Revista de Matemática: Teoría y Aplicaciones 2009 **16**(1) : 159–177
cimpa – ucr issn: 1409-2433

EVALUACIÓN DE UN ALGORITMO DE RECOCIDO SIMULADO CON SUPERFICIES DE RESPUESTAS

Maria Beatriz Bernábe Loranca* José E. Espinosa Rosales†
Javier Ramírez‡

Recibido/Received: 20 Feb 2008 — Aceptado/Accepted: 8 Dic 2008

Resumen

En la solución al problema de conglomerado geográfico está implícito un proceso de clasificación combinatorio sobre unidades geográficas. La agregación propuesta en este trabajo considerara como función objetivo la minimización de distancias entre los objetos a agrupar con el fin de lograr la compacidad geográfica (tan deseable en problemas de diseño geográfico). Este problema es NP duro [1], por lo que es necesario el uso de métodos heurísticos para obtener una solución satisfactoria tanto en la bondad de las soluciones como en tiempo de cómputo en problemas grandes. La discusión se centra en evaluar la calidad de las soluciones obtenidas bajo procedimientos sistemáticos. Este trabajo presenta la modelación del problema de conglomerado geográfico, el uso de un algoritmo de Recocido Simulado en el algoritmo de particionamiento con el fin de obtener soluciones aproximadas y finalmente, para evaluar la calidad de las soluciones generadas, la aplicación de un Diseño de Experimentos Box-Behnken y Superficies de Respuestas para encontrar un balance y adecuación de los valores de los parámetros de Recocido Simulado en el control de la obtención de buenas soluciones.

Palabras clave: conglomerado geográfico, evaluación de parámetros, superficies de respuestas.

Abstract

*Departamento de Sistemas, DEPFI, Universidad Nacional Autónoma de México, México D.F.; y Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Puebla, México. E-Mail: beatriz.bernabe@gmail.com

†Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla. E-Mail pepe-espinosa@hotmail.com.

‡Universidad Autnoma Metropolitana – Unidad Azcapotzalco, Departamento de Sistemas, Avenida San Pablo 180, 02200 México D.F., México. E-Mail: jararo@correo.azc.uam.mx

The solution of the geographical clustering problem includes a combinatorial classification of the geographical units. The aggregation proposed in this work requires an objective function that minimizes the distance between the objects that will be clustered together, in order to achieve geo-graphical compactness (a desirable goal in problems of geographical design). Because this problem is NP hard [10], it is usually solved with heuristic methodologies that can proportionate satisfactory so-lutions in a reasonable amount of computational time, even for large problems. The main purpose of this research, it is to propose a Box-Behnken experimental design applied into the response's surface, in order to evaluate the quality of the generated solutions. The balance and adequacy of Simulated Annealing's parameters would help to control and direct the heuristic method to obtain good solutions for the partitioning problem.

Keywords: Geographical Clustering, Experimental Design, Response's Surface, Simulated Annealing.

Mathematics Subject Classification: 62H30, 62K20.

1 Introducción

El problema de Conglomerado Geográfico (CG) consiste en la clasificación de unidades geográficas (UG) sujetas al cumplimiento de ciertos criterios como el de compacidad geométrica, que es el nos ha ocupado en los últimos trabajos [3, 4, 5]. Las UG que se han considerado corresponden a AGEBSs (reas Geoestadísticas Básicas) [24].

Dada la complejidad combinatoria del problema de CG [1, 13, 14, 18, 23], en este trabajo se presenta una propuesta matemática y computacional para plantear y resolver la tarea específica de agrupación geográfica bajo el cumplimiento de una medida de disimilitud como función objetivo. El problema se centra entonces en minimizar dicha función de costo entendida como compacidad sobre AGEBS. Para optimizar esta función objetivo se utiliza un método de gran eficiencia en la resolución heurística de problemas difíciles de optimización combinatoria: Recocido Simulado (RS). Con el propósito de cuantificar la calidad de las soluciones generadas se ha aplicado una metodología estadística factorial [15].

1.1 Aspectos generales de CG

El problema de CG cae en la categoría de Diseño Territorial (DT) de donde se desprende la cual desprende una gran diversidad de problemas que han sido abordados desde diferentes ángulos [1, 8, 9, 13, 16, 23].

En términos generales, DT puede ser visto como un problema de agrupación de áreas geográficas pequeñas (áreas básicas o unidades geográficas básicas) en grupos geográficos más grandes llamados territorios, de tal forma que la agrupación aceptable es aquella última que cumpla con criterios predeterminados del problema que ocupa [23]. Estos criterios a cumplir obedecen a la naturaleza de un particular problema donde restricciones espaciales son muy demandadas [1, 7, 17, 21].

La condición NP-duro de un problema de DT implica resolver un gran número de tareas geográficas donde destaca el proceso de clasificación sujeto al cumplimiento de una función

de costo que minimice distancias entre los objetos a agrupar [1]. A nivel internacional han existido esfuerzos similares encaminados a generar de manera automática agrupaciones geográficas. Sin embargo, y hasta donde sabemos, ninguno ha abordado la agregación del territorio utilizando un método de optimización combinatoria como apoyo para la generación de grupos considerando como unidades territoriales a los AGEBS.

En México, se cuenta con importantes contribuciones pioneras para esta línea de investigación, como lo son parcelación de territorio nacional y distritación electoral [18, 24]. En ambos casos consideran a las manzanas como las unidades geográficas a agrupar (lo que facilita establecer la compacidad geométrica entre manzanas recurriendo a la geometría computacional como una excelente herramienta). Sin embargo, al considerar AGEBS como unidades geográficas para clasificar, los métodos de adyacencia conocidos para obtener tal compacidad, no facilitan el proceso dado que los AGEBS están separadas por distancias no uniformes y su estructura espacial es heterogénea entre cada UG. Justamente esta es la naturaleza espacial de los AGEBS en México.

Debido al carácter combinatorio del problema CG, la propuesta de este trabajo se sitúa en el diseño, desarrollo e implementación de un algoritmo de particionamiento sobre unidades geográficas AGEBS de una zona metropolitana. Para evitar la generación de mínimos locales, en este algoritmo se hace necesaria la inserción de métodos heurísticos, donde la función de costo considera los aspectos fundamentales de agregación territorial: compacidad para ubicación geográfica de los datos.

