Revista Ciencias Matemáticas

Volumen 32, No. 2, Noviembre 2018

A los Lectores

La revista Ciencias Matemáticas está dirigida a investigadores, profesores y estudiantes interesados en resultados recientes en la Matemática o la Ciencia de la Computación. Consecuentemente en la revista se publican artículos originales sobre todas las áreas de la Matemática o la Ciencia de la Computación así como aquellos en los cuales se discuten valoraciones didácticas sobre la impartición, en el nivel universitario, de estas ciencias. Se divulgan también ensayos histórico-críticos que expongan una reflexión sobre el desarrollo de los conocimientos matemáticos y computacionales.

La revista Ciencias Matemáticas publica un nuevo volumen cada año distribuido en dos números. La fecha de salida de cada número no está predeterminada aunque usualmente el primer número sale en el primer semestre del año, el otro número en el segundo semestre.

ISSN: 0256 - 5374
RNPS No: 0228, **Folio:** 76, **Tomo:** I

Dirección Postal:
Universidad de La Habana
Facultad de Matemática y Computación
San Lázaro y L, Edificio Felipe Poey
Vedado, La Habana, Cuba, CP 10400

Sitio Web: http://www.matcom.uh.cu

[1] https://creativecommons.org/licenses/by-nc-nd/4.0/

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 81-88, 2018
Recibido 02-2018

Empleo de técnicas de regularización para la solución del problema inverso unidimensional en la Tomografía Óptica Difusa

Use of regularization technics for the solution of the one-dimensional Difusse Optical Tomography inverse problem

Carmen Tejeda Toledo[1]*, Luis Orlando Castellanos Pérez[2]

Resumen Los problemas inversos se han convertido en una herramienta de gran potencial dentro de las Matemáticas. En la presente investigación se abordarán elementos relacionados con la teoría general de los problemas inversos y su aplicabilidad a la tomografía óptica difusa. Se hará uso de la pseudoinversa de Moore-Penrose como caso particular de inversa generalizada. Se describirán elementos relacionados con el método de descomposición en valores singulares de una matriz, la solución de sistemas lineales en el sentido de mínimos cuadrados y la teoría general de la regularización, abordada a través de tres métodos fundamentales: la descomposición en valores singulares truncada, la regularización de Tikhonov y el método de las iteraciones de Landweber. Además se formulará y resolverá el problema inverso unidimensional en la tomografía óptica difusa por los métodos ya mencionados, realizando una comparación entre ellos para seleccionar el que brinda mejor comportamiento ante diferentes variaciones de los parámetros involucrados.

Abstract Inverse problems have become a tool of great potential in Mathematics. In the current research, elements related to the general theory of inverse problems and their applicability to the diffuse optical tomography are exposed. The Moore-Penrose pseudoinverse as particular case of generalized inverse is also presented. Furthermore, elements related to the singular value decomposition of a matrix, the solution of linear systems in the sense of least square, and the general theory of regularization, approached through three fundamental methods: the truncated singular value decomposition, the Tikhonovs regularization, and the Landwebers iterative method are described. The methods already mentioned are used to numerically solve the one dimensional version of the inverse problem in diffuse optical tomography, which model is also described in this paper, making a comparison between them to select the one that provides better behavior against different variations of the parameters involved.

Palabras Clave
pseudoinversa — tomografía — problema inverso — regularización

[1] *Departamento de Licenciatura en Matemática, Universidad de Holguín, Holguín, Cuba, ctejedat@uho.edu.cu*
[2] *Departamento de Licenciatura en Matemática, Universidad de Holguín, Holguín, Cuba, locp@uho.edu.cu*
***Autor para Correspondencia**

Introducción

La tomografía óptica difusa construye imágenes a partir del comportamiento de un grupo de fotones al atravesar un cuerpo utilizando luz casi infrarroja. Ofrece varias ventajas: es una modalidad de examen no invasiva, da posibilidad de ser usada para exámenes neonatales, el sistema de obtención de imágenes es seguro, de bajo costo, relativamente simple y, en la mayoría de los casos, incluso transportable; además la obtención de informaciones precisas sin causar daños a los pacientes es de gran valor para un diagnóstico certero. El enfoque más usado de este problema inverso responde a la versión tridimensional, para la cual se han planteado varios métodos de regularización y modelos computacionales. La complejidad que encierra esta versión tridimensional la hace inaccesible para un país subdesarrollado como Cuba donde todavía no se ha estudiado esta técnica, hasta donde pudo conocer esta investigación, ni se cuenta con los recursos materiales para implementarla; entonces se impone la necesidad de desarrollar y resolver la versión unidimensional de este problema inverso como base para comprender y poder lle-

var a la práctica el problema tridimensional. Lo planteado anteriormente, genera la siguiente **problemática** expresada en el hecho de que desde el punto de vista matemático, en la actualidad, todavía no hay una teoría totalmente completa que permita apoyar las investigaciones médicas dedicadas a la detección del cáncer de manera no invasiva. Es por ello que esta investigación se propone como **objetivo**: determinar cuál método de regularización lineal brinda mejor precisión para resolver el problema inverso unidimensional en la tomografía óptica difusa..

1. Problemas Inversos

En las últimas dos décadas los problemas inversos se han posicionado como una de las áreas de mayor crecimiento en Matemática Aplicada. En [4] se define como problema inverso aquel que tiene el objetivo de determinar causas a través de efectos observados. Las teorías físicas permiten hacer predicciones, es decir, dada una descripción completa de un sistema físico (modelo), se puede predecir el resultado de algunas mediciones, éste es el llamado problema directo. El problema inverso consiste en utilizar el resultado real de algunas medidas para deducir los valores de los parámetros que caracterizan el sistema. En general, se considera a los Problemas Inversos como rama de las Matemáticas a raíz de la aparición de los trabajos de Tikhonov [9], con la introducción de los métodos de regularización para problemas mal planteados: el argumento básico es que sus ideas permitieron a la comunidad científica romper con lo que ahora se considera un prejuicio histórico y que tiene su origen en un concepto que, por otro lado, ha hecho avanzar grandemente las Ecuaciones en Derivadas Parciales: el concepto de problema bien propuesto o bien planteado (*well-posed*). Hadamard [5] afirmó que los problemas de interés físico son aquellos que tienen una solución única que depende continuamente de los datos, es decir, es estable con respecto a toda perturbación en los datos. Los problemas que no satisfacen alguna de estas condiciones, se denominan mal planteados (*ill-posed*).

2. Preliminares Matemáticos. Inversa generalizada

La inversa generalizada es una generalización de la noción de inversa de matrices la cual se aplica a matrices no cuadradas, como caso particular de inversa generalizada se encuentra la pseudoinversa de Moore-Penrose [1], descubierta por Eliakim Hastings Moore en 1920 y reinventada por Sir Roger Penrose en 1955.

Definición 2.1 ***Pseudoinversa de Moore-Penrose.*** *Sean m y $n \in \mathbb{N}$ y sea $A \in \mathbb{R}^{m\times n}$. Una matriz $A^+ \in \mathbb{R}^{n\times m}$ se dice que es la pseudoinversa de Moore-Penrose de A si satisface las condiciones siguientes:*

1. *Condición general: $AA^+A = A$*
2. *Condición reflexiva: $A^+AA^+ = A^+$*
3. *Condición normalizada: $AA^+ \in \mathbb{R}^m$ es autoadjunta, es decir $AA^+ = (AA^+)^t$*
4. *Condición normalizada reversa: $A^+A \in \mathbb{R}^n$es autoadjunta, es decir $A^+A = (A^+A)^t$*

2.1 Descomposición en valores singulares

La descomposición en valores singulares (SVD) es una importante factorización de una matriz rectangular análoga a la diagonalización de matrices simétricas. Los siguientes teoremas [1] garantizan la SVD de cualquier matriz A y una expresión para su pseudoinversa.

Teorema 2.1 *Sea $A \in \mathbb{R}^{m\times n}$ y $rang\,(A) = r = min\{m,n\}$. Entonces existen matrices unitarias $U \in \mathbb{R}^m$ y $V \in \mathbb{R}^n$ tales que*

$$A = USV^t \tag{1}$$

donde $S \in \mathbb{R}^{m\times n}$ es una matriz que tiene como r primeros elementos de la diagonal los valores singulares positivos de A y todos los otros elementos iguales a cero.

Teorema 2.2 *La pseudoinversa de Moore-Penrose $A^+ \in \mathbb{R}^{n\times m}$ de A está dada por*

$$A^+ = VS^+U^t \tag{2}$$

donde S^+ es la pseudoinversa de S, cuyos elementos son los recíprocos de los valores singulares no nulos de A, y U y V son las matrices ortogonales que resultan de descomponer a A en valores singulares.

2.2 Solución de sistemas lineales en el sentido de mínimos cuadrados

Se parte del siguiente problema. Sean $A \in \mathbb{R}^{m\times n}$ y $y \in \mathbb{R}^m$ dados. Determinar $x \in \mathbb{R}^n$ tal que $y = Ax$. Cuando este problema no tiene solución lo que se hace es relajar el concepto de solución considerando soluciones aproximadas; en mínimos cuadrados se toma $\hat{x} \in \mathbb{R}^n$ que cumple $\|A\hat{x} - y\| = min_{x\in\mathbb{R}^n}\|Ax - y\|$ [1].

Teorema 2.3 *Sea $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$. El problema de minimización*

$$\widehat{x} = argmin_{x\in\mathbb{R}^n}\|Ax - y\|$$

tiene la misma solución $\hat{x} \in \mathbb{R}^n$ que la ecuación $A^tA\hat{x} = A^ty$. Además es única si y solo si $\mathcal{N}(A^tA) = \{0\}$.

3. Métodos de regularización para la solución de problemas mal planteados

El mal planteamiento de los problemas inversos es resuelto usando técnicas conocidas como de regularización que consisten en introducir alguna clase de información a priori acerca de la solución deseada para estabilizar el problema. En términos matemáticos, su objetivo consiste en aproximar la solución x de la ecuación $y = Ax$, a partir del conocimiento del dato directo perturbado y^δ con un nivel de error dado:

$\|y-y^{\delta}\| \leq \delta$. Las definiciones y los teoremas que se brindan a continuación reflejan los aspectos esenciales dentro de la teoría de la regularización [6].

Definición 3.1 *Una estrategia de regularización es una familia de matrices $R_\alpha : \mathbb{R}^m \longrightarrow \mathbb{R}^n$, $\alpha \geq 0$, de forma que*

$$\lim_{\alpha \to 0} R_\alpha A x = x, \forall x \in \mathbb{R}^n \tag{3}$$

El parámetro α es llamado parámetro de regularización y $x^{\alpha,\delta} = R_\alpha y^{\delta}$ es la aproximación de la solución x de $y = Ax$.

Teorema 3.1 *Sea $A \in \mathbb{R}^{m\times n}$ una matriz con sistema singular (u_i, σ_i, v_i), y sea q una función, $q : (0,\infty) \times [\sigma_r, \sigma_1] \longrightarrow \mathbb{R}$, tal que para cada $\alpha \geq 0$ existe una constante $C(\alpha)$ de forma que:*

1. $|q(\alpha,\sigma)| \leq C(\alpha)\sigma$, $\sigma \in [\sigma_r, \sigma_1]$
2. $\lim_{\alpha\to 0} q(\alpha,\sigma) = 1$, $\sigma \in [\sigma_r, \sigma_1]$

Entonces la familia de matrices $R_\alpha \in \mathbb{R}^{n\times m}$, $\alpha > 0$, definidas como

$$R_\alpha y := \sum_{i=1}^{r} \frac{1}{\sigma_i} q(\alpha, \sigma_i)(y, u_i) v_i, y \in \mathbb{R}^m \tag{4}$$

describe una estrategia de regularización con $\|R_\alpha\| \leq C(\alpha)$. La función q es llamada una función filtro regularizadora para A.

Teorema 3.2 *Supongamos que la primera hipótesis del teorema **3.1** se cumple entonces, si la segunda hipótesis del teorema **3.1** se sustituye por:*

1. *Existe $C_1 > 0$ tal que $|q(\alpha,\sigma) - 1| \leq C_1 \dfrac{\sqrt{\alpha}}{\sigma}$, $\forall \alpha > 0$ y además $x \in \mathscr{R}(A^t)$ entonces $\|R_\alpha A x - x\| \leq C_1 \sqrt{\alpha}\|z\|$ donde $x = A^t z$.*
2. *Existe $C_2 > 0$ tal que $|q(\alpha,\sigma) - 1| \leq C_2 \dfrac{\alpha}{\sigma^2}$, $\forall \alpha > 0$ y además $x \in \mathscr{R}(A^t A)$ entonces $\|R_\alpha A x - x\| \leq C_2 \alpha \|z\|$ donde $x = A^t A z$.*

3.1 Descomposición en valores singulares truncada

La descomposición en valores singulares truncada (TSVD) es un método para mejorar el mal condicionamiento del problema reemplazando los más pequeños valores singulares no nulos de A por ceros [1].

Teorema 3.3 *Sea $A \in \mathbb{R}^{m\times n}$ no nula con descomposición en valores singulares $A = USV^t$ donde $S_{11} \geq S_{22} \geq \ldots \geq S_{rr} > 0$ y $S_{ij} = 0 \ \forall \ i \neq j$. La descomposición en valores singulares truncada (TSVD) de A es la matriz*

$$A_{(k)} = U S_{(k)} V^t$$

donde $k \in \{1, 2, \ldots, r-1\}$ y $\left(S_{(k)}\right)_{ii} = S_{ii}$ cuando $i < k$ y $\left(S_{(k)}\right)_{ij} = 0$ en otro caso.

3.2 Regularización de Tikhonov

Este método puede ser introducido de dos formas [6]: a través de un problema de minimización o como un caso especial del Teorema **3.1**. Para el primer caso Tikhonov sugirió, para superar el mal planteamiento, transformar la ecuación normal a una ecuación de la forma $(A^tA + \alpha\mathbb{I})x = A^t y$ que tiene la misma solución que el problema de minimización:

$$x_\alpha = argmin \left\{ \|Ax - y\|^2 + \alpha \|x\|^2 \right\}, \ \ \alpha > 0.$$

Tomando en cuenta que se trabaja con datos perturbados y^{δ}, se define el funcional de Tikhonov como:

$$J_\alpha(x^{\delta}) = \|Ax^{\delta} - y^{\delta}\|^2 + \alpha \|x^{\delta}\|^2, \alpha = \alpha(\delta) > 0. \tag{5}$$

Además el valor ínfimo de este funcional $x^{\alpha,\delta}$ debe satisfacer: $\|Ax^{\alpha,\delta} - y^{\delta}\| = \delta$. Una descripción formal de esta técnica como un problema de minimización se refleja en el siguiente teorema [6]:

Teorema 3.4 *Sea $A \in \mathbb{R}^{m\times n}$ y $\alpha > 0$. Entonces para cada $y \in \mathbb{R}^m$ existe un único $x^\alpha \in \mathbb{R}^n$ tal que*

$$J_\alpha(x^\alpha) = \min_{x\in\mathbb{R}^n} J_\alpha(x)$$

El minimizador x^α coincide con la solución única de la ecuación normal

$$\left(A^tA + \alpha\mathbb{I}\right) x^\alpha = A^t y.$$

Una descripción formal de la regularización de Tikhonov como caso particular del Teorema **3.1** es la siguiente [6]:

Teorema 3.5 *Sea $A \in \mathbb{R}^{m\times n}$ una matriz dada. Entonces, para cada $\alpha > 0$ la matriz $A^tA + \alpha\mathbb{I}$ es inversible. Más aún, la familia $R_\alpha := (A^tA + \alpha\mathbb{I})^{-1}A^t$ describe una estrategia de regularización con $\|R_\alpha\| \leq \dfrac{1}{2\sqrt{\alpha}}$.*

Es importante preguntarse en cualquier problema que se desee resolver ¿cómo elegir el parámetro de regularización α? En este trabajo se escogió el principio de discrepancia de Morozov [6] como criterio de elección de dicho parámetro.

Teorema 3.6 *Sea $A \in \mathbb{R}^{m\times n}$ con rango completo por columna. Sea $y = Ax$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $y^{\delta} \in \mathbb{R}^m$ y $\|y^{\delta} - y\| \leq \delta \|y^{\delta}\|$. Sea $x^{\alpha(\delta),\delta}$ la solución del método de Tikhonov que satisface $\|Ax^{\alpha(\delta),\delta} - y^{\delta}\| = \delta$, $\forall \delta \in (0, \delta_0)$. Entonces: $x^{\alpha(\delta),\delta} \to x$ para $\delta \to 0$, es decir, el Principio de discrepancia es admisible.*

3.3 Iteraciones de Landweber

Los algoritmos para la regularización de Tikhonov tienden a destruir la estructura específica de la matriz de los coeficientes cuando esta tiene grandes dimensiones. En 1951, Landweber [7] realizó la sugerencia de escribir para el sistema $y = Ax$ la ecuación normal $A^tAx = A^ty$ en la forma de una ecuación de punto fijo $x = (\mathbb{I} - \lambda A^tA)x + \lambda A^t y$ para algún $\lambda > 0$, llamado parámetro de relajación. El método iterativo clásico de Landweber [6] tiene el siguiente algoritmo:

$$x^0 = 0; x^{k+1} = x^k - \lambda A^t(Ax^k - y) = (\mathbb{I} - \lambda A^tA)x^k + \lambda A^t y \tag{6}$$

donde $k \in \mathbb{N}$, $\lambda \in \mathbb{R}$. Por inducción con respecto a k se obtiene que $x^k = R_k y$, donde

$$R_k := \lambda \sum_{i=0}^{k-1} \left(\mathbb{I} - \lambda A^t A\right)^i A^t, k = 1,2,...$$

Si (u_j, σ_j, v_j) es el sistema singular de A, entonces R_k admite la siguiente representación:

$$R_k y = \sum_{j=1}^{r} \frac{q(k,\sigma_j)}{\sigma_j}(y,u_j)v_j \tag{7}$$

donde $q(k,\sigma) = 1-(1-\lambda\sigma^2)^k$ es una función filtro que cumple las hipótesis del Teorema **3.2** para $C(\alpha) = \dfrac{1}{2\sqrt{\alpha}}$, $C_1 = \dfrac{1}{2}, C_2 = 1$.

Teorema 3.7 *Sea $A \in \mathbb{R}^{m\times n}$ una matriz dada, y sea $0 < \lambda < \dfrac{1}{\|A\|^2}$. Entonces la familia de matrices $R_k \in \mathbb{R}^{n\times m}$, definidas por (7) describen una estrategia de regularización con parámetro $\alpha = \dfrac{1}{k}$ y $\|R_k\| \leq \sqrt{k\lambda}$.*

La iteración de Landweber es un método de regularización lineal siempre que la iteración esté truncada en algún índice finito k^* [6]. Existen dos criterios de parada para la determinación de k^*, uno a priori donde solo depende del nivel de ruido $k^* = k^*(\delta)$ y otro a posteriori donde además también depende de los datos perturbados $k^* = k^*(\delta, y^\delta)$. En este trabajo se emplea el criterio a posteriori dado por el Principio de discrepancia de Morozov que realiza la iteración siempre que $\|Ax^{k,\delta} - y^\delta\| > \eta\delta$ se cumpla con $\eta > 1$.

4. Problema inverso en la Tomografía Óptica Difusa (DOT)

El objetivo de la obtención de imágenes ópticas usando luz difusa casi infrarroja, es obtener información cuantitativa acerca de cambios en las propiedades ópticas dentro del tejido usando mediciones de frontera, es decir, el objetivo fundamental de la DOT es básicamente reconstruir los coeficientes de absorción y dispersión de un medio macroscópico para mediciones en la frontera. El problema en cuestión parte de los siguientes elementos [2]. Sea $\Omega = \{x : x \geq 0\}$. La densidad de energía Φ obedece a la versión unidimensional de la ecuación de la difusión de time-dependence:

$$\frac{\partial}{\partial t}\Phi(x,t) = D\frac{\partial^2}{\partial x^2}\Phi(x,t) - c\mu_a(x)\Phi(x,t) \tag{8}$$

donde $x \in \Omega$ y el coeficiente de difusión D se toma constante, μ_a se denomina coeficiente de absorción. La densidad de energía debe satisfacer las condiciones inicial y de frontera siguientes:

$$\Phi(x,0) = \delta(x-x_1) \quad \Phi(0,t) - l_{ext}\frac{d}{dt}\Phi(0,t) = 0 \tag{9}$$

Como Φ decrece exponencialmente, se considera para $k \geq 0$ la transformada de Laplace:

$$\Phi(x,k) = \int_0^\infty e^{-k^2Dt}\Phi(x,t)\,dt \tag{10}$$

que satisface la ecuación siguiente, donde la dependencia de Φ en k se suprime para simplificar notación:

$$-\frac{d^2}{dx^2}\Phi(x) + k^2(1+\eta(x))\Phi(x) = \frac{1}{D}\delta(x-x_1) \tag{11}$$

donde k es el número de onda difusa y $\eta(x)$ es la parte espacialmente variante de la absorción. El **problema directo** consiste en dadas las distribuciones de fuentes emisoras de fotones en la frontera del dominio y dado el valor de los parámetros ópticos relacionados, determinar el flujo de fotones resultante en la frontera. La solución del problema directo está dada por la siguiente ecuación integral

$$\Phi(x) = \Phi_i(x) - k^2\int_\Omega G(x,y)\Phi(y)\eta(y)\,dy \tag{12}$$

donde $G(x,y)$ es una función de Green de la forma

$$G(x,y) = \frac{1}{2Dk}\left(e^{-k|x-y|} + \frac{1-kl_{ext}}{1+kl_{ext}}e^{-k|x+y|}\right) \tag{13}$$

y $\Phi_i(x)$ es el campo incidente que satisface **(11)** con $\eta = 0$. La ecuación integral **(12)** puede ser linealizada con respecto a $\eta(x)$ reemplazando Φ en la parte derecha por Φ_i, la cual es una aproximación precisa cuando el soporte de η ($supp(\eta)$) y η son pequeños. Si además se introduce el dato de dispersión $\Phi_s = \Phi_i - \Phi$, se obtiene:

$$\Phi_s(x_1,x_2) = k^2\int_\Omega G(x_1,y)G(y,x_2)\eta(y)\,dy \tag{14}$$

Aquí $\Phi_s(x_1,x_2)$ es proporcional al cambio en la intensidad debido a un punto fuente en x_1 que es medido por un detector en x_2. En la geometría de retrodispersión, la fuente y el detector están ubicados en el origen ($x_1 = x_2 = 0$); utilizando esto, junto a la ecuación **(13)** y omitiendo constantes totales, **(14)** se convierte en la siguiente ecuación integral de Fredholm de primera especie:

$$\Phi_s(k) = \int_0^\infty e^{-kx}\eta(x)dx \tag{15}$$

Lo que permite reformular el **problema directo** como: dada la parte espacialmente variante del coeficiente de absorción $\eta(x)$ y el núcleo $K = e^{-kx}$, determinar el dato de dispersión $\Phi_s(k)$. Entonces el **problema inverso** puede definirse como: dado el núcleo $K = e^{-kx}$ y el dato de dispersión $\Phi_s(k)$ determinar la parte espacialmente variante del coeficiente de absorción $\eta(x)$. Dicho problema inverso podría verse como el de invertir la Transformada de Laplace el cual es un problema exponencial mal planteado y su resultado es una integral de gran complejidad:

$$\eta(x) = \int_0^\infty dk\int_{-\infty}^\infty dsR(\frac{1}{\sigma_s})f_s(x)g_s^*(k)\Phi_s(k) \tag{16}$$

donde el regulador R es introducido para controlar la contribución de los valores singulares pequeños.

Otra vía de solución pudiera ser tomar dicha integral desde cero hasta un valor finito L, es decir, transformar el intervalo de $[0;\infty)$ a un intervalo finito $[0;L]$, lo cual es posible pues solo se necesita considerar que la señal x desaparece fuera de este intervalo que constituye el soporte de dicha variable. Una vía eficiente de resolver este problema inverso mal planteado sería discretizar dicha ecuación integral y luego aplicar los diferentes métodos abordados en la sección anterior.

4.1 Discretización de la ecuación integral

En esta investigación se utilizan dos reglas de discretización de ecuaciones integrales: la regla de Simpson compuesta y la trapezoidal compuesta. La primera se emplea para generar la matriz de los coeficientes y así resolver el problema directo. La segunda regla se emplea para generar la matriz de los coeficientes y así resolver el problema inverso. La **regla de Simpson** compuesta y la **regla Trapezoidal compuesta** constituyen dos de las fórmulas de Newton-Cotes cerradas para la integración numérica, ambas consisten en dividir el intervalo $[0,L]$ en n subintervalos de longitud h. Las hipótesis de la regla de Simpson son las siguientes [3]:

Teorema 4.1 *Sea $f \in C^4[a,b]$, n par, $h=\dfrac{b-a}{n}$ y $x_j = a+jh$, $j=0,1,...,n$. Entonces existe $\mu \in (a,b)$ tal que la regla de Simpson compuesta para n subintervalos puede ser escrita como:*

$$\int_a^b f(x)dx = \frac{h}{3}[f(a)+2\sum_{j=1}^{\frac{n}{2}-1} f(x_{2j})+$$

$$+4\sum_{j=1}^{\frac{n}{2}} f(x_{2j-1})+f(b)]-\frac{b-a}{180}h^4 f^{IV}(\mu) \qquad (17)$$

Utilizando las mismas hipótesis anteriores pero cambiando el hecho de que solo es necesario que $f \in C^2[a,b]$, la regla trapezoidal puede enunciarse como:

$$\int_a^b f(x)dx = \frac{h}{2}[f(a)+2\sum_{j=1}^{n-1}(f(x_j))+f(b)]-$$

$$-\frac{b-a}{12}h^2 f''(\mu) \qquad (18)$$

La función en cuestión es $f(x)=e^{-k_i x}\eta(x)$, $i=1,...,m$, $h=\dfrac{L}{n}$ y $x_j = jh$ por tanto aplicando Simpson:

$$\Phi_s(k_i)=\int_0^L e^{-k_i x}\eta(x)dx \approx \tfrac{h}{3}[e^{-k_i x_0}\eta(x_0)+$$

$$+2\sum_{j=1}^{\frac{n}{2}-1} e^{-k_i x_{2j}}\eta(x_{2j})+4\sum_{j=1}^{\frac{n}{2}} e^{-k_i x_{2j-1}}\eta(x_{2j-1})+e^{-k_i x_n}\eta(x_n)] \qquad (19)$$

Aplicando trapezoidal compuesta:

$$\Phi_s(k_i)=\int_0^L e^{-k_i x}\eta(x)dx \approx$$

$$\frac{h}{2}[e^{-k_i x_0}\eta(x_0)+2\sum_{j=1}^{n-1} e^{-k_i x_j}\eta(x_j)+e^{-k_i x_n}\eta(x_n)] \qquad (20)$$

Si se expresa la sumatoria **(19)** en forma matricial, se obtiene el sistema de ecuaciones lineales $y=Bx$, donde

$$y=\begin{bmatrix} \Phi_s(k_1) \\ \Phi_s(k_2) \\ \vdots \\ \Phi_s(k_m) \end{bmatrix}$$

$$B=\frac{h}{3}\begin{bmatrix} e^{-k_1x_0} & 4e^{-k_1x_1} & 2e^{-k_1x_2} & \cdots & e^{-k_1x_n} \\ e^{-k_2x_0} & 4e^{-k_2x_1} & 2e^{-k_2x_2} & \cdots & e^{-k_2x_n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e^{-k_mx_0} & 4e^{-k_mx_1} & 2e^{-k_mx_2} & \cdots & \end{bmatrix}$$

$$x=\begin{bmatrix} \eta(x_0) \\ \eta(x_1) \\ \eta(x_2) \\ \vdots \\ \eta(x_{n-2}) \\ \eta(x_{n-1}) \\ \eta(x_n) \end{bmatrix}$$

Si se expresa la sumatoria **(20)** en forma matricial, se obtiene el sistema de ecuaciones lineales $y=Ax$, donde x y y se mantiene iguales pero el núcleo B cambia y se transforma en la matriz A dada a continuación:

$$A=\frac{h}{2}\begin{bmatrix} e^{-k_1x_0} & 2e^{-k_1x_1} & 2e^{-k_1x_2} & \cdots & e^{-k_1x_n} \\ e^{-k_2x_0} & 2e^{-k_2x_1} & 2e^{-k_2x_2} & \cdots & e^{-k_2x_n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e^{-k_mx_0} & 2e^{-k_mx_1} & 2e^{-k_mx_2} & \cdots & e^{-k_mx_n} \end{bmatrix}$$

Se concluye que: y es un vector de orden $m\times 1$, A y B son matrices de orden $m\times(n+1)$ y x es de orden $(n+1)\times 1$, es válido aclarar que también se tendrá en cuenta el caso en que $m \neq n+1$.

5. Principales resultados

Como parte de este trabajo se elaboraron códigos de programación en el software MATLAB de ambas reglas de discretización abordadas anteriormente, además de los códigos de la descomposición en valores singulares de una matriz y de los tres métodos de regularización, con sus distintos enfoques. Posteriormente se resuelve un problema test [6], similar al problema inverso unidimensional en la tomografía óptica difusa que viene dado por la siguiente ecuación integral de Fredholm de primera especie:

$$\int_0^1 (1+ts)e^{ts}x(s)ds = e^t,\ 0\le t\le 1,$$

la cual tiene solución única $x(t) = 1$. Con los resultados obtenidos se verificó que los métodos implementados en MATLAB son válidos para resolver un problema inverso mal planteado.

Para resolver el problema inverso unidimensional en la tomografía óptica difusa se emplean los siguiente datos:

$\Omega = \{x \in \mathbb{R} : 0 \leq x \leq 9\}$, se emplean los valores reales de los parámetros ópticos en las longitudes de onda entre $750\,nm$ y $830\,nm$ reportados en [8], $0{,}02 \leq \mu_a \leq 0{,}08$ y $4 \leq \mu_s' \leq 14$, expresados en cm^{-1}, el factor de anisotropía $g = 0{,}9$, el número de onda difusa, expresada en cm, varía de modo decreciente en el rango de 0,157079632679 a 0,04487989505128. Para obtener el rango en que varía la parte espacialmente variante de la absorción $\eta(x)$ se elaboraron en MATLAB dos funciones que muestran el comportamiento de los coeficientes de absorción y dispersión reducido, se creó además un programa principal que se utilizará para el procesamiento de los datos y la obtención de los resultados.

En el mismo se utiliza $m = 26$ y $n = 16$ para la discretización, se determina el rango de variación de la solución exacta $\eta(x)$ dada por un vector de orden 17×1, se genera la matriz B para resolver el problema directo por la regla de Simpson lo cual permite conocer los datos, es decir, la parte derecha del sistema y, la cual es un vector de orden 26×1, se determina por la regla trapezoidal el núcleo A de la ecuación integral para resolver el problema inverso, el cual es una matriz mal condicionada pues su número de condicionamiento $\kappa(A) = 2{,}1256 \cdot 10^{17}$ y sus valores singulares tienden a cero, luego este problema inverso es mal planteado y es necesario aplicar métodos de regularización para su resolución. Se considera que los datos y elaborados anteriormente están sujetos a ciertos ruidos, lo que ocurre frecuentemente debido a que, en general, estos datos provienen de la discretización de una función continua o porque, como es el caso, es un dato obtenido experimentalmente y por tanto está sujeto a errores de medición y aproximación. El vector con ruidos y^δ se generó utilizando el comando de MATLAB *rand* a través de la siguiente expresión $y^\delta = y + \delta(-ones(size(y)) + 2rand(size(y)))$. Al utilizar las funciones implementadas en MATLAB en el programa principal se determina el error entre la solución exacta y la solución con ruidos, sin emplear regularización, y se determinaron los errores para diferentes valores de δ. Las figuras 1 y 2 demuestran el resultado de cualquier intento por resolver el sistema sin regularización. Se realizaron, para cada método, tablas y figuras con valores para el error cuando hay presentes distintos niveles de ruido δ y diferentes variaciones del parámetro de regularización α. Esto permitió realizar comparaciones entre el método de Tikhonov en sus dos variantes y la *TSVD*, en donde se concluyó que Tikhonov como problema de minimización es el mejor de los tres, y al comparar las diferentes variantes del método iterativo de Landweber se concluyó que tanto la que emplea el número de iteraciones *a priori* como la que emplea la función filtro son las más adecuadas, es válido aclarar que su uso depende de la disponibilidad de memoria y de tiempo para realizar la *SVD* que es

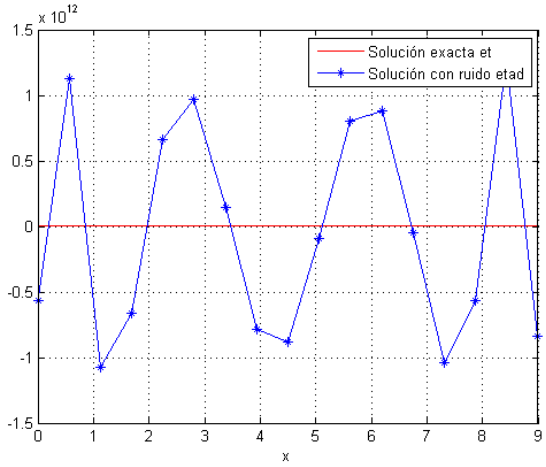


Figura 1. Comparación entre solución exacta y con ruido $\delta = 0{,}5$

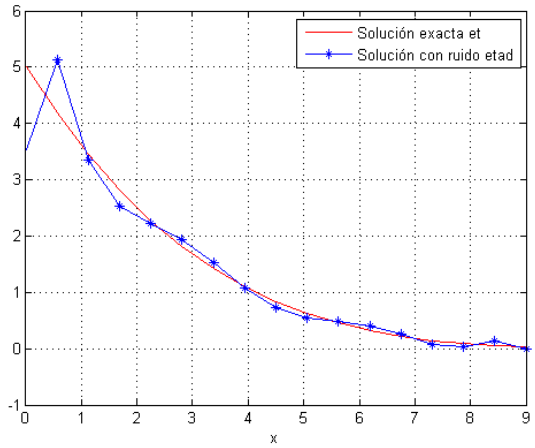


Figura 2. Comparación entre solución exacta y con ruido $\delta = 0$

una operación costosa en ambos sentidos.

Las figuras 3, 5, 4 y 6 reflejan el comportamiento de la solución exacta y el de las soluciones regularizadas, para distintos niveles de ruido ($\delta = 0{,}1$ y $\delta = 0$), obtenidas para Tikhonov, como problema de minimización y para Landweber, con el número de iteraciones dado *a priori*. A partir de ellas se puede notar que los valores obtenidos por Landweber son mayores que los obtenidos por Tikhonov, aunque el primero es más estable con respecto a perturbaciones del lado derecho, inclusive para δ grandes. Para dar cumplimiento al objetivo de esta investigación, que es determinar cuál de los métodos abordados brinda mejor precisión para resolver el problema inverso unidimensional en la tomografía óptica difusa, se concluye que el método de Tikhonov brinda una solución aproximada más precisa que la obtenida por el método iterativo de Landweber.

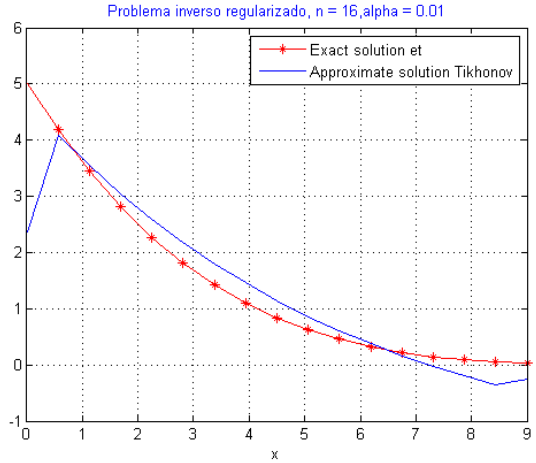


Figura 3. Exacta y Tikhonov $\delta = 0{,}1$

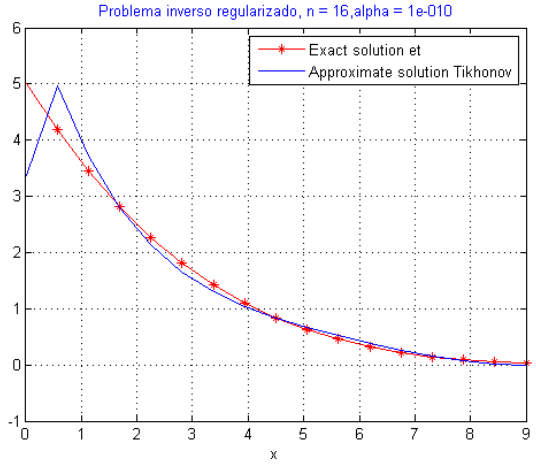


Figura 5. Exacta y Tikhonov $\delta = 0$

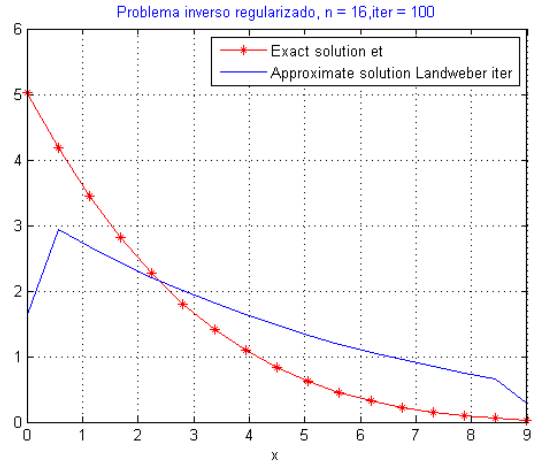


Figura 4. Exacta y Landweber $\delta = 0{,}1$

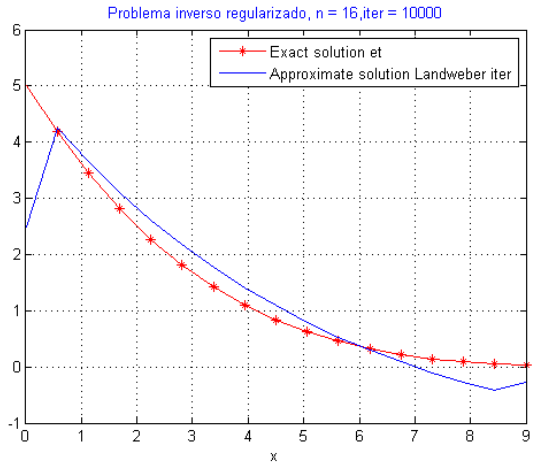


Figura 6. Exacta y Landweber $\delta = 0$

6. Conclusiones

Los resultados obtenidos permiten concluir que:

- Los fundamentos teóricos-metodológicos estudiados sobre la teoría general de los problemas inversos y la teoría general de la regularización, resultaron de vital importancia para un correcto análisis del problema inverso unidimensional en la tomografía óptica difusa.
- La *SVD* de una matriz constituye la base para la realización de una familia importante de métodos de regularización abordados en el trabajo.
- El problema inverso unidimensional en la *DOT* es reducible a la discretización de una ecuación integral de Fredholm de primera especie, el cual es un problema inverso mal planteado, obteniéndose un sistema de ecuaciones lineales al cual resulta imprescindible aplicar los métodos de regularización descritos.
- Las simulaciones numéricas realizadas de los problemas abordados, corroboran las estimaciones teóricas y muestran que el método de Landweber brinda estabilidad con respecto a distintos niveles de ruido y que el método de Tikhonov es el que brinda la solución más precisa.
- La implementación se realizó sobre MATLAB, software que resultó efectivo, por sus atributos: exactitud y confiabilidad.

Agradecimientos

Los autores desean agradecer a todas las personas que de un modo u otro contribuyeron a la realización de este trabajo, sin su dedicación y paciencia esto no hubiera sido posible.

Referencias

[1] Joao Carlos Alves Barata. Notas para un Curso de Física-Matemática. *Departamento de Física-Matemática, USP, Sao Paulo, Brasil*, 2016.

[2] Simon R. Arridge and John C. Schotland. Optical tomography: forward and inverse problems. 2009.

[3] Ricahrd L. Burden and J. Douglas Faires. *Numerical Analisis*. Brooks/Cole, Cengage Learning, USA, 2011.

[4] H.W. Engl and A. Neubauer. Regularization of inverse problems.

[5] J. Hadamard. *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. Yale University Press, New York,, 1923.

[6] Andreas Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Second Edition, Springer, New York,, 2011.

[7] L. Landweber. *An iteration formula for Fredholm integral equations of the first kind*. Am.J.Math., 1951.

[8] Elder Rubens Silveira Rampazzo Filho and Marcelo Idel Vasserman. Tomografía por Óptica Difusa-Protótipo de 16 canales. *Escuela Politécnica de la Universidad de Sao Paulo. Departamento de Ingeniería Mecánica*, 2010.

[9] A.N. Tikhonov. *Solution of incorrectly formulated problems and the regularization method*. 1963.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 89-95, 2018
Recibido 02-2018

Un modelo del déficit de presión en un pozo petrolero usando derivadas Caputo de orden fraccionario

A model of pressure deficit in an oil well using fractional order caputo derivatives

Benito Fernando Martínez Salgado[1], Fernando Brambila Paz[1], Rolando Rosas-Sampayo[1], Carlos Fuentes[2]*

Resumen En esta ponencia se presenta un sistema de ecuaciones de flujo, con derivada temporal fraccionaria, donde se considera el medio como un todo, además de las ecuaciones de flujo incorporadas a un medio de triple porosidad y triple permeabilidad con la misma variante en la derivada temporal, se da una solución semianálitica

Abstract In this paper we present a system of flow equations, with fractional temporal derivative, which considers the medium as a whole, in addition to the flow equations incorporated into a medium of triple porosity and triple permeability with the same variant in the temporal derivative , a semi-analytic solution is given.

