Representación y clasificación de datos geoespaciales: comparación entre mapas autoorganizativos de Kohonen y el método Gas Neuronal*

Marly Esther De Moya Amarís**

Representation and classification of geospatial data: a comparison of Kohonen's maps and the Neural Gas method

Resumen

Aproximadamente el 80% de toda la información existente en el mundo corresponde a información georreferenciada. Esto crea una creciente necesidad de disponer de herramientas más flexibles, precisas y fáciles de usar para la visualización, exploración y clasificación de grandes volúmenes de datos geoespaciales. En esta investigación preliminar se comparan diferentes técnicas para representar y clasificar datos georreferenciados utilizando dos tipos de redes neuronales: mapas autoorganizativos de Kohonen (SOM) y el método Gas Neuronal (NG). El estudio incluye dos tipos de análisis: visualización y clasificación. Para el estudio correspondiente a visualización se escogieron dos tipos de datos: En primer lugar se seleccionó una muestra de 23000 coordenadas (x, y, z) de una zona montañosa de Colombia con el objetivo de analizar la capacidad de cada uno de los métodos para modelar el terreno, es decir, para presentar visualmente la forma del relieve. El segundo conjunto de datos corresponde a la población de cada uno de los 1090 municipios de Colombia (coordenadas x, y, y población total). El objetivo poder visualizar geográficamente la densidad poblacional de cada una de las regiones. Para el análisis de clasificación igualmente se seleccionaron dos conjuntos de datos: el primero corresponde a la codificación climática de los municipios de Colombia; el segundo, a la clasificación de los municipios en su respectivo departamento. Para los casos de visualización, SOM mostró un mejor desempeño que NG, dándose el caso contrario para los ejemplos de clasificación.

PALABRAS CLAVE

Sistemas de Información Geográfica (SIG), minería de datos espaciales, redes neuronales, Mapas Autoorganizativos (SOM), Kohonen, Gas Neuronal

Abstract

Approximately 80% of all the existing information in the world correspond to geo-referenced information. This creates an increasing necessity to have tools more flexible, precise and easy to use to the visualization, exploration and classification of great volumes of geospatial data. Additionally its necessary achieve smaller times to process this kind of information. In this preliminary investigation, different techniques are compared to visualize and to classify geo-referenced data using two types of neuronal networks: Kohonen's maps (SOM) and the Neural Gas method (NG). For the visualization cases, SOM showed a better performance than NG, occurring the opposite case for the classification examples.

KEY WORDS

Geographic Information System (GIS), spatial data mining, neural networks, Self Organizing Map (SOM), Kohonen, Neural Gas.

* Las gráficas de este artículo tienen la mejor calidad que se pudo lograr, a partir del archivo suministrado por el autor (nota del editor).

** Ingeniera de sistemas, estudiante del Magíster en Ingeniería de Sistemas de la Universidad Nacional de Colombia. Administradora del Sistema de Información Geográfica del IGAC. Tel.: 3694006, Ext. 4410. Correo electrónico: me_demoya@igac.gov.co.

1. Introducción

La creciente cantidad de información georeferenciada obliga a que los sistemas de información geográfica (GIS) tengan que almacenar y manejar volúmenes de información cada día más grandes, lo cual se traduce en tiempos de ejecución bastante altos y gran consumo de recursos de hardware y de software.

Un aporte importante para los GIS sería poder brindar herramientas para visualizar y analizar estos grandes volúmenes de información con un menor consumo de recursos y con tiempos de procesamiento mucho más bajos que los ofrecidos por los métodos tradicionales. Un camino para lograr esto consiste en encontrar un conjunto de datos prototipos que actúen como representantes de los datos originales, es decir, un conjunto de datos que, siendo mucho más pequeño en cantidad, represente el mismo comportamiento de los datos originales. Entonces, en lugar de procesar el conjunto masivo de datos, únicamente se procesa el conjunto de datos prototipo, también conocidos como "vectores prototipo".

