Búsqueda de entorno variable multiobjetivo para resolver el problema de particionamiento de datos espaciales con características poblacionales

María Beatríz Bernábe Loranca y Carlos Guillén Galván

Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Puebla, Puebla,
México

{beatriz.bernabe, carlosguillen.galvan }@gmail.com

Resumen. El problema de particionamiento siendo un problema NP difícil, ha sido ampliamente estudiado debido a varias razones, en particular, por su vulnerabilidad al obtener óptimos locales de los criterios que optimiza. Para problemas de particionamiento en optimización combinatoria, existen diversos trabajos que han propuesto la inclusión de heurísticas con el fin de lograr óptimos globales. Muchos han sido los esfuerzos para resolver el particionamiento y encontrar buenas soluciones cuando en el proceso de optimización discreta se optimiza un solo objetivo, sin embargo, ha sido poco atendido el problema de particionamiento con más de un objetivo debido a la dificultad de obtener el conjunto de soluciones eficientes, óptimas y no dominadas. En este trabajo se expone el problema de multiobjetivo en particionamiento para datos espaciales con dos objetivos: minimización de distancias y de variables censales. El algoritmo de particionamiento que se ha diseñado es una extensión del grupo geográfico que optimiza solo un objetivo. En este trabajo para escapar de óptimos locales se ha hecho uso de Búsqueda por Entorno Variable (VNS) y para obtener el conjunto de soluciones no dominadas se han aprovechado las propiedades del conjunto Máxima.

Palabras Clave. Algoritmos heurísticos, Máxima, particionamiento multiobjetivo.

Multi-Objective Variable Neighborhood Search to Solve the Problem of Partitioning of Spatial Data with Population Characteristics

Abstract. The problem of partitioning is NP hard and has been studied extensively for several reasons including vulnerability to obtain local optima. For partitioning problems in combinatorial optimization, several works have proposed the inclusion of heuristics in order to achieve global optima. There have been made many efforts to solve the partitioning problem and find good solutions when the discrete optimization process optimizes a single objective. However, the partitioning problem with more than one goal has not been addressed due to the difficulty of obtaining the set of efficient optimal and non-dominated solutions. This paper presents the multi-objective partitioning problem with two objectives: minimization of distances and of census variables. The designed partitioning algorithm is an extension of the geographic cluster that optimizes only one objective. In this work, we used Variable Neighborhood Search (VNS) to escape local optima, and to obtain the set of non-dominated solutions, our methodology takes advantage of the properties of the set Maxima.

Keywords. Heuristics algorithms, Maxima, multi-objective partitioning.

1 Introducción

Los problemas de particionamiento han sido ampliamente estudiados en la literatura, sin embargo, su carácter combinatorio hace difícil su resolución mediante métodos exactos debido a la complejidad computacional asociada a la categoría NP duro. Cuando solo un objetivo es tratado, una respuesta a estos problemas es utilizar heurísticas que intenten resolver el problema mediante la aplicación de criterios de búsqueda, de este modo es posible obtener buenos resultados en tiempo computacional aceptable [10]. Aun cuando el particionamiento es estudiado con el fin de obtener mejores

soluciones, un reto importante en esta área es proponer esquemas para resolver el problema de particionamiento multiobjetivo. La dificultad reside en resolver al menos dos aspectos: (a) la generación de un conjunto de soluciones subóptimas y no dominadas y (b) el diseño de un algoritmo de particionamiento que construya grupos (clusters), satisfaciendo simultáneamente dos funciones objetivo.

La existencia de problemas sobre agrupamiento para datos espaciales implica proponer soluciones en el marco del particionamiento con el uso de métodos heurísticos, incluso con la incorporación de técnicas multiobjetivo cuando se trata más de un objetivo. En este sentido y especialmente para problemas de agrupamiento de carácter censal, la resolución se centra en la obtención de agrupaciones compactas y homogéneas para escenarios de carácter poblacional/censal.

La agrupación de datos censales con sus respectivas variantes, presenta serios problemas dada la selección de variables que intervendrán en la agrupación. Por otro lado, la agrupación también debe ser compacta considerando la ubicación espacial de estos datos. Esto da lugar a proponer un método que agrupe datos censales con características poblacionales para responder a homogeneidad y paralelamente se logren agrupaciones geométricamente compactas tomando la posición espacial en R^2.

Para solucionar este problema hemos aprovechado resultados previos sobre particionamiento geográfico [3], teoría básica multiobjetivo [7], la heurística VNS [8], y propiedades de Maxima [7].

En este punto, la aportación principal de este trabajo es integrar dentro del algoritmo de particionamiento multiobjetivo tanto a VNS para lograr soluciones aproximadas como al método multiobjetivo basado en la teoría del orden, que escogerá de estas soluciones las que sean no dominadas.

El presente trabajo se encuentra organizado como sigue: sección 1 como introducción. En la sección dos se aborda un marco teórico sobre particionamiento y teoría multiobjetivo. En la sección tres se presenta el método que hemos construido para obtener el Frente de Pareto. En la sección 4 se muestra un conjunto de pruebas y finalmente las conclusiones en la sección 5.

2 Marco teórico

Distintos métodos clásicos en análisis multivariado de datos encuentran óptimos locales de los criterios que optimizan. El caso, que en este trabajo se retoma, corresponde a la clasificación por particiones. Bajo la aplicación de distintas metaheurísticas en particionamiento, muchos han sido los esfuerzos para encontrar óptimos globales en problemas de optimización discreta.

2.1 Particionamiento

En los métodos de particionamiento, se busca una única partición de los objetos en estudio en *k* clases disjuntas. La teoría tradicional de los métodos de particionamiento son fundamentalmente *k*-medias, nubes dinámicas y algoritmos de transferencias, entre otros.

En la clasificación por particionamiento se tiene $\{x_1, \ldots, x_n\}$ el conjunto finito de *n* objetos a clasificar y $k<n$ el número de clases en los cuales se desea clasificar a los objetos. Una partición P = $\{C_1, \ldots, C_k\}$ de Ω en *k* clases, $C_1, \ldots, C_k$, está caracterizada por las siguientes condiciones:

1. $\Omega = \bigcup_{i=1}^{k} C_i$
2. $C_i \cap C_j = \varnothing$, para todo $i \neq j$

El número *k* es el tamaño de la partición. Es posible permitir eventualmente que algunas de las clases C_i sea vacía, de manera que en realidad las particiones $P = \{C_1,\ldots, C_k\}$ que se consideran son particiones Ω de en *k* o menos clases. Sin embargo, las particiones óptimas de acuerdo al criterio de inercia contienen exactamente *k* clases no vacías aceptables [9]. En general, se quieren obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas.