Palabras Clave
Ley de Darcy — Difusión anómala — Derivada Fraccionaria

[1] *Departamento de Matemáticas, Facultad de Ciencias, UNAM, Ciudad de México, México, masabemx@yahoo.com.mx, fernandobrambila@gmail.com, rolasmat@ciencias.unam.mx*
[2] *Instituto Mexicano de Tecnología del Agua, Jiutepec, Morelos, México, cbfuentesr@gmail.com*
***Autor para Correspondencia**

Introducción

La correcta modelación de un yacimiento petrolero es de trascendente importancia, ya que permite la toma decisiones que pueden mejorar la obtención de hidrocarburos. A lo largo de los años, diversos enfoques se han utilizado para modelar de manera más completa el comportamiento del fluido dentro del yacimiento a través de la interpretación del déficit de presión. Warren & Root en [14] propusieron ecuaciones en las que consideraban que la matriz del medio y las fracturas tenían una estructura euclidiana. A partir de ese planteamiento Chang y Yortsos en [7] presentan una formulación en la que se considera una fractura fractal en una matriz euclidiana. Camacho-Velazquez et al. en [3] retoman, en parte, esta idea y proponen un modelo de doble porosidad en yacimientos vugulares naturalmente fracturados, en su modelo hacen uso de una derivada de orden fraccionaria tipo Caputo, la cual ya había propuesta en modelos de flujo por Metzler-Glökle-Nonenmacher como se puede ver en [11]. Camacho et al. en su artículo [4] han generalizado una ecuación de flujo clásica a una ecuación que considere el medio como unión de dos o tres medios porosos (medio fracturado, medio vugular y matriz del medio para el último de los casos). Los modelos clásicos se construyen a partir del principio de conservación de la masa de cada uno de los fluidos involucrados en el medio poroso y la ley de Darcy para el fluido en medios porosos como se ilustra por Peaceman en [12]. La ecuación con derivada fraccionaria utiliza una ley de Darcy fraccionaria deducida por Le Mehaute como se ve en [10] en la forma que aparece en el artículo de Raghavan, [13], donde el orden de las derivadas fraccionarias se expresa en términos de la dimensión de Hausdorff del medio.

1. Métodos

El modelo clásico presupone que las propiedades de roca y fluidos son estables, la hidrodinámica del flujo de fluidos en el medio poroso es adecuadamente descrita por la ley de Darcy, la geometría del yacimiento es del tipo euclidiano. La base del modelo se encuentra en la ecuación de continuidad y la ley de Darcy para un flujo a través de un medio poroso, como se ilustra en [2] y en [12], dichas ecuaciones se pueden expresar como:

$$\frac{\partial(\rho\theta)}{\partial t} + \nabla \cdot p(\rho q) = \rho\Upsilon, \tag{1}$$

$$q = -\frac{1}{\mu}k(p)(\nabla p - \rho g \nabla D), \tag{2}$$

donde θ es el contenido volumétrico del fluido;
$q = (q_1, q_2, q_3)$ es el flujo de Darcy, con sus componentes

espaciales (x,y,z), t es el tiempo; ρ es la densidad del fluido; μ es la viscosidad dinámica del fluido; g es la aceleración gravitatoria , Υ es un término de fuente y representa un volumen aportado de fluido por unidad de volumen de medio poroso en la unidad de tiempo; p es la presión; D es la profundidad como una función de coordenadas espaciales, generalmente asimilada a la coordenada vertical z; k es el tensor de permeabilidad del medio poroso $\theta(p)$ y $k(p)$ son características de la dinámica de los fluidos del medio. La ecuación general de transferencia de fluidos se obtiene combinando las ecuaciones como en [12] y [8]:

$$\frac{\partial(\rho\theta)}{\partial t} = \nabla \cdot p[\frac{\rho}{\mu}k(p)(\nabla p - \rho g \nabla D)] + \rho\Upsilon. \tag{3}$$

Esta ecuación diferencial contiene dos variables dependientes, a saber el contenido de humedad y la presión del fluido, pero están relacionadas. Por esta razón, la saturación $S(p)$ está definida así

$$\theta(p) = \phi(p)S(p), \tag{4}$$

donde ϕ es la porosidad total del medio. La capacidad específica está definida por

$$C(p) = \frac{d(\rho\phi S)}{dp} = \phi S\frac{d\rho}{dp} + \rho S\frac{d\phi}{dp} + \rho\phi\frac{dS}{dp}, \tag{5}$$

en consecuencia

$$\frac{\partial(\rho\theta)}{\partial t} = C(p)\frac{\partial p}{\partial t}. \tag{6}$$

En algunos yacimientos las estructuras y/o el comportamiento de los fluidos no son los ideales, así que surge el concepto de memoria, con éste el comportamiento del fluido depende de su trayectoria espacio-temporal y no el clásico markoviano, este concepto ha sido desarrollado por varios investigadores, entre ellos Caputo en [5], también se ha tratado de reflejar una estructura fractal de los medios en el modelo, en nuestro caso usaremos una derivada de tipo Caputo para una versión de la ley de Darcy, descrita por Le Mehaute en [10] y que Raghavan en [13] reescribe como:

$$q(x,t) = -\frac{K_\gamma}{\mu}\frac{\partial^{\gamma-1}}{\partial t^{\gamma-1}}\frac{\partial p(x,t)}{\partial x}, \tag{7}$$

$\gamma = \dfrac{1}{d_f}$, con d_f dimensión fractal de Hausdorff del medio, junto con la ecuación de conservación en coordenadas rectangulares:

$$\frac{\partial}{\partial x_i}q_i(\bar{x};t) = \phi c\frac{\partial}{\partial t}p(\bar{x},t), \tag{8}$$

al combinar las dos anteriores ecuaciones se obtiene con un sistema con simetría radial:

$$\frac{1}{r^{n-1}}\frac{\partial}{\partial r}\left[r^{n-1}\lambda(r)\frac{\partial p(r,t)}{\partial r}\right] = \phi c\frac{\partial^{2-\gamma}}{\partial t^{2-\gamma}}p(r,t), \tag{9}$$

con n la dimensión euclidiana del medio, en nuestro caso $n = 2$,

1.1 Cálculo Fraccional

Existen varias definiciones de derivada fraccionaria: la más difundida es la de Riemann-Liouville, solo daremos la definición de la derivada Caputo por ser la que utilizaremos, una referencia muy completa en el área es el libro de Baleanu et al. [1]. Se define la integral fraccionaria de Riemann-Liouville de orden $\alpha > 0$ como:

$${}_tJ^\alpha f(t) := \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}f(\tau)d\tau,\ \alpha > 0. \tag{10}$$

Con la convención ${}_tJ^0 = I$ (operador identidad) y la propiedad de semigrupo:

$${}_tJ^\alpha {}_tJ^\beta = {}_tJ^\beta {}_tJ^\alpha = {}_tJ^{\alpha+\beta},\ \alpha,\beta \geq 0. \tag{11}$$

Definimos la derivada fraccionaria Caputo de orden $\alpha > 0$ como el operador ${}_tD^\mu_*$ tal que ${}_tD^\mu_* f(t) := {}_tJ^{m-\mu}{}_tD^m f(t)$, de aquí que

$${}_tD^{\mu_*} = \begin{cases} \dfrac{d^m}{dt^m}\left[\dfrac{1}{\Gamma(m-\mu)}\displaystyle\int_0^t \frac{f(\tau)d\tau}{(t-\tau)^{\mu+1-m}}\right], m-1 < \mu < m \\ \dfrac{d^m}{dt^m}f(t),\ \mu = m \end{cases}. \tag{12}$$

La derivada fraccionaria Caputo satisface la propiedad de que es cero cuando se aplica a una constante. Otra propiedad importante es que se le puede aplicar una transformada de Laplace:

$$\mathscr{L}\{{}_tD^\mu_* f(t);s\} = s^\mu\tilde{f}(s) - \sum_{k=0}^{m-1} s^{\mu-1-k}f^{(k)}(0^+), \tag{13}$$
$$m-1 < \mu < k$$

donde $\tilde{f}(s) = \mathscr{L}\{f(t);s\} = \displaystyle\int_0^\infty e^{-st}f(t)dt$, $s \in \mathbb{C}$, y $f^{(k)}(0^+) := \lim_{t\to 0^+} f(t)$.

1.2 Funciones de Bessel

La siguiente ecuación diferencial de segundo orden

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} - (z^2+\nu^2)y = 0, \tag{14}$$

donde ν es una constante real se llama ecuación de Bessel modificada, las soluciones a la ecuación anterior son llamadas funciones de Bessel modificadas las cuales toman la siguiente forma:

$$K_\nu(z) = \left(\frac{\pi}{2}\right)\frac{I_{-\nu}(z) - I_\nu(z)}{\text{sen}(\nu\pi)}, \tag{15}$$

donde $I_\nu(z)$ son las funciones de Bessel modificadas de primer tipo, se hace notar que I_ν y $I_{-\nu}$ forman un conjunto de soluciones para la ecuación (14) y la ecuación (15) es conocida como la función de Bessel modificada de segundo tipo.

Algunas propiedades de la función de Bessel modificada de segundo tipo son:

$$\frac{d}{dz}K_\nu(\alpha z) = -\alpha K_{\nu-1}(\alpha z) - \frac{\nu}{z}K_\nu(\alpha z), \tag{16}$$

$$\frac{d}{dz}K_\nu(\alpha z) = -\alpha K_{\nu+1}(\alpha z) + \frac{\nu}{z}K_\nu(\alpha z). \tag{17}$$

1.3 Ecuación de flujo con derivada temporal fraccionaria

Podemos simplificar la ecuación (9) que representa el fluido, donde el medio es un todo, así tenemos:

$$\phi c_\alpha \frac{\partial^\alpha p}{\partial t^\alpha} = \frac{k}{\mu}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p}{\partial r}\right), \tag{18}$$

donde la expresión de derivada de la izquierda denota la derivada fraccionaria Caputo de orden $\alpha \in \mathbb{R}$, con variables adimensionales, la ecuación (18) es

$$\phi c_{D\alpha}\frac{\partial^\alpha p_D}{\partial t_D^\alpha} = \kappa_D\frac{1}{r}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_D}{\partial r_D}\right), \tag{19}$$

donde

$$p_D = \frac{2\pi hk(p_i - p)}{Q_0 B_0 \mu}, t_D = t\frac{k}{\phi c r_w^2 \mu}, r_D = \frac{r}{r_w}, \tag{20}$$

donde ϕ representa la porosidad del medio medida en unidades de $\frac{m^3}{m^3}$, c representa la compresibilidad del medio en unidades de Pa^{-1}, k representa la permeabilidad del medio con unidades de m^2, p representa la presión del fluido en el medio con unidades de Pa, μ es la viscosidad del fluido con unidades de $Pa \cdot s$, t representa el tiempo en unidades de s, r representa la distancia del pozo en unidades de m, r_w es un parámetro de referencia: radio del pozo con unidades de m, h es el espesor del pozo medido en m, p_i es la presión inicial del yacimiento, el valor de Q_0 es el caudal con unidades de $m^3 s^{-1}$ y B_0 es el factor del fluido (adimensional).

1.4 Transformada de laplace

La transformada de Laplace aplicada a la ecuación (19) da el siguiente resultado usando la ecuación (13):

$$u^\alpha \bar{p}_D = \frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial \bar{p}_D}{\partial r_D}\right),\ u > 0, \tag{21}$$

debido a que $\bar{p}_D(t_0) = 0$, debido a que $p = p_i$, en $t = t_0$.

1.4.1 Funciones de Bessel

Las derivadas espaciales al ser desarrolladas en la ecuación (21) muestran la siguiente forma:

$$r_D^2\frac{\partial^2 \bar{p}_D}{\partial r_D^2} + r_D\frac{\partial \bar{p}_D}{\partial r_D} - r_D^2 u^\alpha \bar{p}_D = 0, \tag{22}$$

la cual es una ecuación de Bessel, por tanto su solución es:

$$\bar{p}_D = AK_0(\beta r_D). \tag{23}$$

Al sustituir la ecuación (23) en la ecuación (21) y teniendo en mente las ecuaciones (16) y (17) para encontrar el valor de β se tiene que:

$$\beta = \pm\sqrt{u^\alpha},\ u > 0. \tag{24}$$

La ecuación (23) al considerar el valor de β, ecuación (24) es

$$\bar{p}_D = AK^0(r_D\sqrt{u^\alpha}). \tag{25}$$

En la ecuación (25) se descarta $\beta = -\sqrt{u^\alpha}$ debido a que la función de Bessel modificada segundo tipo no está definida para valores negativos.

1.4.2 Condiciones de frontera

Para encontrar la solución de la ecuación (19), se considera la siguiente condición de frontera:

$$r_D\frac{\partial \bar{p}_D}{\partial r_D}\Big|_{r_D=1} = -\frac{1}{u}. \tag{26}$$

La sustitución de la ecuación (25) en (26) genera lo siguiente:

$$A = \frac{1}{u}[\sqrt{u^\alpha}K_1(\sqrt{u^\alpha})]^{-1}, \tag{27}$$

$$\bar{p}_D = \frac{1}{u}[\sqrt{u^\alpha}K_1(\sqrt{u\alpha})]^{-1}K_0(r_D\sqrt{u^\alpha}). \tag{28}$$

Por lo tanto , el valor de la presión en la frontera del pozo ($r_D = 1$) es en el espacio de Laplace:

$$\bar{p}_D|_{r_D=1} = \frac{1}{u}[\sqrt{u^\alpha}K_1(\sqrt{u\alpha})]^{-1}K_0(\sqrt{u^\alpha}). \tag{29}$$

.

2. Ecuación de flujo con triple porosidad y triple permeabilidad con derivada temporal fraccionaria

A partir de las ecuaciones de transferencia clásica, Carlos Fuentes en [8] propone un sistema de ecuaciones de flujo acoplados con triple porosidad y triple permeabilidad, las cuales tienen la siguiente forma:

$$\begin{aligned}\phi_m c_m\frac{\partial p_m}{\partial t} &= \frac{k_m}{\mu}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p_m}{\partial r}\right)\\ &\quad + a_{mf}(p_f - p_m) + a_{mv}(p_v - p_m),\\ &\quad \text{con } c_m = \frac{1}{\phi_m}\frac{\partial \phi_m}{\partial p_m},\end{aligned} \tag{30}$$

$$\begin{aligned}\phi_f c_f\frac{\partial p_f}{\partial t} &= \frac{k_f}{\mu}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p_f}{\partial r}\right)\\ &\quad - a_{mf}(p_f - p_m) + a_{fv}(p_v - p_f),\\ &\quad \text{con } c_f = \frac{1}{\phi_f}\frac{\partial \phi_f}{\partial p_f},\end{aligned} \tag{31}$$

$$\begin{aligned}\phi_v c_v\frac{\partial p_v}{\partial t} &= \frac{k_v}{\mu}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p_v}{\partial r}\right)\\ &\quad - a_{mv}(p_v - p_m) - a_{fv}(p_v - p_f),\\ &\quad \text{con } c_v = \frac{1}{\phi_v}\frac{\partial \phi_v}{\partial p_v}.\end{aligned} \tag{32}$$

donde ϕ_m, ϕ_f, ϕ_v representan las porosidades de la matriz del suelo, el medio fracturado y el medio vugular respectivamente en unidades de m^3/m^3; c_m, c_f, c_v representan la compresibilidad en cada medio poroso en unidades de Pa^{-1}; k_m, k_f, k_v representan la permeabilidad de cada medio poroso con unidades de m^2; p_m, p_f, p_v representan la presión del fluido en cada medio poroso con unidades de Pa; μ es la viscosidad del fluido con unidades de $Pa \cdot s$; a_{mf}, a_{mv}, a_{fv} son los términos de transferencia en las interfaces matriz-fractura, matriz-vúgulo y fractura-vúgulo respectivamente con unidades de $Pa^{-1}s^{-1}$: t representa el tiempo en unidades de s y r representa la distancia al pozo en unidades de m.

2.1 Adimensionalización de las ecuaciones de flujo

Con el fin de manejar las ecuaciones (30), (31) y (32) de una manera más fácil, se aplica la adimensionalización de variables. La adimensionalización es una técnica utilizada comúnmente para hacer que los parámetros o variables en una ecuación no tengan unidades, llevar a un rango los posibles valores de una variable o una constante con el fin de que su valor sea conocido y de esta manera más manipulable. El sistema de ecuaciones (30), (31) y (32) toma, después de aplicar la adimensionalización, la siguiente forma:

$$(1-\omega_f-\omega_v)\frac{\partial p_{Dm}}{\partial t_D} = (1-\kappa_f-\kappa_v)\frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Dm}}{\partial r_D}\right) + \lambda_{mf}(p_{Df}-p_{Dm}) + \lambda_{mv}(p_{Dv}-p_{Dm}), \quad (33)$$

$$\omega_f\frac{\partial p_{Df}}{\partial t_D} = \kappa_f\frac{1}{r_d}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Df}}{\partial r_D}\right) - \lambda_{mf}(p_{Df}-p_{Dm}) + \lambda_{fv}(p_{Dv}-p_{Df}), \quad (34)$$

$$\omega_v\frac{\partial p_{Dv}}{\partial t_D} = \kappa_v\frac{1}{r_d}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Dv}}{\partial r_D}\right) - \lambda_{mv}(p_{Dv}-p_{Dm}) - \lambda_{fv}(p_{Dv}-p_{Df}), \quad (35)$$

donde

$$\omega_f = \frac{\phi_f c_f}{\phi_m c_m + \phi_f c_f + \phi_v c_v}, \quad (36a)$$

$$\omega_g = \frac{\phi_v c_v}{\phi_m c_m + \phi_f c_f + \phi_v c_v}, \quad (36b)$$

$$r_D = \frac{r}{r_w}, \quad (37a)$$

$$\kappa_f = \frac{k_f}{k_m + k_f + k_v}, \quad (37b)$$

$$\kappa_g = \frac{k_v}{k_m + k_f + k_v}, \quad (37c)$$

$$\lambda_{mf} = \frac{a_{mf}\mu r_w^2}{k_m + k_f + k_g}, \quad (38a)$$

$$\lambda_{mv} = \frac{a_{mv}\mu r_w^2}{k_m + k_f + k_g}, \quad (38b)$$

$$\lambda_{fv} = \frac{a_{fv}\mu r_w^2}{k_m + k_f + k_g}, \quad (38c)$$

$$p_{Dj} = \frac{2\pi h(k_m + k_f + k_v)(p_i - p_j)}{Q_0 B_0 \mu}, \quad (39a)$$

$$t_D = \frac{t(k_m + k_f + k_v)}{\mu r_w^2(\phi_m c_m + \phi_f c_f + \phi_v c_v)}. \quad (39b)$$

Las ecuaciones (36)-(39) representan las variables adimensionalizadas, se puede verificar que estas variables no tienen unidades; en la ecuación (38) el valor de r_w es un parámetro de referencia, en este caso el radio del pozo, con el fin de que la variable r_D tenga el valor ínfimo igual a 1, con unidades de m, en la ecuación (39) el valor de h representa el espesor del yacimiento petrolero con unidades de m; p_j son las presiones en los diferentes medios porosos, donde $j = m, f, v$ y p_i es la presión inicial en el yacimiento; el valor de Q_0 es el caudal con unidades de $m^3 s^{-1}$ y B_0 es el factor de formación de fluido (adimensional).

2.2 El sistema con derivada fraccionario

A partir del sistema de ecuaciones con variables adimensionales (33)-(35), usando la ecuación de flujo con derivada temporal fraccionaria (19), expresamos un sistema con derivada temporal fraccionaria:

$$(1-\omega_f-\omega_v)\frac{\partial^\beta p_{Dm}}{\partial t_D^\beta} = (1-\kappa_f-\kappa_v)\frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Dm}}{\partial r_D}\right) + \lambda_{mf}(p_{Df}-p_{Dm}) + \lambda_{mv}(p_{Dv}-p_{Dm}), \quad (40)$$

$$\omega_f\frac{\partial^\beta p_{Df}}{\partial t_D^\beta} = \kappa_f\frac{1}{r_d}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Df}}{\partial r_D}\right) - \lambda_{mf}(p_{Df}-p_{Dm}) + \lambda_{fv}(p_{Dv}-p_{Df}), \quad (41)$$

$$\omega_v\frac{\partial^\beta p_{Dv}}{\partial t_D^\beta} = \kappa_v\frac{1}{r_d}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial p_{Dv}}{\partial r_D}\right) - \lambda_{mv}(p_{Dv}-p_{Dm}) - \lambda_{fv}(p_{Dv}-p_{Df}), \quad (42)$$

donde las variables mostradas en las ecuaciones (40)-(42) tienen el mismo significado que las ecuaciones (36)-(38). Por medio de la transformada de Laplace y con el uso de la ecua-

ción (13) se llega al siguiente sistema :

$$(1-\omega_f-\omega_v)u^\beta \bar{p}_{Dm} = (1-\kappa_f-\kappa_v)\frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial \bar{p}_{Dm}}{\partial r_d}\right) + \lambda_{mf}(\bar{p}_{Df}-\bar{p}_{Dm})+\lambda_{mv}(\bar{p}_{Dv}-\bar{p}_{Dm}), \quad (43)$$

$$\omega_f u^\beta \bar{p}_{Df} = \kappa_f\frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial \bar{p}_{Df}}{\partial r_d}\right) - \lambda_{mf}(\bar{p}_{Df}-\bar{p}_{Dm})+\lambda_{fv}(\bar{p}_{Dv}-\bar{p}_{Df}), \quad (44)$$

$$\omega_v u^\beta \bar{p}_{Dv} = \kappa_v\frac{1}{r_D}\frac{\partial}{\partial r_D}\left(r_D\frac{\partial \bar{p}_{Dv}}{\partial r_d}\right) - \lambda_{mv}(\bar{p}_{Dv}-\bar{p}_{Dm})-\lambda_{fv}(\bar{p}_{Dv}-\bar{p}_{Df}), \quad (45)$$

donde $\bar{p}_{Dm}, \bar{p}_{Df}$ y $\bar{p}_{Dv}$ representan las transformadas de Laplace de las variables p_{Dm}, p_{Df} y p_{Dv}. Al desarrollar las ecuaciones (43)-(45) es fácil ver que cumplen con la forma de una ecuación de Bessel y por tanto sus soluciones, al igual que en el caso $\beta = 1$, son

$$\bar{p}_{Dm} = AK_0(\alpha r_D), \quad (46)$$

$$\bar{p}_{Df} = BK_0(\alpha r_D), \quad (47)$$

$$\bar{p}_{Dv} = CK_0(\alpha r_D), \quad (48)$$

con el fin de simplificar las sucesivas ecuaciones, se definen los siguientes términos:

$$m_1(u) = u^\beta(1-\omega_f-\omega_v)+\lambda_{mf}+\lambda_{mv}, \quad (49a)$$

$$m_2 = \lambda_{mf}, \quad (49b)$$

$$m_3 = \lambda_{mv}, \quad (49c)$$

$$m_4(u) = u^\beta\omega_f+\lambda_{mf}+\lambda_{fv}, \quad (50a)$$

$$m_5 = \lambda_{fv}, \quad (50b)$$

$$m_6(u) = u^\beta\omega_v+\lambda_{mv}+\lambda_{fv}. \quad (50c)$$

Como resultado de sustituir las ecuaciones (46)-(48) en el sistema mostrado en (43)-(45) y haciendo uso de las definiciones mostradas por (49)-(50), se tiene las siguientes:

$$K_0(\alpha r_D)\{A[(1-\kappa_f-\kappa_v)\alpha^2-m_1]+Bm_2+Cm_3\}=0, \quad (51)$$

$$K_0(\alpha r_D)\{Am_2+B[\kappa_f\alpha^2-m_4]+Cm_5\}=0, \quad (52)$$

$$K_0(\alpha r_D)\{Am_3+Bm_5+C[\kappa_v\alpha^2-m_6]\}=0. \quad (53)$$

Puesto que las funciones de Bessel modificadas de segunda especie tienen un comportamiento asintótico, es decir nunca toman el valor de cero, entonces el sistema mostrado en las ecuaciones (51)-(53) puede expresarse como sigue:

$$\begin{bmatrix}(1-\kappa_f-\kappa_v)\alpha^2-m_1 & m_2 & m_3\\ m_2 & \kappa_f\alpha^2-m_4 & m_5\\ m_3 & m_5 & \kappa_v\alpha^2-m_6\end{bmatrix}\begin{bmatrix}A\\B\\C\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix} \quad (54)$$

La ecuación (54) es utilizada para encontrar los valores de A, B y C, con esto en mente hay que notar dos casos principales: el determinante de la matriz 3×3 tiene valor igual a cero o el determinante de dicha matriz es diferente a cero. El primer caso nos da la solución trivial $A = B = 0$. El segundo caso, donde el determinante es igual a cero, se obtiene la ecuación de grado seis que sigue:

$$\begin{aligned}&(1-\kappa_f-\kappa_v)\kappa_f\kappa_v\alpha^6\\ &-[(1-\kappa_f-\kappa_v)(\kappa_f m_6+\kappa_v m_4)+\kappa_f\kappa_v m_1]\alpha^4\\ &+[(1-\kappa_f-\kappa_v)m_4m_6-(1-\kappa_f-\kappa_v)m_5^2\\ &+(\kappa_f m_6+\kappa_v m_4)m_1-\kappa_v m_2^2-\kappa_f m_2^3]\alpha^2\\ &-m_1m_4m_6+m_1m_5^2+m_2^2m_6+2m_2m_3m_5\\ &+m_3^2m_4=0.\end{aligned} \quad (55)$$

En la ecuación anterior las potencias de α son pares, se puede por tanto resolver como una ecuación de grado 3. Esta ecuación tiene tres raíces reales. Las soluciones generales de las ecuaciones (43)-(45) tienen al incorporar las tres raíces reales, de la siguiente forma:

$$\bar{p}_{Dm} = A_1D_1K_0(\alpha_1r_D)+A_2D_2K_0(\alpha_2r_D)+A_3D_3K_0(\alpha_3r_D), \quad (56)$$

$$\bar{p}_{Df} = B_1D_1K_0(\alpha_1r_D)+B_2D_2K_0(\alpha_2r_D)+B_3D_3K_0(\alpha_3r_D), \quad (57)$$

$$\bar{p}_{Dv} = D_1K_0(\alpha_1r_D)+D_2K_0(\alpha_2r_D)+D_3K_0(\alpha_3r_D), \quad (58)$$

donde los términos $A_i, B_i, i = 1,2,3$, tienen la forma:

$$A_1 = \frac{m_3(\kappa_f\alpha_1^2-m_4)-m_2m_5}{m_2^5[(1-\kappa_f-\kappa_v)\alpha_1^2-m_1][\kappa_f\alpha_1^2-m_4]}, \quad (59)$$

$$B_1 = \frac{-m_3-A_1[(1-\kappa_f-\kappa_v)\alpha_1^2-m_1]}{m_2}, \quad (60)$$

$$A_2 = \frac{m_3(\kappa_f\alpha_2^2-m_4)-m_2m_5}{m_2^5-[(1-\kappa_f-\kappa_v)\alpha_2^2-m_1][\kappa_f\alpha_2^2-m_4]}, \quad (61)$$

$$B_2 = \frac{-m_3-A_2[(1-\kappa_f-\kappa_v)\alpha_2^2-m_1]}{m_2}, \quad (62)$$

$$A_3 = \frac{m_3(\kappa_f\alpha_3^2-m_4)-m_2m_5}{m_2^5-[(1-\kappa_f-\kappa_v)\alpha_3^2-m_1][\kappa_f\alpha_3^2-m_4]}, \quad (63)$$

$$B_3 = \frac{-m_3-A_3[(1-\kappa_f-\kappa_v)\alpha_3^2-m_1]}{m_2}, \quad (64)$$

donde los términos D_1, D_2 y D_3 son obtenidos a partir de las condiciones de frontera y $\alpha_1, \alpha_2, \alpha_3$ son las raíces positivas de $\alpha_1^2, \alpha_2^2, \alpha_3^2$.

Los valores de D_1, D_2, D_3 se obtienen a partir de las condiciones de frontera y son iguales a

$$D_1 = \frac{1}{u}\left\{\alpha_1E_1K_1(\alpha_1)+\alpha_3E_3K_1(\alpha_3)\frac{(B_1-1)K_0(\alpha_1)}{(1-B_3)K_0(\alpha_3)} + \frac{\left[(1-A_1)K_0(\alpha_1)+(1-A_3)\frac{(B_1-1)}{(1-B_3)}K_0(\alpha_1)\right]}{\left[(A_2-1)K_0(\alpha_2)+(A_3-1)\frac{(B_2-1)}{(1-B_3)}K_0(\alpha_2)\right]}\alpha_2E_2K_1(\alpha_2) + \alpha_3E_3K_1(\alpha_3)\frac{(B_2-1)K_0(\alpha_2)}{(1-B_3)K_0(\alpha_3)}\right\}^{-1}, \quad (65a)$$

$$D_2 = \frac{1}{u}\left\{ \frac{\left[(A_2-1)K_0(\alpha_2)+(A_3-1)\dfrac{(B_2-1)}{(1-B_3)}K_0(\alpha_2)\right]}{\left[(1-A_1)K_0(\alpha_1)+(1-A_3)\dfrac{(B_1-1)}{(1-B_3)}K_0(\alpha_1)\right]} + \left[\alpha_1E_1K_1(\alpha_1)+\alpha_3E_3K_1(\alpha_3)\frac{(B_1-1)K_0(\alpha_1)}{(1-B_3)K_0(\alpha_3)}\right] + \left[\alpha_2E_2K_1(\alpha_2)+\alpha_3E_3K_1(\alpha_3)\frac{(B_2-1)K_0(\alpha_2)}{(1-B_3)K_0(\alpha_3)}\right]\right\}^{-1}, \tag{65b}$$

$$D_3 = \frac{1}{u}\left\{\alpha_1E_1K_1(\alpha_1)\frac{(1-B_3)K_(\alpha_3)}{(B_1-1)K_0(\alpha_1)}+\alpha_3E_3K_1(\alpha_3) + \left[\frac{(1-A_1)(1-B_3)+(1-A_3)(B_1-1)}{(A_2-1)(1-B_3)+(A_3-1)(B_2-1)}\right]\times \left[\frac{\alpha_2E_2(1-B_3)K_1(\alpha_2)K_0(\alpha_3)+\alpha_3E_3(B_2-1)K_1(\alpha_3)K_0(\alpha_2)}{(B_1-1)K_0(\alpha_2)}\right]\right\}^{-1} + \frac{1}{u}\left\{\alpha_2E_2K_1(\alpha_2)\frac{(1-B_3)K_0(\alpha_3)}{(B_2-1)K_0(\alpha_2)}+\alpha_3E_3K_1(\alpha_3)+ + \left[\frac{(1-A_2)(1-B_3)+(1-A_3)(B_2-1)}{(A_1-1)(1-B_3)+(A_3-1)(B_1-1)}\right]\times \left[\frac{\alpha_1E_1(1-B_3)K_1(\alpha_1)K_0(\alpha_3)+\alpha_3E_3(B_1-1)K_1(\alpha_3)K_0(\alpha_1)}{(B_2-1)K_0(\alpha_1)}\right]\right\}, \tag{65c}$$

donde

$$E_1 = \left[(1-\kappa_f-\kappa_v)A_1+\kappa_fB_1+\kappa_v\right],$$
$$E_2 = \left[(1-\kappa_f-\kappa_v)A_2+\kappa_fB_2+\kappa_v\right],$$
$$E_3 = \left[(1-\kappa_f-\kappa_v)A_3+\kappa_fB_3+\kappa_v\right].$$

Se llega a las siguientes ecuaciones como resultado de sustituir las ecuaciones (56)-(58) en las condiciones de frontera.

$$\alpha_1K_1(\alpha_1)D_1[(1-\kappa_f-\kappa_v)A_1+\kappa_fB_1+\kappa_v] + \alpha_2K_1(\alpha_2)D_2[(1-\kappa_f-\kappa_v)A_2+\kappa_fB_2+\kappa_v] + \alpha_3K_1(\alpha_3)D_3[(1-\kappa_f-\kappa_v)A_3+\kappa_fB_3+\kappa_v] = \frac{1}{u}, \tag{66}$$

$$(A_1-1)D_1K_0(\alpha_1)+(A_2-1)D_2K_0(\alpha_2) + (A_3-1)D_3K_0(\alpha_3) = 0, \tag{67}$$

$$(B_1-1)D_1K_0(\alpha_1)+(B_2-1)D_2K_0(\alpha_2) + (B_3-1)D_3K_0(\alpha_3) = 0. \tag{68}$$

Con el fin de simplificar las ecuaciones (66)-(68) se definen los siguientes términos:

$$P_i = \alpha_iK_1(\alpha_i)[(1-\kappa_f-\kappa_v)A_i+\kappa_fB_i+\kappa_v], \tag{69}$$
$$Q_i = (A_i-1)K_0(\alpha_i),\ R_i = (B_i-1)K_0(\alpha_i), \tag{70}$$

donde $i = 1,2,3$. La ecuación matricial asociada al sistema de ecuaciones (66)-(68) tiene la forma

$$\begin{bmatrix} P_1 & P_2 & P_3 \\ Q_1 & Q_2 & Q_3 \\ R_1 & R_2 & R_3 \end{bmatrix}\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} 1/u \\ 0 \\ 0 \end{bmatrix}. \tag{71}$$

Se define

$$m = Q_1R_2P_3 - Q_1P_2R_3 - R_1Q_2P_3 - R_2P_1Q_3 + P_2R_1Q_3 + P_1Q_2R_3. \tag{72}$$

La solución de la ecuación matricial es

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} \dfrac{(Q_2R_3-Q_3R_2)}{m} \\ \dfrac{-(Q_1R_3-Q_3R_1)}{m} \\ \dfrac{(Q_1R_2-Q_2R_1)}{m} \end{bmatrix}. \tag{73}$$

Ahora que se han obtenido todos los términos que definen la solución del sistema de ecuaciones (43)-(45) resta mostrar el valor de la presión en la frontera del pozo, dicho valor es:

$$\begin{aligned}\bar{p}_w|_{r_d=1} &= D_1K_0(\alpha_1)+D_2K_0(\alpha_2)+D_3K_0(\alpha_3)\\ &= B_1D_1K_0(\alpha 1)+B_2D_2K_0(\alpha_2)+B_3D_3K_0(\alpha_3).\end{aligned} \tag{74}$$

3. Conclusiones

El uso de la derivada Caputo en la ecuación de flujo, permitirían reflejar la fractalidad del medio o los medios en el comportamiento del fluido, un requisito sería tener las dimensiones de los medios, algunos resultados al respecto son de Casar-Gonzalez en [6] y de Hewett en [9].

Referencias

[1] D. Baleanu, K. Diethelm, E. Scalas, and J. J. Trujillo. *Fractional Calculus: Models and Numerical Methods*, volume 3 of *Series on Complexity, Nonlinearity and Chaos*. World Scientific, 2012.

[2] J. Bear. Dynamics of fluids in porous media. 1972.

[3] R. Camacho Velazquez, G. Fuentes-Cruz, and M. A. Vasquez-Cruz. Decline-curve analysis of fractured reservoirs with fractal geometry. *Society of Petroleum Engineers*, 11(3):606–619, 2008.

[4] R. Camacho-Velazquez, M. A. Vasquez-Cruz, R. Castrejon-Aivar, and V. Arana-Ortiz. Pressure transient and decline curve behaviors in naturally fractured vuggy carbonate reservoirs. *Society of Petroleum Engineers*, 8(2), 2005.

[5] M. Caputo and W. Plastino. Diffusion in porous layers with memory. *Geophysical Journal International*, 158(1):385–396, 2004.

[6] R. Casar-Gonzalez and V. Suro-Perez. Stochastic imaging of vuggy formations. SPE International Petroleum Conference and Exhibition in Mexico, 1-3 February, Villahermosa, Mexico, Society of Petroleum Engineers, 2000.

[7] J. Chang and Y. C. Yortsos. Pressure transient analysis of fractal reservoirs. *SPE Formation Evaluation*, 5(1):31–38, 1990.

[8] C. Fuentes. Yacimiento petrolero como un reactor fractal: Un modelo de triple porosidad y permeablidad del medio fracturado. Fondo sectorial CONACYT-SENER-HIDROCARBUROS S0018-2011-11, Proyecto de Grupo. Informe IV, Facultad de Ingeniería, Universidad Autónoma de Querétaro, 2013.

[9] T. A. Hewett. *Stochastic Modeling and Geostatistics: Principles, Methods, and Case Studies*, volume 3 of *AAPG Computer Applications in Geology*, chapter 19. Fractal Methods for Fracture Characterization, pages 249–260. 1994.

[10] A. Le Mehaute. Transfer processes in fractal media. *Journal of Statistical Physics*, 36(5):665–676, 1984.

[11] R. Metzler, W. G. Glöckle, and T. F. Nonnenmacher. Fractional model equation for anomalous diffusion. *Physica A: Statistical Mechanics and its Applications*, 211(1):13–24, 1994.

[12] D. W. Peaceman. *Fundamentals of Numerical Reservoir Simulation*, volume 6 of *Developments in Petroleum Science*. Elsevier Science, 1st edition, 1977.

[13] R. Raghavan. Fractional derivatives: application to transient flow. *Journal of Petroleum Science and Engineering*, 80(1):7–13, 2011.

[14] J. E. Warren and P. J. Root. The behavior of naturally fractured reservoirs. *Society of Petroleum Engineers Journal*, 3(3):245–255, 1963.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 97-106, 2018
Recibido 02-2018

Un esquema spline cónico de Hermite *fair*. A *fair* Hermite quadratic spline scheme.

D. García Pérez[1], J. Estrada Sarlabous[1]*, S. Behar Jequín[2], W. Morales Lezca[2]

Resumen En este trabajo presentamos un nuevo esquema para la interpolación de Hermite de un conjunto de puntos en el plano por medio de una curva spline racional cuadrática. El spline cónico es representado como una curva racional cuadrática de Bézier, el cual depende de un parámetro de tensión local que controla la forma de cada sección. Definimos una familia de funcionales de *fairness* como el conjunto las combinaciones lineales de la longitud de arco y de la energía elástica de la sección cónica. El valor del parámetro de tensión que minimiza el funcional corresponde a la curva *fair*. Aplicando el esquema de subdivisión para splines cónicos propuesto en [5] obtenemos buenas aproximaciones numéricas del funcional y sus derivadas. Se demuestra además que el funcional alcanza su valor mínimo en cada uno de los segmentos del spline y utilizamos un algoritmo numérico para hallarlo. Escogiendo una determinada combinación lineal en el funcional se demuestra que el esquema spline cónico de Hermite propuesto es invariante bajo transformaciones rígidas y homotecias, reproduce arcos de circunferencia y satisface las condiciones presentadas en [7]. El esquema ha sido implementado en MatLab y se presenta una galería de salidas gráficas del código.

Abstract In this work we present a new scheme for Hermite interpolation of a given set of planar points with a conic spline curve. The conic spline is represented as a piecewise rational quadratic Bezier curve, which depends on local tension parameters in order to control the shape of each section. We define a *fairness* functional family as the set of linear combinations of the arc length and the bending energy of the conic section. The values of the tension parameters minimizing this energy functional determine the *fairest* curve. Applying a subdivision scheme for conic splines introduced in [5], we obtain good approximations of the functional and its derivative, which are used for an efficient numerical computation of its minimum value. If we choose an specific linear combination of the functional we can show that the proposed Hermite conic spline scheme is invariant to rigid chages of coordinates and uniform scalings, reproduces arcs of circles and satisfies the *fairness* requirements listed in [7]. The *fair* Hermite conic spline scheme has been implemented in MatLab and a gallery of results is shown.

Palabras Clave
Fairness, G^1-continuidad, sección cónica, Spline racional de Bézier, Interpolación de Hermite.
MSC: 41A15, 97N50, 65D05

[1] *Departamento de Matemática, Instituto de Cibernética, Matemática y Física, La Habana, Cuba, dayron@icimaf.cu, jestrada@icimaf.cu*
[2] *Departamento de Matemática, Universidad de la Habana, La Habana, Cuba, sofia@matcom.uh.cu, wilfre@matcom.uh.cu*
***Autor para Correspondencia**

Introducción

Una técnica muy popular para los diseñadores de curvas es construir un spline que interpole una secuencia de puntos sobre el plano. Pero la interpolación no solamente se limita a éstos, sino que también se puede especificar que sea tangente a ciertas direcciones en cada uno de los puntos (interpolación de Hermite). Existen infinitas curvas que satisfacen estas condiciones de interpolación, por lo que es de esperar que el diseñador seleccione la que mejor se ajusta a los datos. En este sentido se introduce la noción de lo que es una curva *fair*.

Según el cálculo variacional [4], considerando a las componentes de la parametrización de la curva $\mathbf{C}(t) = (x(t), y(t))$ como funciones suaves de t, si el funcional de *fairness* incluye a las derivadas de orden n de $x(t)$ y $y(t)$ respecto a t, entonces la curva $\mathbf{C}(t)$ que minimiza al funcional es solución de un sistema de ecuaciones diferenciales de Euler de orden $2n$. La solución numérica de este sistema de ecuaciones diferenciales puede ser computacionalmente costosa y desde el punto de vista teórico es difícil garantizar que la curva óptima sea acotada, conexa y no singular en la región de interés. Por este motivo nos restringiremos a buscar el mínimo del funcional de *fairness* en un espacio de funciones G^1-continuas de dimensión finita que satisfagan de forma natural la interpolación de Hermite y cuya graficación sea poco costosa: las curvas spline racionales cuadráticas de Bézier.

