in several amounts and directions. Then, in order to actually achieve fuel savings, a control system that determines the power split among the energy sources at each time according to the driver requirements, must be defined. This control, usually electronically implemented, is called ”supervisory control”. Each one of the vehicle devices has in turn its own control that obeys the supervisory commands at a lower level. The supervisory control objective is to minimize fuel consumption along a given mission. In the case of public transport, where the routes are repetitive or in vehicles equipped with positioning systems, it is possible to know in advance the future power requirements. In the general case, there is the drawback that the future power requirements are not known. The solution that has been usually used is to solve the problem in typical situations under the assumption that the power requirements are known and so, get knowledge for the design of on-line algorithms.

Under the assumption that the power requirements are known, the minimizing control may be obtained by solving a non-linear constrained optimal control problem. In this work, this optimal control problem is solved through the numerical solution of the optimality conditions given by Pontryagin Maximum Principle. To solve the resulting boundary value problem we use a software tool named PASVA4([1],[2]), which is an ODE solver that manages right hand side and solution discontinuities and additional unknown parameters, two features that appear in constrained optimal control problems. It is observed that the adjoint state represents a factor that allows translating the electric power supplied by the batteries into an equivalent fuel consumption. Once this factor has been computed, it is used for the design of a real time algorithm. Previous publications that present a similar approach ([3],[4],[5]) compute initial values for the adjoint state

by performing a dichotomic search using the results of successive solutions of the optimality conditions. PASVA4 solves the problem directly, even in the case where the required power severely changes as in the case of urban missions. This is precisely our aim since we seek for an on-line supervisory control strategy for the urban vehicle prototype that is being developed in our group. We tested our algorithms using the data from this vehicle.

2 MODEL AND PROBLEM STATEMENT

The model used is a simplified one that only takes into account the energy sources and the power flows among them ([6]). The power supplied by the engine is taken as the control variable and is named $u(t)$. We set that at each time the sum of the power from the engine plus the power from the batteries equals the power required by the driver which is called $r(t)$. Hence the power from the batteries is $r(t) - u(t)$. Concerning the energy drawn from the sources, it has to be considered that, in order to compensate for the losses that occur in each one of the intermediate energy conversion processes between the source and the power summing junction (e.g., chemical to mechanic between the fuel tank and the engine, mechanical to electrical between the engine and the generator), to actually supply a power $u(t)$, the source must deliver a greater amount. This effect will be represented by a function f_C that depends on the power that is being delivered and that in practice is experimentally determined by tests performed on the vehicle. Then, the fuel consumption can be expressed by the energy drawn from the fuel tank: $\int_0^T f_C(u)dt$.

A similar situation occurs in the electrical path, and therefore the energy within the batteries at each time is given by

$$x(t) = x_0 - \int_0^t f_B(x(s), r(s) - u(s))ds \tag{1}$$

where x_0 is the initial value and f_B is the function that represents the compensation for losses. The energy in the batteries is the state variable and will be denoted $x(t)$. We choose a "charge sustaining" operation mode for the vehicle by imposing that at the end of the mission the energy in the batterries be the same as in the beginning. There are in addition constraints in the flows that naturally have physical bounds. The problem statement is the following.

Find a piecewise continuous control u that maximizes $-\int_0^T f_C(u)dt$*, subject to*

$$\begin{aligned}
\dot{x} &= -f_B(x(t), r(t) - u(t)) \; \forall t \in [0, T] && (2)\\
x(0) &= x_0, \quad x(T) = x_0 && (3)\\
0 &\leq u(t) \leq u_{max} && (4)\\
K_{min} &\leq r(t) - u(t) \leq K_{max} \; \forall t \in [0, T] && (5)
\end{aligned}$$

where K_{max}, K_{min}, u_{max} and x_0 are known. The constraint on the state, i. e.,$x_{min} \leq x \leq x_{max}$ will be considered later.

3 OPTIMALITY CONDITIONS

To solve this problem using the optimality conditions given by Pontryagin Maximum Principle, we define the corresponding Hamiltonian:

$$H(t, x, \lambda, u) = -f_C(u(t) - \lambda(t) f_B(x(t), r(t) - u(t))),$$

where λ is the *adjoint* state (For brevity we skip here the algebraic constraints that are taken into account in the so called *augmented Hamiltonian* or *Lagrangian*).

The optimality conditions were derived in previous work ([7]). Essentially, they state that

$$
\begin{aligned}
u &= \arg\max_{\underline{U} \le u \le \bar{U}} H && (6)\\
\dot{x} &= -f_B(x(t), r(t) - u(t)) \; \forall t \in [0, T] && (7)\\
\dot{\lambda} &= \lambda \frac{\partial f_B}{\partial x}\\
x(0) &= x_0, \; x(T) = x_T.
\end{aligned}
$$

where $\underline{U}$ and $\bar{U}$, defined by $\underline{U}(t) = max(0, r(t) - K_{max}), \bar{U} = min(u_{max}, r(t) - K_{min})$ sum up the constraints on the flows.

The optimal solution is the function $u(t)$, $t \in [0, T]$, that maximizes H or equivalently minimizes $-H$, over the set of all functions satisfying the constraints. That means that the optimal u minimizes the sum of the power (including losses) that has to be supplied by the fuel-path plus the complementary power that has to be supplied by the electrical path times the adjoint state. Hence $\lambda(t)$ can be thought of as a weighting factor that scales the "cost" of using power from the electrical path against the "cost" of using power from the fuel path. If this "equivalent consumption factor" ([8])were known in advance we could compute the optimal control just by an instantaneous minimization of the Hamiltonian. Unfortunately, the optimality conditions must be solved globally in the whole interval $[0, T]$ (note that there are no boundary conditions for the adjoint state; they derive from the evolution of the state). Nevertheless, the computation off-line of $\lambda(t)$ over intervals with typical forms of $r(t)$ or over short time intervals, previous to the current one, may allow the design of algorithms to compute suboptimal control functions by an instantaneous optimization. This is our purpose.

A strong difficulty arises from the form of the function $r(t)$ that, for the case of a neighborhood vehicle like ours, changes according to the successive accelerations and decelerations that such a vehicle must perform in any urban mission. As a consequence, the bounds for the control function, though continuous, are highly time varying.

3.1 SOLUTION

We set up this problem to be solved using PASVA4. The input u to the state equation comes from the maximization of the Hamiltonian and then in general depends on the state and on the adjoint state, i.e., $u = u(x, \lambda)$. As the maximization is done in a bounded interval, u may be piecewise defined since, depending on the values of x and λ, the maximum may be interior to the interval or one of its extremes. This introduces switchings or even discontinuities in the right hand side of the equations. If, in addition, there are bounds on the state, the discontinuities occur at the junction points, i.e., the points where the state constraint becomes active or becomes inactive. These junctions points are not known in advance, since they depend on the state trajectory. Suitable time transformations may be done in such a way that the junction points become parameters of the differential equations ([2]). They need to be obtained together with the solution adding the corresponding additional multipoint boundary conditions. Because of the former features, we chose a software tool like PASVA4 that is capable of automatically managing those situations. It implements a method based in finite differences. The basic discretization is the trapezoidal rule, enhanced by deferred corrections. It includes an automatic procedure to choose the suitable mesh, based on the equi-distribution of an estimate of the local truncation error. It provides in addition an asimptotic estimate of the global truncation error ([1], [2]).

4 EXAMPLE OF A REAL TIME ALGORITHM

We now propose the following simple algorithm to be used on-line. Basically it consists of dividing the whole interval in subintervals of equal length (200 seconds, say) and solve the problem in each subinterval, using as required power the data of the previous 200 seconds.

Array $r[1...200]$

Initialize λ

For $interval = 1, N\_interval$
For $t = 1, ..., 200$
$s = t + interval * 200$
read $r[s]$
Compute $u* = argmin_u(f_C(u) + \lambda * f_B(r[s] - u))$
end for
Compute λ *using* $r[1, ..., 200]$
end for

5 DISCUSSION AND CONCLUSIONS

In the above sections we solved the control constrained problem without state constraints. That allowed us to compute the equivalent consumption factor λ over a short horizon and under the condition that the final state were equal to the initial state. For short horizons and for many driving cycles this condition is sufficient to ensure that the state trajectory does not deviate very much from that initial value and so the bounds for the state are not reached during the whole cycle.

Nevertheless, there may be driving cycles for which including the state constraints in the formulation is necessary to ensure that the trajectory will remain within the bounds. It is in that case where the facilities of PASVA4 will be most useful. We are currently working in this problem where it is necessary to estimate the state binding intervals. Tests made using PASVA4 for the classical linear time invariant fuel optimal control and state constrained control problem([9], which has an analytical solution to be used as a reference, has been successful.

Concerning the real time algorithm, it is clear that many variants of the one presented may be designed. Although the control function given by the algorithm is suboptimal, several test performed for the case of our vehicle yielded a small increase in the consumption compared with the optimal value (less than 3.57%). Model predictive control approaches can also be tested.

Summarizing, is spite this work is still in progress, we think we have done a first step to the design of a real time control strategy for optimizing power management in hybrid vehicles. In addition, we gained experience in the numerical solution of constrained control problems, by means of solving the boundary value problem resultant from the statement of Pontryagin Maximum Principle optimality conditions.

REFERENCES

[1] M. LENTINI, AND V. PEREYRA,*PASVA4: An O.D.E boundary solver for problems with discontinuous interfaces and algebraic parameters*, Mat. Aplic. Comp., V.2. N2 (1983), pp. 103-118.

[2] V.L. PEREYRA,*Solución numérica de ecuaciones diferenciales con valores de frontera*, Acta Científica Venezolana.30(1979), pp. 7-22.

[3] S. DELPRAT, T. M. GUERRA, G. PAGANELLI, J. LAUBER AND M. DELHOM,*Control strategy optimization for an hybrid parallel powertrain*, Proc. of the American Control Conference(2001), pp. 1315-1320.

[4] S. DELPRAT, J. LAUBER, T. M. GUERRA AND J. RIMAUX,*Control of a Parallel Hybrid Powertrain: Optimal Control*, IEEE Transactions on Vehicular Technology, V53, No 3 (2004), pp. 872-881.

[5] L. SERRAO, SIMONA ONORI AND G. RIZZONI,*ECMS as a realization of Pontryagin Minimum Principle for HEV control*, Proc. of the American Control Conference (2009), pp. 3964-3969.

[6] A. BRAHMA, Y. GUEZENNEC AND G. RIZZONI,*Dynamic optimization of mechanical/electric power flow in parallel hybrid electric vehicles*,in Proc. of AVEC 2000, 5th. Intern. Symp. on Advanced Vehicle Control, Ann Arbor, Michigan (2000).

[7] L.V. PÉREZ, AND G.O. GARCÍA,*State constrained optimal control applied to supervisory control in hybrid electric vehicles*, Oil & Gas Science and Technology, Revue de l'Institut Francais du Pétrole, Vol. 65, No. 1 (2010), pp. 191-201. DOI: 10.2516/ogst/2009040.

[8] A. SCIARETTA, AND L. GUZZELLA, *Control of hybrid electric vehicles*, IEEE Control Systems Magazine, Vol.27, No 2 (2007), pp. 60-70, Digital Object Identifier 10.1109/MCS.2007.338280.

[9] H.P. GEERING, *Optimal control with engineering applications*, Springer-Verlag, 2007.

EXISTENCE AND UNIQUENESS OF DISTRIBUTED OPTIMAL CONTROL PROBLEMS GOVERNED BY PARABOLIC VARIATIONAL INEQUALITIES OF THE SECOND KIND

Mahdi Boukrouche[♭] and Domingo A. Tarzia[†]

[♭]*Lyon University, F-42023 Saint-Etienne, Laboratory of Mathematics, University of Saint-Etienne, LaMUSE EA-3989, 23 Docteur Paul Michelon 42023 Saint-Etienne Cedex 2, France. E-mail: Mahdi.Boukrouche@univ-st-etienne.fr*

[†]*Departamento de Matemática-CONICET, FCE, Univ. Austral, Paraguay 1950, S2000FZF Rosario, Argentina. E-mail: DTarzia@austral.edu.ar*

Abstract: Let u_g be the unique solution of a parabolic variational inequality of second kind, with a second member g. Using a regularization method, we prove, for all g_1 and g_2, a monotony property between $\mu u_{g_1} + (1-\mu)u_{g_2}$ and $u_{\mu g_1+(1-\mu)g_2}$ for $\mu \in [0,1]$. This allowed us to prove the existence and uniqueness results to a family of distributed optimal control problems governed by parabolic variational inequalities of second kind over g for each heat transfer coefficient $h > 0$, associated to the Newton law, and of another distributed optimal control problem associated to a Dirichlet boundary condition.

Keywords: *Parabolic variational inequalities of the second kind, convex combination of solutions, monotony property, regularization method, strict convexity of cost functional, optimal control problems, free boundary problem*
2000 AMS Subject Classification: 35R35, 35B37, 35K85, 49J20, 49K20

1 INTRODUCTION

Let consider the following problem governed by the parabolic variational inequality of the second kind

$$\langle \dot{u}(t)\,,\, v-u(t)\rangle + a(u(t)\,,\, v-u(t)) + \Phi(v) - \Phi(u(t)) \geq < g(t)\,,\, v-u(t) > \quad \forall v \in K, \tag{1}$$

a.e. $t \in]0,T[$, with the initial condition

$$u(0) = u_b, \tag{2}$$

where, a is a symmetric, continuous and coercive bilinear form (with λ its positive coerciveness constant) on the Hilbert space $V \times V$, Φ is a proper and convex function from V into $\mathbb{R}$ and is lower semi-continuous for the weak topology on V, $< \cdot, \cdot >$ denotes the duality brackets between V' and V, K is a closed convex non-empty subset of V, u_b is an initial value in another Hilbert space H with V being densely and continuously imbedded in H, and g is a given function in the space $L^2(0,T,V')$. It is well known [4, 5] that, there exists a unique solution

$$u \in \mathcal{C}(0,T,H) \cap L^2(0,T,V) \quad \text{with} \quad \dot{u} = \frac{\partial u}{\partial t} \in L^2(0,T,H)$$

to problem (1)-(2). So we can consider $g \mapsto u_g$ as a function from $L^2(0,T,H)$ to $\mathcal{C}(0,T,H) \cap L^2(0,T,V)$.

In Section 2, we establish in Theorem 1, the error estimate between $u_3(\mu)$ and $u_4(\mu)$, where

$$u_3(\mu) = \mu u_{g_1} + (1-\mu)u_{g_2}, \qquad u_4(\mu) = u_{g_3(\mu)}, \quad \text{with} \quad g_3(\mu) = \mu g_1 + (1-\mu)g_2. \tag{3}$$

This result generalizes our previous result obtained in [3] for the elliptic variational inequalities. We deduce in Corollary 1 a condition on the data to get $u_3(\mu) = u_4(\mu)$ for all $\mu \in [0,1]$.

Let Ω an open bounded set in $\mathbb{R}^N$ with its regular boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2$. We suppose that $\Gamma_1 \cap \Gamma_2 = \emptyset$, and $mes(\Gamma_1) > 0$. Let given a time interval $[0,T]$ for some $T > 0$. Let consider the following free boundary

problem

$$\frac{\partial u}{\partial t} - \Delta u = g, \quad in \quad \Omega\times]0,T[, \tag{4}$$

$$\left.\begin{array}{l} \left|\dfrac{\partial u}{\partial n}\right| < q \Longrightarrow u = 0, \\ \left|\dfrac{\partial u}{\partial n}\right| = q \Longrightarrow \exists k > 0 : \quad u = -k\dfrac{\partial u}{\partial n} \end{array}\right\} \quad on \quad \Gamma_2\times]0,T[, \tag{5}$$

$$u = b \quad on \quad \Gamma_1\times]0,T[, \tag{6}$$

with the initial boundary condition (2) and the compatibility condition

$$u_b = b \quad on \quad \Gamma_1\times]0,T[, \tag{7}$$

where n is the unit outward normal to Γ_2, g is the external force, b is given on Γ_1, (5) is the Tresca type boundary condition (for more description see [1],[2],[5]) and q is the Tresca friction coefficient on Γ_2.

We consider a family of free boundary problems (4)-(5) with the intitial condition (2), where the Dirichlet boundary condition (6) is replaced by the following Robin condition which depends on a parameter $h > 0$

$$-\frac{\partial u}{\partial n} = h(u - b) \quad on \quad \Gamma_1\times]0,T[. \tag{8}$$

Let $H = L^2(\Omega)$, $V = H^1(\Omega)$. Let

$$V_0 = \{v \in V : \quad v_{|\Gamma_1} = 0\}, \quad and \quad K = \{v \in V : \quad v_{|\Gamma_1} = b\}.$$

So we consider the following variational problems:

Problem (P) Let given $b \in L^2(]0,T[\times\Gamma_1)$, $g \in L^2(0,T,H)$ and $q \in L^2(]0,T[\times\Gamma_2)$, $q > 0$. Find u in $\mathcal{C}([0,T],H) \cap L^2(0,T,K)$ solution of the parabolic variational inequality (1), where $< \cdot,\cdot >$ is the scalar product $(\cdot,\cdot)$ in H, with the initial condition (2), where $a(u,v) = \int_\Omega \nabla u \nabla v dx$ and $\Phi(v) = \int_{\Gamma_2} q|v|ds$.

Problem (P_h) Let given $b \in L^2(]0,T[\times\Gamma_1)$, $g \in L^2(0,T,H)$ and $q \in L^2(]0,T[\times\Gamma_2)$, $q > 0$. For all coefficient $h > 0$, find $u \in \mathcal{C}(0,T,H) \cap L^2(0,T,V)$ solution of the parabolic variational inequality

$$\begin{aligned} \langle \dot{u}(t)\,,\, v - u(t)\rangle + a_h(u(t)\,,\, v - u(t)) + \Phi(v) - \Phi(u(t)) \geq (g(t), v - u(t)) \\ + h\int_{\Gamma_1} b(t)(v - u(t))ds \quad \forall v \in V, \end{aligned} \tag{9}$$

and the initial condition (2), where $a_h(u,v) = a(u,v) + h\int_{\Gamma_1} uvds$.

We know that there exists $\lambda > 0$ such that $\lambda\|v\|_V^2 \leq a(v\,,\, v), \forall v \in V_0$. Moreover, it follows from [11] that there exists $\lambda_1 > 0$ such that $a_h(v,v) \geq \lambda_h\|v\|_V^2, \forall v \in V,$ with $\lambda_h = \lambda_1 \min\{1\,,\, h\}$ so a_h is a bilinear, continuous, symmetric and coercive form on V. Therefore, there exists an unique solution to each of the two problems (P) and (P_h). Then we can consider [7, 8] the cost functional J defined by

$$J(g) = \frac{1}{2}\|u_g\|^2_{L^2(0,T,H)} + \frac{M}{2}\|g\|^2_{L^2(0,T,H)}, \tag{10}$$

where M is a positive constant, and u_g is the unique solution to (1)-(2), corresponding to the control g. One of our main purposes is to prove the existence and uniqueness of the distributed optimal control problem:

$$\text{Find } g_{op} \in L^2(0,T,H) \quad \text{such that} \quad J(g_{op}) = \min_{g\in L^2(0,T,H)} J(g). \tag{11}$$

This can be reached if we prove the strictly convexity of the cost functional J, which follows from the following monotony property : for any two control g_1 and g_2 in $L^2(0,T,H)$,

$$u_4(\mu) \leq u_3(\mu) \qquad \forall \mu \in [0,1]. \tag{12}$$

In Section 3, by using a regularization method, we prove in Theorem 2 this monotony property (12), for the solutions of the two problems (P) and (P_h). This result with a new proof and simplified, generalizes that obtained by [10] for elliptic variational inequalities.

In Section 4, we also consider the family of distributed optimal control problems $(P_h)_{h>0}$:

$$\text{Find } g_{op_h} \in L^2(0,T,H) \quad \text{such that} \quad J(g_{op_h}) = \min_{g \in L^2(0,T,H)} J_h(g), \tag{13}$$

with the cost functional

$$J_h(g) = \frac{1}{2}\|u_{g_h}\|^2_{L^2(0,T,H)} + \frac{M}{2}\|g\|^2_{L^2(0,T,H)}, \tag{14}$$

where u_{g_h} is the unique solution of (9) and (2), corresponding to the control g for each $h > 0$.

We prove the strict convexity of the cost functional (10) and (14), associated to the distributed optimal control problems (11) and (13) respectively by using the crucial property of monotony (12) (see Theorem 2). Then, the corresponding existence and uniqueness of solutions to optimal control problems (11) and (13) follows from [8].

This paper generalizes the results obtained in [6, 10] for elliptic variational equalities, and in [9] for parabolic variational equalities, to the case of parabolic variational inequalities of second kind.

2 Error estimate for convex combinations of solutions

Theorem 1 *Let u_1 and u_2 be two solutions of the parabolic variational inequality* (1) *with the same initial condition, and corresponding to the two control g_1 and g_2 respectively. We have the following estimate*

$$\begin{aligned}\frac{1}{2}\|u_4(\mu) - u_3(\mu)\|^2_{L^\infty(0,T,H)} + \lambda\|u_4(\mu) - u_3(\mu)\|^2_{L^2(0,T,V)} + \mu\mathcal{I}_{14}(\mu)(T) + (1-\mu)\mathcal{I}_{24}(\mu)(T)\\ +\mu\Phi(u_1) + (1-\mu)\Phi(u_2) - \Phi(u_3(\mu)) \leq \mu(1-\mu)(\mathcal{A}(T,g_1) + \mathcal{B}(T,g_2)) \quad \forall \mu \in [0,1],\end{aligned}$$

where

$$\mathcal{I}_{j4}(\mu)(T) = \int_0^T I_{j4}(\mu)(t)dt \quad \text{for } j = 1,2, \quad \mathcal{A}(T,g_1) = \int_0^T \alpha(t)dt, \quad \mathcal{B}(T,g_2) = \int_0^T \beta(t)dt,$$

$$I_{j4}(\mu) = \langle \dot{u}_j\,,\, u_4(\mu) - u_j\rangle + a(u_j\,,\, u_4(\mu) - u_j) + \Phi(u_4(\mu)) - \Phi(u_j) - \langle g_j, u_4(\mu) - u_j\rangle \geq 0,$$

$$\alpha = \langle \dot{u}_1\,,\, u_2 - u_1\rangle + a(u_1\,,\, u_2 - u_1) + \Phi(u_2) - \Phi(u_1) - \langle g_1, u_2 - u_1\rangle \geq 0, \tag{15}$$

$$\beta = \langle \dot{u}_2\,,\, u_1 - u_2\rangle + a(u_2\,,\, u_1 - u_2) + \Phi(u_1) - \Phi(u_2) - \langle g_2, u_1 - u_2\rangle \geq 0. \tag{16}$$

Corollary 1 *From Theorem* 1 we get $a.e.\, t \in [0,T]$:

$$\mathcal{A}(T,g_1) = \mathcal{B}(T,g_2) = 0 \Rightarrow \begin{cases} u_3(\mu) = u_4(\mu) & \forall \mu \in [0,1], \\ I_{14}(\mu) = I_{24}(\mu) = 0 & \forall \mu \in [0,1], \\ \Phi(u_3(\mu)) = \mu\Phi(u_1) + (1-\mu)\Phi(u_2) & \forall \mu \in [0,1]. \end{cases}$$

Lemma 1 *Let u_1 and u_2 be two solutions of the parabolic variational inequality of second kind* (1) *with respectively as second member g_1 and g_2, then we get*

$$\|u_1 - u_2\|^2_{L^\infty(0,T,H)} + \lambda\|u_1 - u_2\|^2_{L^2(0,T,V)} \leq \frac{1}{\lambda}\|g_1 - g_2\|^2_{L^2(0,T,V')}. \tag{17}$$

3 ON THE PROPERTY OF MONOTONY

Theorem 2 *For any two control g_1 and g_2 in $L^2(0,T,H)$, it holds that*

$$u_4(\mu) \leq u_3(\mu) \quad in \quad \Omega \times [0,T], \quad \forall \mu \in [0,1]. \tag{18}$$

Here $u_4(\mu) = u_{\mu g_1 + (1-\mu) g_2}$, $u_3(\mu) = \mu u_{g_1} + (1-\mu) u_{g_2}$, $u_1 = u_{g_1}$ and $u_2 = u_{g_2}$ are the unique solutions of the variational problem P, with $g = g_1$ and $g = g_2$ respectively, and for the same q, and the same initial condition (2)*. Moreover, it holds also that*

$$u_{h4}(\mu) \leq u_{h3}(\mu) \quad in \quad \Omega \times [0,T], \quad \forall \mu \in [0,1]. \tag{19}$$

Here $u_{4h}(\mu) = u_{\mu g_{1h} + (1-\mu) g_{2h}}$, $u_{3h}(\mu) = \mu u_{g_{1h}} + (1-\mu) u_{g_{2h}}$, $u_{1h} = u_{g_{1h}}$ and $u_{h2} = u_{g_{h2}}$ are the unique solutions of the variational problem P_h, with $g = g_1$ and $g = g_2$ respectively, and for the same q, h, b and the same initial condition (2).

Proof. Because the functional Φ is not differentiable we use the regularization method by considering for $\varepsilon > 0$ the following approach $\Phi_\varepsilon(v) = \int_{\Gamma_2} q\sqrt{\varepsilon^2 + |v|^2} ds, \forall v \in V$, which is Gateaux differentiable and then we consider the limit when $\varepsilon \to 0$ by using [5, 12].

□

4 OPTIMAL CONTROL PROBLEMS

Lemma 2 *Assume that $g \geq 0$ in $\Omega \times]0,T[$, $b \geq 0$ on $\Gamma_1 \times]0,T[$, $u_b \geq 0$ in Ω. Then as $q > 0$, we have $u_g \geq 0$. Assuming again that $h > 0$, then $u_{g_h} \geq 0$ in $\Omega \times]0,T[$.*

Theorem 3 *Assume the same hypotheses of Lemma 2. Then J and J_h, defined by (10) and (14) respectively, are strictly convex applications on $L^2(0,T,H)$, so there exist unique solutions g_{op} and g_{op_h} in $L^2(0,T,H)$ respectively to the distributed optimal control problems* (11) *and* (13) *for all $h > 0$.*

ACKNOWLEDGEMENTS

This work was realized while the second author was a visitor at Saint Etienne University (France) and he is grateful to this institution for his hospitality . It was partially supported by PIP No. 0460 from CONICET-UA and Grant FA9550-10-1-0023, Rosario, Argentina.

REFERENCES

[1] M. BOUKROUCHE AND I. CIUPERCA *Asymptotic behaviour of solutions of lubrication problem in a thin domain with a rough boundary and Tresca fluid-solid interface law.* Quart. Appl. Math. 64 (2006), pp. 561-591.

[2] M. BOUKROUCHE AND F. SAIDI *Non-isothermal lubrication problem with Tresca fluid-solid interface law. Partie I.* Nonlinear Anal. Real World Applications 7 (2006), pp. 1145-1166.

[3] M. BOUKROUCHE AND D. A. TARZIA, *On a convex combination of solutions to elliptic variational inequalities.* Electro. J. Differential Equations 2007(2007), No. 31, pp. 1-10.

[4] H. BRÉZIS, *Problèmes unilatéraux.* J.Math. Pure et Appl. 51 (1972), pp. 1-162.

[5] G. DUVAUT AND J.L. LIONS, *Les Inéquations en Mécanique et en Physique.* Dunod, 1972.

[6] C.M. GARIBOLDI AND D.A. TARZIA, *Convergence of distributed optimal controls on the internal energy in mixed elliptic problems when the heat transfer coefficient goes to infinity.* Appl. Math. Optim. 47 (3) (2003), pp. 213-230. See also, *A new proof of the convergence of distributed optimal controls on the internal energy in mixed elliptic problems*, MAT - Serie A, 7 (2004), pp. 31-42.

[7] S. KESAVAN AND T. MUTHUKUMAR, *Low-cost control problems on perforated and non-perforated domains*, Proc. Indian Acad. Sci. (Math. Sci.) 118, No. 1, (2008) pp. 133-157.

[8] J.L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles.* Dunod Paris (1968).

[9] J.L. MENALDI AND D. A. TARZIA, *A distributed parabolic control with mixed boundary conditions.* Asymptotic Analysis 52 (2007), pp. 227-241.

[10] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques.* J. Functional Analysis 22 (1976), no. 2, pp. 130-185.

[11] D. A. TARZIA, *Una familia de problemas que converge hacia el caso estacionario del problema de Stefan a dos fases*, Math. Notae 27 (1979), pp. 157-165.

[12] D. A. TARZIA, *Etude de l'inéquation variationnelle proposée par Duvaut pour le problème de Stefan à deux phases, I*, Boll. Unione Mat. Italiana 1B (1982), pp. 865-883.

MÉTODOS DE HACES APLICADO A LA COORDINACIÓN HIDROTÉRMICA DE CORTO PLAZO CONSIDERANDO RESTRICCIONES AC.

Aldo J. Rubiales†‡, Pablo A. Lotito†‡, Lisandro Parente‡ y Fernando J. Mayorano†‡

† PLADEMA , Universidad Nacional del Centro de la Provincia de Buenos Aires
‡CONICET

Resumen: En el presente trabajo se muestra una nueva metodología para abordar el problema de coordinación hidrotérmica a corto plazo (STHTC). La resolución de este tipo de problemas comprende tanto el pre-despacho (Unit Commitment), como el despacho económico de las unidades térmicas e hidráulicas en forma integral para un horizonte de tiempo semanal o diario con paso horario. Para la resolución de este problema se propone utilizar la Descomposición Generalizada de Benders para descomponer el problema original en un problema maestro y un subproblema, de manera que el primero proponga los despachos considerando variables enteras y el segundo controle la factibilidad eléctrica del despacho propuesto considerando una linealización de las restricciones que surgen al considerar las características de la red. A su vez se propone una mejora a esta técnica utilizando la metodología de haces introducida por Lemarechal y Sagastizábal.

Palabras claves: *Coordinación Hidrotérmica a corto plazo, Unit Commitment, Método de Haces, Descomposición Generalizada de Benders*

1. INTRODUCCIÓN

En el presente trabajo se muestra una nueva metodología para abordar el problema de coordinación hidrotérmica a corto plazo (STHTC). Generalmente en este tipo de problemas, la red de transmisión no es considerada o se encuentra modelada de manera simplificada. Esto hace que en países con redes extensas y/o débilmente malladas (como la de la mayoría de los países de Latinoamérica), la solución a este problema luego debe ser corregida mediante un flujo de potencia AC para poder ser aplicable. En la metodología aplicada en el presente trabajo, se realiza un modelado mucho más exacto de la parte eléctrica evitando de esta manera las correcciones post-depacho y logrando mejores soluciones que las propuestas por la resolución desacopladas de los problemas de pre-despacho, despacho económico y flujo de potencia AC.
Este problema ha sido estudiado a lo largo de los años considerando distintas definiciones y aplicando distintos métodos de resolución. Las formulaciones más sencillas de este problema y que fueron el punto de partida para este campo de investigación, consideraban modelos sencillos que no se correspondían con las características reales de los sistemas eléctricos.
En se presenta uno de los primeros enfoques (que por su simpleza es sólo de índole académica) el cual solo utiliza unidades térmicas y se basa en lista de orden de mérito. Es decir, las unidades se despachaban en orden creciente de costos por unidad de energía producida. Este procedimiento se diferencia bastante de la realidad ya que no se tienen en cuenta restricciones intertemporales (como los tiempos mínimos de operación de las unidades térmicas o la consideración de los costos de arranque), o el hecho de que no siempre las unidades térmicas de generación operan a potencia constante.

La utilización de Programación Dinámica aplicada a este problema fue también mencionado en y a pesar de que permite modelar problemas no-lineales, no-convexos, por su naturaleza combinatorial solo se puede considerar un número reducido de unidades térmicas si se desea tener tiempos razonables de cálculo. En se presenta este problema considerando la aplicación a sistemas que poseen centrales hidroeléctricas de bombeo. En este articulo se menciona el problema de la dimensionalidad y sugiere para su resolución el enfoque presentado en.

Uno de los primeros enfoques que se utilizó para descomponer el problema de STHTC fue la Relajación Lagrangeana permitiendo dividir el problema original en muchos problemas de menor dimensión y de

resolución más sencilla. Más adelante se describirá en detalle los problemas a los que se aplicaron este método de resolución y los principales inconvenientes que se encontraron al aplicarla.

2. DESCOMPOSICIÓN PROPUESTA

Para la posible minimización de la función f (y) y debido a la complejidad que el problema que esta representa, el problema de minimización debe ser descompuesto y queda definido de la siguiente manera:

Donde fm(ym) representa la función objetivo del problema maestro y ϕ(ym) el modelo que aproxima el subproblema evaluado en el punto ym. Cabe destacar que se definen ym e ysp como el conjunto de valores candidatos asociados a las variables del problema maestro y del subproblema respectivamente. A medida que suceden las iteraciones, el algoritmo va agregando cortes al modelo del subproblema y va generando una solución mas aproximada al valor real de la función objetivo del subproblema.

En el esquema de descomposición definido para este problema se intenta balancear las comple-jidades del problema maestro y del subproblema definiendo que restricciones se asocian a cada uno de ellos. Como se observa en la función objetivo del problema maestro est´a asociada a la suma de los costos de encendido de las centrales térmicas y al coeficiente independiente de los costos cuadráticos asociados a la generación de las centrales térmicas
Donde ym representa el conjunto de variables que se fijan en el problema maestro en cada iteración y son pasadas al subproblema. En el presente trabajo este conjunto de variables está dado por la potencia activa hidráulica de cada unidad pht,i, las variables binarias asociadas al estado de ambos tipos de unidades utt,i y uht,i, y las variables asociadas a los prendidos y apagados de las unidades térmicas stt,i y ett,i. O sea:

$$f_{maestro}(y_m) = \sum_t \sum_i C_i ut_{t,i} + D_i st_{t,i}$$

de las cuales utt,i y pht,h se pasan al subproblema, y uht,h, stt,i y ett,i solo se usan en el maestro.

A continuación se detallan todas las restricciones asociadas al problema maestro:

$$ut_{t,i} - ut_{t-1,i} = st_{t,i} - et_{t,i}$$
$$st_{t,i} + et_{t,i} \leq 1$$
$$ut_{t,i} + ut_{t-1,i} + ... + ut_{t+on_i^{Low}-1,i} \geq st_{t,i} on_i^{LOW}$$
$$(1 - ut_{t,i}) + (1 - ut_{t-1,i}) + ... + (1 - ut_{t+off_i^{Low}-1,i}) \geq et_{t,i} off_i^{LOW}$$
$$uh_{t,i} ph_j^{LOW} \leq ph_{t,j} \leq uh_{t,i} ph_j^{UP}$$
$$ph_{t,j} = q_{t,j}^T \beta_j$$
$$a_{t+1,r} = a_{t,r} + \Delta T (q_{t,r}^I - q_{t,r}^T - q_{t,r}^S)$$
$$a_r^{LOW} \leq a_{t,r} \leq a_r^{UP}$$

Debido a la distinta naturaleza de las variables que forman parte del problema maestro, se decidió generar un τ para cada variable.