El objetivo de esta investigación es analizar y comparar dos métodos para encontrar estos vectores prototipo utilizando redes neuronales, aplicado específicamente a visualización y clasificación de datos geoespaciales. Se comparará el método de mapas autoorganizativos de Kohonen (SOM) [1] con el método Gas Neuronal (NG) [2].

En la sección 2 se presenta una descripción de los mapas autoorganizativos de Kohonen y del método Gas Neuronal; en la sección 3 se presenta el marco experimental, que describe los conjuntos de datos utilizados en los análisis, así como su correspondiente preprocesamiento. En la sección 4 se analizan los resultados obtenidos tanto para visualización como para clasificación utilizando ambas técnicas. En la sección 5 se analizarán algunos indicadores para comparar la eficiencia de los SOM y del método NG en la generación de vectores prototipos como representantes del conjunto completo de datos. Finalmente, en la sección 6 se resume el trabajo realizado y se presentan las conclusiones.

2. Redes competitivas: mapas autoorganizativos y Gas Neuronal

Las redes competitivas son redes uni o multicapa cuyo común denominador es postular algún tipo de competición entre unidades con el fin de conseguir que una de ellas quede activada y el resto no. Esto se consigue mediante aprendizaje no supervisado, presentando algún patrón de entrada y seleccionando la unidad cuyo patrón de pesos incidentes se parezca más al patrón de entrada, reforzando dichas conexiones y debilitando las de las unidades perdedoras.

La competición entre unidades se puede conseguir simulando una característica neurofisiológica del córtex cerebral, llamada *inhibición lateral*. Esto se logra postulando la existencia de conexiones inhibitorias intracapa y conexiones excitatorias intercapa, de manera que la presentación de un patrón de entrada tenderá a producir la activación de una única unidad y la inhibición del resto.

Al final se consigue que cada unidad responda frente a un determinado patrón de entrada y, por generalización, que cada unidad responda frente a patrones de entrada similares, de modo que los pesos aferentes de esa unidad converjan en el centro del grupo de patrones con características similares.

Es usual que haya una capa de neuronas de entrada y una capa de salida. Se usan tantas entradas como dimensiones tenga el espacio vectorial de los patrones de entrada (espacio real o binario), y tantas salidas como clases o categorías se quieren utilizar para clasificar los patrones de entrada, de manera que cada nodo de salida representa una categoría.

Para nuestro caso de estudio, se analizarán dos tipos de redes neuronales competitivas: los mapas autoorganizativos de Kohonen y el método Gas Neuronal.

2.1 Mapas autoorganizativos de Kohonen

Los mapas de Kohonen permiten convertir relaciones estadísticas complejas no lineales (establecidas entre un conjunto de datos de muchas dimensiones) en una sencilla relación geométrica en un espacio de pocas dimensiones. Además permiten la compresión de información preservando las relaciones topológicas y métricas más

importantes. Estos dos aspectos, visualización y abstracción, pueden ser utilizados en muchas tareas complejas; por ejemplo la visualización y clasificación de datos geoespaciales.

El aprendizaje del modelo de Kohonen es no supervisado de tipo competitivo. Las neuronas de la capa de salida compiten por activarse y sólo una de ellas permanece activa ante determinada información de entrada a la red. Los pesos que las conexiones se ajustan en función de la neurona que haya resultado vencedora.

El algoritmo de aprendizaje utilizado para establecer los valores de los pesos de las conexiones entre las N neuronas de entrada y las M de salida es el siguiente:

1. Iniciar los pesos (w_{ij}) con valores aleatorios pequeños y fijar la zona inicial de vecindad entre las neuronas de salida.

2. Presentar una entrada en forma de vector $E_k = (e_1^{(k)},..., e_N^{(k)})$, donde los componentes $e_i^{(k)}$ serán valores continuos.