El particionamiento también es conocido como grupo y su planteamiento puede ser visto como uno de optimización de la siguiente manera: Dados un conjunto de *n* objetos $X = \{x_1, x_2,..., x_n\}$

donde $x_i \in R^D$, y k un número entero positivo conocido *apriori*, el problema del clustering (agrupamiento), consiste en encontrar una partición $P = \{C_1, C_2, ..., C_k\}$ de X, siendo C_j un cluster conformado por objetos similares, satisfaciendo una función objetivo

$$f: \boldsymbol{P} \rightarrow \mathrm{R},$$

donde $\boldsymbol{P}$ es la colección de todas las particiones de Ω.

Para medir la similaridad entre dos objetos $\boldsymbol{x}_a$ y $\boldsymbol{x}_b$ se usa una función de distancia denotada por $d(\boldsymbol{x}_a, \boldsymbol{x}_b)$, siendo la distancia euclidiana la más usada para medir la similaridad. Así la distancia entre dos diferentes elementos $\boldsymbol{x}_i = (x_{i_1}, \ldots, x_{i_D})$ $\boldsymbol{x}_j = (x_{j_1}, \ldots, x_{j_D})$ es

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{l=1}^{D} (x_{i_l} - x_{j_l})^2}$$

Los objetos de un cluster son similares cuando las distancias entre ellos es mínima; esto permite formular la función objetivo como

$$D(P) = \sum_{c \in P} \sum_{x_i \in C} d(\boldsymbol{x}_i, \boldsymbol{x}_C)^2, P \in \boldsymbol{P} \quad (1)$$

esto es, se desea minimizar (1); donde $\boldsymbol{x}_C$, conocido como elemento representativo del cluster C, es la media de los elementos del cluster C,

$$\boldsymbol{x}_C = \frac{1}{|C|} \sum_{x_i \in C} \boldsymbol{x}_i \quad (2)$$

y corresponde al centro del cluster. Bajo esas características, el agrupamiento es un problema de optimización combinatoria, y ha sido demostrado que es NP-difícil, entonces este problema puede ser abordado como un problema de optimización combinatoria y es similar decir que el agrupamiento no jerárquico (particionamiento), es un problema combinatorio. Esto significa que cuando se quiere obtener una partición en k clases de un conjunto con n individuos, no tiene sentido examinar todas las posibles particiones del conjunto de individuos en k clases. De hecho, se puede probar que el número $S(n,k)$ de particiones diferentes de un conjunto de n individuos en k clases, cumple la ecuación de recurrencia $S(n,k) = S(n-1,k-1) + kS(n-1,k)$. Esto lleva a que

$$S(n,k) = \frac{1}{k!} \sum_{1}^{k} (-1)^{k-i} \binom{k}{i}^{i^n} \quad (3)$$

En particular, el estudio exhaustivo del particionamiento clásico, ha sido útil en el desarrollo de un algoritmo de particionamiento que agrupa unidades geográficas para dar respuesta al problema de agrupamiento geográfico que minimiza la distancia entre los objetos [4]. Este problema que es de elevado costo computacional es bien conocido en diseño territorial [2].

Cuando solo optimiza el objetivo de compacidad geométrica es llamado Grupo Geográfico (CG), el cual, ha sido tratado con dos metaheurísticas: Búsqueda por Entorno Variable [3], y Recocido Simulado [4]. Para elegir la heurística que mejor aproxime el criterio de minimización de distancias en CG, se ha diseñado un experimento que nos permite calibrar los parámetros de ambas heurísticas bajo un mismo tiempo de ejecución para estas. Este resultado ha proporcionado información estadística para proponer a VNS como método de aproximación en el problema de particionamiento multiobjetivo que nos ocupa [4].

2.2 Multiobjetivo

La optimización multiobjetivo, puede ser definida como un problema de optimización que presenta dos o más funciones objetivo. El inconveniente principal en este tipo de problemas en relación a un modelo de objetivo único reside en la subjetividad de la solución encontrada. Un problema multiobjetivo no tiene una solución óptima única, más bien, genera un conjunto de soluciones que no pueden ser consideradas diferentes entre los objetivos que optimiza. De esta manera el conjunto de soluciones óptimas es denominado Frente de Pareto (FP). Esta frontera de soluciones contiene todos los puntos que no son superados en todos los objetivos por otra solución. Este concepto lleva el nombre de dominancia, por esta razón el FP consiste solo de

soluciones no dominadas. Una solución domina a otra si y sólo si, es al menos tan buena como la otra en todos sus objetivos y es mejor en al menos uno de ellos [7].

Muchos son los ejemplos que se pueden citar en esta área, sin embargo, estos problemas pueden ser más claros identificando las relaciones entre las características del problema, sus restricciones y los objetivos principales que se desean mejorar en conjunto. Para este tipo de problemas, es posible tener una expresión como una función matemática. Al referirse a la mejora en conjunto, se dice que se debe optimizar a todas las funciones de manera simultánea, definiéndose entonces un problema del tipo descrito a continuación:

Definición 1. Un problema multiobjetivo (MOP) puede definirse en el caso de minimización (y análogamente para el caso de maximización) como:

Minimizar $f(\boldsymbol{x})$

dado que $f : F \subseteq \mathrm{R}^n \longrightarrow \mathrm{R}^q$, $q \geq 2$ con región factible en

$$A = \{\boldsymbol{a} \in F: g_i(\boldsymbol{a}) \leq 0, i = 1,\ldots,m\} \neq \emptyset$$

El conjunto A es llamado región factible y se dice que el problema se encuentra sujeto a las restricciones g_i: $\mathrm{R}^n \longrightarrow \mathrm{R}$ que son funciones cualesquiera.

En el ámbito de la optimización multiobjetivo se tiene que decidir un cierto esquema de mejoría de una solución sobre otra, es decir cuáles soluciones se elegirán para ser más aptas; a esta relación de mejoría de un individuo sobre otro se le conoce como esquema de dominación y su definición se basa principalmente en que la solución de un problema multiobjetivo no es única y por lo tanto el tomador de decisiones debe elegir de entre una gama de posibles soluciones que no se pueden mejorar entre sí, es decir que no se dominan. Este concepto es más claro si pensamos que dentro del campo de los números reales se encuentra definido el orden de manera natural. Para R^n podemos extender el concepto mediante la siguiente definición.