La función objetivo del subproblema comprende el termino cuadrático y lineal de los costos de generación de potencia activa y considera los costos asociados a las penalizaciones por déficit o exceso de potencia activa o reactiva en cada una de las barras.
Las restricciones que se consideran en el subproblema, se corresponden con:

- Las de caja de potencia activa para todas las centrales térmicas
- Las de caja de potencia reactiva para todas las centrales
- Las de balance de potencia activa y reactiva en cada barra

- Las capacidades límites de las líneas de transmisión
- Los niveles de tensión requeridos en cada barra
- Las asociadas a los compensadores de potencia reactiva

Como se puede observar en este caso el problema maestro es computacionalmente más caro que el subproblema. El problema maestro se corresponde con un problema de programación cuadrática con restricciones lineales que también tiene restricciones asociadas a variables enteras. Este tipo de problemas ha sido estudiado por mucho tiempo y en la actualidad existen varios solvers comerciales con probada eficiencia en resolver este tipo de problemas. En este trabajo se utiliza el lenguaje algebraico de modelado GAMS y los solvers que el mismo provee para resolver tanto el problema maestro como el subproblema subproblema.

REFERENCIAS

[1] AUBIN, J. P., MATHEMATICAL METHODS OF GAME AND ECONOMIC THEORY. ELSEVIER, 1980.

[2] CONTRERAS. J, KLUSCH, M., J.B. KRAWCZYK, J.B., NUMERICAL SOLUTIONS TO NASH-COURNOT EQUILIBRIA IN COUPLED CONSTRAINT ELECTRICITY MARKETS. IEEE TRANSACTIONS ON POWER SYSTEMS, V.19, N.1, 2004, P. 195-206.

[3] KRAWCZYK, J. B., URYASEV, S., RELAXATION ALGORITHMS TO FIND NASH EQUILIBRIA WITH ECONOMIC APPLICATIONS. ENVIRONMENTAL MODELING AND ASSESSMENT, VOL. 5, 2000, P. 63–73.

[4] MAIORANO, A., SONG, Y. H., AND TROVATO, M., DYNAMICS OF NONCOLLUSIVE OLIGOPOLISTIC ELECTRICITY MARKETS". PROC. IEEE POWER ENG. SOC. WINTER MEETING, SINGAPORE, 2000, P.838-844.

[5] MOITRE, D., NASH EQUILIBRIA IN COMPETITIVE ELECTRIC ENERGY MARKETS. INTERNATIONAL JOURNAL OF ELECTRIC POWER SYSTEMS RESEARCH, ELSEVIER, U.K. VOL 60/3, 2002, P. 153-160.

[6] MOITRE, D., SAUCHELLI, V., Y GARCÍA, G, OPTIMIZACIÓN DINÁMICA BINIVEL DE CENTRALES HIDROELÉCTRICAS DE BOMBEO EN UN POOL COMPETITIVO - PARTE I: MODELO Y ALGORITMO. REVISTA IEEE AMÉRICA LATINA. V.3, N.2, 2005A, P. 62 – 67.

[7] MOITRE, D., SAUCHELLI, V., Y GARCÍA, G., OPTIMIZACIÓN DINÁMICA BINIVEL DE CENTRALES HIDROELÉCTRICAS DE BOMBEO EN UN POOL COMPETITIVO – PARTE II: CASOS DE ESTUDIO. REVISTA IEEE AMÉRICA LATINA. , V.3, N.2, 2005B, P.68 – 74.

[8] VEGA, MAURICIO A. Y VILLENA, MAURICIO G. "EL MERCADO HIDROTÉRMICO CHILENO: UN ENFOQUE DE TEORÍA DE JUEGOS", CUADERNOS DE ECONOMÍA, V. XXV, N. 45, BOGOTÁ, 2006, PÁGINAS 155-203.

[9] J. P. AUBIN, MATHEMATICAL METHODS OF GAME AND ECONOMIC THEORY. AMSTERDAM, THE NETHERLANDS: ELSEVIER, 1980.

[10] NUMERICAL SOLUTIONS TO NASH-COURNOT EQUILIBRIA IN COUPLED CONSTRAINT ELECTRICITY MARKETS

[11] CONTRERAS, J. KLUSCH, M. KRAWCZYK, J.B.

[12] S. URYASEV AND R. Y. RUBINSTEIN, "ON RELAXATION ALGORITHMS IN COMPUTATION OF NONCOOPERATIVE EQUILIBRIA," IEEE TRANS. AUTOMAT. CONTR., VOL. 39, PP. 1263–1267, JUNE 1994.

[13] J. B. KRAWCZYK AND S. URYASEV, "RELAXATION ALGORITHMS TO FIND NASH EQUILIBRIA WITH ECONOMIC APPLICATIONS," ENVIRONMENTAL MODELING AND ASSESSMENT, VOL. 5, PP. 63–73, 2000.

[14] [10] J. B. ROSEN, J. B., , "EXISTENCE AND UNIQUENESS OF EQUILIBRIUM POINTS FOR CONCAVE N-PERSON GAMES," . ECONOMETRICA, VOL. 33, 1965, PP. 520–534,. 1965.

ESTUDIO DE LA NO NEGATIVIDAD DE SISTEMAS SINGULARES DE CONTROL VIA REALIMENTACIONES

Alicia Herrero y Néstor Thome

Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, España. {*aherrero,njthome*}*@mat.upv.es*

Resumen: En este artículo se analizan sistemas singulares de control atendiendo a las propiedades de regularidad y de no negatividad. Este tipo de sistemas involucran una matriz de estados singular, que en este trabajo se supondrá de índice 1. Esta hipótesis garantiza la existencia de una forma específica de las matrices de estado en el caso en que el proyector de grupo asociado sea no negativo. Esta forma de las matrices de estado permitirá abordar el objetivo principal del trabajo que consiste en buscar una realimentación de estados de modo que el sistema en lazo cerrado cumpla las condiciones de regularidad y de no negatividad. A tal efecto, se diseña un algoritmo y se da la justificación teórica del mismo.

Palabras clave: *Sistema de control, no negatividad, realimentación*
2000 AMS Subject Classification: 15A09 - 93C05

1. INTRODUCCIÓN Y PRELIMINARES

En este artículo se trabajará con sistemas singulares de control del tipo

$$\left\{ \begin{array}{rcl} Ex(k+1) & = & Ax(k)+Bu(k) \\ y(k) & = & Cx(k) \end{array} \right. \tag{1}$$

donde $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $x(k) \in \mathbb{R}^{n\times 1}$, $u(k) \in \mathbb{R}^{m\times 1}$, e $y(k) \in \mathbb{R}^{p\times 1}$ con $\mathrm{rg}(E) = r < n$. En general, este sistema se denota por (E, A, B, C) o bien por (E, A, C) cuando $B = O$. Un amplio análisis de los sistemas singulares se puede encontrar en el libro de Duan [2] en el que se recogen desde sus propiedades estructurales hasta un estudio sobre la regularización y la asignación de polos.

Además se supondrá que la matriz E tiene índice 1 y que su proyector de grupo es no negativo. Se recuerda que una matriz $E \in \mathbb{R}^{n\times n}$ tiene índice 1 cuando $\mathrm{rg}(E) = \mathrm{rg}(E^2)$. Este hecho garantiza la existencia de la matriz inversa de grupo de E, representada por $E^{\#}$, que es la única solución del sistema matricial: $EXE = E$, $XEX = X$ y $EX = XE$. También es conocido que el proyector de grupo de una matriz $E \in \mathbb{R}^{n\times n}$ viene dado por $EE^{\#}$ suponiendo que $E^{\#}$ existe. Por otra parte, se dice que una matriz es no negativa si todos sus elementos son no negativos y se denota por $E \geq O$.

Bajo las condiciones indicadas anteriormente, la matriz E se puede escribir de la forma:

$$PEP^t = \left[\begin{array}{c} I \\ O \\ N \end{array} \right] XTY \left[\begin{array}{ccc} I & M & O \end{array} \right] =: \Psi(XTY, M, N) \tag{2}$$

donde P es una matriz de permutación y P^t su traspuesta, $T \in \mathbb{R}^{r\times r}$ es una matriz invertible, $X = \mathrm{diag}(x_1, x_2, \ldots, x_r)$, $Y = \mathrm{diag}(y_1^t, y_2^t, \ldots, y_r^t)$, x_i e y_i son vectores unitarios positivos tales que $YX = I$, M, N son matrices no negativas de tamaños $q \times s$ y $t \times q$, respectivamente [4, 5]. Se observa que la división en bloques se realiza de acuerdo a la partición $n \times n = (q + s + t) \times (q + s + t)$.

De este modo, el sistema singular (E, A, B, C) puede ser transformado mediante el cambio de variables $z(k) = Px(k)$ en el sistema equivalente $(\widetilde{E}, \widetilde{A}, \widetilde{B}, \widetilde{C})$ donde

$$\widetilde{E} = PEP^t, \quad \widetilde{A} = PAP^t, \quad \widetilde{B} = PB, \quad \widetilde{C} = CP^t. \tag{3}$$

Claramente alguna propiedades de E son heredadas por la matriz $\widetilde{E}$, por ejemplo, $\widetilde{E}$ tiene índice 1 y su proyector de grupo es no negativo.

Sobre este último sistema se realizará una realimentación de estados del tipo $u(k) = Fz(k)$ de modo que el sistema en lazo cerrado $(\widetilde{E}, \widetilde{A}+\widetilde{B}F, \widetilde{C})$ cumpla la condición de regularidad y sea no negativo. Se recuerda que un sistema (E, A, B, C) satisface la condición de regularidad si existe un escalar λ tal que $\det(\lambda E - A) \neq 0$, lo que garantiza la existencia y unicidad de su solución [6]. Además un sistema (E, A, B, C) se dice que es no negativo cuando a una condición inicial y controles no negativos corresponden estados y salidas no negativas.

2. Algoritmo para la construcción de la realimentación

En esta sección se presenta un algoritmo en el que se analiza la existencia de la realimentación y su construcción. Para ello, se realiza una partición en bloques de la matriz

$$\widetilde{C} = CP^t = \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \tag{4}$$

teniendo en cuenta los tamaños de los bloques de la matriz $\Psi(XTY, M, N)$ dada en (2).

Algoritmo.

Entrada: Sistema singular (E, A, B, C) que satisface $EE^{\#} \geq O$.

Salidas: La matriz F tal que el sistema en lazo cerrado $(\widetilde{E}, \widetilde{A}+\widetilde{B}F, \widetilde{C})$ es no negativo y satisface la condición de regularidad y la solución de dicho sistema.

Paso 1: Transformar el sistema original (E, A, B, C) en el sistema equivalente $(\widetilde{E}, \widetilde{A}, \widetilde{B}, \widetilde{C})$ dado en (3).

Paso 2: Si $(C_1 + C_3N)X < O$ entonces *'No existe una matriz F para que el sistema en lazo cerrado sea no negativo'*. Ir al Fin.

Paso 3: Elegir un escalar β tal que $T^{-1} - \beta I \geq O$ y una $\{1\}$-inversa generalizada $\widetilde{B}^-$.

Paso 4: Si $(I - \widetilde{B}\widetilde{B}^-)(I - \beta\widetilde{E} - \widetilde{A}) \neq O$ entonces volver al Paso 3 o ir al Fin.

Paso 5: Construir la matriz F como $F = \widetilde{B}^-(I - \beta\widetilde{E} - \widetilde{A}) + (I - \widetilde{B}^-\widetilde{B})Y$ con Y una matriz arbitraria.

Paso 6: Por tanto *'El sistema en lazo cerrado es no negativo y satisface la condición de regularidad'*. Las salidas del sistema $(\widetilde{E}, \widetilde{A} + \widetilde{B}F, \widetilde{C})$ vienen dadas por

$$y(k) = (C_1 + C_3N)X(T^{-1} - \beta I)^k Y \begin{bmatrix} I & M & O \end{bmatrix} z(0).$$

Fin

Si al elegir un escalar β y una matriz $\widetilde{B}^-$ en el paso 3, la condición $(I - \widetilde{B}\widetilde{B}^-)(I - \beta\widetilde{E} - \widetilde{A}) = O$ no se cumple, se debe realizar una nueva elección de ellos. Es posible fijar uno de ellos y cambiar el otro o bien cambiar los dos simultáneamente. En cualquier caso, un criterio de parada puede ser que este paso se realice un número finito de veces.

Por otro lado, la matriz Y del paso 5 se puede elegir de manera arbitraria lo que produce todas las posibles realimentaciones para el valor de β y la matriz $\widetilde{B}^-$ fijados. Si sólo se requiere calcular una realimentación basta tomar $Y = O$, pero si se quieren hallar otras realimentaciones puede hacerse variando esta matriz Y.

Por último, la condición inicial $z(0)$ que aparece en la solución del sistema realimentado en el paso 6 debe elegirse de entre las condiciones admisibles del sistema, como se indicará en la próxima sección.

3. Justificación del algoritmo

Como se ha indicado anteriormente, el sistema original (E, A, B, C) dado en (1) se puede convertir en el sistema equivalente $(\widetilde{E}, \widetilde{A}, \widetilde{B}, \widetilde{C})$ mediante la transformación $z(k) = Px(k)$ indicada en (3). De este modo, mediante la realimentación $u(k) = Fz(k)$, se obtiene el sistema en lazo cerrado $(\widetilde{E}, \widetilde{A} + \widetilde{B}F, \widetilde{C})$.

Con el objetivo de encontrar una matriz adecuada F, ésta se elegirá de manera que

$$\widetilde{A} + \widetilde{B}F = I - \beta\widetilde{E} \tag{5}$$

para que el sistema en lazo cerrado cumpla la condición de regularidad. La ecuación matricial $\widetilde{B}F = I - \beta\widetilde{E} - \widetilde{A}$ tiene solución si y sólo si [1]

$$\widetilde{B}\widetilde{B}^{-}(I - \beta\widetilde{E} - \widetilde{A}) = I - \beta\widetilde{E} - \widetilde{A} \qquad \Longleftrightarrow \qquad (I - \widetilde{B}\widetilde{B}^{-})(I - \beta\widetilde{E} - \widetilde{A}) = O.$$

Cuando esta última condición se satisface entonces la forma general de la matriz F viene dada por

$$F = \widetilde{B}^{-}(I - \beta\widetilde{E} - \widetilde{A}) + (I - \widetilde{B}^{-}\widetilde{B})Y$$

siendo Y una matriz arbitraria de tamaño adecuado.

Además, se podrá asegurar la no negatividad de este sistema mediante las condiciones: $\widetilde{E}\widetilde{E}^{\#} \geq O$, $\widetilde{E}^{\#}(\widetilde{A} + \widetilde{B}F) \geq O$ y $\widetilde{C}\widetilde{E}\widetilde{E}^{\#} \geq O$ [3]. La primera de estas tres condiciones se verifica por hipótesis, con lo que sólo es necesario analizar las otras dos. Para estudiar la segunda utilizaremos (2), (3) y (5). Se puede comprobar que $\widetilde{E}^{\#} = \Psi(XT^{-1}Y, M, N)$ con lo que la segunda condición resulta

$$\Psi(XT^{-1}Y, M, N)[I - \beta\Psi(XTY, M, N)] \geq O. \tag{6}$$

A continuación se indican propiedades generales de la función

$$\Psi : \mathbb{R}^{q\times q} \times \mathbb{R}^{q\times s} \times \mathbb{R}^{t\times q} \to \mathbb{R}^{n\times n}$$

definida a partir de (2) por

$$\Psi(K, M, N) = \begin{bmatrix} I \\ O \\ N \end{bmatrix} K \begin{bmatrix} I & M & O \end{bmatrix}$$

donde las matrices M, N y K tienen tamaños adecuados de modo que las operaciones entre ellas estén bien definidas. A partir de esta definición y de las propiedades de las operaciones de matrices por bloques se deducen las propiedades enunciadas en el siguiente resultado.

Lema 1 *La función Ψ cumple las siguientes propiedades:*

(a) $\Psi(K_1 + K_2, M, N) = \Psi(K_1, M, N) + \Psi(K_2, M, N)$.

(b) $\Psi(\alpha K, M, N) = \alpha\Psi(K, M, N)$ *para cualquier* $\alpha \in \mathbb{R}$.

(c) $\Psi(K_1K_2 \ldots K_l, M, N) = \Psi(K_1, M, N)\Psi(K_2, M, N) \ldots \Psi(K_l, M, N)$.

A partir de las propiedades indicadas en el Lema, la desigualdad (6) se reduce a

$$\Psi(X(T^{-1} - \beta I)Y, M, N) \geq O.$$

donde se ha tenido en cuenta que $YX = I$. Esta desigualdad implica que el bloque $(1,1)$ de $\Psi(X(T^{-1} - \beta I)Y, M, N)$ ha de ser no negativo, es decir que $X(T^{-1} - \beta I)Y \geq O$. Puesto que $X, Y \geq O$ y también $YX = I$ se tiene que $T^{-1} - \beta I \geq O$. Notar que es el único bloque que aporta información sobre la no negatividad al ser M y N matrices no negativas.

La tercera condición que se debe analizar para asegurar la no negatividad del sistema es $\widetilde{C}\widetilde{E}\widetilde{E}^{\#} \geq O$, lo que implica

$$\widetilde{C}\Psi(XTY, M, N)\Psi(XT^{-1}Y, M, N) = \widetilde{C}\Psi(XY, M, N) \geq O.$$

Particionando en bloques la matriz $\widetilde{C}$ de acuerdo con los tamaños de los bloques de $\Psi(XY, M, N)$ como en (4) se tiene que

$$\begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \begin{bmatrix} I \\ O \\ N \end{bmatrix} XY \begin{bmatrix} I & M & O \end{bmatrix} \geq O.$$

De nuevo el bloque $(1,1)$ de la última matriz es el que contiene la información necesaria para asegurar la no negatividad de toda la matriz. Esta condición se transforma en $(C_1 + C_3N)X \geq O$ teniendo en cuenta que $X \geq O$ y que $YX = I$.

Hasta este momento se han encontrado las condiciones necesarias y suficientes para que el sistema en lazo cerrado $(\widetilde{E}, \widetilde{A} + \widetilde{B}F, \widetilde{C})$ cumpla la condición de regularidad y sea no negativo, justificando de esta manera del Paso 1 hasta el Paso 5 del algoritmo.

Por último se presenta la solución explícita de dicho sistema mostrando además que efectivamente es no negativa bajo las condiciones encontradas. Como $\widetilde{E}$ tiene índice 1, la inversa de Drazin se reduce a la inversa de grupo, y la solución del sistema realimentado es $y(k) = \widetilde{C}z(k)$ donde $z(k)$ queda de la siguiente manera [6]:

$$z(k) = (\widetilde{E}^{\#}(I - \beta\widetilde{E}))^k \widetilde{E}^{\#}\widetilde{E}z(0)$$

siendo $z(0) \in \mathrm{Im}\left[\begin{array}{cc} \widetilde{E}^{\#}\widetilde{E} & (I - \widetilde{E}^{\#}\widetilde{E})(I - \beta\widetilde{E})^D \end{array}\right]$ donde $(I - \beta\widetilde{E})^D$ representa la inversa de Drazin de la matriz $I - \beta\widetilde{E}$ [1]. Este último conjunto es el subespacio de las condiciones iniciales admisibles del sistema. Ahora, reescribiendo $z(k)$ mediante propiedades de la inversa de grupo y poniéndola en términos de la función Ψ, se tiene

$$\begin{aligned} z(k) &= ((I - \beta\widetilde{E})\widetilde{E}^{\#})^k z(0) \\ &= \Psi(X(T^{-1} - \beta I)^k Y, M, N)z(0) \end{aligned}$$

donde se han utilizado las propiedades de Ψ indicadas en el Lema. De la definición de la función Ψ se observa que los estados $z(k)$ son no negativos cuando $T^{-1} - \beta I \geq O$ y $z(0) \geq 0$.

Por tanto, la salida del sistema viene dada por:

$$\begin{aligned} y(k) &= \left[\begin{array}{ccc} C_1 & C_2 & C_3 \end{array}\right] \Psi(X(T^{-1} - \beta I)^k Y, M, N)z(0) \\ &= (C_1 + C_3N)X(T^{-1} - \beta I)^k Y \left[\begin{array}{ccc} I & M & O \end{array}\right] z(0) \end{aligned}$$

donde se ve claramente que si $(C_1 + C_3N)X \geq O$ y $T^{-1} - \beta I \geq O$ entonces $y(k) \geq 0$ para una condición inicial admisible no negativa.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por el Proyecto DGI MTM2010-18228 y por el Proyecto de la Universidad Politécnica de Valencia, PAID-06-09, Ref.: 2659.

REFERENCIAS

[1] A. BEN-ISRAEL AND T. GREVILLE, *Generalized inverses: Theory and applications*, Wiley, 1974.

[2] G.R. DUAN, *Analysis and design of descriptor linear systems*, Springer, 2010.

[3] A. HERRERO, A. RAMÍREZ AND N. THOME, *An algorithm to check the nonnegativity of singular systems*, Applied Mathematics and Computation, Vol. 189 (2007), pp. 355-365.

[4] A. HERRERO, F.J. RAMÍREZ AND N. THOME, *Characterization of matrices with nonnegative group-projector*, Lecture Notes in Control and Information Sciences, Vol. 389 (2009), pp. 315-320.

[5] S.K. Jain, J. Tynan, Nonnegative matrices A with $AA^{\#} \geq O$. *Linear Algebra and its Applications* 379:381–394 (2004).

[6] T. KACZOREK, *Linear control systems*, Vol. I and II, Wiley, 1992.

MODELADO MATEMÁTICO PARA EL CONTROL ÓPTIMO DE LA POLIOMIELITIS

Alvaro Andrés Quintero Orrego, Anibal Muñoz Loaiza y Leonardo Duvan Restrepo Alape

Facultad de Educación, Programa de Matemáticas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia
aquintero76@hotmail.com, anibalml@hotmail.com, www.uniquindio.edu.co

Resumen: Se modela el control óptimo por prevención de la poliomielitis, mediante un funcional objetivo de costos indirectos y directos ligado a un sistema de ecuaciones diferenciales no lineales que interpreta la dinámica de transmisión de la bacteria en la población humana. Se analiza por el principio máximo de Pontryagin obteniendo un problema de contorno que se resuelve por MATLAB utilizando valores hipotéticos para los parámetros.

Palabras clave: *Modelado Matemático, Control óptimo, Poliomielitis, Funcional de costos, Principio Máximo de Pontryagin, Problema de contorno.*

1. INTRODUCCIÓN

La poliomielitis es una enfermedad viral que afecta principalmente el sistema nervioso y que puede llevar a parálisis total o parcial. Es causada por la infección con el poliovirus y es una enfermedad altamente contagiosa que se propaga fácilmente de persona a persona por contacto con moco, flema o con heces infectadas. En las áreas donde hay un brote, las personas con mayor vulnerabilidad para contraer la enfermedad son los niños, las mujeres embarazadas y los ancianos. Una vez que la enfermedad se ha declarado, no hay un tratamiento que la cure. En el periodo agudo, el tratamiento persigue controlar la fiebre y aliviar el dolor.

En relación al control óptimo y aplicación del principio del máximo de Pontryagin, se encuentran aplicaciones deterministas en epidemiología matemática como en Cancer, VIH-SIDA, Tuberculosis, Cólera, Dengue, Malaria, Influenza [1], [3], [5], [2], [4].

2. FORMULACIÓN DEL PROBLEMA DE CONTROL ÓPTIMO

Se propone el siguiente problema de control óptimo de la poliomielitis mediante control por vacunación y medidas preventivas para evitar el contagio con la bacteria diseminada en el medio ambiente. La dinámica corresponde a un proceso estocástico, continuo de nacimiento - muerte homogéneo con estados discretos y tasas de flujos de Poisson. Se interpreta mediante un sistema de ecuaciones diferenciales no lineales para las magnitudes promedio.

Las variables y parámetros del modelo son: x: número promedio de personas susceptibles en el grupo de riesgo desde que nacen a los veinte años; w: número promedio de personas infectadas asintomáticas en el grupo de riesgo desde que nacen a los veinte años; y: número promedio de personas infecciosas en el grupo de riesgo desde que nacen a los veinte años; z; número promedio de personas inmunes en el grupo de riesgo desde que nacen a los veinte años; N: población total en el grupo de riesgo desde que nacen a los veinte años en un tiempo t, respectivamente. c: concentración del poliovirus en el medio ambiente; $u_1(t)$: control dependiente del tiempo que indica la fracción de recién nacidos vacunados; $u_2(t)$: control dependiente del tiempo que indica la fracción de personas que toman medidas preventivas al contagio con el ambiente contaminado; $\beta_y,\ \beta_c$: probabilidades de transmisión directa e indirecta; θ: tasa de evolución de la infección; γ: tasa de infecciosos que adquieren inmunidad; $\sigma,\ \delta$: tasas de descarga del poliovirus al medio ambiente por las personas infecciosas y personas infectadas sintomáticos y ϵ: tasa de eliminación del poliovirus del medio ambiente; τ: tiempo fijo para el control óptimo; $\rho_1,\ \rho_2$: pesos.

Se plantea el funcional objetivo de costos directos e indirectos:

$$J\big(u_1(t),u_2(t)\big)=\int_0^{\tau}L\big(\boldsymbol{x}(t),\boldsymbol{u}(t)\big)dt=\int_0^{\tau}\Big\{w(t)+y(t)+\frac{\rho_1}{2}u_1^2(t)+\frac{\rho_2}{2}u_2^2(t)\Big\}\,dt$$

ligado al sistema de ecuaciones diferenciales ordinarias no lineales que interpretan la dinámica:

$$\begin{aligned}
\frac{dx}{dt} &= \big(1-u_1(t)\big)\mu N-\beta_y\frac{y}{N}x-\beta_c\big(1-u_2(t)\big)cx-\mu x\equiv f_1(\boldsymbol{x},\boldsymbol{u})\\
\frac{dw}{dt} &= \beta_y\frac{y}{N}x+\beta_c\big(1-u_2(t)\big)cx-(\theta+\mu)w\equiv f_2(\boldsymbol{x},\boldsymbol{u})\\
\frac{dy}{dt} &= \theta w-(\gamma+\mu)y\equiv f_3(\boldsymbol{x},\boldsymbol{u})\\
\frac{dz}{dt} &= \mu N u_1(t)+\gamma y-\mu z\equiv f_4(\boldsymbol{x},\boldsymbol{u})\\
\frac{dc}{dt} &= \sigma w+\delta y-\epsilon c\equiv f_5(\boldsymbol{x},\boldsymbol{u})
\end{aligned}$$

con condiciones iniciales, $x(0)=x_0$, $w(0)=w_0$, $y(0)=y_0$, $z(0)=z_0$, $c(0)=c_0$.

Se trata de hallar un control óptimo $\big(\tilde{u}_1(t),\tilde{u}_2(t)\big)$ tal que:

$$J\big(\tilde{u}_1(t),\tilde{u}_2(t)\big)=\min_{\Gamma}J\big(u_1(t),u_2(t)\big)$$

en donde,

$$\Gamma=\Big\{\big(u_1(t),u_2(t)\big)\in L^2(0,\tau):0\le u_1(t),u_2(t)\le 1\Big\}.$$

2.1. Análisis del problema de control óptimo

Dado el problema de control óptimo:

$$\begin{cases}
J\big(\boldsymbol{x}(t),\boldsymbol{u}(t)\big)=\displaystyle\int_0^{\tau}L\big(\boldsymbol{x}(t),\boldsymbol{u}(t)\big)dt\\
\dfrac{d\boldsymbol{x}}{dt}=\boldsymbol{F}(\boldsymbol{x},\boldsymbol{u},t),\quad\forall\,t\ge 0,\quad\forall\,\boldsymbol{u}\in\Omega\\
\boldsymbol{x}(0)=\boldsymbol{x}_0
\end{cases}$$

La solución existe si las siguientes hipótesis se cumplen:

i) El conjunto de controles y variables de estado es no vacío.

ii) El conjunto de controles admisibles Ω es cerrado y convexo.

iii) Cada f_i del sistema de ecuaciones de estado son continuas, están contenidas y acotadas superiormente por una suma de controles y estados contenidos, y puede ser escrita como una función lineal de u con coeficientes que dependen del estado y control.

iv) Existen constantes $\alpha_1,\alpha_2>0$ y $\rho>1$ tal que el Lagrangiano (el integrando) $L(\boldsymbol{x},\boldsymbol{u},t)$ del funcional objetivo J es cóncava y satisface:

$$L(\boldsymbol{x},\boldsymbol{u},t)\le\alpha_2-\alpha_1\Big(|u_1(t)|^2+|u_2(t)|^2\Big)^{\rho/2}$$

Al respecto se formula el siguiente teorema:

Teorema 1 *Dado el funcional objetivo* $J(u_1,u_2)=\int_0^\tau L\big(\boldsymbol{x}(s),\boldsymbol{u}(s)\big)ds$ *donde*

$$\Omega=\Big\{\boldsymbol{u}=(u_1,u_2):\ u_i \textit{ es medible},\ \ 0\le u_i\le 1,\ \ t\in[0,\tau]\quad \textit{para}\quad i=1,2\Big\}$$

sujeto a las ecuaciones de variables de estado con $\boldsymbol{x}(0)=\boldsymbol{x}_0$ *y* $\lambda(\tau)=0$*, entonces existe un control óptimo* $\bar{\boldsymbol{u}}=(\bar u_1,\bar u_2)$ *tal que* $\max_{\boldsymbol{u}\in\Omega}J(u_1,u_2)=J(\bar u_1,\bar u_2)$.

La función Hamiltoniana o (función de Pontryagin) es de la forma $H(\boldsymbol{x},\boldsymbol{u},\lambda)=L(\boldsymbol{x},\boldsymbol{u})+\sum_{i=1}^5\lambda_i f_i$, donde $\boldsymbol{x}$ es el vector de variables de estado, $\boldsymbol{u}$ el vector de controles, λ el vector de variables adjuntas o conjugadas y L es el Lagrangiano. Es decir,

$$\begin{aligned}H(\boldsymbol{x},&\boldsymbol{u},\lambda)=\\ &w+y+\frac{\rho_1}{2}u_1^2(t)+\frac{\rho_2}{2}u_2^2(t)+\lambda_1\Big[\big(1-u_1(t)\big)\mu N-\beta_y\frac{y}{N}x-\beta_c\big(1-u_2(t)\big)cx-\mu x\Big]+\\ &\lambda_2\Big[\beta_y\frac{y}{N}x+\beta_c\big(1-u_2(t)\big)cx-(\theta+\mu)w\Big]+\lambda_3\big(\theta w-(\gamma+\mu)y)\big)+\lambda_4\big(\mu N u_1(t)+\gamma y-\mu z\big)+\\ &\lambda_5(\sigma w+\delta y-\epsilon c)+\nu_1u_1(t)+\nu_2\big(1-u_1(t)\big)+\nu_3u_2(t)+\nu_4\big(1-u_2(t)\big)\end{aligned}$$

donde $\nu_i(t)\quad i=1,...,4$ son multiplicadores de penalización tales que:

$$\nu_1u_1=0,\quad \nu_2(1-u_1)=0\quad \text{y}\quad \nu_3u_2=0,\quad \nu_4(1-u_2)=0 \tag{1}$$

Aplicando la condición de primer orden $\frac{\partial H}{\partial \boldsymbol{u}}=0$ en particular $\frac{\partial H}{\partial u_1}=0$, $\frac{\partial H}{\partial u_2}=0$ se obtiene los controles óptimos:

$$\bar u_1(t)=\text{mín}\left(\text{máx}\left(0,\frac{1}{\rho_1}(\lambda_1\mu N-\lambda_4\mu N-\nu_1+\nu_2)\right),1\right)$$

y

$$\bar u_2(t)=\text{mín}\left(\text{máx}\left(0,\frac{1}{\rho_2}(\lambda_2\beta_c cx-\lambda_1\beta_c cx+\nu_4-\nu_3\right),1\right)$$

El sistema conjugado (o sistema adjunto) tiene la forma $\frac{d\lambda}{dt}=-H_{\boldsymbol{x}}(\boldsymbol{x},\lambda,\boldsymbol{u})$. Es decir,

$$\begin{aligned}\frac{d\lambda_1}{dt}&=\lambda_1\Big[\beta_y\frac{y}{N}+\beta_c\big(1-\bar u_2(t)\big)c+\mu\Big]-\lambda_2\Big[\beta_y\frac{y}{N}+\beta_c\big(1-\bar u_2(t)\big)c\Big]\equiv g_1(\boldsymbol{x},\lambda,\boldsymbol{u})\\ \frac{d\lambda_2}{dt}&=-1+\lambda_2(\mu+\theta)-\lambda_3\theta-\lambda_5\sigma\equiv g_2(\boldsymbol{x},\lambda,\boldsymbol{u})\\ \frac{d\lambda_3}{dt}&=-1+\lambda_1\beta_y\frac{x}{N}-\lambda_2\beta_y\frac{x}{N}+\lambda_3(\gamma+\mu)-\lambda_4\gamma-\lambda_5\delta\equiv g_3(\boldsymbol{x},\lambda,\boldsymbol{u})\\ \frac{d\lambda_4}{dt}&=\lambda_4\mu\equiv g_4(\boldsymbol{x},\lambda,\boldsymbol{u})\\ \frac{d\lambda_5}{dt}&=\lambda_1\beta_c\big(1-\bar u_2(t)\big)x-\lambda_2\beta_c\big(1-\bar u_2(t)\big)x+\lambda_5\epsilon\equiv g_5(\boldsymbol{x},\lambda,\boldsymbol{u})\end{aligned}$$

con condiciones de transversalidad $\lambda_i(\tau)=0, i=1,...,5$.