Con la inclusión de RS es posible escapar favorablemente de mínimos locales y al mismo tiempo mejorar el desempeño del algoritmo de particionamiento que hemos diseñado. Por otro lado de Experimentos Box-Benhken y Superficies de Respuestas [15] para obtener condiciones favorables de ajuste de parámetros de la heurística y contar con valores que posibiliten la obtención de soluciones subóptimas de calidad en problemas pequeños.

Dado que actualmente no se disponen de metodologías claras para determinar cómo calibrar parámetros de una heurística para lograr calidad de soluciones, nuestra aportación se centra justamente en este punto. Conscientes de que RS tiene propiedades de parámetros que la definen y que el control de estos bajo procesos sistemáticos permiten encontrar bondad en los resultados, en este trabajo estamos presentando una técnica para balancear estos parámetros que orienten a la generación de soluciones buenas y cercanas al óptimo para CG.

Se han considerado trabajos sobre clasificación bajo criterios de minimización de distancias que han sido de apoyo en este artículo pero sin ofrecer métodos sistemáticos que demuestren cómo la variación de sus parámetros hacen que sus instancias garanticen buenas soluciones. En particular PAM (Partitioning Around Mediods) propuesto por Kaufman y Rousseeuw (1987) [10, 19], es un buen algoritmo de particionamiento exacto con la desventaja de tener alto costo computacional [19]. Sin embargo, ha sido necesario implementar PAM para clasificar AGEBS con el fin de obtener una solución exacta y comparar las soluciones generadas por RS para problemas pequeños y los hemos utilizado para calibrar los parámetros de la heurística.

Los datos que hemos considerado a clasificar corresponden a los AGEBS de la Zona Metropolitana del Valle de Toluca (ZMVT) [24]. Las variables de clasificación están conformadas por 57 variables socioeconómicas disponibles para dichas áreas.

Se ha integrado a RS al algoritmo de particionamiento que presentamos en la sección 2. La estrategia consiste en elegir k AGEBS como centroides de manera aleatoria para identificar el número de grupos (conglomerados). Aquellos AGEBS que no son centroides serán parte de un determinado grupo si la distancia hacia el centroide es menor que la distancia hacia otro centroide. Considerada así una solución inicial, se crea una solución vecina de la misma manera eligiendo nuevos k centroides. Se compara esta solución vecina (solución actual) con la solución inicial para determinar que tan buena es con respecto a la anterior.

Una vez que se ha obtenido una solución final se hace necesario proponer métodos para validar la calidad de la solución [2]. Para ello, bajo la aplicación de Box-Benhken, hemos encontrado un conjunto de instancias para ser evaluadas y a su vez con la aplicación de la metodología de Superficies de Respuestas se obtuvieron valores para la calibración de los parámetros de RS que facilitan la generación de soluciones hacia un mínimo global.

En congruencia con lo descrito anteriormente, el documento se encuentra organizado como sigue: esta introducción como sección 1, se describe el diseño de un modelo de optimización para cluster geográfico en la siguiente sección. Para dar inicio a la validación de los parámetros, en el apartado 3 se presentan las instancias y validación del modelo estadístico experimental. En la sección 4 se concluye la validación de los resultados y finalmente en la sección 5 presentamos las conclusiones y trabajo futuro.

2 Un modelo matemático para cluster geográfico

Existen diversas propuestas para resolver problemas de agregación geográfica, una de ellas es el diseño de zonas donde los autores lo implementaron con un algoritmo genético [1]. De acuerdo con esta propuesta, el modelo para el problema del CG para AGEBS se presenta en esta sección (Modelo CG). En el problema de CG las UG son AGEBS, cada AGEB está separado por distancias diferentes de estructura geométrica no uniforme debido a que las AGEBS son datos espaciales [6, 7], su ubicación geográfica está dada por latitud y longitud lo que ha facilitado el cálculo de distancias entre las AGEBS.

Se resuelve la agrupación de AGEBS de tal forma que las AGEBS que componen los grupos estén entre ellas muy cercanas geográficamente donde se requiere el uso de una función de costo que minimice distancias entre estas. Básicamente, la estrategia se basa en elegir aleatoriamente AGEBS como centroides que determinan 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. Definir formalmente compacidad no es simple [21], sin embargo, en la definición 1 se plantea la compacidad para UG [6, 22]:

Definición 1. Compacidad
Si denotamos por $Z = \{1, 2, \ldots, n\}$ al conjunto de n objetos a clasificar, se trata de dividir Z en k grupos $\{G_1, G_2, \ldots, G_k\}$ con $k < n$, de tal forma que:

$$\bigcup_{i=1}^{k} G_i = Z$$
$$G_i \cap G_j = \emptyset, i \neq j$$

$$|G_i| \geq 1, i = 1, 2, \ldots, k$$

Un grupo G_m con $|G_m| > 1$ es compacto si para cada objeto $t \in G_m$ cumple:

$$\min_{i \in G_m} d(t,i) < \min_{j \in Z - G_m} d(t,j),\ i \neq t. \qquad (CV1)$$

Un grupo G_m con $|G_m| = 1$ es compacto si su objeto t cumple:

$$\min_{i \in Z - \{t\}} d(t,i) > \min_{j,l \in G_f} d(j,l), \forall f \neq m.$$

El criterio de vecindad entre objetos para lograr la compacidad está dado por los pares de distancias descritos en (CV 1).
Con la idea de la definición 1 y con el fin de resolver el problema de CG, se presenta la siguiente modelación:

2.1 Modelo para conglomerado geográfico (Modelo CG)

Sea UG el número total de AGEBS. Sea el conjunto inicial de n UG, $UG = \{x_1, x_2, \ldots, x_n\}$, donde: x_i es la $i-$ésima unidad geográfica, (i es el índice de UG), y k es el número de zonas (grupos). Dado que se desean formar grupos y para referirnos a éstos, definimos: Z_i como el conjunto de las UG que pertenecen a la zona i, C_t es el centroide, y $d(i,j)$ es la distancia euclidiana del nodo i al nodo j (de un AGEB a otro). Entonces se tienen como restricciones: $Z_i \neq \emptyset$ para $i = 1, \ldots, k$ (los grupos no son vacíos), $Z_i \cap Z_j = \emptyset$ para $i \neq j$ (no existen AGEBS repetidos en distintos grupos), y $\bigcup_{i=1}^{k} Z_i = UG$ (la unión de todos los grupos son todos los AGEBS).