Definición 2. Dados $\boldsymbol{x}, \boldsymbol{y}$ vectores en R^n $\boldsymbol{x} \preceq \boldsymbol{y}$ si y solo si $x_k \leq y_k$ para todo $k \in \{1, \ldots, n\}$ y $\boldsymbol{x} \prec \boldsymbol{y}$ si y solo si $\boldsymbol{x} \preceq \boldsymbol{y}$ con $\boldsymbol{x} \neq \boldsymbol{y}$, donde ≤ es el orden usual en R.

2.2.1 Frente de Pareto

Una opción común a usarse como relación de dominación es la conocida dominación de Pareto definida como sigue:

Definición 3. Dado el problema multiobjetivo, minimizar $f(x)$, donde $f : F \subseteq \mathrm{R}^n \longrightarrow \mathrm{R}^q$ $q \geq 2$ con $A \subseteq F$ la región factible. Decimos que un vector $\boldsymbol{x}^* \in A$ es no dominado o un óptimo Pareto si no existe un vector $\boldsymbol{x} \in A$ tal que $\boldsymbol{x} \prec \boldsymbol{x}^*$.

Así, la respuesta al problema de hallar las mejores soluciones (las soluciones no dominadas, como quiera que se defina la dominación dentro de la técnica) en un problema multiobjetivo es a lo que se le llama el conjunto solución del problema y el conjunto de valores de la función objetivo con dominio restringido a los vectores del conjunto solución (es decir, los vectores no dominados) es lo que conocemos como Frente de Pareto.

En este sentido pensar en el conjunto de vectores no dominados conduce lógicamente al concepto de conjunto parcialmente ordenado.

Definición 4. El conjunto $E(A; f)$ de soluciones de Pareto eficientes (también conocido como conjunto de óptimos de Pareto) se define de la manera siguiente:

$$E(A, f) := \{\boldsymbol{a} \in A: \nexists\, \boldsymbol{b} \in A \text{ que cumpla } f(\boldsymbol{b}) \prec f(\boldsymbol{a})\}$$

Es decir, el conjunto de todos los vectores no dominados bajo el esquema de Pareto.

En resumen, el conjunto de óptimos de Pareto es el espacio solución del problema, y el Frente de Pareto es su imagen con respecto a la función a optimizar

$$f: F \subseteq \mathrm{R}^n \longrightarrow \mathrm{R}^q, q \geq 2$$

Un concepto íntimamente relacionado con el Frente de Pareto es el de óptimo de Pareto. Tanto el óptimo de Pareto como el Frente de Pareto son el marco sobre el que se trabaja dentro de la toma de decisiones multicriterio.

El conjunto de óptimos Pareto para un problema multiobjetivo dado, es un conjunto parcialmente ordenado (poset) visto formalmente. En los problemas multiobjetivo se buscan los

elementos minimales del espacio de solución R^n visto como un poset con la relación $\preceq$ dada en la definición 2.

En este punto los elementos no dominados poseen propiedades como el conjunto Máxima (conjunto de minimales o maximales). Las propiedades de este conjunto han sido de singular importancia en la identificación de soluciones no dominadas en el método que exponemos en este trabajo.

3 Metodología

El problema que se enfrenta en este artículo debe resolver la agrupación de dos objetivos clásicos en Diseño Territorial: compacidad y homogeneidad. Para la agrupación, se han elegido las propiedades de los algoritmos de particionamiento y sabiendo que este es de carácter combinatorio, se ha escogido a VNS como método de aproximación. Por otra parte, para hallar de entre todas las soluciones que se generan aquellas que sean no dominadas, se introduce un método que se apoya de la teoría del orden donde además se hace uso de las características del conjunto Máxima [0].

3.1 Planteamiento del problema

En término generales, el problema que aquí se trata es sobre Multiobjetivo en Particionamiento para Datos Espaciales (MUPDE), y el conflicto principal es optimizar simultáneamente las funciones de compacidad para la ubicación geográfica y homogeneidad para variables censales. Los datos a agrupar son áreas geoestadísticas básicas (AGEBs) y MUPDE busca respuestas a diferentes problemas poblacionales de concentración o distribución de un valor censal para una determinada zona metropolitana. Esta distribución tiene asociado un valor que debe estar balanceado para cada uno de los grupos de toda la extensión territorial (homogeneidad) y a la vez respetar la propiedad de cercanía física entre las AGEBs que forman un grupo (compacidad).

La solución a MUPDE consiste en resolver la agrupación de los dos objetivos mencionados que son exigidos en problemas de diseño territorial [12]. Cada partición es representada por un par de soluciones compuesto por compacidad y homogeneidad (C,H). Estas soluciones son revisadas bajo una variante de la dominancia Pareto con el fin de obtener un subconjunto soluciones no dominadas y no comparables (C,H). Este subconjunto es el Frente de Pareto.

El particionamiento implícito debe establecer a través de una función bi-objetivo, que cada grupo sea compacto y que la suma de alguna variable poblacional se obtenga tan homogénea como sea posible. La aproximación de la función de costo se realiza con VNS y se combina con un método que se apoya de la teoría del orden para encontrar un conjunto de soluciones no-dominadas y no comparables a través de los puntos minimales que forman al conjunto Máxima [0].

Finalmente, el proceso de desarrollo para MUPDE se describe informalmente de manera simple:

Inicio

1. Obtener una solución inicial de particionamiento compacto y homogéneo con VNS, la cual se denota como (C_i, H_i) (solución actual).

repetir

2. Generar siguiente solución (C_{i+1}, H_{i+1})
3. ¿Es no comparable y no dominada (C_{i+1}, H_{i+1}) contra (C_i, H_i)? si, entonces hacer
   3.1 Etiquetar a (C_{i+1}, H_{i+1}) como minimal. (C_{i+1}, H_{i+1}) es ahora solución actual
4. Almacenar minimal

hasta

estructuras de vecindad de VNS = 0

Fin

Se tiene entonces una historia de todas las soluciones que son candidatas a minimales. Este proceso se repite hasta que los parámetros de la VNS lo permitan recogiendo así el conjunto de soluciones minimales que forman el Frente de Pareto.