3. Resultados numéricos

El problema de contorno esta formado por el sistema de variables de estado de la dinámica de transmisión de la poliomielitis, con sus respectivas condiciones iniciales, el sistema conjugado y las condiciones terminales y el control óptimo:

$$\begin{cases}\frac{d\boldsymbol{x}}{dt}=F(\boldsymbol{x},\bar{\boldsymbol{u}},\lambda)\\ \frac{d\lambda}{dt}=G(\boldsymbol{x},\bar{\boldsymbol{u}},\lambda)\\ x(0)=x_0,\quad \lambda(\tau)=0\\ \bar u_1(t)=\text{mín}\left(\text{máx}\left(0,\frac{1}{\rho_1}(\lambda_1\mu N-\lambda_4\mu N-\nu_1+\nu_2)\right),1\right)\\ \bar u_2(t)=\text{mín}\left(\text{máx}\left(0,\frac{1}{\rho_2}(\lambda_2\beta_c cx-\lambda_1\beta_c cx+\nu_4-\nu_3\right),1\right).\end{cases}$$

se resolvió utilizando el programa MATLAB con las condiciones iniciales, condiciones terminales y valores hipotéticos de los parámetros.

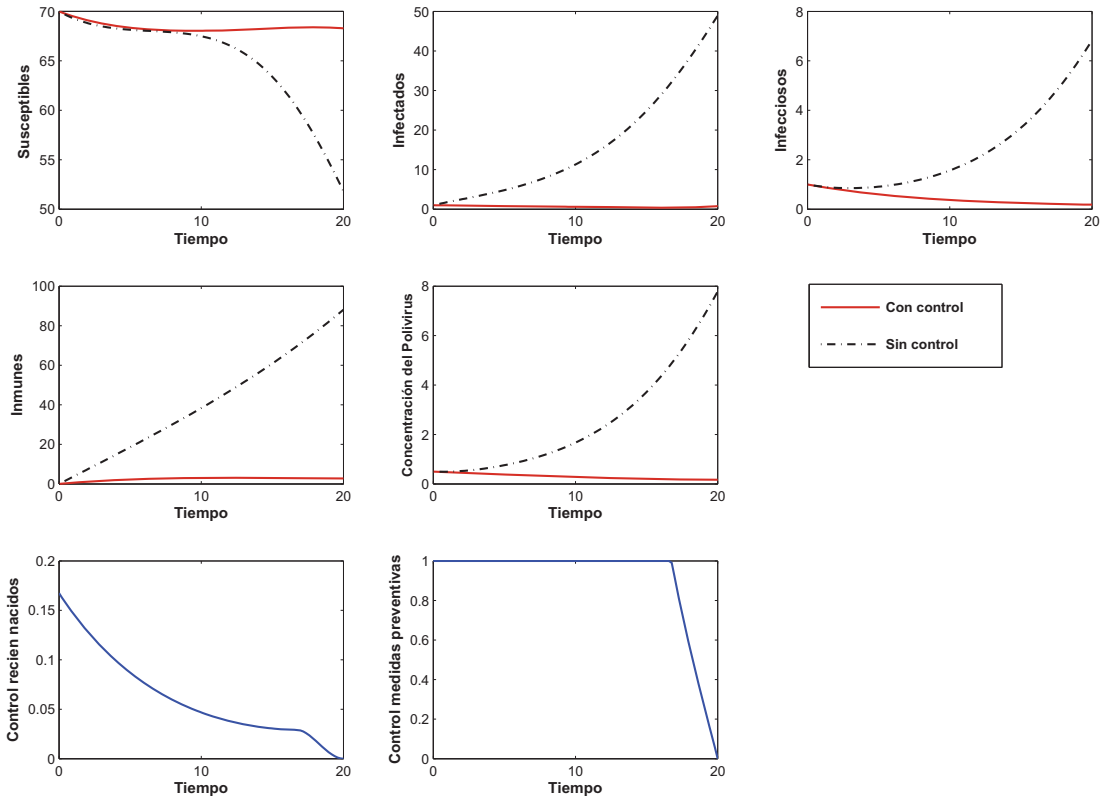


Figura 1: Comportamiento de las personas susceptibles x, personas infectadas asintomáticos w, personas infecciosas y, personas inmunes z y concentración del poliovirus en el medio ambiente c.

AGRADECIMIENTOS

A la Facultad de Ciencias, Universidad Nacional de Colombia, Sede Manizales-Caldas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, GMME, Universidad del Quindío.

REFERENCIAS

[1] S. BOWONG, *Optimal control of the transmission dynamics of tuberculosis*, Nonlinear Dyn, 21 March 2010.

[2] A.B. GUMEL, O. SHAROMI, *Curtailing smoking dynamics: A mathematical modeling approach*, Applied Mathematics and Computation, 19(2008)475-499.

[3] J. KARRAKCHOU, M. RACHIK, S. GOURARI, *Optimal control and infectiology: application to an HIV-AIDS model*, Applied Mathematics and Computation 177(2006)807-818.

[4] C. KAYA, *Time-optimal switching control for the US cocaine epidemic*, Socio-Economic Planning Sciences 38(2004)57-72.

[5] N.R.L. MILLER, E. SCHAEFER, H. GAFF, R. K. FISTER, S. LENHART, *Modeling Optimal Intervention Strategies for Cholera*, Bulletin of Mathematical Biology (2010).

[6] Caetano M.A., Yoneyama T. *Optimal and sub-optimal control in Dengue epidemics*, Optim. Control Appl. Math. 2001;22:63-73.

MODELO PARA EL CONTROL ÓPTIMO DEL DENGUE CON PERIODICIDAD

Luis Eduardo López M[1], Anibal Muñoz Loaiza [2], Gerard Olivar Tost[3] y Jose Betancourt Betancourt [4]

[1,3]*Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Colombia, Sede Manizales-Caldas;* [2]*Facultad de Educación, Programa de Matemáticas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia;* [4]*Centro de Medicina y Complejidad, Universidad Medica de Camaguey Carlos J. Finlay, Camaguey-Cuba; luisedu178@yahoo.com, anibalml@hotmail.com, www.uniquindio.edu.co*

Resumen: Se formula un problema de control óptimo por prevención del dengue clásico a un serotipo, mediante un funcional objetivo de costos indirectos y directos ligado a un sistema de ecuaciones diferenciales no lineales que interpreta la dinámica de transmisión periódica en la población humana acoplada a la dinámica periódica de incidencia del virus en los mosquitos *Aedes aegypti*, el cual se analiza por el principio máximo de Pontryagin obteniendo un problema de contorno que se resuelve por MATLAB utilizando valores hipotéticos para los parámetros.

Palabras clave: *Control óptimo, Dengue, Funcional de costos, Principio Máximo de Pontryagin, Problema de contorno, Aedes aegypti.*

1. INTRODUCCIÓN

El dengue es una enfermedad viral transmitida al hombre por mosquitos vectores del género *Aedes* siendo la especie *aegypti* la más importante en la transmisión de la enfermedad. El ciclo de vida del mosquito comprende cuatro fases: el huevo, cuatro estados larvales, un estado de pupa y el mosquito adulto; en las tres primeras el mosquito es inmaduro y en la última pasa a ser maduro. El dengue es endémico en las regiones tropicales y temporal o esporádico en las subtropicales y neárticas. Se reconoció clínicamente en América, Africa y Asia desde el siglo XVIII y se descubrió su transmisión por el *A. aegypti* en 1906, la cual fué comprobada en 1918. El aislamiento de su agente causal se logró en 1944 (3). El mosquito vector *A. aegypti* es hematófago y antropofílico (1), (2).

Según la Organización Panamericana de la Salud (Colombia), a la fecha del 10 de septiembre de 2010 se han encontrado 132906 casos totales de dengue de los cuales le $7\,\%$ presentan complicaciones de atención. Además se han confirmado 164 muertes, 22 se encuentran aún en estudio y 66 han sido descartadas.

En relación al control óptimo y aplicación del principio del máximo de Pontryagin, se encuentran aplicaciones deterministas en epidemiología matemática como en Cancer, VIH-SIDA, Tuberculosis, Cólera, Dengue, Malaria, Influenza (1), (2).

2. EL PROBLEMA DE CONTROL ÓPTIMO

Se propone el siguiente problema de control óptimo del *Aedes aegypti* mediante control químico del estado maduro y del estado inmaduro (huevos, larvas y pupas), integrando la dinámica de la población humana variable y el efecto de la estacionalidad, considerando unas tasas de contacto periódicas $\beta(t)$ y $\sigma(t)$. La dinámica corresponde a dos procesos estocásticos acoplados, continuos de nacimiento - muerte no homogéneos con estados discretos y tasas de flujos de Poisson. Se interpreta mediante un sistema de ecuaciones diferenciales para las magnitudes promedio.

Las variables y parámetros del modelo son x_1 : número promedio de personas susceptibles, x_2 : número promedio de personas infecciosas, x_3 : número promedio de personas inmunes temporalmente a un serotipo y x : población total humana variable en un tiempo t, respectivamente. y_1 : número promedio de mosquitos maduros no portadores del virus, y_2 : número promedio de mosquitos maduros portadores del virus, y : población total de mosquitos maduros y z : número promedio de estados inmaduros (huevos, larvas, pupas)

en un tiempo t, respectivamente. Δ : flujo constante de personas susceptibles que ingresan, μ : tasa de muerte natural de las personas en cada subpoblación, θ : tasa de recuperación temporal a un serotipo del virus, ω : tasa de desarrollo de los estados inmaduros a estado de mosquito maduro, ϵ : tasa de muerte natural de los mosquitos maduros (susceptibles y portadores del virus), π : tasa de muerte natural de los estados inmaduros, ϕ : tasa de ovoposición de los mosquitos, k : capacidad de carga de los criaderos, $\beta(t), \sigma(t)$: tasas de contacto periódicas, $u_1(t)$: control por medidas preventivas de la población susceptible, $u_2(t)$: control químico del mosquito maduro (no portadores y portadores).

Se plantea un funcional objetivo de costos directos e indirectos:

$$J(u_1(t), u_2(t)) = \int_0^{\tau} L(\mathbf{x}(t), \mathbf{u}(t))dt = \int_0^{t_f} \left\{ x_2(t) + z(t) + \frac{\eta_1}{2}u_1^2(t) + \frac{\eta_2}{2}u_2^2(t) \right\} dt$$

Ligado al sistema de ecuaciones diferenciales no lineal:

$$\begin{aligned}
\frac{dx_1(t)}{dt} &= \Delta - \beta(t)\frac{y_2(t)}{y_1(t)+y_2(t)}x_1(t) - \mu x_1(t) \equiv f_1(\mathbf{x}, \mathbf{u}) \\
\frac{dx_2(t)}{dt} &= \beta(t)\frac{y_2(t)}{y_1+y_2(t)}x_1(t) - (\mu+\theta)x_2(t) \equiv f_2(\mathbf{x}, \mathbf{u}) \\
\frac{dx_3(t)}{dt} &= \theta x_2(t) - \mu x_3(t) \equiv f_3(\mathbf{x}, \mathbf{u}) \\
\frac{dy_1(t)}{dt} &= \omega z(t) - \sigma(t)\frac{x_2(t)}{x_1(t)+x_2(t)+x_3(t)}y_1(t) - (\epsilon + u_1(t))y_1(t) \equiv f_4(\mathbf{x}, \mathbf{u}) \\
\frac{dy_2(t)}{dt} &= \sigma(t)\frac{x_2(t)}{x_1+x_2(t)+x_3(t)}y_1(t) - (\epsilon + u_1(t))y_2(t) \equiv f_5(\mathbf{x}, \mathbf{u}) \\
\frac{dz(t)}{dt} &= \phi\left(y_1(t)+y_2(t)\right)\left(1 - \frac{z(t)}{k}\right) - (\pi + \omega + u_2(t))z(t) \equiv f_6(\mathbf{x}, \mathbf{u})
\end{aligned}$$

donde, $\Delta, \mu, \theta, \omega, \epsilon, \phi, k, \pi > 0$; $x_1(0) = x_{10}$, $x_2(0) = x_{20}$, $x_3(0) = x_{30}$, $y_1(0) = y_{10}$, $y_2(0) = y_{20}$, $z(0) = z_0$. $\beta(t) = \beta_0 + \beta_1 \cos \delta t$, $\beta_0, \beta_1, \delta > 0$. $\sigma(t) = \rho_0 + \rho_1 \cos \alpha t$, $\rho_0, \rho_1, \alpha > 0$.

Se trata de hallar un control óptimo $(\tilde{u}_1(t), \tilde{u}_2(t))$ tal que:

$$J\left(\tilde{u}_1(t), \tilde{u}_2(t)\right) = \min_{\Omega} J\left(u_1(t), u_2(t)\right)$$

donde,

$$\Omega = \left\{(u_1, u_2) \in L^2(0, t_f) : 0 \leq u_1, u_2 \leq 1\right\}.$$

2.1. Análisis del problema de control óptimo

Dado el problema de control óptimo:

$$\begin{cases} J(\mathbf{x}(t), \mathbf{u}(t)) = \int_0^{\tau} L(\mathbf{x}(t), \mathbf{u}(t))dt \\ \frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}, \mathbf{u}, t), \quad \forall \quad t \geq 0, \quad \forall \quad \mathbf{u} \in \Omega \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases}$$

La solución existe si las siguientes hipótesis se cumplen:

i) El conjunto de controles y variables de estado es no vació.

ii) El conjunto de controles admisibles Ω es cerrado y convexo.

iii) Cada f_i del sistema de ecuaciones de estado son continuas, están contenidas y acotadas superiormente

por una suma de controles y estados contenidos, y puede ser escrita como una función lineal de u con coeficientes que dependen del estado y control.

iv) Existen constantes $\alpha_1, \alpha_2 > 0$ y $\rho > 1$ tal que el Lagrangiano (el integrando) $L(\mathbf{x}, \mathbf{u}, t)$ del funcional objetivo J es cóncava y satisface:

$$L(\mathbf{x}, \mathbf{u}, t) \leq \alpha_2 - \alpha_1 \left(|u_1(t)|^2 + |u_2(t)|^2\right)^{\rho/2}$$

Al respecto se formula el siguiente teorema:

Teorema 1 *Dado el funcional objetivo* $J(u_1, u_2) = \int_0^\tau L(\boldsymbol{x}(s), \boldsymbol{u}(s))ds$ *donde*

$$\Omega = \{\boldsymbol{u} = (u_1, u_2) : u_i \quad es \quad medible, \quad 0 \leq u_i \leq 1, t \in [0, \tau] \quad para \quad i = 1, 2\}$$

sujeto a las ecuaciones de variables de estado con $\boldsymbol{x}(0) = \boldsymbol{x}_0$ *y* $\lambda(\tau) = 0$*, entonces existe un control óptimo* $\bar{u} = (\bar{u}_1, \bar{u}_2)$ *tal que* $\max_{\boldsymbol{u} \in \Omega} J(u_1, u_2) = J(\bar{u}_1, \bar{u}_2)$.

La función Hamiltoniana o (función de Pontryagin) es de la forma $H(\mathbf{x}, \mathbf{u}, \lambda) = L(\mathbf{x}, \mathbf{u}) + \sum_{i=1}^{6} \lambda_i f_i$, donde **x** es el vector de variables de estado, **u** el vector de controles, λ el vector de variables adjuntas o conjugadas y L es el Lagrangiano. Es decir,

$$H(\mathbf{x}, \mathbf{u}, \lambda) = x_2 + z + \tfrac{\eta_1}{2} u_1^2(t) + \tfrac{\eta_2}{2} u_2^2(t) + \lambda_1 \left[\Delta - \beta(t) \tfrac{y_2}{y_1 + y_2} x_1 - \mu x_1\right] + \lambda_2 \left[\beta(t) \tfrac{y_2}{y_1 + y_2} x_1 - (\mu + \theta) x_2\right]$$

$$+ \lambda_3(\theta x_2 - \mu x_3) + \lambda_4 \left[\omega z - \sigma(t) \tfrac{x_2}{x_1 + x_2 + x_3} y_1 - (\epsilon + u_1(t)) y_1\right] + \lambda_5 \left[\sigma(t) \tfrac{x_2}{x_1 + x_2 + x_3} y_1 - (\epsilon + u_1(t)) y_2\right] +$$

$$+ \lambda_6 \left[\phi\,(y_1 + y_2)\left(1 - \tfrac{z}{k}\right) - (\pi + \omega + u_2(t)) z\right].$$

Aplicando la condición de primer orden $\frac{\partial H}{\partial \mathbf{u}} = 0$ en particular $\frac{\partial H}{\partial u_1} = 0$, $\frac{\partial H}{\partial u_2} = 0$ se obtiene los controles óptimos:

$$\bar{u}_1(t) = \min\left(\max\left(0, \frac{1}{\eta_1}(\lambda_4 y_1 + \lambda_5 y_2)\right), 1\right)$$

y

$$\bar{u}_2(t) = \min\left(\max\left(0, \frac{\lambda_6 z}{\eta_2}\right), 1\right)$$

El sistema conjugado (o sistema adjunto) tiene la forma $\frac{d\lambda}{dt} = -H_{\mathbf{x}}(\mathbf{x}, \lambda, \mathbf{u})$. Es decir,

$$\frac{d\lambda_1}{dt} = \lambda_1 \left[\beta(t) \tfrac{y_2}{y_1 + y_2} + \mu\right] - \lambda_2 \beta(t) \tfrac{y_2}{y_1 + y_2} - \lambda_4 \sigma(t) \tfrac{x_2}{(x_1 + x_2 + x_3)^2} y_1 + \lambda_5 \sigma(t) \tfrac{x_2}{(x_1 + x_2 + x_3)^2} y_1$$

$$\frac{d\lambda_2}{dt} = -1 + \lambda_2(\mu + \theta) - \lambda_3 \theta + \lambda_4 \sigma(t) \tfrac{x_1 + x_3}{(x_1 + x_2 + x_3)^2} y_1 - \lambda_5 \sigma(t) \tfrac{x_1 + x_3}{(x_1 + x_2 + x_3)^2} y_1$$

$$\frac{d\lambda_3}{dt} = \lambda_3 \mu - \lambda_4 \sigma(t) \tfrac{x_2}{(x_1 + x_2 + x_3)^2} y_1 + \lambda_5 \sigma(t) \tfrac{x_2}{(x_1 + x_2 + x_3)^2} y_1$$

$$\frac{d\lambda_4}{dt} = -\lambda_1 \beta(t) \tfrac{y_2}{(y_1 + y_2)^2} x_1 + \lambda_2 \beta(t) \tfrac{y_2}{(y_1 + y_2)^2} x_1 + \lambda_4 \left[\sigma(t) \tfrac{x_2}{x_1 + x_2 + x_3} + (\epsilon + \bar{u}_1(t))\right] - \lambda_5 \sigma(t) \tfrac{x_2}{x_1 + x_2 + x_3} -$$
$$\lambda_6 \phi \left(1 - \tfrac{z}{k}\right)$$

$$\frac{d\lambda_5}{dt} = \lambda_1 \beta(t) \tfrac{y_1}{(y_1 + y_2)^2} x_1 - \lambda_2 \beta(t) \tfrac{y_1}{(y_1 + y_2)^2} x_1 + \lambda_5(\epsilon + \bar{u}_1(t)) - \lambda_6 \phi \left(1 - \tfrac{z}{k}\right)$$

$$\frac{d\lambda_6}{dt} = -1 - \lambda_4 \omega + \tfrac{\lambda_6 \phi}{k}(y_1 + y_2) + \lambda_6(\pi + \omega + \bar{u}_2(t))$$

con condiciones de transversalidad $\lambda_i(\tau) = 0, i = 1, ..., 5$.

3. Resultados numéricos

El problema de contorno esta formado por el sistema de variables de estado de la dinámica de transmisión del dengue, con sus respectivas condiciones iniciales, el sistema conjugado y las condiciones terminales y el control óptimo:

$$\begin{cases} \frac{d\mathbf{x}}{dt} = F(\mathbf{x}, \bar{\mathbf{u}}, \lambda) \\ \frac{d\lambda}{dt} = G(\mathbf{x}, \bar{\mathbf{u}}, \lambda) \\ x(0) = x_0 \quad , \quad \lambda(\tau) = 0 \\ \bar{u}_1(t) = \text{mín}\left(\text{máx}\left(0, \frac{1}{\eta_1}(\lambda_4 y_1 + \lambda_5 y_2)\right), 1\right) \\ \bar{u}_2(t) = \text{mín}\left(\text{máx}\left(0, \frac{\lambda_6 z}{\eta_2}\right), 1\right) \end{cases}$$

se resolvió utilizando el programa MATLAB con las condiciones iniciales, condiciones terminales y valores hipotéticos de los parámetros.

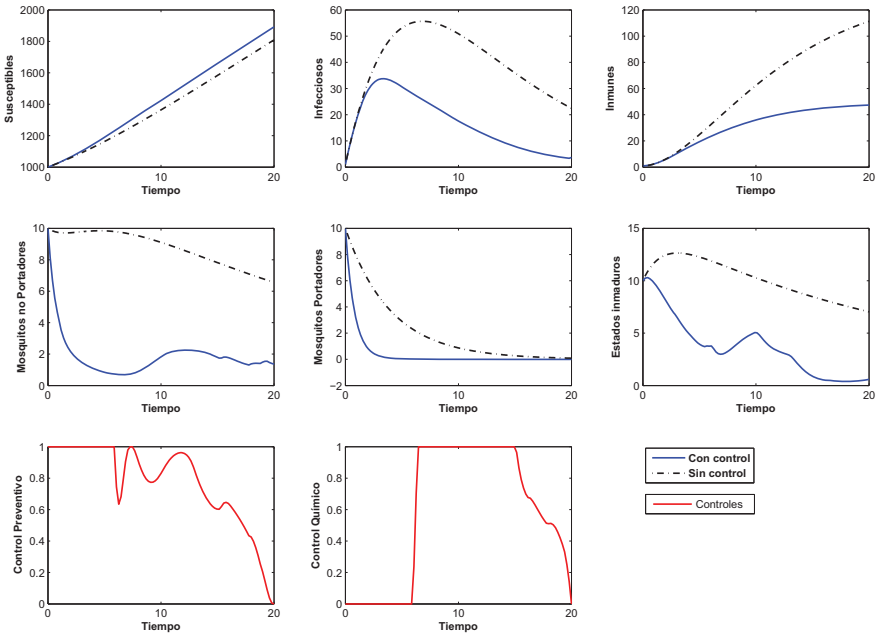


Figura 1: Comportamiento de las personas susceptibles x_1, personas infecciosas x_2, personas inmunes a un serotipo x_3, mosquitos no portadores del virus y_1, mosquitos portadores del virus y_2 y estados inmaduros del mosquito (huevos, larvas y pupas) z.

Agradecimientos

A la Facultad de Ciencias, Universidad Nacional de Colombia, Sede Manizales-Caldas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, GMME, Universidad del Quindío.

Referencias

[1] N.R.L. Miller, E. Schaefer, H. Gaff, R. K. Fister, S. Lenhart, *Modeling Optimal Intervention Strategies for Cholera*, Bulletin of Mathematical Biology (2010).

[2] Caetano M.A., Yoneyama T. *Optimal and sub-optimal control in Dengue epidemics*, Optim. Control Appl. Math. 2001;22:63-73.

[3] Secretaría de Salud. *Manual para la vigilancia epidemiológica del dengue*, (1984).

CONTROL ÓPTIMO DEL TABAQUISMO

Nini Johana Fiallo Rendon, Anibal Muñoz Loaiza y Leonardo Duvan Restrepo Alape

Facultad de Educación, Programa de Matemáticas, Laboratorio de Matemáticas y Biología Teórica L-MyBT, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia, ninis-428@hotmail.com, anibalml@hotmail.com, www.uniquindio.edu.co

Resumen: Se formula un problema de control óptimo por prevención para la adicción al tabaquismo, mediante un funcional objetivo de costos indirectos y directos ligado a un sistema de ecuaciones diferenciales no lineales que interpreta la dinámica, el cual se analiza por el principio máximo de Pontryagin obteniendo un problema de contorno que se resuelve por MATLAB utilizando valores hipotéticos para los parámetros.

Palabras clave: *Control óptimo, Tabaquismo, Funcional de costos, Principio Máximo de Pontryagin, Problema de contorno.*

1. INTRODUCCIÓN

El tabaquismo es una enfermedad causada por el consumo excesivo de tabaco, no sólo es un problema de salud pública sino también es un problema social ya que tiene efectos nocivos a la salud, no solo para las personas que lo consumen, sino de las que conviven con ellas. Esta enfermedad, considerada como una adicción de riesgo voluntario, es muy difícil de abandonar y controlar, por lo que una vez iniciado el hábito es muy difícil de dejarlo, ya que pasa a ser parte de la vida de una persona, quien a veces a pesar de saber el daño que hace, no se da cuenta que a cambio de un rato de placer; de forma lenta, pero efectiva, el tabaco va ocasionando daños irreversibles en la mayoría de los órganos del cuerpo, generando varias enfermedades crónicas y degenerativas y es causa de muerte prematura.

En relación al control óptimo y aplicación del principio del máximo de Pontryagin, se encuentran aplicaciones deterministas en epidemiología matemática como en Cancer, VIH-SIDA, Tuberculosis, Cólera, Dengue, Malaria, Influenza (1), (3), (5) y estudios en drogadición (2), (4), (6), (7), (8).

2. FORMULACIÓN DEL PROBLEMA DE CONTROL ÓPTIMO

Se formula y analiza un problema para el control óptimo de la dinámica del tabaquismo con población variable. Se plantea un funcional objetivo de costos directos e indirectos ligado a un sistema de ecuaciones diferenciales ordinarias no lineales que interpretan dicha dinámica. Dicho problema se analiza aplicando el principio máximo de Pontryagin. Los supuestos del modelo son: grupo de riesgo de personas al consumo de tabaquismo desde la edad promedio de 10 años, población total variable, seudo principio de acción de masas y mortalidad por consumo de tabaquismo en el caso de que sea crónico.

Las variables y parámetros son S : número promedio de personas mayores de 10 años susceptibles a ser fumadores activos, I_a : número promedio de personas mayores de 10 años fumadores activos, I_p : número promedio de personas mayores de 10 años fumadores pasivos, C : número promedio de personas mayores de 10 años fumadores crónicos en un tiempo t respectivamente, N : flujo de personas que cumplen la edad continuamente y que ingresan al grupo de riesgo de la población susceptible, α : tasa de personas que dejan de fumar y vuelven hacer susceptibles, β : coeficiente de encuentros efectivos, para personas que se vuelven fumadoras, δ : fuerza de infección de fumadores pasivos que adquieren el habito de fumar, μ : tasa de mortalidad natural, σ : flujo de personas que cumplen la edad continuamente y que ingresan al grupo de fumadores pasivos, θ : tasa de fumadores que recaen a fumadores crónicos, ξ: tasa de incremento de las personas susceptibles, ϑ: tasa de muerte por la infección crónica, u_1, u_2: controles preventivos dependientes del tiempo, τ: tiempo fijo y η_i, i=1,2,3,4: pesos de los costos directos e indirectos.

Se plantea el funcional objetivo de costos indirectos y directos:

$$J(\mathbf{x},\mathbf{u}) = \int_0^\tau L(\mathbf{x},\mathbf{u})dt = \int_0^\tau \left\{\eta_1 I_a + \eta_2 I_p + \frac{\eta_3}{2}u_1^2 + \frac{\eta_4}{2}u_2^2\right\} dt$$

ligado al sistema de ecuaciones diferenciales:

$$\begin{aligned}
\frac{dS}{dt} &= \xi N - \beta(1-u_1)\frac{I_a}{N}S - \sigma(1-u_2)\frac{I_a}{N}S + \alpha I_a - \mu S \equiv f_1(\mathbf{x},\mathbf{u}) \\
\frac{dI_a}{dt} &= \beta(1-u_1)\frac{I_a}{N}S + \delta I_p - (\alpha+\mu+\theta)I_a \equiv f_2(\mathbf{x},\mathbf{u}) \\
\frac{dI_p}{dt} &= \sigma(1-u_2)\frac{I_a}{N}S - (\delta+\mu)I_p \equiv f_3(\mathbf{x},\mathbf{u}) \\
\frac{dC}{dt} &= \theta I_a - (\mu+\vartheta)C \equiv f_4(\mathbf{x},\mathbf{u}) \\
\frac{dN}{dt} &= (\xi-\mu)N - \vartheta C \equiv f_5(\mathbf{x},\mathbf{u})
\end{aligned}$$

con condiciones iniciales, $\mathbf{x}(0) = (S(0), I_a(0), I_p(0), C(0)) = (s_0, i_{a0}, i_{p0}, c_0)$

Se trata de hallar un control óptimo $(\tilde{u}_1(t), \tilde{u}_2(t))$ tal que:

$$J\left(\tilde{u}_1(t), \tilde{u}_2(t)\right) = \min_{\Gamma} J\left(u_1(t), u_2(t)\right)$$

en donde,

$$\Gamma = \left\{(u_1,u_2) \in L^2(0,\tau) : 0 \le u_1, u_2 \le 1\right\}.$$

2.1. Análisis del problema de control óptimo

Dado el problema de control óptimo:

$$\begin{cases} J(\mathbf{x}(t),\mathbf{u}(t)) = \int_0^\tau L(\mathbf{x}(t),\mathbf{u}(t))dt \\ \frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x},\mathbf{u},t), \quad \forall \quad t \ge 0, \quad \forall \quad \mathbf{u} \in \Gamma \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases}$$

La solución existe si las siguientes hipótesis se cumplen:

i) El conjunto de controles y variables de estado es no vació.

ii) El conjunto de controles admisibles Γ es cerrado y convexo.

iii) Cada f_i del sistema de ecuaciones de estado son continuas, están contenidas y acotadas superiormente por una suma de controles y estados contenidos, y puede ser escrita como una función lineal de u con coeficientes que dependen del estado y control.

iv) Existen constantes $\alpha_1, \alpha_2 > 0$ y $\rho > 1$ tal que el Lagrangiano (el integrando) $L(\mathbf{x},\mathbf{u},t)$ del funcional objetivo J es cóncava y satisface:

$$L(\mathbf{x},\mathbf{u},t) \le \alpha_2 - \alpha_1\left(|u_1(t)|^2 + |u_2(t)|^2\right)^{\rho/2}$$

Al respecto se formula el siguiente teorema:

Teorema 1 *Dado el funcional objetivo* $J(u_1,u_2) = \int_0^\tau L(\boldsymbol{x}(s),\boldsymbol{u}(s))ds$ *donde*

$$\Omega = \{\boldsymbol{u} = (u_1,u_2) : u_i \quad es \quad medible, \quad 0 \le u_i \le 1, t \in [0,\tau] \quad para \quad i = 1,2\}$$

sujeto a las ecuaciones de variables de estado con $\boldsymbol{x}(0) = \boldsymbol{x}_0$ *y* $\lambda(\tau) = 0$*, entonces existe un control óptimo* $\bar{u} = (\bar{u}_1, \bar{u}_2)$ *tal que* $\max_{\boldsymbol{u}\in\Gamma} J(u_1,u_2) = J(\bar{u}_1,\bar{u}_2)$.

La función Hamiltoniana o (función de Pontryagin) es de la forma $H(\mathbf{x},\mathbf{u},\lambda) = L(\mathbf{x},\mathbf{u}) + \sum_{i=1}^{5} \lambda_i f_i$, donde $\mathbf{x}$ es el vector de variables de estado, $\mathbf{u}$ el vector de controles, λ el vector de variables adjuntas o conjugadas y L es el Lagrangiano. Es decir,

$$H(\mathbf{x},\mathbf{u},\lambda) = \lambda_1\left[\xi N - \beta(1-u_1)\tfrac{I_a}{N}S - \sigma(1-u_2)\tfrac{I_a}{N}S + \alpha I_a - \mu S\right]$$

$$+\lambda_2\left[\beta(1-u_1)\tfrac{I_a}{N}S + \delta I_p - (\alpha+\theta+\mu)I_a\right] + \lambda_3\left[\sigma(1-u_2)\tfrac{I_a}{N}S - (\delta+\mu)I_p\right] + \lambda_4\left[\theta I_a - (\mu+\vartheta)C\right]$$

$$+\lambda_5\left[(\xi-\mu)N - \vartheta C\right] + \phi\left[\eta_1 I_a + \eta_2 I_p + \tfrac{\eta_3}{2}u_1^2 + \tfrac{\eta_4}{2}u_2^2\right] + \upsilon_1 u_1 + \upsilon_2(1-u_1) + \upsilon_3 u_2 + \upsilon_4(1-u_2).$$

donde $\upsilon_i(t) \quad i = 1, ..., 4$ son multiplicadores de penalización tales que:

$$\upsilon_1 u_1 = 0, \quad \upsilon_2(1-u_1) = 0 \quad \text{y} \quad \upsilon_3 u_2 = 0, \quad \upsilon_4(1-u_2) = 0 \tag{1}$$

donde, $\phi > 0$.

Aplicando la condición de primer orden $\frac{\partial H}{\partial \mathbf{u}} = 0$ en particular $\frac{\partial H}{\partial u_1} = 0$, $\frac{\partial H}{\partial u_2} = 0$ y las condiciones de los multiplicadores de penalización, se obtiene los controles óptimos:

$$\bar{u}_1(t) = \text{mín}\left(\text{máx}\left(0, \left(\frac{\lambda_2 - \lambda_1}{\phi\eta_3}\right)\beta\frac{I_a}{N}S\right), 1\right)$$

y

$$\bar{u}_2(t) = \text{mín}\left(\text{máx}\left(0, \left(\frac{\lambda_3 - \lambda_1}{\phi\eta_4}\right)\sigma\frac{I_a}{N}S\right), 1\right)$$

El sistema conjugado (o sistema adjunto) tiene la forma $\frac{d\lambda}{dt} = -H_{\mathbf{x}}(\mathbf{x},\lambda,\mathbf{u})$. Es decir,

$$\frac{d\lambda_1}{dt} = \lambda_1\left[\beta(1-u_1)\tfrac{I_a}{N} + \sigma(1-u_2)\tfrac{I_a}{N} + \mu\right] - \lambda_2\beta(1-u_1)\tfrac{I_a}{N} - \lambda_3\sigma(1-u_2)\tfrac{I_a}{N}$$

$$\frac{d\lambda_2}{dt} = \lambda_1\left[\beta(1-u_1)\tfrac{S}{N} + \sigma(1-u_2)\tfrac{S}{N} - \alpha\right] - \lambda_2\left[\beta(1-u_1)\tfrac{S}{N} - (\alpha+\theta+\mu)\right]$$

$$-\lambda_3\sigma(1-u_2)\tfrac{S}{N} - \theta\lambda_4 - \phi\eta_1$$

$$\frac{d\lambda_3}{dt} = -\delta\lambda_2 + \lambda_3(\delta+\mu) - \phi\eta_2$$

$$\frac{d\lambda_4}{dt} = \lambda_4(\mu+\vartheta) + \vartheta\lambda_5$$

$$\frac{d\lambda_5}{dt} = \lambda_1\left[-\xi - \beta(1-u_1)\tfrac{I_a S}{N^2} - \sigma(1-u_2)\tfrac{I_a S}{N^2}\right] + \lambda_2\beta(1-u_1)\tfrac{I_a S}{N^2} + \lambda_3\sigma(1-u_2)\tfrac{I_a S}{N^2} + \lambda_5(\xi-\mu)$$

con condiciones de transversalidad $\lambda_i(\tau) = 0, i = 1, ..., 5$.