Una vez que se ha decidido el número k de centroides c_t, $t = 1, \ldots, k$, a utilizar hay que seleccionarlos en forma aleatoria y enseguida asignar los AGEBS a los centroides de la siguiente manera: para cada AGEB i

$$\min_{t=1,\ldots,k} \{d(i, c_t)\}$$

cada AGEB es asignado al centroide más cercano c_t.

Para cada valor de k se calcula la suma de las distancias de los AGEBS asignados a cada centroide y se escoge el mínimo y nit es el número de iteraciones. Esto puede expresarse como:

$$\min_{k=1,\ldots,nit} \left\{ \min \left\{ \sum_{t=1}^{k} \sum_{i \in c_t} d(i, c_t) \right\} \right\}. \qquad (1)$$

2.2 Algoritmo de recocido simulado para la obtención de soluciones subóptimas en CG

Para garantizar la generación de buenas soluciones, se requiere de la inclusión de una heurística dentro del algoritmo de particionamiento de CG. Se ha considerado a RS por ser conocido como un método eficiente que escapa satisfactoriamente de óptimos locales [11, 12]. RS es un método de búsqueda por entornos caracterizado por un criterio de

aceptación de soluciones vecinas que se adapta a lo largo de su ejecución. Hace uso de las variables ya conocidas: Temperatura inicial T_i, Temperatura final T_f, alfa (α) y $L(t)$. Estos parámetros son los que se evalúan en la sección siguiente.

En RS, el criterio de Metropolis que permite aceptar soluciones vecinas se define como sigue.

Definición 2. Criterio de Metropolis.
Sea (S, C) un caso de un Problema de Optimización Combinatorio e i y j dos soluciones con costo $C(i)$ y $C(j)$ respectivamente. Entonces el criterio, de Metropolis, para j es:

$$P_T\{\text{aceptar } j\} = \begin{cases} 1 & \text{si } C(j) < C(i) \\ \exp\left(\frac{C(i)-C(j)}{T}\right) & \text{si } C(j) > C(i) \end{cases}$$

$T > 0$ es el parámetro de control, denominado generalmente temperatura. Si se permite que T alcance valores suficientemente pequeños ya no habrá más movimientos a peores soluciones y la convergencia será a un óptimo local [11, 12].

El siguiente algoritmo de RS puede ser adaptable prácticamente a cualquier problema de optimización combinatoria.

Algoritmo de RS (RS 1)
Sean $C(s)$ el costo de la solución actual y $V(s)$ una vecindad
Seleccionar una solución inicial s_n
Seleccionar una temperatura inicial $T_i > 0$
Seleccionar función de reducción de temperatura
Seleccionar un número de iteraciones *nrep*
Seleccionar un *criterio de parada*

REPETIR
  REPETIR
    seleccionar aleatoriamente una solución $s \in V(s_n)$
    $\delta = C(s) - C(s_0)$
    si $\delta < 0$ entonces $s_n = s$
    si no
      generar aleatoriamente $x \in U(0, 1)$
      si $x < \exp(-\delta/t)$, $s_n = s$
    fin si no
    hasta que cuenta-iteraciones = *nrep*
    $t = \alpha(t)$
Hasta criterio de parada

Por otro lado, el algoritmo (RS 1), permite adecuarse al problema de CG. Se implementará la función de Costo 1 del Modelo CG con una variante sencilla del algoritmo anterior y se ha escrito en pseudococódigo con el fin de que sea adaptable al algoritmo de particionamiento para CG.

Pseudocódigo de RS (RS 2)
INPUT $(T_0, \alpha, L(t), T_f)$

$T \leftarrow T_0$ (Valor inicial del parámetro de control)
$Sact \leftarrow$ Genera solución inicial
WHILE $T \geq T_f$ DO (Condición de parada)
BEGIN
FOR cont $\leftarrow$ 1 TO $L(T)$ DO (Velocidad de Enfriamiento (T))
BEGIN
Scand $\leftarrow$Selecciona solución N(Sact) (Generación de una nueva solución)
$\delta \leftarrow$ costo(Scand) $-$ costo(Sact) (Cálculo de la diferencia de costos)
IF $U(0,1) < e^{(-\delta/T)}$ OR (Aplicación del criterio de aceptación)
END
$T \leftarrow \alpha(T)$ (Mecanismo de enfriamiento)
END
{Escribe como solución la mejor de las Sact visitadas}

Finalmente el algoritmo de particionamiento para AGEBS con la inclusión de RS2 para CG queda integrado de la manera descrita en la sección siguiente.

2.3 Algoritmo de recocido simulado y particionamiento para cluster geográfico (RS-CG)

Sea n el número de objetos a clasificar.
UG_{ij} denota que el objeto i está asignado al centroide j
$i = 1, \ldots, n; j = 1, \ldots, k$
Sea $M = \{M_1, M_2, \ldots, M_k\}$ una solución de K centroides
T_0 es la temperatura inicial
T_f es la temperatura final
$L(t)$ es el número de iteraciones que se van a realizar con la misma temperatura

1. Inicio
Obtiene Solución inicial
Generar aleatoriamente centroides iniciales $M = M_1, M_2, \ldots, M_k$
Cualquier ageb puede ser centroide obtenido de forma aleatoria
costo_act $\leftarrow$ Costo(M)$*$
Esta asignación representa ya una Solución inicial, es una Solución propuesta generada por el paso anterior. En los siguientes pasos se genera otra Solución (Solución vecina) para determinar qué tan buena es con respecto a la actual y decidir si se cambia o no la Solución actual.
Mientras $T \geq T_f$
mientras el sistema No esté frio
Para *cont* $= 1$ hasta $L(t)$ hacer
número de ciclos a realizar con la misma temperatura (parametro de RS)
$C \leftarrow$ Genera una Solución aleatoria
*se genera la Solución que se compara con **
costo_cand $\leftarrow$ Costo(C)

se obtiene el costo de Solución candidata que se ha generado
$\delta \leftarrow$ costo_cand − costo_act
diferencia de costos para obtener el valor de probabilidad de aceptación de la solución candidata

