Self Organizing Maps en el Análisis de Datos Espaciales

Self Organizing Maps in Spatial Data Analysis

Jose Franco, *Estudiante, Universidad EAFIT, Alejandro Betancourt, Estudiante, Universidad EAFIT,*

***Abstract*—Given that the area of Spatial Data Analysis is a field where the application of Artificial Intelligence techniques can yield very interesting results, in this work a methodology for georeferenced data clustering based in the Neural Network model known as Kohonen Maps is developed. To do this, the networks are adapted in such a way that they receive naturally data that takes in account its position in space, and the competitive training algorithm is adapted so that the spatial contiguity restriction inherent to spatial clustering techniques is respected. The resulting algorithm yields good results, is fast and flexible. Results are shown and then applications are presented using socio-demographic data from Accra (Ghana) and China.**

***Resumen* - Dado que el área del análisis de datos espaciales es un campo donde la aplicación de las técnicas de inteligencia artificial puede entregar resultados muy interesantes, en este trabajo se desarrolla una metodología para la agrupación de datos georeferenciados basada en las redes neuronales conocidas como Mapas de Kohonen. Para hacer esto se adaptan las redes de manera que reciban naturalmente datos que tienen en cuenta su posición en el espacio, y se adapta el algoritmo de entrenamiento competitivo de los mapas de Kohonen para que respete la restricción de contigüidad espacial inherente a los algoritmos de clustering espacial. El algoritmo resultante entrega buenos resultados, es rápido y flexible. Se muestran los resultados y luego se presenta una aplicación utilizando datos socio-demográficos de Accra (Ghana) y China.**

***Index Terms*—Self Organizing Maps, Mapas de Kohonen, Análisis de Datos Espaciales, Clustering, Clustering Espacial, Agrupación**

I. Introducción y marco teórico

Últimamente, el volumen de datos que tienen en cuenta el contexto espacial, es decir, datos georeferenciados, se ha visto en rápido aumento, llegando a un punto en el que se ha convertido en una prioridad para muchas empresas y organizaciones implementar técnicas que permitan explotar la información adicional que se puede vislumbrar en esta nueva fuente de conocimiento.

Sin embargo, el hecho de trabajar tanto con coordenadas espaciales como datos de alta dimensionalidad, hace que la tarea de visualización y análisis resulte altamente compleja. Debido a esta complejidad los investigadores han incrementado sus esfuerzos para el diseño de metodologías y algoritmos que permitan agrupar las áreas de los mapas en regiones homogéneas, buscando así obtener mapas con una menor cantidad de áreas pero que no pierdan la representación de los fenómenos que este capturaba en su distribución original.

Los autores agradecen al profesor Juan Carlos Duque, PhD. por su invaluable colaboración en término de materiales y experiencia en el desarrollo de este trabajo y por permitir la utilización del software clusterPy como base para el desarrollo e implementación de la técnica acá presentada.

A. Área de aplicación

La ciencia regional es una disciplina que combina elementos de la estadística, la geografía, la geometría y la cartografía para realizar un análisis más completo de cualquier información disponible susceptible de ser representada por medio de un mapa. Es por esta razón que los temas en los que se puede aplicar es bastante extenso y va desde análisis de patrones criminales [1] hasta la epidemiología [2], pasando así por temas como, el riesgo incendiario, planteamiento de políticas de votación, localización de nuevas sedes para una empresa, etc.

Sin embargo la disminución de la gran dimensionalidad de los datos no es la única utilidad de las distintas metodologías de agregación espacial. Una de las ventajas mas importantes de esta ciencia es la búsqueda de regiones que incluyen áreas que comparten características similares en su interior, que permitan analizar fenómenos restringidos a regiones bien delimitadas.

Es tanta la importancia del análisis de datos espaciales que actualmente es fácil encontrar información sobre el tema, ya que son muchos los investigadores que actualmente están interesados en el análisis de este tipo de información. Para comenzar, se tiene una amplia gama de libros que plantean las bases de la estadística y la econometría espacial, que establece las bases necesarias para realizar un estudio comprensivo de datos con estas características [3],[4].

B. Metodologías de agregación

Desde el punto de vista computacional, son muchos los algoritmos y metodologías propuestas para agregación espacial. En [5] se hace una revisión comprensiva sobre los distintos tipos de algoritmos de agregación que se encuentran en la literatura. Particularmente, [6] propone un algoritmo para la detección de clusters espaciales por medio de un algoritmo combinatorio que luego en [7] se reformula de una manera computacionalmente eficiente.