Dado un conjunto ordenado de puntos del plano y vectores asociados a éstos, nos proponemos construir un spline racional cuadrático de Bézier tal que los interpole y que la tangente del spline en esos puntos tenga la misma dirección que el vector

que se le asocia. Esta la curva spline debe ser G^1-continua y *fair*. Se estudia un funcional de *fairness* definido en términos de la energía elástica y longitud de arco de una sección del spline y se demuestran algunas de sus propiedades, como son su invarianza a transformaciones rígidas de coordenadas y homotecias y la reproducción de arcos de circunferencias.

1. Algunos Resultados Relativos a las Curvas de Bézier

Un spline de interpolación está constituido por secciones de curvas de bajo grado que interpolan un conjunto de puntos ordenados del plano. El hecho de que las secciones interpolantes no tengan un grado tan elevado garantiza que no presente oscilaciones indeseadas. Se dice que un spline de interpolación es de Hermite si además de interpolar un conjunto de puntos del plano, también lo hace para ciertas direcciones tangentes sobre cada uno de los puntos. En este trabajo cada uno de éstos segmentos es una curva racional cuadrática de Bézier, las cuales por sus propiedades, son adecuadas para la interpolación.

En esta sección introduciremos algunas definiciones y resultados básicos de las curvas de Bézier tomados de [5].

1.1 Polinomios de Bernstein

Definición Los polinomios de Bernstein de grado n son de la forma

$$B_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}, \qquad i = 0, ..., n.$$

Una propiedad de los polinomios de Bernstein es que satisfacen la siguiente fórmula de recursión

$$B_i^n(t) = (1-t)B_i^{n-1} + tB_{i-1}^{n-1}(t).$$

Otra propiedad importante es que forman una *partición de la unidad*, dado que

$$\sum_{j=0}^{n} B_j^n(t) = \sum_{j=0}^{n} \binom{n}{i} t^j (1-t)^{n-j} = [t + (1-t)]^n = 1.$$

1.2 Curvas de Bézier

De acuerdo a lo anterior, los polinomios de Bernstein B_i^n de grado n forman una base del espacio vectorial de los polinomios de grado menor o igual que n. De este modo, toda curva polinómica $\mathbf{C}(t)$ de grado menor o igual que n posee una **representación de Bézier** única

$$\mathbf{C}(t) = \sum_{i=0}^{n} \mathbf{b}_i B_i^n(t), \tag{1}$$

donde los coeficientes $\{\mathbf{b}_i\}_{i=0}^n \subset \mathbb{R}^m, m = 2, 3, ...$ son llamados puntos de control (o puntos de Bézier). Si los puntos $\{\mathbf{b}_i\}_{i=0}^n$, forman un polígono convexo (uniendo los vértices en el orden dado), entonces la curva polinómica en la forma de Bernstein-Bézier (1), se obtiene como una combinación convexa (baricéntrica) de los puntos de control para cada $t \in [0,1]$ y está contenida en la envoltura convexa del polígono de control.

Las curvas en la forma Bernstein-Bézier ofrecen la ventaja de poder manipular su geometría a través de los puntos de control. Estas curvas tienen una generalización al caso racional ofreciendo más flexibilidad a sus propiedades geométricas y aplicaciones.

Definición Una curva racional de Bézier de grado n es una curva paramétrica descrita por los puntos de control $\{\mathbf{b}_i\}_{i=0}^n$, los pesos $\{\omega\}_{i=0}^n$ y el parámetro $t \in [0,1]$, con la forma

$$\mathbf{C}(t) = \frac{\sum_{i=0}^n \omega_i \mathbf{b}_i B_i^n(t)}{\sum_{i=0}^n \omega_i B_i^n(t)}. \tag{2}$$

Algunas de las propiedades de las curvas racionales de Bézier son la *invariancia afín*, la *invariancia bajo transformaciones paramétricas afines*, la *propiedad de la envoltura convexa* y la *interpolación de los puntos extremos y las aristas incidentes del polígono de control*.

1.3 Secciones Cónicas

Las secciones cónicas (de forma abreviada: cónicas) han recibido la mayor atención a lo largo de los siglos. A continuación mostraremos algunos de los conceptos básicos relacionados con estas curvas, que serán parte del objeto de estudio de este trabajo. Para las secciones cónicas usaremos la siguiente definición tomada de [6]:

Definición Una sección cónica en $\mathbb{R}^2$ es la proyección de una parábola en $\mathbb{R}^3$ sobre un plano.

Por esto es natural ver las cónicas como curvas racionales en el plano. En particular, su representación en la forma de Bernstein-Bézier (2) es

$$\mathbf{C}(t) = \frac{\sum_{i=0}^2 \omega_i \mathbf{b}_i B_i^2(t)}{\sum_{i=0}^2 \omega_i B_i^2(t)}.$$

Llamamos a los puntos $\mathbf{b}_i$ *puntos de control* de la cónica $\mathbf{C}$, al polígono que se construye uniendo dos puntos de control consecutivos *polígono de control* y a los parámetros de tensión ω_i se les denominan *pesos* correspondientes a los vértices del polígono de control. Esta curva puede ser parametrizada a la *forma estándar* de manera que $\omega_0 = \omega_2 = 1$ y $\omega_1 = \omega$, por tanto la curva se expresa como

$$\mathbf{C}(t) = \frac{\mathbf{b}_0 B_0^2(t) + \omega \mathbf{b}_1 B_1^2(t) + \mathbf{b}_2 B_2^2(t)}{B_0^2(t) + \omega B_1^2(t) + B_2^2(t)}, \tag{3}$$

donde el parámetro ω controla la forma de la curva de manera monótona, permitiendo la siguiente clasificación: Si $0 < \omega < 1$ se tiene una elipse, con el círculo como caso particular. Si $\omega = 1$ se tiene una parábola. Si $\omega > 1$ se tiene una hipérbola.

Aunque tiene sentido hablar de $\omega < 0$, a partir de ahora se considerará $\omega \geq 0$ a menos que se indique lo contrario; de este modo se garantiza que se cumpla la propiedad de envoltura convexa, la cual será de especial importancia de ahora en adelante.

Como las cónicas son curvas racionales de grado 2, se necesitan 3 puntos de control. Al triángulo cuyos vértices son tales puntos se le llama *triángulo de control*. La cónica correspondiente a (3), interpola a los vértices $\mathbf{b}_0$ y $\mathbf{b}_2$ y es tangente a los lados que unen dichos vértices con $\mathbf{b}_1$.

A veces es más útil conocer la ecuación implícita de una curva en vez de la expresión paramétrica, en particular, como un modo de comprobar si un punto se encuentra sobre la curva o no. Toda cónica $\mathbf{C}(t)$ tiene una representación implícita de la forma: $f(x,y) = 0$, donde f es un polinomio cuadrático de x y y. Haciendo uso de las coordenadas baricéntricas podemos plantear el siguiente teorema.

Teorema Sean $(u, v, 1-u-v)$ las coordenadas baricéntricas de un punto perteneciente a la cónica con parámetro ω correspondiente al triángulo de control, se cumple entonces que u y v satisfacen la siguiente ecuación implícita

$$v^2 - 4\omega^2 u(1-u-v) = 0. \tag{4}$$

Tomando la representación de la cónica en la forma de Bernstein-Bézier (3), se denomina *shoulder point* al punto $\mathbf{S} = \mathbf{C}(\frac{1}{2})$, cumpliéndose además que es la intersección de la recta que une al vértice $\mathbf{b}_1$ del triángulo de control con el punto medio de la arista que une $\mathbf{b}_0$ con $\mathbf{b}_2$. Este punto de la cónica juega un papel fundamental en la regla de subdivisión que veremos a continuación.

2. Regla de Subdivisión

Dado un conjunto de puntos $\{\mathbf{Q}_i, i = 1,..,m\}$ y vectores asociados a éstos $\{\vec{\mathbf{v}}_i, i = 1,..,m\}$, si calculamos los puntos $\mathbf{M}_i$ de intersección de la recta que pasa por $\mathbf{Q}_i$ con tangente $\vec{\mathbf{v}}_i$ con la recta que pasa con por $\mathbf{Q}_{i+1}$ con tangente $\vec{\mathbf{v}}_{i+1}$, entonces podemos construir un polígono inicial de subdivisión $\mathbf{P}^0$ con vértices $\{\mathbf{P}^0_i, i = 1,..,2m-1\}$ definidos como $\mathbf{P}^0_{2i-1} = \mathbf{Q}_i$ y $\mathbf{P}^0_{2i} = \mathbf{M}_i$. A partir de este polígono inicial de control se construyen recursivamente refinamientos $\mathbf{P}^j$, que son el resultado de aplicar j-veces una regla de subdivisión a $\mathbf{P}^0$.

Este proceso de refinamiento de $\mathbf{P}^j$ consiste en lo siguiente. Dados tres puntos consecutivos de $\mathbf{P}^j$, $\mathbf{P}^j_k$, $\mathbf{P}^j_{k+1}$ y $\mathbf{P}^j_{k+2}$, con k impar, en $\mathbf{P}^{j+1}$ se conservan el primero y el último ($\mathbf{P}^j_k$ y $\mathbf{P}^j_{k+2}$) y el punto intermedio, $\mathbf{P}^j_{k+1}$, se sustituye por tres nuevos puntos, que son ciertos puntos interiores de los segmentos $\mathbf{P}^j_k\mathbf{P}^j_{k+1}$ y $\mathbf{P}^j_{k+1}\mathbf{P}^j_{k+2}$ y el sholuder point del arco de cónica con parámetro de tensión ω^j_i que interpola a $\mathbf{P}^j_{k+1}$ y es tangente en $\mathbf{P}^j_{k+1}$ al segmento $\mathbf{P}^j_k\mathbf{P}^j_{k+1}$ e interpola a $\mathbf{P}^j_{k+2}$ y es tangente en $\mathbf{P}^j_{k+2}$ al segmento $\mathbf{P}^j_{k+1}\mathbf{P}^j_{k+2}$. Este arco de cónica queda dividido en 2 subarcos de cónica racional de Bézier, con triángulos de control determinados por los vértices de $\mathbf{P}^{j+1}$ y sus parámetros de tensión se calculan a partir de ω^j_i.

Los detalles del esquema de subdivisión se dan a continuación. La demostración de los resultados de esta sección pueden verse en la tesis de R. Díaz [5].

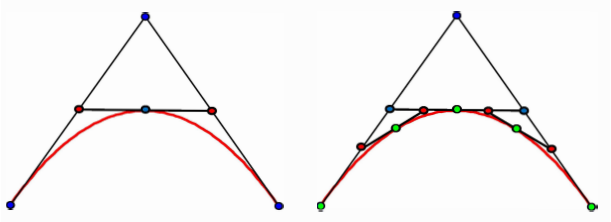


Figura 1. Proceso de Subdivisión.

En efecto, dados 3 puntos consecutivos de $\mathbf{P}^j$, $\mathbf{P}^j_i$, $\mathbf{P}^j_{i+1}$, $\mathbf{P}^j_{i+2}$, con i impar, y el parámetro ω^j_i del arco de cónica asociado, se calculan a partir de éstos 5 puntos del polígono refinado $\mathbf{P}^{j+1}_{2i-1}$,..,$\mathbf{P}^{j+1}_{2i+3}$ como sigue.

$$\mathbf{P}^{j+1}_{2i-1} = \mathbf{P}^j_i, \tag{5}$$

$$\mathbf{P}^{j+1}_{2i+3} = \mathbf{P}^j_{i+2}. \tag{6}$$

Tomamos los puntos $\mathbf{P}^j_i, \mathbf{P}^j_{i+1}$ de la arista i y los puntos $\mathbf{P}^j_{i+1}, \mathbf{P}^j_{i+2}$ de la arista $i+1$ para calcular nuevos puntos sobre ellas

$$\mathbf{P}^{j+1}_{2i} = \frac{\mathbf{P}^j_i + \omega^j_i \mathbf{P}^j_{i+1}}{1+\omega^j_i}, \tag{7}$$

$$\mathbf{P}^{j+1}_{2i+2} = \frac{\mathbf{P}^j_{i+2} + \omega^j_i \mathbf{P}^j_{i+1}}{1+\omega^j_i}. \tag{8}$$

Como ya habíamos visto anteriormente, el shoulder point puede ser calculado mediante

$$\mathbf{S}^j_i = \mathbf{S}^{j+1}_{2i} = \mathbf{P}^{j+1}_{2i+1} = \frac{\mathbf{P}^{j+1}_{2i} + \mathbf{P}^{j+1}_{2i+2}}{2}. \tag{9}$$

Si sustituimos las ecuaciones (7) y (8) en (9) se tiene finalmente que

$$\mathbf{P}^{j+1}_{2i+1} = \frac{\mathbf{P}^j_i + 2\omega^j_i \mathbf{P}^j_{i+1} + \mathbf{P}^j_{i+2}}{2(1+\omega^j_i)}. \tag{10}$$

En cada iteración, los puntos de la poligonal de control con subíndice impar pertenecen a la curva. A medida que el algoritmo realiza un mayor número de iteraciones se genera una mayor cantidad de puntos sobre la curva.

Teorema Sea ω_i^j el parámetro de tensión asociado a la cónica i-ésima en el paso j-ésimo y sean ω_{2i-1}^j y ω_{2i}^j los parámetros asociados a los subarcos de esta cónica, entonces se cumple que

$$\omega_{2i-1}^{j+1} = \omega_{2i}^{j+1} = \sqrt{\frac{1+\omega_i^j}{2}}. \tag{11}$$

Está claro que los puntos con subíndice impar insertados en cada paso pertenecen a un arco de cónica racional de Bézier que corresponde a una curva spline que interpola los puntos dados como datos y las tangentes asociadas a cada uno de ellos, pero no queda claro que la sucesión de puntos que se va generando cubre todo este arco de cónica y no se acumulan alrededor de un número finito de puntos sobre la cónica. Se puede demostrar que la sucesión del máximo de las normas de las diferencias entre dos puntos consecutivos en cada refinamiento tiende a cero, véase [5].

3. Estudio del *Fairness*

Por *fair* se entiende, de forma intuitiva, una curva cuya gráfica de curvatura es continua, con muy pocas oscilaciones y con valores extremos no muy grandes. Tal *definición*, a pesar de ser subjetiva, es sin embargo muy práctica, pues una curva *fair* es, estéticamente, lo que más desea un diseñador. El ploteo de curvatura será usado por un diseñador de experiencia como una herramienta cotidiana e imprescindible pues, en general, de todas las posibles curvas que puedan utilizarse para interpolar un conjunto de datos, suele ser la curva *fair* quien mejor lo hace.

Como requisitos importantes que debe satisfacer una curva *fair* podemos citar los siguientes (que fueron tomados de [7]):

1. **Extensionalidad**, entendida a partir de que si se adicionan como nuevos datos puntos que están sobre el spline original, el nuevo spline no varía significativamente.

2. **Redondez**, entendida como reproducción de arcos de círculo.

3. **Alto orden de continuidad**.

4. **Curvatura monótona**.

3.1 Un cambio de coordenadas adecuado

En esta sección introduciremos un cambio de coordenadas que posibilita una representación más sencilla de la parametrización de una sección cónica en la base de Bernstein-Bézier.

Los segmentos del spline se describen mediante la fórmula paramétrica dada por (3). El segmento de cónica determinado por los puntos $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2$ puede ser descrito por la parametrización

$$\begin{aligned} x(t) &= \frac{P_{0,x}B_0^2(t) + \omega P_{1,x}B_1^2(t) + P_{2,x}B_2^2(t)}{B_0^2(t) + \omega B_1^2(t) + B_2^2(t)}, \\ y(t) &= \frac{P_{0,y}B_0^2(t) + \omega P_{1,y}B_1^2(t) + P_{2,y}B_2^2(t)}{B_0^2(t) + \omega B_1^2(t) + B_2^2(t)}. \end{aligned}$$

donde $\mathbf{P}_i = (P_{i,x}, P_{i,y}), i = 0,1,2$.

Realizando una adecuada traslación y rotación de los ejes coordenados de modo que el segmento $\overline{\mathbf{P}_0\mathbf{P}_2}$ esté incluido en el eje de las abscisas (ver Figura 2) se logra una parametrización más sencilla de la sección cónica interior al triángulo de control con vértices $\mathbf{P}_0, \mathbf{P}_1$ y $\mathbf{P}2$.

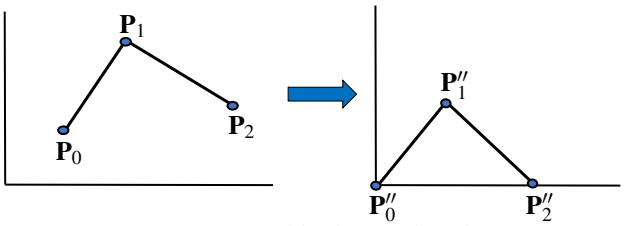


Figura 2. Cambio de coordenadas

La parametrización de la curva con respecto a las nuevas coordenadas queda de la siguiente manera

$$x(t) = \frac{2a(1-t)t\omega + Lt^2}{(1-t)^2 + 2(1-t)t\omega + t^2}, \tag{12}$$

$$y(t) = \frac{2b(1-t)t\omega}{(1-t)^2 + 2(1-t)t\omega + t^2}, \tag{13}$$

donde L es la longitud del segmento $\overline{\mathbf{P}_0\mathbf{P}_2}$ y (a,b) son las nuevas coordenadas del segundo vértice del triángulo, $\mathbf{P}_1{}''$.

Con esta nueva parametrización de las curvas racionales de Bézier debemos dejar claro en qué dominio o espacio de parámetros están ellas definidas. Puesto que la parametrización depende de a, b, L, ω y t, definimos entonces el siguiente espacio de parámetros.

Definición Se define como el espacio de parámetros $\mathbb{E} \subset \mathbb{R}^5$ a los puntos que cumplen las siguientes condiciones

$$\begin{aligned} \mathbb{E} = \{&(a,b,L,\omega,t) \in \mathbb{R}^5 \text{ tales que} \\ &a > 0, b > 0, L > 0, \omega \geq 0, 0 \leq t \leq 1\}. \end{aligned} \tag{14}$$

El motivo por el que debemos especificar qué valores puede tomar cada parámetro de la parametrización dada por (12) y (13) es que para valores negativos de a, b y L el triángulo de control es degenerado, y las cónicas de Bézier están definidas para valores de $\omega \geq 0$ y $t \in [0,1]$ como se mostró en la sección 1.2.

3.2 Longitud de arco de una cónica

La longitud de arco de una curva paramétrica $C(t)$ se puede calcular por

$$S = \int_0^1 ||C'(t)||dt, \tag{15}$$

donde sabemos que $||C'(t)|| = \sqrt{(x'(t))^2 + (y'(t))^2}$ y las expresiones de $x'(t)$ y $y'(t)$ se obtienen de derivar (12) y (13) respecto al parámetro t.

Dados tres puntos del plano $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2$, éstos definen tres valores para los parámetros a, b y L de la parametrización de la curva dada por las ecuaciones (12) y (13). Si consideramos fijos los parámetros a, b y L entonces de (12) y (13) se obtiene la siguiente expresión para la longitud de arco.

$$S(\omega) = \int_0^1 \frac{2\sqrt{b^2\omega^2(2t-1)^2 + (-2at\omega + a\omega + Lt - Lt^2 + Lt^2\omega)^2}}{(-1+2t-2t^2-2t\omega+2t^2\omega)^2}dt \tag{16}$$

donde los parámetros a, b y L determinan la geometría del triángulo de control.

El siguiente lema argumenta la continuidad de la función $S(\omega)$:

Lema Restringiendo los parámetros al espacio $\mathbb{E}$, la función $S(\omega)$ en (16) y su derivada respecto a ω son funciones continuas de ω.

Demostración

Es suficiente hallar los valores de los parámetros que anulan el denominador del integrando de (16) y comprobar que dichos valores no están en $\mathbb{E}$.

En efecto, el denominador del integrando en (16) se anula para $\pm t = \frac{-1+\omega+\sqrt{-1+\omega^2}}{2(-1+\omega)}$. Es por tanto inmediato verificar que si $0 \leq \omega < 1$ o $\omega > 1$, los ceros del denominador del integrando no pertenecen al intervalo $[0,1]$ previsto para $t \in \mathbb{E}$. Por otra parte, si $\omega = 1$, el denominador es idénticamente igual a 1 y no se anula. ■

Cuando $\omega \to 0$, la cónica que describe este parámetro de tensión se aproxima al segmento de recta que une a los puntos de interpolación $\mathbf{P}_0$ y $\mathbf{P}_2$, es decir

$$\lim_{\omega \to 0} S(\omega) = ||\mathbf{P}_0 - \mathbf{P}_2||.$$

Si el parámetro de tensión toma valores muy elevados, el arco de curva (que es convexo e interpola a $\mathbf{P}_0$ y $\mathbf{P}_2$) tiende a acercarse al segundo vértice del triángulo $\mathbf{P}_1$, lo que trae como consecuencia que la longitud de arco se aproxima a la suma de las longitudes de los lados del triángulo $\overline{\mathbf{P}_0\mathbf{P}_1}$ y $\overline{\mathbf{P}_1\mathbf{P}_2}$, esto es

$$\lim_{\omega \to \infty} S(\omega) = ||\mathbf{P}_0 - \mathbf{P}_1|| + ||\mathbf{P}_1 - \mathbf{P}_2||.$$

De lo anterior se infiere que la longitud de arco de una cónica es una función acotada del parámetro ω.

3.3 Energía Elástica de una Cónica

La energía elástica o energía de deformación es un concepto que viene de la Física y se define como el aumento de energía interna acumulada en el interior de un sólido deformable como resultado del trabajo realizado por las fuerzas que provocan la deformación. El valor de la energía elástica de una curva se define como el valor de la siguiente integral:

$$E = \int_0^l k^2(s)ds,$$

donde l es la longitud de la curva y $k(s)$ es la curvatura del punto que describe una longitud de arco s. Si hacemos un cambio de parámetros podemos decir que

$$E = \int_0^l k^2(s)ds = \int_0^1 \left(k^2(t)\frac{ds}{dt}(t)\right)dt.$$

Sabemos que dentro de una misma familia de cónicas de Bézier, cada miembro de la familia está caracterizado por un determinado valor del parámetro ω. Si procedemos igual que en la sección 3.2, fijando los valores de a, b y L, entonces la energía elástica de una cónica puede representarse por la siguiente función

$$E(\omega) = \int_0^1 \frac{b^2\omega^2L^2(-1+2t-2t^2-2t\omega+2t^2\omega)^4}{2\left(b^2\omega^2(2t-1)^2 + (-2at\omega+a\omega+Lt-Lt^2+Lt^2\omega)^2\right)^{5/2}}dt. \tag{17}$$

Lema Restringiendo los parámetros al espacio $\mathbb{E}$, se cumple que

- **a)** la función $E(\omega)$ en (17) y su derivada respecto a ω son funciones continuas de ω,
- **b)** si $\omega \to 0$ entonces $E(\omega) \to \infty$,
- **c)** si $\omega \to \infty$, entonces $E(\omega) \to \infty$.

Demostración

a) El denominador del integrando en (17) es igual a una constante multiplicada por la raíz de una suma de cuadrados, que no se anulan simultáneamente si $t, \omega, L \in \mathbb{E}$.

b) Para t y ω muy pequeños, la derivada del integrando de (17) respecto a t y evaluada en $t = 0$ es igual a $-\frac{5b^2L^3a}{2\omega^4} + O\left(\omega^{-3}\right)$. Respectivamente, la derivada del integrando de (17) respecto a ω y evaluada en $t = 0$ igual a $-\frac{3b^2L^2}{2\omega^4\left(a^2+b^2\right)^{5/2}}$. Por lo tanto, ambas derivadas son negativas y existe entonces una vecindad V de $(t,\omega) = (0,0)$ para la que el integrando de (17) es estrictamente decreciente. Si $\delta > 0$ es suficientemente pequeño, podemos suponer que $\{(t,\omega)/\,\text{máx}(t,\omega) \leq \delta\} \subset V$ y además se cumple que la evaluación de este integrando en $(t,\omega) = (\delta,\delta)$ es igual a $\frac{b^2L^2}{2\left(2aL+b^2+L^2+a^2\right)^{5/2}\delta^3} + O\left(\delta^{-2}\right)$. Debido a la monotonía del integrando restringido a V resulta que la integral para el subintervalo $t \in [0,\delta]$ es mayor que $O\left(\delta^{-2}\right)$

Puesto que el integrando es no negativo, la integral para $t \in [0,1]$ en (17) es mayor a la integral para el subintervalo $t \in [0,\delta]$, consecuentemente, la integral en (17) es estrictamente mayor que $O\left(\delta^{-2}\right)$. Si δ (y consecuentemente ω) tiende a 0, esta última integral tiende a ∞ .

c) Si ω es muy grande, el integrando de (17) es igual a $\frac{8b^2L^2t^4(-1+t)^4\omega}{\left(b^2(2t-1)^2+\left(-2at+a+Lt^2\right)^2\right)^{5/2}}+O(1)$, o sea, es $O(\omega)$, para $\delta > 0$ suficientemente pequeño y $t \in [\delta, 1-\delta]$. Por lo tanto, la integral en (17) restringida al subintervalo $t \in [\delta, 1-\delta]$ tiende a ∞ si ω tiende a ∞. ■

3.4 Funcional de fairness

Existen diferentes funcionales que nos brindan una idea del *fairness* de una curva. Entre los más empleados se encuentran el de minimizar la energía elástica y la longitud de arco de una curva. El funcional de *fairess* más comúnmente empleado en la literatura es una combinación lineal de la energía elástica y de la longitud de arco (véase [2], [9], [10], [11]).

$$\begin{aligned} F_\lambda(\omega) &= E(\omega)+\lambda S(\omega) \qquad (18)\\ &= \int_0^1 \left(k^2(\omega,t)\frac{ds}{dt}(\omega,t)\right)dt+\lambda\int_0^1 \frac{ds}{dt}(\omega,t)dt. \end{aligned}$$

Este funcional describe la energía elástica que almacena una curva racional cuadrática de Bézier más la longitud de arco multiplicada por un cierto $\lambda \geq 0$, donde el usuario decide su valor en dependencia a cuál de las dos energías le quiere dar mayor peso.

Como ya se ha demostrado en los **lemas** de las secciones 3.2 y 3.3 tanto el funcional $F_\lambda(\omega)$ como su derivada respecto al parámetro ω son continuos en el espacio de parámetros $\mathbb{E}$. Otra propiedad demostrada en la sección 3.3, es que toma valores muy elevados cuando $\omega \to 0$ y $\omega \to \infty$.

La derivada de $F_\lambda(\omega)$ respecto a ω , $\frac{\partial F_\lambda(\omega)}{\partial \omega}$, es igual a

$$F'_\lambda(\omega)=\int_0^1 \frac{\partial\left(k^2(\omega,t)\frac{ds}{dt}(\omega,t)\right)}{\partial\omega}dt+\lambda\int_0^1\frac{\partial\frac{ds}{dt}(\omega,t)}{\partial\omega}dt. \quad (19)$$

Se calcularon expresiones explícitas para los integrandos en la formula anterior, que no se incluyen por su complejidad.

3.4.1 Existencia de mínimo

El valor del parámetro ω que describe la curva *fair* como habíamos definido desde el inicio, es el siguiente:

$$\arg\{\min_{\omega} F_\lambda(\omega)\}.$$

El siguiente lema demuestra que el funcional de *fairness* alcanza su valor mínimo en el espacio de parámetros $\mathbb{E}$.

Lema Si en el espacio de parámetros $\mathbb{E}$ establecemos valores fijos de a, b y L, y además fijamos el valor de λ en la combinación lineal del funcional (18), con $\lambda \in [0,+\infty)$; entonces el funcional $F_\lambda(\omega)$ alcanza su mínimo en el intervalo $(0,+\infty)$ para ω.

Demostración

El funcional de *fairness* toma valores tan elevados como uno quiera cuando $\omega \to 0$ y $\omega \to \infty$. Lo anterior permite afirmar que, para al menos dos valores diferentes de ω, el funcional alcanza un mismo valor. Supongamos que ello ocurre para $\omega = \varepsilon$ y $\omega = \xi$, o sea, $F_\lambda(\varepsilon) = F_\lambda(\xi)$.

Asumiendo $\varepsilon < \xi$ y restringiendo ahora el funcional al intervalo $[\varepsilon,\xi]$ en el que ya se había demostrado su continuidad entonces, por el teorema de Weierstrass, se puede afirmar que los valores extremos en este intervalo se alcanzan. En nuestro caso en particular, nos interesa su valor mínimo.

Sabemos por el teorema de Fermat (ver [8]) que si un extremo se alcanza en un punto interior de un intervalo donde la función es derivable, la derivada en dicho extremo se anula. Dado que ya se probó que el funcional es derivable en el intervalo $[\varepsilon,\xi]$ y además $F_\lambda(\varepsilon) = F_\lambda(\xi)$ entonces por el teorema de Rolle (ver [8]) la derivada de la función se anula al menos una vez en el intervalo y, el mínimo buscado se halla, por tanto, entre los puntos que anulan la derivada del funcional. ■

El lema anterior solo establece la existencia de un mínimo del funcional (18) en el intervalo $(0,+\infty)$ para ω, pero no establece su unicidad. Demostrar la unicidad directamente a partir del análisis de la segunda derivada del funcional resulta ser muy engorroso. Sin embargo, en todos los ejemplos numéricos calculados, este funcional resulta ser convexo. Por lo tanto, declaramos una sección de cónica como *fair*, si su parámetro ω es un mínimo local del funcional (18) y por extensión, un spline cónico es *fair*, si todas sus secciones lo son.

3.5 Aproximaciones del funcional de fairness

El proceso de subdivisión (sección 2) genera una secuencia de poligonales convergentes a la curva $\{\mathbf{P}^0,\mathbf{P}^1,...,\mathbf{P}^j,...\}$. Tomamos entonces la suma de las longitudes de los segmentos que componen la poligonal $\mathbf{P}^j$ como aproximaciones de la longitud de arco. Se tiene entonces la aproximación

$$S(\omega)=\int_0^1 ||C'(\omega,t)||dt \approx \sum_{i=1}^{n}||\mathbf{P}_i^j-\mathbf{P}_{i-1}^j|| = \sum_{i=1}^{n} l_i, \quad (20)$$

donde $l_i = ||\mathbf{P}_i^j - \mathbf{P}_{i-1}^j||$ es la distancia entre los puntos $\mathbf{P}_i^j$ y $\mathbf{P}_{i-1}^j$, por tanto, la longitud del segmento i-ésimo de la poligonal de aproximación $\mathbf{P}^j$.

La integral que define la energía elástica puede ser reducida a una integral elíptica, pero el proceso de su reducción a la forma normal de Legendre resulta ser complicado e

inestable numéricamente, por lo que utilizamos el hecho de que podemos generar una muestra suficientemente grande de puntos sobre la curva y, con poco costo adicional, aproximar las integrales utilizando un método de integración numérica como el método de los trapecios.

Usando el cambio de coordenadas dado en la sección 3.1 para la i-ésima sección del spline, si tenemos una muestra de puntos $\{\mathbf{P}_0,...,\mathbf{P}_n\}$ sobre la i-ésimasección del spline con coordenadas $\mathbf{P}_k = (x_k, y_k), k = 0,..,n$, a cada punto $\mathbf{P}_k$ se le hace corresponder un único valor del parámetro t_k en la parametrización de la curva, dado por la fórmula de inversión

$$t_k = \frac{2\omega_i(bx_k - ay_k)}{2\omega_i(bx_k - ay_k) + Ly_k}. \tag{21}$$

Una vez calculado el valor t_k es posible calcular el valor de curvatura en el punto $\mathbf{P}_k$ utilizando la parametrización en (12) y (13):

$$k(t_k) = \frac{|x'(t_k)y''(t_k) - y'(t_k)x''(t_k)|}{((x'(t_k))^2 + (y'(t_k))^2)^{3/2}}. \tag{22}$$

Es posible entonces obtener buenas aproximaciones del valor de la función de la energía elástica definida basándonos en la muestra de puntos generada sobre la i-ésima sección de la curva spline. Como ya habíamos visto en la sección 2, la muestra de puntos sobre la curva que se genera en el proceso de subdivisión no se acumulan alrededor de un número finito de ellos y a mayor cantidad de iteraciones que se realicen, menor será la distancia entre cada par de puntos consecutivos. Esto garantiza que la longitud de los intervalos de la partición del intervalo de integración $[0,1]$ sea tan pequeña como se desee, permitiendo que el método de los trapecios aporte una buena aproximación del funcional.

Es posible obtener también aproximaciones para la derivada de $F_\lambda(\omega)$ respecto a ω , $\frac{\partial F_\lambda(\omega)}{\partial \omega}$, calculando aproximadamente las integrales en (19).

3.5.1 Hallar el valor mínimo

Ya tenemos entonces tanto aproximaciones del funcional y de su derivada respecto a ω, nos resta entonces hallar el valor del parámetro ω que minimiza el valor de este funcional. Como ya habíamos planteado en la sección 3.4.1 el funcional alcanza su valor mínimo en el espacio de parámetros en el que está definido. Hasta ahora tenemos aproximaciones tanto del funcional de *fairness* como de su derivada, por lo que resta aplicar un algoritmo numérico para hallarlo. En este trabajo empleamos dos métodos para hallarlo.

El primero de los métodos es el *Método de Newton - Raphson* (véase [3]). En la sección 3.4.1 vimos que la derivada del funcional se anula al menos una vez en el intervalo donde el parámetro de tensión ω toma sus valores, por lo que la ecuación $F'_\lambda(\omega) = 0$ tiene solución donde $F_\lambda(\omega)$ es el funcional de *fairness*. Sin embargo, chequear las condiciones de convergencia de éste método es tan complicado como resolver analíticamente la ecuación. Debemos destacar que la expresión analítica de la derivada del funcional de *fairness* $F'_\lambda(\omega)$ es mucho más complicada que la del propio funcional, no obstante el algoritmo desarrollado evalúa directamente su expresión. La expresión analítica de la segunda derivada del funcional $F''_\lambda(\omega)$, resulta tan complicada que es más eficiente obtener aproximaciones de sus valores por medio de la diferenciación numérica de la primera derivada

$$F''_\lambda(x) = \lim_{h\to 0} \frac{F'_\lambda(x+h) - F'_\lambda(x)}{h}.$$

El segundo método es conocido en la literatura como *Método de la Sección de Oro* o *Método de la Sección Áurea* (Ver [1]). Este método básicamente va reduciendo el intervalo de incertidumbre donde se encuentra el valor mínimo de la función hasta obtener un intervalo donde el error de la aproximación no sea mayor que un cierto valor dado por el usuario. Usualmente el método brinda buenas aproximaciones en el caso de que la función objetivo sea convexa, sin embargo, este resultado no se demostró en el presente trabajo por lo complicado que resulta trabajar con el funcional de *fairness*.

No obstante, ambos métodos brindan igual resultado para los mismos parámetros como se mostrará en la sección 4.

4. Resultados numéricos

Debemos destacar además en esta sección que pese a que no se obtuvieron soluciones exactas de las integrales que definen los funcionales de energía elástica y longitud de arco en las secciones 3.2 y 3.3, las aproximaciones numéricas realizadas muestran buenos resultados, los cuales proponemos a continuación.

La experimentación numérica realizada para que el funcional de *fairness* presente las características expuestas en el inicio de la sección 3, motivó a proponer como valor del parámetro λ en la expresión (18) el siguiente

$$\lambda = \frac{\text{sen}^2(\beta)}{\cos^2(\frac{\alpha+\gamma}{2})L^2} \tag{23}$$

donde $\alpha = \sphericalangle \mathbf{M}_i\mathbf{Q}_i\mathbf{Q}_{i+1}$, $\beta = \sphericalangle \mathbf{Q}_i\mathbf{M}_i\mathbf{Q}_{i+1}$, $\gamma = \sphericalangle \mathbf{Q}_i\mathbf{Q}_{i+1}\mathbf{M}_i$ y $L = ||\mathbf{Q}_{i+1} - \mathbf{Q}_i||$ como se muestra en la Figura 3.

Ahora bien, ¿por qué escoger el valor de λ que aparece en (23) y no otro? Como ya habíamos dicho en la sección 3.4, para cada valor de λ en la combinación lineal de la longitud de arco y energía elástica se definía un nuevo funcional de *fairness*. Precisamente al tomar este valor en la combinación lineal del funcional, este último manifiesta buenas propiedades para el diseño geométrico como son las siguientes:

- El valor de

$$\arg\{\min_{\omega} F_\lambda(\omega)\}$$

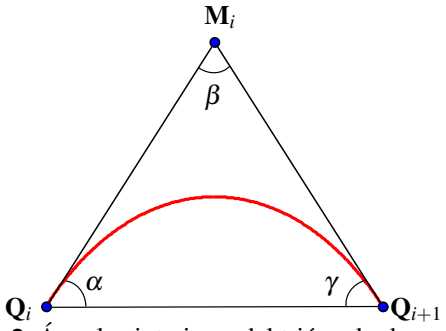


Figura 3. Ángulos interiores del triángulo de control.

es invariante bajo transformaciones rígidas de coordenadas y homotecias.

- Reproduce arcos de circunferencia (**Redondez**).

4.1 Reproducción de arcos de circunferencia

Como parte de este trabajo también se encuentra reproducir arcos de circunferencias, pues como habíamos visto al inicio de esta sección, si situamos los puntos de interpolación sobre una circunferencia entonces la curva más *fair* que los interpola es precisamente el arco de circunferencia que contiene a dichos datos. Para que la curva interpolante sea un arco de circunferencia el triángulo de control debe ser isósceles por la simetría del círculo. En dicho caso, debemos demostrar que el círculo de interpolación es único. El parámetro de tensión asociado a dicha cónica debe cumplir la siguiente condición.

Lema Denotemos por $\alpha = \sphericalangle \mathbf{M}_i\mathbf{Q}_i\mathbf{Q}_{i+1}$, $\beta = \sphericalangle \mathbf{Q}_i\mathbf{M}_i\mathbf{Q}_{i+1}$, $\gamma = \sphericalangle \mathbf{Q}_i\mathbf{Q}_{i+1}\mathbf{M}_i$ y $L = ||\mathbf{Q}_{i+1} - \mathbf{Q}_i||$ como se muestra en la Figura 3. Entonces se cumple que:

i. Los datos $\mathbf{Q}_i$, $\vec{\mathbf{v}}_i$ y $\mathbf{Q}_{i+1}$, $\vec{\mathbf{v}}_{i+1}$ provienen de un círculo si y solo si $\alpha = \gamma$; en otras palabras, si el triángulo con vértices $\mathbf{Q}_i$, $\mathbf{Q}_{i+1}$ y $\mathbf{M}_i$ es isósceles.

ii. Si los datos $\mathbf{Q}_i$, $\vec{\mathbf{v}}_i$ y $\mathbf{Q}_{i+1}$, $\vec{\mathbf{v}}_{i+1}$ provienen de un círculo entonces el círculo descrito en **i.** puede representarse como una curva cónica de Bézier (Véase la sección 1.2) tomando como triángulo de control al formado por los puntos $\mathbf{Q}_i$, $\mathbf{M}_i$ y $\mathbf{Q}_{i+1}$ y parámetro $\omega = \cos(\alpha) = \cos(\gamma)$.

La demostración de éste resultado puede verse en [6].

Un círculo completo puede obtenerse uniendo piezas de un spline cerrado, donde cada segmento sea un arco de circunferencia. Por ejemplo, podemos representar un círculo utilizando tres arcos iguales (Figura 4). Con todos los ángulos $\alpha_j = \gamma_j = 60^0$ y los pesos $\omega_j = \frac{1}{2}$ se obtiene una representación exacta de un círculo.

Partiendo del **Lema** anterior podemos argumentar las consecuencias de tomar el valor que proponemos para el parámetro λ al inicio de esta sección y que se exponen en el siguiente teorema:

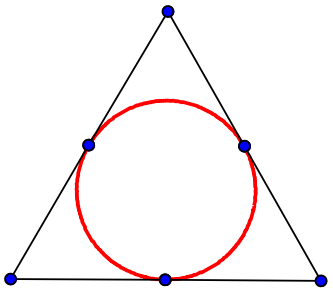


Figura 4. Reproducción de un círculo.

Teorema Sean $\mathbf{Q}_i$ y $\mathbf{Q}_{i+1}$ dos puntos con direcciones tangentes asociadas $\vec{\mathbf{v}}_i$ y $\vec{\mathbf{v}}_{i+1}$ respectivamente, denotemos por $\mathbf{M}_i$ el punto de intersección de las rectas que pasan por los puntos $\mathbf{Q}_i$ y $\mathbf{Q}_{i+1}$ con direcciones tangentes dadas por los vectores $\vec{\mathbf{v}}_i$ y $\vec{\mathbf{v}}_{i+1}$. Se define λ_C como

$$\lambda_C = \frac{\text{sen}^2(\beta)}{\cos^2(\frac{\alpha+\gamma}{2})L^2}$$

donde $L = ||\mathbf{Q}_{i+1} - \mathbf{Q}_i||$, $\alpha = \sphericalangle \mathbf{M}_i\mathbf{Q}_i\mathbf{Q}_{i+1}$, $\gamma = \sphericalangle \mathbf{M}_i\mathbf{Q}_{i+1}\mathbf{Q}_i$ y $\beta = \sphericalangle \mathbf{Q}_i\mathbf{M}_i\mathbf{Q}_{i+1}$.

Si $\mathbf{Q}_i$, $\vec{\mathbf{v}}_i$, $\mathbf{Q}_{i+1}$ y $\vec{\mathbf{v}}_{i+1}$ son datos que provienen de un círculo $\mathbf{C}_i$, entonces la cónica de Bézier que interpola a $\mathbf{Q}_i$, $\vec{\mathbf{v}}_i$, $\mathbf{Q}_{i+1}$, $\vec{\mathbf{v}}_{i+1}$ y minimiza el funcional

$$F_{\lambda_C}(\omega) = E(\omega) + \lambda_C S(\omega)$$

es el arco $\widehat{\mathbf{Q}_i\mathbf{Q}_{i+1}}$ del círculo $\mathbf{C}_i$.