Si $U(0,1) < e^{-\delta - T}$ ó $\delta < 0$ hacer

si la probabilidad de aceptación aún es alta
M ← C
si se acepta la Solución candidata
costo_act ← costo_cand
Fin Si
Fin para
$T \leftarrow \alpha(T)$
se está enfriando el sistema
Fin Mientras
Fin

2. Función Costo (Sol)
determina qué tan buena es la solución SOL, es decir, qué tanto minimiza el objetivo
$i \leftarrow 1$
inicializa primer objeto
$cost \leftarrow 0$
Mientras $i \leq n$
para cada objeto en Ug hacer
si Ug_i no es centroide entonces
$dmin \leftarrow dist(Sol_1, Ug_i)$
representa la distancia del objeto$_i$ hacia Sol_1 (primer centroide donde Sol representa al conjunto de todos los centroides. Se calcula la distancia cada objeto a su centroide más cercano, (distancia de un objeto i que no es centroide hacia Sol1 que es el centroide 1)
$j \leftarrow 2$
paso al segundo centroide
Mientras $j \leq k$
Si $dist\,(Sol_j, Ug_i) < dmin$
se calcula la distancia del objeto i hacia Sol_j (otro centroide)
$dmin \leftarrow dist(Sol_j, Ug_i)$
Fin si
$j \leftarrow j + 1$
paso al siguiente centroide
Fin Mientras
$cost \leftarrow cost + dmin$
Fin si
$i \leftarrow i + 1$
Fin Mientras

$Costo\,(Sol) \leftarrow cost$

Una vez implementado $(RS - CG)$ y para observar la diferencia entre un óptimo y las soluciones que genera dicho algoritmo, estas se han comparado con los resultados de PAM dado que su proceso de clasificación agota todas las combinaciones posibles y crea un valor exacto en problemas pequeños pero con alto costo computacional [10, 20].

3 Ajuste de parámetros

Una de las pruebas que es importante realizar sobre los resultados obtenidos es evaluar la calidad de los resultados usando para esto un método sistemático que permita identificar el efecto de los parámetros de control sobre el valor de la función de costo, modelar la dependencia de esta función respecto a los parámetros y finalmente poder hacer un estudio sobre la influencia de los parámetros en la búsqueda por encontrar mínimos ya sea locales o generales de la función [2]. Para ello hemos considerado un diseño experimental de superficies de respuestas que nos ha permitido observar los efectos descritos en el párrafo anterior. Este tipo de experimento es una prueba o serie de pruebas en las cuales se inducen cambios deliberados en algunas variables de entrada del sistema mientras otras se mantienen fijas, de tal forma que es posible identificar las fuentes de los cambios en las variables de salida [15].

3.1 Diseño de un experimento que permita modelar los resultados del efecto de los predictores de la función de costo

La metodología de superficies de respuesta es una combinación de técnicas de diseño y análisis de experimentos que, utilizadas en forma secuencial, permiten determinar condiciones de operación que son óptimos locales para el problema a tratar. Una función compleja suave puede aproximarse localmente (es decir, en zonas "pequeñas" de la región de operación) mediante polinomios de orden bajo. Si la zona donde se realiza la aproximación local está "lejos" de la zona donde se encuentra un máximo local entonces un polinomio de primer orden deberá ser una buena aproximación. En cambio, si la zona está "cerca" del máximo local será necesario utilizar un polinomio de segundo orden para describir a la función [15].

El análisis sistematizado que hemos mencionado se desarrolló utilizando un diseño tipo Box-Behnken (BB), este tipo de diseño por sus características es fácil de llevar a cabo definiendo niveles adecuados de los parámetros de diseño, además de que es un diseño rotable o sea con igual varianza para todos los puntos de experimentación que se encuentran a la misma distancia del centro del diseño, y por otro lado es posible hacer experimentos secuenciales para estudiar los efectos individuales de los parámetros de control y los efectos combinados de los mismos de manera simultanea. Otra de las ventajas de este diseño es que permite modelar los resultados con una función de segundo orden y por lo tanto desarrollar un análisis del comportamiento de la función de costo utilizando la metodología de superficies de respuesta. Los diseños BB se forman combinado factoriales

Parámetro	Nivel Alto	Nivel Central	Nivel bajo
T_i	5500	5250	5000
T_f	0.1	0.055	0.01
A	0.99	0.985	0.98
$L(t)$	5	4	3
Grupos	24	18	12

Tabla 1: Niveles y parámetros utilizados en el experimento BB para el problema CG.

2^k con diseño de bloques incompletos. Los diseños resultantes suelen ser más eficientes en términos del número de corridas facilitando su generación [15].

Para nuestro problema se ha utilizado un diseño BB con cinco parámetros de control que giran alrededor de 24 grupos, el cual es un número que habla sobre un buen punto de inflexión en la agrupación [3, 5]. Con esta información introducida a BB, el experimento resultante ha sido una muestra de 46 corridas significativas dado se han utilizado cuatro puntos centrales [15]. La elección de los niveles de los parámetros usados en la construcción del experimento obedece a los resultados obtenidos por el método heurístico, lo que ha permitido definir una región de experimentación. Los niveles integrados se muestran en la tabla 1.

Con estos nivles y el diseño BB se han llevado a cabo las 46 corridas experimentales que se muestran en la tabla 2. La nomenclatura utilizada en la tabla es: C (corrida), T_i (Temperatura Inicial), T_f (Temperatura final), α (alpha), Lt ($L(t)$), G (Grupos), FC (Función Objetivo). En esta tabla, para la corrida 29 se obtiene el óptimo 14.12 para 12 grupos y para la corrida 31 se obtiene el óptimo 9.279 para 24 grupos.

Las figuras 1 y 2 muestran el comportamiento de la heurística para dos casos: 24 grupos que se ha identificado como el experimento más confiable y 12 grupos como el menos adecuado, es decir, al comparar la diferencia de la función de costo contra las instancias de la tabla 2 para 24 grupos, notamos que el valor es menor que la diferencia de otra función de costo hacia las corridas restantes. Para las corridas asociadas a 12 grupos se observa que la diferencia que existe hacia el valor exacto es mayor que las corridas respectivas para 18 y 24 grupos. En dichas figuras se observa el costo de la función objetivo contra el número de iteraciones. Cada caso se ha extraído de la tabla anterior eligiendo la corrida 36 como aquella que mejor se ha acercado al óptimo siendo el principal parámetro de referencia el número de grupos. En la corrida 36 observamos que con 24 grupos y con los parámetros de $T_i = 5500$, $T_f = .055$, $\alpha = .985$, $L(t) = 4$, se generó un costo de la función objetivo de 11.2403, el más cercano al óptimo obtenido por PAM que es de 9.279. En contraste con el tiempo que logra PAM para generar la solución exacta que fue de 17 horas [6], RS con 3049 iteraciones, 2183 soluciones aceptadas, reduce el costo computacional a un segundo.