3. Resultados numéricos

El problema de contorno esta formado por el sistema de variables de estado de la dinámica del tabaquismo con sus respectivas condiciones iniciales, el sistema conjugado y las condiciones terminales y el control óptimo:

$$\begin{cases} \frac{d\mathbf{x}}{dt} = F(\mathbf{x},\bar{\mathbf{u}},\lambda) \\ \frac{d\lambda}{dt} = G(\mathbf{x},\bar{\mathbf{u}},\lambda) \\ x(0) = x_0 \quad , \quad \lambda(\tau) = 0 \\ \bar{u}_1 = \text{mín}\left(\text{máx}\left(0, \left(\frac{\lambda_2-\lambda_1}{\phi\eta_3}\right)\beta\frac{I_a}{N}S\right), 1\right) \\ \bar{u}_2 = \text{mín}\left(\text{máx}\left(0, \left(\frac{\lambda_3-\lambda_1}{\phi\eta_4}\right)\sigma\frac{I_a}{N}S\right), 1\right). \end{cases}$$

se resolvió utilizando el programa MATLAB con las condiciones iniciales, condiciones terminales y valores hipotéticos de los parámetros.

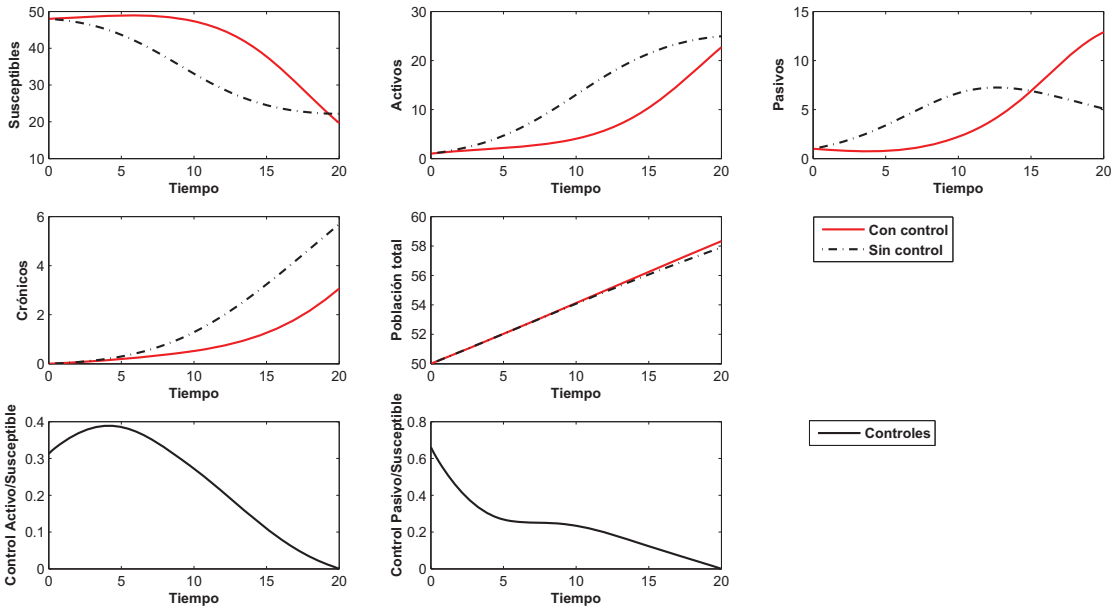


Figura 1: Comportamiento de los susceptibles S, fumadores activos I_a, fumadores pasivos I_p y fumadores crónicos C, población total y controles.

AGRADECIMIENTOS

Al Programa de matemáticas y Facultad de Educación, Universidad del Quindío.

REFERENCIAS

[1] S. BOWONG, *Optimal control of the transmission dynamics of tuberculosis*, Nonlinear Dyn, 21 March 2010.

[2] A.B. GUMEL, O. SHAROMI, *Curtailing smoking dynamics: A mathematical modeling approach*, Applied Mathematics and Computation, 19(2008)475-499.

[3] J. KARRAKCHOU, M. RACHIK, S. GOURARI, *Optimal control and infectiology: application to an HIV-AIDS model*, Applied Mathematics and Computation 177(2006)807-818.

[4] C. KAYA, *Time-optimal switching control for the US cocaine epidemic*, Socio-Economic Planning Sciences 38(2004)57-72.

[5] N.R.L. MILLER, E. SCHAEFER, H. GAFF, R. K. FISTER, S. LENHART, *Modeling Optimal Intervention Strategies for Cholera*, Bulletin of Mathematical Biology (2010).

[6] G. MULONE, B. STRAUGHAN, *A note on heroin epidemics*, Mathematical Biosciencies, 218(2009) 138-141.

[7] F. NYABADZA, D. MUSEKWA-HOVE, *From heroin epidemics to methamphetamine epidemics: Modelling substance abuse in a South African Province*, Mathematical Biociences 225(2010)132-140.

[8] B. OLSSON, G. CARLSSON, M. FANT, T. JOHANSSON, O. OLSSON, C. ROTH, *Heavy drug abuse in sweden 1979-a national casefinding*, Drug and Alcohol Dependence, 7(1981)273-283.

CONTROL ÓPTIMO EN UN MODELO HOSPEDERO - PARASITOIDE

Monica J. Mesa M., Oscar E. Molina D., Hernán D. Toro Z. y Anibal Muñoz L.

Facultad de Educación, Programa de Matemáticas, Grupo de Modelación Matemática en Epidemiología (GMME), Universidad del Quindío, Armenia, Quindío, Colombia monicamesa83@hotmail.com, oscarmolina2908@hotmail.com, torozapatahd@hotmail.com, anibalml@hotmail.com, www.uniquindio.edu.co

Resumen: Se formula un problema de control óptimo para el control químico de un insecto plaga en una relación Hospedero - Parasitoide, mediante un funcional de costos cuadrático ligado a un sistema de ecuaciones diferenciales que interpreta la dinámica poblacional, que incluye el ciclo de vida del hospedero y una fracción de escape al parasitismo de acuerdo a la distribución binomial negativa. Se aplica el principio del máximo de Pontryagin en el análisis del modelo con control óptimo.

Palabras clave: *Hospedero - Parasitoide, Distribución Binomial Negativa, Control óptimo, Principio del Máximo de Pontryagin.*

1. INTRODUCCIÓN

El parasitismo es una interacción biológica entre organismos de diferentes especies, en la que uno de los organismos (el parásito) consigue la mayor parte del beneficio de una relación estrecha con otro, el huésped, además, es un proceso por el cual una especie amplía su capacidad de supervivencia utilizando a otras especies para que cubran sus necesidades básicas y vitales. El parasitismo puede ser considerado un caso particular de depredación. Un parásito que mata al organismo donde se hospeda es llamado parasitoide.

La estrategia aplicada es el Principio del Máximo de Pontryagin para resolver el problema de control óptimo general. Este principio establece un conjunto de condiciones necesarias para que una curva sea solución del problema de control óptimo y se expresa en forma sencilla en términos de una función llamada Hamiltoniano. Además se plantea una situación biológica particular, donde se tiene una relación hospedero-parasitoide, en este caso el huésped es una población de insectos plaga de agrocultivos los cuales son parasitados en su estado larval y se pretende conocer su dinámica poblacional cuando la plaga es controlada químicamente.

2. MODELO MATEMÁTICO HOSPEDERO-PARASITOIDE

Se considera un problema de control óptimo que describe la dinámica hospedero-parasitoide, donde X_1 es el número promedio de insectos plaga adultos, los cuales serán afectados debido al control químico U, también se incluye el ciclo de vida del hospedero que son número promedio de huevos viables de la plaga, número promedio de larvas del insecto plaga, pupas del insecto plaga, representadas con X_2, X_3, X_4 respectivamente en un tiempo t, además se tiene que las larvas son afectadas por parasitoides adultos donde X_5 es el número promedio de ellos y X_6 el número promedio de larvas parasitadas en un tiempo t, se considera una fracción de escape al parasitismo de acuerdo a la distribución binomial negativa la cual se representa como $f(0) = \left(1 + \frac{aX_5}{d}\right)^{-d}$.

La dinámica poblacional tendrá sentido biológico en la siguiente región:

$$\Omega_{\mathbf{X}} = \left\{\mathbf{X} \in \mathbb{R}^6 : X_1 > 0, 0 < X_2 < k, 0 < X_3 < X_2, 0 \leq X_4 \leq X_3, X_5 > 0, 0 \leq X_6 \leq X_3\right\}$$

donde $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6) \in \Omega_{\mathbf{X}} \subset \mathbb{R}^6$ son las variables de estado y el control $\mathbf{U}$ está definido en el conjunto $\Omega_{\mathbf{U}}$ dado por,

Tabla 1: Interpretación de los parámetros.

Símbolo	*Interpretación*
ϕ	tasa de ovoposición del insecto plaga.
ϵ, ρ, ω	tasa de muerte natural del insecto plaga, larvas y pupas.
α	tasa de desarrollo de huevo a larva.
ψ	tasa de eliminación natural de huevos del insecto plaga.
θ	tasa de desarrollo de larva a pupa del insecto plaga.
δ	tasa de desarrollo de pupa a insecto plaga adulto.
$f(0)$	fracción de escape al parasitismo.
β	tasa de muerte natural de los parasitoides adultos.
μ	tasa de muerte natural de larvas parasitadas.
η	tasa de desarrollo de larvas parasitadas a parasitoides adultos.
a	la eficiencia de búsqueda del parasitoide.
d	indica el nivel espacial.
k	capacidad de carga de los huevos del insecto plaga

$$\Omega_{\mathbf{U}} = \{u \in L^1(0, t_f) : 0 \leq U(t) \leq 1\}$$

La dinámica es modelada mediante el sistema de ecuaciones diferenciales,

$$\begin{cases} \frac{dX_1}{dt} = \delta X_4 - \epsilon X_1 - U X_1 \\ \frac{dX_2}{dt} = \phi X_1 \left(1 - \frac{X_2}{k}\right) - (\psi + \alpha) X_2 \\ \frac{dX_3}{dt} = \alpha X_2 - \left(1 - \left(1 + \frac{aX_5}{d}\right)^{-d}\right) X_3 - (\theta + \rho) X_3 \\ \frac{dX_4}{dt} = \theta X_3 - (\omega + \delta) X_4 \\ \frac{dX_5}{dt} = \eta X_6 - \beta X_5 \\ \frac{dX_6}{dt} = \left(1 - \left(1 + \frac{aX_5}{d}\right)^{-d}\right) X_3 - (\eta + \mu) X_6 \\ X_i(0) = X_i^0, \text{para } i = 1, \ldots, 6. \end{cases} \tag{1}$$

donde $\delta, \epsilon, \phi, \pi, \alpha, a, d, \theta, \rho, \omega, \eta, \beta, \mu > 0$

3. Problema de control óptimo

Dado el sistema dinámico anterior, donde $X \in \Omega_X \subset \mathbb{R}^6$ son las variables de estado, $U \in \Omega_U$ es el control y sea $J(X(t), U(t))$ una función la cual describe los costos (directos e indirectos) definida de la siguiente manera:

$$J(X(t), U(t)) = \frac{1}{2} \int_0^{t_f} \{\xi_1 X_2^2(t) + \xi_2 U^2(t)\} dt$$

para un t_f fijo, $\xi_i \geq 0$ e $i = 1, 2$.

Con lo anterior se tiene un problema de control y para darle solución se requiere encontrar un control óptimo $\tilde{U}(t)$ tal que minimice el funcional de costos cuadrático $J\left(\tilde{U}(t)\right) = \min_{\Omega_u} J\left(U(t)\right)$.

Para lograr esto se define primero el Hamiltoniano $H(X, U, t)$ así:

$$
\begin{aligned}
H(\mathbf{X},\mathbf{U}) &= \eta_1\left[\delta X_4 - \epsilon X_1 - UX_1\right] + \eta_2\left[\phi X_1\left(1-\tfrac{X_2}{k}\right) - (\psi+\alpha)X_2\right] + \\
&\quad \eta_3\left[\alpha X_2 - \left(1-\left(1+\tfrac{aX_5}{d}\right)^{-d}\right)X_3 - (\theta+\rho)X_3\right] + \\
&\quad \eta_4\left[\theta X_3 - (\omega+\delta)X_4\right] + \eta_5\left[\eta X_6 - \beta X_5\right] + \\
&\quad \eta_6\left[\left(1-\left(1+\tfrac{aX_5}{d}\right)^{-d}\right)X_3 - (\eta+\mu)\,X_6\right] + \\
&\quad \sigma\left[\tfrac{1}{2}\left(\varepsilon_1 X_2^2 + \varepsilon_2 U^2\right)\right] + Z_1U + (1-U)Z_2.
\end{aligned}
$$

donde Z_1, Z_2 son positivos y satisfacen $Z_1U = 0, \quad (1-U)Z_2 = 0.$

Teorema 1 *Dado el control óptimo $U \in \Omega_u$ y la solución para el sistema $\overline{X_i}$, $i = 1, 2, ..., 6$ de el sistema, existen variables adjuntas η_i para $i = 1, 2, ..., 6$ que satisfacen $\frac{d\eta}{dt} = -H_X(\overline{X}, \overline{U}, \eta)$, es decir,*

$$
\begin{aligned}
\frac{d\eta_1}{dt} &= \eta_1(\epsilon+\mu) - \eta_2\phi\left(1-\frac{X_2}{k}\right) \\
\frac{d\eta_2}{dt} &= \eta_2\frac{\phi X_1}{k} + \eta_1(\psi+\alpha) - \eta_3\alpha - \sigma\varepsilon_1 X_2 \\
\frac{d\eta_3}{dt} &= \eta_3\left[1-\left(1+\frac{aX_5}{d}\right)^{-d}\right] + \eta_3(\theta+\rho) - \theta\eta_4 - \eta_6\left[1-\left(1+\frac{aX_5}{d}\right)^{-d}\right] \\
\frac{d\eta_4}{dt} &= -\eta_1\delta + \eta_4(\omega+\delta) \\
\frac{d\eta_5}{dt} &= a\eta_3 X_3\left(1+\frac{aX_5}{d}\right)^{-d-1} + \eta_5\beta - a\eta_6 X_3\left(1+\frac{aX_5}{d}\right)^{-d-1} \\
\frac{d\eta_6}{dt} &= -\eta_5\eta + \eta_6(\eta+\mu) \\
\eta_i(\tau) &= 0, \text{ para } i = 1,2,3,4,5,6,
\end{aligned}
$$

y el control óptimo dado por:

$$\overline{U} = \text{mín}\left(\text{máx}\left(0, \frac{\eta_1 X_1}{\sigma\varepsilon_2}\right), 1\right)$$

4. Resultados numéricos

Las simulaciones del problema de contorno se desarrollaron en ambiente Matlab en el intervalo $[0, t_f]$, los valores de los parámetros para la simulación son: $X_1(0) = 5000$; $X_2(0) = 10000$; $X_3(0) = 6000$; $X_4(0) = 5000$; $X_5(0) = 100$; $X_6(0) = 100$; $\phi = 0{,}7$; $\epsilon = 0{,}1$; $\rho = 0{,}1$; $\omega = 0{,}1$; $\alpha = 0{,}7$; $\psi = 0{,}1$; $\theta = 0{,}8$; $\delta = 1$; $\beta = 0{,}1$; $\mu = 0{,}1$; $\eta = 0{,}2$; $k = 10000$; $\sigma = 1$; $a = 1$; $d = 1$; $\zeta = 0{,}01$.

Las figuras muestran el comportamiento de las variables de estado sin control en linea punteada y con control en la linea continua.

En la figura 1 se observa una gran diferencia entre la dinámica del modelo con control y sin control; la población de insectos adultos disminuye considerablemente, esto se logra debido a que el costo directo es bajo, y hace que se pueda aplicar el control químico a la plaga de agrocultivos que es atacada por un parasitoide larval de una manera más fuerte.

Para este nuevo escenario se ha incrementado el valor del peso directo ξ_2 el cual representa un costo económico alto, se puede observar que el máximo control es del $42\,\%$ durante los primeros días, esto se tiene debido a que la aplicación del químico es costosa, además se aprecia en las figuras correspondientes a la dinámica con control que los insectos adultos son afectados de una manera significativa y esto se logra en un periodo de tiempo mayor comparado con la figura 1.

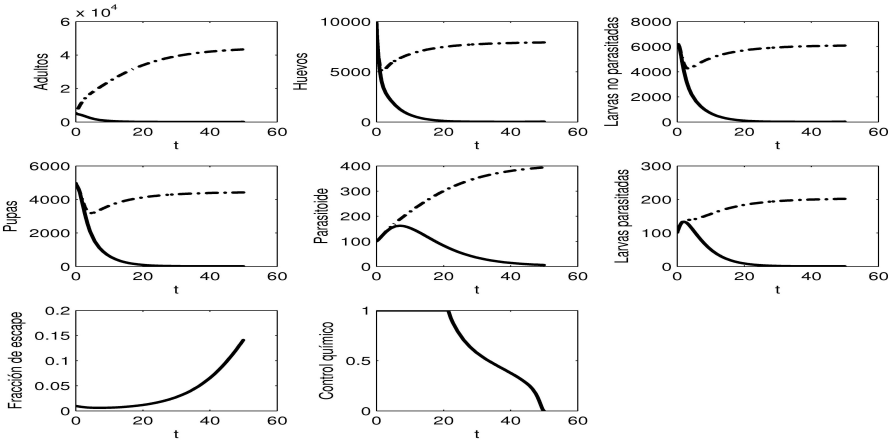


Figura 1: Se considera $\xi_1 = 0{,}01$, $\xi_2 = 100$.

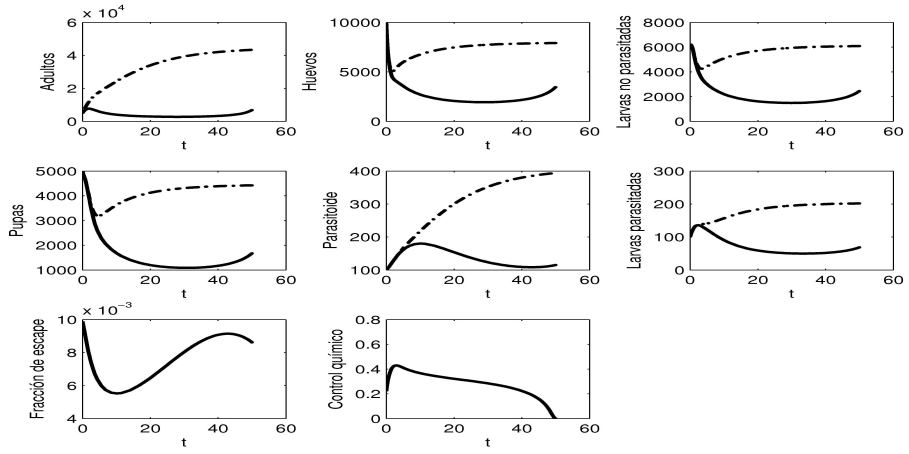


Figura 2: Se consideran los valores $\xi_1 = 0{,}01$, $\xi_2 = 1000000$.

5. CONCLUSIONES Y RESULTADOS

Se puede observar que el modelo presenta muchos parámetros relevantes y cuya variación da origen a escenarios diferentes cada vez, pero en términos generales se puede concluir que el control químico es muy efectivo para controlar la población de insectos plaga. Además, al comparar los escenarios obtenidos al variar el costo directo ξ_2, se puede apreciar que una buena estrategia de control se tiene cuando ξ_2 es bajo, es decir, la aplicación de control químico es económicamente favorable y por lo tanto seria podría ser más aplicada. Efectivamente puede verse como la población de insectos plaga del agrocultivo es eliminada aunque requiera más tiempo que cuando los costos directos son altos.

6. REFERENCIAS

Caetano M. A., Yoneyama T. Optimal and sub-optimal control in Dengue epidemics, Optim. Control Appl. Math. 2001; 22:63-73.

May Robert M. *Host-Parasitoid Systems in Patchy Enviroments: A Phenomenoligical Model*. Biology Department. Princeton University, Princeton. Journal of Animal Ecology (1978).

N.R.L Miller, E. SCHAEFER, H. GAFF, R.K. FISTER, S. LENHART. Modeling Optimal Intervention Strategies for Cholera. Bulletin of Mathematical Biology (2010).

MODELADO DE LA EXTRACCIÓN DE ACEITES VEGETALES Y SUS COMPUESTOS MINORITARIOS

Erica R. Baümler, Amalia A. Carelli, Guillermo H. Crapiste y María E. Carrín.

PLAPIQUI (UNS-CONICET). La Carrindanga Km 7, C.C. 717, 8000-Bahía Blanca, Argentina. Tel./Fax: (54-291) 4861700, email: mcarrin@plapiqui.edu.ar

RESUMEN

Un modelo matemático transitorio en dos dimensiones desarrollado previamente fue expandido y aplicado para representar la extracción de aceite y sus compuestos minoritarios en un extractor industrial tipo De Smet. Las difusividades y parámetros de equilibrio fueron obtenidos de datos experimentales. El modelo fue resuelto numéricamente usando diferencias finitas para discretizar las derivadas espaciales y sumas finitas para reemplazar términos integrales. Las ecuaciones diferenciales ordinarias fueron resueltas utilizando Runge-Kutta de cuarto orden con MatLab 7.5®. Esto permitió predecir la concentración de los compuestos en la miscela y collets a lo largo del extractor. Se confirmó que la extracción de fosfolípidos y ceras cristalizables es más acentuada al final del proceso. Las simulaciones numéricas del proceso ponen de manifiesto que sería posible obtener un alto rendimiento de aceite con elevado contenido de tocoferoles e inferiores de fosfolípidos y ceras a 60ºC utilizando menos etapas.

Palabras Clave: *Extracción por solvente, difusión, tocoferoles, fosfolípidos, ceras*

2000 AMS Subjects Classification: 90-02

INTRODUCCIÓN

Desde el punto de vista de la extracción, los aceites vegetales pueden ser considerados como un sólo componente, debido a que todos los glicéridos son solubles en hexano, sin embargo otros componentes se extraen junto con los triglicéridos. Algunos de estos compuestos minoritarios tienen propiedades pro o antioxidantes (ej. tocoferoles, metales y ácidos grasos libres), mientras que otros deben ser removidos en el proceso de refinación (ej. ácidos grasos libres, fosfolípidos y ceras). Hoy en día, un modelo de extracción continua que tenga en cuenta la extractabilidad de los componentes menores no está disponible en la literatura, conocimiento que podría contribuir al diseño efectivo de la extracción de aceite por solvente.

Modelos matemáticos con ecuaciones simplificadas han sido presentados por varios autores [1-3], quienes analizaron diferentes sistemas con flujo cruzado en contracorriente, considerando las zonas de carga y descarga. El modelo desarrollado por Carrín y Crapiste [3] representa el proceso de extracción mediante un modelo en estado transitorio en dos dimensiones introduciendo el concepto de diferentes categorías de aceites en términos de disponibilidad y cinética de extracción.

El objetivo del presente trabajo es introducir modificaciones a un modelo existente [3], extendiendo el análisis a la extracción diferencial de algunos compuestos minoritarios de interés industrial, tales como ceras, fosfolípidos y tocoferoles. Como sólo se encuentran disponibles datos de extractabilidad de fosfolípidos y tocoferoles [4], estudios cinéticos de la extracción de ceras a partir de collets de girasol en un reactor batch utilizando n-hexano como solvente fueron realizados.

MODELO DE EXTRACCIÓN

Las modificaciones fueron introducidas sobre un modelo desarrollado sobre un extractor industrial tipo De Smet [3], considerando disponibilidad de aceite en los collets, flujo en contracorriente en los poros del sólido y en el lecho, transferencia de masa entre los collets y la miscela, transporte de la miscela entre las distintas secciones de percolación, zonas de carga y descarga y operación en régimen transitorio del extractor. Las condiciones asumidas para el desarrollo del modelo fueron mantenidas en este trabajo, las ecuaciones fueron extendidas a los compuestos minoritarios.

Ecuaciones de transferencia de aceite y compuestos minoritarios

Los componentes dentro de los collets están compuestos por dos categorías [5]: el componente libre, fácilmente extraíble de las partículas de sólido y equivalente a la etapa de lavado, C_1, y el componente ligado, difícil de extraer que corresponde a la etapa de difusión, C_2. El contenido del componente crítico, C_{scr}, corresponde a la concentración máxima de C_2 en el sólido. Aplicando balances de masa a las partículas de collets en estado estacionario, cuando la transferencia de masa convectiva ocurre entre la superficie del collet y el seno de la fase fluida el flujo difusivo para cada componente que abandona la matriz porosa (c) en su i-categoría se

obtiene como [7]: $n_{ci} = \frac{d_p}{6} \frac{\rho}{\left(\frac{d_p}{6\,k_L} + \frac{1}{K_x E_f}\right)} (K_c\, C_{sci} - C_{ci})$; con: $E_f = \frac{3}{\varphi}\left(\coth\varphi - \frac{1}{\varphi}\right)$, $\varphi = \frac{d_p}{2}\left(\frac{K_x}{D_{oh}}\right)^{0.5}$, $K_x = \frac{D_i\,\xi}{\rho\, K\left(d_{ip}/2\right)}$, $\xi = \frac{4\,\rho_s}{d_{po}\,\varepsilon_p}$, donde E_f es el coeficiente de difusión efectiva y φ es el módulo de Jüttner , el cual relaciona la difusión en la partícula (D_i) y la difusión en el líquido (D_{oh}).

Balances de masa

La expresión del balance de masa del componente en la fase líquida y en el collets (incluyendo la miscela ocluida) en una porción del lecho, expresado en forma adimensional resulta:

a) Seno de la fase líquida o "bulk":
$$\frac{\partial(\rho\, C_{ci})}{\partial t^*} = -\frac{\partial\left(v_z^*\, C_{ci}\, \rho\right)}{\partial z^*} - \left(\frac{L}{W}\right)\frac{\partial\left(v_y^*\, C_{ci}\, \rho\right)}{\partial y^*} + \left(\frac{d_p}{L}\right)\left(\frac{\partial}{\partial z^*}\left(\frac{v_z^*\,\rho}{Pe}\frac{\partial C_{ci}}{\partial z^*}\right)\right) + \left(\frac{d_p}{L}\right)\left(\frac{L}{W}\right)^2\left(\frac{\partial}{\partial y^*}\left(\frac{v_y^*\,\rho}{Pe}\frac{\partial C_{ci}}{\partial y^*}\right)\right) + \left(\frac{1-\varepsilon_b}{\varepsilon_b}\right)\left(a_p\,\frac{L}{v_m}\right) n_{ci}$$

En esta ecuación, la densidad de miscela (ρ) se consideró como función de la concentración de aceite, simplificación que se basa en la hipótesis de que los compuestos minoritarios tienen baja incidencia en estas propiedades, debido a sus bajas concentraciones.

b) Collets y miscela ocluida: $\frac{\partial C_{sci}}{\partial t^*} = -\frac{a_p}{\rho_s}\frac{L}{v_m} n_{ci} - \frac{m_s}{\rho_s\, dz^*\, X\, W\, L\,(1-\varepsilon_b)}\left(\frac{L}{v_m}\right)\frac{\partial(C_{sci})}{\partial y^*}$

c) Bandejas: La miscela es esparcida en las bandejas con una distribución no uniforme de la concentración de componente de la categoría i. Considerando una mezcla rápida de la miscela en el interior de las bandejas, puede asumirse que esta concentración en la bandeja es uniforme, pero dependiente del tiempo para el régimen transitorio. Aplicando la ley de conservación a una bandeja de volumen V_t, la ecuación que tiene en cuenta el cambio de concentración del componente de la categoría i en la miscela en las bandejas intermedias (C_{mcis}) es:

$$\frac{d(C_{mcis}\,\rho_{ms})}{dt^*} = \left(\varepsilon_b\, L\, X / V_t\right)\left(W \int_{y_{fds}^*}^{y_{lds}^*} (C_{ciL}\,\rho_L)\, dy^* - W_e\, C_{mcis}\,\rho_{ms}\right)$$

La primer bandeja (1) recibe la miscela que sale de la primera etapa, y la miscela que corresponde a la zona de drenaje. La ecuación correspondiente (C_{mci1}) es:

$$\frac{d(C_{mci1}\,\rho_{m1})}{dt^*} = \left(\varepsilon_b\, L\, X / V_t\right)\left(W \int_{y_{fd1}^*}^{y_{ldd}^*} (C_{ciL}\,\rho_L)\, dy^* - (W_e + W_d)\, C_{mci1}\,\rho_{m1}\right)$$

La última bandeja (PN), la concentración másica de la miscela de salida se establece por integración sobre la zona de drenaje de esta bandeja. La expresión resultante es la siguiente (C_{mciPN}):

$$C_{mciPN}\,\rho_{mPN} = \frac{\int_{y_{fdPN}^*}^{y_{ldPN}^*} (C_{ciL}\,\rho_L)\, dy^*}{\int_{y_{fdPN}^*}^{y_{ldPN}^*} dy^*}$$

La concentración promedio del componente en la miscela ocluida en los poros, en una posición horizontal fija del lecho $\left(C_{sci}^a\right)$ es evaluada a través de la siguiente expresión: $C_{sci}^a = \int_0^1 C_{sci}\, dz^*$

El set de ecuaciones, con las condiciones iniciales y de borde constituyen el modelo propuesto que fue resuelto utilizando el método de diferencias finitas para discretizar las derivadas espaciales, y el de sumas finitas para remplazar los términos que contienen integrales. Las ecuaciones diferenciales ordinarias fueron resueltas mediante el método numérico de Runge-Kutta de cuarto orden, con Mat-Lab 7.5®.

RESULTADOS Y DISCUSIÓN

Propiedades del medio poroso (collets de girasol): ε_p= 0.24 ± 0.05, ε_b= 0.40 ± 0.08, d_p= 0.0189 ± 0.0003 m, d_{ip} =6.18 10^{-4} m, ρ_s =893 ± 81 kg/m^3. Los valores de difusividad de la etapa difusiva (D_{2c}) y los valores de las constantes de equilibrio (K_c) del aceite, tocoferoles y fosfolípidos fueron extraídos de Baümler *et al.* [4], los parámetros relacionados con las ceras se obtuvieron a partir de datos originales de manera similar (Tabla 1).

Las ceras exhibieron una importante etapa de lavado, seguida inmediatamente por una absorción en el sólido. En consecuencia los valores de C_{scr} obtenidos son significativamente menores que su concentración inicial C_{s0}. Por otro lado los fosfolípidos no mostraron la etapa de lavado a 40 y 50°C. Las simulaciones fueron realizadas a 40, 50 y 60°C.

La partícula sometida a extracción a 60ºC se encuentra prácticamente agotada en aceite a partir de la bandeja número 4 a la 1 (C_{so} < 0.0005), mientras que a temperaturas menores el sólido abandona el extractor (etapa 1) con concentraciones residuales de aceite (C_{so}=0.031 y C_{so}=0.012 a 40 y 50ºC, respectivamente), Figura 1. La concentración de fosfolípidos en el sólido decrece lentamente a medida que el mismo avanza en el extractor (desde la bandeja 9 hacia la bandeja 1), confirmando que la extracción de los mismos es más acentuada al final del proceso (etapas 1-3), Figura 1. Los valores de concentración de tocoferoles en el sólido mostraron la misma tendencia observada para el aceite, siendo extraídos más rápidamente a 60ºC. La concentración de ceras en el sólido disminuye lentamente desde la entrada del sólido al extractor hasta la zona de descarga del sólido, donde comienzan a extraerse a mayor velocidad. A 60ºC la disminución de la concentración de ceras es más acentuada. Puede observarse a esta temperatura, Figura 1, una primera zona que corresponde a la etapa de lavado (entre la bandeja 9, ingreso del sólido al extractor, y la bandeja 7), luego la disminución de la concentración de ceras en el sólido es más lenta.

Tabla 1. Difusividades (D_{1c} y D_{2c}), concentraciones críticas (C_{sccr}) y constantes de equilibrio (K_c) del aceite y los compuestos minoritarios.

	D_{1c} (m^2/s)	D_{2c} (m^2/s)	C_{sccr} (kg/sólido inerte kg)	**K_c (C/C_s)**
Aceite – C_{Sc0}=0.307 (kg/sólido inerte kg)				
40ºC	$1.880\ 10^{-6}$	$1.356\ 10^{-8}$	0.2016	1.622
50ºC	$6.549\ 10^{-6}$	$1.683\ 10^{-8}$	0.1934	2.109
60ºC	$9.463\ 10^{-6}$	$2.247\ 10^{-8}$	0.1853	4.207
Fosfolípidos – C_{Sc0}=$2.548\ 10^{-3}$ (kg/sólido inerte kg)				
40ºC	------	$1.323\ 10^{-9}$	$2.548\ 10^{-3}$	0.412
50ºC	------	$1.498\ 10^{-9}$	$2.548\ 10^{-3}$	0.721
60ºC	$2.903\ 10^{-5}$	$4.229\ 10^{-9}$	$1.841\ 10^{-3}$	0.848
Tocoferoles – C_{Sc0}=$2.072\ 10^{-4}$ (kg/sólido inerte kg)				
40ºC	$1.360\ 10^{-7}$	$7.006\ 10^{-9}$	$1.822\ 10^{-4}$	0.450
50ºC	$1.218\ 10^{-7}$	$3.066\ 10^{-8}$	$1.880\ 10^{-4}$	0.456
60ºC	$2.981\ 10^{-6}$	$1.324\ 10^{-7}$	$1.777\ 10^{-4}$	1.305
Ceras – C_{Sc0}=$1.529\ 10^{-4}$ (kg/sólido inerte kg)				
40ºC	$3.832\ 10^{-7}$	$4.989\ 10^{-9}$	$7.671\ 10^{-5}$	0.649
50ºC	$2.944\ 10^{-7}$	$1.040\ 10^{-9}$	$4.076\ 10^{-5}$	1.228
60ºC	$5.724\ 10^{-6}$	$6.975\ 10^{-10}$	$5.068\ 10^{-5}$	1.738

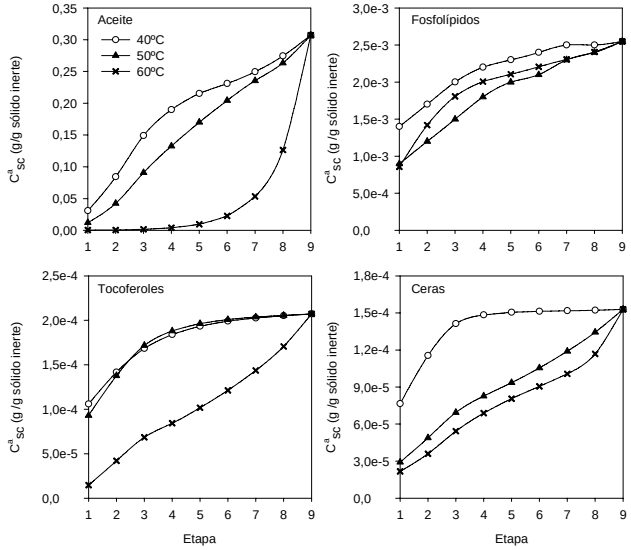


Figura 1. Concentraciones promedio del aceite y los compuestos minoritarios en la fase sólida como función de la temperatura de extracción. La etapa 9 corresponde a la entrada del sólido en el extractor.