Demostración

Si imponemos a los parámetros a, b, L que el triángulo sea isóceles (como corresponde a una sección de circunferencia) y que $\omega = \cos(\alpha) = \cos(\gamma)$, como se establece en el lema anterior, entonces es posible calcular exactamente las integrales que aparecen en (19) y despejar el valor λ_C de λ que anula a $\frac{\partial F_\lambda(\omega)}{\partial \omega}$. ■

Observación. Si los datos a interpolar provienen de un arco de círculo $\mathbf{C}_i$, pero el valor de λ en la combinación lineal del funcional no es igual a λ_C, entonces el mínimo del funcional $F_\lambda(\omega)$ no necesariamente reproduce a $\mathbf{C}_i$.

4.2 Implementación Numérica

El algoritmo para generar puntos sobre una curva cónica de Bézier descrito en la sección 2 fue implementado en MatLab. Ya una vez teniendo a mano un algoritmo bastante eficiente para generar puntos sobre las cónicas se implementó también en el mismo software varias funciones que tienen como objetivo calcular, de manera aproximada y como se indica en la sección 3.5, el valor del funcional de *fairness* y de su derivada con respecto al parámetro ω para una combinación de datos específica. Además se programaron

ambos métodos para hallar el valor mínimo del funcional. Algunos de los resultados obtenidos por el programa se muestran a continuación.

En lo adelante, expondremos los resultados de nuestro algoritmo para el siguiente conjunto de puntos en el plano $\mathbf{P}_0 = (0,0), \mathbf{P}_1 = (1,1)$ y $\mathbf{P}_2 = (2,0)$, tomados como triángulo de control. En la Figura 5 mostramos una comparación entre diferentes cónicas de Bézier definidas en el mismo triángulo de control en cuanto al gráfico de los valores de curvatura.

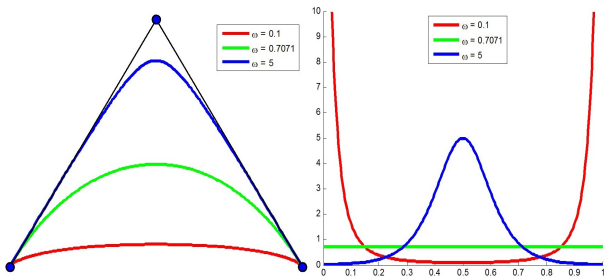


Figura 5. Izquierda: Gráfico de las cónicas de Bézier. Derecha: Gráfico de los valores de curvatura.

En el Cuadro 1 se muestran valores aproximados de la longitud de arco para distintos valores del parámetro de tensión. Con 4 iteraciones del algoritmo de subdivisión se obtienen buenas aproximaciones para $S(\omega)$. Debemos observar que cuando $\omega \to 0$ se tiene que $S(\omega) \to 2$ y cuando $\omega \to \infty$ entonces $S(\omega) \to 2\sqrt{2}$ como se ya se planteó en las secciones 3.2 y 3.3.

	$S(\omega)$		
ω	2 iteraciones	4 iteraciones	8 iteraciones
0.01	2.0002	2.0003	2.0004
0.3	2.0767	2.0875	2.0884
0.5	2.1445	2.1571	2.1580
0.7071	2.2125	2.2259	2.2268
1	2.2808	2.2947	2.2956
10	2.7119	2.7206	2.7213
50	2.8020	2.8047	2.8050
100	2.8149	2.8164	2.8166

Cuadro 1. Aproximaciones de la longitud de arco.

En el Cuadro 2 muestra que para obtener buenas aproximaciones de la energía elástica es necesario realizar una mayor cantidad de iteraciones del proceso de subdivisión. La razón por la que hay que realizar más subdivisiones está en el hecho de que como utilizamos el método de los trapecios para obtener las aproximaciones, mientras mayor sea el número de puntos generados sobre la curva, menor será el paso de la integral (ver sección 3.5) con lo que se obtienen mejores resultados.

	$E(\omega)$		
ω	4 iteraciones	8 iteraciones	12 iteraciones
0.01	3.7908	1132.3	1547.0
0.3	2.6969	2.7709	2.7712
0.5	1.8095	1.8134	1.8134
0.7071	1.6658	1.6654	1.6654
1	1.7528	1.7524	1.7524
10	9.4707	9.4450	9.4449
50	44.3944	44.4090	44.4071
100	86.9953	88.1138	88.1085

Cuadro 2. Aproximaciones de la energía elástica.

Otra observación importante que podemos realizar de los datos obtenidos es que para valores muy pequeños del parámetro de tensión ω y pocas iteraciones no se obtienen buenas aproximaciones del funcional, sin embargo, mientras mayor sea la cantidad de iteraciones en el proceso de subdivisión, mejores serán los resultados en la aproximación.

4.3 Aplicaciones

Las curvas spline actualmente en el diseño geométrico asistido por computadora juegan un papel fundamental. En esta sección se mostrarán algunos de sus usos en la vida diaria. Como motivo de visualizar una curva *fair* en el diseño geométrico se ha desarrollado una aplicación con interfaz gráfica de MatLab donde el diseñador puede introducir los puntos y tangentes de interpolación en un área destinada para el diseño.

Una aplicación inmediata de las curvas spline es el diseño de figuras planas y de carreteras. En la figura 8 se presenta un ejemplo de objeto real que fue modelado utilizando la interfaz gráfica desarrollada.

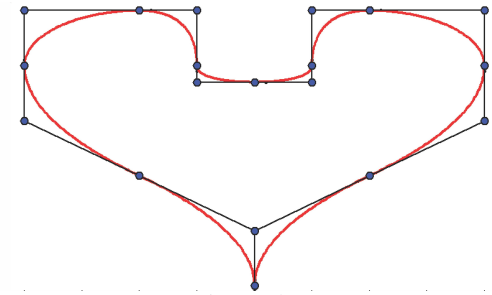


Figura 6. Diseño de un corazón.

5. Conclusiones

Se da solución al problema de encontrar un spline G^1-continuo racional cuadrático de Bézier que interpole un conjunto de puntos del plano con vectores tangentes asociados a cada uno de ellos y también cumple la propiedad de ser *fair*, ya que cada segmento que lo compone es la curva que minimiza el funcional de *fairness* que se propone.

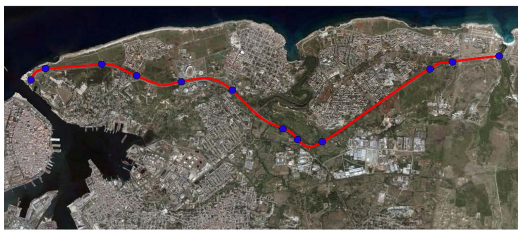


Figura 7. Diseño de una carretera.

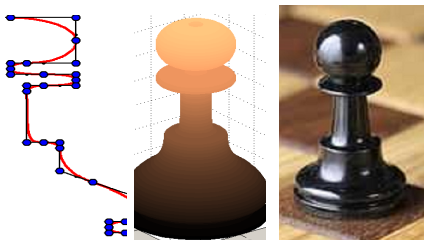


Figura 8. Modelo de una pieza de ajedrez. (a) Curva generatriz. (b) Diseño tridimensional. (c) Objeto real.

Para la minimización del funcional de *fairness* se utilizaron aproximaciones numéricas. En la experimentación numérica se muestra que se obtuvieron buenos valores de aproximación con pocas iteraciones.

Desde el punto de vista teórico se demuestra que el funcional de *fairness* tiene buenas propiedades como son su invarianza a transformaciones rígidas de coordenadas y homotecias y la reproducción de arcos de circunferencias.

Referencias

[1] Alikhani, J., Hosseini, S., M., M., Maalek, F., M., *A new optimization algorithm based on chaotic maps and golden section search method*. Engineering Applications of Artificial Intelligence, 50, 201–214, 2016.

[2] Bajaj, Chandrajit L., *Energy formulations for A-splines*. Computer Aided Geometric Design, 16, 39–59, 1999.

[3] Boor C., Conte S. D., *Elementary Numerical Analisys An Algorithmic Approach*. Editorial Félix Varela, La Habana, 2004.

[4] Courant, R.; Hilbert, D., *Methods of mathematical physics*, I, New York, N.Y.: Interscience Publishers, Inc., ISBN 0-471-50447-5, 1953.

[5] Díaz, R., *Esquema de subdivisión interpolatorio con parámetros de tensión local basado en spline cónico*. Tesis presentada en opción del grado de Licenciado en Ciencias Matemáticas, Universidad de la Habana, 2010.

[6] Farin, G., *Curves and Surfaces for Computer Aided Geometric Design: a practical guide*. Academic Press Inc, 1997.

[7] Levien, R., Séquin C., *Interpolating Splines: Which is the fairest of them all?*. Computer-Aided Design & Applications, 6(1), 91–102, 2009.

[8] Rudin, W., *Principles of Mathematical Analysis* (3rd ed.) New York: McGraw-Hill. ISBN 978-0-07-054235-8. 1976.

[9] Yang, Xunnian., *Curve fitting and fairing using conic splines*. Computer-Aided Design 36, 461–472, 2004.

[10] Yong, J-H., Cheng F., *Geometric Hermite curves with minimum strain energy*. Computer-Aided Design 21, 281-301, 2004.

[11] Y.J. Ahn, C. Hoffmann, P. Rosen, *Geometric constraints on quadratic Bézier curves using minimal length and energy*, Journal of Computational and Applied Mathematics (2013).

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 107-114, 2018
Recibido 02-2018

Modelo de Programación Lineal Difusa Multiobjetivo para la evacuación óptima de personas bajo amenaza de desastres naturales
Fuzzy Multiobjective Linear Programming Model for the optimal evacuation of people under threat of natural disasters

Ernesto Parra Inza[1]*, Carlos Segura Vidal[1], José María Sigarreta Almira[2], Juan Carlos Hernández Gómez[2]

Resumen La evacuación de personas es un proceso sustantivo dentro de la gestión operativa de desastres. En Cuba, y en especial en la provincia de Holguín, dicho proceso incluye el control sistemático de los datos poblacionales de cada municipio, y la confección de un plan de respuesta de acuerdo a diversos criterios. Esto último se realiza tradicionalmente de forma manual, con las correspondientes limitaciones en la toma de decisiones involucradas. En ese sentido, la presente investigación tiene por objetivo modelar mediante un enfoque de programación lineal difusa multiobjetivo, el problema de transporte asociado a la evacuación de personas ante desastres naturales en la provincia de Holguín. El enfoque propuesto permite obtener de manera rápida y eficiente una propuesta de evacuación que facilitará la toma de decisiones teniendo en cuenta múltiples criterios y la presencia de incertidumbre en los datos. Se han considerado cuatro casos de estudios relacionados con posibles escenarios de evacuación de personas en la provincia de Holguín. Los resultados muestran que el enfoque propuesto resulta eficaz y lo suficientemente pertinente para ser aplicado en escenarios reales.

Abstract The evacuation of people is a substantive process within the operational management of disasters. In Cuba, and especially in the province of Holguín, this process includes the systematic control of the population data of each municipality, and the preparation of a response plan according to various criteria. The latter is traditionally done manually, with the corresponding limitations in the decision-making involved. In this sense, the present research aims to model the problem of transport associated with the evacuation of people to natural disasters in the province of Holguin using a fuzzy multiobjective linear approach. The proposed approach makes it possible to obtain a fast and efficient evacuation proposal that would facilitate the decision making taking into account multiple criteria and the presence of uncertainty in the data. We have considered four cases of studies related to possible scenarios of evacuation of people in the province of Holguin. The results show that the proposed approach is effective and relevant enough to be applied in real scenarios.

Palabras Clave
Programación lineal difusa multiobjetivo — gestión operativa de desastres

[1] *Departamento de Licenciatura en Matemática, Universidad de Holguín, Holguín, Cuba, eparrainza@gmail.com, csegurav@uho.edu.cu*
[2] *Unidad Académica de Matemáticas, Universidad Autónoma de Guerrero, Guerrero, México, josemariasigarretaalmira@hotmail.com, jcarloshguagro@gmail.com*
***Autor para Correspondencia**

1. Introducción

Responder ante los desastres de manera eficiente no es una tarea fácil, pues los factores que intervienen en estos procesos son numerosos. Por ejemplo, el ambiente post-ayuda en caso de desastre es caótico, existe el pánico público, el transporte se pierde así como la infraestructura de comunicación; el número y la variedad de actores involucrados es alto (donantes, medios de comunicación, gobierno, militares, organizaciones humanitarias, entre otros); además de la falta de recursos suficientes para proveer una respuesta adecuada ante la situación.

Una ayuda humanitaria eficiente pero flexible es un tema clave en caso de desastres, del cual se está hablando con mayor auge en el mundo académico actual [9] como una extensión de esta,

la logística humanitaria es una de las disciplinas de mayor importancia dentro del manejo de desastres [14]. Uno de los grandes obstáculos para superar en cadenas de suministro de ayuda humanitaria, es la enorme incertidumbre y los múltiples objetivos en la demanda, los suministros y la siempre existente presión que ejerce el tiempo. La logística humanitaria es un proceso considerado de alto nivel de complejidad, pues constituye la parte más costosa de la mitigación de desastres [24]; esta se encarga en la etapa de recuperación de minimizar los efectos del desastre, realizando la búsqueda y rescate de víctimas, y la provisión de víveres y servicios de emergencias. La Defensa Civil mantiene un estricto control de las vías de acceso terrestre a una gran cantidad de puntos estratégicos del territorio de Holguín, que le permite realizar las labores logísticas humanitarias. Como consecuencia, durante o luego de la crisis, toda esta información varía de manera vertiginosa, por lo que se hace prácticamente imposible mantener de manera manual la información del estado real de las vías de acceso, así como proponer rutas óptimas que minimicen los gastos de recursos y maximicen la ayuda brindada. De esta forma, el procedimiento antes mencionado incurre en gastos excesivos de recursos y requiere de mucho tiempo para consolidar la información. Por tanto se aprecia que es muy difícil detectar posibles errores, lo que dificulta la toma de decisiones.

En el caso de desastres, uno de los trabajos de la Defensa Civil es el transporte de la población a zonas de menos riesgo, así como minimizar el tiempo de este trabajo y el costo que incurre en el mismo. Teniendo en cuenta los anteriores objetivos, el escenario para dar respuesta a tal situación se muestra plagado de incertidumbres. Lo anteriormente descrito introduce el siguiente problema: ¿cómo planificar de manera óptima la evacuación de personas bajo amenaza de desastres naturales en la provincia de Holguín, de manera que se tenga en cuenta la incertidumbre de los datos, así como el cumplimiento de los objetivos humanitarios y económicos?

Dada la posibilidad de modelar estos escenarios de decisión como problemas de programación lineal difusa y la necesidad del cumplimiento de múltiples objetivos, la presente investigación se propuso como objetivo: resolver mediante un enfoque de programación lineal difusa multiobjetivo (PLDM) el problema de transporte asociado a la evacuación de personas durante la amenaza de huracanes en la provincia de Holguín. Con este trabajo se espera mejorar la toma de decisiones en casos de desastres, que permita salvar vidas y economizar los gastos a los que se enfrenta la provincia en estas situaciones.

2. Gestión operativa de desastres

Se entiende como desastre el acontecimiento o suceso que destruye las estructuras básicas y el funcionamiento normal de una sociedad o comunidad. Ocasiona pérdidas y afectaciones humanas, a la economía, la infraestructura, los servicios esenciales o medios de sustento, más allá de la capacidad normal de las comunidades afectadas para dar una respuesta. Los peligros de desastres, que potencialmente pueden afectar al país, han sido clasificados, atendiendo a su origen en: naturales, tecnológicos y sanitarios.

En los últimos 20 años una nueva disciplina ha emergido en el contexto de la Investigación de Operaciones y las Ciencias de la Administración aplicadas en la gestión de desastres naturales. Altay y otros (2006) [1] la llamaron Gestión operativa de desastres (DOM), además la definen como el conjunto de actividades realizadas previamente, durante y después a un desastre con el objetivo de prevenir la pérdida de vidas humanas, reduciendo su impacto en la economía, y retornar a un estado de normalidad.

3. Programación lineal difusa multiobjetivo

La extensión difusa de la programación lineal trata la incertidumbre de determinados elementos del modelo mediante la teoría de la lógica difusa[25]. Conviene por tanto definir primero que es un conjunto difuso.

Conjunto difuso[3]: Sea $Y \subseteq \mathbb{R}$ un conjunto no vacío. Un subconjunto difuso de Y es una función $\mu : Y \to [0,1]$.

El valor de y en la función μ expresa una medida o grado para el cual y está en Y. Si $\mu(y) = 1$ entonces $y \in Y$. Por otro lado, si $\mu(y) = 0$ se puede decir que $y \notin Y$. En el caso de $\mu(y) \in (0,1)$, dicho valor proporciona el grado de pertenencia de y en Y. La función $\mu(x)$ recibe el nombre de función de pertenencia.

Cuando la función objetivo $z = c^T x$ es difusa, se supone que existe un valor de aspiración para la función objetivo. Dicho valor se denotará $d_0 \in \mathbb{R}$.

Es decir, se espera encontrar un $x^* \in X$ tal que $z(x^*) \leq d_0$. En muchos casos no es posible encontrar una solución que satisfaga esta condición, por lo cual se permite que la función objetivo pueda alcanzar valores mayores a d_0.

Para esto, se fija un valor p_0 que define el grado mínimo de cumplimiento o pertenencia al nivel de aspiración. De esta forma, si $z(x) \geq d_0 + p_0$ se dice que tiene un grado de cumplimiento de 0. Si $z(x) \leq d_0$, el grado de cumplimiento o pertenencia es 1. Además, si $d_0 \leq z(x) \leq d_0 + p_0$ entonces el grado o porcentaje de cumplimiento está dado por $1 - \frac{z(x)}{p_0}$. Lo descrito anteriormente se puede expresar por medio de la siguiente función de pertenencia trapezoidal para la función objetivo [15].

$$\mu_z(z(x)) = \begin{cases} 1, & \text{si} \quad z(x) \leq d_0 \\ 1 - \frac{z(x)-d_0}{p_0}, & \text{si} \quad d_0 \leq z(x) \leq d_0 + p_0 \\ 0, & \text{si} \quad z(x) \geq d_0 + p_0 \end{cases} \quad (1)$$

Luego introduciendo una variable auxiliar , el modelo de programación lineal del problema de minimización con función objetivo difusa se expresa de la siguiente forma:

$$\begin{aligned} &\text{máx} \quad \lambda \\ &s.a: \ \mu_z(\sum_{i=1}^{n} c_i x_i) \geq \lambda \\ &x \in X, \ \ \lambda \in [0,1] \end{aligned} \tag{2}$$

Como puede verse en [15], el problema anterior es equivalente al siguiente problema de optimización paramétrica:

$$\begin{aligned} &\text{máx} \quad \lambda \\ &s.a: \ \sum_{i=0}^{n} c_i x_i \geq d_0 - p_0(1-\lambda) \\ &x \in X, \lambda \in [0,1] \end{aligned} \tag{3}$$

cuya solución óptima λ^*, x^* se considera la solución del problema original con función objetivo $z = c_i x$ difusa.
Si se está analizando un problema donde no solo la función objetivo, sino que además las restricciones son difusas, entonces se puede tomar la decisión difusa [2] considerando que no existe diferencia entre función objetivo difusa $c_i x \leq d_0$ y restricciones difusas $Ax \sim\leq d_0$, este modelo se puede expresar de la forma siguiente [26]: $Bx \sim\leq b^{'}$, donde

$$B = \begin{bmatrix} c \\ A \end{bmatrix} \quad y \quad b^{'} = \begin{bmatrix} d_0 \\ b \end{bmatrix}$$

Aplicando (1) a Bx, el modelo (3) quedaría de la siguiente forma

$$\begin{aligned} &\text{máx} \quad \lambda \\ &s.a: \ \sum_{i=0}^{n} Bx_i \geq b_0^{'} - p_0(1-\lambda) \\ &x \in X, \lambda \in [0,1] \end{aligned} \tag{4}$$

En 1987, Zimmermann, extendió su enfoque sobre la programación lineal difusa al problema de programación lineal multiobjetivo con K funciones objetivo $z_i = c_j x$, $i \in \{1,...,K\}$. Para cada una de las funciones objetivo $z_i = c_j x$ de este problema, se asume que aquella persona especialista o capacitada para llevar a cabo la toma de decisiones, tenga un objetivo difuso tal como $z_i(x)$ debe ser menor que o igual a cierto valor d_i . Entonces la correspondiente función de pertenencia de un problema de minimización puede ser escrita como [17]:

$$\mu_i^L(z_i(x)) = \begin{cases} 0, & \text{si} \quad z_i(x) \leq z_i^0 \\ \frac{z_i(x) - z_i^0}{z_i^1 - z_i^0}, & \text{si} \quad z_i^0 \geq z_i(x) \geq z_i^1 \\ 1, & \text{si} \quad z_i(x) \leq z_i^1 \end{cases} \tag{5}$$

donde z_i^0 y z_i^1 denotan los valores de la función objetivo $z_i(x)$ tales que los grados de la función de pertenencia sean 0 y 1, respectivamente.
Usando esta función de pertenencia y siguiendo las reglas de decisión difusa por [2], el modelo puede interpretarse como:

$$\begin{aligned} &\text{máx} \ \ \text{mín}_{i=1,...,k} \ \mu i^L(z_i(x)) \\ &s.a: \ Ax \leq b, x \geq 0 \end{aligned}$$

Téngase en cuenta que las restricciones se consideran rígidas. Este problema puede ser reducido al siguiente problema de programación lineal convencional:

$$\begin{aligned} &\text{máx} \ \ \lambda \\ &s.a: \quad \lambda \leq \mu_i^L(z_i(x)) \\ &\qquad Ax \leq b, x \geq 0 \end{aligned}$$

Asumiendo la existencia de la solución óptima x^{i0} del problema de minimización de las funciones objetivos individuales bajo las restricciones se define por: $\text{mín}_{x \in X} \ z_i(x), \ \ i = 1,...,k$ donde $X = \{x \in \mathbb{R}^n | Ax \leq b, x \geq 0\}$, en [27] el autor sugiere una forma para determinar la función lineal de pertenencia $\mu_i^L(z_i(x))$. Para ser específicos, usando el mínimo individual $\forall i = 1,..,k$

$$z_i^{min} = z_i(x^{i0}) = \min_{x \in X} z_i(x)$$

junto con

$$z_i^m = \text{máx}\{z_i(x^{10}),...,z_i(x^{i-1,0}, z_i(x^{i+1,0}),...,z_i(x^{k,0}))\}$$

determinó la función lineal de pertenencia como (5) pero escogiendo $z_i^1 = z_i^{min}$ y $z_i^0 = z_i^m$.
En el caso donde no solo las funciones objetivos sean difusas sino también las restricciones lo sean, utilizando igual función de pertenencia un análisis similar puede ser empleado [17]. Aunque existen otros métodos, este es el que se emplea en el desarrollo de la investigación.

4. Modelos de evacuación en desastres naturales

La mayoría de los modelos de evacuación disponibles definen su objetivo como minimizar el flujo del tráfico o el tiempo total de transportación [20],[23],[13]. En [5], [21], [18] consideraron seguridad en sus modelos, pero hacen esto penalizando o prohibiendo soluciones que dejen evacuados detrás, al final de la evacuación. De manera similar, en localizar refugios, [10] minimizan el peso promedio de las demandas no logradas por los refugios y el tiempo de transportación.
La función objetivo en [21] minimiza una vez más el tiempo total de evacuación, pero incluyen restricciones sobre demostraciones y costos operativos para ocuparse de esos objetivos adicionales.
Varias estrategias mejoradas de evacuación han sido consideradas en la literatura, incluyendo: determinar la ruta de evacuación óptima y/o destino asignado (ej. [5],[21],[13]); planificar en etapas (también conocido como escenificar) la evacuación (ej. [19], [5],[4],[11]); y otras estrategias de control de tráfico [6], [22],[12]. Hay muchos aspectos de una evacuación: quién se queda, quién se va, cuándo se va, donde va, que ruta tomará para llegar ahí, y que camino y modo de transporte está disponible y cuándo. Cada modelo de evacuación asume

cada uno de estos aspectos del problema como un aporte incontrolable o algo que potencialmente puede controlarse a fin de mejorar la evacuación. Muchos de los primeros modelos son meramente descriptivos, asumiendo que cada dimensión brinda un aporte. MASSVAC [7], [8], OREMS [16], y NETVAC [20] son ejemplos de estos tipos de modelos. Estos tipos de modelos pueden ser usados de dos formas:
a) para estimar los tiempos de realización que ayuden a decidir cuando deberían ser emitidas las órdenes de evacuación a fin de asegurar el tiempo adecuado para la ejecución o
b) desarrollar planes de evacuación a través de un proceso por tanteo en el cual las suposiciones diferentes (los planes propuestos) de aportes son manejadas y el modelo se usa para evaluarlas.
Otros modelos son preceptivos, asumiendo que uno o más aspectos del problema son potencialmente controlables y teniendo la intención de determinar la forma óptima para controlarla. Estos modelos varían en que los aspectos están predeterminados por el comportamiento del sistema de transporte existente o del evacuado, y que son considerados controlables, así se someten a la mejora a través de la implementación de una estrategia de evacuación.
En realidad, muchos aspectos de una evacuación son parcialmente controlables. Por ejemplo, no se puede asumir realmente que las personas cumplan con el momento exacto en que se les dice deben partir; de cualquier forma, a través de las órdenes de evacuación obligatorias, las autoridades pueden ejercer cierto control sobre quién se va y cuándo. Favorece, incluso si el proceso óptimo descrito por un modelo preceptivo no es enteramente factible en realidad, este puede ser usado para prever cuán bien puede ir un proceso de evacuación. El componente uno de la evacuación que, para el conocimiento de los autores, ha sido siempre considerado aporte es quién se va. Los modelos disponibles por consiguiente no permiten la posibilidad que la mejor estrategia para las personas podría ser quedarse donde ellos están. Esta suposición está probablemente relacionada con la declaración del objetivo, como minimizar tiempo de despeje del tráfico, puesto que un modelo que deja a las personas quedarse cuando el único objetivo es minimizar el despeje del tránsito, está aconsejando que todo el mundo se quede dónde está.

5. Descripción del problema

Antes de exponer los detalles del enfoque propuesto conviene describir el escenario de decisión que se pretende resolver. El mismo se puede definir informalmente de la forma siguiente:
Dado un conjunto de centros de evacuación, con capacidades y costos de utilización conocidos, y de localidades con personas a evacuar, determinar el esquema de transporte que cumpla con las demandas de las localidades y minimice simultáneamente el tiempo total de evacuación y el costo de utilización de los centros.
Téngase en cuenta que los tiempos de transportación, así como las capacidades de los centros de evacuación están definidos de manera difusa, esto es, sin seguir una distribución estadística conocida.
Una vez las personas han sido llevadas a los centros de evacuación, éstas se deben atender desde el punto de vista logístico y de salud. Para realizar dichas atenciones, existen centros de recursos (almacenes) y centros de salud (hospitales, policlínicos, etc.). Por lo tanto, sería importante tener en cuenta las capacidades de atención de estos centros, y que el tiempo de transportación total, desde los centros de evacuación a estos, sea mínima. Es decir, a partir de la solución encontrada para el problema de evacuación, sería conveniente determinar cómo atender de manera efectiva a estas personas evacuadas, teniendo en cuenta que el tiempo total de transportación, así como la capacidad de estos centros son igualmente difusos.
Notación de los datos de entrada

- M_k: capacidad de atención del centro logístico k, $k \in \{1,...,K\}$ siendo K el número de centros logísticos.
- N_l: capacidad de atención del centro de salud l, $l \in \{1,...,L\}$ siendo L el número de centros de salud.
- P_i: parte entera de la división del número de personas que deben ser evacuados en la localidad i, por la capacidad promedio de los ómnibus a emplear; $i \in \{1,...,I\}$ siendo I el número de localidades a evacuar.
- t_{ij}: tiempo de transportación de un medio de transporte desde la localidad i hacia el centro j.
- t^1_{kj}: tiempo de transportación de un medio de transporte desde el centro logístico k hacia el centro de evacuación j.
- t^2_{jl}: tiempo de transportación de un medio de transporte desde el centro de evacuación j hacia el centro de salud l.
- c_j: costo de utilización del centro j en \$.

Con el fin de dar solución al problema antes descrito se propone emplear el enfoque de programación lineal difusa multiobjetivo analizado con anterioridad; para esto se crea el siguiente modelo de programación lineal, del cual se describen las variables de decisión así como las funciones objetivos y las restricciones asociadas. Téngase en cuenta que este modelo se encuentra en su fase inicial de desarrollo.

5.1 Variables de decisión

- x_{ij}: número de viajes a realizar desde la localidad i hasta el centro de evacuación j.
- y_j: si se utiliza (1) o no (0) el centro de evacuación j.
- z^1_{kj}: número de viajes a realizar desde el centro logístico k hasta el centro de evacuación j.
- z^2_{jl}: número de viajes a realizar desde el centro de evacuación j hasta el centro de salud l.

5.2 Funciones objetivos

Las funciones objetivo que se presentan son difusas pues t_{ij}, t^1_{kj}, t^2_{jl} y c_j son valores difusos que dependen del criterio de los expertos en la toma de decisión que intervienen en el problema analizado.

- Minimizar el tiempo de transportación de la evacuación:

$$\text{mín } Z_{te} = \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}x_{ij}$$

- Minimizar el costo de utilización de los centros de evacuación:

$$\text{mín } Z_{costo} = \sum_{j=1}^{J} c_j y_j$$

- Minimizar el tiempo de transportación de los recursos logísticos:

$$\text{mín } Z_{tl} = \sum_{k=1}^{K}\sum_{j=1}^{J} t^1_{kj}z^1_{kj}$$

- Minimizar el tiempo de transportación a los centros de salud:

$$\text{mín } Z_{ts} = \sum_{j=1}^{J}\sum_{l=1}^{L} t^2_{jl}z^2_{jl}$$

5.3 Restricciones

Restricciones de capacidad de los centros de evacuación, logísticos y salud en este orden. Considérese que E_j, M_k y N_l son elementos difusos que también dependen de las circunstancias y las situaciones existente en los centros.

$$\begin{aligned}
\sum_{i=1}^{I} x_{ij} &\preceq E_j y_j, \quad \forall j = \{1,...,J\}\\
\sum_{j=1}^{J} z^1_{kj} &\preceq M_k, \quad \forall k = \{1,...,K\}\\
\sum_{j=1}^{J} z^2_{jl} &\preceq N_l, \quad \forall l = \{1,...,L\}
\end{aligned}$$

Restricciones sobre el número de viajes a realizar teniendo en cuenta el número de personas a evacuar en las localidades, su atención logística y médica.

- Evacuación:

$$\sum_{j=1}^{J} x_i j = P_i, \quad \forall i = \{1,...,I\}$$

- Atención logística:

$$\sum_{k=1}^{K} z^1_{kj} = \sum_{i=1}^{I} x_{ij}, \quad \forall j = \{1,...,J\}$$

- Atención médica:

$$\sum_{l=1}^{L} z^2_{jl} = \sum_{i=1}^{I} x_{ij}, \quad \forall j = \{1,...,J\}$$

$$x_{ij}, z^1_{kj}, z^2_{jl} \geq 0; \quad x_{ij}, z^1_{kj}, z^2_{jl} \in \mathbb{Z}^+; \quad y_j \in \{0,1\}$$

6. Solución mediante MATLAB

Para la solución de este problema se utilizó el enfoque sugerido por Zimmermann, descrito con anterioridad, aplicado a la función de pertenencia (5). Siguiendo el enfoque de (Bellman et al.,) para decisión difusa, se considera que no existe diferencia entre funciones objetivo difusas $c_i x \preceq d_0$ y restricciones difusas $Ax \preceq b$, se aplica entonces de igual forma la función de pertenencia tanto a las funciones objetivos difusos como a las restricciones difusas. De tal manera el modelo propuesto sería transformado en un modelo de tipo (4), añadiendo las restricciones no difusas.

Al aplicar la función de pertenencia (5) a las funciones objetivo difusas estas toman la siguiente forma:

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}x_{ij} - (z^1_{te} - z^0_{te})\lambda &\leq z^0_{te}\\
\sum_{j=1}^{J} c_j y_j - (z^1_{costo} - z^0_{costo})\lambda &\leq z^0_{costo}\\
\sum_{k=1}^{K}\sum_{j=1}^{J} t^1_{kj}z^1_{kj} - (z^1_{tl} - z^0_{tl})\lambda &\leq z^0_{tl}\\
\sum_{j=1}^{J}\sum_{l=1}^{L} t^2_{jl}z^2_{jl} - (z^1_{ts} - z^0_{ts})\lambda &\leq z^0_{ts}
\end{aligned}$$

Donde z^0 y z^1 expresan los máximos y mínimos que alcanzan dichas funciones de forma independiente y asociadas a las restricciones no difusas.

```
for i=1:cantfundif
    [~,fvalmin(i)=linprog(zdif(i,:),[],[],Aeq,beq,
        lb,up,[],options)];
    [~,fval(i)=linprog(-zdif(i,:),[],[],Aeq,beq,lb
        ,up,[],options)];
    z0d(i)=-fvalm(i);
    zrestd(i,:)=[zdif(i,:)-(fvalmin(i)+fvalm(i))];
end
```

En el caso de las restricciones de desigualdad se sigue la decisión difusa [2] considerando que no existe diferencia entre función objetivo difusa y restricción difusa, por lo que se aplica de igual forma la función de pertenencia; considerando que *bm* y *bmin* son los valores máximos y mínimos permitidos a tales restricciones, estas quedarían de la siguiente forma:

$$\begin{aligned}
\sum_{i=1}^{I} x_{ij} - E_j y_j - (b^1_{min} - b^1_m)\lambda &\leq b^1_m\\
\sum_{j=1}^{J} z^1_{kj} - M_k - (b^2_{min} - b^2_m)\lambda &\leq b^2_m\\
\sum_{j=1}^{J} z^2_{jl} - N_l - (b^3_{min} - b^3_m)\lambda &\leq b^3_m
\end{aligned}$$

```
for i=1:cantfundif
    bdifa(i)=bm(i);
    Adifa(i,:)=[Adif(i,:)-(bmin(i)-bm(i))];
end
```

Solo queda plantear el modelo (4), uniendo las funciones objetivos con las restricciones de desigualdad, y crear la nueva función objetivo. Luego este es resuelto mediante *linprog*, algoritmo de optimización para programación lineal que utiliza el método Simplex o Punto Interior.

```
fun=[zeros(1,cantvar)-1];
Bx=[Adifa;zrestd];
bpri=[bdifa,z0d];
Bxeq=[Aeq,zeos(cantrestndifeq,1)];
bprieq=beq;
[x, solucion]=linprog(fun,Bx,bpri,Bxeq,bprieq,[lb
    ,-inf],[up,inf],[],options);
```

6.1 Caso de estudio

Debido a la seguridad con la que se manejan estos datos, no están disponibles al público, por lo que no fue posible obtener los datos reales que se manejan en el proceso de evacuación de personas bajo amenaza de desastres en la provincia de Holguín. No obstante, se utilizó para los casos de estudio (CE), datos diseñados, que están cerca de la realidad.

	Con atención médica	Sin atención médica
Datos I	CE 1	CE 3
Datos II	CE 2	CE 4

Para la elaboración de los casos de estudio se consideraron dos casos, uno en el cual se analiza el enfoque propuesto, en el que se incluye atención médica y en el otro no se incluye atención médica, cada uno evaluado en dos instancias del problema.

Estas instancias presentan las siguientes características:

Considere que se desea evacuar 14 municipios con poblaciones P_i y se cuenta con 47 centros de evacuación con capacidad K_j. Estos, a su vez, reciben atención logística desde 33 centros con capacidades M_k y atención médica desde 33 centros con capacidades N_l. El costo de utilización de los centros de evacuación es de c_i . Los factores que se variaron fueron los tiempos de transportación desde los municipios hasta los centros de evacuación, y de estos hasta los centros de atención logística y médica.

En la siguiente tabla se presentan los resultados de los casos de estudio, en la que se muestran los tiempos totales de transportación y el costo de utilización de los centros. También se propone la solución de estos casos de estudio, resueltos como problemas sin un enfoque difuso a través del método minimax ponderado. Además de los tiempos de ejecución de los algoritmos en cada instancia.

	Enfoque difuso (Zimmermann, 1976)	Tiempo (s)
CE1	(7701,218,5851,7401)	4.489904
CE2	(8234,218,4184,5703)	4.496354
CE3	(7701,218,5851)	3.330397
CE4	(8240,218,4179)	3.011402
	Enfoque no difuso Minimax ponderado (Bowman, 1976)	Tiempo (s)
CE1	(7218,218,4149,5615)	21.411992
CE2	(7662,218,4149,5615)	22.149234
CE3	(7218,218,5439)	7.964351
CE4	(7662,218,4145)	7.907791

Centrando el análisis solo en la parte cuantitativa de los resultados obtenidos en los casos de estudio presentados, se puede plantear que el modelo difuso arroja resultados más desfavorables que el modelo no difuso. De estos resultados no podemos limitar al método difuso, ni absolutizar la superioridad del enfoque no difuso, pues en el aspecto cualitativo, el modelo difuso se considera superior, al no difuso, ya que logra incorporar y aglutinar el criterio de múltiples expertos; que en la vida real sin el análisis difuso, cada opinión de un experto representaría un modelo diferente, debido a que es poco probable que todos tengan la misma opinión o pensamiento lógico. Por otra parte si se analiza el tiempo de ejecución de los algoritmos implementados, aquel que brinda solución al enfoque no difuso puede tardar hasta cuatro veces el tiempo de ejecución del otro. Para la implementación de los algoritmos se utilizó el MATLAB, un lenguaje de programación de alto nivel, con un enfoque predominantemente matemático; que permite obtener resultados rápidos y confiables.

7. Conclusiones

En este artículo se hace un breve análisis de los fundamentos teóricos que sustentan la programación lineal difusa multi-objetivo, así como la gestión operativa de desastres y algunos modelos que tratan el tema de la evacuación ante desastres.

El modelo de programación lineal difuso y multi-objetivo propuesto contribuye a la gestión operativa de desastres, ya que facilita la toma de decisiones, además de tener en cuenta múltiples criterios y la presencia de incertidumbre en los datos.

Se implementaron los algoritmos en MATLAB y los resultados arrojados permitieron el análisis comparativo de los casos de estudio y demostraron la funcionalidad del modelo.

Referencias

[1] Nezih Altay and Walter G Green. Or/ms research in disaster operations management. *European journal of operational research*, 175(1):475–493, 2006.

[2] Richard E Bellman and Lotfi Asker Zadeh. Decision-making in a fuzzy environment. *Management science*, 17(4):B–141, 1970.

[3] James J Buckley and Esfandiar Eslami. *An introduction to fuzzy logic and fuzzy sets*, volume 13. Springer Science & Business Media, 2002.

[4] X Chen and FB Zhan. Agent-based modeling and simulation of urban evacuation: relative effectiveness of simultaneous and staged evacuation strategies. In *Agent-Based Modeling and Simulation*, pages 78–96. Springer, 2014.

[5] Yi-Chang Chiu, Hong Zheng, Jorge Villalobos, and Bikash Gautam. Modeling no-notice mass evacuation using a dynamic traffic flow optimization model. *IIE Transactions*, 39(1):83–94, 2007.

[6] Thomas J Cova and Justin P Johnson. A network flow model for lane-based evacuation routing. *Transportation research part A: Policy and Practice*, 37(7):579–604, 2003.

[7] Antoine G Hobeika and Bahram Jamei. Massvac: A model for calculating evacuation times under natural disasters. *Emergency Planning*, pages 23–28, 1985.

[8] Antoine G Hobeika and Changkyun Kim. Comparison of traffic assignments in evacuation modeling. *IEEE transactions on engineering management*, 45(2):192–198, 1998.

[9] Gyöngyi Kovács and Karen M Spens. Humanitarian logistics in disaster relief operations. *International Journal of Physical Distribution & Logistics Management*, 37(2):99–114, 2007.

[10] Anna CY Li, Ningxiong Xu, Linda Nozick, and Rachel Davidson. Bilevel optimization for integrated shelter location analysis and transportation planning for hurricane events. *Journal of Infrastructure Systems*, 17(4):184–192, 2011.

[11] Yue Liu, Gang-Len Chang, Ying Liu, and Xiaorong Lai. Corridor-based emergency evacuation system for washington, dc: system development and case study. *Transportation Research Record: Journal of the Transportation Research Board*, (2041):58–67, 2008.

[12] Qiang Meng and Hooi Ling Khoo. Optimizing contraflow scheduling problem: model and algorithm. *Journal of Intelligent Transportation Systems*, 12(3):126–138, 2008.

[13] ManWo Ng, Junsik Park, and S Travis Waller. A hybrid bilevel model for the optimal shelter assignment in emergency evacuations. *Computer-Aided Civil and Infrastructure Engineering*, 25(8):547–556, 2010.

[14] Ehsan Nikbakhsh and Reza Zanjirani Farahani. Humanitarian logistics planning in disaster relief operations. *Logistics Operations and Management: Concepts and Models*, 291, 2011.

[15] Jaroslav Ramík. Soft computing: overview and recent developments in fuzzy optimization. *Ostravská univerzita, Listopad*, pages 33–42, 2001.