C	T_i	T_f	α	Lt	G	FC	C	T_i	T_f	α	Lt	G	FC
2	5500	0.01	0.985	4	18	13.588	25	5000	0.055	0.985	3	18	13.660
3	5000	0.1	0.985	4	18	14.034	26	5500	0.055	0.985	3	18	13.535
4	5500	0.1	0.985	4	18	14.122	27	5000	0.055	0.985	5	18	14.026
5	5250	0.055	0.98	3	18	13.917	28	5500	0.055	0.985	5	18	13.067
6	5250	0.055	0.99	3	18	14.129	29	5250	0.055	0.98	4	12	16.850
7	5250	0.055	0.98	5	18	13.235	30	5250	0.055	0.99	4	12	17.108
8	5250	0.055	0.99	5	18	13.893	31	5250	0.055	0.98	4	24	12.215
9	5250	0.01	0.985	4	12	16.216	32	5250	0.055	0.99	4	24	11.728
10	5250	0.1	0.985	4	12	16.55	33	5000	0.055	0.985	4	12	16.696
11	5250	0.01	0.985	4	24	11.539	34	5500	0.055	0.985	4	12	16.783
12	5250	0.1	0.985	4	24	12.029	35	5000	0.055	0.985	4	24	11.884
13	5000	0.055	0.98	4	18	16.302	36	5500	0.055	0.985	4	24	***11.240***
14	5500	0.055	0.98	4	18	14.110	37	5250	0.01	0.985	3	18	13.558
15	5000	0.055	0.99	4	18	13.916	38	5250	0.1	0.985	3	18	13.211
16	5500	0.055	0.99	4	18	13.955	39	5250	0.01	0.985	5	18	13.700
17	5250	0.055	0.985	3	12	***15.635***	40	5250	0.1	0.985	5	18	14.760
18	5250	0.055	0.985	5	12	16.084	41	5250	0.055	0.985	4	18	13.927
19	5250	0.055	0.985	3	24	12.331	42	5250	0.055	0.985	4	18	13.822
20	5250	0.055	0.985	5	24	11.638	43	5250	0.055	0.985	4	18	13.583
21	5250	0.01	0.98	4	18	13.520	44	5250	0.055	0.985	4	18	13.989
22	5250	0.1	0.98	4	18	14.304	45	5250	0.055	0.985	4	18	13.639
23	5250	0.01	0.99	4	18	13.3445	46	5250	0.055	0.985	4	18	12.901

Tabla 2: Corridas experimentales determinadas por el experimento BB.

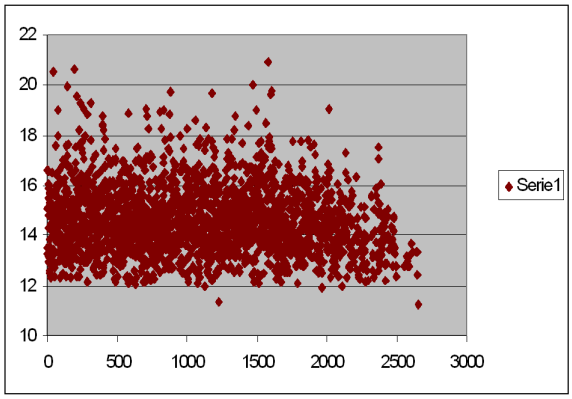

Figura 1: Corrida 36 con 24 grupos.

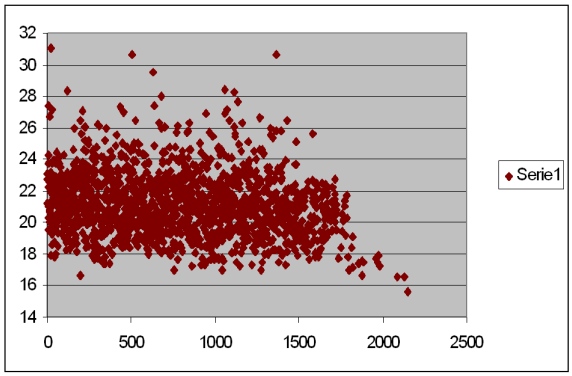

Figura 2: Corrida 17 con 12 grupos.

3.2 Verificación del modelo experimental

En la figura 3 se observan los resultados de verificación del modelo, concluyendo que los datos se comportan normalmente, que el modelo de segundo orden es adecuado y que no existen efectos de una corrida a otra en el experimento.

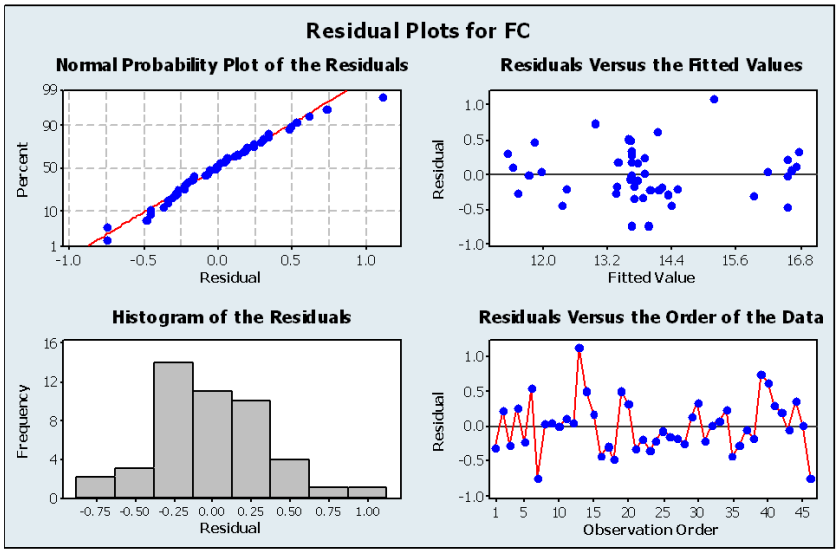

Figura 3: Verificación del modelo.