La evolución de la concentración de aceite en la miscela y collets a través del extractor muestran un comportamiento en forma de olas en la parte superior del extractor (z*=0), debido al cambio en la calidad de la

miscela entre dos bandejas (Figuras 2a y 2b). Este comportamiento desaparece en el fondo del lecho (z*=1), donde se observa una disminución monótona de la concentración del aceite. En estado estacionario y a 40-50ºC este comportamiento se observa en casi todas las bandejas, mientras que a 60ºC sólo se observa en las bandejas cercanas a la zona de carga, donde ocurre la extracción de aceite libre (categoría C_1). Desde la mitad del lecho hasta la zona de drenaje (y*=0.5-1) se obtiene una meseta, debido a la lenta transferencia del aceite de difícil extracción (C_2). Esto sugiere un posible sobredimensionamiento del extractor cuando se opera a 60ºC al no producir las últimas etapas un incremento apreciable del rendimiento en aceite.

Las predicciones del modelo indican que sería posible trabajar a 60ºC utilizando menos etapas sin afectar el rendimiento de aceite. En estas condiciones las concentraciones de fosfolípidos y ceras en aceite que se obtendrían serían menores, efecto que podría resultar en un proceso de refinado menos exigente con inferiores costos y pérdidas de aceite. Además, el modelo ampliado presentado en este trabajo ha demostrado ser útil para predecir los perfiles de concentración de aceite y compuestos minoritarios y para representar los complejos fenómenos que tienen lugar durante el proceso de extracción en diferentes condiciones de funcionamiento. Este modelo puede ser empleado para diseñar y manipular extractores utilizando varios tipos de solventes y sólidos porosos. Sin embargo, datos experimentales obtenidos en condiciones de laboratorio son necesarios para la utilización del modelo.

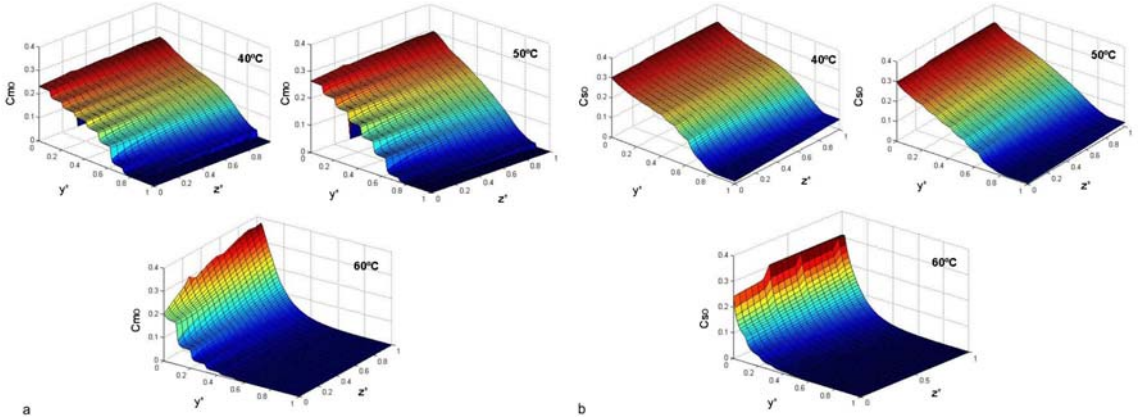


Figura 2. Perfiles de concentración: a)- aceite/miscela, b)- aceite/sólido, a lo largo del extractor operación en estado estacionario.

AGRADECIMIENTOS

Este trabajo fue posible gracias al financiamiento del CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica) y de la Universidad Nacional del Sur, Argentina.

REFERENCIAS

[1] VELOSO, G., KRIOUKOV, V. & VIELMO, H. (2005). *Mathematical Modeling of Vegetable Oil Extraction in a Countercurrent Crossed Flow Horizontal Extractor.* Journal of Food Engineering, 66(4), pp. 477–486.

[2] THOMAS, G., KRIOUKOV, V. & VIELMO H. (2005). *Simulation of Vegetable Oil Extraction in Counter-Current Crossed Flows using the Artificial Neuronal Network.* Chemical Engineering Processes, 44, pp.581-592.

[3] CARRÍN, M & CRAPISTE, G. (2008). *Mathematical Modeling of Vegetable Oil Solvent Extraction in a Multistage Horizontal Extractor.* Journal of Food Engineering, 85, pp. 418-425.

[4] BAÜMLER, E., CRAPISTE, G. & CARELLI, A. (2010). *Oil Solvent Extraction: Kinetic Study of Major and Minor Compounds.* Journal of the American Oil Chemists Society, 87(12), pp.1489–1495.

[5] FAN, H., MORRIS, J. & WAKEHAM, H. (1948). *Diffusion Phenomena* in Solvent Extraction of Peanut Oil. Industrial & Engineering Chemistry, 40, pp. 195–199.

[6] WAKAO, N. & KAGUEI, S. (1982). *Heat and Mass Transfer in Packed Beds*. London: Gordon and Breach Science Publishers.

[7] CARELLI, A., FRIZZERA, L., FORBITO, P. & CRAPISTE, G. (2002). *Wax Composition of Sunflower Seed Oils.* Journal of the American Oil Chemists Society, 79, pp. 763-768.

ANÁLISIS EXERGÉTICO DEL GENERADOR DE VAPOR DE 350 MW A CARGAS PARCIALES

Juan A. Jiménez[1]†, Guillermo Jarquin[2] ‡ María D. Durán[3] †‡ y Javier García Gutiérrez[1]

† Unidad Académica Profesional Nezahualcóyotl, Universidad Autónoma del Estado de México, Av. Bordo de Xochiaca s/n Col. Benito Juárez, 57000 Cd. Nezahualcóyotl. Edo. De México, jjimenez@uaemex.mx, www.uaemex.mx

‡ Instituto Politécnico Nacional, SEPI-ESIME-Culhuacan, Av. Santa Ana No. 1000, Edif.2, tercer piso, Colonia San Francisco Culhuacán, Coyoacán, C.P. 04430. México, D.F., gjarquin@ipn.mx, www.esimecu.ipn.mx

†‡ Facultad de Ingeniería, Universidad Autónoma del Estado de México, Cerro de Coatepec s/n, Ciudad Universitaria C.P. 50100, Toluca, Estado de México, mduran@fi.uaemex.mx, www.fi.uaemex.mx

Resumen: Se presenta el análisis Exergético de generador de vapor de 350 MW, instalado en la central termoeléctrica "Villa de Reyes" ubicada en San Luis Potosí, México, operando a diferentes regímenes de carga (100%,75%, 50% y 25%). Para tal propósito se realizó un modelo de balance de exergía considerando el generador de vapor como un volumen de control .La exergía de cada uno de los flujos participantes en el modelo (combustible, aire, agua de alimentación, extracción de vapor primario, secundario y gases de escape) fueron calculados para diferentes regímenes de carga. Los resultados de la destrucción de exergía son presentados; mostrando que la generación de irreversibilidades se maximiza cuando el generador de vapor es operado a regímenes de carga menores (50% y 25%).

Palabras claves: *Análisis Exergético, generador de vapor, cargas parciales*
2000 AMS Subjects Classification: 80A20

1. INTRODUCCIÓN

El generador de vapor de 350 MW instalado en la central termoeléctrica "Villa de Reyes", es un generador de vapor de circulación forzada, recalentador radiante – Convectivo, tipo intemperie que quema combustóleo [1].

El objetivo de este estudio es determinar el grado de destrucción de exergía de este generador de vapor operando a cargas parciales. Para tal propósito se emplea la metodología exegética [2], que supera las limitaciones que presenta un análisis basado en la primera ley (análisis energético) y puede claramente, identificar la degradación de la energía en numerosos procesos [3].

Se realiza un balance de exergía basado en las características termodinámicas de los flujos de entrada y salida del generador de vapor considerándolo como un volumen de control (Fig. 1) se determina la cantidad de exergía destruida en cada uno de los regímenes de operación (100%,75%, 50% y 25%).

2. METODOLOGÍA

El análisis exergético es una técnica basada en el concepto de exergía, que busca el uso eficiente de los recursos energéticos, provee una medida para evaluar la magnitud de la energía suministrada en relación a la energía proporcionada o transformada en una planta o elemento analizado[4].

En ausencia de efectos nucleares, magnéticos, eléctricos y efectos de tensión superficial, la Exergía E puede definirse en función de cuatro componentes; exergía física (E^{PH}), exergía cinética (E^{K}), exergía potencia (E^{P}) y exergía química (E^{CH}), esto es [5]:

$$E = E^{PH} + E^{K} + E^{P} + E^{CH} \quad (1)$$

Para un volumen de control la exergía física de flujo o corriente (ψ) (kJ/kg) se denota con la expresión [6]:

$$\psi = (h - h_0) - T_0 (s - s_0) + \frac{v^2}{2} + gz \qquad (2)$$

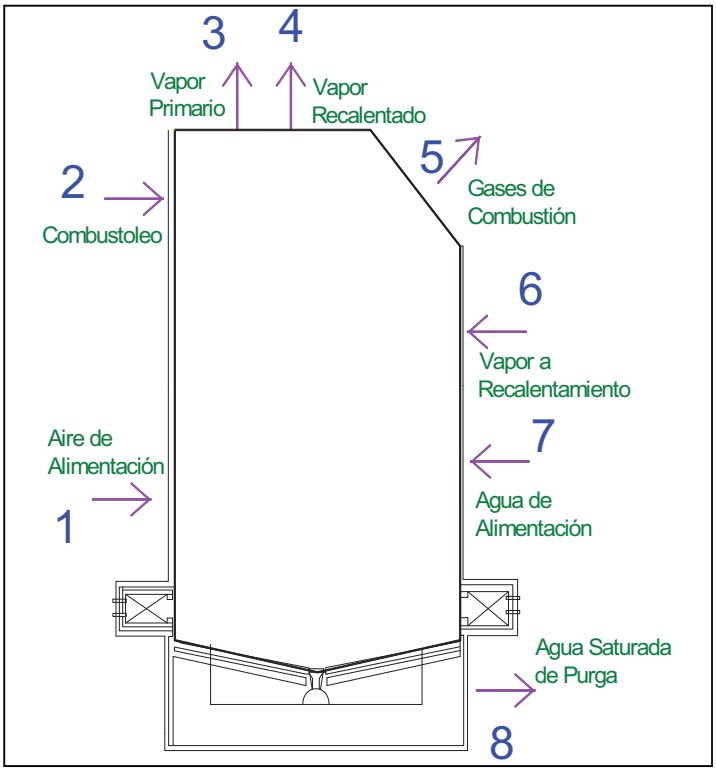


Figura 1: Volumen de control asociado al generador de vapor

El cambio de entalpía para un gas ideal durante un proceso que pasa de un estado 1 al 2 se determina con la siguiente expresión [6]:

$$h - h_0 = C_p (T - T_0) \qquad (3)$$

Donde $C_p (kJ/kgK)$, es el calor específico a presión constante y T, T_0 son las temperaturas del estado 2 y 1 respectivamente.

La variación de entropía para gas ideal considerando la variación de calores específicos constante se define como [6]:

$$s - s_0 = C_p \ln\left(\frac{T}{T_0}\right) - R \ln\left(\frac{P}{P_0}\right) \qquad (4)$$

Sustituyendo las ecuaciones (6) y (7) en (5) se obtiene la expresión de la Exergía física de flujo ψ (KJ/Kg) para gas ideal.

$$\psi = C_p\left(T - T_0 - T_0 \ln\left(\frac{T}{T_0}\right)\right) + RT_0 \ln\left(\frac{P}{P_0}\right) \quad (5)$$

Para el presente análisis se considerara la exergía química del combustóleo. La expresión que define la exergía química de combustibles líquidos es [3]:

$$E^{CH} = \dot{m}\left(1.0401 + 0.1728\frac{H}{C} + 0.043\frac{O}{C} + 0.2169\frac{S}{C}\left(1 - 2.0628\frac{H}{C}\right)\right)\cdot PCS \quad (6)$$

Donde E^{CH} (KW) es la exergía química del combustible, $\dot{m}$ (Kg/s) es el flujo másico de combustible, H,O,C y S son la composición química molar del hidrogeno, Oxigeno, Carbón y Azufre respectivamente presentes en la mezcla del combustible.

3. CALCULO EXERGÉTICO DEL GENERADOR DE VAPOR A CARGAS PARCIALES

El grado de destrucción de exergía (ED) del generador de vapor a cargas parciales, fue calculado aplicando un balance de exergía para el volumen de control mostrado en la figura 1:

$$ED = \left(\dot{m}_1 e_1 + \dot{m}_2 e_2 + \dot{m}_6 e_6 + \dot{m}_7 e_7\right) - \left(\dot{m}_3 e_3 + \dot{m}_4 e_4 + \dot{m}_5 e_5 + \dot{m}_8 e_8\right) \quad (7)$$

Donde $\dot{m}_n$ (Kg/s) son los flujos másicos en cada uno de los nodos del volumen de control., e_n (KJ/Kg) es la exergía de flujo en cada uno de los nodos del volumen de control, donde

$$e_n = e^{PH} + e^{CH} \quad (8)$$

La exergía química del aire y de los productos de la combustión no son considerados para el presente cálculo, la exergía física de los gases de combustión fueron calculados como gas ideal. Se considera que el aire de alimentación se inyecta a presión y temperatura de referencia. El combustible considerado es combustóleo con una composición química molar de 83.64%C, 11.3%H, 4.2% S, un poder calorífico superior de 41830 KJ/kg y se inyecta a la cámara de combustión a una temperatura de 135°C.

Las variables termodinámicas de operación y el cálculo de la Exergía física y química para régimen de carga 100% del generador de vapor, para el volumen de control de la fig. 1 se muestran en la tabla 1.

Tabla 1 Variables termodinámicas y Exergía del generador de vapor operando a 100% de carga.

No.	Característica del Flujo	Masa (Kg/s)	Presión (Bar)	Temperatura (K)	Entalpia (KJ/Kg)	Entropía (S) (KJ/Kg°C)	Exergía Física (KW)	Exergía Química (KW)
1	Aire de Alimentación	338.4	1.013	293.15	-	-	0	0
2	combustóleo liquido	21.507	18.495	408.15	-	-	0	965347.267
3	Gases de combustión	367.2	1.013	419.15	-	-	7818.257	-
4	Vapor sobrecalentado	288.3	171.126	814.15	3400.664	6.4087	921054.503	-
5	Vapor recalentado	259.2	32.166	814.15	3545.818	7.316	860790.670	-
6	Vapor a recalentamiento	259.2	33.833	587.15	3097.35	7.6961	742595.6382	-

7	Agua de alimentación	301.2	182.698	511.15	1030.66	2.653	270992.363	-
8	liquido saturado de Purga	2.9	179.952	630.08	1734.6186	3.8763	4562.103361	-

Usando la ecuación (10) y los datos de la exergía calculada para cada una de los regímenes de carga del generador de vapor, el grado de destrucción de exergía se muestra en la tabla 2.

Tabla 2 Grado de Exergía destruida del generador de vapor operando a cargas parciales

Carga Térmica (%)	Exergía Destruida (KW)	Exergía Destruida (%)
100%	184709.7	19.1
75%	130200.5	17.6
50%	190584.7	36.4
25%	88778.2	34.5

4. CONCLUSIÓN

El análisis exergético del generador de vapor operando a cargas parciales fue realizado, encontrándose que la destrucción de Exergía se incrementa cuando el generador de vapor opera a cargas parciales bajas, siendo de 190,584.7 KW cuando se opera al 50% y de 88,778.2 KW cuando se opera al 25% de carga. Estos valores respecto a la cantidad de Exergía suministrada al volumen de control son máximos y representan los regímenes de operación con mayor generación de irreversibilidades en el ciclo.

5. AGRADECIMIENTOS

Los autores agradecen el apoyo brindado por el Consejo Mexiquense de Ciencia y Tecnología (COMECYT) y a la SIEA de la Universidad Autónoma del Estado de México.

REFERENCIAS

[1] COMISIÓN FEDERAL DE ELECTRICIDAD, *Manual de operación y Parámetros del proceso térmico del generador de vapor de la central termoeléctrica Villa de Reyes*, CFE, 1997.

[2] BEJAN A., TSATSARONIS G., MORAN M., *Thermal Desing and Optimization. New York*, John Wiley & Sons, 1996.

[3] PALMA, R. SILVIA, *Análisis Exergético, Termoeconómico y ambiental de un sistema de generación de energía, estudio del caso de la central termoeléctrica de rio de Janeiro*, Departamento de Ingeniería Mecánica, Universidad de Brasil, 2007.

[4] TSATSARONIS, G. *Thermoeconomic Analysis and Optimization of Energy System*. Prog. Energy Combustion Science, vol. 19 (1993), pp. 227-257.

[5] FREDERICK J, BARCLAY, *Combined Power And Process- an Exerfy Approach*, Professional Engineering Publishing, 1998.

[6] YUNUS A. CENGEL & MICHAEL A. BONES, *Termodinámica*, Quinta edición, McGraw-hill. México, 2000.

SOLUCIONES ANALÍTICAS USANDO FUNCIONES DE GREEN PARA EL MODELO DE PENNES DE DIFUSIÓN DE CALOR EN MEDIOS BIOLÓGICOS

Mauricio A. Giordano[1], Gustavo Gutiérrez[2], Julio C. Massa[1]

[1]*Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de Córdoba, Casilla de correo 916, 5000 Córdoba, Argentina, mauricioagiordano@gmail.com, http://www.efn.uncor.edu*

[2]*Department of Mechanical Engineering, University of Puerto Rico Mayagüez Campus, PO Box 9000 (00680) Mayagüez, Puerto Rico - E-mail: ggutierr@me.uprm.edu*

Resumen: Se derivan soluciones analíticas para el modelo de Pennes de difusión de calor en medios biológicos para predecir los perfiles de temperatura que se pueden obtener en tumores durante el tratamiento del cáncer por hipertermia usando fluidos magnéticos. Éstos fluidos magnéticos son suspensiones coloidales, en agua o aceite, de nano-partículas que bajo la acción de un campo magnético variable en el tiempo, disipan energía en el medio por efectos de relajación magnética. Se muestra que es posible elevar la temperatura del tejido a niveles terapéuticos (42~45 °C) utilizando, para el campo magnético, intensidades y frecuencias seguras y tolerables por el organismo y que es crucial lograr que las nano-partículas se distribuyan en la superficie del tumor para obtener un perfil de temperaturas más próximo al ideal que es aquel que mantiene la temperatura terapéutica constante dentro del tumor y la temperatura normal del cuerpo (37 °C) en el tejido sano.

Palabras claves: *Nano-partículas magnéticas, Cáncer, Hipertermia local, Ecuación Bio-heat.*
2000 AMS Subjects Classification: 35K05- 80A99

1. INTRODUCCIÓN

Durante el tratamiento del cáncer por hipertermia se eleva la temperatura de las células hasta alcanzar valores comprendidos en el rango de 42 a 45 ºC que producen alteraciones de ciertas proteínas, induciendo la muerte celular. La utilización de partículas magnéticas y la subsecuente aplicación de un campo electromagnético variable para generar calor en tejidos cancerosos se remonta al año 1957 [1]. El advenimiento de la nano-tecnología y el desarrollo de ferro-fluidos permitió utilizarlos como portadores de las nano-partículas a ser depositadas en el tejido canceroso dando lugar a lo que hoy se conoce en la literatura como *Magnetic Fluid Hyperthermia* (**MFH**). Dichos fluidos son suspensiones coloidales de partículas de diámetros en el rango de 6 a 100 nm.

El modelo más utilizado para representar la difusión de calor en sistemas biológicos es el propuesto por Pennes [2], también conocido como 'bio-heat equation'. Entre los trabajos previos utilizando funciones de Green para resolver el modelo de Pennes se puede citar, entre otros, el trabajo de Deng y Liu [3] donde se presentan resultados interesantes pero con poca aplicación en el campo de MFH. El objetivo del presente trabajo es cuantificar, utilizando el modelo de Pennes, los perfiles de temperatura posibles de ser alcanzados durante el tratamiento del cáncer por hipertermia local al considerar dos tipos de distribuciones de las nano-partículas que pueden presentarse en el interior de un tumor de forma esférica.

2. FORMULACIÓN MATEMÁTICA

En este trabajo se adopta el modelo de Pennes para la difusión de calor en tejidos debido a: *i*) su simpleza matemática comparada con otros modelos disponibles en la literatura y *ii*) su buena correlación con mediciones experimentales. Este modelo es la ecuación de difusión de calor estándar para medios estacionarios en la cual se incluyen dos nuevos términos en el balance de energía en un elemento diferencial para tener en cuenta los efectos del metabolismo. La expresión general, válida en cualquier sistema de coordenadas, es:

$$\rho c \frac{\partial T_{(\mathbf{X},t)}}{\partial t} = k \nabla^2 T_{(\mathbf{X},t)} + \rho_s c_s \omega_s \left[T_a - T_{(\mathbf{X},t)} \right] + q_{met(\mathbf{X},t)} + Q_{gen(\mathbf{X},t)} \quad (1)$$

donde $\mathbf{X} = (x_1, x_2, x_3)$ es un vector posición, $T_{(\mathbf{X},t)}$ es la temperatura del tejido, ρ, c y k son respectivamente la densidad, calor específico y conductividad del tejido, ρ_s, y c_s son la densidad y calor específico de la sangre, ω_s es el coeficiente de perfusión sanguínea, T_a es la temperatura arterial, $q_{met(\mathbf{X},t)}$ es la tasa de calor generada por el metabolismo y $Q_{gen(\mathbf{X},t)}$ es el término de generación debido a las nano-partículas.

En este trabajo se analizan dos casos de interés práctico en los que se asume generación de calor axi-simétrica. De esta manera la Ecuación (1) se reduce a su versión unidimensional en coordenadas esféricas:

$$\frac{\partial \Theta_{(r,t)}}{\partial t} = \frac{\alpha}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \Theta_{(r,t)}}{\partial r} \right) - \gamma^2 \Theta_{(r,t)} + \frac{q_{met(r,t)}}{\rho c} + \frac{Q_{gen(r,t)}}{\rho c} \quad (2)$$

donde $\Theta_{(r,t)} = T_{(r,t)} - T_a$, $\gamma^2 = \rho_s c_s \omega_s / (\rho c)$ y $\alpha = k/(\rho c)$ es la difusividad térmica del tejido. $q_{met\,(r,t)}$ presenta pequeñas variaciones según el órgano dónde se evalúe y será considerado como constante en la formulación.

2.1. MODELO DE GENERACIÓN

El modelo de generación de calor adoptado en este trabajo fue desarrollado por Rosensweig [4] para estimar los perfiles de temperatura obtenibles en aplicaciones de MFH. La expresión resultante cuantifica la potencia por unidad de volumen en función de las propiedades físicas de las partículas y del fluido base, y de los parámetros del campo magnético (intensidad y frecuencia).

Para una muestra monodispersa de partículas, la disipación está dada por la expresión:

$$P = \mu_0 \, \pi \, \chi_0 \; f \; H_0^2 \; \frac{2\pi \, f \, \tau}{1+\left(2\pi \, f \, \tau\right)^2} \tag{3}$$

donde P es la potencia en $[\mathrm{W/m^3}]$, $\mu_0 = 4\pi \times 10^{-7}$ $[\mathrm{Tm/A}]$ es la permeabilidad en el vacío, χ_0 es la susceptibilidad magnética, $f = \omega/2\pi$ $[\mathrm{Hz}]$ es la frecuencia del campo, H_0 $[\mathrm{A/m}]$ es la amplitud del campo magnético variable y τ [s] es la constante de tiempo efectiva del proceso de relajación magnética y viene dada por:

$$\tau = \frac{\tau_B \; \tau_N}{\tau_B + \tau_N} \qquad \text{donde:} \quad \tau_B = \frac{3\eta \, V_H}{k_B T} \quad \text{y} \quad \tau_N = \tau_0 \sqrt{\frac{\pi k_B \, T}{2 \, K V_M}} \; e^{\left(\frac{KV_M}{k_B T}\right)} \tag{4}$$

τ_B es el tiempo característico de la relajación Browniana, τ_N es el tiempo característico de la relajación de Néel, τ_0 es la constante de tiempo de relajación magnética, η es la viscosidad del fluido portador, V_H es el volumen hidrodinámico, V_M es el volumen del núcleo magnético de las partículas, $k_B = 1{,}38 \times 10^{-23}$ J/K es la constante de Boltzmann, T es la temperatura absoluta y K es la constante anisotrópa.

Inspeccionando la Ecuación (3) puede mostrarse que para una amplitud y frecuencia fija del campo magnético, la disminución en la tasa de generación provocada por el aumento de temperatura y subsecuente cambio en las constantes de tiempo τ_B y τ_N es despreciable dentro del rango de temperaturas válidas para MFH. Por lo tanto, se obtiene una buena aproximación evaluando la disipación P para la temperatura de referencia T_a y suponiendo que ese valor se mantiene constante durante el tratamiento.

3. SOLUCIONES ANALÍTICAS

Las soluciones analíticas pueden obtenerse utilizando la función de Green, G, que corresponde a la Ec. (2) para un dominio semi-infinito y que resulta de resolver la siguiente ecuación diferencial:

$$\frac{\partial G_{(r,t|r',\tau)}}{\partial t} \; = \; \alpha \, \frac{\partial^2 G_{(r,t|r',\tau)}}{\partial r^2} \; + \; \frac{2\alpha}{r} \; \frac{\partial G_{(r,t|r',\tau)}}{\partial r} \; - \gamma^2 \; G_{(r,t|r',\tau)} + \; \frac{\delta_{(r-r')} \; \delta_{(t-\tau)}}{4\pi r^2} \tag{5}$$

donde δ es el delta de Dirac. La Ecuación (5) pone de manifiesto el significado físico de la función de Green: es la distribución de temperatura en el medio debido a un pulso de calor instantáneo actuando en $r = r'$, en el tiempo $t = \tau$. Los detalles de la derivación de la solución de la Ecuación (5) se encuentran disponibles en libros de texto de conducción de calor, e.g. [5] y serán omitidos. La solución, en términos de la función de Green, de un problema de difusión de calor gobernado por la Ec. (2) viene dada en forma general por [4]:

$$\Theta_{(r,t)} = T_{(r,t)} - T_a = \int_{r'=0}^{\infty} r'^2 \, G_{(r,t|r',\tau)}\big|_{\tau=0} \; F_{(r')} dr' \; + \; \frac{\alpha}{k} \int_{\tau=0}^{t} \int_{r'=0}^{\infty} r'^2 \, G_{(r,t|r',\tau)} \; q_{(r',\tau)} \; dr' \, d\tau \tag{6}$$

donde $F_{(r)}$ es la condición inicial del problema y $q_{(r,t)}$ representa todas las fuentes/sumideros de calor, en este caso $q_{(r,t)} = q_{met(r,t)} + Q_{gen(r,t)}$. Debido a que se trabaja con dominios semi-infinitos las condiciones de borde exigen que las integrales converjan y que:

$$\Theta_{(0,\,t)} = \text{ finito} \qquad\qquad \Theta_{(r\to\infty,t)} \to 0 \tag{7}$$

Además, la condición inicial establece que todo el dominio se encuentra a la temperatura basal del cuerpo ($T_{(r,0)} = T_a$), entonces

$$\Theta_{(r,0)} = 0 \tag{8}$$

La función de Green, solución de la ecuación diferencial (5) que satisface condiciones iniciales y condiciones de borde, es

$$\left[\left.\frac{\partial G}{\partial r}\right|_{r=0} = 0, \quad G\big|_{t=0} = 0 \right] \quad \rightarrow \quad G_{(r,t|r',\tau)} \; = \; \frac{H_{(t-\tau)} \; e^{-\gamma^2 (t-\tau)}}{2 r r' \sqrt{\pi \alpha \left(t-\tau\right)}} \left[e^{-\frac{(r-r')^2}{4\alpha(t-\tau)}} - e^{-\frac{(r+r')^2}{4\alpha(t-\tau)}} \right] \tag{9}$$

3.1. GENERACIÓN DEBIDA A UNA DISTRIBUCIÓN VOLUMÉTRICA DE PARTÍCULAS MAGNÉTICAS

En el modelo matemático para una fuente con forma esférica de radio *R*, el término fuente es

$$Q_{gen(r,t)} = g\left[1 - H_{(r-R)}\right] \tag{10}$$

donde g [W/m^3] es la intensidad de la fuente y $H_{(r\text{-}R)}$ es la función de Heaviside o escalón unitario. El factor [1-$H_{(r\text{-}R)}$] en el término fuente implica que la integración que involucra la generación debida a las partículas debe efectuarse entre $r' = 0$ y $r' = R$. Reemplazando el valor de $Q_{gen(r,t)}$ dado en la Ec. (10) en el término correspondiente a $q_{(r,t)}$ en la Ec. (6) y teniendo en cuenta las Ecuaciones (8) y (9), se obtiene

$$\Theta_{(r,t)} = \frac{\alpha}{k\gamma^2} q_{met}\left(1 - e^{-\gamma^2 t}\right) + \frac{\alpha g}{2kr}\int_{\tau=0}^{t}\int_{r'=0}^{R} \frac{r' H_{(t-\tau)} e^{-\gamma^2(t-\tau)}}{\sqrt{\pi\alpha(t-\tau)}}\left[e^{-\frac{(r-r')^2}{4\alpha(t-\tau)}} - e^{-\frac{(r+r')^2}{4\alpha(t-\tau)}}\right] dr'\,d\tau \tag{11}$$

El primer término de la Ecuación (11) es el resultado de realizar las integraciones correspondientes en las variables r' y τ. Las integrales del segundo término de la Ecuación (11) se deben evaluar numéricamente.

3.2. GENERACIÓN DEBIDA A UNA DISTRIBUCIÓN SUPERFICIAL DE PARTÍCULAS MAGNÉTICAS

El modelo matemático para una fuente con forma de cascarón esférico de radio *R* y de espesor infinitesimal liberando calor en forma continua con intensidad g_P [W] constante está dado por:

$$Q_{gen(r,t)} = \frac{g_P}{4\pi r^2}\delta_{(r-R)} \tag{12}$$

Igual que en el caso de generación volumétrica, la distribución de temperatura en el medio se obtiene reemplazando la Ec. (12) en la Ec. (6) y teniendo en cuenta las Ecuaciones (8) y (9):

$$\Theta_{(r,t)} = \frac{\alpha}{k\gamma^2} q_{met}\left(1 - e^{-\gamma^2 t}\right) + \frac{\alpha g_P}{8\pi r k}\int_{\tau=0}^{t} \frac{e^{-\gamma^2(t-\tau)}}{\sqrt{\pi\alpha(t-\tau)}}\left[e^{\frac{-(r-r_0)^2}{4\alpha(t-\tau)}} - e^{\frac{-(r+r_0)^2}{4\alpha(t-\tau)}}\right] d\tau \tag{13}$$

En la Ecuación (13), el primer término resulta de evaluar las integrales que involucran la generación del metabolismo; y el segundo corresponde al término de generación Q_{gen}, donde se ha utilizado la propiedad de muestreo de la función delta de Dirac, a saber $\int_0^\infty f_{(r)}\delta_{(r-R)}dr = f_{(R)}$.

4. RESULTADOS Y DISCUSIÓN

Para presentar gráficamente las distribuciones de temperatura predichos por las soluciones de la sección anterior, se consideran los valores de las propiedades termo-físicas del tejido publicados en [6]: $k = 0{,}5$ [W/mK], $\rho = \rho_s = 1000$ kg/m^3, $c = c_s = 3800$ J/kgK, $q_{met} = 700$ W/m^3, $\omega_s = 0{,}0005$ 1/s. Por otra parte, el valor de la tasa de generación corresponde a un ferro-fluido de partículas de magnetita en base agua.

Considerando nano-partículas de 12 nm de diámetro, una concentración de 0,01 g de magnetita por gramo de tejido, un campo magnético de intensidad igual a 6,5 kA/m y una frecuencia de 70 kHz, el modelo de Rosensweig, Ec. (3), predice que la potencia a ser disipada en el medio es 7,5x10^5 W/m^3. El valor asignado a la concentración de material magnético está por debajo del límite admisible para la magnetita (por toxicidad) y los parámetros del campo magnético pueden aplicarse en forma segura y ser tolerados por el paciente. Atkinson *et al.* [7] establecieron como criterio de seguridad que el producto (Hf) no exceda el valor 4,85x10^8 A/ms.

Si en las Ecuaciones (10) y (11) se anula la contribución del término de generación debido al material magnético, el término restante define el estado basal correspondiente a la condición metabólica del cuerpo de auto-regulación de la temperatura. Un aumento en la tasa de generación q_{met} se traduce en un aumento de la temperatura. Por otra parte, un incremento del coeficiente de perfusión ω_s implica una disminución de la temperatura ya que ω_s cuantifica la remoción de calor por convección e interviene en la solución a través de γ.