[16] Ajay K Rathi and Rajendra S Solanki. Simulation of traffic flow during emergency evacuations: a microcomputer based modeling system. In *Simulation Conference Proceedings, 1993. Winter*, pages 1250–1258. IEEE, 1993.

[17] Masatoshi Sakawa, Hitoshi Yano, Ichiro Nishizaki, and Ichiro Nishizaki. *Linear and multiobjective programming with fuzzy stochastic extensions*. Springer, 2013.

[18] Fatemeh Sayyady and Sandra D Eksioglu. Optimizing the use of public transit system during no-notice evacuation of urban areas. *Computers & Industrial Engineering*, 59(4):488–495, 2010.

[19] Hayssam Sbayti and Hani Mahmassani. Optimal scheduling of evacuation operations. *Transportation Research Record: Journal of the Transportation Research Board*, (1964):238–246, 2006.

[20] Yosef Sheffi, Hani Mahmassani, and Warren B Powell. A transportation network evacuation model. *Transportation research part A: general*, 16(3):209–218, 1982.

[21] Qian Tan, Guo H Huang, Chaozhong Wu, Yanpeng Cai, and Xinping Yan. Development of an inexact fuzzy robust programming model for integrated evacuation management under uncertainty. *Journal of Urban Planning and Development*, 135(1):39–49, 2009.

[22] Gregoris Theodoulou and Brian Wolshon. Alternative methods to increase the effectiveness of freeway contraflow evacuation. *Transportation Research Record: Journal of the Transportation Research Board*, (1865):48–56, 2004.

[23] Suleyman Tufekci and Thomas M Kisko. Regional evacuation modeling system (rems): A decision support system for emergency area evacuations. *Computers & industrial engineering*, 21(1-4):89–93, 1991.

[24] Luk N Van Wassenhove. Humanitarian aid logistics: supply chain management in high gear. *Journal of the Operational research Society*, 57(5):475–489, 2006.

[25] Lofti Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE transactions on Automatic Control*, 8(1):59–60, 1963.

[26] Hans-J Zimmermann. Description and optimization of fuzzy systems. *International Journal of General System*, 2(1):209–215, 1975.

[27] Hans-Jürgen Zimmermann. *Fuzzy sets, decision making, and expert systems*, volume 10. Springer Science & Business Media, 2012.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 115-123, 2018
Recibido 02-2018

Método de las diferencias finitas aplicado a un problema elíptico unidimensional con condiciones de discontinuidad
Finite differences method applied to a one-dimensional elliptic problem with discontinuity conditions

Frank Ernesto Alvarez Borges[1], Julián Bravo Castillero[1]*, Angela León Mecías[2], Raúl Guinovart Díaz[1], Reinaldo Rodriguez Ramos[1]

Resumen En el presente trabajo es planteada la ecuación del calor con condiciones de contacto imperfecto, para luego obtener un esquema en diferencias finitas que aproxime al problema previamente presentado. Es analizada la consistencia, estabilidad y convergencia del método utilizado, para luego discutir algunos ejemplos.

Abstract The heat equation with non-perfect contact conditions is stated, and then, a finite difference scheme is obtained in order to approximate it. Consistency, stability and convergence is studied and some examples are discussed.

Palabras Clave
Problema elíptico — Contacto Imperfecto — Diferencias finitas

[1] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, La Habana, Cuba, jb_castillero@yahoo.com.mx,*
[2] *Departamento de Matemática Aplicada, Facultad de Matemática y Computación, Universidad de la Habana, La Habana, Cuba, angela@matcom.uh.cu*
***Autor para Correspondencia**

Introducción

La ecuación del calor, acoplada con determinadas condiciones de contacto (que puede ser perfecto o imperfecto), aparece en el estudio de la transmisión del calor por regiones compuestas por materiales de distintas propiedades, así como en otros problemas de la mecánica de sólidos.

Esta diferencia de propiedades puede ser modelada a través de condiciones de discontinuidad. Diversos árticulos han estudiado este tipo de problemas mediante un enfoque numérico. En [**?**] y [**?**] se abordan problemas que presentan discontinuidades mediante el método de Elementos Finitos. El mét́odo de las Diferencias Finitas, el cual será abordado en el presente documento, fue utilizado en [**?**] para resolver las ecuaciones de Maxwell en el dominio del tiempo, en [**?**] y [**?**] es considerado un sistema de ecuaciones con coeficientes dependientes de variables temporales y espaciales, con condiciones de discontinuidad, en una y dos dimensiones respectivamente.

A continuación será planteada la ecuación que modela un problema estacionario de conducción del calor, en el caso unidimensional con condiciones de contacto imperfecto. Más adelante, en la Sección 1, el método de las diferencias finitas es aplicado al problema previamente obtenido. En la Sección 2 se analizará la consistencia y estabilidad del método, así como la convergencia de la solución numérica a la solución exacta. Finalmente en la Sección 3 son presentados y discutidos algunos ejemplos.

0.1 Obtención de la ecuación del calor

La ecuación del calor con convección viene dada por la expresión

$$d\frac{\partial u}{\partial t} - \nabla\cdot(a\nabla u) + b\cdot\nabla u + cu = f \text{ en } \Omega, \text{ donde } c = \nabla\cdot b. \tag{1}$$

La deducción de esta ecuación a partir de una ley de conservación y varias relaciones constitutivas puede ser encontrada en la introducción de [3]. Para el caso de un problema estacionario (independiente del tiempo), unidimensional (por ejemplo, el problema de conducción del calor sobre una barra metálica),

donde $b = 0$, la ecuación (1) queda reducida a

$$-\frac{d}{d\xi}\Big(a(\xi)\frac{d}{d\xi}u\Big) = f, \text{ en } \Omega. \tag{2}$$

En dicho caso Ω sería un intervalo. Si además se considera que la barra está conformada por secciones, cada una de diferente material, y por lo tanto, con diferentes propiedades, en cada uno de los puntos de enlace de dichas secciones deben cumplirse condiciones del tipo Robin

$$a\frac{du}{d\xi} + \kappa(u^+ - u^-) = 0,$$

donde κ es un coeficiente de transferencia de calor. Es decir, sobre los enlaces se considerará el flujo proporcional a la diferencia entre las temperaturas de las secciones en cuestión. Además se considerará que el flujo se mantiene constante sobre los puntos de enlace. Denotando como ξ_k los puntos de encuentro de dos secciones consecutivas, estas condiciones quedan como

$$a(\xi)\frac{d}{d\xi}u(\xi)\Big|_{\xi_k^-} = \beta_k[\![u(\xi)]\!]_{\xi_k}, \tag{3}$$

y

$$\Big[\!\Big[a(\xi)\frac{d}{d\xi}u(\xi)\Big]\!\Big]_{\xi_k} = 0. \tag{4}$$

Las constantes β_k son conocidas como números de Biot. Para las condiciones de frontera, que en el caso unidimensional serán dadas sobre los puntos extremos de Ω, serán seleccionadas condiciones de Dirichlet, $u(\xi_0) = u_0$ y $u(\xi_f) = u_f$. De este modo, con la ecuación (2) junto con las condiciones (3) y (4), además de las condiciones de frontera, queda establecido el problema a tratar numéricamente por el método de las diferencias finitas.

Bajo determinadas condiciones de diferenciabilidad de las funciones $a(\xi)$ y $f(\xi)$, y de positividad de $a(\xi)$, el problema (2)-(4) posee solución diferenciable hasta determinado orden, cuya expresión analítica, para $\xi \in (\xi_j, \xi_{j+1})$, es

$$u(\xi) = -\int_{\xi_0}^{\xi}\frac{1}{a(s)}\Big(\int_{\xi_0}^{s} f(t)dt - c\Big)ds - \sum_{i=1}^{j}\frac{1}{\beta_j}\Big(\int_{\xi_0}^{\xi_j} f(\xi)d\xi - c\Big) + u_0, \tag{5}$$

con

$$c = \frac{\Big[u_f - u_0 + \langle\frac{1}{a(\xi)}\Big(\int_{\xi_0}^{\xi} f(s)ds\Big)\rangle + \sum_{i=1}^{l}\frac{1}{\beta_j}\int_{\xi_0}^{\xi_j} f(\xi)d\xi\Big]}{\Big(\langle\frac{1}{a(\xi)}\rangle + \sum_{i=1}^{l}\frac{1}{\beta_j}\Big)},$$

donde

$$\langle F(\xi)\rangle = \int_{\xi_0}^{\xi_f} F(\xi)d\xi.$$

Esta expresión de $u(\xi)$ garantiza su acotación, así como la de sus derivadas.

Nótese que las integrales presentes en (5) podrían no ser expresables en términos de funciones elementales, lo cual representa otro factor para analizar numéricamente el problema (2)-(4).

1. Obtención del sistema de ecuaciones lineales para el problema con condiciones de discontinuidad

1.1 Formulación del problema

Sean $\{\xi_k\}_{k=1}^{l}$ números reales tales que

$$\xi_0 = 0 < \xi_1 < \xi_2 < \ldots < \xi_l < \xi_{l+1} = 1.$$

Sean además $f(\xi)$ y $a(\xi)$ funciones reales infinitamente diferenciables para todo $\xi \in [\xi_i, \xi_{i+1}]$, $i = 0, \ldots, l$, con $a(\xi)$ satisfaciendo la desigualdad

$$0 < \alpha_0 \leq a(\xi) \leq \alpha_1$$

(donde α_0 y α_1 son constantes reales). Considérese también el operador contraste definido como

$$[\![F(\xi)]\!]_{\xi_i} = \lim_{\xi\to\xi_i^+} F(\xi) - \lim_{\xi\to\xi_i^-} F(\xi).$$

El objetivo perseguido es aplicar el método de las diferencias finitas al problema

$$-\frac{d}{d\xi}\Big(a(\xi)\frac{du}{d\xi}\Big) = f(\xi),\ \xi \in \Omega^* = (0,1)\setminus\{\xi_k\}, \tag{6}$$

$$a(\xi)\frac{du}{d\xi}\Big|_{\xi_k^-} = \beta_k[\![u(\xi)]\!]_{\xi_k}, \tag{7}$$

$$[\![a(\xi)\frac{du}{d\xi}]\!]_{\xi_k} = 0, \tag{8}$$

con las condiciones de frontera $u(0) = t_0$ y $u(1) = t_1$, donde $t_0, t_1, \beta_i \in \mathbb{R}$, $\beta_k > 0$ para todo $k = 1, \ldots, l$.

1.2 Simplificación del problema

Para abordar el problema numéricamente, resultaría provechoso que los puntos de discontinuidad formaran parte de la malla que será seleccionada, para de este modo utilizar la información brindada por las condiciones (7)-(8). Esto impediría en muchos casos que dicha malla forme una red uniforme. La demostración de este hecho será dada a continuación:

Sea $\hat{\xi}$ un punto del intervalo $(0,1)$. Suponiendo que existe una red uniforme de dicho intervalo, que contiene a $\hat{\xi}$, entonces existen naturales p y q, y un número real h, tal que $ph = \hat{\xi}$ y $qh = 1$, consecuentemente, $\hat{\xi}$ puede ser expresado como $\xi_0 = p/q$. Esta expresión conduce a una contradicción en el caso en que $\hat{\xi}$ sea irracional. En conclusión, dado un conjunto de números reales en el intervalo $(0,1)$, no siempre es posible encontrar una red uniforme que contenga a dichos puntos como nodos.

Trabajar con una malla uniforme, simplifica enormemente las expresiones obtenidas para las aproximaciones de las derivadas de las funciones en cuestión. Por esta razón será mostrado a continuación como transformar el problema (6)-(8), donde los puntos de discontinuidad no son equidistantes entre si, en un problema equivalente donde los puntos de discontinuidad formen una red uniforme.

Para ello considérese la función $g(x)$, lineal por tramos y continua, obtenida de unir los puntos $x_i = \frac{i}{l+1}$ con los puntos ξ_i, para $i = 0,\ldots,l+1$. La expresión analítica para esta función es

$$g(x) = (\xi_{i+1}-\xi_i)(l+1)x+\xi_i(i+1)-i\xi_{i+1} \text{ si } \frac{i}{l+1} < x < \frac{i+1}{l+1},$$

para $i = 0,\ldots,l$. Esta función es positiva, diferenciable en todo el intervalo $(0,1)$, excepto por los puntos $\frac{i}{l+1}$, $i = 1,\ldots,l$, y en los puntos donde la derivada existe, es mayor que cero. Definiendo las funciones $a_1(x) = \frac{a(g(x))}{g'(x)}$ y $u_1(x) = u(g(x))$, se tiene entonces que

$$\begin{aligned}
&-\frac{d}{dx}\Big(a_1(x)\frac{d}{dx}u_1(x)\Big)\\
&=-\frac{d}{dx}\Big(\frac{a(g(x))}{g'(x)}\frac{d}{d\xi}u(\xi)\Big|_{\xi=g(x)}g'(x)\Big)\\
&=-\frac{d}{dx}\Big(a(g(x))\frac{d}{d\xi}u(\xi)\Big|_{\xi=g(x)}\Big)\\
&=-\frac{d}{d\xi}\Big(a(\xi)\frac{d}{d\xi}u(\xi)\Big)\Big|_{\xi=g(x)}g'(x).
\end{aligned}$$

Consequentemente, sabiendo que $u(\xi)$ es solución de (6), entonces $u_1(x)$ es solución de la ecuación

$$-\frac{d}{dx}\Big(a_1(x)\frac{d}{dx}u_1(x)\Big) = f(g(x))g'(x) = f_1(x).$$

Además, como $\xi_i = g(x_i)$ también se cumple que

$$\begin{aligned}
\Big(a_1(x)\frac{d}{dx}u_1(x)\Big)\Big|_{x=x_i^-} &= \Big(\frac{a(g(x))}{g'(x)}\frac{d}{d\xi}u(\xi)\Big|_{\xi=g(x)}g'(x)\Big)\Big|_{x=x_i^-}\\
&= \Big(a(g(x))\frac{d}{d\xi}u(\xi)\Big|_{\xi=g(x)}\Big)\Big|_{x=x_i^-}\\
&= \Big(a(g(x_i^-))\frac{d}{d\xi}u(\xi)\Big|_{\xi=g(x_i^-)}\Big)\\
&= \Big(a(\xi_i^-))\frac{d}{d\xi}u(\xi)\Big|_{\xi=\xi_i^-}\Big)\\
&= \beta_i[\![u(\xi)]\!]_{\xi_i}\\
&= \beta_i[\![u_1(x)]\!]_{x_i}.
\end{aligned}$$

De igual modo puede ser comprobado que

$$[\![a_1(x)\frac{du}{dx}]\!]_{x_i^-} = [\![a(\xi)\frac{du}{d\xi}]\!]_{\xi_k^-} = 0.$$

De este modo, ha sido demostrado que el problema (6)-(8) siempre puede ser reducido a un problema de la forma

$$-\frac{d}{dx}\Big(a_1(x)\frac{du_1}{dx}\Big) = f_1(x),\ x\in\Omega^* = (0,1)\setminus\{x_i = \frac{i}{l+1}\}, \tag{9}$$

$$a_1(x)\frac{du_1}{dx}\Big|_{x_i^-} = \beta_i[\![u_1(x)]\!]_{x_i}, \tag{10}$$

$$[\![a_1(x)\frac{du_1}{dx}]\!]_{x_i} = 0, \tag{11}$$

con las condiciones de frontera $u_1(0) = u(g(0)) = u(0) = t_0$ y $u_1(1) = t_1$, donde los coeficientes del nuevo problema satisfacen todas las condiciones de diferenciabilidad por tramos (y positividad en el caso de $a_1(x)$), que el problema original. Además, las discontinuidades de los coeficientes del nuevo problema conforman una red uniforme en el intervalo $(0,1)$, de paso $\frac{1}{l+1}$. Esto justifica que, sin perdida de generalidad, el análisis numérico se realice solo para problemas similares a (9)-(11).

Observación 1 *Una vez obtenida la solución $u_1(x)$ de (9)-(11), es posible recuperar la solución $u(\xi)$ de (6)-(8) a través de la relación $u(\xi) = u_1(g^{-1}(\xi))$, donde $g^{-1}(\xi)$ representa la función inversa de $g(x)$, la cual, como fue comentado con anterioridad, existe y al ser una función lineal por tramo, es bastante sencilla de determinar.*

Observación 2 *La simplificación realizada es equivalente a considerar en el problema original, en cada intervalo (ξ_i,ξ_{i+1}), un paso distinto $h_i = \frac{\xi_{i+1}-\xi_i}{n}$, $n\in\mathbb{N}$ al elegir la malla.*

1.3 Selección de la malla y obtención del esquema en diferencias finitas

Sean $\{x_i\}_{i=1}^l$ los puntos de discontinuidad de los coeficientes del problema (9)-(11), los cuales conforman una red uniforme de paso $\frac{1}{l+1}$. Resulta conveniente entonces elegir una malla uniforme que contenga a los puntos x_i entre sus elementos. Para ello basta con elegir un paso $h = \frac{1}{n(l+1)}$, con $n\in\mathbb{N}$. De modo general, la expresión para los elementos de la malla es

$$\eta_j = \frac{j}{n(l+1)},$$

así, cuando $j = kn$, $k = 1,\ldots,l$, entonces η_j será el k-ésimo punto de discontinuidad. Como en dichos puntos la solución tendrá valores distintos a la derecha y a la izquierda, es conveniente entonces, para estos valores de j, definir 2 nodos, en vez de 1, denotados como η_j^- y η_j^+. Esta selección de h, junto con las consideraciones adicionales hechas para los puntos de discontinuidad, dará como resultado una cantidad de nodos igual a $m = n(l+1)+l-1$ (sin incluir los extremos del intervalo $\eta_0 = 0$ y $\eta_{n(l-1)} = 1$). Estos nodos en lo adelante serán llamados nodos principales. Para simplificar la notación, en lo adelante se considerará $u_j = u_1(\eta_j)$, si $j\neq kn$, $u_j^- = u_1(\eta_j^-)$ y $u_j^+ = u_1(\eta_j^+)$ si $j = kn$ para $k = 1,\ldots,l$. Para abordar numéricamente el problema se definirán además los

nodos auxiliares $\eta_{j+1/2}$, como los puntos medios de los nodos definidos con anterioridad, es decir

$$\eta_{j+1/2} = \frac{\eta_j + \eta_{j+1}}{2}, j = 0, \ldots, n(l+1) - 1.$$

Estos nodos son meramente simbólicos, ya que no estarán incluidos en las ecuaciones que serán obtenidas. Su inclusión se debe a que servirán de apoyo para calcular las aproximaciones de las derivadas.
La ecuación (9) puede ser reescrita como

$$-(J(u_1)(x))' = f_1(x), x \in \Omega^*,$$

donde $J(u_1)(x)$ es el flujo asociado a $u_1(x)$, definido como

$$J(u_1)(x) = a_1(x)u_1'(x).$$

Para abreviar, en lo adelante se considerará

$$J(u_1)(\eta_{j+1/2}) = J_{j+1/2}.$$

Esto permite aproximar la derivada del flujo en los puntos η_j, si $j \neq kn$, a través de la fórmula en diferencias centrales para la derivada:

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} + \frac{h^2}{12}(u'''(x+\theta_2) - u'''(x-\theta_1)), \tag{12}$$

con $0 < \theta_1, \theta_2 < h$, y con las fórmulas en diferencias hacia adelante y hacia atrás para las derivadas en los puntos η_j^- y η_j^+ respectivamente, cuando $j = kn, k = 1, \ldots, l$:

$$u'(x^-) = \frac{u(x^-) - u(x-h)}{h} - \frac{h}{2}u''(x-\theta_1), \tag{13}$$

$$u'(x^+) = \frac{u(x+h) - u(x^+)}{h} + \frac{h}{2}u''(x+\theta_2). \tag{14}$$

De este modo

$$-(J(u_1)(\eta_j))' \approx -\frac{J_{j+1/2} - J_{j-1/2}}{h}, \text{ si } j \neq kn, \tag{15}$$

$$-(J(u_1)(\eta_j^-))' \approx -\frac{J(u_1)(\eta_j^-) - J_{j-1/2}}{h/2}, \text{ si } j = kn, \tag{16}$$

$$-(J(u_1)(\eta_j^+))' \approx -\frac{J_{j+1/2} - J(u_1)(\eta_j^+)}{h/2}, \text{ si } j = kn. \tag{17}$$

Utilizando nuevamente la expresión (12) para aproximar el flujo $J_{j+1/2}$, y las condiciones (10) y (11) para $J(u_1)(\eta_j^-)$ y $J(u_1)(\eta_j^+)$ respectivamente se tiene que

$$J_{j+1/2} = a_1(\eta_{j+1/2})u_1'(\eta_{j+1/2}) \approx a_{j+1/2}\frac{u_{j+1} - u_j}{h}, \tag{18}$$

$$J(u_1)(\eta_j^-) = J(u_1)(\eta_j^+) = a_1(\eta_j^-)u_1'(\eta_j^-) = \beta_k(u_j^+ - u_j^-). \tag{19}$$

Sustituyendo (18) y (19) en (15), (16) y (17) se llega a las ecuaciones lineales buscadas

$$h^{-2}(-a_{j-\frac{1}{2}}u_{j-1} + (a_{j-\frac{1}{2}} + a_{j+\frac{1}{2}})u_j - a_{j+\frac{1}{2}}u_{j+1}) = f(\eta_j), \tag{20}$$

$$-\frac{a_{j-\frac{1}{2}}}{h^2}u_{j-1} + (\frac{a_{j-\frac{1}{2}}}{h^2} + \frac{\beta_k}{h})u_j^- - \frac{\beta_k}{h}u_j^+ = f(\eta_j^-)/2, \tag{21}$$

$$-\frac{\beta_k}{h}u_j^- + (\frac{a_{j+\frac{1}{2}}}{h^2} + \frac{\beta_k}{h})u_j^+ - \frac{a_{j+\frac{1}{2}}}{h^2}u_{j+1} = f(\eta_j^+)/2. \tag{22}$$

La matriz asociada a este sistema es tridiagonal, simétrica y en la próxima sección será visto que también es definida positiva. Nótese que los nodos auxiliares introducidos solo aparecen para evaluar la función $a_1(x)$, la cual es conocida. Las componentes del vector independiente $b^T = (b_1, \ldots, b_{n(l+1)})$ son

$$\begin{aligned} b_1 &= f(\eta_1) + h^{-2}a_{1/2}t_0, \\ b_j &= f(\eta_j), \text{ si } j \neq kn, \\ b_j^- &= f(\eta_j^-)/2, \text{ si } j = kn, \\ b_j^+ &= f(\eta_j^+)/2, \text{ si } j = kn, \\ b_{n(l+1)} &= f(\eta_{n(l+1)}) + h^{-2}a_{n(l+1)+1/2}t_1. \end{aligned}$$

1.4 Un ejemplo

A continuación serán aplicados los resultados de la subsección anterior a un ejemplo específico. Considérese la función $a(x)$ definida como

$$a(x) = \begin{cases} a_1 & \text{si} \quad 0 < x < 1/2, \\ a_2 & \text{si} \quad 1/2 < x < 1, \end{cases}$$

con $a_1, a_2 > 0$. El problema a analizar numéricamente es

$$\begin{aligned} -\frac{d}{dx}\Big(a(x)\frac{du}{dx}\Big) &= 0, \ x \in (0,1) \setminus \{1/2\}, \\ a(x)\frac{du}{dx}\Big|_{(1/2)^-} &= \beta[\![u(x)]\!]_{1/2}, \\ [\![a(x)\frac{du}{dx}]\!]_{1/2} &= 0, \end{aligned}$$

donde $\beta > 0$. Además se considerarán las condiciones de frontera $u(0) = t_0$ y $u(1) = t_1$. La malla utilizada para dar solución al problema tendrá paso $h = 1/4$. De este modo, dicha malla estará conformada por los valores $(0, 1/4, (1/2)^-, (1/2)^+, 3/4, 1)$. Los nodos auxiliares son los puntos

$$\eta_{1/2} = 1/8, \ \eta_{3/2} = 3/8, \ \eta_{5/2} = 5/8, \ \eta_{7/2} = 7/8,$$

de modo que

$$\begin{aligned} a_{1/2} &= a(1/8) = a_1, \\ a_{3/2} &= a(3/8) = a_1, \\ a_{5/2} &= a(5/8) = a_2, \\ a_{7/2} &= a(7/8) = a_2. \end{aligned}$$

De acuerdo a esta información, el sistema de ecuaciones lineales derivado de aplicar el método de las diferencias finitas es

$$\begin{pmatrix} 8a_1 & -4a_1 & 0 & 0 \\ -4a_1 & 4a_1+\beta & -\beta & 0 \\ 0 & -\beta & 4a_2+\beta & -4a_2 \\ 0 & 0 & -4a_2 & 8a_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 4a_1t_0 \\ 0 \\ 0 \\ 4a_2t_1 \end{pmatrix}$$

Nótese que la primera y la última ecuación podrían ser divididas entre $4a_1$ y $4a_2$ respectivamente, pero no se ha realizado esta simplificación para conservar la simetría de la matriz. Los cálculos necesarios para dar solución a este sistema son bastante extensos, no obstante conducen a la solución

$$\begin{aligned} u_1 &= \frac{c}{4a_1}+t_0, \\ u_2 &= \frac{c}{2a_1}+t_0, \\ u_3 &= \frac{c}{2a_2}+c\Big(\frac{1}{\beta}+(\frac{1}{a_1}-\frac{1}{a_2})\frac{1}{2}\Big)+t_0, \\ u_4 &= \frac{3c}{4a_2}+c\Big(\frac{1}{\beta}+(\frac{1}{a_1}-\frac{1}{a_2})\frac{1}{2}\Big)+t_0, \end{aligned}$$

donde

$$c = \frac{t_1-t_0}{\frac{1}{2}(a_1^{-1}+a_2^{-1})+\beta^{-1}}.$$

Aun más, es posible comprobar que estos valores coinciden con los de la solución exacta, la cual es

$$u(x) = \begin{cases} \frac{c}{a_1}x+t_0 & \text{si} \quad 0<x<1/2, \\ \frac{c}{a_2}x+c\Big(\frac{1}{\beta}+(\frac{1}{a_1}-\frac{1}{a_2})\frac{1}{2}\Big)+t_0 & \text{si} \quad 1/2<x<1. \end{cases}$$

La coincidencia se debe a que la solución exacta está conformada por rectas, por lo que las aproximaciones para sus derivadas coinciden con las derivadas reales, haciendo que todos los valores obtenidos sean también exactos.

2. Consistencia, estabilidad y convergencia

En lo adelante, el problema (9)-(11) será representado como

$$L(u_1,f_1) = -(J(u_1)(x))' - f_1(x) = 0,$$

mientras que el problema obtenido por el método de las diferencias finitas, para un determinado valor de h será denotado como

$$L_h(u_h,f_h) = -A_{df}u_h - f_h = 0, \tag{23}$$

donde $-A_{df}$ es la matriz del sistema y f_h es el vector independiente. Esta nueva notación ayudará en el análisis de la consistencia y la estabilidad del método de las diferencias finitas aplicado al problema de valores en la frontera con condiciones de discontinuidad.

2.1 Estabilidad

En el segundo capítulo de [1] es dado el siguiente concepto de *estabilidad*

Definición 1 *Un método numérico está bien planteado, (o es estable) si para cualquier h, existe una única solución, dependiente continuamente de los datos.*

Según esta definición, para demostrar la estabilidad del método aquí utilizado, es necesario demostrar que la solución u_h de (23) es única, y además que depende continuamente de los datos del problema. Nótese que la matriz del sistema, $-A_{df}$, es tridiagonal, simétrica, $a_{ii} > 0$ para todo i,

$$a_{i(i-1)} = a_{(i-1)i} < 0$$

para todo $i \geqslant 2$ y $a_{(i+1)i}+a_{ii}+a_{i(i+1)} = 0$ para $i = 2,\ldots,m-1$, donde $m = n(l+1)+l-1$ es la dimensión de la matriz $-A_{df}$. Además $a_{11}+a_{12} > 0$ y $a_{m(m-1)}+a_{mm} > 0$. Será demostrado a continuación que una matriz con estas condiciones, es definida positiva, es decir,

$$(v_h, -A_{df}v_h) > 0, \ \forall v_h \neq 0,$$

donde $(\cdot,\cdot)$ es el producto escalar usual en $\mathbb{R}^m$ y v_h una función definida sobre los nodos de la malla del problema. En efecto, al ser A_{df} tridiagonal y simétrica, entonces

$$(v_h, -A_{df}v_h) = \sum_{i=1}^{m} a_{ii}v_i^2 + \sum_{i=1}^{m-1} 2a_{i(i+1)}v_iv_{i+1}. \tag{24}$$

Utilizando la propiedad $a_{(i+1)i}+a_{ii}+a_{i(i+1)} = 0$ para $i = 2,\ldots,m-1$, el miembro derecho de (24) puede reescribirse como

$$\begin{aligned} &(v_h, -A_{df}v_h) \\ &\quad = (a_{11}+a_{12})v_1^2 + (a_{m(m-1)}+a_{mm})v_m^2 \\ &\qquad + \sum_{i=1}^{m-1}(-a_{i(i+1)}v_i^2 + 2a_{i(i+1)}v_iv_{i+1} - a_{i(i+1)}x_{i+1}^2) \\ &\quad = (a_{11}+a_{12})v_1^2 + (a_{m(m-1)}+a_{mm})v_m^2 \\ &\qquad - \sum_{i=1}^{m-1} a_{i(i+1)}(v_i-v_{i+1})^2. \end{aligned} \tag{25}$$

Finalmente, usando las condiciones $a_{i(i+1)} < 0$, $a_{11}+a_{12} > 0$ y $a_{m(m-1)}+a_{mm} > 0$ se puede concluir que $-A_{df}$ es definida positiva. Esto implica que es inversible, por lo tanto la solución de (23) existe y es única para cada valor de h. Además, sustituyendo los valores de a_{11}, a_{mm} y $a_{i(i+1)}$ en (25) se obtiene la siguiente expresión

$$\begin{aligned} (v_h, -A_{df}v_h) =& h^{-2}a_{1/2}v_1^2 + h^{-2}a_{m+1/2}v_m^2 \\ &+ h^{-2}\sum_{i\in I}^{m-1} a_{i+1/2}(v_i-v_{i+1})^2 \\ &+ h^{-1}\sum_{k=1}^{l} \beta_k(v_{kn}^+ - v_{kn}^-)^2, \end{aligned} \tag{26}$$

donde $I=\{i : i\neq kn \wedge x_i = x_{nk}^+\}$. La expresión (26) permite definir una norma en V_h, el espacio de las funciones definidas sobre los nodos de la malla. Dicha norma será denotada en lo adelante como $\|\cdot\|_A$ y viene dada por la expresión

$$\|v_h\|_A^2 = (v_h, -A_{df}v_h).$$

Considérese además la siguiente norma en V_h

$$\|v_h\|_h^2 = (v_h, v_h) = \sum_{i=1}^{m} v_i^2.$$

Entonces existe una constante c, independiente de h, tal que

$$\|v_h\|_h \leqslant c\|v_h\|_A,\ \ \forall v_h \in V_h.$$

En efecto, teniendo en cuenta que $a_{j+1/2} = a_1(\eta_{j+1/2})$, que $a_1(x)$ es mayor que una constante positiva α_0 para cualquier valor de x, que las constantes β_k son positivas y por lo tanto tienen un mínimo β_0, entonces, denotando $c_0 = (\text{mín}\{\alpha_0, \beta_0\})^{1/2}$, se tiene que, para cualquier valor de i

$$\begin{aligned} c_0 v_j =& c_0\Big(v_1 + \sum_{i=1}^{j-1}(v_{i+1}-v_i)\Big) \\ =& c_0\Big(v_1 + \sum_{i\in I'}(v_{i+1}-v_i) + \sum_{i\in K'}(v_i^+ - v_i^-)\Big), \end{aligned}$$

donde $I' = \{i \in I, i \leqslant j-1\}$ y $K' = \{i = kn, \leqslant j-1\}$. Elevando al cuadrado ambos miembros y utilizando la desigualdad de Minkowski

$$\Big(\sum_{i=1}^{d} p_i\Big)^2 \leqslant d\sum_{i=1}^{d} p_i^2,$$

la cual se cumple para cualquier entero $d \geqslant 1$ y cualquier sucesión de números reales $\{p_i\}_{i=1}^{d}$, se tiene que

$$\begin{aligned} & c_0^2 v_j^2 \\ & \leqslant 3\Big(c_0^2 v_1^2 + j\sum_{i\in I'}^{j-1} c_0^2(v_{i+1}-v_i)^2 + l\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big) \\ & \leqslant 3\Big(a_{1/2}v_1^2 + j\sum_{i\in I} a_{i+1/2}(v_{i+1}-v_i)^2 + l\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big). \end{aligned}$$

Sumando desde $j=1$ hasta m

$$\begin{aligned} c_0^2\sum_{j=1}^{m} v_j^2 \leqslant& 3\Big(ma_{1/2}v_1^2 + \sum_{i=1}^{m} j\sum_{i\in I} a_{i+1/2}(v_i - v_{i+1})^2 \\ & + ml\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big) \\ =& 3\Big(ma_{1/2}v_1^2 + \frac{m(m+1)}{2}\sum_{i\in I} a_{i+1/2}(v_i - v_{i+1})^2 \\ & + ml\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big). \end{aligned}$$

Teniendo en cuenta que $m = n(l+1)+l-1$ y que $h = \frac{1}{n(l+1)}$ y por lo tanto $hm = 1 + \frac{l-1}{n(l+1)} \leqslant 2$, para todo $n \in \mathbb{N}$

$$\begin{aligned} & c_0^2\sum_{j=1}^{m} v_j^2 \\ & \quad \leqslant 12l\Big(h^{-1}a_{1/2}v_1^2 + h^{-2}\sum_{i\in I} a_{i+1/2}(v_i - v_{i+1})^2 \\ & \qquad + h^{-1}\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big) \\ & \quad \leqslant 12l\Big(h^{-2}a_{1/2}v_1^2 + h^{-2}a_{m+1/2}v_m^2 \\ & \qquad + h^{-2}\sum_{i\in I} a_{i+1/2}(v_i - v_{i+1})^2 + h^{-1}\sum_{k=1}^{l}\beta_k(v_{kn}^+ - v_{kn}^-)^2\Big). \end{aligned}$$

Consecuentemente, para toda $v_h \in V_h$

$$\|v_h\|_h^2 \leqslant c\|v_h\|_A^2, \tag{27}$$

con $c = \frac{12l}{\text{mín}\{\alpha_0,\beta_0\}}$, independiente de h. Finalmente, para demostrar la estabilidad del método, basta con tomar la ecuación

$$-A_{df}u_h = f_h,$$

y multiplicarla escalarmente por u_h. De este modo se obtiene la relación

$$\|u_h\|_A^2 = (u_h, -A_{df}u_h) = (f_h, u_h).$$

Utilizando la desigualdad (27) y la desigualdad de Cauchy-Schwarz, se obtiene

$$\|u_h\|_h^2 \leqslant c\|f_h\|_h\|u_h\|_h,$$

y luego de simplificar un factor $\|u_h\|_h$

$$\|u_h\|_h \leqslant c\|f_h\|_h,$$

por lo tanto el método es estable, ya que la solución además de ser única, depende continuamente de los datos del problema.

2.2 Consistencia

Para demostrar la consistencia del método, es necesario probar que

$$L_h(u_1, f_1) = L_h(u_1, f_1) - L(u_1, f_1) \longrightarrow 0, \text{ si } h \to 0.$$

A partir de la aproximación (12), se tiene que

$$\begin{aligned} u_1'(x) =& \frac{u_1(x+h/2) - u_1(x-h/2)}{h} \\ & + \frac{h^2}{48}(u_1'''(x+\theta_2) + u_1'''(x-\theta_1)) \end{aligned}$$

con $0 < \theta_1, \theta_2 < h/2$, por lo tanto, la acotación de la solución exacta del problema y de sus derivadas implican que

$$\frac{u_1(x+h/2) - u_1(x-h/2)}{h} = u_1'(x) + h^2 R^1,$$

con $|R^1| \leqslant \frac{1}{24}\sup_{[0,1]}|u_1'''(x)|$. Sea $j \neq kn$, entonces

$$\begin{aligned}
&L_h(u_1,f_1)(\eta_j)\\
=&-\frac{a_{j+\frac{1}{2}}\frac{u_1(\eta_{j+1})-u_1(\eta_j)}{h}-a_{j-\frac{1}{2}}\frac{u_1(\eta_j)-u_1(\eta_{j-1})}{h}}{h}-f(\eta_j)\\
=&-\frac{a_{j+\frac{1}{2}}u_1'(\eta_{j+\frac{1}{2}})-a_{j-\frac{1}{2}}u_1'(\eta_{j-\frac{1}{2}})}{h}-f(\eta_j)\\
&-h(a_{j+\frac{1}{2}}R^1_{j+1}-a_{j-\frac{1}{2}}R^1_j)\\
=&-\frac{J(u_1)(\eta_{j+\frac{1}{2}})-J(u_1)(\eta_{j-\frac{1}{2}})}{h}-f(\eta_j)\\
&-h(a_{j+\frac{1}{2}}R^1_{j+1}-a_{j-\frac{1}{2}}R^1_j).
\end{aligned}$$

Utilizando nuevamente la aproximación (12), esta vez para el flujo

$$\begin{aligned}
&L_h(u_1,f_1)(\eta_j)\\
=&-\frac{J(u_1)(\eta_{j+\frac{1}{2}})-J(u_1)(\eta_{j-\frac{1}{2}})}{h}-f(\eta_j)\\
&-h(a_{j+\frac{1}{2}}R^1_{j+1}-a_{j-\frac{1}{2}}R^1_j)\\
=&-\Big(J(u_1)(\eta_j)\Big)'-f(\eta_j)-h^2R^2_j-h(a_{j+\frac{1}{2}}R^1_{j+1}-a_{j-\frac{1}{2}}R^1_j)\\
=&-h^2R^2_j-h(a_{j+\frac{1}{2}}R^1_{j+1}-a_{j-\frac{1}{2}}R^1_j),
\end{aligned}$$

con

$$|R^2_j| \leqslant \frac{1}{24}\sup_{[0,1]}|\Big(J(u_1)\Big)'''| = \frac{1}{24}\sup_{[0,1]}|f''|,$$

de modo que

$$|L_h(u_1,f_1)(\eta_j)| \leqslant \frac{h^2}{24}\sup_{[0,1]}|f''| + \frac{h\alpha_1}{12}\sup_{[0,1]}|u_1'''(x)|,$$

o equivalentemente, para h suficientemente pequeño

$$|L_h(u_1,f_1)(\eta_j)| \leqslant \frac{h\alpha_1}{6}\sup_{[0,1]}|u_1'''(x)|,$$

donde α_1 es el supremo de la función $a(x)$.
De igual manera, las aproximaciones (13) y (14) implican que

$$\begin{aligned}
\frac{u_1(x^-)-u_1(x-h/2)}{h/2} &= u_1'(x^-)+hR^3,\\
\frac{u_1(x+h/2)-u_1(x^+)}{h/2} &= u_1'(x^+)-hR^3.
\end{aligned}$$

con

$$|R^3| \leqslant \frac{1}{4}\sup_{[0,1]}|u_1''(x)|.$$

Luego, para $j = kn$

$$\begin{aligned}
&2L_h(u_1,f_1)(\eta_j^-)\\
=&-\frac{\beta_k(u_1(\eta_j^+)-u_1(\eta_j^-))-a_{j-1/2}\frac{u_1(\eta_j^-)-u_1(\eta_{j-1})}{h}}{h/2}-f(\eta_j^-)\\
=&-\frac{J(u_1)(\eta_j^-)-J(u_1)(\eta_{j-1/2})-a_{j-1/2}h^2R^1_j}{h/2}-f(\eta_j^-)\\
=&-\frac{J(u_1)(\eta_j^-)-J(u_1)(\eta_{j-1/2})}{h/2}-f(\eta_j^-)+2a_{j-1/2}hR^1_j\\
=&-\Big(J(u_1)(\eta_j^-)\Big)'-f(\eta_j^-)-2hR^3_j-2a_{j-1/2}hR^1_j\\
=&-2hR^3_j-2a_{j-1/2}hR^1_j.
\end{aligned}$$

Una aproximación similar puede hacerse para $L_h(u_1,f_1)(\eta_j^+)$, asegurando que

$$\begin{aligned}
|L_h(u_1,f_1)(\eta_j^-)| &\leqslant h(\frac{1}{4}\sup_{[0,1]}|u_1''(x)| + \frac{\alpha_1}{24}\sup_{[0,1]}|f''|),\\
|L_h(u_1,f_1)(\eta_j^+)| &\leqslant h(\frac{1}{4}\sup_{[0,1]}|u_1''(x)| + \frac{\alpha_1}{24}\sup_{[0,1]}|f''|).
\end{aligned}$$

Por lo tanto, se ha demostrado el resultado requerido, ya que, tomando la norma $\|\cdot\|_h$ definida en el análisis de la estabilidad

$$\|v_h\|_h^2 = \sum_{i=1}^m v_i^2,$$

se tiene que

$$\begin{aligned}
&\|L_h(u_1,f_1)\|_h^2\\
=&\sum_{i=1}^m (L_h(u_1,f_1)(\eta_i))^2\\
=&\sum_{j\neq nk}\Big(L_h(u_1,f_1)(\eta_j)\Big)^2\\
&+\sum_{j=nk}\Big(L_h(u_1,f_1)(\eta_j^-)\Big)^2+\Big(L_h(u_1,f_1)(\eta_j^+)\Big)^2\\
\leqslant& m\frac{h^2\alpha_1^2}{36}\sup_{[0,1]}|u_1'''(x)|^2\\
&+4lh^2(\frac{1}{16}\sup_{[0,1]}|u_1''(x)|^2+\frac{\alpha_1^2}{576}\sup_{[0,1]}|f''|^2).
\end{aligned}$$

Nuevamente, teniendo en cuenta que $mh \leqslant 2$,

$$\begin{aligned}
\|L_h(u_1,f_1)\|_h^2 \leqslant& \frac{h\alpha_1^2}{18}\sup_{[0,1]}|u_1'''(x)|^2\\
&+4lh^2\Big(\frac{1}{16}\sup_{[0,1]}|u_1''(x)|^2+\frac{\alpha_1^2}{576}\sup_{[0,1]}|f''|^2\Big),
\end{aligned}$$

o equivalentemente para h suficientemente pequeño

$$\|L_h(u_1,f_1)\|_h^2 \leqslant \frac{h\alpha_1^2}{9}\sup_{[0,1]}|u_1'''(x)|^2.$$

Por lo tanto, el método es consistente, ya que.

$$\|L_h(u_1,f_1)\|_h \longrightarrow 0 \text{ si } h \to 0.$$

2.3 Convergencia

Introduciendo la función de error global e_h, definida sobre los nodos de la malla como

$$e(\eta_j) = u_1(\eta_j) - u_h(\eta_j),$$

se puede apreciar que dicha función satisface las relaciones

$$-A_{df}e_h = -A_{df}u_1 - (-A_{df}u_h) = -A_{df}u_1 - f_h = L_h(u_1, f_1).$$

Teniendo en cuenta que el método es estable, la función e_h, también cumple que

$$\|e_h\|_h \leqslant c\|L_h(u_1, f_1)\|_h.$$

Finalmente la consistencia del método y la independencia de h de la constante c implican

$$\|e_h\|_h^2 \leqslant c^2 \frac{h\alpha_1^2}{9} \sup_{[0,1]} |u_1'''(x)|^2 \longrightarrow 0 \text{ si } h \to 0.$$

Concluyendo, la solución u_h converge hacia la solución exacta u_1 cuando h tiende a cero, siempre y cuando esta última sea al menos tres veces continuamente diferenciable por tramos en el intervalo $(0,1)$.