3.2 Modelación

Como estamos interesados en hallar particiones de Ω (AGEBs) que minimicen la

compacidad y la homogeneidad se requieren hacer algunas adaptaciones mínimas a las definiciones 1, 3 y 4. Para esto consideramos la colección de todas las particiones de Ω:

$$\boldsymbol{P} = \{P: P \text{ es una partición de } \Omega\}$$

Sea $f : \boldsymbol{P} \longrightarrow \mathrm{R}^2$ la función tal que $f(P) = (C(P), H(P))$ donde C y H son las funciones compacidad y homogeneidad respectivamente, ambas con dominio en $\boldsymbol{P}$ y valores en R.

La función de compacidad C es dada por:

$$C(P) = \sum_{C \in \boldsymbol{P}} \sum_{i,j \in C} d(i,j) \quad (4)$$

Analíticamente la función de homogeneidad H será descrita mas adelante en el segundo objetivo de la sección 3.2.2.
En nuestro caso la definición (1) se reduce al siguiente problema multiobjetivo:

Minimizar $f(P)$

dado que $f : \boldsymbol{P} \subset 2^{\Omega} \longrightarrow \mathrm{R}^2$, con región factible en

$$\boldsymbol{P} = \{P \in 2^{\Omega} : P \text{ es partición de } \Omega\}$$

donde 2^{Ω} es el conjunto potencia de Ω y

$$f(P) = (C(P), H(P))$$

Dado el problema multiobjetivo anterior podemos inducir un orden parcial $\preceq_{\boldsymbol{P}}$ sobre el conjunto de particiones $\boldsymbol{P}$ de la siguiente manera*:*

$$P \preceq_{\boldsymbol{P}} P' \text{ si y solo si } f(P) \preceq f(P'),$$

donde $\preceq$ es el orden dado en la definición 2. De manera análoga a la definición 3, decimos que una partición $P^* \in \boldsymbol{P}$ es no dominada o un óptimo Pareto si no existe una partición $P \in \boldsymbol{P}$ tal que $P \prec_{\boldsymbol{P}} P^*$, donde $\prec_{\boldsymbol{P}}$ es el orden estricto inducido por el orden parcial $\preceq_{\boldsymbol{P}}$.

Entonces el conjunto de óptimos Pareto $\mathrm{E}(\boldsymbol{P}, f)$ en nuestro caso queda definido como:

$$\mathrm{E}(\boldsymbol{P}, f) = \{P \in \boldsymbol{P} : \nexists\, P' \in \boldsymbol{P} \text{ que cumpla } P' \preceq_{\boldsymbol{P}} P\}.$$

Observe que el conjunto de las particiones $\boldsymbol{P}$ se genera a partir del conjunto finito Ω entonces la imagen (Frente Pareto) de la función objetivo f es finita en consecuencia el Frente de Pareto es un conjunto discreto.

La bondad del mecanismo de nuestro trabajo para buscar soluciones de mejor compromiso radica en la manera en que se resuelve la agrupación: devuelve un conjunto diverso de particiones con el uso VNS. Por otra parte, para encontrar el subconjunto de soluciones eficientes y no dominadas se evalúan las soluciones que se van generando revisando que sean no dominadas y no comparables. En la siguiente sección se muestra el algoritmo en pseudocódigo.

3.2.1 Representación

La meta es obtener un subconjunto finito de particiones óptimas de datos espaciales AGEBs cuya composición está dada por 2 componentes: coordenadas geográficas en el plano R^2 y un vector de 144 características descriptivas de carácter censal.

La primera componente permite que se obtenga una matriz de distancias para el proceso de cálculo de la compacidad geométrica, una de las funciones de objetivo a minimizar.

El vector de descripción se utiliza para optimizar la segunda función objetivo y consiste en minimizar alguna de las variables censales con el fin de calcular homogeneidad o equilibrio para dicha variable. Para la selección de las variables censales disponibles se tiene un procedimiento de consultas.

3.2.2 Mecanismo de agrupamiento

La estrategia de particionamiento consiste en elegir aleatoriamente AGEBs como centroides que identifican el número de grupos. Aquellos AGEBs no centroides que tengan la distancia más corta hacia un determinado centroide-AGEB, son los integrantes de un grupo. Esta idea informal es la que se entiende como compacidad geométrica.

Una vez formados los grupos bajo la minimización de distancias, se calcula la homogeneidad a la agrupación creada dado que el dominio en este problema es una partición, es decir, en problemas multiobjetivo la función a optimizar tiene el mismo dominio para todos los objetivos [10]. De este modo sobre una misma partición se optimiza la compacidad y homogeneidad.

Para aclarar esta situación, basta observar que en los problemas de optimización multiobjetivo se elige la mejor alternativa *x* de un conjunto X respecto a un objetivo amplio y poco preciso. Entonces se introduce una jerarquía de objetivos siendo los *m* de más bajo nivel lo suficientemente precisos para permitir medir los resultados de cada alternativa. Se tiene así, que elegir la mejor alternativa respecto a los *m* objetivos.

Para introducirnos a la modelación de MUPDE, primero se presenta la notación básica. Los elementos que forman un grupo territorial (GT) son áreas geoestadísticas básicas AGEBs.

Componentes descriptivas para cada AGEB:

a) ubicación geográfica y b) vector de parámetros de características asociadas a la AGEB (variables censales cuantificadas).

Se asume que *k* es el número de grupos territoriales y *n* el número total de AGEBs $k \leq n$.

Restricciones:

1. Cada AGEB debe pertenecer a un único grupo.
2. En un grupo el valor de cada parámetro en el valor de la variable censal.
3. Los grupos son disjuntos.
4. No existen grupos vacíos.
5. Las variables poblaciones pueden o no estar acotadas.
6. Pueden estar en la agrupación todas las variables o un subconjunto de ellas.
7. Los AGEBs asignados a cada grupo deben formar un grupo compacto.
8. Los grupos deben estar balanceados con respecto a una meta de equilibrio para alguna característica medible.

Objetivos. Las restricciones 7 y 8 se traducen en dos objetivos, el primero que los grupos sean lo más compactos posible y el segundo que los grupos sean homogéneos para una variable censal. A continuación se describen cada uno de ellos.

Primer objetivo. Minimización de distancias.

Este objetivo se ha resuelto de un algoritmo de particionamiento basado en el método de nubes de puntos [9].