Al estimar el orden de magnitud del espesor necesario del cascarón esférico que albergará las nano-partículas hay que tener en cuenta que el valor máximo de disipación de energía está limitado por las cotas impuestas al campo magnético por cuestiones de seguridad y tolerancia. Para los valores adoptados se obtiene un espesor del orden de 1 mm.

$$\begin{bmatrix} R = 1\,\text{cm} \\ g_P = 0{,}9\,\text{W} \end{bmatrix} \quad P = \frac{g_P}{V} = \frac{g_P}{4\pi\left(r_2^3 - R^3\right)/3} = 7{,}5\times10^5\,\text{W/m}^3 \;\rightarrow\; r_2 \approx 1{,}1\,\text{cm} \;\rightarrow\; e = r_2 - R \approx 1\,\text{mm} \tag{14}$$

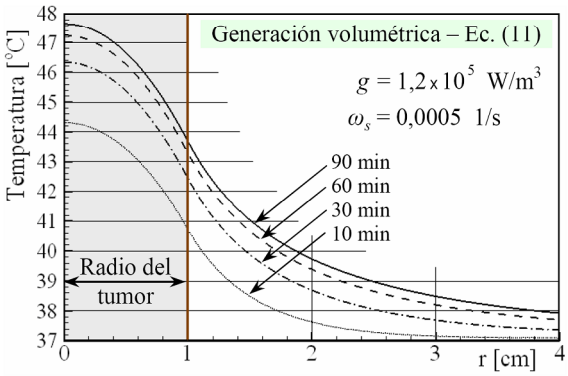


Figura 1: Perfil de temperatura - generación volumétrica

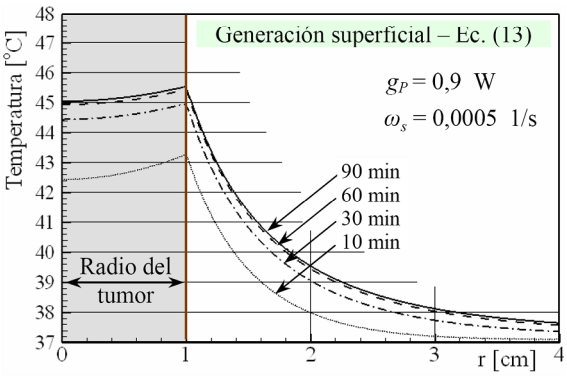


Figura 2: Perfil de temperatura - generación superficial

En la Figura 1 se muestra el perfil de temperatura producido por una distribución uniforme de partículas en un tumor de radio R = 1 cm para distintos tiempos de aplicación del campo magnético. $T_{(r,0)} = T_a = 37$ °C.

La Figura 2 muestra el perfil para el caso de una distribución superficial de nano-partículas magnéticas alrededor del tumor. Según la Ecuación (14), para alcanzar niveles de temperatura dentro de los valores terapéuticos para un tumor de R = 1 cm debe ubicarse una concentración de 0,01 g de material magnético en un casquete esférico de espesor $e \approx$ 1 mm. Se observa un decremento en la temperatura desde la superficie hacia el centro del mismo. Esto se debe a la remoción de energía por convección causada por el flujo sanguíneo, cuantificada en la ecuación diferencial gobernante por el término que contiene el coeficiente de perfusión ω_s. Si se considerara el caso $\omega_s = 0$, el perfil en la región $r = R$ sería constante. Es importante destacar que según la predicción del perfil de temperatura en estado estacionario (t = 90 min) mostrado en la Figura 1b, este decremento no resulta en temperaturas debajo de los niveles terapéuticos por lo que no afectaría el éxito del tratamiento.

La comparación de los resultados para ambas distribuciones de partículas sugiere que, si bien se presenta una penetración similar de la temperatura terapéutica en el tejido sano, una distribución superficial produce un perfil aproximadamente plano y de valor terapéutico dentro del tejido canceroso y, a la vez, un decaimiento un tanto más acentuado desde la superficie del tumor (r = 1 cm) hacia el tejido sano, siendo estas características del perfil más favorables para el tratamiento. Por ejemplo, para r = 3 cm se aprecia una diferencia de temperatura en los perfiles de 0,5 ºC. Por otra parte, para el caso de distribución volumétrica, es necesario sobrepasar los límites terapéuticos de temperatura en el centro de manera que temperaturas comprendidas entre 42 y 45 ºC estén presentes en toda la región que ocupa el tumor. Temperaturas superiores a los 45 ºC inducen un tipo de muerte celular conocido como necrosis que viene acompañado de inflamación de la zona necrótica razón por la cual es preferible evitar esta condición en beneficio del paciente.

5. CONCLUSIONES

Los resultados numéricos aquí obtenidos abren expectativas promisorias en cuanto a la utilización de ferrofluidos en el tratamiento de cáncer por hipertermia local. Las tasas de generación alcanzadas por las nano-partículas capaces de ser sintetizadas en la actualidad son suficientes para elevar la temperatura del medio a los valores que demanda la aplicación. Por otro lado, muestran que una distribución superficial de las partículas produce perfiles de temperaturas más próximos al ideal, que es aquel que presenta una temperatura constante y de valor terapéutico en la zona a tratar y, al mismo tiempo, temperatura normal del cuerpo en el tejido sano.

REFERENCIAS

[1] R. GILCHRIST, R. MEDAL, W. SHOREY, R. HANSELMAN, J. PARROTT AND C. TAYLOR, *Selective inductive heating of lymph*, Ann Surg, Vol. 146 (1957), pp. 596-606.

[2] H.H. PENNES, *Analysis of tissue and arterial blood temperatures in the resting human forearm*, J App Physiol, Vol. 85(1), (1948), pp. 5-34.

[3] Z.S. DENG AND J. LIU, *Analytical study on bioheat transfer problems with spatial or transient heating on skin surface or inside biological bodies*, Journal of Biomechanical Engineering, Vol. 124(6) (2002), pp. 638-649.

[4] R.E. ROSENSWEIG, *Heating magnetic fluid with alternating magnetic field.* Journal of Magnetism and Magnetic Materials, Vol. 252 (2002), pp. 370-374.

[5] M.N. OZISIK, *Heat Conduction*. John Wiley & Sons, New York, 1993.

[6] K.R. DILLER, J.W. VALVANO AND J.A. PEARCE, *Bioheat transfer*, The CRC Handbook of Thermal Engineering, F. Kreith *et al*., Editors. CRC Press LLC: Boca Raton. (2000), pp. 114-187.

[7] W.J. ATKINSON, I.A. BREZOVICH AND D.P. CHAKRABORTY, *Usable frequencies in hyperthermia with thermal seeds. IEEE Transactions on Biomedical Engineering*, Vol. 31(1) (1984), pp. 70-75.

APROXIMACIÓN NUMÉRICA DE LA SOLUCIÓN DE UN PROBLEMA DE TRANSFERENCIA DE CALOR Y MASA EN UN MEDIO POROSO

M. C. Olguin[♭], E. A. Santillan Marcus[♭,†] y M. C. Sanziel[♭,♯]

[♭]*Fac.Cs.Exactas, Ing. y Agrimensura, Universidad Nacional de Rosario, Av.Pellegrini 250, Rosario, Argentina*
[♯]*CIUNR, FCEIA - UNR*
[†]*Dpto. de Matemática, FCE, Univ. Austral, Paraguay 1950, Rosario, Argentina*
mcolguin@fceia.unr.edu.ar, esantillan@austral.edu.ar, sanziel@fceia.unr.edu.ar

Resumen: Se proponen diferentes abordajes numéricos para calcular los valores de la solución aproximada de un problema de transferencia de calor y masa con cambio de fase, en un medio poroso. Luikov [3], [4] estableció la formulación matemática del problema físico, en tanto que en [7] se realizó un análisis matemático teórico que permitió demostrar la existencia de solución única local en el tiempo, para el caso de un medio finito. En el presente trabajo se aplican diferentes métodos numéricos: aproximación de punto fijo, diferencias finitas con inmovilización del dominio y método cuasi-estacionario, a fin de encontrar la solución aproximada del problema, es decir se encuentran valores aproximados de la temperatura, la humedad y la frontera libre, y se comparan los resultados obtenidos.

Palabras clave: *transferencia de calor y masa, frontera libre, métodos numéricos*
2000 AMS Subject Classification: 35R35 - 45G15 - 65R20 - 80A22

1. INTRODUCCIÓN

Los problemas de transferencia de calor y masa que suceden en medios porosos, tales como evaporación, condensación, congelamiento, derretimiento, sublimación y desublimación, tienen una gran aplicación en procesos de separación, tecnología de alimentos, migración de calor, mezclas en suelos, etc. [1] [2].

La formulación matemática de la transferencia de calor y masa en cuerpos de capilares porosos fue establecida por A. V. Luikov [3]. Dos modelos diferentes fueron presentados por M. D. Mikhailov para resolver el problema de la evaporación de humedad líquida de un medio poroso [5], en tanto que para el problema de congelamiento (o desublimación) de un semiespacio húmedo poroso, Mikhailov también presentó una solución exacta [6]. Debido a la no-linearidad de los problemas, las soluciones usualmente involucran dificultades matemáticas. Sólo unas pocas soluciones exactas fueron halladas para casos ideales.

En [7] se realizó un análisis matemático teórico del congelamiento (desublimación) de humedad en un medio poroso finito, con una condición de flujo de calor en $x = 0$.

1.1. PRESENTACIÓN DEL PROBLEMA

Se considera el flujo de calor y humedad a través de un medio poroso finito durante el congelamiento. La posición del frente de cambio de fase al tiempo t está dado por $x = s\,(t)$. Este frente divide al cuerpo poroso en dos regiones.

Se nota con $u = u(x,t)$ la distribución de temperatura en la región de congelamiento, donde no hay movimiento de humedad, y con $v = v(x,t)$, y $w = w(x,t)$ la distribución de temperatura y la distribución de humedad en la región donde el calor y la humedad fluyen acoplados. Por conveniencia en la resolución del problema, se introduce una nueva función incógnita que acopla a v y w (a_1 y a_2 son las difusividades termales, a_m es la difusividad de humedad, δ es el coeficiente del gradiente termal):

$$z\,(x,t) = v\,(x,t) + \left[\frac{1}{\delta}\left(1 - \frac{a_2}{a_m}\right)\right] w\,(x,t)\,, \qquad s\,(t) < x < 1, 0 < t < T$$

Las ecuaciones que modelan el problema físico son

$$\frac{\partial u}{\partial t}\,(x,t) = a_1 \frac{\partial^2 u}{\partial x^2}\,(x,t)\,, \qquad 0 < x < s\,(t)\,, 0 < t < T \tag{1}$$

$$\frac{\partial v}{\partial t}\,(x,t) = a_2 \frac{\partial^2 v}{\partial x^2}\,(x,t)\,, \qquad s\,(t) < x < 1, 0 < t < T \tag{2}$$

$$\frac{\partial z}{\partial t}(x,t) = a_m \frac{\partial^2 z}{\partial x^2}(x,t), \qquad s(t) < x < 1, 0 < t < T \tag{3}$$

$$u(x,0) = \theta(x) \leq 0 \quad , \quad 0 < x < s(t) \tag{4}$$

$$v(x,0) = \Phi(x) \geq 0 \quad , \quad s(t) < x < 1 \tag{5}$$

$$v(1,t) = h(t) > 0 \quad , \quad 0 < t < T \tag{6}$$

$$z(x,0) = \eta(x) \quad , \quad s(t) < x < 1, 0 < t < T \tag{7}$$

$$z(1,t) = \chi(t) \quad , \quad s(t) < x < 1, 0 < t < T \tag{8}$$

$$k_1 \frac{\partial u}{\partial x}(0,t) = j(t) \quad , \quad 0 < t < T \tag{9}$$

$$u(s(t),t) = v(s(t),t) = 0, \qquad 0 < t < T \tag{10}$$

$$k_1 \frac{\partial u}{\partial x}(s(t),t) - k_2 \frac{\partial v}{\partial x}(s(t),t) = \nu z(s(\tau),\tau) \frac{ds}{dt}(t), \qquad 0 < t < T \tag{11}$$

$$\frac{\partial z}{\partial x}(s(t),t) + \frac{a_2}{a_m} \frac{\partial v}{\partial x}(s(t),t) = 0, \qquad 0 < t < T \tag{12}$$

donde $k_i, i = 1,2$ son las conductividades térmicas, $\nu = \frac{\delta \rho_2 r a_m}{a_m - a_2}$, $\rho_i, i = 1,2$ son las densidades de masa del cuerpo poroso , y r es el calor latente de congelamiento.

1.2. RESULTADOS TEÓRICOS

En [7] se reformuló el problema anterior en forma integral. Llamando $X(t) = u_x(s(t),t), Y(t) = v_x(s(t),t), S(t) = z(s(t),t)$ se obtuvo el siguiente sistema de ecuaciones integrales de Volterra:

$$\begin{aligned} X(t) &= 2\int_0^b G_1(s(t),t;\xi,0)\,\theta'(\xi)d\xi + 2\int_0^t a_1 N_{1x}(s(t),t;s(\tau),\tau)X(\tau)d\tau \\ &\quad -2\int_0^t \frac{a_1}{k_1} N_{1x}(s(t),t;0,\tau)j(\tau)d\tau \end{aligned} \tag{13}$$

$$\begin{aligned} Y(t) &= 2\int_b^1 N_2(s(t),t;\xi,0)\,\varphi'(\xi)d\xi + 2\int_0^t N_2(s(t),t;1,\tau)\,h'(\tau)d\tau \\ &\quad -2\int_0^t a_2 G_{2x}(s(t),t;s(\tau),\tau)Y(\tau)d\tau \end{aligned} \tag{14}$$

$$\begin{aligned} S(t) &= 2\int_b^1 G_m(s(t),t;\xi,0)\eta(\xi)d\xi + 2\int_0^t a_m N_{mx}(s(t),t;1,\tau)\chi(\tau)d\tau \\ &\quad +2\int_0^t \left(\frac{k_2}{\nu} - a_2\right) G_m(s(t),t;s(\tau),\tau)Y(\tau)d\tau \\ &\quad -2\int_0^t a_m N_{mx}(s(t),t;1,\tau)S(\tau)d\tau \\ &\quad -2\int_0^t \frac{k_1}{\nu} G_m(s(t),t;s(\tau),\tau)X(\tau)d\tau \end{aligned} \tag{15}$$

donde $G_i = G_i(x,t;\xi,\tau), N_1 = N_1(x,t;\xi,\tau), i = 1,2$ son las funciones de Green y Neumann.

Usando el Teorema de Punto Fijo de Banach, bajo adecuadas hipótesis sobre las funciones datos iniciales y de contorno, se probó en [7] que el sistema de ecuaciones integrales de Volterra tiene única solución local en el tiempo.

Además se obtuvo que la posición de la frontera libre viene dada por:

$$s(t) = b + \int_0^t \frac{k_1 X(\tau) - k_2 Y(\tau)}{\nu S(\tau)} d\tau \tag{16}$$

2. MÉTODOS NUMÉRICOS

A fin de obtener valores aproximados de las temperaturas, la humedad y la posición de la frontera libre, se aplican diferentes métodos numéricos y se comparan los resultados obtenidos [8].

2.1. MÉTODO DE INMOVILIZACIÓN DEL DOMINIO

Se realiza la transformación de coordenadas [9], [10]:

$$\zeta = \frac{s(t) - x}{s(t)} \quad 0 < x < s\left(t\right), \quad 0 < t < T$$

$$\xi = \frac{x - s(t)}{1 - s(t)} \quad s\left(t\right) < x < 1, \quad 0 < t < T$$

a fin de inmovilizar el dominio. El problema queda planteado en términos de las nuevas funciones incógnita

$$U(\zeta, t), \qquad 0 < \zeta < 1 \quad 0 < t < T$$

$$V(\xi, t), \qquad Z(\xi, t) \qquad 0 < \xi < 1 \quad 0 < t < T$$

siendo las ecuaciones (1)-(3) reemplazadas por

$$\frac{\partial U}{\partial t} s^2 + \frac{\partial U}{\partial \zeta}(1 - \zeta) s \frac{ds}{dt} = a_1 \frac{\partial^2 U}{\partial \zeta^2}, \qquad 0 < \zeta < 1, \qquad 0 < t < T \tag{17}$$

$$\frac{\partial V}{\partial t}(1 - s)^2 + \frac{\partial V}{\partial \xi}(\xi - 1)(1 - s) \frac{ds}{dt} = a_2 \frac{\partial^2 V}{\partial \xi^2}, \qquad 0 < \xi < 1, \qquad 0 < t < T \tag{18}$$

$$\frac{\partial Z}{\partial t}(1 - s)^2 + \frac{\partial Z}{\partial \xi}(\xi - 1)(1 - s) \frac{ds}{dt} = a_m \frac{\partial^2 Z}{\partial \xi^2}, \qquad 0 < \xi < 1, \qquad 0 < t < T \tag{19}$$

con las correspondientes condiciones iniciales y de contorno. Las condiciones (11) y (12) sobre la frontera libre, se transforman en

$$\frac{k_1}{s} \frac{\partial U}{\partial \zeta} - \frac{k_2}{1 - s} \frac{\partial V}{\partial \xi} = \nu Z \frac{ds}{dt} \qquad \zeta = \xi = 0 \quad 0 < t < T \tag{20}$$

$$\frac{\partial Z}{\partial \xi} + \frac{a_2}{a_m} \frac{\partial V}{\partial \xi} = 0, \qquad \xi = 0 \qquad 0 < t < T \tag{21}$$

Aplicando esquemas de diferencias finitas usuales en el espacio y explícitos en el tiempo se calculan los valores aproximados de la solución.

2.2. MÉTODO CUASI-PUNTO FIJO

Se basa en la discretización del sistema de ecuaciones integrales de Volterra (13)-(15). considerando pasos espacial y temporal fijo.

2.3. MÉTODO CUASI-ESTACIONARIO

Para el caso en que el número de Stefan es suficientemente chico, es decir, cuando al calor latente del material es grande con respecto al calor específico, el método cuasi-estacionario brinda una buena aproximación de la solución [11]. Se proponen las siguientes expresiones para las distribuciones de temperatura y humedad:

$$U(x, t) = A(t) + B(t)x, \quad 0 < x < s\left(t\right), t > 0$$

$$V(x, t) = D(t) + E(t)x, \quad s(t) < x < 1, t > 0$$

$$Z(x,t) = F(t) + G(t)x, \quad s(t) < x < 1, t > 0.$$

Reemplazando estas expresiones en el problema P se tiene que las nuevas funciones incógnita $A(t), B(t)$, $D(t), E(t), F(t)$ y $G(t)$, pueden obtenerse a partir de los datos de borde e iniciales y de la función $s(t)$. Esta última función es la solución del siguiente problema de valores iniciales:

$$s'(t) = \Big[j(t) - k_2 \frac{h(t)}{1-s(t)}\Big] \frac{1}{\nu[\chi(t) + \frac{a_2}{a_m} h(t)]}, \quad t > 0,$$
$$s(1) = b$$

3. CONCLUSIONES

Los resultados teóricos exhibidos en [7] demuestran la existencia y unicidad de solución del problema, para tiempos en un intervalo restringido de tiempo, cuya amplitud depende de los valores de los datos. En el presente trabajo, a través de la aplicación de los distintos esquemas numéricos fue posible determinar rangos más amplios de posibles valores para los coeficientes físicos del problema y, consecuentemente, ampliar el intervalo temporal.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Proyecto "Tratamiento numérico de Problemas de Frontera Libre" (ING305-UNR).

REFERENCIAS

[1] M. FARID, *The moving boundary problem from melting and freezing to drying and frying of food*, Chem.Eng.Process, 41 (2002), 1-10.

[2] L.A. CAMPAÑONE, L.A. ROCHE, V.O. SALVADORI AND R.H. MASCHERONI *Structural studies on unpackaged food during their freezing and storage*, J.Food Sci. 71 (2006) E218-E226.

[3] A.V.LUIKOV,*Heat and Mass Transfer in Capillary-porous bodies*, Pergamon Press, Oxford, 1966.

[4] A.V.LUIKOV,*Heat and Mass Transfer*, MIR Publishers, Moscow, 1978.

[5] M.D. MIKHAILOV, *Exact solution of temperature and moisture distribution in a porous half-space with moving evaporation front*, Int.J.Heat Mass Transfer. 18(1975)797-804.

[6] M.D. MIKHAILOV, *Exact solution for freezing of humid porous half space*, Int.J.Heat Mass Transfer. 19(1976) 651-655.

[7] E.A. SANTILLAN MARCUS, AND A.C. BRIOZZO, *On freezing of a finite humid porous medium with a heat flux condition*, Non Linear Analysis 67 (2007), 1919-1937.

[8] , J.CRANK, *Free and moving boundary problems*, Clarendon Press, Oxford, 1984.

[9] S.L.MITCHELL, AND M.VYNNYCKY, *Finite-difference methods with increased accuracy and correct initialization for one-dimensional Stefan problems* App.Math.Comp. 215(2009) 1609-1621.

[10] J. CALDWELL, AND S. SAVOVIC, *Numerical solution of Stefan problem by variable space grid and boundary inmobilization method*, J.Math.Sci. 13(2002),67-79.

[11] V.J. LUNARDINI (ED.),*Heat Transfer with Freezing and Thawing*, Elsevier, London, 1991.

DESARROLLO DE UNA HERRAMIENTA COMPUTACIONAL PARA EL DISEÑO AERODINÁMICO DE PALAS DE AEROGENERADORES DE EJE HORIZONTAL

Uribe Gustavo†

†Estudiante de Ingeniería Mecánica, Universidad Nacional del Comahue, Buenos Aires 1400, 8400 Neuquén, Argentina, gustfuribe@hotmail.com, www.uncoma.edu.ar

Resumen: En este trabajo se presenta un modelo matemático para analizar el comportamiento aerodinámico de turbinas eólicas basado en la teoría BEM (*Blade-Element Momentum*) y su implementación en un programa computacional que consta de dos módulos. El primero devuelve la geometría óptima de la turbina según un criterio de máxima eficiencia aerodinámica. El segundo módulo requiere como parámetros de entrada la geometría de la turbina y devuelve las curvas características de performance, para un amplio rango de velocidades de viento U y velocidades de rotación Ω. Estas curvas concretamente son los clásicos coeficientes de potencia C_P, de torque C_Q y de empuje C_T, en función de la relación de velocidades Ω·R/U. También proporciona curvas de potencia P, torque Q y empuje T versus velocidad del viento U. Se presentan resultados de algunos ejemplos propuestos y se obtienen conclusiones de ellos.

Palabras claves: *energía eólica, aerogenerador, teoría de cantidad de movimiento del elemento de pala (BEM Theory)*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

La capacidad de una turbina eólica para convertir la energía del viento en energía útil se define por el coeficiente de potencia C_P, que es la razón entre la potencia útil obtenida en el eje de la turbina y la potencia disponible del viento, de velocidad U, que atraviesa un área igual al área barrida por las aspas.

Existen diferentes enfoques para el diseño aerodinámico de aerogeneradores: teoría BEM (*Blade Element Momentum*), métodos de estela de vórtices, CFD (*Computational Fluid Dynamic*), etc. En este trabajo la metodología que se aplica es la conocida como **teoría BEM** o *Teoría de la cantidad de movimiento del elemento de pala*.

Sus ecuaciones son implementadas en el desarrollo de un código computacional, denominado **ASPA** [5], el cual consta de dos subprogramas o módulos:

1. Módulo de diseño: devuelve la geometría óptima de la pala (distribución de cuerda y alabeo) que maximiza el coeficiente de potencia para una dada velocidad de diseño. El usuario debe especificar los coeficientes aerodinámicos del perfil a utilizar, el número de palas, la velocidad de diseño, etc.

2. Módulo de análisis de performance: a partir de una geometría de pala y los datos de sustentación y arrastre del perfil (o los perfiles) seleccionado, calcula los parámetros de performance de la turbina.

La fundamentación de utilizar esta teoría representa un compromiso entre precisión y costo computacional. Si bien es un método simple (si se lo compara con algunos complejos métodos de CFD), es rápido computacionalmente hablando, y la experiencia demuestra su excelente aptitud y practicidad en el diseño de palas de aerogeneradores.

2. MODELO MATEMÁTICO

Para la deducción del modelo matemático se utiliza el esquema representado en la Figura 1, en el cual se muestra un tubo de corriente con forma anular de espesor *dr*, entre las superficies *d* y *e*, en la parte superior, y sus homólogas en la parte inferior. El tubo comienza desde la sección transversal a_0

(considerada a una gran distancia del rotor aguas arriba), hasta la estela lejana en la sección a_1 (considerada a una gran distancia aguas abajo). Todas las secciones a_0, b, a, c y a_1 se consideran anulares. Se asume que existe simetría de revolución en el flujo.

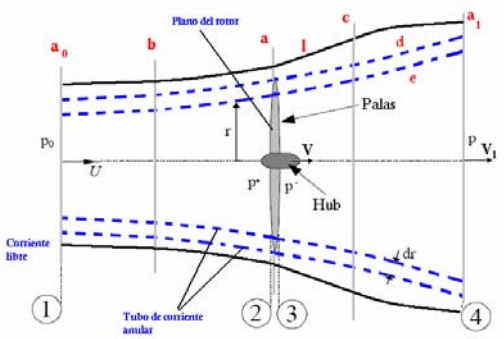


Figura 1: Geometría del modelo de un tubo anular de corriente a través de la hélice, considerando rotación de la estela

En cada sección las componentes de las velocidades son:

- **Corriente libre (sección a_0):** componente axial U.
- **Rotor (sección a):**
  - En el lado inmediatamente aguas arriba: componente axial V y componente radial v_r. No existe componente tangencial.
  - En el lado inmediatamente aguas abajo: componente axial V, componente radial v_r y componente tangencial u.
- **Estela lejana (sección a_1)**: componente axial V_1, componente radial v_{r1} y componente tangencial u_1.

2.1. HIPÓTESIS

A continuación se presentan las hipótesis que se asumen para la deducción del modelo:

- Se considera al viento libre como una corriente uniforme y estacionaria.
- Los efectos de la viscosidad sólo se toman en cuenta en el arrastre de los perfiles aerodinámicos (estela no viscosa).
- El aire se asume incompresible.
- Las aspas giran en un plano perpendicular al viento.
- Cualquier tubo de corriente a través del rotor, entre *r* y *r+dr*, es independiente de los tubos adyacentes.
- El número de palas se considera a través del factor de pérdidas por desprendimientos de vórtices.

2.2. ECUACIONES

A los fines de ahorrar espacio para la presentación de los algoritmos implementados, no se detallan aquí las ecuaciones utilizadas. Además, dichas ecuaciones pueden encontrarse desarrolladas en la bibliografía referente a la temática [1], [2], [3], [4]. Se desea enfatizar que las ecuaciones de la teoría BEM son producto de la combinación de ecuaciones provenientes de la teoría de cantidad de movimiento y de la teoría del elemento de pala (Figura 2).

Figura 2: Origen de la teoría BEM

3. ALGORITMOS IMPLEMENTADOS

Se presenta sólo el algoritmo implementado en el módulo de análisis de performance, ya que el mismo es una extensión del utilizado para el módulo de diseño. Se hace uso del método iterativo de cálculo propuesto por [6].

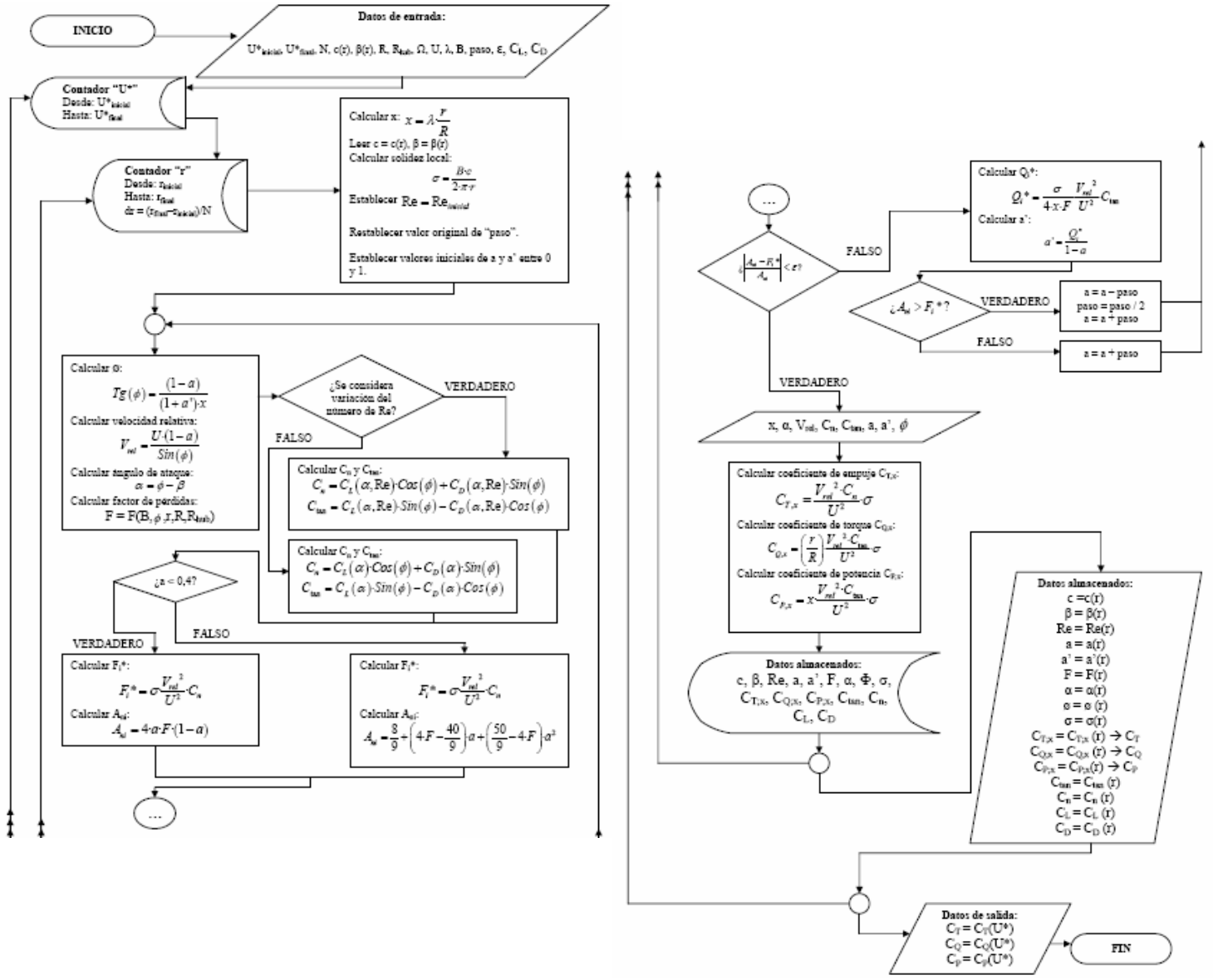
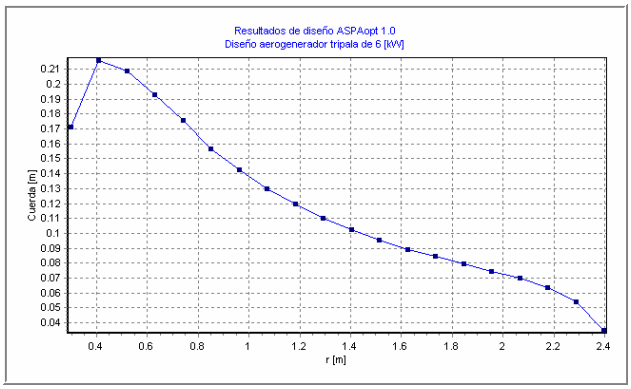


Figura 3: Diagrama de flujo del módulo de análisis de performance

4. RESULTADOS

A manera de ejemplificar el uso del código desarrollado, se propone efectuar el diseño de un aerogenerador tripala que sea capaz de entregar 6 kW a una velocidad de diseño U = 11 m/s utilizando el módulo de diseño. Posteriormente se lo analiza con el módulo de análisis de performance. Los resultados entregados por el programa se presentan en forma completa en el póster de la presentación.

4.1. RESULTADOS MÓDULO DE DISEÑO



r [m]	r/R_{pala}	Cuerda [m]	Alabeo [°]
0.3002	0.122531	0.17151	23.59
0.4104	0.16751	0.21588	18.6
0.5206	0.21249	0.20859	14.86
0.6309	0.25751	0.19305	11.74
0.7412	0.302531	0.17553	9.33
0.8514	0.34751	0.15671	7.65

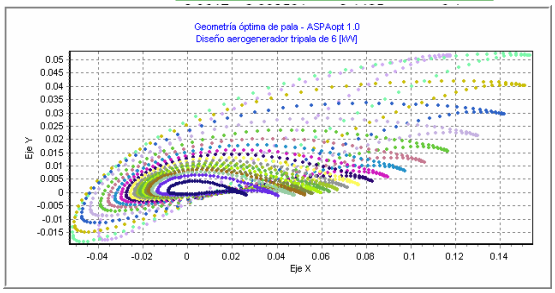


Figura 4: Algunos resultados del módulo de diseño

4.2. RESULTADOS MÓDULO DE ANÁLISIS DE PERFORMANCE

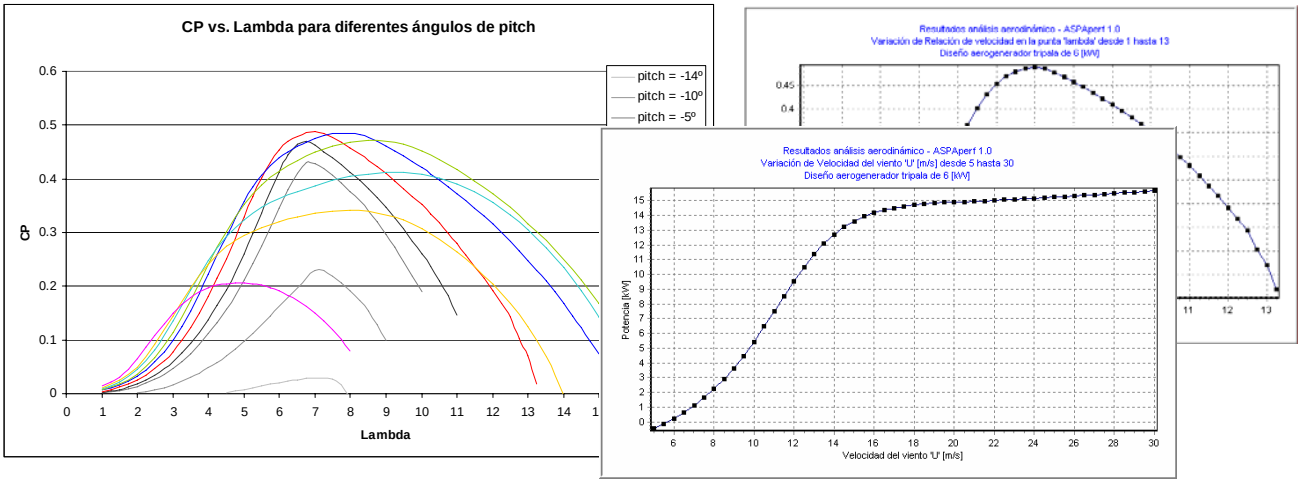


Figura 5: Algunos resultados del módulo de análisis de performance

AGRADECIMIENTOS

A mi tutor de este Trabajo Final de Carrera, el Ing. Pablo Álvarez. A la Universidad Nacional del Comahue. A Mauricio Schneebeli, de INVAP Ingeniería S.A. A mis seres queridos.