3. Algunos ejemplos de la aplicación del método

Considérese la función

$$a(x) = \begin{cases} 1 & \text{si} \quad 0 < x < 1/8, \\ x^2 & \text{si} \quad 1/8 < x < 3/5, \\ x^3 & \text{si} \quad 3/5 < x < 1. \end{cases}$$

Entonces, el problema

$$-\frac{d}{dx}\left(a(x)\frac{du}{dx}\right) = 0,\ x \in \Omega^* = (0,1) \setminus \{1/8, 3/5\}, \quad (28)$$

$$a(x)\frac{du}{dx}\Big|_{x_i^-} = [\![u(x)]\!]_{x_i}, \quad (29)$$

$$[\![a(x)\frac{du}{dx}]\!]_{x_i} = 0, \quad (30)$$

con las condiciones de frontera $u(0) = 1$ y $u(1) = 2$, posee como solución exacta a

$$u(x) = \begin{cases} \frac{72}{673}x + 1 & \text{si} \quad 0 < x < 1/8, \\ -\frac{72}{673}\frac{1}{x} + \frac{1330}{673} & \text{si} \quad 1/8 < x < 3/5, \\ -\frac{36}{673}\frac{1}{x^2} + \frac{1382}{673} & \text{si} \quad 3/5 < x < 1. \end{cases}$$

La Figura 1 muestra el gráfico de dicha solución.

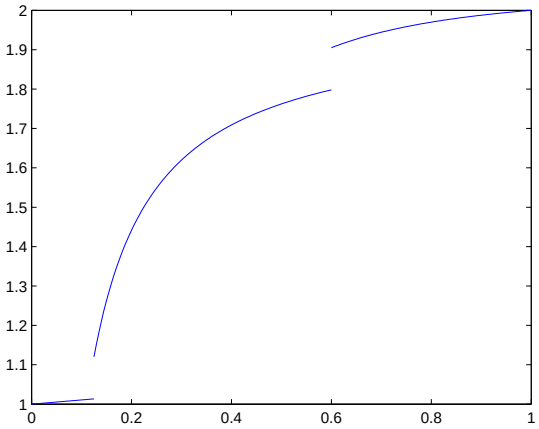


Figura 1. Gráfico de la solución exacta del problema (28)-(30).

Por otro lado, la Figura 2 muestra la comparación de $u(x)$ (en rojo) con las soluciones numéricas (en azul para $n = 2$, $n = 5$, $n = 10$ y $n = 100$, es decir, para un tamaño de paso, en el problema simplificado, $h = 0{,}25$, $h = 0{,}1$, $h = 0{,}05$ y $h = 0{,}005$.

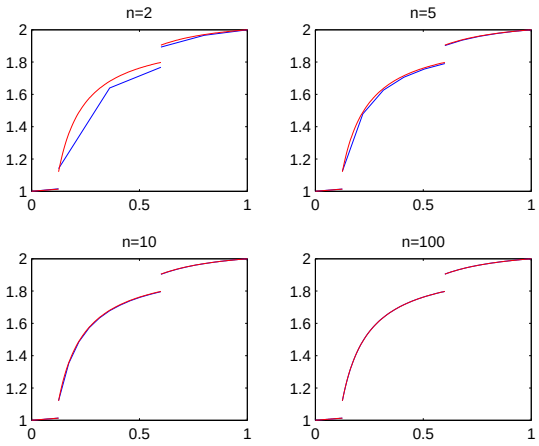


Figura 2. Comparación entre las soluciones numéricas para distintos valores de n y la solución exacta.

La convergencia de las soluciones numéricas hacia la solución exacta a medida que crece n es evidente.
Considérese ahora el mismo problema (28)-(30), pero con $f(x) = \sin x + x\cos x = (x\sin x)'$. En este caso en la fórmula de la solución general dada en la introducción, aparecerán términos del tipo

$$\int_{1/8}^{x} \frac{\sin x}{x},$$

que no pueden ser expresados en términos de funciones elementales. La Figura 3 muestra las distintas aproximaciones de la solución para este nuevo problema, a partir de distintos valores de n.

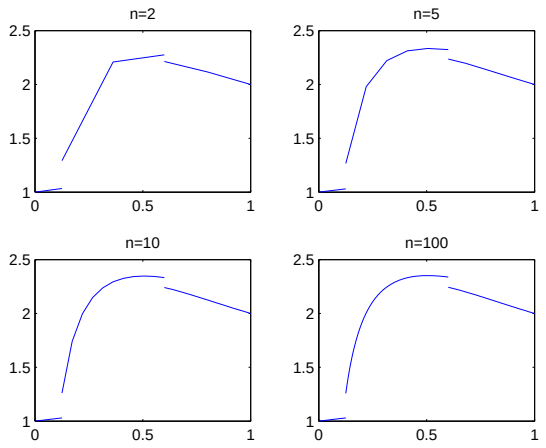


Figura 3. Soluciones numéricas para distintos valores de n del problema (28)-(30) modificado.

4. Conclusiones

El problema de conducción del calor con condiciones de contacto resulta de especial interés en la teoría de la homogeneización, ya sea simple o reiterada, donde aparecen problemas de este tipo, que al depender de un parámetro rapídamente oscilante, dificulta su tratamiento numérico. Sin embargo, durante el proceso de homogeneización aparecen los llamados problemas locales, además del problema homogeneizado, los cuales presentan una estructura similar al problema original, pero son independientes del parámetro pequeño, por lo tanto si resulta conveniente tratarlos numéricamente.

En el presente informe, luego de simplificar el problema original para que cumpliera determinadas condiciones, fue aplicado el método de las diferencias finitas a la ya mencionada ecuación, provando finalmente, la estabilidad y consistencia del método y por lo tanto la convergencia de las soluciones numéricas a la solución exacta. las funciones de MATLAB implementadas fueron discutidas e incluidas en el anexo y algunos ejemplos fueron discutidos.

Otra método numérico factible para abordar la ecuación aquí estudiada es el método de elementos finitos, puesto que es posible encontrar una formulación variacional para el problema de conducción del calor con condiciones de discontinuidad.

Agradecimientos

Los autores desean agradecer al Proyecto Nacional de Ciencias Básicas de Cuba PNCB-UH-57/14 “Modelación matemática de fenómenos dependientes de varias escalas microestructurales”; y a la Cátedra IIMAS-UNAM 2016.

Referencias

[1] A. Quarteroni, R. Sacco, F. Saleri, Numerical Mathematics, Springer-Verlag New York, ISBN 0-387-98959-5, 2007.

[2] L. Beilina, Hybrid Discontinuous Finite Element/Finite Difference Method for Maxwell’s Equations, AIP Conference Proceedings 1281, 324 (2010);

[3] De-Kang Mao, A Treatment of Discontinuities for Finite Differences Methods, Journal of Computational Physics 103,359-369 (1992)

[4] De-Kang Mao, A Treatment of Discontinuities for Finite Differences Methods in the Two-Dimensional Case, Journal of Computational Physics 104,377-397 (1993)

[5] Jirásek M., Comparative study on fnite elements with embedded discontinuities, Computer Methods in Applied Mechanics and Engineering, 188(1-3), pp. 307-330, 2000.

[6] Juarez-Luna G., Ayala A., Elmentos Finitos con Discontinuidades Interiores Mejorados para Simular el Daño en Sólidos, Concreto y Cemento. Investigación y Desarrollo, Vol. 6 Núm. 1, 15-35, 2014.

Vol. 27, 1976

[7] S. Larsson and V. Thomhée, Partial Differential Equations with Numerical Methods, Springer-Verlag Berlin Heidelberg, 2009.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 125-129, 2018
Recibido 02-2018

Tabla de Evaluación de Stress para Escalamiento Multidimensional
Stress Evaluation Table for Metric Multidimensional Scaling

Lic. Javier Alejandro Quintero Roba [1], Lic. María Esther Reyes Calzado [2*], Dra. Elina Miret Barroso [3], Alejandro Javier Quintero Roba [4]

Resumen El Escalamiento Multidimensional es una técnica de exploración multivariada que permite, mediante la minimización de una función de pérdida STRESS, la reducción de la dimensión de los datos para la búsqueda de patrones en la estructura de los mismos. Para medir la bondad de ajuste de la representación a partir de su valor de STRESS se han planteado numerosos criterios empíricos, teniendo en cuenta que, cuando se estudia un número grande de objetos, los valores de STRESS tienden a crecer junto a la dimensión del problema. En esta investigación se presenta una tabla probabilística de evaluación de STRESS con la intención de contar en la práctica de cotas superiores para dicha magnitud. Esta tabla se obtuvo a partir generación aleatoria de matrices de disimilitud y su procesamiento por varios métodos de Escalamiento Multidimensional.

Abstract Multidimensional Scaling is a multivariate exploration technique that allows to reduce dimension on data by minimizing a loss function called STRESS so that we can find patterns on the data structure. To measure the goodness of the adjusted representation from its STRESS value it has been studied a number of empirical criteria, knowing that STRESS tends to increase along with the dimension of the problem when the number of objects is large. This investigation presents a probabilistic table for STRESS evaluation, obtained from study of random dissimilarity matrices, intended to find heights for that magnitude.

Palabras Clave
Escalamiento Multidimensional — STRESS — Tabla de Evaluación de Stress

[1] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, jjquinte5693@gmail.com*
[3] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, m.reyes@matcom.uh.cu*
[2] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, elina@matcom.uh.cu*
[4] *Estudiante de Licenciatura en Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba*
[*] **Autor para Correspondencia**

Introducción

El Escalamiento Multidimensional (Multidimensional Scaling, MDS por sus siglas en inglés) es una técnica exploratoria para el análisis de similitudes y disimilitudes entre objetos; en general, medidas de relación entre varios individuos de estudio. La técnica intenta modelar los datos como distancias entre puntos en un espacio geométrico, usualmente que sea reconocible de manera visual.

Para realizar MDS es necesario un algoritmo computacional que a partir de los datos iniciales sea capaz de encontrar una representación fiel a la realidad. Para medir la bondad de ajuste de una solución se utiliza la función STRESS o alguna de sus variantes.

En la evaluación de la calidad de la representación del MDS a partir de su valor de STRESS se han planteado numerosos criterios empíricos teniendo en cuenta que, cuando se estudia un número grande de objetos, los valores de STRESS tienden a crecer junto a la dimensión del problema.

En esta investigación se presentan tablas probabilísticas de evaluación de STRESS con la intención de contar en la práctica de cotas superiores para dicha magnitud. Esta tabla se obtuvo a partir generación aleatoria de matrices de disimilitud y su procesamiento mediante métodos de Escalamiento Métrico y No Métrico.

1. Perspectiva de Escalamiento Multidimensional

En el MDS se representan medidas de similitud o disimilitud entre pares de objetos (individuos), como distancias entre espacios de baja dimensión, usualmente 2 ó 3. La representación gráfica que provee el Escalamiento Multidimensional permite literalmente explorar su estructura visualmente, lo que revela regularidades que permanecen ocultas cuando se estudian los números indistintamente. [9]

En dependencia de la transformación utilizada para el escalamiento de las proximidades iniciales, los modelos de MDS se clasifican en Métrico y No Métrico. El modelo más utilizado en la literatura es el no métrico u ordinal que se basa en la premisa de que las disparidades están en escala ordinal, lo que significa que para la construcción de las distancias finales solo importa el orden inicial de las proximidades y no se tiene en cuenta su valor. El modelo no métrico posee varias ventajas respecto al métrico, ya que los valores de bondad de ajuste son mejores que su contraparte, también el proceso de optimización es más flexible, lo que permite mover con mayor libertad las disparidades. Las desventajas de este modelo radican en la aparición de soluciones degeneradas y que no se tiene en cuenta como información relevante las nociones de cercanía o lejanía. [5] [6]

Se trabaja con MDS Métrico cuando la función de disparidades es continua, paramétrica y monótona. En este caso se desea conservar, además de los rangos ordinales entre las proximidades, la noción de magnitud entre ellas. Estos modelos poseen la ventaja de que a partir de las características de la función de disparidad es posible buscar propiedades matemáticas deseables para los algoritmos. Además, el modelo métrico evita la degeneración al imponer una estructura menos flexible que el modelo ordinal. Su desventaja radica en la bondad de ajuste, pues la libertad de movimiento de las disparidades es controlada por una función continua y para lograr una buena representación se requiere datos bastante precisos. Por lo anterior, los valores de STRESS en el caso del modelo métrico serán generalmente mayores, aun cuando las representaciones sean aproximadamente equivalentes. [3] [4]

A partir del planteamiento matemático del problema del MDS, los algoritmos pueden dividirse en dos ramas: técnicas algebraicas y algoritmos iterativos. Dentro de las primeras se encuentra el Escalamiento Clásico, cuya función de ajuste es el STRAIN y su optimización se basa en la teoría de la descomposición espectral. [8] [5] Dentro de los iterativos están, por ejemplo: los cuasi-Newton, máximo descenso, ALSCAL, SMACOF, entre otros; así también como métodos heurísticos aproximados.

Los algoritmos iterativos son un poco más flexibles que el MDS Clásico, pues permiten el re-escalamiento óptimo de los datos. Usualmente estos algoritmos parten de una configuración inicial, mueven los puntos para reducir el STRESS, y posteriormente reescalan las disimilitudes iniciales de forma óptima para generar nuevas disparidades, dentro de los límites de los datos. Este proceso de modificar la configuración del MDS (manteniendo fijas las disparidades), y re-escalar las disparidades (manteniendo fijas las distancias), se repite hasta lograr convergencia.[3] [4]

Sin embargo, no en todos los casos los resultados han sido del todo satisfactorios especialmente cuando la dimensión 1 (i.e cantidad de individuos) es alta. Además, la implementación de una herramienta a partir del material disponible resulta una tarea que en ocasiones se hace imposible debido a que las explicaciones en la literatura no son del todo claras y resultan insuficientes.[11]

1.1 STRESS y calidad de la representación

Para medir la bondad de ajuste de una solución se utiliza la función STRESS o alguna de sus variantes. Sea $X_{(n\times p)}$ una configuración en $\mathbb{R}^p$, la expresión de X como solución del MDS en p dimensiones se halla generalmente minimizando la función de pérdida $Stress-1$ ó de Kruskal, dada por:

$$Stress-1=\sqrt{\frac{\sum_{i<j}(d_{ij}-\widehat{\delta}_{ij})^2}{\sum_{i<j}d_{ij}^2}}$$

donde los d_{ij} son las nuevas distancias euclídeas, obtenidas de las nuevas coordenadas para los puntos de Ω.[8] [6]

Esta función toma valores en el intervalo $[0,1]$ y expresa si la representación obtenida refleja correctamente las relaciones de distancias originales. Otras variantes de la función de *Stress* son: el *Stress-normado* y el *S-Stress*.[3] [4]

Para evaluar la calidad de la representación del MDS a partir del STRESS existen varias opiniones. Esta investigación sigue dos líneas de esta teoría: los criterios empíricos de Kruskal de 1964, y el de bondad de ajuste bajo selección aleatorizada planteado por Borg, Groenen, & Mair en 2013. [10] [4]

El criterio empírico de Kruskal se basa en la experiencia del investigador, y establece las siguientes clasificaciones para una representación de puntos:

Stress	Ajuste
0.20	Pobre
0.10	Justo
0.05	Bueno
0.025	Excelente
0	"Perfecto"

Borg, Groenen, & Mair indican que una solución de MDS perfecta es aquella que tiene STRESS = 0, pues en este caso las distancias de la configuración representan los datos de manera precisa (en el sentido deseado). Esto conlleva a la pregunta de cuándo un valor de STRESS es suficientemente bueno.

Primeramente ha de analizarse que el algoritmo utilizado debe ser capaz de reconocer patrones ocultos en los datos, i.e. captar la topología de estos, reconocer que los datos poseen cierta estructura, y que no son solo un conjunto de datos aleatorios. Según los autores, la evaluación de un valor de STRESS específico es una cuestión compleja, que involucra un gran número de parámetros y consideraciones.

En general, un valor de STRESS es suficientemente bueno si es menor que el valor esperado de STRESS a partir de aleatorización de las matrices de disimilitud. Si esto no se cumple, es imposible interpretar significativamente en ningún sentido las distancias que se obtienen mediante el algoritmo, ya que estas no están realmente relacionadas con los datos. [12]

En [12] se establecen límites de aleatorización del STRESS, o sea el límite que indica que un valor por debajo de él tiene probabilidad 1 % de ser obtenido a partir de configuraciones aleatorias, o sea el percentil de la distribución empírica bajo simulaciones.

Otras medidas, como el STRESS por punto, y los diagramas de Shepard se reportan que pueden ser útiles. No obstante, en la práctica, los valores de STRESS obtenidos con datos reales son menores que los que provienen de selección aleatoria, esto se debe a la estructura propia de dichos datos.

2. Evaluación del STRESS para matrices aleatorias

La evaluación del STRESS es costosa cuando el número de individuos es elevado, pero cuando la dimensión crece es importante obtener indicadores de la calidad de la representación.

Teniendo en cuenta que, una matriz de disimilitudes que representa datos reales contiene una estructura subyacente al escalamiento, pero una matriz de disimilitudes generada aleatoriamente carece de estructura predefinida, es acertado pensar que los valores de STRESS que se obtienen de escalar matrices aleatorias son mayores que aquellos que se obtienen de matrices reales. Por lo cual, los valores de ajuste obtenidos del escalamiento de matrices aleatorias proveen una cota superior en la práctica para la función de STRESS.

En 1969, Klahr muestra que para 8 objetos los valores de STRESS obtenidos del escalamiento de matrices aleatorias eran superiores a los de matrices con estructura real. Posteriormente, en 1973, Spence and Ogilvie reportan experimentación con matrices aleatorias de 12-48 objetos y de 1-5 dimensiones.

En 1978, Levine produce una tabla que refleja la probabilidad de encontrar un valor específico de STRESS para una configuración dada. Esta tabla muestra valores de STRESS de 1–5 dimensiones para 6, 8, 10, 12, 16, 20 y 24 objetos. Esta idea fue retomada por Sturrock & Rocha, en 2000, quienes ofrecen una tabla de valores aleatorios de STRESS para 5-100 objetos escalados por métodos no métricos.[12]

Borg & Groenen plantean la necesidad de obtener el valor esperado bajo selección aleatoria de la función de STRESS como posible cota superior. Sin embargo, no se cuenta en la literatura con cifras de este tipo para el modelo métrico. [4]

El objetivo de esta investigación es elaborar una tabla que recopile los percentiles de nivel 1 de una muestra de valores de STRESS obtenidos a partir del escalamiento de matrices de disimilitud generadas aleatoriamente para contar en la práctica con cotas superiores de bondad de ajuste.

Para ello se generaron matrices de disimilitud de 5-25 objetos a partir de una distribución uniforme [0,1] (800 matrices en cada caso) que fueron escaladas mediante MDS No métrico con Evolución Diferencial y MDS Métrico con CMA-ES absoluto.

La selección de estos algoritmos se basa en trabajos previos de los autores donde, al aplicar Escalamiento Multidimensional empleando diferentes Metaheurísticas a ejemplos de la literatura y reales, se obtuvieron los mejores resultados para los métodos antes mencionados. [11] [10]

2.1 Evolución Diferencial en MDS No Métrico

Los algoritmos evolutivos son métodos de búsqueda dirigida basada en probabilidad. Estos algoritmos establecen una analogía entre el conjunto de soluciones de un problema y el conjunto de individuos de una población natural, codificando la información de cada solución en un string a modo de cromosoma. A tal efecto se introduce una función de evaluación de los cromosomas, que llamaremos calidad ("fitness") y que está basada en la función objetivo del problema. Igualmente se introduce un mecanismo de selección de manera que los cromosomas con mejor evaluación sean escogidos para "reproducirse" más a menudo que los que la tienen peor.[13]

La Evolución Diferencial se caracteriza por el uso de vectores de prueba, los cuales compiten con los individuos de la población actual a fin de sobrevivir. El algoritmo asume que las variables del problema a optimizar están codificadas como un vector de números reales y que el dominio de las variables del problema está restringido por ciertas cotas definidas para cada variable. Dado que se opera con una población en cada iteración, se espera que el método converja de modo que al final del proceso la población sea muy similar, y en el infinito se reduzca a un solo individuo. [7]

2.2 CMA-ES en MDS Métrico

La estrategia evolutiva de adaptación de la matriz de covarianzas o Covariance Matrix Adaptation Evolution Estrategy (CMA-ES, por sus siglas en inglés) es uno de los algoritmos de optimización continua más exitoso de los últimos años. Este algoritmo controla mediante la adaptación de la matriz de covarianzas los pasos individuales en cada dirección y las relaciones entre las coordenadas.[2]

Mediante una distribución normal se generan las mutaciones, creando la nueva población. CMA-ES adapta la matriz de covarianzas de la distribución normal multivariante de mutaciones, captando las relaciones de dependencia entre las variables, ya que la matriz de covarianzas define la dependencia por pares entre las variables de la distribución.

CMA-ES está especialmente orientada para el escenario provisto de problemas "black-box" y puede sobrellevar problemas de alta dimensionalidad de forma rápida. Está diseñada para tomar ventaja de espacios escarpados, problemas no separables, no linealidad, no suavidad y multimodalidad del STRESS. El STRESS es considerado una función de "black-box", pues su dominio no se conoce explícitamente, pero el valor de cada representación puede calcularse, siendo esta la única información disponible.[1]

3. Resultados

Se generaron 20000 matrices de disimilitud aleatoriamente (distribuidas como se describe anteriormente), y se procesaron por MDS Métrico con CMA-ES Absoluto (disimilitudes no escaladas) y MDS No Métrico con Evolución Diferencial.

El procesamiento de las muestras permitió la elaboración de tablas para cada algoritmo [Figuras 1 y 3] que representan los valores Mínimos y Máximos, Medias y percentiles de nivel 1 para cada cantidad de objetos estudiada (5-25).

No. de Objetos	Stress Mínimo	Media	Stress Máximo	Percentil Nivel 1
5	0,0748	0,2777	0,5628	0,092
6	0,1057	0,3113	0,5620	0,117
7	0,1615	0,3363	0,5218	0,182
8	0,2247	0,3586	0,5306	0,234
9	0,2246	0,3733	0,5340	0,259
10	0,2540	0,3934	0,5378	0,278
11	0,2998	0,4104	0,5461	0,308
12	0,3217	0,4274	0,5524	0,329
13	0,3298	0,4425	0,5637	0,345
14	0,3535	0,4554	0,5674	0,357
15	0,3623	0,4687	0,5669	0,387
16	0,3838	0,4815	0,5908	0,399
17	0,3917	0,4902	0,5651	0,406
18	0,4155	0,5008	0,5816	0,426
19	0,4308	0,5128	0,5898	0,440
20	0,4243	0,5224	0,6038	0,461
21	0,4487	0,5303	0,6238	0,460
22	0,4470	0,5365	0,6129	0,460
23	0,4514	0,5440	0,6183	0,476
24	0,4849	0,5506	0,6221	0,490
25	0,4879	0,5579	0,6269	0,496

Figura 1. Tabla Stress. MDS con CMAES

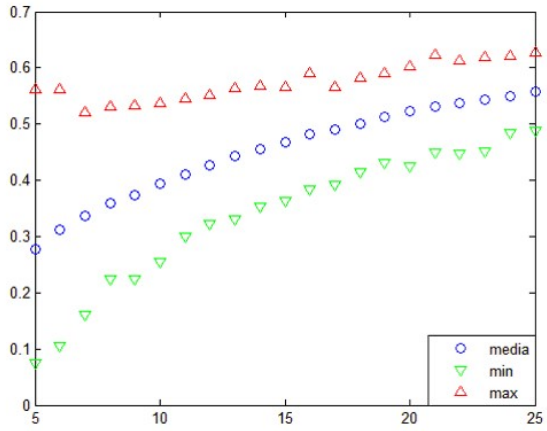


Figura 2. Curva de crecimiento CMAES

El percentil de nivel 1 aporta una cota superior de los valores esperados de STRESS para un número específico de objetos cuando se trabaje con datos reales.

Como era esperado, los valores de STRESS crecen al crecer el número de objetos. En las Figuras 2 y 4 puede apreciarse el comportamiento de los valores de STRESS al crecer la dimensión.

Los resultados obtenidos para los dos métodos empleados coinciden con los reportados por Sturrock & Rocha (2000).

No. de Objetos	Stress Mínimo	Media	Stress Máximo	Percentil Nivel 1
5	0.0351	0.2428	0.8371	0.066
6	0.0661	0.2746	0.6281	0.110
7	0.1252	0.3117	0.5659	0.155
8	0.1571	0.3317	0.5657	0.199
9	0.1874	0.3503	0.5493	0.224
10	0.2352	0.3692	0.5848	0.259
11	0.2568	0.3844	0.5627	0.281
12	0.2855	0.4012	0.5533	0.313
13	0.2972	0.4085	0.5276	0.325
14	0.3216	0.4220	0.5536	0.339
15	0.3218	0.4318	0.5494	0.354
16	0.3562	0.4421	0.5563	0.368
17	0.3695	0.4533	0.5510	0.385
18	0.3680	0.4606	0.5482	0.395
19	0.3842	0.4679	0.5688	0.400
20	0.3788	0.4742	0.5742	0.414
21	0.3817	0.4813	0.5833	0.425
22	0.4048	0.4873	0.5670	0.434
23	0.4269	0.4933	0.5701	0.441
24	0.4190	0.4993	0.5701	0.449
25	0.4338	0.5058	0.5824	0.455

Figura 3. Tabla Stress. MDS con Evolución Diferencial

Conclusiones

Se generaron 20000 matrices de disimilitud aleatoriamente, y se procesaron por MDS No Métrico con Evolución Diferencial y MDS Métrico con CMA-ES absoluto. El procesamiento de las muestras permitió la elaboración de una tabla que representa los percentiles de nivel 1. La tabla elaborada ofrece cotas superiores de calidad para la evaluación del STRESS en estos métodos. Las cifras obtenidas sirven como herramienta al investigador para juzgar a partir del valor de STRESS de su representación si el método empleado reconoce la estructura original de los datos. Se trabaja en el incremento del número de objetos y la validación de los resultados mediante su comparación con ejemplos reales.

Referencias

[1] A Auger and N. Hansen. Tutorial cma-es: evolution strategies and covariance matrix adaptation. *In GECCO (Companion).*, 2012.

[2] A. Bolufé Röhler. Búsqueda de población mínima–un algoritmo de optimización escalable para problemas multimodales. *Tesis de Doctorado.*, 2015.

[3] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling. Theory and Applications.* Springer., 2005.

[4] Ingwer Borg, Patrick J. F. Groenen, and Patrick Mair. *Applied multidimensional scaling.* Springer., 2013.

[5] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling.* Chapman & Hall., 1994.

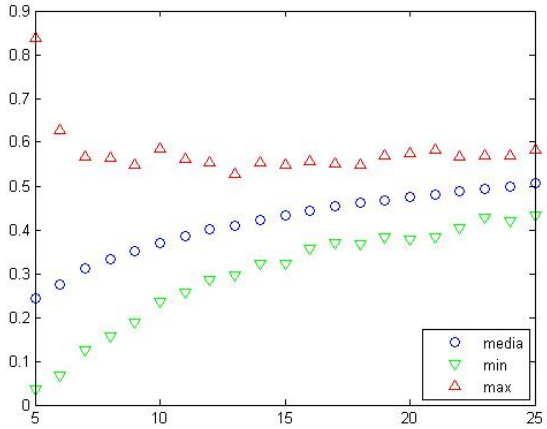


Figura 4. Curva de crecimiento Evolución Diferencial

[6] Carles M. Cuadras. *Nuevos Métodos de Análisis Multivariante.* CMC Editions., 2008.

[7] Stefan Etschberger and Andreas Hilbert. Multidimensional scaling and genetic algorithms : A solution approach to avoid local minima. *Arbeitspapiere zur mathematischen Wirtschaftsforschung.*, No. 181, 2003.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.*, volume Tenth printing.1995. Academic Press Inc., 1979.

[9] Elina Miret. Un enfoque unificado para técnicas de representación euclidiana. *Tesis de Doctorado.*, 2005.

[10] Javier A. Quintero. Algoritmo de escalamiento multidimensional métrico con el empleo de la estrategia evolutiva de adaptación de la matriz de covarianzas. *Tesis de Licenciatura.*, 2016.

[11] María E. Reyes. Escalamiento multidimensional empleando metaheurísticas. *Tesis de Licenciatura.*, 2015.

[12] Kenneth Sturrock and Jorge Rocha. A multidimensional scaling stress evaluation table. *Field Methods*, 2000.

[13] El-Ghazali Talbi. *Metaheuristics From Design to Implementation.* John Wiley & Sons., 2009.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 131-134, 2018
Recibido 02-2018

Soluciones de onda viajera de un modelo continuo de transmisión de energía en medios no homogéneos

Traveling-wave solutions of a continuous model of energy transmission in non-homogeneous media

Jorge E. Macías-Díaz

Abstract En este trabajo, se investigará la existencia de soluciones de un model continuo de transmisión de ondas en medios no homogéneos. Una extensión del método clásico de integrales directas (también conocido como el método de la ecuación de prueba) será empleada en este trabajo para derivar soluciones analíticas del modelo bajo estudio. Todas las soluciones que se obtendrán son ondas viajeras descritas de forma exacta.

Abstract In this work, we investigate the existence of exact solutions of a continuous model of wave transmission in nonhomogeneous media. An extension of the classical method of direct integrals (also known as the trial equation method) is used in this work in order to derive some analytical solutions of the model under study. All the solutions derived in this work are actually traveling waves that are presented in exact form.

Ecuaciones hiperbólicas alineales — Coeficientes temporalmente variables — Soluciones de onda viajera
Nonlinear hyperbolic equations — Time-dependent coefficients — Traveling-wave solutions

Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes, México, jemacias@correo.uaa.mx

Introduction

The classical Klein–Gordon equation and its (continuous and discrete) nonlinear generalizations are models that describe a wide range of phenomena in modern physics. Among many other recent reports, equations of the Klein–Gordon and sine-Gordon types have been employed to obtain a vast family of localized solutions in exact form [1], to study the behavior of a massive scalar field in a charged rotating hairy black hole [2], to explore the dynamics of genuine continuum breathers and the conditions under which it persists in a $\mathscr{P}\mathscr{T}$-symmetric medium [3], as a model to construct non-equilibrium steady states in arbitrary spatial dimensions under a local quench [4], as an example of a discrete model for which phase-shift breathers cannot be supported [5] and in the investigation of the phenomenon of nonlinear supratransmission in nonlinear regimes [6] among other applications [7, 8].

Most of the Klein–Gordon models investigated in the literature consider the presence of constant coefficients. Indeed, most of the relevant models are particular cases or (continuous or discrete) extensions of the nonlinear hyperbolic partial differential equation

$$u_{tt}+du_t-a^2u_{xx}+bu+cu^3=0, \tag{1}$$

where u is a function of the real variables x and t. Meanwhile, the symbols u_t, u_{tt} and u_{xx} represent the partial derivatives of u with respect to t, t^2 and x^2, respectively, and a, b, c and d are usually fixed real numbers. This last requirement physically implies that the media described by those Klein–Gordon models are homogeneous in both space and time.

Motivated by these facts, we will consider here an inhomogeneous generalization of the damped Klein–Gordon equation in which the coefficients are functions of time. The goal of this work is to propose expressions for traveling-wave solutions of that model, as well as conditions on the (non-constant) coefficients and the wave velocities that guarantee the existence of such solutions. To that end, we will employ the extension of the method of direct integrals (also known as the trial equation method) proposed in [1]. Such approach is a generalization of the method designed by W.-X. Ma and B. Fuchssteiner for partial differential equations with constant coefficients [9]. It is important to mention that different approaches have been employed in the literature to determine traveling-wave solutions of nonlinear models in physics [10]. However, the generalized trial equation technique is particularly interesting in view of its simplicity and the fact that various families of solutions are fully characterized through the values of discriminant systems (for instance, see [11, 12] and references therein).

This note is sectioned as follows. In Section 1, we present the damped Klein–Gordon equation with time-dependent co-

efficients that motivates this work. The method of the direct integral is used then to determine traveling-wave solutions of the model. The expressions of the constant parameters required by the direct integral method are derived, along with necessary conditions for the existence of the solutions. In Section 2, we consider various particular cases and, in each of them, we derive traveling-wave solutions of the parametric Klein-Gordon equation of interest. We will note in that section that some of the traveling-wave solutions that we will obtain are complex-valued functions. Finally, we close this work with a section of concluding remarks.

1. Klein–Gordon equation

In this work, we will use the symbol $\overline{\mathbb{R}^+}$ to represent the set of nonnegative real numbers, and we will suppose that a, b, c and d are real-valued and Riemann-integrable functions defined on $\overline{\mathbb{R}^+}$. Throughout we will assume that u is a nonnegative real function on the vector $(x,t) \in \mathbb{R} \times \overline{\mathbb{R}^+}$ with continuous partial derivatives up to the second order, satisfying the partial differential equation

$$u_{tt} + d(t)u_t - a^2(t)u_{xx} + b(t)u + c(t)u^3 = 0. \tag{2}$$

Note that this model is a generalization of the Klein–Gordon equation (1), which considers now non-constant coefficients that depend on the temporal variable.

In the present work, we will use an extension of the method of direct integrals (or trial equation method) proposed by W.-X. Ma and B. Fuchssteiner in [9] in order to account for the presence of time-dependent coefficients. Such parametric approach has been used recently for simpler diffusive models [1, 13]. Following the approach described in [1], we assume the existence of traveling-wave solutions of the model (2) which take the form

$$u(x,t) = u(\xi), \tag{3}$$

where

$$\xi = \xi(x,t) = \kappa(t)x + \omega(t), \tag{4}$$

for suitable functions $\kappa, \omega : \overline{\mathbb{R}^+} \to \mathbb{R}$ with continuous derivatives up to the second order.

Using these assumptions and the chain rule, the following identities readily follow:

$$u_t = (\kappa'(t)x + \omega'(t))\dot{u}, \tag{5}$$

$$u_{tt} = (\kappa'(t)x + \omega'(t))^2\ddot{u} + (\kappa''(t)x + \omega''(t))\dot{u}, \tag{6}$$

$$u_{xx} = \kappa^2(t)\ddot{u}. \tag{7}$$

Here, $\dot{u}$ represents the derivative of $u = u(\xi)$ with respect to ξ, and $\ddot{u}$ denotes the respective second derivative of u. Suppose that there exist a nonnegative integer m and real constants α_i for each $i = 0, 1, \ldots, m$, such that

$$(\dot{u})^2 = \sum_{k=0}^{m} \alpha_k u^k. \tag{8}$$

As a consequence,

$$\ddot{u} = \frac{1}{2}\sum_{k=1}^{m} k\alpha_k u^{k-1}. \tag{9}$$

Substituting the formulas (5)–(7) as well as the expansion (8) into our model, we obtain the ordinary differential equation

$$\Theta(x,t)\dot{u} + \Lambda(x,t)\sum_{k=1}^{k} k\alpha_k u^{k-1} + b(t)u + c(t)u^3 = 0, \tag{10}$$

where

$$\Theta(x,t) = [k''(t) + d(t)\kappa'(t)]x + \omega''(t) + d(t)\omega'(t), \tag{11}$$

$$\Lambda(x,t) = \frac{1}{2}\left[(\kappa'(t)x + \omega'(t))^2 - a^2(t)\kappa^2(t)\right]. \tag{12}$$

Note that the left-hand side of (10) is a polynomial in the variables u and $\dot{u}$ which must be equal to zero. Applying he balancing principle, it follows that $m = 4$. Moreover, as additional consequences of the balancing principle we obtain that

$$\Theta(x,t) = 0, \tag{13}$$

$$2\alpha_2\Lambda(x,t) + b(t) = 0, \tag{14}$$

$$4\alpha_4\Lambda(x,t) + c(t) = 0, \tag{15}$$

for every $(x,t) \in \mathbb{R} \times \overline{\mathbb{R}^+}$. Note also that $\alpha_1 = \alpha_3 = 0$ must be satisfied.

Several simplifications readily follow from these last identities. Indeed, the facts that both α_2 and α_4 are constants together with Equations (14) and (15) imply that Λ depends only on the temporal variable. This and the expression of Λ yield that $\kappa'(t) = 0$. Equivalently, $\kappa(t)$ must be equal to a constant $\kappa \in \mathbb{R}$. Therefore both Θ and Λ are functions exclusively of t, and the conditions (13)–(14) can be rewritten as

$$\omega''(t) + d(t)\omega'(t) = 0, \tag{16}$$

$$\alpha_2 = \frac{b(t)}{\kappa^2 a^2(t) - (\omega'(t))^2}, \tag{17}$$

$$\alpha_4 = \frac{c(t)}{2\left[\kappa^2 a^2(t) - (\omega'(t))^2\right]}, \tag{18}$$

for each $t \in \overline{\mathbb{R}^+}$. Observe that κ must be a real number such that the right-hand sides of the last two equations are constants.

Before we close this section, it is important to note that the solution (16) can be readily expressed in analytic form. Moreover, constant functions $\omega(t) = \omega$ satisfy this equation. Also, in the absence of the damping term the function ω adopts the form $\omega(t) = \omega_1 t + \omega_2$, where ω_1 and ω_2 are real constants.

2. Exact solutions

Following the results in Section 1, let $m=4$, let $\alpha_1=\alpha_3=0$, and suppose that $\alpha_0\in\mathbb{R}$ and that the expressions in (17) and (18) are constant real numbers. Using the change of variable $u=v^{1/2}$ and some algebraic simplifications, Equation (8) becomes

$$(\dot{v})^2=4\alpha_4v^3+4\alpha_2v^2+4\alpha_0v. \tag{19}$$

To simplify this expression we let $w=(4\alpha_4)^{1/3}v$. Expressing the identity (19) in terms of w, separating variables in the resulting equation and integrating both sides, we obtain

$$\int\frac{dw}{\sqrt{w(w^2+b_1w+b_0)}}=\pm(4\alpha_4)^{1/3}(\xi-\xi_0), \tag{20}$$

where $b_1=4\alpha_2(4\alpha_4)^{-2/3}$ and $b_0=4\alpha_0(4\alpha_4)^{-1/3}$, and ξ_0 is an arbitrary real number.

Let F be the polynomial inside the radical in the integrand of the last equation as a function of w. Note that the discriminant of $F(w)$ is given by $\Delta=b_1^2-4b_0$. Accordingly, various cases can be considered now depending on the discriminant and the constants in the expansion (8). Following the approach used in [1], we will consider those cases next.