Una vez desarrollado este experimento y analizando la información obtenida, se ha ajustado los resultados con un modelo de regresión de segundo orden, obteniéndose la ecuación de predicción mostrada en la tabla 3.

Término	Coeficiente	SE Coef	T	P
Constant	16963.0	6707.0	2.529	0.018
TI	-0.5	0.2	-2.360	0.026
TF	358.1	1106.2	0.324	0.749
alfa	-31921.4	13443.1	-2.375	0.026
l(t)	-16.2	49.8	-0.326	0.747
grupos	6.2	8.3	0.742	0.465
TI*TI	0.0	0.0	1.301	0.205
TF*TF	-0.4	84.0	-0.005	0.996
alfa*alfa	15020.8	6801.6	2.208	0.037
L(t)*L(t)	-0.1	0.2	-0.692	0.495
grupos*grupos	0.0	0.0	2.323	0.029
TI*TF	0.0	0.0	0.028	0.978
TI*alfa	0.4	0.2	2.220	0.036
TI*l(t)	0.0	0.0	-0.831	0.414
TI*grupos	0.0	0.0	-0.727	0.174
TF*alpha	-395.8	1116.3	-0.355	0.726
TF*l(t)	7.8	5.6	1.401	0.173
TF*grupos	0.1	0.9	0.152	0.880
alfa*L(t)	22.3	50.2	0.445	0.660
alfa*grupos	-6.2	8.4	-0.742	0.465
l(t)*grupos	0.0	0.0	-1.138	0.266

Tabla 3: Regresión de segundo orden, con $S = 0.5023$, $R^2 = 93.8\%$ y $\tilde{R}^2 = 88.8\%$.

4 Validación de la variación de los parámetros

En esta sección mostramos las gráficas de superficies de respuestas y de predicción que son obtenidas con el modelo descrito en la sección anterior. Se presentan los gráficos de contorno que a su vez son generadas por conclusiones que responden al análisis de las superficies de respuesta. Esto es, al identificarse en que regiones se alcanzan valores cercanos al óptimo de la función objetivo, y con el fin de observar gráficamente este comportamiento, se han graficado contornos que revelan como esta función se ajusta para regiones donde los parámetros de predicción de la misma son los adecuados.

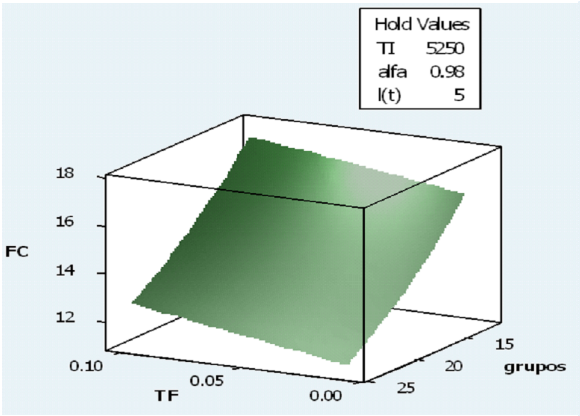

Figura 4: Función de costo 1.

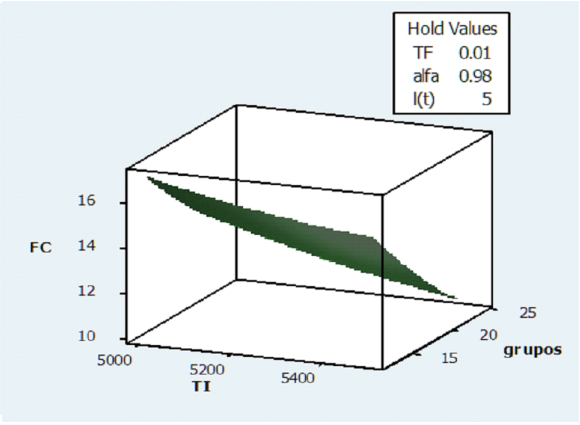

Figura 5: Función de costo 2.

En la función de costo 1 se ha mantenido fija la temperatura inicial, alfa y el número de iteraciones. Como se puede observar al cambiar la temperatura final y el número de grupos, se logra un mínimo de la función de costo para temperaturas finales pequeñas y grandes números de grupos. La función de costo 2 muestra el efecto sobre la función de costos de variar la temperatura inicial y el número de grupos considerados en el experimento, se ha mantenido en niveles fijos la temperatura final, α y el número de iteraciones para $L(t)$. En este punto se concluye que se logra un mínimo de la función de costo para una temperatura inicial alta y el mayor número de grupos posibles (Figuras 4 y 5).

La figura 6 revela el resultado de la modelación de la función de costo manteniendo fijas la temperatura inicial y final con α. Considerando el comportamiento antes observado, el mínimo continúa apareciendo para el mayor numero de grupos, sin embargo en este caso el mejor mínimo corresponde a un número bajo de iteraciones, incrementando su valor al crecer el número de iteraciones. Este comportamiento es contrastante. En la función de costo-4 se ha mantenido fija la temperatura inicial, la temperatura final y el número de iteraciones, nuevamente encontramos que esta función es mínima para un numero grande de grupos, en este caso además observamos que alfa debe ser grande para lograr el mejor mínimo; este comportamiento es consistente con lo observado en las anteriores figuras.

Del análisis de las gráficas de modelación antes mostradas podemos concluir lo siguiente:

1. La función de costo siempre tiene un mínimo para el número mayor de grupos.

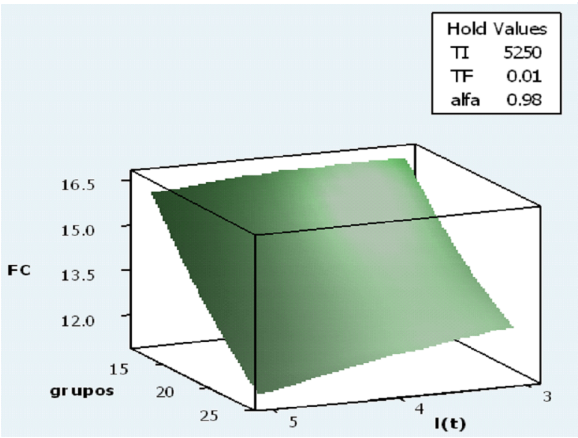

Figura 6: Función de costo 3.