Por otra parte, respecto a los Mapas de Kohonen, áreas previamente abordadas incluyen generación de imágenes médicas, análisis del mercado de valores, organización de exhibiciones en un museo [8], minería de datos [9], reconocimiento de patrones [10], análisis de datos espaciales [11], análisis de datos estadísticos [12], entre otros.

La búsqueda de regiones geográficas con características homogéneas puede aprovechar la metodología de los Mapas de Kohonen debido a la capacidad inherente de las redes neuronales con aprendizaje no supervisado de reconocer y agrupar patrones en los datos de entrada. En [13] se aplica un algoritmo modificado para la detección de estos patrones regionales.

C. Self Organizing Maps

Los Self Organizing Maps son un tipo de red neuronal artificial, que se utiliza para producir visualizaciones en dos dimensiones de conjuntos de datos multi-dimensionales, utilizando entrenamiento no supervisado con ciertas características que permiten preservar las propiedades topológicas de los datos originales. En análisis de datos espaciales, esta propiedad los hace idóneos para la búsqueda de conjuntos de datos homogéneos. Sin embargo no se garantiza el cumplimiento de la restricción de contigüidad espacial. En este artículo se desarrolla una aplicación que implementa esta metodología para el análisis de datos espaciales.

Los self organizing maps son una red neuronal compuesta de dos capas, una de entrada y una de salida, que utiliza aprendizaje no supervisado como mecanismo de entrenamiento para encontrar las ponderaciones adecuadas para clasificar los elementos de entrada, generando así una visualización apropiada. En la figura 1 se puede observar la estructura de esta red.

Para el entrenamiento de los Self Organizing Maps, el algoritmo clásico propuesto por Kohonen es el de aprendizaje competitivo con una modificación que da cuenta de la ubicación espacial de las neuronas de la capa de salida. Este procedimiento se observa en el algoritmo 1.

Algorithm 1 Entrenamiento de un Self Organizing Map ($\eta(t)$: Tasa de entrenamiento, $h_t(i,j)$: Función de vecindad)

```
Inicializar los pesos aleatoriamente
for i=0 hasta i=MaxIt do
    for x en vectores de entrada do
        Elegir el nodo de salida i tal que sus pesos minimicen la distancia euclídea a x
        Actualizar los pesos: Para toda j, W_j(t+1) = W_j(t) + η(t)h_t(i,j)(x − W_j(t))
    end for
    Actualizar η(t), v_j(t)
end for
```

Cada neurona de la capa de entrada recibe un vector X_i en el que se almacena el valor de la variable i para cada área del mapa. Dicho vector está dado por:

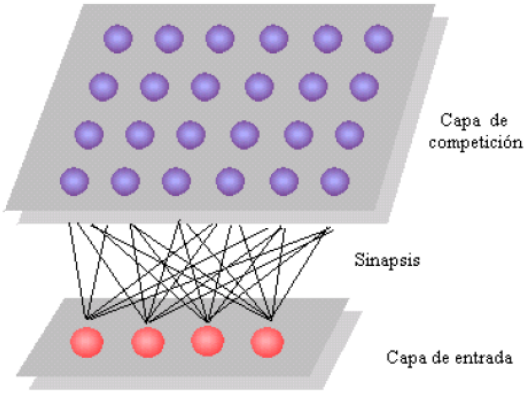

Fig. 1. Estructura de la red neuronal

$$X_i = [X_{i1}, X_{i2}, X_{i3}, \ldots, X_{iN}]$$

Esta primera capa está conectada con cada una de las neuronas de salida por medio de un conjunto de pesos W_{ij} que están encargados de enlazar la neurona de entrada i con la neurona de salida j. Por último, la capa post-sináptica o capa de salida está conformada por m neuronas ubicadas sobre una superficie bidimensional teselada regularmente de forma rectangular o hexagonal. Esta última será la encargada de mostrar de manera resumida la correlación espacial entre las variables de entrada y de sugerir las regiones que cuentan con áreas homogéneas en su interior.

Se han realizado comparaciones entre esta metodología y otras como agrupación utilizando K-means. Los Self Organizing Maps resultaron menos propensos a caer en mínimos locales [14]. Adicionalmente, el algoritmo K-means no tiene en cuenta explícitamente la restricción de contigüidad espacial.