REFERENCIAS

[1] D. M. EGGLESTON, F. STODDART, *Wind turbine engineering design*, Van Nostrand Reinhold Company, 1987.

[2] H. GLAUERT, *Air plane propellers, aerodynamic theory*, Div. L, Capítulo XI, Springer Verlag, 1935.

[3] G. INGRAM, *Wind turbine blade analysis using the blade element momentum method*, School of Engineering, Durham University, 2005.

[4] D. LE GOURIERES, *Energía eólica. Teoría, concepción y cálculo práctico*, Masson S.A., 1983.

[5] G. URIBE, *Desarrollo de un software para diseño de palas de aerogeneradores de eje horizontal*, Proyecto Integrador Profesional, Universidad Nacional del Comahue, 2010.

[6] A. J. VITALE, A. P. ROSSI, *Método iterativo de cálculo para diseño y simulación de hélices de turbinas eólicas de pequeña escala*, Mecánica Computacional, Vol. XXVII, pp. 2457-2467, Asociación Argentina de Mecánica Computacional, 2008.

APROXIMACIÓN NUMÉRICA DE UN PROBLEMA DE FRONTERA LIBRE QUE DESCRIBE LA INTERFACE ENTRE DOS GRUPOS DE ANIMALES DE UNA MISMA ESPECIE.

Oscar A. Ramírez[♭] y Deccy Y. Trejos[†]

[♭] *Universidad Distrital Francisco José de Caldas, Proyecto Curricular de Matemáticas, Bogotá,Colombia, oscarexud@hotmail.com, www.udistrital.edu.co*

[†] *Universidad Distrital Francisco José de Caldas, Proyecto Curricular de Matemáticas, Bogotá,Colombia, dytrejosa@udistrital.edu.co, www.udistrital.edu.co*

Resumen: En este trabajo se presenta una aproximación numérica de un Problema de Frontera Libre (PFL) asociado con la evolución de la interface entre dos grupos de animales de la misma especie. Se tiene en cuenta la dinámica local del sistema, el esquema implícito de diferencias finitas es utilizado y algunas simulaciones numéricas se exhiben en diferentes escenarios.

Palabras clave: *Problema de Frontera Libre, Interface, Aproximación numérica, ecología.*
2000 AMS Subject Classification: 21A54 - 55P54

1. INTRODUCCIÓN

Un fenómeno interesante en la ecología de poblaciones es la aparición de una partición regional de múltiples especies. Como es el caso de dos grupos de individuos de una misma especie que están luchando en un punto (interface [1]) para obtener sus propios hábitats.

Para modelar el caso unidimensional de la anterior situación, se asume que los grupos con funciones de densidad poblacional u_1 y u_2 , se encuentran sometidos a la dispersión (k_1 y k_2 coeficientes de dispersión) y a un crecimiento logístico. Además están localizados en $0 < x < h(t)$ y $h(t) < x < l$ respectivamente, donde la función $h(t)$ que separa ambas regiones es la frontera libre del problema, las densidades en ese punto son iguales y es a priori totalmente desconocido.

Para avanzar desde $h(t)$ a $h(t + \Delta t)$ habrá fluido[2] a través de $x = h(t)$ una cierta cantidad de densidad poblacional, de modo que por la ley de conservación de la masa resulta que:

$$\int_t^{t+\Delta t} \varphi(h(t),\tau)d\tau = \begin{array}{c}\text{Cant. de densidad}\\ \text{poblacional } u_1 \text{ que}\\ \text{pasó por } x = h(t)\end{array} = \int_{h(t)}^{h(t+\Delta t)} u_1(x, t + \Delta t)dt,$$

luego, mediante el Teorema el Valor Medio se obtiene:

$$\Delta t\varphi(h(t),\hat{\tau}) = (h(t + \Delta t) - h(t))\, u_1(\hat{x}, t + \Delta t),$$

donde $h(t) \leq \hat{x} \leq h(t + \Delta t)$ y $t \leq \hat{\tau} \leq t + \Delta t$.

Ahora usando la Ley de Fick ,resulta que:

$$-k_1\,(\Delta t)\, u_{1x}(h(t),\hat{\tau}) = (h(t + \Delta t) - h(t))\, u(\hat{x}, t + \Delta t).$$

[1] Interface: frontera intermedia.

[2] La función $\varphi(x, t)$ representa el flujo de la población u_1 a través de la sección x y en el tiempo t.

De este modo, cuando $\Delta t \to 0$,

$$-k_1 u_{1x}(h(t),t) = h'(t)u_1(h(t),t).$$

De forma similar para u_2, se tiene

$$-k_2 u_{2x}(h(t),t) = h'(t)u_2(h(t),t).$$

Quedando formulado el siguiente Problema de Frontera Libre que describe la interface entre dos grupos de animales de una misma especie, asumiendo que no existe entrada ni salida de flujo por las fronteras fijas ($x = 0$ y $x = l$),

$$\begin{cases} u_{1t} - k_1 u_{1xx} = a_1 u_1(1 - u_1/K_1), & 0 < x < h(t) \\ u_{2t} - k_2 u_{2xx} = a_2 u_2(1 - u_2/K_2), & h(t) < x < l \\ u_1 = u_2 \quad u_1 h'(t) = -k_1 u_{1x} = -k_2 u_{2x}, & x = h(t) \\ u_{1x}(x,0) = u_{1x}(l,t) = 0, & t > 0 \\ u(x,0) = u_o(x) > 0, & 0 \le x \le l \\ h(0) = b & (0 < b < l) \end{cases} \tag{1}$$

Donde $u_0 \in C^2[0,b] \cap C^2[b,l]$, $ku_0' \in C^1[0,l]$ y $u_0'(0) = u_0'(l) = 0$. Kwang Ik Kim y Zhigi Lin mostraron en el articulo *"A free boundary problem for a parabolic system describing an ecological model"* que este modelo está bien planteado, demostrando la existencia y unicidad de la solución [1].

2. Aproximación Numérica

Sea

$$v(x,t) = \begin{cases} u_1(x,t) & \text{si} \quad 0 < x < h(t) \\ u_1(x,t) = u_2(x,t) & \text{si} \quad x = h(t) \\ u_2(x,t) & \text{si} \quad h(t) < x < l \end{cases}$$

y $v_0(x) = u_0(x)$ y $0 < h(t) < l$, luego el sistema (1) se puede reescribir como:

$$\begin{cases} v_t - k_1 v_{xx} = a_1 v_(1 - v/K_1), & 0 < x < h(t) \\ v_t - k_2 v_{xx} = a_2 v_(1 - v/K_2), & h(t) < x < l \\ v h'(t) = -k_1 v_x = -k_2 v_x & 0 < x = h(t) < l \\ v_x(x,0) = v_x(l,t) = 0, & t > 0 \\ v(x,0) = v_o(x) > 0, & 0 \le x \le l \\ h(0) = b & (0 < b < l) \end{cases} \tag{2}$$

donde, $v_0 \in C^2[0,b] \cap C^2[b,l]$, $kv_0' \in C^1[0,l]$ y $v_0'(0) = v_0'(l) = 0$.

Para discretizar (2) se usa el esquema de diferencias finitas implícito que es incondicionalmente estable. Se realiza una partición del rectángulo $R = \{(x,t) : 0 < x < l, 0 < t < T_f\}$ en una malla que consta de $2M + 1 \times N + 1$ nodos de la forma (x_i, t_j), donde

$$x_i = (i-1)h, \quad i = 0; 2M+1; \quad \text{con} \quad h = \frac{l}{2M}, \quad M \in \mathbb{N}.$$

$$t_j = (j-1)k, \quad i = 0, ..., N+1; \quad \text{con} \quad k = \frac{T}{N}, \quad N \in \mathbb{N}.$$

Si se designa $v_{(i)}^{(j)} \approx v(x_i, t_j)$ y se reliza las aproximaciones a las derivadas por series de Taylor se tiene de (2),

Frontera en el paso t_{j+1}.
Si $1 < \overline{h(j)} < 2M + 1$ entonces,

$$\overline{h_{(j+1)}} \approx \frac{k}{2hv_{(h(j))}^{(j)}} \left[-k_2 v_{(h(j)+1)}^{(j)} + (k_2 - k_1) v_{(h(j))}^{(j)} + k_1 v_{(h(j)-1)}^{(j)} \right] + \overline{h_{(j)}},$$

donde $x_{h(j)}$ es el nodo más cercano de la malla a $\overline{h_{(j)}}$, es decir, $x_{h(j)} \approx \overline{h_{(j)}}$ ($\overline{h_{(j)}} \approx h(t_j)$).

Modelo Discreto

$$\begin{cases} -h^2 v_{(2)}^{(j)} + \theta_1 v_{(2)}^{(j+1)} + \alpha_1 v_{(3)}^{(j+1)} + \omega_1 \left[v_{(2)}^{(j+1)} \right]^2 \approx 0 \\ -h^2 v_{(i)}^{(j)} + \theta_2 v_{(i)}^{(j+1)} + \alpha_1 v_{(i+1)}^{(j+1)} + \alpha_1 v_{(i-1)}^{(j+1)} + \omega_1 \left[v_{(i)}^{(j+1)} \right]^2 \approx 0 \qquad i = 3; h(j+1) - 2 \\ -h^2 v_{(h(j+1)-1)}^{(j)} + \theta_3 v_{(h(j+1)-1)}^{(j+1)} + \alpha_2 v_{(h(j+1)+1)}^{(j+1)} + \alpha_1 v_{(h(j+1)-2)}^{(j+1)} + \omega_1 \left[v_{(h(j+1)-1)}^{(j+1)} \right]^2 \approx 0 \\ -h^2 v_{(h(j+1)+1)}^{(j)} + \theta_4 v_{(h(j+1)+1)}^{(j+1)} + \alpha_2 v_{(h(j+1)-1)}^{(j+1)} + \alpha_3 v_{(h(j+1)+2)}^{(j+1)} + \omega_2 \left[v_{(h(j+1)+1)}^{(j+1)} \right]^2 \approx 0 \\ -h^2 v_{(i)}^{(j)} + \theta_5 v_{(i)}^{(j+1)} + \alpha_3 v_{(i+1)}^{(j+1)} + \alpha_3 v_{(i-1)}^{(j+1)} + \omega_2 \left[v_{(i)}^{(j+1)} \right]^2 \approx 0 \qquad i = h(j+1) + 2; 2M - 1 \\ -h^2 v_{(2M)}^{(j)} + \theta_6 v_{(2M)}^{(j+1)} + \alpha_3 v_{(2M-1)}^{(j+1)} + \omega_2 \left[v_{(2M)}^{(j+1)} \right]^2 \approx 0 \end{cases} \tag{3}$$

donde,

$$\begin{array}{llll} \theta_1 = h^2 + k_1 k - a_1 k h^2 & \theta_4 = \theta_5 - \frac{k_2^2 k}{k_1 + k_2} & \alpha_1 = -k_1 k & \gamma_1 = \frac{a_1 k h^2}{K_1} \\ \theta_2 = h^2 + 2k_1 k - a_1 k h^2 & \theta_5 = h^2 + 2k_2 k - a_2 k h^2 & \alpha_2 = -\frac{k_1 k_2 k}{k_1 + k_2} & \gamma_2 = \frac{a_2 k h^2}{K_2} \\ \theta_3 = \theta_2 - \frac{k_1^2 k}{k_1 + k_2} & \theta_6 = h^2 + k_2 k - a_2 k h^2 & \alpha_3 = -k_2 k. & \end{array}$$

Valor aproximado de v en la frontera libre $h(j+1)$ en el instante t_{j+1}. Si $h(j+1) = 2 : 2M$ entonces,

$$v_{(h(j+1))}^{(j+1)} \approx \frac{k_2 v_{(h(j+1)+1)}^{(j+1)} + k_1 v_{(h(j+1)-1)}^{(j+1)}}{k_2 + k_1}.$$

El modelo discreto (3) es un sistema no lineal de ecuaciones para cada paso de tiempo que se resuelve haciendo uso del metodo de Newton-Raphson [2].

3. Resultado de las Simulaciones Numéricas

La siguientes simulaciones numéricas se han realizado utlizando el software MATLAB, bajo los siguientes parámetros hipotéticos: $P_0 = 10sen(x)+2$, $l = \pi$, $T_f = 4$, $k_1 = k_2 = 0.5$, $a_2 = a_1 = 0.01$, $K_1 = 5 = 2K_2$, $M = 100$, $N = 40$, $maxit = 100$ y $eps = 0.01$, donde $maxit$ es el máximo de iteraciones para el método de Newton-Raphson y eps el criterio de tolerancia.

En las figuras (Fig 1.),(Fig 2.) y (Fig 3) pueden compararse los resultados obtenidos manteniendo fijos las capacidades de carga, las tasas de crecimiento natural, los coeficientes de dispersión y la condición inicial, pero variando la posición inicial de la frontera libre. Puede observarse como la posición de ésta influye significativamente en la distribución de las poblaciones, cabe resaltar que bajo estas hipótesis ningún grupo esta condenado a la extinción dado que la rapidez con que distribuyen uniformemente implica que la variación de tal frontera dismimuya.

Dinámica de la distribución de las poblaciones
para diferentes valores iniciales de la frontera libre.

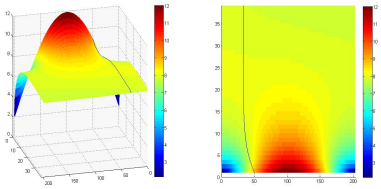


Fig 1. $h(0) = \pi/4$

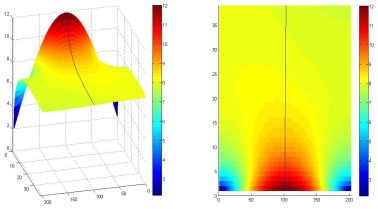


Fig 2.$h(0) = \pi/2$

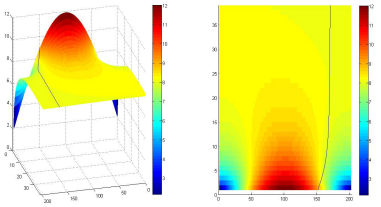


Fig 3. $h(0) = 3\pi/4$

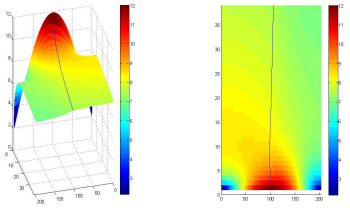


Fig 4. $h(0) = \pi/2$ y $K_1 = 5K_2$

4. Resultados y Conclusiones

Este trabajo se centró en la aproximación numérica de un problema de frontera libre que describe la interface entre dos grupos de animales de una misma especie, mediante un esquema de diferencias finitas implícito que es incondicionalmente estable, obteniendo así, un sistema algebraico no lineal de ecuaciones, que se resolvió mediante el método de Newton-Raphson.

Para el modelo considerado, se asumió que la interface estaba determinada por la interacción de los grupos y que cerca de ésta la densidad y el flujo eran continuos, condiciones similares a la frontera libre se pueden encontrar en modelos que describen el fenómeno de la segregación de hábitat en ecología de poblaciones [3].

Para trabajos futuros se intentará discretizar el problema por el método de Crank-Nicolson, manipular la frontera libre para extender el modelo 2D y 3D, incentivar un trabajo interdisciplinar conforme al tema y desarrollar métodos óptimos para aproximar la solucion de este tipo de sistemas.

Referencias

[1] K.Kim, Z. and Z. Lin, *A free boundary problem for a parabolic system describing an ecological model*, Linear Analysis: Real World Applications 10 (2009) $p428 : 436$.

[2] J. Mathews and K. Fink, *Métodos Numéricos con Matlab.*, 3 Edición.

[3] M. Mimura, Y. Yamada, and S. Yotsutani, *A free boundary problem in ecology*, Japan J. Appl. Math. 2 (1985) p,151 : 186.

MODELIZACIÓN MATEMÁTICA DEL PROCESO DE OBTENCIÓN DE BIOETANOL UTILIZANDO VARIABLES DE ESTADO.

Pablo Javiers†, Ornella Antonelli†, Pablo Mendez†, Guillermo Cocha‡

†Alumno de grado. Cátedra de Control Automático de Procesos. Ingeniería Química., U.T.N.F.R.L.P.
‡ Docente Cátedra de Control Automático de Procesos. Ingeniería Química, U.T.N.F.R.L.P.
gcocha@frlp.utn.edu.ar

Resumen: La forma tradicional en que se lleva a cabo el control de procesos químicos se basa todavía en la Transformada de Laplace. Esta metodología permite controlar de manera sencilla e intuitiva procesos con retardos pero su uso está limitado a los procesos lineales y lazos simples. En la actualidad, los requerimientos de calidad de producto y versatilidad en procesos batch marcan una tendencia hacia el uso de modelos en variables de estado que permiten el uso de herramientas matemáticas lineales y no lineales, además de multivariables. Además, en muchos procesos no están disponibles la totalidad de las variables de control por lo cual se deben desarrollar "sensores virtuales" que permiten estimar el valor de las variables que no son accesibles de manera física. Esto sólo es posible planteando el modelo en variables de estado. El presente trabajo muestra la metodología para modelizar un proceso de producción de etanol.

Palabras claves: *variables de estado, bioreactor, etanol.*

2000 AMS: 00A06 - 19 - Pósteres de estudiantes de grado.

1. INTRODUCCIÓN

El control de procesos químicos se basa fuertemente en modelos basados en la transformada de Laplace para sistemas de entrada salida de una variable. Como el método es apto para sistemas lineales, el proceso se opera en forma manual hasta el entorno de un punto de equilibrio, donde opera el proceso, y allí se conmuta al modo automático. La ventaja de este método es que permite un manejo intuitivo del proceso y es muy útil para operadores familiarizados con las características dinámicas del proceso.
En los últimos años, el aumento en los requerimientos de calidad y la versatilidad necesaria para que una misma planta maneje distintos productos, hace que el método de variables de estado sea más adecuado que el de función de transferencia. Para introducir el tema de variables de estado en problemas de control de procesos químicos se analiza el proceso utilizado comúnmente en el mundo para la obtención de etanol, a partir de caña de azúcar.
El corazón de dicho proceso es el bioreactor, que transforma los distintos azúcares de una corriente, en este caso Jugo de Caña, mediante la acción de un microorganismo, llamado Saccharomyces Cerevisiae, obteniendo como producto de esta reacción etanol y dióxido de carbono según la siguiente ecuación

$$C_6H_{12}O_6 + Microorganismo \rightarrow 2CH_3CH_2OH + 2CO_2.$$

Luego de la obtención se produce a la separación del dióxido y a la concentración de etanol que es de alrededor del 7% luego de la fermentación. El proceso se lleva a cabo en un reactor químico [2] como el que se muestra en la Figura 1.

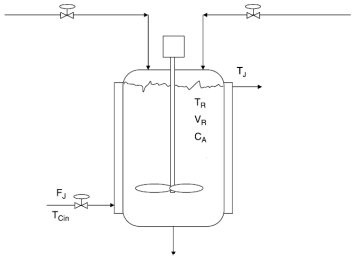


Fig. 1: Esquema básico de un reactor químico

Dada la naturaleza exotérmica de la reacción, se hace imprescindible un sistema de enfriamiento que se encargará de mantener la temperatura al nivel óptimo para la reproducción del microorganismo, situado entre 32 y 34°C. La carga de microorganismos intrínseca de la corriente de jugo de azúcar se anula mediante un proceso de calentamiento, de modo de que no haya ningún tipo de competencia por los azucares y sea el Saccheromyces quien se encargue de consumirlas. El objetivo del control es, entonces mantener esta temperatura ideal dentro del reactor para lo cual se opera una válvula que maneja el flujo de fluido en la camisa del reactor.

2. EL MODELO MATEMÁTICO DEL PROCESO

La forma más general de representación por variable de estado [1] de un sistema continuo está dada por la (Ec.1.a) que define la variación temporal de las variables de estado en función de estas mismas variables, las entradas $u(t)$; y la (Ec.1.b) que define la salida en función de las variables de estado, las entradas $u(t)$ y el tiempo. Así tenemos

$$\dot{x}(t) = f(x(t), u(t), t) \qquad \text{Ec.1.a}$$
$$y(t) = g(x(t), u(t), t) \qquad \text{Ec.1.b}$$

Aquí consideramos que *x*, *y* y *u* son vectores (columnas) de *n*, *p* y *m* componentes respectivamente. Esta forma de representación es válida para los sistemas continuos no-lineales y variantes en el tiempo en forma general.

Si el sistema es invariante en el tiempo, las funciones *f* y *g* dejan de depender explícitamente del tiempo y si el sistema es de coeficientes constantes, la ecuación de estado se expresa en la forma siguiente

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad \text{Ec.2.a}$$
$$y(t) = Cx(t) + Du(t) \qquad \text{Ec.2.b}$$

A es una matriz de *n* x *n*, *B* es una matriz de *n* x *m* (*n* filas x *m* columnas), *C* es una matriz de *p* x *n*, y *D* una matriz de *p* x *m*, que pueden ser dependientes del tiempo.

Si además de lineal, el sistema es invariante en el tiempo, las matrices A, B, C y D pasan a ser de coeficientes constantes.

Del balance de materia, concentraciones y energía [3] surgen las ecuaciones matemáticas que describen al modelo:

Balance de Masa:

$$dV = Fi - Fe$$

Balance por Componentes:

- Microorganismo

$$dCx = \frac{m_{iux} \cdot Cx \cdot Cs}{Ks + Cs} \cdot \exp(-kp \cdot Cp) - \left(\frac{Fe}{V}\right) \cdot Cx$$

- Etanol

$$dCp = \frac{m_{iup} \cdot Cx \cdot Cs}{Ks + Cs} \cdot \exp(-kp \cdot Cp) - \left(\frac{Fe}{V}\right) \cdot Cp$$

- Glucosa

$$dCps = -\frac{m_{iux} \cdot Cx \cdot Cs}{Ks + Cs} \cdot \frac{\exp(-kp \cdot Cp)}{Rsx} - \frac{m_{iup} \cdot Cx \cdot Cs}{Ks1 + Cs} \cdot \frac{\exp(-kp \cdot Cp)}{Rsp} + \frac{Fi}{V} \cdot C\sin - \left(\frac{Fe}{V}\right) \cdot Cp$$

- Oxigeno

$$dCo_2 = kla \cdot (Cst - Co_2) - r_{O2} - \frac{Fe}{V} \cdot Co_2$$

Balance de Energía en el Reactor

$$dT = \frac{1}{32 \cdot V \cdot r_{O2} \cdot \Delta H} - kT \cdot AT \cdot (T - Tag) + Fi \cdot r_0 \cdot ccal \cdot (Tin + 273) - \frac{Fe \cdot r_0 \cdot ccal \cdot (T + 273)}{r_0 \cdot ccal \cdot V}$$

Balance de Energía en la Chaqueta de enfriamiento

$$dTag = Fag \cdot ccalag \cdot roag \cdot (Tiag - Tag) + \frac{kT \cdot AT \cdot (T - Tag)}{Vm \cdot roag \cdot ccalag}$$

Cuando el sistema es no lineal como en este caso, se debe tomar un punto de operación para linealizar el proceso alrededor del mismo. Utilizando la serie de Taylor de una función y truncándola en un polinomio de orden 1, resulta

$$y \approx f(x_{10}, x_{20}, ..., x_{n0}) + \left[\frac{\partial f}{\partial x_1}\right] \cdot (x_1 - x_{10}) + \left[\frac{\partial f}{x_2}\right] \cdot (x_2 - x_{20}) + ... + \left[\frac{\partial f}{\partial x_n}\right] \cdot (x_n - x_{n0})$$

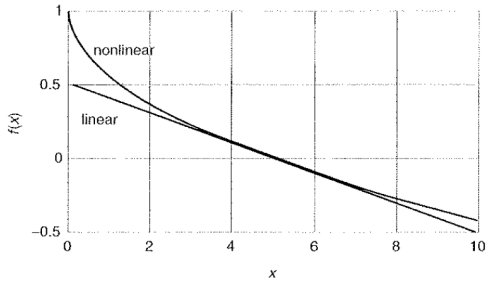


Fig. 2: Idea básica de la linealización

Del procedimiento de linealización y la Figura 2, pueden deducirse las consecuencias más importantes del uso de este proceso:

a. El valor de la constante (la derivada) depende del punto de funcionamiento elegido, y por lo tanto el modelo linealizado depende también de dicho punto.
b. La aproximación entre la curva original y la recta es tanto más exacta cuanto más cerca estemos del punto elegido, o lo que es lo mismo, la linealización es válida en un entorno del punto de funcionamiento.
c. Una vez linealizado el modelo, las variables originales se sustituyen por las variables incrementales respecto del punto de funcionamiento.

El modelo del bioreactor, luego de haber realizado el procedimiento de linealización nos queda como

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad [Ec.3.a]$$
$$y(t) = Cx(t) \qquad [Ec.3.b]$$

donde las matrices A, B, surgen como

$$A = \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial x} \end{bmatrix}, \quad B = \begin{bmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial u_m} \end{bmatrix}.$$

Una vez que tenemos el modelo matemático lineal, tenemos varias alternativas para plantear la estrategia de control.

Aplicando transformada de Laplace a la ecuación de estado, podemos obtener la siguiente relación $Y(s) = [C(sI - A)^{-1}B + D]U(s)$, donde *A, B, C y D* son las matrices vistas e *I* es la matriz identidad.

Obtenemos así la función matricial de transferencia del sistema al cual se lo puede controlar aplicando cualquiera de los métodos conocidos, por ejemplo el control PID.

En ecuaciones de estado, si se toma como señal de control a $u = -Kx$, la señal de control se determina por un estado instantáneo. A este tipo de control se lo denomina realimentación de estados y los valores del vector K se pueden encontrar usando la fórmula de Ackerman.

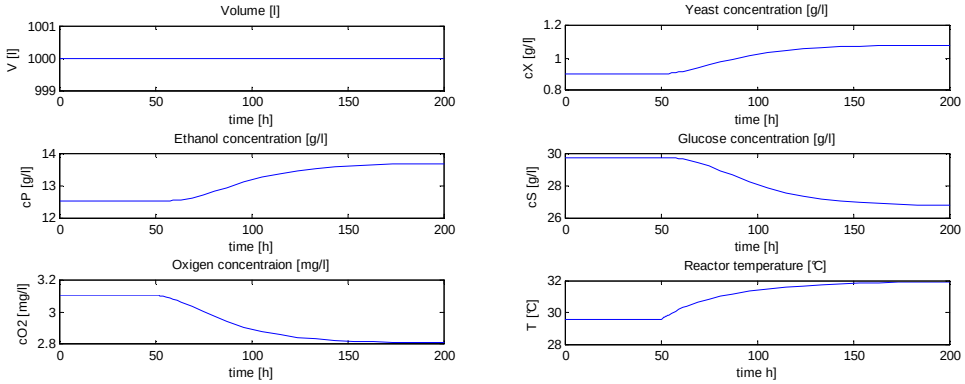


Figura 3: Comportamiento dinámico del proceso

3 CONCLUSIONES

El planteo del modelo matemático en variables de estado tiene la ventaja de poder elegir la estrategia de control de una manera más amplia que si utilizamos el modelo de función de transferencia, permite controlar sistemas multivariables, y estar preparado para utilizar de manera natural las herramientas del control avanzado.

El uso actual del software en clase permite, una vez obtenido el modelo en variables de estado pasar al modelo de función de transferencia de manera sencilla.

En la Figura 3 se muestran algunos resultados de las simulaciones efectuadas al aplicar el control al modelo aquí analizado.

4. REFERENCIAS

[1] K. OGATA, *Ingeniería de Control Moderna, Tercera Edición,* Prentice Hall, 2007.

[2] W. LUYBEN, *Chemical Reactor, Design and Control,* John Wiley & Sons, 2007.

[3] W. BEQUETTE, *Process Dynamics, Modeling, Analysis and Simulation,* Prentice Hall, 1998.

SIMULACIÓN NUMÉRICA – PERFIL NACA 2411 MODELOS DE TURBULENCIA

Adotti, Marcelo I.†

† Laboratorio de Aerodinámica. Facultad de Ingeniería, Universidad Nacional del Nordeste (UNNE), Avenida las Heras 727, (3500) Resistencia, Chaco, Argentina.

Resumen: En el presente trabajo se analiza la simulación de un perfil aerodinámico NACA 2411, empleando procedimientos de Dinámica de Fluidos Computacional o (CFD). Las ecuaciones, que describen el comportamiento del flujo, derivadas de la Mecánica de los Fluidos, son las de Continuidad, Cantidad de Movimiento (Navier – Stokes), Energía y la de los gases ideales. La velocidad de simulación es de 0,8 Mach, con aplicaciones de modelos de turbulencia dentro del dominio de RANS (Navier – Stokes promedio Reynolds). A fin de comparar resultados de los distintos modelos de turbulencia simulados, se analizan los coeficientes de sustentación y de arrastre del perfil

Palabras claves: *Simulación Numérica, Modelos de Turbulencia, Perfil aerodinámico, Coeficiente Sustentación, Coeficiente de Arrastre, Flujo Compresible.*

1 INTRODUCCIÓN

En este trabajo se analiza la simulación de un perfil aerodinámico NACA 2411, utilizando la Dinámica de Fluidos Computacional o (CFD). Las ecuaciones que describen el escurrimiento de los fluidos las proporciona la Mecánica de los Fluidos a partir de la ley de Continuidad, Cantidad de Movimiento (Navier – Stokes) y en el caso de flujos compresibles la de Energía, adicionando como ecuación complementaria la ecuación de los gases ideales. Describir analíticamente el escurrimiento de un fluido real mediante estas ecuaciones resulta complejo, desembocando en sistemas de ecuaciones diferenciales parciales con términos no lineales. Hay casos donde éstas no se pueden despreciar o anular, por ello se recurre a métodos numéricos para obtener soluciones aproximadas. [1]

Para la simulación se utilizará el software "Fluent". Éste utiliza el método volumen finito de control o celdas. Así, el dominio o volumen de control que rodea al perfil es reemplazado por un sistema de celdas Figuras (1 y 2), también conocido como discretización del dominio o mallado. El programa utilizado para generar la malla es el "Gambit". En cada celda se aplican las ecuaciones de Navier – Stokes (N-S) con valores promedios del Número Adimensional de Reynolds y simultáneamente las ecuaciones antes mencionadas, resultando en un sistema de ecuaciones.

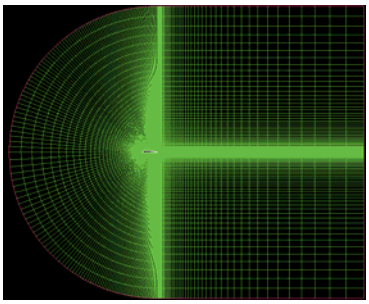


Figura 1: Malla del Perfil

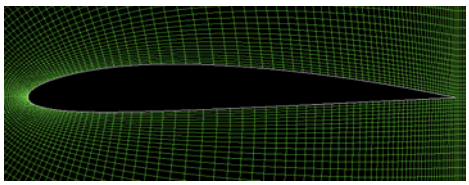


Figura 2: Mallado más denso cerca de la capa límite

2.1 FLUJO COMPRESIBLE Y MODELOS DE TURBULENCIA

La simulación del perfil alar se realizó con cuatro diferentes modelos de turbulencia. Estos modelos representan propiedades del flujo turbulento mediante ecuaciones de transporte, como la intensidad de

mezclado o difusión de los torbellinos turbulentos. La mayor parte de los modelos de turbulencia utilizados en la simulación se sostienen en la hipótesis de que la turbulencia es isotrópica, independientes del sistema de coordenadas, y poseen comportamiento estadístico estable y semejante. El flujo simulado es compresible, lo que requiere que además de aplicar las ecuaciones de continuidad, N-S y modelos de turbulencia, se deba incluir la ecuación de Energía, al considerar el intercambio de calor en la capa límite del perfil. Al simular a velocidades mayores a 0,3 de Mach, la densidad del aire varía sensiblemente con la temperatura, debiéndose agregar la ecuación de estado de los gases ideales, donde la temperatura pasa a ser una incógnita del campo de flujo. [2]

2.2 MODELO SPALART ALLMARAS

Este modelo de turbulencia consta de una ecuación para modelar y resolver la ecuación de transporte para la energía cinética viscosa de los torbellinos. Posee la característica de no tener que calcular una escala de longitud relacionada con los esfuerzos cortantes locales en relación al espesor de la capa límite. Demuestra buenos resultados para simular capas límites sometidas a gradientes de presión adversos. [3]

2.3 MODELO STANDARD K – EPSILON

Este modelo de turbulencia consta de dos ecuaciones de transporte que determinan por separado el campo de velocidades turbulentas y la escala turbulenta. La primera refiere al transporte de la energía cinética turbulenta (k) y la segunda incorpora la tasa de disipación de la turbulencia (ε). Es un modelo semi-empírico y es válido únicamente para flujos completamente turbulentos. [3]

2.4 MODELO STANDARD K – OMEGA

El modelo κ-ω que se utilizó en la simulación, posee la capacidad de reproducir de manera eficiente flujos entre paredes y flujos cortantes libres.

Es una ecuación semi-empírica, basada en la modelación de las ecuaciones de transporte para la energía cinética turbulenta (κ), y la tasa de disipación de la vorticidad (ω). [3]

2.5 MODELO TENSORES DE REYNOLDS (RSM)

Abandona la hipótesis de la viscosidad turbulenta isotrópica.

Integra las ecuaciones de Navier – Stokes, resolviendo los tensores de Reynolds en conjunto con una ecuación de tasa de disipación. Para el caso en 2D además de las ecuaciones de continuidad, N-S, ecuación de estado y energía, se necesitan además cinco ecuaciones de transporte para resolver el campo de flujo turbulento. Este modelo posee un gran potencial para describir flujos complejos. [3]

3.1 CONDICIONES DE CONTORNO DEL MALLADO

Volumen de Control: Pressure Far_Field (Presión Campo Lejano)

Paredes del Perfil: Wall. Shear Condition = No_ Slip

3.2 CONDICIONES DE BORDE SIMULACIÓN (BOUNDARY CONDITIONS)

Velocidad del Aire 0.8 Mach (Flujo Compresible). 0° Ángulo de Ataque; Temperatura (300°K); Aire (gas – ideal). Presión (atm) = 101325 Pa; Se considera la ecuación de Energía. Criterio de Convergencia de los residuos $1x10^{-06}$. Intensidad de turbulencia del 10% y un radio hidráulico igual a la cuerda del perfil, o un radio de viscosidad turbulenta de 10 en otros casos. El software resuelve las ecuaciones de N-S, las de Energía Mecánica y Térmica en conjunto con la de continuidad, basada en la Presión, formulación implícita, espacio 2D, tiempo estable.