3. Determinar la neurona vencedora de la capa de salida. Esta será aquella que tenga el valor más parecido al patrón de entrada E_k. Para ello, se calculan las distancias o diferencias entre ambos vectores, considerando una por una todas las neuronas de salida:

$$d_j = \sum_{i=1}^{8} (e_i^{(k)} - w_{ij})^2 \quad 1<=j<=M$$

donde:

$e_i^{(k)}$: Componente i-ésimo del vector k-ésimo de entrada.

w_{ij}: Peso de la conexión entre la neurona i de la capa de entrada y la neurona j de la capa de salida.

4. Una vez localizada la neurona vencedora (j^*), se actualizan los pesos de las conexiones entre las neuronas de entrada y dicha neurona, así como los de las conexiones entre las de entrada y las neuronas vecinas de la vencedora.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)\left[e_i^{(k)} - w_{j^*i}(t)\right]$$

donde:

$\alpha(t)$: Parámetro de ganancia o coeficiente de aprendizaje, con un valor entre 0 y 1, el cual decrece con cada iteración.

5. El proceso se debe repetir presentando todo el juego de patrones de aprendizaje un mínimo de 500 veces.

El mapa construido (SOM) es una representación plana de los vectores prototipo, imaginados como puntos localizados en el espacio de datos. La eficiencia de esta representación es medida por dos índices:

1. *Error de cuantización* q_1. Este error corresponde al promedio de la distancia euclidiana de los vectores de datos a sus representantes más cercanos, es decir, al *codebook vector* más cercano.

2. *Error topológico* q_2. Este error indica la fracción de vecinos en el mapa, los cuales no tienen regiones de Voronoi en el espacio de datos.

2.2 El método Gas Neuronal

El método *Neural Gas* (NG), propuesto por Martinetz [2], se orienta hacia la cuantización del espacio, al igual que el SOM. El método NG es más general comparado con los mapas de Kohonen.

Al igual que el SOM, el método NG usa una función de vecindad. Esta función depende de un parámetro de afinamiento conocido como l. El proceso de aprendizaje en este método es similar al utilizado por el método de Kohonen, pero complementado por la facilidad para manejar el parámetro λ en la función de vecindad. Además, se considera que una ventaja de este método es el hecho de permitir la reducción de las variables de entrada y que éstas vayan creciendo (*growging neural gas*) hasta alcanzar una representación óptima del espacio de entrada.

La mayor diferencia entre los métodos SOM y NG está en que, en el modelo NG, la colección de neuronas no se halla conectada por una red: cada neurona puede moverse libremente a través del espacio de datos. Las coordenadas de las neuronas son llamadas tradicionalmente *pesos*. Si consideramos una colección de *m* neuronas, cada neurona puede ser imaginada como un punto en el espacio de datos. Las coordenadas de la i^{th} neurona será denotada como $w_i = (w_{i1}, ...,w_{ip})$. Al inicio de los cálculos, la colección de neuronas se distribuye aleatoriamente en el espacio de datos. En las siguientes iteraciones, las neuronas cambian su posición y se adaptan a la nube de datos. El proceso de adaptación es lla-

mado *aprendizaje* o *entrenamiento*. El entrenamiento es realizado en ciclos llamados épocas. Durante cada época *n* vectores de datos (elegidos en orden aleatorio del espacio de entrada) son presentados secuencialmente al conjunto de neuronas. Si se suponen *h* épocas, el máximo número de iteraciones de ajuste permitidas es $K_{máx} = n * h$. En cada iteración *k* ($k=0,1,..., K_{máx}$), se presenta un vector de datos aleatorio **x** al conjunto de neuronas. Para cada vector de datos **x** expuesto en la k-ésima iteración se encuentra la neurona más cercana (cercana en el sentido euclidiano). Esta neurona es llamada ganadora y obtiene el índice *w*. El vector de pesos de la neurona ganadora satisface la siguiente relación:

$$d(x, w_w) = \min\ d(x, w_i)\ 1<=i<=m$$

En el paso siguiente, se establece el vecindario de la neurona ganadora. La magnitud (diámetro) del vecindario decrece exponencialmente con (*k*), el número actual de la presentación. Para cada *k*, todas las neuronas pertenecientes al vecindario de la neurona ganadora cambian su posición para ubicarse más cerca del vector **x**, actualmente expuesto. El cambio es descrito por la fórmula (1):

$$w_i = w_i + \alpha(k) G(i, k, x, w, \lambda)(x - w_i)$$

donde la función G describe el vecindario de la neurona ganadora, es decir, la vecindad del vector $\mathbf{w}_w$ (2):

$$G(i, k, x, w, \lambda) = \exp\left\{-\frac{d^2(x, w_i)}{2\lambda^2(k)}\right\}$$

y el índice *i* corre sobre todas las neuronas pertenecientes al vecindario establecido para la ganadora *w*.

Analizando la fórmula (1), se observa que los cambios de posición de la *i*-ésima neurona depende de dos factores: a, y el valor de la función G, el cual depende del parámetro λ. Ambos coeficientes *a* y el parámetro λ dependen de *k*, el número de la actual iteración, lo cual significa:

α(k) : define el coeficiente de aprendizaje. Éste define que tan grande puede ser el cambio de posición. El coeficiente *a* decrece usualmente con el número de iteraciones: tiene un valor inicial α_0 y decrece gradualmente hasta un valor α_{min} alcanzado al final de todas las iteraciones. El valor de decrecimiento de α(*k*) es descrito por la función (3):

$$\alpha(k) = \alpha_0 \left(\frac{\alpha_{min}}{\alpha_0}\right)^{k / k_{max}}$$

λ(*k*): define el diámetro del área de vecindad en la *k*-ésima presentación. Normalmente esta decrece con *k*, el número actual de presentación: para un valor inicial λ_0, decreciendo gradualmente hasta un valor final λ_{min} al final de las presentaciones para $k = k_{máx}$ (4):

$$\lambda(k) = \lambda_0 \left(\frac{\lambda_{min}}{\lambda_0}\right)^{k / k_{max}}$$

Puede decirse que la fórmula anterior expresa la adaptación de la neurona ganadora y sus vecinas en la dirección del vector de datos **x** expuesto.

3. Marco experimental

Se realizarán dos tipos de análisis: visualización y clasificación de datos geoespaciales. Para el estudio correspondiente a visualización se escogieron dos tipos de datos. En primer lugar se seleccionó una muestra de datos tridimensionales de una zona montañosa de Colombia (23368 coordenadas *x, y, z*) (figura 1). Los datos fueron obtenidos en el Instituto Geográfico Agustín Codazzi (IGAC). El objetivo es analizar la capacidad de cada uno de los métodos para modelar el terreno, es decir, para presentar visualmente la forma del relieve.

El segundo conjunto de datos corresponde a la población de cada uno de los municipios de Colombia (1090 en total, cada uno con coordenadas *x*, y, y población total). Los datos fueron obtenidos en la página web del Departamento Nacional de Estadística (DANE). El objetivo es visualizar la densidad poblacional en cada una de las regiones.

Para el análisis de clasificación se seleccionaron dos conjuntos de datos: el primero corresponde a la codificación climática (figura 2) de cada uno de los municipios de Colombia; el segundo, a la clasificación de cada

uno de los municipios en su correspondiente departamento.

Figura 1. Zona de trabajo.

Se comparará la eficiencia de cada uno de los dos métodos (SOM y NG), tanto para visualización como para clasificación de datos geoespaciales, utilizando el software SOMTOOLBOX para Matlab 5.3.

Figura 2. Clasificación climática.

3.1 Preprocesamiento de datos

El primer paso es construir los conjuntos de datos y dejarlos en un formato apropiado para su procesamiento. Los cuatro conjuntos de datos utilizados para los análisis de visualización y de clasificación son almacenados en estructuras de datos. En la figura 3 podemos apreciar dichos conjuntos.