II. Estado del arte

En [5] se presenta una revisión sobre los métodos de agrupación espacial que han emergido a través de los tiempos. La principal motivación para el desarrollo de estas técnicas han sido problemáticas como la preservación de la confidencialidad al aprovechar grandes volúmenes de información desagregada, la reducción de los efectos de outliers o errores en la toma de datos, y la facilitación de la visualización e interpretación de la información en los mapas.

A. Agregación jerárquica

Los métodos más simples de agregación espacial consisten en el clustering tradicional de los datos utilizando métodos jerárquicos. Posteriormente, se impone la restricción de contigüidad espacial separando las regiones que no cumplan con ella. Otro tipo de algoritmos que no tienen en cuenta la contigüidad espacial son los de maximización de la compacidad regional, que intenta obtener regiones lo más parecidas posibles a un círculo. Estos algoritmos, aunque fáciles de implementar y relativamente rápidos, no tienen en cuenta directamente las características espaciales del problema, lo que afecta la calidad de las soluciones y limita sus campos de aplicación.

B. Agregación exacta

Otros métodos mucho más sofisticados emplean optimización lineal y problemas de programación entera mixta para encontrar regiones homogéneas analíticas de manera exacta. Algunos métodos utilizan conceptos de la teoría de grafos para tener en cuenta la coordenada espacial de los datos. Los inconvenientes de esta aproximación tienen que ver con el tamaño de los problemas: Aunque las soluciones entregadas son muy buenas, la optimización exacta de estas formulaciones toma mucho tiempo.

C. Agregación heurística

Una tercera agrupación de métodos para agrupar áreas teniendo en cuenta la restricción espacial es la gran batería de algoritmos heurísticos que resuelven problemas similares utilizando diferentes aproximaciones. Estos modelos tienen en cuenta la restricción de contigüidad espacial de diferentes formas, según el modelo heurístico subyacente a cada método. Una última categoría consiste en los modelos mixtos heurísticos, que tienen como objetivo combinar la eficiencia de los heurísticos con la calidad de los modelos exactos.

D. Agregación no exhaustiva

Una categoría diferente de métodos de agregación son aquellos donde no se intenta asignar absolutamente todas las áreas a agrupaciones determinadas. Estos métodos usualmente se utilizan para la detección de regiones donde las variables tienen comportamiento atípico, como AMOEBA [**?**], [7], que asigna clusters de valores altos y bajos, pero no se dice nada más sobre las áreas que no resulten en ningun cluster.

III. Desarrollo del modelo

Como se mencionó anteriormente, los Mapas de Kohonen son un modelo muy eficiente para encontrar patrones entre los datos, o en el contexto particular del análisis de datos espaciales, encontrar grupos homogéneos de áreas. Sin embargo, el hecho que sean homogéneos no quiere decir que conformen regiones contiguas, por lo que se debe modificar el algoritmo de manera que la restricción de contigüidad espacial se cumpla.

A. Modificación de la retícula de salida del Self Organizing Map

Para comenzar, se pensó en proyectar las neuronas de salida del mapa de Kohonen sobre el mapa estudiado, es decir, se tomó la topografía del mapa donde se encuentran las áreas para la red neuronal. Esto tiene la ventaja que la función de vecindad pasa a tener un sentido muy concreto: Dos neuronas son vecinas cuando las áreas correspondientes en el mapa lo son. Esto se puede visualizar en la figura 2. Adicionalmente, se implementó la restricción de que en el proceso de entrenamiento, un área puede activar únicamente una neurona de la que sea vecina; De esta manera cualquier par de áreas que queden agrupadas estarán también vinculadas espacialmente.