Case 1. $\Delta=0$, $w>0$, $\alpha_2<0$ **and** $a_4>0$
In this case, the following functions are solutions of (2):

$$u=\pm\sqrt{\frac{|\alpha_2|}{2\alpha_4}}\tanh\left(\sqrt{\frac{|\alpha_2|}{2}}(\xi-\xi_0)\right), \tag{21}$$

$$u=\pm\sqrt{\frac{|\alpha_2|}{2\alpha_4}}\coth\left(\sqrt{\frac{|\alpha_2|}{2}}(\xi-\xi_0)\right). \tag{22}$$

Case 2. $\Delta=0$, $w>0$, $\alpha_2>0$ **and** $a_4>0$
Under these conditions, the solution is given by

$$u=\pm\sqrt{\frac{\alpha_2}{2\alpha_4}}\tan\left(\sqrt{\frac{\alpha_2}{2}}(\xi-\xi_0)\right). \tag{23}$$

Case 3. $\Delta=0$, $w>0$ **and** $b_1=0$
In this case, the corresponding solution is

$$u=\pm\frac{\sqrt[3]{2}}{\sqrt[3]{\alpha_4}(\xi-\xi_0)}. \tag{24}$$

Case 4. $\Delta>0$, $b_0=0$, $w>-b_1$, $\alpha_2>0$ **and** $\alpha_4>0$
In this case, the integrations yield the solutions

$$u=\pm\sqrt{\frac{\alpha_2}{\alpha_4}}\left[\frac{1}{2}\tanh^2\left(\sqrt{\frac{\alpha_2}{2}}(\xi-\xi_0)\right)-1\right]^{1/2}, \tag{25}$$

$$u=\pm\sqrt{\frac{\alpha_2}{\alpha_4}}\left[\frac{1}{2}\coth^2\left(\sqrt{\frac{\alpha_2}{2}}(\xi-\xi_0)\right)-1\right]^{1/2}. \tag{26}$$

It is worth mentioning here that, unfortunately, these functions take on values in the system of the complex numbers and they do not provide meaningful real solutions of (2). However, more relevant solutions are given by (see [13])

$$u=\pm\sqrt{\frac{\alpha_2}{\alpha_4}}\left(\frac{\operatorname{sech}(\sqrt{\alpha_2}(\xi-\xi_0))}{1-\sqrt{2}\tanh(\sqrt{\alpha_2}(\xi-\xi_0))}\right). \tag{27}$$

Case 5. $\Delta>0$, $b_0=0$, $w>-b_1$, $\alpha_2<0$ **and** $\alpha_4>0$
The following function is a solution of the damped Klein-Gordon equation:

$$u=\pm\sqrt{\frac{|\alpha_2|}{\alpha_4}}\left[\frac{1}{2}\tan^2\left(\sqrt{\frac{|\alpha_2|}{2}}(\xi-\xi_0)\right)-1\right]^{1/2}. \tag{28}$$

Again, it is easy to realize that this solution exists only for values of ξ for which the term in the brackets is a nonnegative number.

Case 6. $\Delta>0$, $b_0=0$, $w>-b_1$, $\alpha_2>0$ **and** $\alpha_4<0$
In this case, we have the following exact solutions of (see [13]):

$$u = \pm\sqrt{\frac{2\alpha_2}{|\alpha_4|}}\left(1+\cosh(2\sqrt{\alpha_2}(\xi-\xi_0)\right)^{-1/2}, \tag{29}$$

$$u = \pm\sqrt{\frac{\alpha_2}{|\alpha_4|}}\operatorname{sech}\left(\sqrt{\alpha_2}(\xi-\xi_0)\right). \tag{30}$$

Case 7. $\Delta>0$, $b_0\neq0$ **and** $\alpha<\beta<\gamma$
In a first stage, we suppose that $\Delta>0$, $b_0\neq0$, $\alpha<\beta<\gamma$ are real numbers with one of them equal to zero, the others are solutions of $w^2+b_1+b_0=0$, and $\alpha<w<\beta$. Under these circumstances,

$$u=\pm\sqrt[6]{4\alpha_4}\left[\alpha+(\beta-\alpha)\operatorname{sn}^2\left(\sqrt[3]{\alpha_4/2}\sqrt{\gamma-\alpha}(\xi-\xi_9),m\right)\right]^{1/2}. \tag{31}$$

Here, sn represents the Jacobi elliptic sine function and

$$m^2=\frac{\beta-\alpha}{\gamma-\alpha}. \tag{32}$$

Secondly, suppose that $\Delta>0$, $b_0\neq0$, $\alpha<\beta<\gamma$ are as before and $w>\gamma$. The solution of this case is provided by the formula

$$u=\pm\sqrt[6]{4\alpha_4}\frac{\left[\gamma+\beta\operatorname{sn}^2\left(\sqrt[3]{\alpha_4/2}\sqrt{\gamma-\alpha}(\xi-\xi_9),m\right)\right]^{1/2}}{\operatorname{cn}^2\left(\sqrt[3]{\alpha_4/2}\sqrt{\gamma-\alpha}(\xi-\xi_9),m\right)}, \tag{33}$$

where m is defined as in Case 6 and cn represents the Jacobi elliptic cosine function.

3. Conclusions

In this letter, we used an extension of the direct integral method or trial equation method to determine exact solutions of a

damped Klein-Gordon equation with time-dependent coefficients. The extension of the direct integral method used here is based on a paper by Liu [1], and it does not require for the coefficients of the trial equation to be constant. Instead, the approach followed here assumes that those coefficients are functions of the temporal variable. This technique is applied to a partial differential equation with non-constant coefficients that generalizes the well-known damped Klein–Gordon model with nonlinear reaction. As a result of our derivations, some traveling-wave solutions of our model have been calculated in exact form, and pertinent conditions on the coefficients of the model and the wave velocity are established in order to guarantee the existence of such solutions.

References

[1] Yang Liu. Exact solutions to nonlinear Schrödinger equation with variable coefficients. *Applied Mathematics and Computation*, 217(12):5866–5869, 2011.

[2] B Pourhassan. The Klein–Gordon equation of a rotating charged hairy black hole in (2+ 1) dimensions. *Modern Physics Letters A*, 31(09):1650057, 2016.

[3] Nan Lu, Panayotis G Kevrekidis, and Jesús Cuevas-Maraver. $\mathscr{PT}$-symmetric sine-Gordon breathers. *Journal of Physics A: Mathematical and Theoretical*, 47(45):455101, 2014.

[4] Benjamin Doyon, Andrew Lucas, Koenraad Schalm, and MJ Bhaseen. Non-equilibrium steady states in the Klein–Gordon theory. *Journal of Physics A: Mathematical and Theoretical*, 48(9):095002, 2015.

[5] Vassilis Koukouloyannis. Non-existence of phase-shift breathers in one-dimensional Klein–Gordon lattices with nearest-neighbor interactions. *Physics Letters A*, 377(34):2022–2026, 2013.

[6] J E Macías-Díaz and A Puri. An application of nonlinear supratransmission to the propagation of binary signals in weakly damped, mechanical systems of coupled oscillators. *Physics Letters A*, 366(4):447–450, 2007.

[7] Abdul-Majid Wazwaz and SA El-Tantawy. Gaussian soliton solutions to a variety of nonlinear logarithmic Schrödinger equation. *Journal of Electromagnetic Waves and Applications*, 30(14):1909–1917, 2016.

[8] AR Prasanna. On photon trajectories and electromagnetics near strongly gravitating cosmic sources. *Journal of Electromagnetic Waves and Applications*, 29(3):283–330, 2015.

[9] Wen-Xiu Ma and Benno Fuchssteiner. Explicit and exact solutions to a Kolmogorov-Petrovskii-Piskunov equation. *International Journal of Non-Linear Mechanics*, 31(3):329–338, 1996.

[10] Abdul-Majid Wazwaz. Solitary waves solutions for extended forms of quantum Zakharov–Kuznetsov equations. *Physica Scripta*, 85(2):025006, 2012.

[11] Cheng-Shi Liu. Applications of complete discrimination system for polynomial for classifications of traveling wave solutions to nonlinear differential equations. *Computer Physics Communications*, 181(2):317–324, 2010.

[12] Cheng-Shi Liu. A new trial equation method and its applications. *Communications in Theoretical Physics*, 45(3):395, 2006.

[13] Houria Triki and Abdul-Majid Wazwaz. Trial equation method for solving the generalized Fisher equation with variable coefficients. *Physics Letters A*, 380(13):1260–1262, 2016.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 135-139, 2018
Recibido 02-2018

Modelación y solución del problema del péndulo doble utilizando RK-4
Modeling and solving the double pendulum problem using RK-4

Miguel Ángel Abreu Terán[1], Alejandro Mesejo Chiong[2], Glicerio Viltres Castro[1]

Resumen En el presente trabajo se aborda el estudio del problema de oscilaciones de un péndulo doble, como un caso aparentemente sencillo de un sistema físico que puede mostrar un comportamiento caótico. Se presenta de manera breve la formulación analítica del problema para obtener dos Ecuaciones Diferenciales de segundo orden, acopladas y altamente no lineales. Estas ecuaciones están asociadas a los ángulos θ_1 y θ_2, por cuanto bajo condiciones iniciales conocidas es posible resolver el problema original desarrollando un algoritmo Runge-Kutta de orden cuatro "RK4" para obtener la solución numérica.Para el cálculo numérico se empleó el SAC Wolfram Mathematica 9.0.Al final se presentan comparaciones entre las soluciones obtenidas por las diferentes alternativas propuestas y diferencias significativas que existen entre la solución del sistema original de Ecuaciones Diferenciales No Lineales y su respectiva linealización.

Abstract In the present work the study of the problem of oscillations of a double pendulum is approached as an apparently simple case of a physical system that can show a chaotic behavior. The analytical formula of the problem is presented briefly to obtain the second order differential equations, coupled and highly nonlinear. These equations are associated with the angles θ_1 and θ_2, because under known initial conditions it is possible to solve the original problem by developing a Runge-Kutta algorithm of order "RK4" to obtain the numerical solution. For the numerical calculation the SAC Wolfram Mathematica 9.0 was used. At the end, comparisons are presented between the solutions obtained for the different proposed alternatives and significant differences that exist between the system solution original of Nonlinear Differential Equations and their respective linearization.

Palabras Clave
modelación matemática — EDOs no lineales — solución de sistemas de EDOs no lineales

[1] *Departamento de Matemática Aplicada, Universidad de Oriente, Cuba, maabreu@uo.edu.cu, glicerio@uo.edu.cu,*
[2] *Departamento de Matemática, Universidad de la Habana, Cuba, mesejo@matcom.uh.cu*

Introducción

La mecánica es la rama de la Física que estudia y analiza el movimiento y reposo de los cuerpos al igual que su evolución en el tiempo, bajo la acción de fuerzas. La mecánica analítica es una formulación abstracta y general de la mecánica que permite el uso en igualdad de condiciones de sistemas inerciales o no inerciales sin que, a diferencia de las leyes de Newton, la forma básica de las ecuaciones de movimiento cambie. Se considera que el rasgo determinante es considerar la exposición y planteamiento de la misma en términos de coordenadas generalizadas.

Si se quiere estudiar un sistema físico es importante en principio conocer los elementos, variables, parámetros y posibles comportamientos que pueden derivarse de un determinado escenario o evento. En general un péndulo doble es un sistema compuesto por dos péndulos, con el segundo colgando del extremo del primero. En el caso más simple, se trata de dos péndulos simples, con el interior colgando de la masa pendular del superior. Normalmente se sobreentiende que nos referimos a un doble péndulo plano. Este sistema físico posee dos grados de libertad y exhibe un rico comportamiento dinámico, su movimiento está gobernado por dos ecuaciones diferenciales de segundo orden ordinarias no lineales acopladas. Por encima de cierta energía su movimiento es caótico. Si bien estas EDOs poseen serias dificultades para obtener la solución analítica es posible obtener una linealización que permite aplicar la Transformada de Laplace bajo ciertas condiciones y obtener una solución analítica para oscilaciones suficientemente pequeñas.

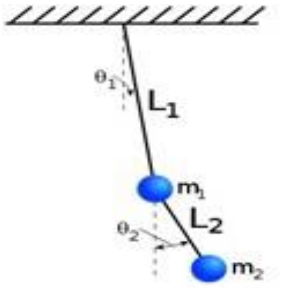


Figura 1. Péndulo Doble.

1. Análisis del movimiento del péndulo doble plano

1.1 Cinemática

La formulación parte de que el sistema oscila en un plano vertical y no existen fuerzas de amortiguamiento en el mismo. En la cinemática solo estamos interesados en encontrar las expresiones de la posición, la velocidad, la aceleración en términos de las variables que especifican el estado del péndulo doble, sin interesarnos por las fuerzas actuantes. Usando las proyecciones de las líneas L_1 y L_2 sobre un sistema de coordenadas localizado en el extremo superior del péndulo 1, es posible establecer las posiciones de las masas m_1 y m_2 de acuerdo con las ecuaciones (1) a (4).

$$x_1 = L_1 \operatorname{sen}\theta_1 \tag{1}$$

$$y_1 = -L_1 \cos\theta_1 \tag{2}$$

$$x_2 = x_1 + L_2 \operatorname{sen}\theta_2 \tag{3}$$

$$y_2 = y_1 - L_2 \cos\theta_2 \tag{4}$$

Si se sitúa el nivel cero de energía potencial en el punto de suspensión del primer péndulo la energía potencial queda expresada de la siguiente forma:

$$E_p = m_1 g y_1 + m_2 g y_2 \tag{5}$$

y sustituyendo (2) y (4) en (5) queda:

$$E_p = -(m_1+m_2) g L_1 \cos\theta_1 + m_2 g L_2 \cos\theta_2 \tag{6}$$

Notemos que la magnitud de la velocidad del primer péndulo se obtiene como el producto entre el radio de la circunferencia y la velocidad angular.

$$\vec{v_1} = L_1 \dot{\theta_1}$$

Su expresión vectorial estará dada por la ecuación

$$\vec{v_1} = L_1 \dot{\theta_1} \cos\theta_1 \vec{i} + L_1 \dot{\theta_1} \operatorname{sen}\theta_1 \vec{j} \tag{7}$$

La magnitud de la velocidad de la segunda partícula en un sistema de referencia que se mueve con velocidad v_1 de la primera partícula es $\vec{v_2} = L_2 \dot{\theta}$ y su expresión vectorial estará dada por la ecuación:

$$\vec{v_2} = L_2 \dot{\theta_2} \cos\theta_2 \vec{i} + L_2 \dot{\theta_2} \operatorname{sen}\theta_2 \vec{j} \tag{8}$$

Ahora podemos plantear la energía cinética como

$$E_c = \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \tag{9}$$

Entonces el lagrangiano del sistema podemos escribirlo

$$L = E_c - E_p$$

En términos de las coordenadas generalizadas podemos escribirlo tras realizar operaciones algebraicas

$$\begin{aligned} L = {} & \frac{1}{2}(m_1+m_2)(L_1)^2(\dot{\theta}_1)^2 + \frac{1}{2}(m_2 L_2^2 \dot{\theta}_2^2) + \\ & + (m_2 L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2)) + \\ & + (m_1+m_2) g L_1 \cos(\theta_1) + m_2 g L_2 \cos(\theta_2) \end{aligned} \tag{10}$$

A partir del lagrangiano se obtienen las ecuaciones del movimiento que conducen a un sistema de EDOs no lineales acopladas. Utilizando las ecuaciones de Euler-Lagrange en este caso particular:

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\theta}_1}\right] = \left[\frac{\partial L}{\partial \theta_1}\right] \tag{11}$$

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\theta}_2}\right] = \left[\frac{\partial L}{\partial \theta_2}\right] \tag{12}$$

Calculando explícitamente las derivadas de las expresiones anteriores se obtiene el sistema (13) de EDOs no lineales

$$\begin{cases} 0 = (m_1+m_2) L_1 \ddot{\theta}_1 + m_2 \ddot{\theta}_2 L_1 L_2 \cos(\theta_1 - \theta_2) + \\ \quad + m_2 L_2 (\dot{\theta}_2)^2 \sin(\theta_1 - \theta_2) + (m_1+m_2) g \sin(\theta_1) \\ \\ 0 = m_2 L_2 \ddot{\theta}_2 + m_2 \ddot{\theta}_1 L_1 \cos(\theta_1 - \theta_2) - \\ \quad - m_2 L_1 (\dot{\theta}_1)^2 \sin(\theta_1 - \theta_2) + m_2 g \sin(\theta_2) \end{cases} \tag{13}$$

Es válido observar que el sistema dado anteriormente no posee solución analítica conocida, sin embargo podemos considerar que si los giros son pequeños, es decir

$$\cos(\theta_1 - \theta_2) \approx 1 \tag{14}$$

$$\sin(\theta_1 - \theta_2) \approx 0 \tag{15}$$

$$\sin\theta_1 \approx \theta_1 \tag{16}$$

$$\sin\theta_2 \approx \theta_2 \tag{17}$$

Considerando que los giros son pequeños el sistema puede ser simplificado a un sistema de EDOs lineales, al cual podremos encontrarle una solución analítica. De otra manera considerar que los giros no son pequeños y el sistema (13) es posible encontrarle una solución numérica por reducción de orden a cuatro ecuaciones diferenciales de primer orden no lineales con condiciones adaptadas, partiendo del sistema original.

1.2 Solución para casos particulares

Consideremos primero las oscilaciones pequeñas libres de un péndulo doble plano y desarrollemos un ejemplo con los siguientes datos

$$L_1 = 16ft,\ \ L_2 = 16ft,\ \ m_1 = 3\ slug$$
$$m_2 = 1\ slug,\ \ g = 32ft/s^2$$
$$\theta_1(0) = 1\ rad,\ \ \theta_2(0) = -1\ rad$$
$$\theta_1(0) = \theta_2(0) = 0\ rad/s$$

para un tiempo final de 30 s.

Sustituyendo en el sistema (13) podemos escribir

$$\begin{cases} 0 = 4\ddot{\theta}_1 + \ddot{\theta}_2 + 8\theta_1 \\ \\ 0 = \ddot{\theta}_1 + \ddot{\theta}_2 + 2\theta_2 \end{cases} \tag{18}$$

Resolvemos el sistema utilizando la transformada de Laplace teniendo en cuenta las condiciones iniciales dadas y apoyándonos en el teorema de los residuos podemos obtener finalmente la solución analítica del sistema:

$$\begin{cases} \theta_1(t) = \dfrac{3}{4}cos(2t) + \dfrac{1}{4}cos(\dfrac{2}{\sqrt{3}}t) \\ \\ \theta_2(t) = -\dfrac{3}{2}cos(2t) + \dfrac{1}{2}cos(\dfrac{2}{\sqrt{3}}t) \end{cases} \tag{19}$$

Podemos ilustrar gráficamente la solución analítica del sistema utilizando el asistente Wolfram Mathematica 9.0.

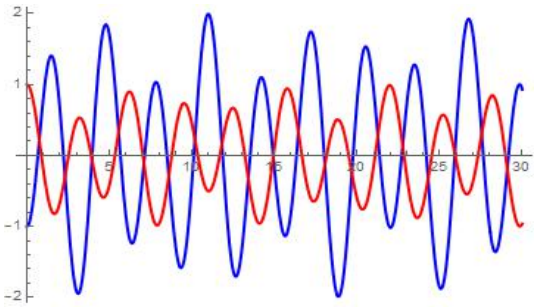


Figura 2. Linealizado utilizando Transformada de Laplace.

En la gráfica se muestra el modelo linealizado, el segundo péndulo (rojo) con amplitudes de giro mayor que el primero (azul). Se puede apreciar la forma de la solución dada en cosenos en donde la amplitud no es periódica a lo largo de todo el intervalo.

1.2.1 Solución del modelo lineal utilizando RK-4

Se utilizarán los mismos datos de la sección 1.2

$$L_1 = 16ft,\ \ L_2 = 16ft,\ \ m_1 = 3\ slug$$
$$m_2 = 1\ slug,\ \ g = 32ft/s^2$$
$$\theta_1(0) = 1\ rad,\ \ \theta_2(0) = -1\ rad$$
$$\theta_1(0) = \theta_2(0) = 0\ rad/s$$

para un tiempo final de 30 s.

El sistema (18) se puede reducir a un sistema de cuatro ecuaciones diferenciales lineales considerando que

$$\frac{d\theta_1}{dt} = w_1 \tag{20}$$
$$\frac{d\theta_2}{dt} = w_2 \tag{21}$$
$$\frac{dw_1}{dt} = \frac{1}{3}(-8\theta_1 + 2\theta_2) \tag{22}$$
$$\frac{dw_2}{dt} = \frac{8}{3}(\theta_1 - \theta_2) \tag{23}$$

Solamente mostraremos en la tabla y en la gráfica los valores de t comprendidos entre 0 y 3 s.

t	θ_1	θ_2
0	0	0
0,1	0,0009978	0,0019956
0,2	0,0019624	0,0039248
0,3	0,0028617	0,0057234
0,4	0,0036658	0,0073317
0,5	0,004348	0,008696
0,6	0,0048855	0,009771
0,7	0,0052605	0,0105209
0,8	0,0054604	0,0109208
0,9	0,0054787	0,0109573
1	0,0053146	0,0106292
1,1	0,0049737	0,0099474
1,2	0,0044673	0,0089347
1,3	0,0038123	0,0076246
1,4	0,0030304	0,0060608
1,5	0,0021476	0,00429513
1,6	0,0011934	0,0023868
1,7	0,0001994	0,0003989
1,8	−0,0008012	−0,0016024
1,9	−0,0017752	−0,0035504
2	−0,0026901	−0,0053802
2,1	−0,0035156	−0,0070311
2,2	−0,0042241	−0,0084481
2,3	−0,004792	−0,009584
2,4	−0,005201	−0,010401
2,5	−0,005436	−0,01087
2,6	−0,005491	−0,010981
2,7	−0,005363	−0,010725
2,8	−0,005056	−0,010112
2,9	−0,004581	−0,009163
3	−0,003954	−0,007908

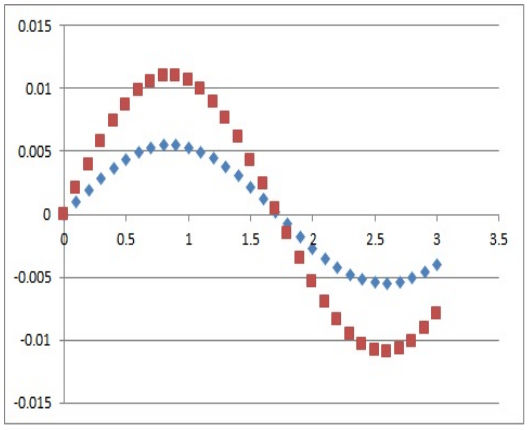


Figura 3. Linealizado utilizando RK-4.

En la gráfica se muestra el modelo linealizado utilizando el Método de RK-4, el segundo péndulo (rojo) con amplitudes de giro mayor que el primero (azul). Se puede apreciar la forma de la solución dada en cosenos en donde la amplitud no es periódica a lo largo de todo el intervalo. Notemos que los gráficos de las soluciones obtenidas en ambos no son significativamente diferentes.

1.2.2 Solución del modelo no lineal utilizando RK-4

Se utilizarán los mismos datos de la sección 1.2

$$L_1 = 16ft,\ \ L_2 = 16ft,\ \ m_1 = 3\ slug$$
$$m_2 = 1\ slug,\ \ g = 32ft/s^2$$
$$\theta_1(0) = 1\ rad,\ \ \theta_2(0) = -1\ rad$$
$$\theta_1(0) = \theta_2(0) = 0\ rad/s$$

para un tiempo final de 30 s.

Partiendo del sistema (13) es posible mediante reducción de orden obtener el siguiente sistema de ecuaciones diferenciales no lineales de primer orden.

$$\frac{d\theta_1}{dt} = w_1 \tag{24}$$

$$\frac{d\theta_2}{dt} = w_2 \tag{25}$$

$$\frac{dw_1}{dt} = \frac{-m_2 l_1 w_1^2 \sin(\theta_1-\theta_2)\cos(\theta_1-\theta_2)}{l_1[(m_1+m_2)-m_2\cos^2(\theta_1-\theta_2)]} + \\ + \frac{m_2 g \sin(\theta_2)\cos(\theta_1-\theta_2)}{l_1[(m_1+m_2)-m_2\cos^2(\theta_1-\theta_2)]} - \\ - \frac{m_2 l_2 w_2^2 \sin(\theta_1-\theta_2)+(m_1+m_2)g\sin(\theta_1)}{l_1[(m_1+m_2)-m_2\cos^2(\theta_1-\theta_2)]} \tag{26}$$

$$\frac{dw_2}{dt} = \frac{-m_2 l_2 w_2^2 \sin(\theta_1-\theta_2)-(m_1+m_2)g\sin(\theta_1)}{m_2 l_2 cos(\theta_1-\theta_2)} - \\ - \frac{(m_1+m_2)l_1\dfrac{dw_1}{dt}}{m_2 l_2 cos(\theta_1-\theta_2)} \tag{27}$$

Este nuevo sistema puede ser resuelto numéricamente mediante una adaptación del Método de RK-4 para sistemas de EDOs con los datos dados anteriormente. Se obtuvieron los valores de θ_1 y θ_2 que aparecen en la siguiente tabla. Solamente mostraremos en la tabla y en la gráfica los valores de t comprendidos entre 0 y 0.2 s.

t	θ_1	θ_2
0	0	0
0,01	0,000251505	0,000103045
0,02	0,00034483	0,000206091
0,03	6,06362	0,000309136
0,04	−0,000450133	0,000412182
0,05	−0,000616787	0,000515227
0,06	5,3009	0,000618273
0,07	0,00123555	0,000721318
0,08	0,001669569	0,000824364
0,09	0,000250582	0,000927409
0,1	−0,002328668	0,001030455
0,11	−0,003281604	0,0011335
0,12	−0,00012773	0,001236545
0,13	0,00564887	0,001339591
0,14	0,007876718	0,001442636
0,15	0,001003321	0,001545682
0,16	−0,011831485	0,001648727
0,17	−0,01695083	0,001751773
0,18	−0,001729951	0,0018548181
0,19	0,027168524	0,00195786355
0,2	0,038270127	0,002060909

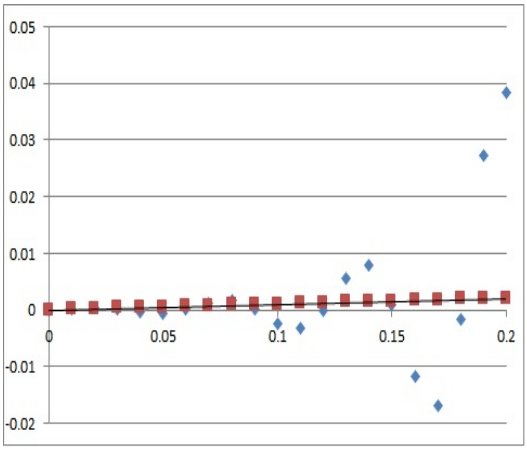


Figura 4. No linealizado utilizando RK-4.

En la **Figura 4** θ_1 aparece en color azul y θ_2 aparece en color rojo.

Según el gráfico podemos apreciar el comportamiento caótico del primer péndulo, cuyo ángulo de giro es θ_1, esto nos permite establecer una comparación entre la solución del modelo

linealizado y el modelo no linealizado, es decir, para oscilaciones no pequeñas existe una diferencia considerable entre ambas soluciones.

2. Resultados y Conclusiones

A partir de los resultados obtenidos queda claro que cuando se modela cualquier fenómeno debemos ser muy cuidadosos, cuando se consideraron oscilaciones pequeñas los resultados obtenidos en el modelo lineal exhiben diferencias no muy significativas entre el modelo numérico y el analítico, no ocurriendo así si quisiéramos establecer comparación entre la solución aproximada del modelo no lineal y las soluciones linealizadas.

Referencias

[1] L. Lifshitz. *Mechanics*, 1976.

[2] M. R. Spiegel. *Theory and problems of theoretical mechanics*, 1967.

[3] J. B. Castillero y R. R. Ramos y B. Valiño. *Apuntes para un curso de mecánica clásica*.

[4] D. Garanin. *Exercises in Classical Mechanics*.

[5] V. H. Mandell. *The Double Pendulum Problem*, 2000.

[6] K. R. Symon. *Mechanics*, 1960.

[7] Stickel. *The Double Pendulum*, 2009.

[8] G. González. *Single and Double plane pendulum*.

[9] P. Valdés Castro.*Sistema de actividades de dinámica no lineal en un curso inicial de Mecánica*,2015.

[10] Seoane, J., Zambrano, S. y SanJuán, M. (2011) *"Teaching Nonlinear Dynamics and Chaos for Beginners"*.

[11] S. C. Chapra y R. P. Canale. *Métodos Numéricos para Ingenieros*. 719-756, 2007.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 141-147, 2018
Recibido 02-2018

Algoritmo Optimizado para la Interpolación Espacial del Krigeado Ordinario
Optimized Algorithm for Spatial Interpolation of Ordinary Kriging

Milenis Fernández Díaz[1]*, José Quintín Cuador Gil[2], César Rául García Jacas[3]

Resumen Los métodos de interpolación espacial proporcionan herramientas para la estimación de valores en localizaciones no muestreadas utilizando las observaciones cercanas. La interpolación de Krigeado Ordinario es uno de los métodos geoestadísticos más frecuentemente usados para la realización de análisis espaciales. Su objetivo consiste en encontrar el Mejor Estimador Lineal Insesgado a partir de los datos disponibles, los cuales generalmente son insuficientes debido al costo de su obtención. Se caracteriza por costosas operaciones de álgebra lineal que repercuten en altos tiempos de ejecución, fundamentalmente la resolución de grandes sistemas de ecuaciones lineales. La reducción del tiempo de ejecución de aplicaciones de interpolación espacial puede ser un objetivo de alta prioridad, por ejemplo, en sistemas que soportan la toma de decisiones rápidas. Con el objetivo de disminuir los tiempos asociados a la interpolación espacial del Krigeado Ordinario, se propuso un algoritmo basado en el uso de técnicas de programación paralela, así como métodos optimizados de búsqueda espacial; que permita resolver los problemas que satisfacen los supuestos matemáticos apropiados en tiempos razonables, fundamentalmente en el campo de las Geociencias. Este algoritmo fue implementado usando C++11 como lenguaje de programación, OpenMP 4.8.2 como biblioteca de programación paralela en memoria compartida, y Atlas CLapack como biblioteca de algebra lineal optimizada para los cálculos matriciales. El algoritmo propuesto permite una mayor rapidez en la interpolación espacial de Krigeado Ordinario, logrando un mejor aprovechamiento de los recursos de cómputo instalados.

Abstract Spatial interpolation methods provide tools for estimating values at unsampled locations using nearby observations. Ordinary Kriging interpolation is one of the most frequently used geostatistical methods for performing spatial analysis. Its objective is find the Best Linear Unbiased Estimator from the data available, which generally are insufficient because of the cost of obtaining it. It was characterized by expensive linear algebra operations affecting high runtimes fundamentally solve large systems of linear equations. Reducting the runtime of spatial interpolation applications can be a high priority target, for example, in systems that support quick decisions. In order to reduce the time associated to spatial interpolation of Ordinary Kriging, it was proposed an algorithm based on the use of parallel programming techniques and optimized search methods; for resolving problems meeting the appropriate mathematical assumptions at reasonable time, mainly in the field of Geosciences. This algorithm was implemented using C++11 programming language, OpenMP 4.8.2 as library of shared memory parallel programming, and Atlas CLAPACK as linear algebra library optimized for matrix calculations. The proposed algorithm allows faster in the spatial interpolation of Ordinary Kriging, making better use of computing resources installed.

Palabras Clave
geoestadística — interpolación espacial — Krigeado Ordinario — programación paralela

[1] *Centro de Geoinformática y Señales Digitales, Universidad de las Ciencias Informáticas, La Habana, Cuba, mfdiaz@uci.cu*
[2] *Departamento de Física, Universidad de Pinar del Río, Pinar del Río, Cuba, cuador@upr.edu.cu*
[3] *Centro de Estudios de Matemática Computacional, Universidad de las Ciencias Informáticas, La Habana, Cuba, crjacas@uci.cu*
***Autor para Correspondencia**

Introducción

La interpolación espacial de Krigeado tiene como objetivo encontrar el Mejor Estimador Lineal Insesgado a partir de los datos disponibles [3], los cuales generalmente son insuficientes debido al costo de su obtención. Se basa en el principio de que las variables espaciales de una determinada población se encuentran correlacionadas en el espacio; es decir que mientras más cercanos estén dos puntos sobre la superficie terrestre,

menor será la variación de los atributos medidos [4]. Se apoya en variogramas como funciones estadísticas que expresan las características de variabilidad y correlación espacial del fenómeno que se estudia a partir de puntos muestreados.

El Krigeado constituye un método de interpolación espacial muy utilizado en la construcción de superficies y cuerpos tridimensionales a partir de nubes irregulares de puntos, en la estimación de variables aleatorias en puntos no muestreados, así como en otras aplicaciones de la geoestadística. Especialmente en el área de las Geociencias, es ampliamente utilizado en la estimación de recursos y reservas minerales útiles, considerando el nivel de precisión y confiabilidad que caracteriza sus resultados en estimaciones locales. Precisamente, es en el campo de la Minería, donde se origina el Krigeado de manos del ingeniero en minas Danie Krige, al explotar la correlación espacial para hacer predicciones en la evaluación de reservas de las minas de oro en Sudáfrica [5]; y gracias a la formulación matemática de George Matheron en la Escuela de Minas de París.

La complejidad de los cálculos matemáticos utilizados en la interpolación de Krigeado, fundamentalmente la resolución de grandes sistemas de ecuaciones, tiene un alto costo computacional confirmado por varios autores: [7, 8, 11, 12, 13]. Se plantea que el algoritmo por cada punto de interés tiene una complejidad cúbica, lo que conduce a una complejidad de $O(MN^3)$, siendo N el número de observaciones disponibles y M el número de puntos a interpolar [13]. Cuando $M \approx N$, la complejidad computacional puede considerarse $O(N^4)$ lo cual no es favorable si se trabaja con grandes volúmenes de datos.

1. Métodos

1.1 Formulación matemática del krigeado ordinario

El problema del Krigeado Ordinario consiste en estimar el valor en el sitio desconocido, expresado matemáticamente mediante la combinación lineal ponderada de los valores muestreados. A través del Krigeado Ordinario se puede estimar tanto el valor desconocido de un punto como el valor promedio de un bloque, conocidos respectivamente como Krigeado puntual y Krigeado de bloques. La estimación se calcula mediante la Ecuación 1:

$$z*_k = \lambda_1 z(x_1) + \lambda_2 z(x_2) + ... + \lambda_n z(x_n) \tag{1}$$

donde:

$z*_k$ es el valor estimado
$z(x_i)$ son los valores de las ensayos
λ_i son los pesos de Krigeado
n es el número de observaciones disponibles.

El Krigeado atribuye un peso λ_i a la ley de cada muestra $z(x_i)$, donde los pesos altos corresponden a las muestras cercanas y los pesos débiles a las alejadas. La ponderación depende del modelo ajustado a los puntos medidos, la distancia a la ubicación de la predicción, y las relaciones espaciales entre los valores medidos alrededor de la ubicación de la predicción. Estos pesos se calculan considerando las características geométricas del problema, de manera que [1] :

1. $\lambda_1 + \lambda_2 + ... + \lambda_n = 1$ se garantice la condición de universalidad (es decir la sumatoria de los pesos debe ser unitaria)
2. $\sigma_{E^2} = var[z^*_k - z_k]$ la varianza del error cometido sea mínima.

Estos elementos conducen a un problema de minimización con restricciones que se resuelve utilizando la técnica denominada multiplicadores de Lagrange. Este método involucra la incógnita auxiliar llamada parámetro de Lagrange (μ) y consiste en igualar a cero las derivadas parciales de la nueva función. Al realizar las *N+1* derivaciones se obtiene un sistema de *N+1* ecuaciones lineales con *N+1* incógnitas. Los valores de los pesos asociados a cada uno de los puntos se calculan mediante la resolución de este sistema de ecuaciones (Ecuación 2) [1].

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(u_i - u_j) + \mu = \gamma(u - u_i) & i = 1...n \\ \sum_{j=1}^{n} \lambda_j = 1 \end{cases} \tag{2}$$

El sistema de ecuaciones también puede ser expresado en forma matricial en función de la covarianza (Ecuación 3). Los términos del miembro izquierdo del sistema de ecuaciones se determinan mediante el cálculo de las covarianzas de cada par de ensayos. Por otra parte, el miembro derecho se determina mediante la covarianza entre el punto o bloque y cada uno de los ensayos. Se observa las propiedades simétricas de la matriz que conforma el miembro izquierdo del sistema de ecuaciones lineales. El aprovechamiento de esta propiedad permitirá reducir los tiempos de construcción de esta matriz.

$$\begin{pmatrix} \gamma(u_1 - u_1) & \cdots & \gamma(u_1 - u_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(u_n - u_1) & \cdots & \gamma(u_n - u_n) & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(u_n - u) \\ \vdots \\ \gamma(u_n - u) \\ 1 \end{pmatrix} \tag{3}$$

Una de las características más importantes del Krigeado Ordinario es que proporciona la varianza del error de estimación, la cual depende del modelo de variograma obtenido y de las localizaciones de los datos originales [2]. La varianza del error puede calcularse mediante la Ecuación 4:

$$\sigma^2_{KO} = \sum_{i=1}^{n} \lambda_i \gamma(u_i - u) - \sigma^2 - \mu \tag{4}$$

1.2 Descripción del Krigeado Ordinario

El proceso de Krigeado involucra 4 pasos fundamentales: búsqueda de los ensayos en la vecindad definida, con el fin de limitar el número de datos a utilizar en la interpolación y evitar el trabajo con grandes sistemas de ecuaciones lineales; el cálculo de los pesos asociados a los ensayos, considerada la operación más costosa computacionalmente; el próximo paso consiste en predecir el valor del punto o el valor promedio del bloque, según sea el caso; y por último calcular la varianza del error de estimación. Este conjunto de pasos se repite para cada uno de los puntos o bloques a estimar.

El algoritmo de Krigeado requiere como entradas:

- $A = \{a_1 \cdots a_n\}$ con $a_i = \{x_i, y_i, z_i, v_i\}$ ensayos o puntos medidos (representados por sus coordenadas espaciales y el valor del atributo medido).
- $M = \{b_1, b_2, \cdots, b_n\}$ modelo de bloques caracterizado por la Expresión 5 (las coordenadas del origen, dimensiones del modelo, dimensiones de los bloques, nivel de discretización de los bloques); donde cada bloque contiene $b = \{(i,j,k), (x_c, y_c, z_c), (a_1, a_2, \cdots, a_n)\}$ (su localización espacial, las coordenadas del centroide y el conjunto de ensayos).

$$p = \left\{(x_0, y_0, z_0), (nb_i * nb_j * nb_k), (lb_i * lb_j * lb_k), (db_i * db_j * db_k)\right\} \tag{5}$$

- $S = \left\{(R_x, R_y, R_z), (\alpha, \beta, \theta), (r_1, r_2, \cdots, r_n)\right\}$ vecindad de estimación delimitada por el elipsoide con radios R_x, R_y, R_z, los ángulos α, β, θ que indican la rotación del elipsoide en cada uno de los ejes, teniendo como restricciones: el número mínimo (r_1) y máximo de datos (r_2), así como el número máximo de datos por octante (r_n); representada por la Ecuación 6: donde x_0, y_0, z_0 constituyen las coordenadas del origen del centroide, y a, b, c los radios en cada uno de los ejes.

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} + \frac{(z-z_0)^2}{c^2} = 1 \tag{6}$$

- $\gamma(h) = \{nst, c_0, (c_1\gamma_1(h) \cdots c_n\gamma_n(h))\}$ variograma que expresa las características de variabilidad y correlación espacial del fenómeno que se estudia a partir de puntos muestreados, siendo *nst* la cantidad de estructuras que conforman el variograma en cada una de las direcciones, c_0 el valor de pepita, c_i las mesetas, y $\gamma_i(h)$ las estructuras de variogramas.

El algoritmo genera como salidas:

- $M = \{b_1, b_2, \cdots, b_n\}$ modelo de recursos, es decir, los puntos o bloques estimados.

1.3 Indexación y búsqueda espacial por rangos

Con el objetivo de optimizar el algoritmo, evitando la realización de búsquedas innecesarias, se propone la búsqueda por rangos. La búsqueda por rangos, esencialmente consiste en buscar los objetos geométricos que contiene una determinada región del espacio de objetos geométricos [10]. La eficiencia de las búsquedas por rangos se sustenta en la previa indexación de los objetos geométricos en una estructura de datos apropiada, en este caso considerando como estructura de datos el propio modelo de bloques.

Para la indexación, por cada posición (i,j,k) del modelo de bloques se guarda una lista con los ensayos pertenecientes a la celda o bloque. Para determinar las celdas en las cuales se encuentran los ensayos, se transforman las coordenadas de los ensayos al espacio de coordenadas de los índices del modelo, teniendo en cuenta el origen (x_0, y_0, z_0) y las dimensiones (nb_i, nb_j, nb_k) de los bloques (Ecuación 7):

$$x_i = \frac{(x-x_0)}{nb_i}; y_j = \frac{(y-y_0)}{nb_j}; z_k = \frac{(z-z_0)}{nb_k}; \tag{7}$$

Luego se obtienen las localizaciones espaciales (i,j,k) de los ensayos redondeando por defecto las coordenadas transformadas: $i = floor(x_i)$, $j = floor(y_j)$ y $k = floor(z_k)$. Después se verifica que las localizaciones espaciales obtenidas estén dentro del rango de índices del modelo de bloques, cumpliéndose que $o \leq i \leq nb_i$, $o \leq j \leq nb_j$ y $o \leq k \leq nb_k$. Una vez indexados los ensayos, para buscar un ensayo en el modelo de bloques simplemente se calcula el índice de localización espacial del punto, siendo n_x la cantidad de columnas del modelo, n_y la cantidad de filas y xyz las coordenadas espaciales del punto en cuestión (Ecuación 8):

$$loc = zn_xn_y + yn_x + x \tag{8}$$

1.4 Técnicas, herramientas y tecnologías

El diseño del algoritmo se basó en los siguientes principios: partición (descomposición de la computación de tareas), comunicación (coordinación en la ejecución de tareas), aglomeración (combinación de los resultados de las tareas) y mapeo (asignación de tareas a los procesadores); descritas en

[6], [9]. Se centró fundamentalmente en la partición y asignación. En el diseño del algoritmo no se presentaron secciones críticas, por lo que no se hacen copias de la lista de puntos a interpolar; todos los procesos pueden leer y escribir en esta lista sin que se generen conflictos de memoria.

El algoritmo de Krigeado Ordinario fue implementado usando C++11 como lenguaje de programación. Para los cálculos matriciales se utilizó la biblioteca de álgebra lineal Atlas CLapack, caracterizada por ser rápida. Esta biblioteca permitió simplificar el cálculo de operaciones matriciales como la resolución de sistemas de ecuaciones lineales y la multiplicación de matrices.