Dada una partición $P \in \boldsymbol{P}$ para cada $C \in P$ elegimos de manera aleatoria un $c \in C$ y definimos la suma

$$S(P) = \sum_{C \in P} \sum_{i \in C} d(i, c)$$

Entonces el número

$$\min \{S(P): P \in \boldsymbol{P}\} \quad (5)$$

minimiza la distancia intra clases entre AGEBs. Se tienen como restricciones:

- $C \neq \emptyset$ (los grupos no son vacíos).
- $C \cap C' = \emptyset$ para $C \neq C'$ (No existen AGEBs repetidos en distintos grupos).
- $\bigcup_{C \in P} C = P$ (la unión de todos los grupos son todos los AGEBs).

La elección aleatoria de los *k* centroides $c_1,\ldots, c_k$ genera una partición $P=\{C_1,C_2\ldots,C_k\}$ donde donde cada c_i es un representante de la clase C_i

Esta partición se construye de la siguiente manera:

1. Se elige un elemento $i \in \Omega$.
2. Se calcula el min $\{d(i,c_t): t=1,\ldots,k\}$.
3. i se ubica en la clase $C_{t'}$ donde $c_{t'}$ es el centroide donde se alcanza el min $\{d(i,c_t): t=1,\ldots,k\}$.

Entonces se generan tantas particiones como elecciones aleatorias se hagan de los centroides. El número de elecciones aleatorias es el número de iteraciones y es denotado por η. Como el número de particiones de Ω puede ser muy grande (ver fórmula 3), las fórmulas (4) y (5) se restringen al subconjuto $\boldsymbol{P'}$ de $\boldsymbol{P}$ de todas las particiones generadas a partir de las diferentes elecciones de los grupos de centroides.

Observe que la cardinalidad de $\boldsymbol{P'}$ es el número de iteraciones η. Dependiendo del tipo de problema es necesario fijar el número de grupos en que se desea particionar una zona geográfica, esto es, cada elemento del conjunto $\boldsymbol{P'}$ tiene el mismo tamaño *k*.

Por lo tanto $\boldsymbol{P'}$ es conjunto de todas las particiones de tamaño *k* formado por η elecciones de grupos de *k* centroides.

Segundo objetivo. **Minimización de homogeneidad para una variable censal**.

El segundo objetivo consiste en encontrar un equilibrio para una variable de interés donde participan algunos conjuntos de variables adecuados con las siguientes combinaciones:

- todas las variables sin restricciones o
- todas las variables acotadas o
- algunas variables sin restricciones y las restantes no participan o
- algunas con restricciones y las restantes no participan.

Para ilustrar esta situación decimos que cuando se desea agrupar AGEBs de una zona metropolitana donde estas AGEBs se seleccionen previamente bajo condiciones en los valores buscando un equilibrio para alguna variable censal, hablamos informalmente del problema de particionamiento para AGEBs bajo criterios de homogeneidad.
Una forma de plantear como elegir variables de población con valores determinados para que puedan ser usadas en un proceso de agrupación posterior, es a través de una matriz de participación. Para formalizar esto a continuación damos la siguiente definición.

Definición 5. Sean $\Omega'=\{AG_1, AG_2, \ldots, AG_n\}$ un conjunto de n AGEBs y $VC=\{X_1, X_2, \ldots, X_r\}$ un conjunto de variables censales que describen a las AGEBs, donde cada variable X_i es una función del conjunto de AGEBs Ω' con valores en los reales positivos R^+. Dados r intervalos $I_j=[\alpha_j,\beta_j]$, $j=1,\ldots r$ y la funciones características $\chi_{[\alpha j,\beta j]}$: $VC\rightarrow$ {0,1},

$$\chi_{[\alpha j,\beta j]}(X)=\begin{cases}1 \; si \; X \in [\alpha_j, \beta_j] \\ \\ 0 \; en \; otro \; caso\end{cases}$$

Entonces definimos la matriz de participación asociada al grupo de AGEBs Ω' con variables VC y condiciones I_j,
$j=1,\ldots r$ como la matriz $M = (v_{ij})$ de $n\times r$ donde

$$v_{ij}= \chi_{[\alpha j,\beta j]}(X_j)\, X_j(AG_i) \qquad (6)$$

Entonces la matriz M contiene todos los valores de las variables participantes en los AGEBs respectivos.
Si $v_{ij}=0$ decimos que la variable X_j no participa en el AGEB AG_i.

Teniendo las variables que participan en la agrupación, para homogeneizar los grupos se calcula lo siguiente:

- se obtiene un promedio ideal para la variable de interés, digamos que la variable de interés es X_j y que su promedio ideal es V_j, esto sucede cuando todos los grupos tengan el mismo valor. Sin embargo esto no es común en la práctica, entonces
- se calcula el promedio real a cada grupo $\frac{1}{n}\sum_{i=1}^{n} v_{ij}$
- y se resta este valor al promedio ideal,

$$V_j - \frac{1}{n}\sum_{i=1}^{n} v_{ij} = \frac{1}{n}\sum_{i=1}^{n}(V_j - v_{ij}) \qquad (7)$$

Al minimizar esta diferencia en valor absoluto se puede obtener el costo de la función objetivo para homogeneidad.

3.2.3 Formulación multiobjetivo

Sea una AGEB un dato espacial definido por sus componentes en el espacio y una descripción de 144 variables censales dada por un vector (INEGI, 2000).

La nomenclatura para la modelación final es:

M= mapa territorial
T= territorio
DT= datos censales
MS= matriz de disimilitud
AG_j= j-ésima unidad geográfica básica (AGEB)
C_i= i-esimo grupo territorial
k = número de grupos territoriales
n= número de AGEBs tal que $k<<n$
i= índice de grupo territorial
j= índice de unidad geográfica básica
c_i= centroide del i-ésimo grupo territorial C_i

El modelo en cuestión es entero mixto y hace uso de las variables binarias para modelos de este tipo. Considerando lo anterior, el modelo para MUPDE es:

Minimizar y= $f(\boldsymbol{x})$= ($f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$)

f_1: es el costo de minimizar la distancia entre AGEBs de acuerdo a la ecuación a) que debe formularse como función, y

f_2: es el costo de minimizar la homogeneizar de una variable censal de las AGEBs. Esta función se debe expresar a partir de la ecuación (7).