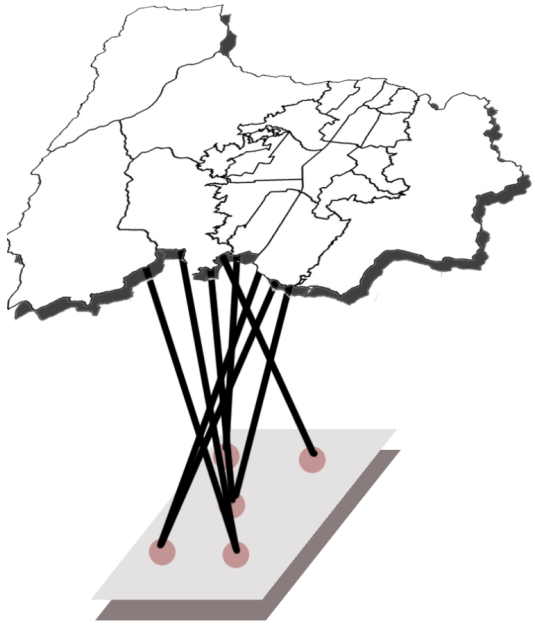

Fig. 2. Topografía de la red neuronal

Toda la información sobre los datos georeferenciados se resume en una matriz Y de datos, donde cada fila es un área y cada columna corresponde a una variable medida en ése área, y la matriz de contacto W, donde cada elemento $w_{i,j}$ es 1 si el área i es contigua al área j, o 0 en otro caso. W tiene la información sobre la vecindad entre las áreas, que básicamente consiste en que dos áreas son vecinas si están conectadas espacialmente por lo menos por un punto (En caso de utilizar el criterio reina) o por una línea (Si se utiliza el criterio torre). Teniendo esto en cuenta, es fácil definir la función de vecindad h:

$$h(i,j) = \begin{cases} 1 & \text{si } i = j \\ \rho & \text{si el área i es contigua al área j} \\ 0 & \text{en otro caso} \end{cases}$$

O alternativamente,

$$h(i,j) = w_{i,j}$$

B. Cumplimiento de la restricción de contigüidad espacial

Es claro que la matriz Y se utilizará como la entrada a la red neuronal. Por cada columna de Y se crea una neurona, y los individuos que se utilizarán para entrenar la red son las filas de esta matriz.

De manera adicional, para que el algoritmo entregara clusters unificados, se implementaron algunas restricciones adicionales sobre qué neurona podía activar cada área durante el entrenamiento de la red. Antes de mencionar estas restricciones es necesario establecer que la neurona lider de un cluster es el área que representa al cluster en cada iteración, es

decir, es el centro del cluster que se está formando durante el entrenamiento. Las restricciones adicionales son las siguientes:

1) Un área i sólo puede activar una neurona j si j es una neurona líder de cluster o j no ha sido procesada aún.
2) Un área i sólo puede activar una neurona de salida j si $w_{i,j} = 1$.
3) Cuando una neurona de entrada i ingresa a un cluster con neurona líder j, se redefine el vecindario de la neurona j como la unión entre los vecindarios de i y j.

Se puede observar como dada la estructura de las neuronas elegibles para ser activadas, es posible que los clusters resultantes no tengan muchas áreas, dependiendo de la cantidad presente en el mapa. Un mapa con pocas áreas puede tener pocos clusters, mientras que uno con muchas áreas normalmente tendrá muchos clusters. Para unificar el número de clusters resultantes en cada mapa, se propone una variación del algoritmo.

C. Algoritmo iterativo

Dadas las restricciones que se agregaron, los clusters que entrega el algoritmo son normalmente pequeños. Para obtener niveles mayores de agregación, se volvió iterativo el algoritmo de manera que se puede repetir hasta llegar a algún número de clusters deseado. Para hacer esto, simplemente se tomó la nueva matriz W como la matriz de contactos de los clusters resultantes, mientras que la nueva Y se toma como los pesos de la neurona que representa el cluster en la iteración anterior.

IV. Experimentos computacionales

Los experimentos computacionales se realizaron ejecutando sobre los mapas de Accra en china el algoritmo de agrupación. La taza de aprendizaje utilizada fue $\eta = 0.9$, máximo de iteraciones 100, 10 iteraciones de agrupación sucesiva y parámetro de vecindad $\rho = 0.9$. Estos parámetros fueron así decididos despues de un período de experimentación.

A. Accra

En las figuras 3 y 4 se observa el comportamiento del algoritmo para agrupar las áreas de Accra, la capital de la República de Ghana, un país situado en Africa. La variable utilizada para realizar la agregación fue la del área de cada polígono, para facilitar la interpretación de los resultados.

El mapa de Ghana tiene una cantidad muy grande de áreas (1717), por lo que se observa que en las primeras iteraciones el nivel de agregación es muy pequeño y tiene una cantidad muy grande de clusters. El número de clusters por cada iteración se encuentra en la tabla I. Dado un número deseado de clusters, es fácil saber que agregación tomar: Si se desean 15 clusters se podría tomar la penúltima, mientras que si se quieren 25 se puede tomar la iteración 7.

El tiempo tomado para ejecutar la agrupación de Accra fue muy pequeño, especialmente a comparación de otros algoritmos de agregación que normalmente toman algún tiempo con un número tan grande de áreas.