Las técnicas de programación paralela y distribuida han demostrado ser una alternativa viable para la solución rápida de este tipo de problemas computacionales. En la presente investigación se aplicaron técnicas de programación paralela en memoria compartida a través de la biblioteca OpenMP 4.8.2.

La biblioteca OpenMP se basa en el modelo *fork-join* para obtener el paralelismo a través de múltiples hilos. Aprovechando la independencia de los datos de entrada se aplica la descomposición del dominio para la división de los datos entre los procesadores. Esta técnica de descomposición consiste en determinar la partición apropiada de los datos, y luego trabajar en los cómputos asociados.

2. Resultados y discusión

2.1 Aceleración mediante OpenMP

El algoritmo propuesto para la interpolación de Krigeado Ordinario consta de 3 etapas de procesamiento paralelo a través del modelo *fork-join* (Figura 1). Este modelo plantea la división del hilo maestro en hilos esclavos que se ejecutan concurrentemente, distribuyéndose las tareas sobre estos hilos. Los hilos acceden a la misma memoria, aunque es posible gestionar estos accesos generando espacios de memoria privada. A continuación se describen las etapas de procesamiento, y se modela el algoritmo de Krigeado Ordinario en forma de pseudocódigo (Algoritmos 1 y 2).

- **Etapa I.** Se realiza la indexación espacial de los puntos en el modelo de bloques y búsqueda de vecinos a utilizar en cada una de las interpolaciones. Los bloques son distribuidos en trozos de aproximadamente igual tamaño entre los procesadores antes de que las iteraciones sean ejecutadas mediante asignaciones estáticas (*shedule static*); todas las iteraciones son repartidas de forma continua antes de ser ejecutadas.
- **Etapa II.** Se realiza el ordenamiento de los bloques a estimar en función de la cantidad de vecinos a utilizar. El modelo de bloques una vez más es particionado en tantas partes iguales como número de procesadores; luego estas partes son ordenadas simultáneamente. Una vez que concluye el proceso de ordenamiento, se mezclan los resultados arrojados por cada procesador en tantas iteraciones como sean necesarias.

Algorithm 1 Pseudocódigo del algoritmo de Krigeado Ordinario (modelo de bloques)

Require: A, M, S, $\gamma(h)$
Ensure: M
1: discretizar (M)
2: indexar (A, M)
3: $numbloques = M.cantidad$
4: {Etapa I}
5: **for** $i = 0 : numbloques$ **do**
6: $b = M.at(i)$
7: $b.vecinos =$buscar ($b.centroide$, A, M, S)
8: **end for**
9: {Etapa II}
10: ordenar (M)
11: {Etapa III}
12: **for** $i = 0 : numbloques$ **do**
13: $b = modelo.at(i)$
14: interpolar (b, A, M, S, $\gamma(h)$)
15: **end for**

Algorithm 2 Pseudocódigo del algoritmo de Krigeado Ordinario (un bloque)

Require: b, A, M, S, $\gamma(h)$
Ensure: M
1: $numvecinos = b.vecinos.cantidad$
2: **if** $numvecinos >= S.cantidadMinimaDeDatos$ **then**
3: $matLM =$ construirMatrizIzquierda (b, A, $\gamma(h)$)
4: $matRM =$ construirMatrizDerecha (b, A, $\gamma(h)$)
5: $matR =$ resolverSEL (LM, RM)
6: $valor = 0$
7: $error = 0$
8: $varianza =$ calcularVarianzaBloque (b, A, $\gamma(h)$)
9: $\mu = R[numvecinos]$
10: **for** $i = 0 : numvecinos$ **do**
11: $valor+ = b.vecinos.at(i).valor$
12: $error+ = R[i] * RM[i]$
13: **end for**
14: $b.valor = valor$
15: $b.error = varianza - error - \mu$
16: **else**
17: $b.valor = NE$ {no estimado}
18: $b.error = NE$ {no estimado}
19: **end if**

- **Etapa III.** Se realiza la interpolación. Los bloques se distribuyen nuevamente entre los procesadores, pero esta vez de forma dinámica (*shedule dynamic*), es decir las iteraciones son asignadas de forma continua a solicitud de los procesadores, hasta que se acaben. En esta etapa se calculan los pesos y los valores asociados a los puntos a interpolación.

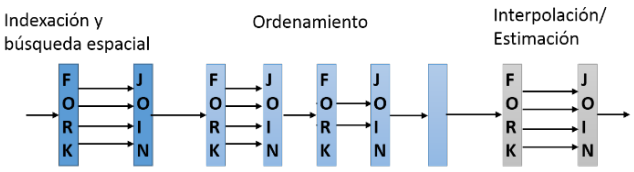


Figura 1. Utilización del modelo *fork-join* para lograr el paralelismo de datos en el algoritmo.

2.2 Experimentos y resultados

Se realizó la evaluación experimental del algoritmo variando el tamaño de entrada *(n)* y el número de procesadores *(p)*, atendiendo al tiempo de ejecución, la ganancia de velocidad (*Speed Up*) y la eficiencia *(E)*. Se entiende por tiempo de ejecución como el tiempo que transcurre desde el comienzo de la ejecución del programa en el sistema paralelo, hasta que el último procesador culmine su ejecución. A la ganancia de velocidad también se le conoce como aceleración, y consiste en la relación entre el tiempo de ejecución sobre un procesador secuencial y el tiempo de ejecución sobre múltiples procesadores. Por eficiencia se entiende el porcentaje de tiempo empleado en proceso efectivo; este indicador mide el grado de utilización de un sistema multiprocesador.

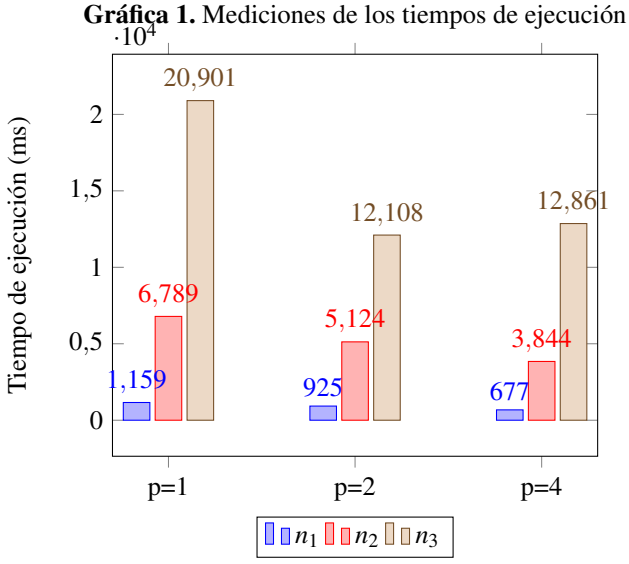
Los experimentos se ejecutaron en una estación de trabajo Acer Aspire 5755 con procesador Intel (R) Core (TM) i5-2430 (compuesta por 2 núcleos físicos y 2 virtuales para un total de 4 procesadores), con una frecuencia de 2.40 GHz, 4 GB de memoria instalada y sistema operativo Ubuntu 14.04 (32 bits). El proceso de Krigeado se realizó de forma puntual considerando los centroides de los bloques como la nube de puntos a estimar. Se realizaron 10 corridas en cada experimento, descartándose los datos atípicos a través del método de los cuartiles.

Se utilizaron 3 juegos de datos generados aleatoriamente variando el número de ensayos. Los juegos de datos estaban compuestos por 1000 bloques y 1000 ensayos (n_1), 1000 bloques y 2000 ensayos (n_2), 1000 bloques y 3000 ensayos (n_3), respectivamente. Se utilizó un modelo de variograma lineal de 1 como valor de pepita, y 10 como valor de la pendiente. Se utilizó una vecindad esférica de 5 metros de radio. Los Cuadros 1, 2 y 3 contienen los resultados de los experimentos realizados usando 1, 2 y 4 procesadores respectivamente, así como la descripción de dichos resultados a partir de estadígrafos como la mediana, cuartil menor y mayor y el promedio.

Cuadro 1. Mediciones de los tiempos de ejecución del algoritmo paralelo para un procesador (ms).

Corrida	t(n_1)	t(n_2)	t(n_3)
1	1150	6687	21023
2	1144	6820	20955
3	1152	6756	20832
4	1176	6884	20799
5	1171	6803	20884
6	1162	6809	20863
7	1155	6829	20918
8	*1215	6789	20841
9	1175	6700	20914
10	1150	6811	20985
Mediana	1158.50	6806.00	20899.00
Cuartil menor	1150	6756	20841
Cuartil mayor	1175	6820	20955
Rango intercuartil	25	64	114
Límite inferior	1112.5	6660	20670
Límite superior	1212.5	6916	21126
Promedio	1159.44	6788.8	20901.4
Tiempo de ejecución	1159	6789	20901

Gráfica 1. Mediciones de los tiempos de ejecución



Los datos arrojados por la evaluación experimental del algoritmo evidencian una disminución de los tiempos requeridos en la interpolación de Krigeado Ordinario (Gráfica 1). La mayor ganancia de velocidad (1.77) se obtuvo para un tamaño de entrada de m=1000 (bloques) y n=2000 (ensayos), donde se logró disminuir el tiempo de ejecución de 6789 ms a 3844 ms al utilizar 4 procesadores. La ganancia de velocidad incrementa al aumentar el tamaño de entrada atendiendo a su valor óptimo cuando se utilizan 2 procesadores (Gráfica 2).

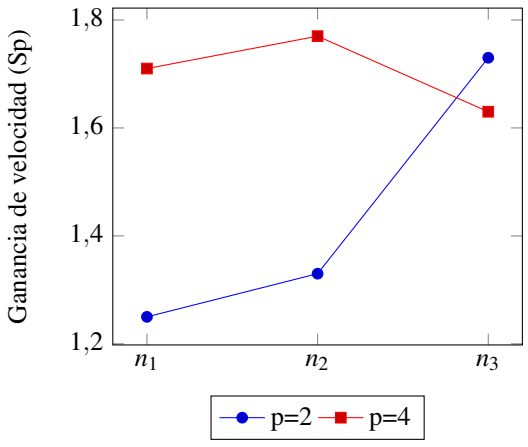
Cuadro 2. Mediciones de los tiempos de ejecución del algoritmo paralelo para 2 procesadores (ms).

Corrida	t(n_1)	t(n_2)	t(n_3)
1	721	4053	12094
2	1190	4227	*12799
3	786	4033	12064
4	760	4081	12093
5	1136	4700	4700
6	1176	6193	12039
7	757	6404	12158
8	1189	4020	12135
9	766	7085	12144
10	772	6445	12155
Mediana	779.00	4463.50	12114.50
Cuartil menor	760	4053	12087
Cuartil mayor	1176	6404	12155
Rango intercuartil	416	2351	68
Límite inferior	136	526.5	11985
Límite superior	1800	9930.5	12257
Promedio	925.3	5124.1	12107.66
Tiempo de ejecución	925	5124	12108
Ganancia de velocidad	1.25	1.33	1.73
Eficiencia	62.5	66.5	86.5

Cuadro 3. Mediciones de los tiempos de ejecución del algoritmo paralelo para 4 procesadores (ms).

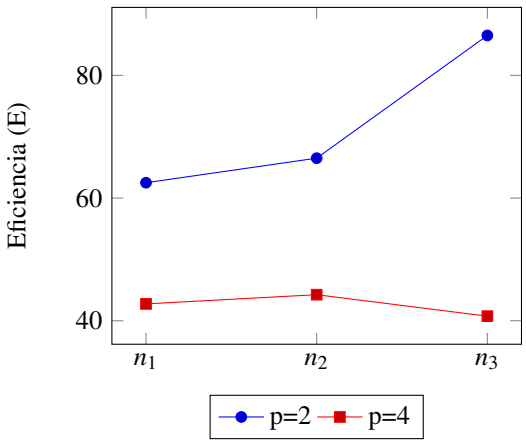
Corrida	t(n_1)	t(n_2)	t(n_3)
1	694	3781	12379
2	*757	3894	12379
3	706	3834	12437
4	672	3878	12352
5	683	3793	13196
6	685	3985	13042
7	662	3821	13434
8	664	3796	12494
9	674	3796	13188
10	657	3860	13708
Mediana	678.50	3827.50	12768.00
Cuartil menor	664	3796	12379
Cuartil mayor	694	3878	13196
Rango intercuartil	30	82	817
Límite inferior	619	3673	11153.5
Límite superior	739	4001	14421.5
Promedio	677.44	3843.8	12860.9
Tiempo de ejecución	677	3844	12861
Ganancia de velocidad	1.71	1.77	1.63
Eficiencia	42.75	44.25	40.75

Gráfica 2. Ganancia de velocidad del algoritmo.



Por otro lado, los valores de eficiencia indican un aprovechamiento del 62.5 % al 86.54 % de las capacidades de procesamiento al utilizar 2 procesadores, y un aprovechamiento cercano a la mitad al utilizar 4 procesadores. Se evidencia un notable incremento de la eficiencia al aumentar los tamaños de entrada con el uso de 2 procesadores (Gráfica 3).

Gráfica 3. Eficiencia del algoritmo (%).



Se observa una disminución de la ganancia de velocidad y la eficiencia al utilizar 4 procesadores para un tamaño de entrada de 1000 bloques y 3000 ensayos, lo cual está influenciado por el procesamiento de ordenamiento de los datos en busca de equilibrar la carga entre los procesadores. Si bien el ordenamiento por mezcla fue implementado de forma paralela, a medida que se incrementan las iteraciones se van desechando procesadores al mezclar los resultados parciales. Esto se va acentuando en la medida que aumentan los tamaños de entrada y el número de procesadores.

2.3 Conclusiones

- La complejidad de los cálculos matemáticos utilizados en la interpolación de Krigeado Ordinario, fundamental-

mente la resolución de grandes sistemas de ecuaciones, repercute en altos tiempos de ejecución que retardan los procesos de estimación.

- El carácter independiente de los datos utilizados en el proceso de Krigeado Ordinario favorece la utilización del paralelismo a nivel de datos como una alternativa eficiente para disminuir los tiempos de respuesta asociados.
- Las técnicas de programación paralela en memoria compartida facilitan la explotación del paralelismo de datos a través del uso de bucles iterativos para la distribución de tareas a los procesadores, propiciando un mejor aprovechamiento de las capacidades de cómputo.
- La subdivisión de los bloques en diferentes procesadores es muy útil cuando se trabaja con grandes cantidades de datos, pero si no se logra una distribución equitativa el aprovechamiento de los procesadores no será óptimo.
- Los métodos de búsqueda espacial por rangos sustentados en la indexación espacial de los objetos geométricos contribuyen a disminuir los tiempos de respuesta en el proceso de Krigeado, al evitar búsquedas innecesarias.
- La reducción de los tiempos asociados a la interpolación de Krigeado Ordinario soportará la toma de decisiones rápidas, por ejemplo: durante la evaluación de la factibilidad de los proyectos, así como la planificación minera de estos en condiciones de rentabilidad económica.

Agradecimientos

Se agradece la colaboración del especialista José Arias de la Oficina Nacional de Recursos Minerales (ONRM).

Referencias

[1] M. A. Alfaro. *Estimación de recursos mineros*. 2007.

[2] M. Armstrong and J. Carignan. *Géostatistique Linéaire, Application au Domaine Minier*. École de Mines de Paris, 1997.

[3] M. Chica. *Análisis Geoestadístico en el Estudio de la Explotación de Recursos Minerales*. PhD thesis, Universidad de Granada, 1997.

[4] J. Deraisme and Ch. De Fouquet. The geostatistical approach for reserves. *Minig Magazine*, 1996.

[5] M. A. Díaz. *Geoestadística Aplicada*. 2002.

[6] I. Foster. *Designing and Building Parallel Programs*. ISBN:0201575949. Addison Wesley, 1995.

[7] K. E. Kerry and K. A. Hawick. Kriging interpolation on high-performace computers. Technical report, Department of Computer Science, Universidad de Adelaide, 1998.

[8] C. D. Lloyd. *Local models for spatials analysis*. ISBN 9780415316811. First Edition CRC Press, 2006.

[9] R. M. Naiouf. *Procesamiento paralelo. Balance de carga dinámico en algortimos de sorting*. PhD thesis, Universidad Nacional de La Plata, 2004.

[10] E. Olinda and G. Hernández. Un enfoque propuesto para las búsquedas por rangos. Technical report, Proyecto de la UPM, 2002.

[11] A. Pesquer, Ll.; Cortés and X. Pons. Parallel ordinary kriging interpolation incorporating automatic variogram fitting. *Computers and Geosciences*, pages 464–473, 2011.

[12] D. Sullivan and D. Unwin. *Geographic Information Analysis*. John Wiley - Sons Hoboken, 2002.

[13] R. Vasan, B.; Duraiswami and R. Murtugudde. Efficient kriging for real-time spatio-temporal interpolation. In *20th Conference on Probability and Statistics in the Atmospheric Sciences*, 2010.

Ciencias Matemáticas, Vol. 32, No. 2, Pag. 149-183, 2018
Recibido 02-2018

Implementación de un Esquema de Aproximación para el Problema de Transferencia de Masas

Miríam G. Báez-Hernández[1], M. Lorena Avendaño-Garrido [1], J. Rigoberto Gabriel-Argüelles[1]

Resumen En este artículo se presentan resultados numéricos de la implementación de un esquema de aproximación para problemas de programación lineal infinita. En particular, este esquema es aplicado al Problema de Transferencia de Masas de Monge-Kantorovich. Se ilustra el esquema con ejemplos en el intervalo $[0,1]$, los cuales tienen solución exacta, lo cual permite comparar los resultados del esquema.

Abstract

Palabras Clave
Esquema de Aproximación — Programación Lineal — Transferencia de Masas

[1] *Facultad de Matemáticas, Univesidad Veracruzana, México, miriam.baez.hdez@gmail.com, maravendano@uv.mx,jgabriel@uv.mx*

Introducción

En 1781 el matemático francés Gaspard Monge planteó el problema de transferencia de masas, el cual consiste en encontrar un plan óptimo de transporte para mover un montículo de tierra a un hoyo. Para ello, Monge dividió el montículo en granos de tierra, así el problema se redujo a encontrar una función que proporcionara la posición del granito en el hoyo. Hoy, esto es conocido como el Problema de Monge y la función de transferencia se conoce como acoplamiento óptimo [19].

Más tarde, en 1942 el matemático ruso Leonid V. Kantorovich estudió el Problema de Translocación de Masas [14], el cual consiste en minimizar el trabajo de translocación del movimiento de una distribución de masa inicial a una distribución de masa final. Para su estudio Kantorovich consideró espacios métricos compactos, conjuntos de Borel y una función de costo no negativa. Posteriormente, en 1948 observó que si en el Problema de Translocación se consideraba a la función de costo como la distancia, este resultaba una generalización del Problema de Monge [15], desde entonces a este problema se conoce como el Problema de Transferencia de Masas de Monge-Kantorovich. Éste es un problema muy conocido en diferentes áreas como: Probabilidad, Análisis Funcional, Geometría Diferencial, Estadística, Economía, Sistema Dinámicos, Computación [20]. Entre las aplicaciones más importantes que tiene el Problema de Transferencia de Masas podemos mencionar a la métrica de Kantorovich [20], en los últimos años la métrica se ha empleado para registro de imágenes [11] y control de radioterapia de cáncer [12].

La métrica de Kantorovich también ha sido empleada para comparar imágenes [13], a una imagen digitalizada se le puede asociar una función de distribución de probabilidad y esta a su vez una medida de probabilidad. Por lo tanto, para comparar imágenes se requiere de una métrica probabilística, sin embargo, calcular numéricamente esta métrica no ha sido una tarea sencilla. Por ejemplo, la métrica de Kantorovich es un problema complicado de resolver, en virtud que es un problema de optimización en espacios de medidas. En la literatura existen algoritmos de aproximación que se pueden emplear para la métrica de Kantorovich, como en [9], donde se desarrolla un algoritmo numérico para aproximar el valor del Problema de Transferencia de Masas en espacios compactos. Diversos autores han estudiado esquemas de aproximación para el Problema de Transferencia de Masas, por ejemplo Hernández-Lerma y Lasserre proponen en [16] un esquema de aproximación general basado en problemas de programación lineal infinita, el cual se puede aplicar a un problema de control de Markov, programación semi-infinita y al problema de transferencia de masas, entre otros esquemas para el Problema de Transferencia de Masas han sido estudiados por Benamou y Brenier [6], Caffarelli [8], Benamou [5] y Guittet [10]. Recientemente por Bocs [7] y Mérigot [18].

1. El Problema de Transferencia de Masas de Monge-Kantorovich

Sean $\mathscr{X}$ y $\mathscr{Y}$ dos espacios métricos compactos, con $\sigma-$álgebras de Borel $\mathbb{B}(\mathscr{X})$ y $\mathbb{B}(\mathscr{Y})$ respectivamente, consideremos además una función medible definida en $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, tomemos medidas de probabilidad ν_1 en $\mathscr{X}$ y ν_2 en $\mathscr{Y}$, y una medida de probabilidad μ en $\mathscr{X} \times \mathscr{Y}$, denotamos por $\Pi_1\mu$ y $\Pi_2\mu$ las marginales o proyecciones de μ en $\mathscr{X}$ y $\mathscr{Y}$, respectivamente. Entonces el problema de transferencia de masas MT está definido por

$$\text{MT}: \text{ Minimizar } \quad \langle \mu, c \rangle := \int c d\mu, \tag{1}$$

sujeto a:

$$\Pi_1\mu = \nu_1, \ \ \Pi_2\mu = \nu_2, \ \ \mu \geq 0. \tag{2}$$

Con su respectivo problema dual

$$\text{MT}^*: \text{ Maximizar } \int f_1 d\nu_1 + \int_Y f_2 d\nu_2$$

sujeto a:

$$f_1(x) + f_2(y) \leq c(x,y), \tag{3}$$

donde $f_1 : \mathscr{X} \to \mathbb{R}$ y $f_2 : \mathscr{Y} \to \mathbb{R}$.

Ver [16].

Diremos que μ es una solución factible para el problema MT si satisface las restricciones (1), además se dice que el programa es consistente si tiene al menos una solución factible. El programa es soluble si existe una solución factible que alcanza su valor óptimo. En la literatura una de las hipótesis requeridas para que el problema sea soluble es la presentado a continuación.

Suposición 1 **(a)** El programa $\mathbb{P}$ es consistente.

(b) El programa $\mathbb{P}$ tiene una solución factible x_0 con $\langle x_0, c\rangle > 0$ y, más aún, el conjunto

$$\mathscr{X}_0 := \{x \in X_+ | \langle x, c\rangle \leq \langle x_0, c\rangle\}$$

es débilmente, secuencialmente compacto.

Ver [16].

Decimos que no hay abertura de dualidad, si el valor del programa primal y el valor del programa dual coinciden. Algunos ejemplos, donde se cumple esta condición se pueden ver en [1].

Teorema 1 Si la Suposición 1 se cumple, entonces

(a) $\mathbb{P}$ es un programa soluble.

(b) No hay abertura de dualidad.

Ver [16].

2. Esquema de Aproximación

El esquema mostrado a continuación es una aplicación del propuesto en [16], el cual consiste de tres pasos: el primero nos asegurará que el programa de programación lineal infinita es soluble, después se agregan restricciones finitas al programa y se relajan las restricciones, finalmente se discretiza la variable de interés.

Denotemos como $\text{sp}(E)$ al espacio generado por el conjunto E, el cual es un subconjunto de un espacio vectorial, y $\text{coc}(E)$ es el cono convexo generado por E.

Requerimos lo siguiente:

Suposición 2

(a) X y Y son espacios métricos compactos.

(b) La función $c(x,y)$ es continua.

Ver [9]

Consideremos $X = Y = [0,1]$, además $(\mathscr{X}, \mathscr{Y})$ y $(\mathscr{Z}, \mathscr{W})$ dos parejas duales, tal que $\mathscr{X} = M([0,1]\times[0,1])$, $\mathscr{Y} = C([0,1]\times[0,1])$, $\mathscr{Z} = (M([0,1])\times M([0,1]))$ y $\mathscr{W} = (C([0,1])\times C([0,1]))$, donde $M([0,1])$ es el espacio vectorial de las medidas signadas en $[0,1]$, y $C([0,1])$ es el espacio de funciones continuas en $[0,1]$.

Denotamos $M_+([0,1])$ el cono positivo definido como

$$M_+([0,1]) := \{\mu \in M([0,1]) | \mu \geq 0\},$$

y

$$M_1([0,1]) := \{\mu \in M([0,1])_+ | \mu([0,1]) = 1\},$$

la familia de medidas de probabilidad.

Como $[0,1]$ es un espacio métrico compacto, $\mathscr{X}$ es el dual topológico de $\mathscr{Y}$, así $M_1([0,1]\times[0,1])$ es secuncialmente compacto en la topología débil $\sigma - (\mathscr{X}, \mathscr{Y})$, lo cual implica que la Suposición 1 se cumple. Además, el espacio $\mathscr{W} = C([0,1])\times C([0,1])$ es separable.

Sea W_∞ un subconjunto denso numerable de $\mathscr{Y}$, y sea $\{W_k\}$ una sucesión creciente de conjuntos finitos tal $W_k \uparrow W_\infty$.

Para cada k, consideremos la agregación

$$MT(W_k): \text{ Minimizar } \langle \mu, c\rangle := \int c d\mu,$$

sujeto a:

$$\langle (\pi_1\mu, \pi_2\mu) - (\nu_1, \nu_2), w_k\rangle = 0 \tag{4}$$

$$\forall w_k := (w_i^k,\ w_j^k) \in W_k, \quad \text{y} \quad \mu \in M_+([0,1]\times[0,1]).$$

Ahora, las restricciones de igualdad serán relajadas a desigualdades. Sea $\{\varepsilon_k\}$ una sucesión de números positivos tal que $\varepsilon_k \downarrow 0$. Entonces para cada $k = 1,2,\ldots$ consideramos el siguiente problema en programación lineal infinita

$$MT(W_k, \varepsilon_k): \text{ Minimizar } \langle \mu, c\rangle,$$

sujeto a:

$$|\langle (\pi_1\mu, \pi_2\mu) - (\nu_1, \nu_2), w_k\rangle| \leq \varepsilon_k \tag{5}$$

$$\forall w_k \in W_k, \quad \text{y} \quad \mu \in M_+([0,1]\times[0,1]).$$

Finalmente, sea X_+^∞ un subconjunto denso numerable de $\mathscr{X}_+ := M([0,1]\times[0,1])_+$, y $\{X_n\}$ una sucesión creciente de conjuntos finitos, tal que $X_n \uparrow X_+^\infty$. Consideremos el programa en programación lineal finita

$$\overline{MT} := MT(X_n, W_k, \varepsilon_k): \text{ Minimizar } \sum_x n^* \lambda_x \langle \mu, c\rangle,$$

sujeto a:

$$\left|\left\langle \sum_x^{n^*} (\pi_1\mu, \pi_2\mu) - (\nu_1, \nu_2), w_k \right\rangle\right| \leq \varepsilon_k \tag{6}$$

$$\forall w_k \in W_k,\ \ \mu \in M_+([0,1]\times[0,1] \quad \text{y} \quad n^* := |X_n|.$$

A continuación se realizará la implementación del esquema. Sea $X = Y = [0,1]$ con la topología usual, y $\nu_1, \nu_2 = m$ la media de Lebesque. Consideremos las siguientes parejas duales $(\mathscr{X}, \mathscr{Y})$ y $(\mathscr{Z}, \mathscr{W})$.

Además, consideremos los siguientes conjuntos

$$W_k = \left\{ (B_i^k, B_j^k) \middle| B_{i,j}^k \in \{B_0^k, B_1^k, \ldots, B_k^k\} \right\},$$

donde B_i^k denota un polinomio de Berstein de grado k, el cual satisface

$$B_i^k(x) = \binom{k}{i} x^i (1-x)^{k-i} \quad \text{y} \quad \int_0^1 B_i^k(x)dx = \frac{1}{k+1}$$

Luego, $W_\infty := \bigcup_{k=1}^{\infty} \text{sp}\{W_k\}$ es débilmente denso en $C([0,1]) \times C([0,1])$, y la cardinalidad del conjunto W_k es $(k+1)^2$.

Para cada $n \in \mathbb{N}$, consideremos el siguiente conjunto

$$X_n = \left\{ \delta_{(a,b)}(\cdot) \right\},$$

de medidas de Dirac, para a y b en

$$\left\{ \frac{j}{2^n} \text{ con } j = 0, \ldots, 2^n \right\},$$

Entonces, $\mathscr{X}_+^\infty := \cup_{n=1}^\infty \text{coc}(X_n)$ es débilmente denso en $\mathscr{X}_+$, y la cardinalidad del conjunto X_n es 2^{2n}. Entonces, el esquema de aproximación es

$$\overline{MT}: \quad \text{Minimizar} \quad \sum_{(a,b)} \lambda_{(a,b)} c_{(a,b)}, \tag{7}$$

sujeto a:

$$\sum_{(a,b)} \lambda_{(a,b)} \left[B_i^k(a) + B_j^k(b) \right] \leq \frac{2}{k+1} + \varepsilon, \tag{8}$$

$$\sum_{(a,b)} \lambda_{(a,b)} \left[B_i^k(b) + B_j^k(b) \right] \geq \frac{2}{k+1} + \varepsilon, \tag{9}$$

donde $\quad c_{(a,b)} = c(a,b).$

Teorema 2 Si la Suposición 1 se cumple, entonces para cada k existe $n(k)$ tal que para cada $n \geq n(k)$, $\text{MT}(X_n, W_k, \varepsilon)$ es soluble y

$$\min \text{MT}(W_k, \varepsilon_k) \leq \min \text{MT}(X_n, W_k, \varepsilon_k) \leq \min \text{MT} + \varepsilon_k.$$

Ver [16].

3. Ejemplos Numéricos

A continuación mostramos ejemplos a los cuales le aplicamos el esquema , estos fueron tomados de [2], de los cuales se conoce el valor óptimo del programa.

Ejemplo 1 Consideremos la función de costo $c(a,b) = ab(a-b)$, en el Problema de Transferencia de Masas, que sabemos tiene valor óptimo $-\frac{9}{256} \approx -0{,}03515625$ y tiene soporte en el conjunto

$$\text{Graph}(g) = \{(t, g(t)) | t \in [0,1]\},$$

con

$$g(t) = \begin{cases} \frac{1}{4} + t & if \quad 0 \leq t < \frac{3}{4}, \\ 1 - t & if \quad \frac{3}{4} \leq t \leq 1, \end{cases}$$

y consideramos $\varepsilon_k^n = \frac{1}{10^{n-2}\sqrt{k}}$.

Variamos el parámetro k desde 2 a 9, y para cada k movemos a n desde 4 a 9. De lo cual obtuvimos los siguientes resultados, los cuales indican los errores de aproximación del esquema al valor óptimo del programa.

	$n=4$	$n=5$
$k=2$	0,017380600	0,0176040700
$k=3$	0,011528550	0,0108222500
$k=4$	0,004827046	0,0026806390
$k=5$	0,004427841	0,0020611890
$k=6$	0,003936912	0,0019239340
$k=7$	0,002954513	0,0009981689
$k=8$	0,002777794	0,0009861557
$k=9$	0,002719209	0,0009109657

	$n=6$	$n=7$
$k=2$	0,01890373000	0,019683010
$k=3$	0,01139729000	0,011758030
$k=4$	0,00247768000	0,002455755
$k=5$	0,00186624500	0,001852651
$k=6$	0,00183779700	0,001826349
$k=7$	0,00081778760	0,0008197949
$k=8$	0,00098615560	0,0008114302
$k=9$	0,00069509160	0,0006775082

	$n=8$	$n=9$
$k=2$	0,0200872300	0,0202909700
$k=3$	0,0119523300	0,0120497000
$k=4$	0,0024566440	0,0024564520
$k=5$	0,0018515210	0,0018524900
$k=6$	0,0018300650	0,0018318280
$k=7$	0,0008302774	0,0008358459
$k=8$	0,0008186530	0,0008224715
$k=9$	0,0006760644	0,0006761160

De los resultados obtenidos por el esquema, por cada fila se muestra el mejor resultado, y por cada conjunto de datos se muestra el mejor resultado del esquema, el cual se obtuvo en $k=9$ y $n=8$. Con un error de 0,0006760644 y cuyo valor de aproximación es $-0{,}03583231$.

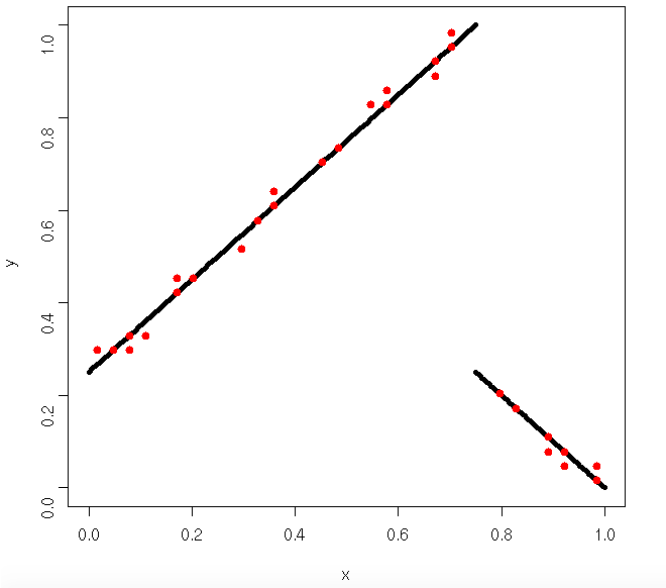


Figura 1. Soporte de la solución óptima a $\overline{MT}$ con $k = 9$ y $n = 8$.

La Figura 1 muestra la gráfica del soporte solución y los puntos rojos muestran el soporte para la medida de aproximación en el caso $n = 8$ y $k = 9$.

Ejemplo 2 Consideremos la función de costo $c(a,b) = (2b - a - 1)^2(2b - a)^2$, en el Problema de Transferencia de Masas, sabemos que tiene valor óptimo 0 y soporte en el conjunto

$$\text{Graph}(h) = \{(t, h(t)) | t \in [0,1]\},$$

con $h : [0,1] \times [0,1]$ tal que

$$h(t) = \begin{cases} \frac{1}{2}t & si \quad 0 \leq t \leq 1, \\ \\ \frac{1}{2}t + \frac{1}{2} & si \quad 0 \leq t \leq 1, \end{cases}$$

y consideramos $\varepsilon_k^n = \dfrac{1}{10(n-2)\sqrt{k}}$.

Variamos el parámetro k desde 2 a 10, y para cada k movemos a n desde 4 a 10. De lo cual obtuvimos los siguientes resultados, los cuales indican los errores de aproximación del esquema al valor óptimo del programa.

	$n = 5$	$n = 6$
$k = 2$	0,0002363199	0,00006007883
$k = 3$	0,0002362977	0,00006007827
$k = 4$	0,0002362751	0,00006007770
$k = 5$	0,0002362534	0,00006007715
$k = 6$	0,0002362328	0,00006007662
$k = 7$	0,0002362132	0,00006007975
$k = 8$	0,0002361945	0,00006007689
$k = 9$	0,0002361778	0,00006007810
$k = 10$	0,0002361673	0,00006007816

	$n = 7$	$n = 8$
$k = 2$	0,00001513965	0,0000037998070
$k = 3$	0,00001513964	0,0000037998060
$k = 4$	0,00001513962	0,0000037998065
$k = 5$	0,00001513961	0,0000038160170
$k = 6$	0,00001516782	0,0000038280000
$k = 7$	0,00001528733	0,0000038113430
$k = 8$	0,00001529179	0,0000038179810
$k = 9$	0,00001516529	0,0000038174210
$k = 10$	0,00001527237	0,0000038147230

	$n = 9$	$n = 10$
$k = 2$	0,0000009518125	0,0000002381858
$k = 3$	0,0000009518126	0,0000002384716
$k = 4$	0,0000009523129	0,0000002382365
$k = 5$	0,0000009530545	0,0000002384016
$k = 6$	0,0000009520117	0,0000002383636
$k = 7$	0,0000009528386	0,0000002383380
$k = 8$	0,0000009531212	0,0000002384158
$k = 9$	0,0000009537839	0,0000002384849
$k = 10$	0,0000009540616	0,0000002384848

De los resultados obtenidos por el esquema, por cada fila se muestra el mejor resultado, y por cada conjunto de datos se muestra el mejor resultado del esquema, el cual se obtuvo en $k = 10$ y $n = 10$. Con un error de 0,0000002384848 y cuyo valor de aproximación es 0,0000002384848.

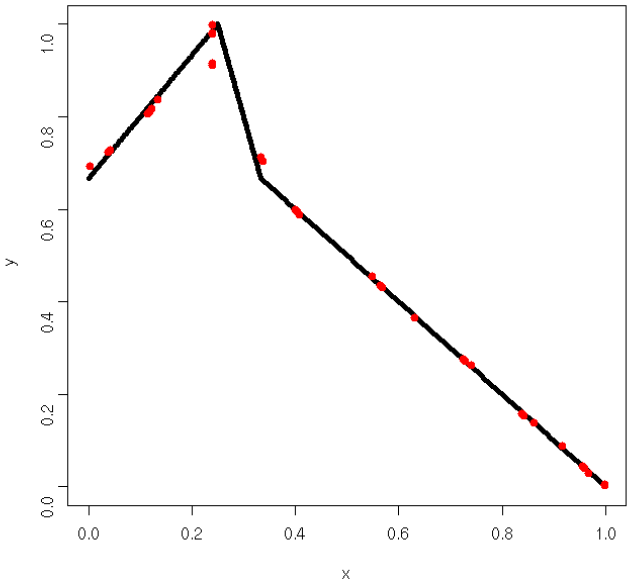


Figura 2. Soporte de la solución óptima a $\overline{MT}$ con $k = 10$ y $n = 10$.

La Figura 2 muestra la gráfica del soporte solución y los puntos rojos muestran el soporte para la medida de aproximación en el caso $n = 10$ y $k = 10$.

4. Conclusiones

La aplicación del esquema propuesto en [16], al Problema de Transferencia de Masas de Monge-Kantorovich es eficiente, y el tiempo computacional es corto, pero los valores de los parámetros están limitados por la capacidad de la computadora, ya que el número de variables es muy grande, incluso si n y k son pequeños. Una continuación a este esquema sería aplicar una meta-heurística con el fin de reducir el número de variables y así aumentar los valores de n y k para obtener una mejor aproximación. Además de encontrar un orden de convergencia.

Referencias

[1] Anderson, E., Nash P. (1987): Linear programming in infinite-dimensional spaces: theory and applications. Wiley, Chichester.

[2] Anderson, E., Philpott, A. (1984):Duality and an algorithm for a class of continuous transportation problems. Mathematics of Operations Research, 9, 222-231.

[3] Avendaño-Garrido M.L., Gabriel J.R., Quintana-Torres L., and González-Hernández J. (2017): An Approximation Scheme for the Kantorovich-Rubistein Problem on Compact Spaces. International Journal of Numerical Methods and Applications, 16, pp 107-125.

[4] Avendaño-Garrido M.L., Gabriel J.R., Quintana-Torres L., and Mezura-Montes E. (2016): A metaheuristic for a numerical approximation to the mass transfer problem. Journal of Optimization Int. J. Appl. Math Computa. Sci., 26(4), pp. 757-766.

[5] Benamou, J.(2003): Numerical resolution of an unbalanced mass transport problem. ESAIM Mathematical Modelling and Numerical Analysis 37, pp. 851-868.

[6] Benamou, J., Brenier, Y.: (2000): A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik 84, pp. 375-393.

[7] Bosc, D. (2010): Numerical approximation of optimal transport maps. SSRN.

[8] Caffarelli, L., Feldman, M., McCann, R.(2002): Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. Journal of the American Mathematical Society 15, pp. 1-26.

[9] Gabriel J.R., González-Hernández J. and López-Martínez R.R. (2010): Numerical aproximations to the mass transfer problem on compact spaces. IMA. J. Numer. Anal., 30(4), pp. 1121-1136.

[10] Guittet, K. (2003): On the time-continuous mass transport problem and its approximation by augmented lagrangian techniques. SIAM Journal on Numerical Analysis 41, pp. 382-399.

[11] Haker S., Zhu L. and Tannenbaum A. (2004): Optimal mass transport for registration and warping. Int. J. Comput. Vis., 63, pp. 225-240.

[12] Hanin, L., Rachev, S.T. and Yakolev A.Y. (1993): On optimal control of cancer radiotherapy for nonhomogeneous cell populations. Adv. Appl. Prob., 25, pp. 1-23.

[13] Kaijset T. (1998): Computing the Kantorovich distance for images. Int. J. Comput. Vis., 9, pp. 173-191.

[14] Kantorovich L.V. (2006): On the translocation of masses. J. Math. Sci. (N.Y.), 133(4), pp. 1383.

[15] Kantorovich L.V. (2006): On a problem of Monge. J. Math. Sci. (N.Y.), 133(4), pp. 1383.

[16] Hernández-Lerma, O., Lasserre, J. (1998): Approximation schemes for infinite linear programs. SIAM Journal on Optimization 8, pp. 973-988.

[17] Levin, V.L. (1975): On the mass transfer problem. Soviet Math. Dokl. 16 (5), pp. 1349-1353.

[18] Mérigot, Q. (2011): A multiscale approach to optimal transport. Computer Graphics Forum 30(5), pp.1583-1592.

[19] Monge, G. (1781): Mémoire sur la théorie des déblais et rémblais. Mém. Acad. Sci., Paris.

[20] Rachev, S.T. and Rüschendorf, L. (1998): Mass Transportation Problems Vol. 1 and 2, Springer, New York.