Ahora, planteando las restricciones para las funciones f_1 y f_2 se tiene:

- $C_i \neq \emptyset$ para i = 1,…, *k*.
- $C_i \cap C_j = \emptyset$ para $i \neq j$ (No existen AGEBs repetidos en distintos grupos).
- $\cup_{i=1}^{k} C_i = \Omega'$ (la unión de todos los grupos son todos los AGEBs).
- $\alpha_k < \beta_k$ (cotas para la variable X_k)
- $\sum_{i=1}^{m} X_{ij} = 1$ es la asignación de AGEBs donde X_{ij} = 1 si $AG_j \in C_i$ o X_{ij} = 0 si $AG_j \notin C_i$ son las variables de decisión

$\boldsymbol{y} = (y_1, y_2) \in Y \subset R^2$ es el vector objetivo.

```
Algoritmo MUPDE

Sea:

n Número de objetos a clasificar.
K Número de grupos
Val_i Valor que tiene el AGEB i para la variable que se va a mantener homogeneidad
Ug Unidad geográfica
MaxVNS Número de veces que se va a recorrer la estructura de vecindades
MaxBL Número máximo de iteraciones para búsqueda local

kVecindad←Generar un número aleatorio entre 1 y n
SolActual←Genera una solución aleatoria que se encuentre en la vecindad kVecindad
costeActual ←getCosteComp(SolActual), getCosteHom(SolActual)
// Este par de soluciones se evalúan iterativamente con la relación de orden de Dominancia Pareto y con la relación de no comparabilidad
cont←1
Mientras cont < MaxVNS hacer
       kVecindad←1
       Mientras kVecindad ≤ n
              SolCand←Genera una solución aleatoria que se encuentre en la vecindad kvecindad
              SolCand←BusquedaLocal (SolCand)

              costeCand← getCosteComp(SolCand,getCosteHom(SolCand)
              Si costeCand < costeActual entonces
                     SolActual←SolCand
                     costeActual← costeCand
              Si no
                     kVecindad← kvecindad + 1
              fin si
      fin Mientras
       Cont←cont+1
fin Mientras
Regresa SolActual
```

Función getCosteComp (Sol)

```
//Regresa un entero indicando que tan buena es la solución Sol en cuanto a compacidad (entre más pequeño sea, la solución es mejor)
i←1
cost←0
Mientras i ≤ n
       Si Ugi no es centroide entonces
              dmin←dist(Sol_1,Ugi )
       //Distancia entre el objeto Sol_1 y el objeto i
              j←2
              Mientras j ≤ k
                     Si dist (Solj,Ugi) < dmin
                            dmin←dist (Solj,Ugi)
                     fin si
                     j←j + 1
              fin Mientras
              cost←cost + dmin
       fin si
       i←i + 1
fin Mientras
getCosteComp(Sol)←cost
```

Función getCosteHom (Sol)

```
//Regresa un entero indicando que tan buena es la solución Sol en cuanto a homogeneidad (entre más pequeño sea, la solución es mejor)
total←0
coste←0
Para i←1 hasta n hacer
       ng←Obtener el número de grupos al cual pertenece el AGEB i
       total←total + Val_i
       totalGrupo_ng←Val_i
fin Para
promedioIdeal←total/k
Para j←1 hasta k hacer
       Coste←coste + |totalGrupoj - promedioIdeal|
fin Para
getCosteHom(Sol)←coste
```

Función BusquedaLocal (Sol)

```
NumItera←0
SolMejorada←Sol
costeSolMejorada← getCosteComp(SolMejorada), getCosteHom(SolMejorada)
Mientras NumItera ≤ MaxBL
```

```
        SolCand←Generar solución aleatoria
vecina de SolMejorada
        costeSolCand← getCosteComp(SolCand),
getCosteHom(SolCand)
        Si costeSolCand < costeSolMejorada
entonces
                SolMejorada←SolCand
                NumItera←MaxBL + 1
        Si no
                NumItera←NumItera + 1
        fin Si
fin Mientras
BusquedaLocal(Sol)←SolMejorada.
```

En el algoritmo que se ha mostrado se calculan las soluciones generadas de manera iterativa con el orden Pareto y a la vez con el orden no comparable.

La lectura del trabajo [0], ha proporcionado información importante sobre el conjunto Maxima, el cual es un conjunto de minimales [0]. En este sentido se han recogido los aspectos básicos de la teoría del orden asociados a los minimales, los cuales tienen la propiedad de ser no dominados y no comparables entre ellos.

Se ha propuesto una variante a la relación de orden Pareto para hallar soluciones no comparables y no dominadas proponiéndose de este modo una estrategia sencilla: calcular que las soluciones generadas por VNS sobre el particionamiento compacto y homogéneo, cumplan el orden Pareto y a la vez que sean no comparables. Es fácil ver que la negación lógica del orden Pareto es una relación de orden no comparable.

Traduciendo el orden Pareto se tiene que:
Dada una solución (a,b) la siguiente solución (a', b') es aceptada si: (*)

$a' > a \wedge b' = b \vee b' > b \wedge a' = a \vee a' > a \wedge b' > b \vee a = a' \wedge b = b'$

La implicación trivial de negar lógicamente las expresiones da lugar a obtener lo siguiente:

Consideremos nuevamente la desigualdad

$$(a, b) < (a', b')$$

Entonces la negación de esta relación produce las siguientes equivalencias:

$$\neg((a,b) < (a',b')) \equiv \neg[(a < a' \vee a = a') \wedge (b < b' \vee b = b')] \equiv \equiv \neg(a < a' \vee a = a') \vee \neg(b < b' \vee b = b') \equiv \equiv (a \geq a' \wedge a \neq a') \vee (b \geq b' \wedge b \neq b') \equiv \equiv (a > a' \vee b > b')$$

De la misma manera tenemos

$$\neg\,((a', b') < (a,b)) \equiv (a' > a \vee b' > b)$$

Por tanto, se concluye que (a,b) y (a', b') son no comparables sii $(a > a' \vee b > b') \wedge (a' > a \vee b' > b)$.

Bajo este orden llamado no comparable ajustado adecuadamente con el orden Pareto se obtienen todos los pares de soluciones minimales satisfaciendo de esta forma el conflicto de las soluciones en compromiso.

Este resultado se confirma cuando NDOMINATED se aplica a todas las soluciones generadas por el algoritmo propuesto [NDOMINATED].

Finalmente este resultado implica que contamos con el Frente de Pareto para el problema MUPDE.

4 Pruebas

Cuando se procesan datos espaciales y censales, se requiere de un análisis exploratorio para su adecuación; de este modo, es posible otorgar calidad a los datos finales además de ser accesibles a diferentes tratamientos posteriores.