Iteracion	Regiones
1	579
2	298
3	160
4	86
5	56
6	35
7	25
8	18
9	15
10	10

TABLE I
NUMERO DE REGIONES POR ITERACION, ACCRA

TABLE II
NUMERO DE REGIONES POR ITERACIÓN, CHINA

Iteracion	Regiones
1	12
2	8
3	4
4	4
5	3
6	2
7	2
8	2
9	2
10	2

B. China

Se estudió el mapa de china con la población por provincia en el año 1998 para realizar agrupaciones de áreas similares. La variable graficada en el mapa se encuentra en 8, donde se observa claramente que existen agrupaciones de áreas similares, notablemente las provincias occidentales que tienen un carácter más rural, mientras que las provincias de la costa están más pobladas, posiblemente por fenómenos relacionados con el impulso a la economía que realiza el comercio. Idealmente, el algoritmo capturaría la costa como un cluster y las demás áreas según el número de áreas.

Las agrupaciones realizadas por el algoritmo iteración por iteración se observan en la figura 5, donde se ve claramente como en cada ejecición del algoritmo se va capturando la dinámica y aumentando el nivel de agregación hasta llegar al punto donde queda resumido el mapa entero en dos regiones: La parte rural de china y la parte costera. Las regiones agrupadas por iteración se encuentran en la tabla IV-B.

China es un mapa con 28 regiones, por lo que a diferencia de Accra unas pocas iteraciones son suficientes para capturar el comportamiento de la variable en el mapa.

C. Análisis de resultados

Se observa que para mapas pequeños unas pocas iteraciones son suficientes para conseguir un buen número de clusters, mientras que para mapas con muchas áreas es necesario iterar repetidamente hasta conseguir la cantidad deseada de regiones. Esto valida la modificación que se realizó sobre el algoritmo.

En ambos casos se observa como la dinámica de la variable analizada fue capturada por las agrupaciones encontradas, esto indica que el algoritmo da buenos resultados en términos de agregación, es decir, encuentra regiones homogéneas que son

1

2

3

4

5

Fig. 3. Iteraciones 1 - 5 del algoritmo de agrupación

1

2

3

4

5

Fig. 4. Iteraciones 6-10 del algoritmo de agrupación

representativas de verdaderos conglomerados cuyas subdivisiones están relacionadas en términos de las variables medidas.

En las figuras 6 y 7 se observa como el número de clusters disminuye rápidamente por iteración, lo que resulta ser una buena herramienta para encontrar el número de clusters deseado.

V. Conclusiones

El algoritmo original de Self Organizing Maps, no garantiza contigüidad espacial. Es por esta razon que en este trabajo se propone una modificacion al algoritmo original por medio de la imposicion de algunas restricciones que garantizan la restricción mencionada.

El algoritmo modificado a pesar de garantizar la contiguidad espacial presentaba deficiencias en el numero de regiones que retornaba, por esta razon se implemento una nueva modificacion que lo convierte en un algoritmo iterativo que reduce en cada ronda el numero de regiones, convirtiendose asi en un proceso adecuado para la agregación espacial.

Los Self Organizing Maps o Mapas de Kohonen resultan ser una herramienta extremadamente útil como algoritmo de agregación espacial, ya que no sólo permiten procesar mapas de manera rápida, sino que además capturan de una manera adecuada las características especificas de la distribución espacial de las variables con las cuales se corre el algoritmo. Es por esta razon que este algoritmo puede llegar a ser muy util para el usuario como herramienta de analisis exploratorio de datos espaciales, pues permitiría rápidamente hacerse una primer idea de el numero de regiones a buscar con un algoritmo mucho mas exacto y con mejor estructura matemática.

Como trabajo futuro se planea investigar mas a fondo en el algoritmo 1, esperando asi evitar tener que convertirlo en iterativo para la reduccion de areas, lo que permitiría tomar deciciones mas justificadas a la hora de delimitar las las regiones.

References

[1] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: A general approach," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 2004.

[2] K. Clarke, McLafferty, P. D, P. D, and Tempalski, "On epidemiology and geographic information systems: A review and discussion of future directions," *Computers, Environment, and Urban Systems*, vol. 2, pp. 85–92, 1996.

[3] S. J. Rey, "Spatial dependence in the evolution of regional income distributions." in *Spatial econometrics and spatial statistics*, A. Getis, J. Múr, and H. Zoeller, Eds. Hampshire: Palgrave, 2002, forthcoming.

[4] S. Magrini, "Regional (di)convergence," in *Handbook of Regional and Urban Economics*, V. Henderson and J. Thisse, Eds. New York: Elsevier, 2004.