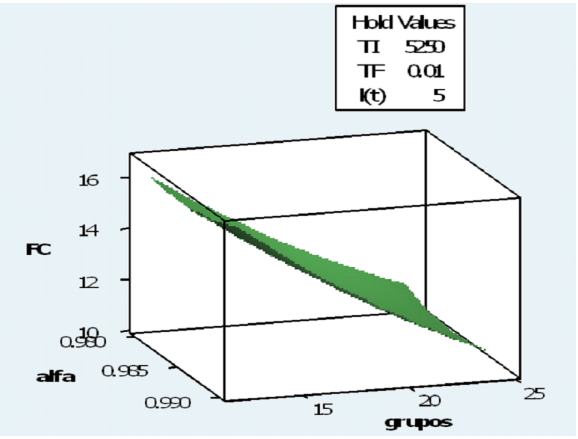

Figura 7: Función de costo 4.

2. El mínimo ocurre para un número de iteraciones pequeñas en el cálculo.

3. El valor de alfa debe ser grande

4. La temperatura final debe ser pequeña.

5. Y la temperatura inicial alta.

Este análisis permite acotar la magnitud de los parámetros de impacto de la función de costo, para buscar que esta sea un mínimo. A continuación presentamos algunas gráficas de contorno donde se obtienen mínimos de la función de costos, cuando hacemos uso de las conclusiones antes obtenidas.

4.1 Gráfico de contornos (curvas de nivel)

La figura 8 representa la curva de nivel de la función de costos ajustada para regiones cercanas al óptimo para 24 grupos.

En el contorno para 24 grupos se ha fijado T_f a .01, α (alfa) a .98 y 24 grupos en la región donde se observan funciones de costo mínimas. Se distingue el comportamiento de la función de costo para valores de T_i y $L(t)$ cercanos al óptimo y al mismo tiempo destaca el mínimo de la FC obtenido con los mejores parámetros.

4.2 Optimización de la función de costo usando el modelo de regresión

Recurriendo al modelo de segundo orden, en la siguiente figura se ha encontrado que con la variación de todos los parámetros en conjunto sin fijar a alguno en particular, es posible obtener un valor muy cercano al óptimo. Siendo el costo real de la función objetivo de 9.27 para 24 grupos, el mínimo alcanzado en este caso es de $y = 10.3597$ y está dado por los parámetros de $T_i = 5477.6723$, $T_f = 0.102$, $\alpha = .980$ y $L(t) = 4.9775$.

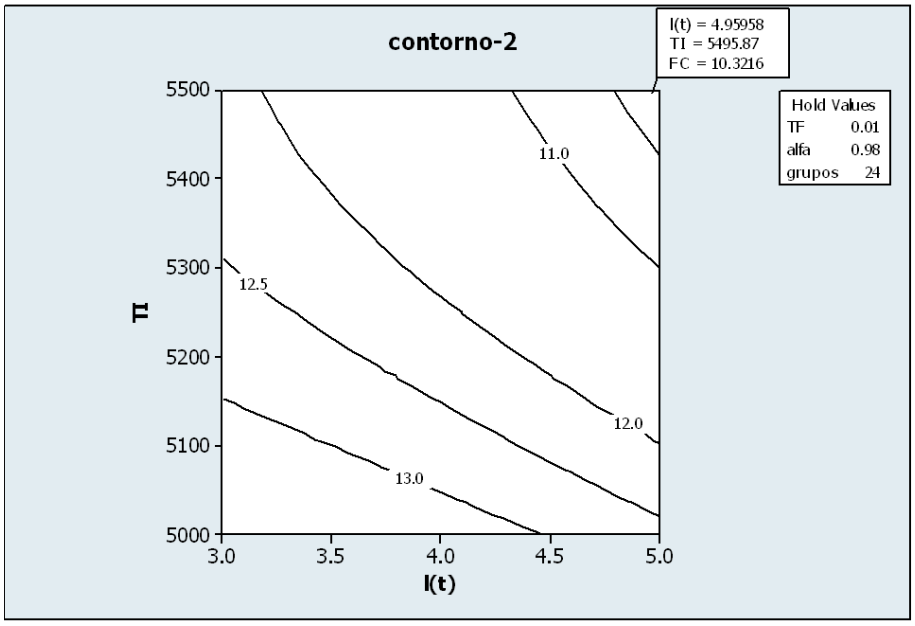

Figura 8: Contorno para 24 grupos.

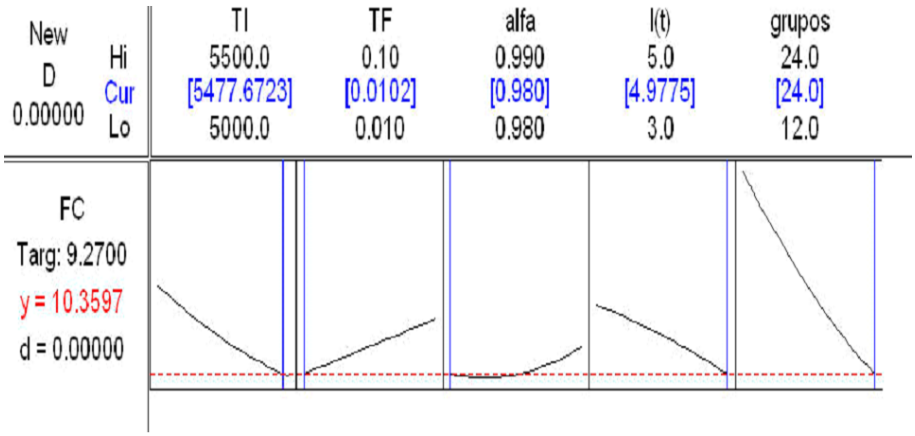

Figura 9: Representación del modelo de segundo orden para 24 grupos.