4 ECUACIONES DE CANTIDAD DE MOVIMIENTO

Modelo Spalart Allmaras

El término *Gv* representa la producción de la viscosidad turbulenta; *Yv* destrucción o difusión de la viscosidad turbulenta. $\sigma v = 0,622$; $Cb2 = 2/3$, son constantes y ν es la viscosidad cinemática molecular.

$$\frac{\partial}{\partial t}(\rho\tilde{\nu}) + \frac{\partial}{\partial x_i}(\rho\tilde{\nu}u_i) = G_v + \frac{1}{\sigma_{\tilde{\nu}}}\left[\frac{\partial}{\partial x_j}\left\{(\mu + \rho\tilde{\nu})\frac{\partial\tilde{\nu}}{\partial x_j}\right\} + C_{b2}\rho\left(\frac{\partial\tilde{\nu}}{\partial x_j}\right)^2\right] - Y\nu$$

Modelo K - Epsilon

Gκ representa la generación o convección de la energía cinemática turbulenta. *Gb* es la generación de la energía cinética debido la flotación. *Ym* representa la contribución a la dilatación fluctuante en turbulencias compresibles. *C1ε =1,44*; *C2ε = 1,92*; *σε = 1,3*. Son las constantes de Prandtl.

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_i}(\rho k u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] + G_k + G_b - \rho\varepsilon - Y_M$$

$$\frac{\partial}{\partial t}(\rho\varepsilon) + \frac{\partial}{\partial x_i}(\rho\varepsilon u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial\varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{k}(G_k + C_{3\varepsilon}G_b) - C_{2\varepsilon}\rho\frac{\varepsilon^2}{k}$$

Modelo K – Omega

Gκ representa la generación o convección de la energía cinemática turbulenta. *Gω* es la generación de la vorticidad. *Γκ* y *Γω* representa la difusión de κ y ω. *Yκ* e *Yω* representan la disipación de la Turbulencia.

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_i}(\rho k u_i) = \frac{\partial}{\partial x_j}\left(\Gamma_k\frac{\partial k}{\partial x_j}\right) + G_k - Y_k$$

$$\frac{\partial}{\partial t}(\rho\omega) + \frac{\partial}{\partial x_i}(\rho\omega u_i) = \frac{\partial}{\partial x_j}\left(\Gamma_\omega\frac{\partial\omega}{\partial x_j}\right) + G_\omega - Y_\omega$$

Modelo Tensores de Reynolds

Analizando el lado izquierdo de la ecuación, el primer término representa la derivada local respecto del tiempo y el segundo la convección de las velocidades medias. Del lado derecho de la ecuación el primer término representa la difusión turbulenta, el segundo la difusión molecular, el tercero la producción de los esfuerzos que componen al tensor de Reynolds, el cuarto la producción de flotación, el quinto el tensor de presión, el sexto la disipación, y el último la producción debido al sistema de rotación.

$$\frac{\partial}{\partial t}\left(\rho\overline{u_i'u_j'}\right) + \frac{\partial}{\partial x_k}\left(\rho u_k\overline{u_i'u_j'}\right) = -\frac{\partial}{\partial x_k}\left[\rho\overline{u_i'u_j'u_k'} + \overline{p\left(\delta_{kj}u_i' + \delta_{ik}u_j'\right)}\right] + \frac{\partial}{\partial x_k}\left[\mu\frac{\partial}{\partial x_k}\left(\overline{u_i'u_j'}\right)\right] +$$

$$\rho\left(\overline{u_i'u_k'}\frac{\partial u_j}{\partial x_k} + \overline{u_j'u_k'}\frac{\partial u_i}{\partial x_k}\right) - \rho\beta\left(g_i\overline{u_j'\theta} + g_j\overline{u_i'\theta}\right) + \overline{p\left(\frac{\partial u_i'}{\partial x_j} + \frac{\partial u_j'}{\partial x_i}\right)} - 2\mu\overline{\frac{\partial u_i'}{\partial x_k}\frac{\partial u_j'}{\partial x_k}} -$$

$$- 2\rho\Omega_k\left(\overline{u_j'u_m'}\varepsilon_{ikm} + \overline{u_i'u_m'}\varepsilon_{jkm}\right)$$

5 CONCLUSIONES

Los primeros tres modelos Spalart-Allamaras, K – Epsilon y K – Omega arrojaron resultados similares de los coeficientes de Arrastre (Cd) y de Sustentación (Cl), Figuras (3 y 4). En cambio con el modelo RSM, en igualdad de condiciones de borde, se obtuvieron valores inferiores del coeficiente de arrastre Cd y el coeficiente de sustentación Cl aumentó, produciendo un mayor rendimiento aerodinámico, expresado como el cociente entre el Cl y el Cd (Figura 5). La diferencia entre RSM y los primeros modelos reside que en el primero abandona la simplificación de la isotropía de los torbellinos, describiendo con mayor detalle el flujo compresible turbulento.

Al analizar la distribución de presiones en el campo de flujo para los modelos de turbulencia Spalart – Allmaras, K-Epsilon y K-Omega, se observan distribuciones de presiones estáticas superficiales similares, aplicada la sustentación en la parte superior del perfil, Figuras (6,7 y 8). En cambio, con el modelo RSM (Figura 9), la magnitud de la sustentación es mayor y corrida hacia el borde de salida del perfil. Esta descripción más detallada del contorno de presión sobre el perfil con el modelo RSM al simular el perfil a 0,8 Mach, se ajusta con lo antes descripto.

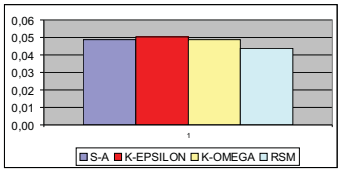


Figura 3: Coeficientes Arrastre (Cd)

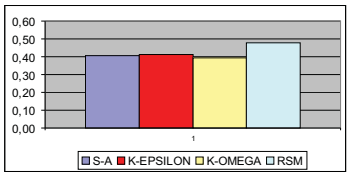


Figura 4: Coeficientes de Sustentación (Cl)

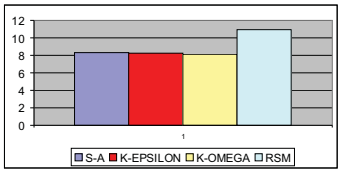


Figura 5: Rendimiento Aerodinámico

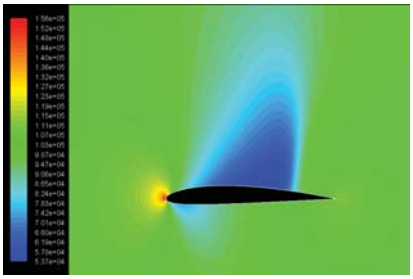


Figura 6: Contorno de Presión Estática S-A (Pa)

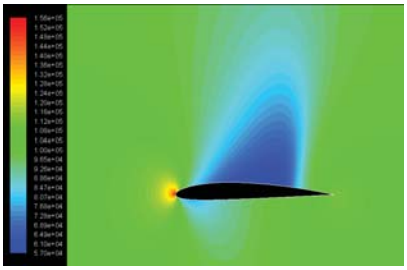


Figura 7: Contorno de Presión Estática k – e

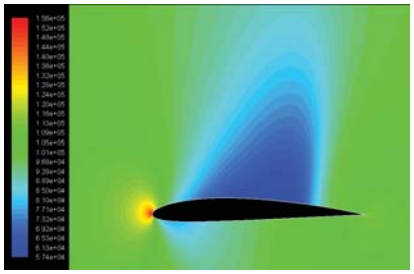


Figura 8: Contorno de Presión Estática k – ω

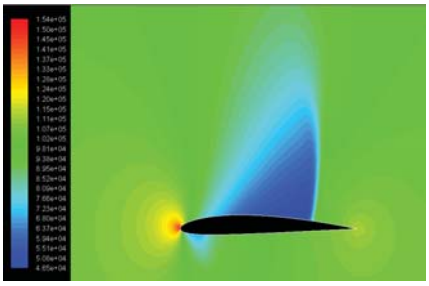


Figura 9: Contorno de Presión Estática RSM

AGRADECIMIENTOS

A los Ingenieros: Castro, Hugo G.; De Bortoli, Mario E.; Marighetti, Jorge O.

REFERENCIAS

[1] MERLE C. POTTER, DAVID C. WIGGERT, *Mecánica de Fluidos.* Tercera Edición (2002), pp.665-706

[2] YUNUS A. ÇENGEL, JOHN M. CIMBALA, *Mecánica de Fluidos. Fundamentos y Aplicaciones* (2006), pp. 840-843, 860-861.

[3] FLUENT USER'S GUIDE. *Chapter 12: Modeling of Turbulence* (2006).

ANALYSIS OF METHOD OF LINES FOR RESOLUTION OF CONVECTION DIFFUSION EQUATION

Marilaine Colnago, Prof.Dr. Messias Meneguette Junior, Prof. Dr. José Roberto Nogueira

Departament of Mathematics, Statistics and Computing, Universidade Estadual Paulista - UNESP, 19.060-900, Presidente Prudente, São Paulo, Brasil, marilainecolnago@yahoo.com.br, {messias,jrnog}@fct.unesp.br

Abstract: This paper is intended to obtain the numerical solution of Burgers equation, known as convection-diffusion equation, by the method of lines. This method is a way of approaching partial differential equations by ordinary differential equations. His main idea is to discretize spacial variables in order to obtain a system of ODE. In this paper it is used finite difference for discretization and the result compared with the analytic solution of the equation.

Key words: *Method of Lines, Bugers Equation*

1. INTRODUCTION

The Method of Lines is a numerical approximation of PDE over time. One approximates the spatial partial derivatives, which in this case was done using finite difference, then solves the resulting ODE system, here done with the help of Matlab. The aim of this study is to compare the numerical solution obtained by the method of lines with the analytic solution of Burgers equation, which is widely used in the teaching of fluid dynamics and engineering as a simplified model for the formation of shock waves and mass transport. It has been studied and applied for many decades. This equation is widely studied because it is nonlinear and yet has a known exact solution (analytical).

2. METHODOLOGY

As previously mentioned, for the Method of Lines, we have to approximate the differential equation, by using finite differences, whose approximation formula comes from the Taylor series. In Matlab there are several tools that can be used for the resultant ODE system. Consider then the Burgers equation, also known as advection-diffusion:

$$u_t + u.u_x - \mu.u_{xx} = 0$$

where $\mu = 0.003$ is called the coefficient of viscosity. We use this equation with initial and boundary conditions given by the exact solution of the equation :

$$u(x,t) = \frac{0.1e^{-A}+0.5e^{-B}+e^{-C}}{e^{-A}+e^{-B}+e^{-C}}$$

where

$$A = \frac{0.05}{\upsilon}(x - 0.5 + 4.95t)$$
$$B = \frac{0.25}{\upsilon}(x - 0.5 + 0.75t)$$
$$C = \frac{0.5}{\upsilon}(x - 0.375)$$

with $0 \le x \le 1$ and $0 \le t \le 1$ with an output interval of 0.1.

The solution of this equation was calculated in Matlab, using 201 points. The spatial derivatives are computed by the finite differences and the system of 201 ODEs is solved by the stiff integrator ode15s. An implementation for this equation can be found at [3].

3. RESULTS AND DISCUSSION

Below are the table with the computational errors and the figures generated by the output file .m.

Table 3.1 – Numerical Errors

Time	Numerical Solution	Analytical Solution	Error
0.0	1.000000	1.000000	0.0
0.0	1.000000	1.000000	0.0
0.0	1.000000	1.000000	0.0
0.0	1.000000	1.000000	0.0
0.0	0.999998	0.999998	0.0
⋮			⋮
0.0	0.100000	0.100000	0.0
0.0	0.100000	0.100000	0.0
0.0	0.100000	0.100000	0.0
0.0	0.100000	0.100000	0.0
0.0	0.100000	0.100000	0.0
0.2	1.000000	1.000000	0.000000
0.2	1.000000	1.000000	0.000000
0.2	1.000000	1.000000	0.000000
0.2	1.000000	1.000000	0.000000
0.2	1.000000	1.000000	0.000000
⋮			⋮
0.2	0.948162	0.944636	0.003526
0.2	0.743099	0.749992	-0.006893
0.2	0.555188	0.555314	-0.000126
0.2	0.508599	0.507371	0.001227
0.2	0.499956	0.499584	0.000372
⋮			⋮
0.2	0.100000	0.100000	0.000000
0.2	0.100000	0.100000	0.000000
0.2	0.100000	0.100000	0.000000
0.4	1.000022	1.000000	0.000022
0.4	1.000016	1.000000	0.000016
0.4	1.000011	1.000000	0.000011
0.4	1.000007	1.000000	0.000007
0.4	1.000004	1.000000	0.000004
⋮			⋮
0.4	0.100000	0.100000	0.000000
0.4	0.100000	0.100000	0.000000
0..4	0.100000	0.100000	0.000000
0.4	0.100000	0.100000	0.000000
0.4	0.10000	0.100000	0.000000
0.6	1.000098	1.000000	0.000098
0.6	1.000086	1.000000	0. 000086
0.6	1.000074	1.000000	0. 000074
0.6	1.000064	1.000000	0.000064
0.6	1.000054	1.000000	0.000054
⋮			⋮
0.6	0.100000	0.100000	0.000000
0.6	0.100000	0.100000	0. 000000
0.6	0.100000	0.100000	0.000000
0.6	0.100000	0.100000	0.000000

0.6	0.100000	0.100000	0.000000
⋮			⋮
1.0	1.000398	1.000000	0.000398
1.0	1.000374	1.000000	0. 000374
1.0	1.000350	1.000000	0. 000350
1.0	1.000327	1.000000	0. 000327
1.0	1.000305	1.000000	0. 000305
⋮			⋮
1.0	1.000076	1.000000	0. 000076
1.0	1.000065	1.000000	0.000065

1.0	1.000056	1.000000	0.000056
1.0	1.000047	1.000000	0.000047
1.0	1.000038	1.000000	0.00008
⋮			⋮
1.0	0.882859	0.856946	0.025914
1.0	0.193920	0.199719	-0.005799
1.0	0.103996	0.102646	0.001350
1.0	0.100164	0.100065	0.000099
1.0	0.100007	0.100002	0.000005

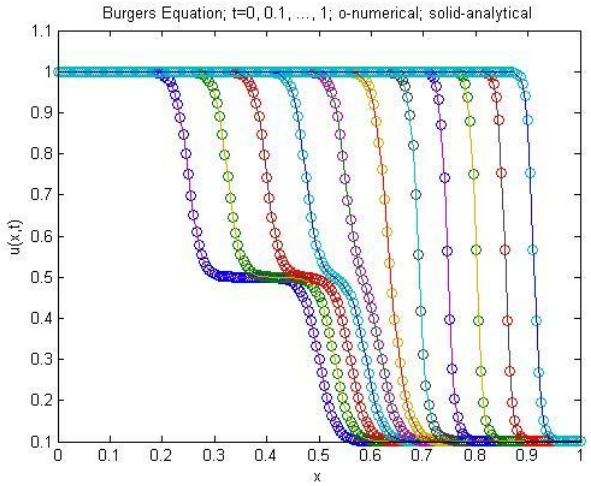


Figure 1 : Burgers Equation Levels

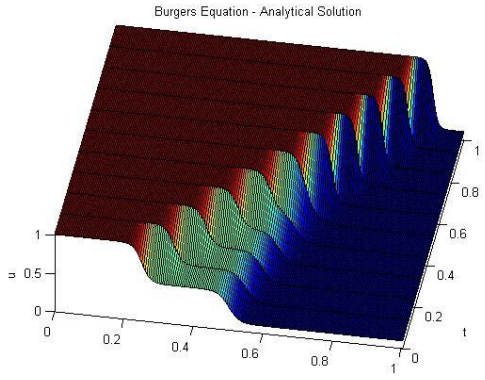


Figure 2 : Analytical Solution

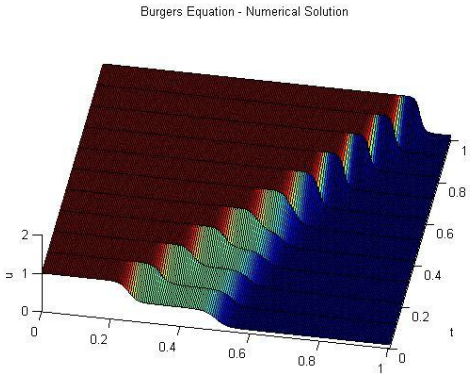


Figure 3 : Numerical Solution

In many physical processes, diffusive and convective phenomenon occurs simultaneously. Despite the presence of diffusion, they can present very high values for cases where convection dominates thus

forming shock, i.e., abrupt and discontinuous fronts. The Burgers' equation is the simplest example of this type.

This method works quite well, and for the spatial grid of 201 points, the agreement between the numerical and analytical solutions is quite satisfactory, the numerical errors did not grow or accumulate with increasing t. The classical methods are oscillatory and the method of lines does not oscillate much. The computational effort is small, and the difficulty of implementation too. In [3] and [4] can be found more details about this.

ACKNOWLEDGEMENTS

Thanks to the PROPG of UNESP for financial support.

REFERENCES

[1] R.J. LEVEQUE, *Finite difference for ordinary and partial differential equations: steady-state and time-dependent problems*, Seatle. Society for Industrial and Applied Mathematics, 2007.

[2] F. SHAMPINE, I. GLADWELL, S. THOMPSON, *Solving ODEs with Matlab*, Cambridge University, (2003), pp. 114-127.

[3] W. E. SHIESSER, G. W. GRIFFITHS, *A compendium of partial differential equation models: Method of Lines analysis with Matlab*, Cambridge University, (2009), pp. 90-114.

[4] A. V. WOUWER, PH. SAUCEZ, W. E. SHIESSER, *Adaptive Method of Lines*, Chapman & Hall/CRC. (2001), pp. 19-24.

DISEÑO ÓPTIMO DE SISTEMAS DE DESTILACIÓN REACTIVA COMO ÚNICO EQUIPO O COMO ETAPA DE "FINISHING"

Juan P. Archenti, M. Soledad Díaz y Patricia M. Hoch

Departamento de Ingeniería Química – Universidad Nacional del Sur - Planta Piloto de Ingeniería Química (PLAPIQUI – UNS – CONICET). Avda. Alem 1253, 8000 Bahía Blanca
{jarchenti,sdiaz,p.hoch}@plapiqui.edu.ar

Resumen: En este trabajo se presenta el diseño óptimo de un proceso de destilación reactiva para la producción de aditivos para combustibles, como reemplazo de los sistemas convencionales reactor/separador. El modelo de la columna reactiva está basado en el clásico modelo de Taylor y Krishna (2000), en el que se incluyeron también ecuaciones que permiten tener en cuenta las restricciones hidráulicas del sistema, de modo de proceder al adecuado dimensionamiento (Stichlmair y Fair, 1998). El problema se plantea en GAMS, y arroja como resultados valores de los números de etapas de reacción y separación requeridos, caudales de operación y relación de reflujo y boilup, así como las dimensiones de la columna para un número fijo de etapas y ubicación de las alimentaciones. Se proponen esquemas que contemplan columnas con diámetros no uniformes de acuerdo a los caudales internos del proceso. Se evalúan los costos de inversión y operativos del sistema reactor/separador y el sistema de destilación reactiva con y sin etapa de pre-reacción, y se comparan para la misma producción y pureza de producto requerida.

Palabras claves: *Destilación reactiva, optimización*
2000 AMS Subjects Classification: 21A54 - 55P5T4

1. INTRODUCCIÓN

La destilación reactiva ha recibido creciente interés durante los últimos años, debido a que se trata de un claro ejemplo de intensificación de procesos. La combinación entre dos procesos que tradicionalmente se llevaban a cabo en unidades separadas trae varias ventajas en lo que respecta al proceso. Por un lado, se disminuye el espacio necesario para la ubicación de los equipos y el consumo energético, y por otro se potencian las ventajas de las dos operaciones unitarias. Al efectuarse la continua separación de los productos de la mezcla reactiva, el equilibrio de la reacción se desplaza hacia el lado de los productos, mientras que muchas veces los azeótropos que los productos forman con los reactivos desaparecen debido a la completa conversión de los mismos.

La síntesis de MTBE a partir de Isobutileno y Metanol es una reacción exotérmica reversible. Existen distintos procesos comerciales disponibles para la producción del MTBE. Los mismos serán agrupados en procesos convencionales y no convencionales. Entre los convencionales se destacan los procesos de Phillips y Hüls (Meyers, 1986) y entre los no convencionales los de UOP y ABB LUMMUS (Fahad y El-Harthi, 2008). Los procesos convencionales de producción de MTBE se dividen en distintas secciones (Meyers, 1986; Fahad y El-Harthi, 2008):

- Síntesis de MTBE (2 etapas de reacción)
- Purificación de MTBE
- Recuperación de Metanol

Los procesos no convencionales combinan la segunda etapa de reacción con la etapa de purificación de MTBE en una unidad de Destilación Reactiva.

2. MODELO UTILIZADO Y ESCENARIOS

Para la primera etapa de diseño de la unidad reactiva se utiliza un "modelo de equilibrio" (Taylor y Krishna, 2000). El mismo supone que el líquido y vapor que abandonan una etapa se encuentran en equilibrio de fases en contraposición con el "modelo de no equilibrio" (Taylor y Krishna, 2000). Las bases para el diseño utilizadas en la primera etapa de diseño se toman del trabajo de Almeida-Rivera (2005). El mismo consiste en utilizar principios fundamentales como balances de masa y energía y relaciones de equilibrio para modelar las distintas etapas.

El trabajo de modelado se lleva a cabo en GAMS y para la optimización del problema no lineal se utiliza el solver CONOPT.

En primera instancia, se pretende reemplazar las secciones de síntesis (ambos reactores) y purificación de MTBE por una única etapa de destilación reactiva.

La columna de destilación reactiva a utilizar contará con 3 platos en la sección de rectificación, 10 platos en la sección reactiva y 6 platos en la sección de despojo.

Se procederá a especificar el caudal y composición de la alimentación de hidrocarburos, la cual se toma de Meyers (1986), así como también el plato en que debe alimentarse. La columna posee un segundo tipo alimentación que es el metanol, reactivo que es alimentado en exceso a las secciones de síntesis de las plantas convencionales. Se trata a la cantidad de metanol a alimentar así como también a la localización de la alimentación como variables de decisión del optimizador, es decir, serán calculadas por el solver utilizado para la optimización. Cabe destacar que, en este contexto, puede existir más de una alimentación de metanol. Las temperaturas de las alimentaciones se fijan en los valores mostrados en la figura 1 mientras que la presión de las mismas será consecuencia del perfil de presión resultante para la columna. Se impone una restricción de 12 Bar(a) como presión máxima operativa de la columna. Se simularán 2 escenarios distintos con la configuración hasta aquí descripta:

- Escenario 1: Se toma como objetivo maximizar la pureza del MTBE en el producto de fondo.
- Escenario 2: Se toma como objetivo maximizar la conversión de MTBE en la unidad de RD.

Tal como se expone en los resultados, los escenarios 1 y 2 no resultan, en primera instancia, completamente satisfactorios. Así, se propone generar 2 nuevos escenarios (3 y 4) en los que se adiciona una etapa de reacción en un reactor de lecho fijo y adiabático antes de procesar en la unidad de destilación reactiva.

Se calcula cuál sería la conversión de equilibrio para ese reactor en las condiciones mostradas en la figura 2 y se toma un 80% de la misma como conversión operativa. Se toma como temperatura de alimentación al reactor el valor de 313 K, para asegurar que es al menos 10 ºC superior a la temperatura del agua de enfriamiento, que se supone 30ºC. Se alimenta la salida del reactor a la columna de destilación reactiva y se permite al solver nuevamente decidir cantidad y localización de metanol a alimentar. Los nuevos escenarios plantean realizar parte de la conversión del isobutileno en un primer reactor y luego terminar de convertir y separar en la columna reactiva con los siguientes objetivos de optimización para el diseño:

- Escenario 3: Se maximiza la pureza del MTBE en el producto de fondo.
- Escenario 4: Se maximiza la conversión global de Isobutileno.

Para cualquiera de los 4 escenarios se impone una restricción a la pureza del MTBE en el producto de fondo de la columna reactiva de 95% de MTBE en fracción en masa, que es el mínimo necesario para utilizarlo como aditivo de combustibles.

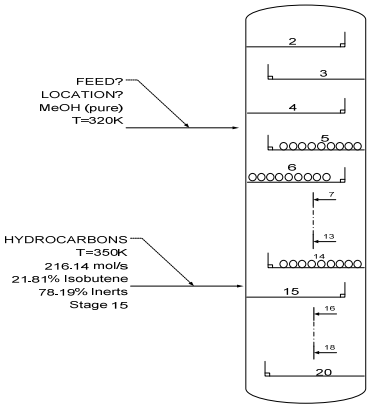


Figura 1: Escenarios 1 y 2

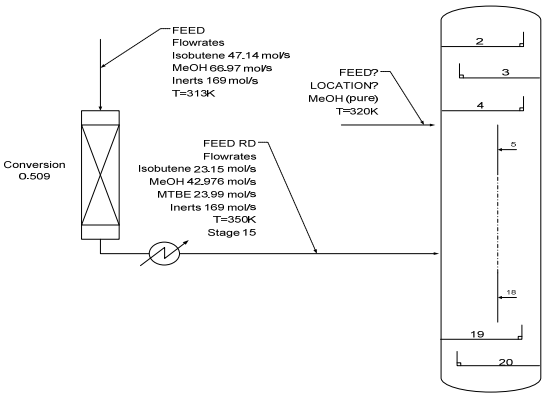


Figura 2: Escenarios 3 y 4

3. RESULTADOS

En la tabla 1 se presentan los resultados obtenidos para las variables de proceso de los 4 escenarios. En la tabla 2 se exponen los detalles de las variables constructivas de la columna correspondiente a cada uno de los escenarios y las estimaciones de costos para cada una de las 3 torres que han sido costeadas de acuerdo a Instituto Francés del Petróleo (1976). Las estimaciones de costo no incluyen rebullidor, condensador o bomba de reflujo. Resulta necesario utilizar un espaciado entre platos mayor en los platos de fondo a fin de poder cumplir con la restricción impuesta por la Ec. (17) de máxima carga líquida tal como se observa en la tabla 2. En la tabla 3 se dan a conocer las variables constructivas que son comunes a todas las columnas de los distintos escenarios.

El "Metanol a recuperar" que aparece en la tabla 1 representa el caudal molar de la especie que es obtenido en el destilado y que será necesario recuperar en la sección de recuperación de Metanol.

El análisis de los datos muestra que el escenario 1 permite alcanzar un muy alto grado de pureza de MTBE pero con una conversión muy lejana al resto de las alternativas o a la reportada por los procesos convencionales (Meyers, 1986) del orden del 99%. Además, el grado de pureza se logra a expensas de una importante relación de reflujo, lo que hace necesario utilizar un valor muy elevado para los diámetros de las secciones. El método de costeo utilizado (Instituto Francés del Petróleo, 1976) indica que la metodología no es aplicable a columnas con más de 5 m de diámetro dado que en esos casos es necesario realizar el montaje en campo lo que puede hacer que los costos reales dupliquen a los estimados. Por este motivo no se provee una estimación de costos para este escenario y se descarta esta opción ya que carece de competitividad frente al resto.

El escenario 2, en cambio, aparenta ser bastante prometedor ya que permite alcanzar la especificación mínima de pureza de MTBE convirtiendo simultáneamente un 98,2%. Este valor, sin embargo, es ligeramente inferior al de los procesos convencionales y, es por esto, que se decide estudiar si la inclusión de una etapa de reacción antes de la destilación reactiva aporta cuantiosas mejorías, es decir, proceder a simular los escenarios 3 y 4.

El escenario 3 permite alcanzar un alto grado de pureza del producto y una conversión del reactivo factible. La relación de reflujo requerida para purificar el MTBE ha disminuido marcadamente respecto del escenario 1. En comparación con el escenario 2 requiere de un mayor consumo energético para operar.

Por último, el escenario 4 permite maximizar la conversión global del proceso adoptando un valor cercano al 99% y alcanzando la mínima pureza requerida al mismo tiempo. Utiliza la menor de las relaciones de reflujo de los 4 escenarios y, por lo tanto, conlleva asociado el menor consumo energético global.

Tabla 1. Variables de Proceso

Escenario Nº	Pureza MTBE % masa	Conv Isobut por paso RD %	Conv Isobut global %	Relac de Reflujo
1	99.5	81.6	81.6	6
2	95	98.2	98.2	2.908
3	99.1	90.3	95.25	3.768
4	95	97.8	98.94	2.415

Duty Cond MW	Duty Reb MW	Pmax operación Bara	Metanol Total Utilizado mol/s	Metanol Alimentado Columna mol/s	Metanol a Recuperar mol/s	Frac Mol Metanol en destil
13.5	8.99	9.24	60.31	60.313	21.71	0.109
6.87	2.24	11.63	71.05	71.05	24.63	0.129
8.81	5.37	9.69	69.21	2.24	21.96	0.114
5.98	2.71	11.63	70.95	3.98	24.33	0.128

Cabe destacar que los valores de diámetros obtenidos en todos los escenarios exceden a los comercialmente usuales. Esto se debe a que en una columna de destilación tradicional los aportes de vapor

a la misma provienen del rebullidor o una alimentación que en esa fase se encuentre. Si todas las alimentaciones son líquidas, el mayor de los caudales de vapor es el del plato inferior que recibe el vapor generado en el rebullidor. En una columna reactiva, en cambio, puede ocurrir que la reacción sea exotérmica (como en el caso de estudio) haciendo que el calor de reacción se consuma en vaporizar parte del líquido que llega al plato. Esto genera un aporte extra de vapor y es por eso, sumado a la necesidad de utilizar relaciones de reflujo moderadas o altas, que se requiere de diámetros menos convencionales. El incremento de los caudales de vapor respecto de los de fondo al ingresar en la zona reactiva es notable. El caudal de vapor de la etapa 15 es de 188 mol/s y el de la etapa 14 (primera etapa reactiva) es de 513 mol/s. Aún cuando 216,14 mol/s de estos 513 mol/s provienen de la alimentación de hidrocarburos, el incremento es marcado. La tendencia se mantiene en la sección reactiva.

Tabla 2. Variables Constructivas de la Unidad de Destilación Reactiva

Escenario Nº	Diámetro etapas 2-14 m	Diámetro etapas 15-20 m	Altura Secc Sup m	Altura Secc Inf m	Espaciado Et 2-17 m	Espaciado Et 18-20 m	Estimac Costo MMU$S
1	5.6	4	13.05	7.56	0.91	1	-
2	4.43	1.71	13.05	6.38	0.91	0.91	1.295
3	4.8	2.9	13.05	6.38	0.91	0.91	1.542
4	4.25	2.2	13.05	7.56	0.91	1	1.327

Tabla 3. Variables Constructivas de la Unidad de Destilación Reactiva

Fracc Hueca Aréa Activa %	Diámetro Orificio Mm	Largo Vertedero	Altura Vertedero cm	Altura Sobres Cat cm	Dens. Lecho Catalíco Kg/m3	Carga Cat Etap Rvas Kg
6	4	0.765 D	10	8	760	294

4. CONCLUSIONES

Se ha presentado un estudio de las posibilidades de procesamiento incluyendo una columna de destilación reactiva, como proceso único y como columna de "finishing" como aplicación de métodos de optimización. Las configuraciones estudiadas en detalle fueron la de columna reactiva como único equipo, y también en serie con un pre-reactor que se utilice para llevar a cabo una parte de la conversión global del proceso. Los escenarios 2 y 4 aparecen como los más prometedores. La performance del proceso en el escenario 4 es levemente superior a expensas de la necesidad de incorporar un pre-reactor con los costos de capital y potenciales dificultades operativas que esto conlleva asociado. La decisión sobre que configuración adoptar deberá tomarse sobre la base de una rigurosa evaluación técnico económica.

REFERENCIAS

[1] ALMEIDA-RIVERA C.P (2005), "Designing Reactive Distillation Processes with Improved Efficiency", Tesis Doctoral, Technische Universitet Delft, Nederlands

[2] BAUR R., TAYLOR R., KRISHNA R.,"Influence of Column Hardware on the performance of reactive Distillation Columns ", Catalysis Today 66 (2001), , pp.225-232.

[3] BENNET D.L., AGRAWALD R., COOK P.J. (1983). "New Pressure Drop Correlation for Sieve Tray Distillation Columns", AIChE Journal 29, pp. 434-442.

[4] ESPINOSA, H.J., AGUIRRE P.A. Y PÉREZ G.(1995). Product composition region of single-feed reactive distillation columns: mixtures containing inerts. Industrial & Engineering Chemistry Research, 34, pp.853-861.

[5] FAHAD S. AL-HARTHI (2008), "Modelling and Simulation of a Ractive Distillation Unit for the Production of MTBE", Tesis de Maestría, King Saud University, Kingdom of Saudi Arabia.

[6] INSTITUTO FRANCÉS DEL PETRÓLEO (1976). "Manual of Economic Analisys of Chemical Processes", McGraw-Hill.

[7] JONES JR E.M. (1985), US Patent 4536373.

[8] MEYERS R. A. (1986). "Handbook of Chemical Production Processes", McGraw-Hill.

[9] PERRY R. H., GREEN D. W (1997). "Perry's Chemical Engineers' Handbook" 7th Edition, McGraw-Hill.

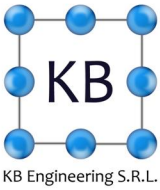
[10] STICHLMAIR J.G., FAIR J.R. (1998). "Distillation: Principles and Practices", J. Wiley and Sons.

[11] TAYLOR R., KRISHNA R. (2000)."Modelling Reactive Distillation", Chemical Engineering Science 55, pp. 5183-5229.

[12] TERMOINT, SOFTWARE DE CÁLCULO DE PROPIEDADES FÍSICAS DE COMPONENTES BASADO EN LA BASE DE DATOS DIPPR 801 (2003) desarrollado por Nuñez D. y Zabaloy M. (PLAPIQUI).

[13] VAN BATEN J.M., KRISHNA R. (2000)."Modelling Sieve Tray Hydraulics using computational Fluid Dynamics", Chemical Engineering Journal 77, pp. 143-151.


BANCOPATAGONIA


KB
KB Engineering S.R.L.


siam
Society for Industrial and Applied Mathematics
Society and Industry Advance with Mathematics


CONICET


asamaci
ASOCIACIÓN ARGENTINA DE MATEMÁTICA APLICADA, COMPUTACIONAL E INDUSTRIAL


AGENCIA


Departamento
de Matemática
UNIVERSIDAD NACIONAL DEL SUR


Departamento de
Ingeniería Química
UNIVERSIDAD NACIONAL DEL SUR
BAHIA BLANCA