[5] J. Duque, R. Ramos, and J. Surinach, "Supervised regionalization methods: A survey," *International Regional Science Review*, 2007.

[6] J. A. A Getis, "Using amoeba to create a spacial weights matrix and identify spacial clusters," *Geographical analysis*, 2005.

[7] J. Franco, A. Betancourt, and J. Duque, "A computationally efficient formulation for the aldstadt and getis amoeba algorithm," in *First congress of the Regional Science Association of the Americas*, 2009.

[8] A. Skupin and S. I. Fabrikant, "Spatialization methods: a cartographic research agenda for non-geographic information visualization," *Cartography and Geographic Information Science*, vol. 30, pp. 95–115, 2003.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, 1996, pp. 1–34.

[10] J. Han and M. Kamber, "Data mining: Concepts and techniques - google book search," 2000.

[11] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[12] L. Kaufman and P. Rousseeuw, *Finding Groups in Data An Introduction to Cluster Analysis*. New York: Wiley Interscience, 1990.

[13] F. Bacao, O. Bacao, V. Lobo, and M. Painho, "Geo-self-organizing map (geo-som) for building and exploring homogeneous regions," 2004.

[14] ——, "Self-organizing maps as substitutes for k-means clustering," in *International Conference on Computational Science (3)*, 2005, pp. 476–483.

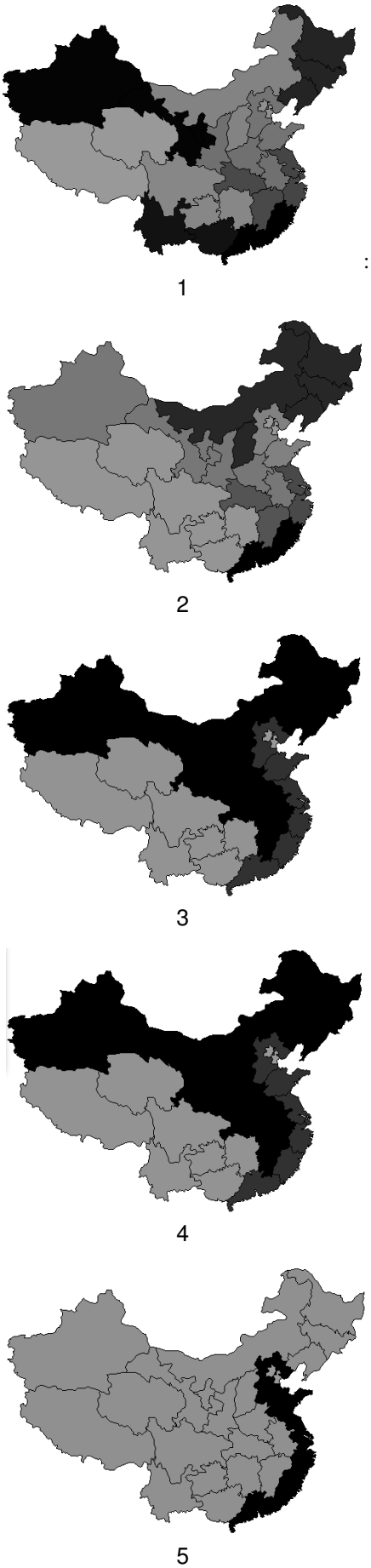

Fig. 5. Iteraciones del algoritmo de agrupación en china, utilizando como variable la población por provincia en 1998

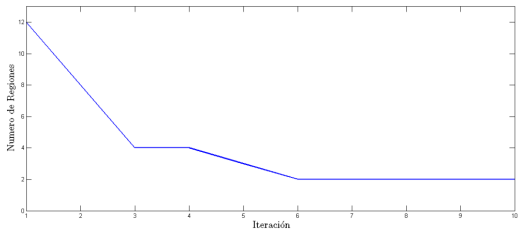

Fig. 6. Convergencia del algoritmo en china

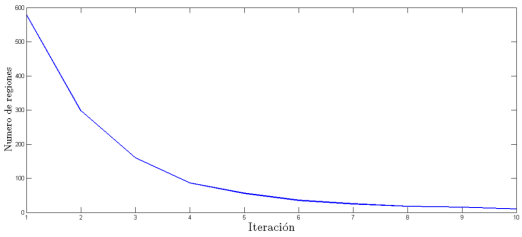

Fig. 7. Convergencia del algoritmo en accra

Fig. 8. Distribución de la población china en 1998

Fig. 9. Resultado del algoritmo