En particular, en el proceso de particionamiento para el problema que nos ocupa, los datos censales originales se han extraído de una base de datos que ofrece el INEGI [6]. Estos datos están definidos de manera descriptiva por un vector de variables y se conocen como AGEBs.

Bajo ciertos procedimientos y el apoyo de un Sistema de Información Geográfica, estos datos se han explorado y transformado para que se proceda al cálculo de distancias entre AGEBs. El resultado que se busca es una matriz básica de distancias (matriz de disimilitud) y una estructura de variables (matriz de homogeneidad). Estas estructuras de distancias y variables se emplean como entrada al algoritmo de MUPDE.

El proceso de particionamiento se compone de la siguiente secuencia:

1. Un proceso de selección de variables adecuadas de acuerdo a un problema particular con el fin de obtener un subconjunto de variables que pueden estar restringidas en sus valores censales.
2. Del grupo de variables conseguido se obtiene una matriz de disimilitud ajustada para las variables seleccionadas.
3. Se hace uso del procedimiento de particionamiento con VNS multiobjetivo. Aquí se escogen los parámetros de VNS incluyendo el número de grupos. El resultado final arroja 3 archivos:
   1) conjunto de iteraciones con el correspondiente valor de la función de costo y el conjunto de minimales.
   2) detalle de los grupos con los AGEBs que le pertenecen y el tiempo de ejecución y
   3) una estructura compatible con un SIG para mostrar los resultados en un mapa geográfico. Este resultado no se presenta en este artículo, sin embargo se estima describir su proceso en trabajo futuro.

Por último se verifica que el conjunto de minimales arrojado por MUPDE corresponda al conjunto de soluciones no dominadas correcto. Para este propósito, se ha hacho uso de NDOMINATED que es una aplicación obtenida del algoritmo propuesto en [7]. Este programa se encuentra disponible de manera gratuita en [NDOMINATED].

Ejemplo:

Asúmase que existe un problema de la población donde se debe conocer la distribución del sector femenino en la Zona Metropolitana del Valle de Toluca (ZMVT). Se requiere además que dicho sector este organizado en 8 agrupaciones.

En congruencia con lo descrito en las secciones anteriores, la respuesta a un problema con este comienza con la selección de las variables de interés. Estas variables tienen la nomenclatura Z + numero natural y son tomadas de [6].

Las variables de interés para el ejemplo hipotético se refieren a la población femenina en los rangos que se observan:
0-4 años (Z006), 0-14 (Z009), 5 años y más (Z012), 6 años y más (Z015), 6-14 años (Z018), 12 años y más (Z021), 15 años y más (Z024), 15-49 años (Z028), 15 a 19 años Z031, 18 años y más Z034, 20 años y más (Z037), 20 a 24 años (Z040), 50 años y más (Z041), 60 años y más (Z044).

```
No. de clusters:           8

Promedio ideal de grupo: 1001,125

Parametros de VNS:

Maximo Iteraciones VNS                :  2
Maximo de iteraciones de Busqueda Local :  15
Solucion inicial               : 90, 415, 133, 46, 430, 88, 181, 26,
Solucion optima encontrada     : 13, 80, 433, 262, 59, 63, 8, 82,
Vecindario de la solucion optima : 13
Numero de iteraciones realizadas :  1881
Numero de soluciones aceptadas   :  1
_____________________________________________________

Tiempos:
Hora inicial   : 20:21:26
Hora final   : 20:24:40
_____________________________________________________

Cluster no 1: 13,1,2,3,4,5,128,129,136,137,138,139,140,141,142,145,146,
Cluster no 2: 80,77,79,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,1
Cluster no 3: 434,130,131,132,133,143,144,161,193,195,196,197,198,199,2
Cluster no 4: 262,110,111,112,113,114,115,255,256,257,258,259,260,261,2
Cluster no 5: 59,45,46,47,48,49,50,51,52,53,54,55,56,57,58,60,116,117,1
Cluster no 6: 63,9,10,11,12,14,15,16,17,18,19,20,21,22,23,24,25,26,27,2
Cluster no 7: 8,6,7,69,70,71,72,73,74,75,151,152,153,158,159,160,162,16
Cluster no 8: 82,65,66,67,68,76,78,81,83,84,85,86,165,176,177,178,179,1

_____________________________________________________
Maximales: (Comp/Hom)


2148008 2997,25
2003081 4493,25
2102931 4337,25
2378834 2368,75
2080697 3737
2063480 3317
2009034 3779,25
2477131 1744,75
2334102 2813
2287480 2967,25
2350957 2313
2303444 2723
```

Fig. 1. Archivo de resultados

Contando con este subconjunto de variables y la matriz de distancias de la ZMVT, se ejecuta el particionamiento con VNS en un valor de 15 para el parámetro de búsqueda local y 2 en estructura de vecindad.

En la figura 1 se muestra la imagen de uno de los archivos resultantes que contiene la información sobre los parámetros de la heurística, los grupos con los AGEBs que le corresponden, tiempo de ejecución y el conjunto de minimales que se etiquetó.

En la figura 2 se han organizado los resultados que proporciona el algoritmo MUPDE. La columna del centro es un extracto de todas las soluciones que produce VNS. Los datos de la izquierda son los minimales que arroja MUPDE (ver figura 1) y la columna de la derecha es el conjunto Máxima que filtra NDOMINATED.

MINIMMALES MUPDE			SOLUCIONES TOTALES			SOLUCIONES NDOMINATED	
compacidad	homogeneidad		compacidad	homogeneidad		compacidad	homogeneid
2148008	2997.25		2378634	5387		2009034	3779.25
2003081	4493.25		2549240	3815.5		2290732	2955
2102931	4337.25		2412015	2549.25		2063480	3317
2378834	2368.75		2744633	4359.5		2287480	2967.25
2080697	3737		3335065	3707.25		2148008	2997.25
2063480	3317		2729560	4917.5		2350957	2313
2009034	3779.25		2935006	3065		2477131	1744.75
2477131	1744.75		2523110	3099		2003081	4493.25
2334102	2813		2807248	4070.75		2303444	2723
2287480	2967.25		3357000	7669.75			
2350957	2313		3639656	6251.5			
2303444	2723		2500273	3431.5			
			2622105	3191.25			
			2290732	2955			
			2884050	2985			
			2246206	3435			
			3230201	4245			
			3130872	5045			
			2931962	6559.25			
			2952301	5795.25			
			2330250	5535.25			
			2823609	5511.5			
			3228096	5477.5			
			2321730	4361			
			3142427	5103.5			
			2176693	3783.25			
			2136972	4607.25	...		
			2692013	5101.25			
			2589329	4275.25			
			2267975	5475.25			
			2608398	4795.25			
			2781303	5181			
			2702075	2799.25			
			2542577	3475.25			
			3231843	5439.25			
			2933123	3861.25			
			3444806	6423.25			
			2042952	5107.5			