5 Conclusiones

De los resultados obtenidos en este trabajo, hemos podido concluir que los parámetros de la heurística de recocido simulado implementada para el problema de cluster geográfico, son sensibles para diferentes condiciones.

a.) En términos generales entre mayor sea el número de grupos más cerca estamos del óptimo.

b.) La temperatura inicial debe estar próxima a 5000 unidades independientemente del número de grupos el costo de la función objetivo converge al óptimo.

c.) Al fijar la temperatura final y alfa en los contornos, la variación de los otros parámetros restantes debe estar bajo control tal y como se muestra en las gráficas de contorno de la sección anterior. Se ha determinado entonces que para los 3 contornos resultantes, fijando .01 para temperatura final y .98 en alfa se logra un buen mínimo en la función de costo.

d.) Cuando se han considerado la variación de todos los parámetros, es claro que un valor de alfa de .980 debe ser exigido mientras que la temperatura final debe ser pequeña con un valor de .01.

Dado que el experimento se inició a partir de los resultados analizados en corridas empíricas donde se determinó que 24 era un buen número de grupos [3, 5], el diseño que hemos presentado en este trabajo fue alimentado tal y como se muestra en la tabla 1. Con estos datos se desarrolló todo el trabajo correspondiente. No se reportó en este artículo el proceso para encontrar un punto estacionario debido a que no pudo ser observable y por tanto no logramos encontrar la ecuación canónica, tan útil en experimentos como el que hemos descrito. Una de las líneas de trabajo a seguir parte de justamente este punto: ampliar el experimento con un mayor número de grupos debido que este fue determinante para alcanzar un mínimo.

Suponemos que al aumentar el valor de los parámetros y generar más instancias, el experimento siendo más extenso daría lugar a encontrar la ecuación canónica.

Por otro lado se está trabajando con la heurística de vecindad variable para el problema de cluster geográfico y construir un diseño de experimentos para esta heurística.

Finalmente estamos reportando la inclusión de un Sistema de Información Geográfica con el fin de revelar los resultados en mapas de tal forma que sea evidente la agrupación geográfica compacta [21].

Referencias

[1] Bação, F.; Lobo, V.; Painho, M. (2004) "Applying genetic algorithms to zone design", *in Springer Verlag.*

[2] Barr R.S.; Golden J.P.; Resende M.G.C.; Stewart W.R. (1995) "Designing and Reporting on Computational Experiments with Heuristics Methods", *Journal of Heuristics*, **1**: 9–32.

[3] Bernábe, L.B.; López, S. (2004) "Statistical classificatory analysis applied to population zones", *8th. World Multiconference on Systemics, Cybernetics and Informatics*, Orlando.

[4] Bernábe, L.B.; Osorio, M.A.; Duque, J.C. (2006) "Clasificación sobre zonas geográficas: un enfoque de optimización combinatoria para el problema de regionalización", *XIII CLAIO Congreso Latino-Iberoamericano de Investigación Operativa*, Montevideo.

[5] Bernábe, L.B.; Aguirre, V.R.; López, S.R. (2004) "Application of non-supervised classification to population data", *ICEEE/CIE2004, International Conference on Electrical and Electronics Engineering*, Acapulco. ISBN 0-7803-8531-4.

[6] Bernábe, L.B. (2006) "Desarrollo de un modelo para la determinación de zonificación óptima", *Proyecto de tesis doctoral en desarrollo*, Posgrado de Ingeniería UNAM, Investigación de Operaciones.

[7] Cliff, A.D.; Haggett, P.; Ord, J.K.; Bassett, K.A.; Davies, R.B. (1975), *Elements of Spatial Structure: a Quantitative Approach*. Cambridge University Press, Cambridge.

[8] Hess S.W.; Samuels S.A. (1971) "Experiences with a sales districting model: criteria and implementation", *Management Science, Series B: Application* **18**: 41–54.

[9] Kalcsics, J.; Nickel, S.; Schröder, M. (2005) *Towards a Unified Territory Design Approach. Applications, Algorithms and GIS Integration*. Universität des Saarlandes, Germany.

[10] Kaufman, L.; Rousseeuw, P. (1987) "Clustering by means of medoids", *Statistical Data Analysis*: 405–416.

[11] Kirkpatrick, S.; Gelatt, D.; Vecchi, M.P. (1983) "Optimization by simulated annealing", *Science* **220**: 671–680.

[12] Lebster, I. (1995) "Adaptative simulated annealing", in: (ASA): lesson learned. Technical Report, Control and Cybernetic, McLean VA.

[13] Macmillan, W.; (2001) "Redistricting in a GIS environment: an optimization algorithm using switching points", *Journal of Geographical Systems* **3**: 167–80.

[14] Mehrotra, A.; Johnson, E.; Nemhauser, G. (1998) "An optimization based heuristic for political districting", *Management Science* **44**: 1100–1114.

[15] Montgomery, D. (1991) *Design and Analisis of Experiments*, 2nd edition. Wiley, New York.

[16] Murtagh F. (1985) "A survey of algorithms for contiguity–constrained clustering and related problems", *Computer Journal* **28**: 82–88.

[17] Openshaw S.; Taylor P. (1981) "The modifiable area unit problem", in: N. Wrigley & R. Bennett(Eds.) *Quantitative Geography*, London: 60–70.

[18] Romero, D.; Burguete, J.; Martínez, E.; Velasco, J. (2004) "Parcelación del territorio nacional: un enfoque de optimización combinatoria para la construcción de marcos de muestreo en hogares", *INEGI*, México.

[19] Rousseeuw, P.J.; Hubert, M.; Struyf, A. (1997) "Clustering in an object-oriented environment", *Journal of Statistical Software* **1**: 2–10.

[20] *MapX Developers Guide*, MapInfo corporation, Troy NY.

[21] Takeshi, S. (2004) "A model of contiguity for spatial unit allocation", *Geographical Analysis*, Institute for Geoinformation, Technical University of Viena, Austria, ISSN 0016-7363.

[22] Zamora, A.E. (2006) "Implementación de un algoritmo compacto y homogéneo para la clasificación de zonas geográficas AGEBs bajo una interfaz gráfica", *Tesis de Ingeniería en Ciencias de la Computación*, BUAP, Puebla.

[23] Zoltners, A.; Sinha, P. (1983) "Towards a unified territory alignment: a review and model", *Management Science* **29**: 1237–1256.

[24] `http://www.inegi.gob.mx`, Instituto Nacional de Estadística, Geografía e Infomática (INEGI), México.