Fig. 2. Evaluación de las soluciones

En la figura 3 puede verse el Frente de Pareto para el ejemplo sobre 8 grupos de la población femenina de la ZMVT. Los puntos azules corresponden a los minimales que MUPDE identifica y los rojos son el conjunto Maxima que NDOMINATED proporciona. Son tres soluciones las que no coinciden para ambas aplicaciones,

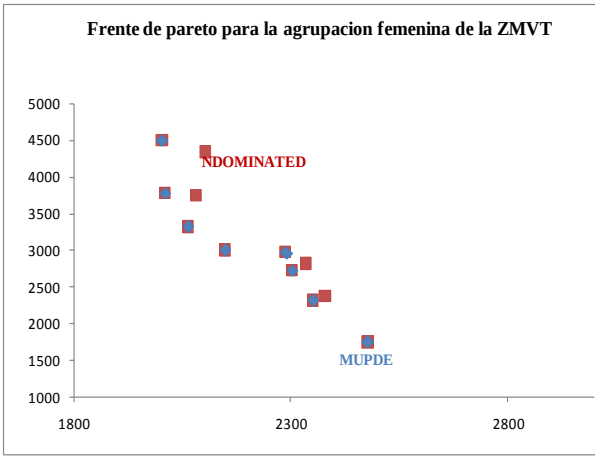

Fig. 3. Frente de Pareto para 8 grupos

sin embargo puede notarse con claridad que nuestro método obtiene todos los elementos que NDOMINATED identifica.

5 Conclusiones

El método desarrollado en este trabajo ha proporcionado buenos resultados en tiempo y calidad obteniéndose adecuadas aproximaciones prácticas al Frente de Pareto. Sin embargo, son pocas las pruebas realizadas para el ejemplo descrito en la sección 4. Para concluir con un resultado más completo, es conveniente que se desarrolle una serie de pruebas con un diseño de experimentos estadístico factorial, lo cual debe atenderse como trabajo posterior.

Por otra parte, una aportación extra en este trabajo reside en que el método desarrollado para la agrupación multiobjetivo, puede ser empleado para otro tipo de datos espaciales con algunas variantes en la implementación. Finalmente logramos obtener soluciones no dominadas mientras la heurística va trabajando, en contraste con otras aplicaciones que filtran soluciones no

dominadas pero no generan las soluciones del problema que se optimiza.

Referencias

1. **Algorithms to identified nondominated solutions in a multi-dimensional set (s.f.).** Retrieved from http://www.cs.cinvestav.mx/~emoobook/nodom/nondominated.html
2. **Altman, M. (1997).** The Computational Complexity of Automated Redistricting: Is Automation the Answer?. *Rutgers Computer and Technology Law Journal,* 23(1), 81-141.
3. **Bernábe, B., Osorio, M.A., Espinosa, J., & Aceves, R. (2009).** An Adjusted Variable Neighborhood Search Algorithm applied to the Geographical Clustering Problem. *Research in Computing Science*, 42, 113–126.
4. **Bernábe, M.B., Espinosa, J.E., & Ramírez, J. (2009).** Evaluación de un Algoritmo de Recocido Simulado con Superficies de Respuestas. *Revista de Matemática:Teoría y Aplicaciones*, 16(1), 159–177.
5. **Bernábe, M.B., Rosales, J., Osorio, M.A., Ramírez, J., & García, R. (2009).** A comparative study of Simulated Annealing and Variable Neighborhood Search for the Geographic Clustering Problem. *The5th International Conference on Data Mining (DMIN'09),* Las Vegas Nevada, USA, 595–599.
6. **Censo General de Población y Vivienda 2000, INEGI (s.f.).** Retrieved from http://www.inegi.org.mx/sistemas/microdatos2/default.aspx?c=14061&s=est
7. **Kung, H.T., Luccio, F., & Preparata, F.P. (1975).** On Finding the Maxima of a Set of Vectors. *Journal of the ACM (JACM),* 22(4), 469–476.
8. **Lara, A. (2003).** *Un estudio de las Estrategias Evolutivas para problemas Multiobjetivo*. Tesis de Maestria en Ciencias, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México, D.F.
9. **Mladenovic, N. & Hansen, P. (1997).** Variable Neighborhood Search, *Computers & Operations Research,* 24(11), 1097–1100.
10. **Pizza, E., Murillo, A., & Trejos, J. (1999**). Nuevas técnicas de particionamiento en clasificación automática, *Revista de Matemática: Teoría y Aplicaciones,* 6(1), 51–66.
11. **Ríos, D. (1987).** Sobre soluciones optimas en problemas de optimización multiobjetivo. *Trabajos de Investigación Operativa,* 2(1), 49–67.
12. **Zoltners, A. & Sinha, P. (1983).** Sales territory alignment: A review and model. *Management Science*, 29 (11), 1237–1256.

María Beatríz Bernábe Loranca *was born in the city of Puebla, Mexico. He received the B.S. degree in Computer Science from Benemérita Universidad Autónoma de Puebla (BUAP), Mexico in 1993 and the M.I. degree in quality engineering from Universidad Iberoamericana (UIA), Mexico in 2003. In January 2010, she received the Doctorate degree in Operations Research from the Universidad Nacional Autónoma de México (UNAM.). Since 1995, she has been a professor at the School of Computer Science of BUAP, where she works in databases and statistics. She belongs to the National System of Researchers with Level Candidate (SNI). Her research interests are: combinatorial optimization, territorial design and multiobjective techniques.*

Carlos Guillén Galván *obtained his Doctorate degree in Computer Science at Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE) and the Master Degree in Mathematics at Facultad de Ciencias Físico Matemáticas BUAP. Is Full Professor in the Facultad de Ciencias de la Computación at Benemérita Universidad Autónoma de Puebla in México since 1995. His research interests are in heuristics methods, combinatorial optimization and image processing.*

Artículo recibido el 31/12/2010; aceptado el 14/11/2011.