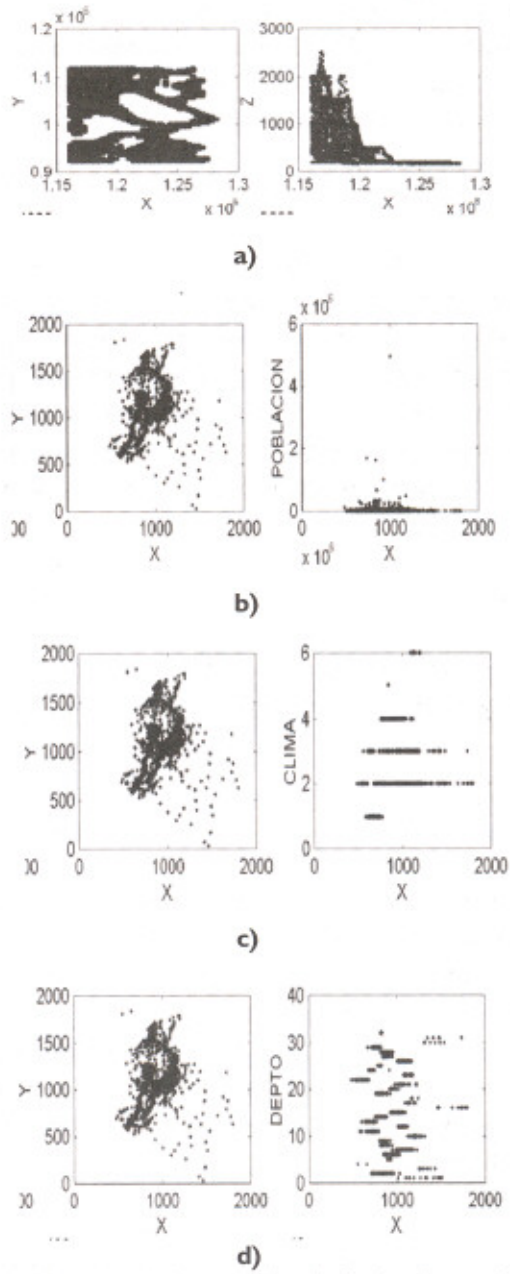

Figura 3. Conjuntos de datos para visualización de relieve y población a) y b) y clasificación climática y departamental c) y d), respectivamente.

El paso siguiente es proceder con la etapa de normalización de los datos, es decir, convertirlos en un rango de valores específico.

Debido a que los algoritmos SOM y NG están basados en las distancias existentes entre los datos, la escala de las variables es muy importante en la determinación de los vectores prototipo. Si el rango de una variable es mucho más grande que el de las otras, probablemente, esta variable dominará por completo la organización de los resultados. Para nuestros casos de análisis, las coordenadas geográficas están en el rango de los cientos de miles, mientras que las alturas y la densidad poblacional están en el rango de miles, y la clasificación departamental y climática en el rango de decenas y unidades respectivamente. Debido a esto se utilizó la función de normalización automática provista por la herramienta SOMTOOLBOX, a fin de dejar los datos en un rango adecuado de valores para su correcto procesamiento.

4. Resultados

A continuación se describen los resultados obtenidos mediante la utilización de los métodos SOM y Gas Neuronal para la visualización y clasificación de información georreferenciada. El análisis fue realizado utilizando el software SOM TOOLBOX para Matlab 5 [4].

4.1 Visualización de datos geoespaciales

En esta sección se presentan los resultados obtenidos para los experimentos de visualización, se analiza en detalle el desempeño de los SOM y del método Gas Neuronal.

4.1.1 Desempeño SOM

En la tabla I se muestran los conjuntos de datos elegidos, el número de ejemplos, los tamaños de los mapas y el desempeño de la red en cuanto a precisión.

Para analizar el comportamiento del SOM se determinaron dos tamaños de mapas. Como se puede observar en la tabla I, la disminución de las dimensiones del mapa en un 50% aumentó en más del 100% el error de cuantización, mientras que el error topológico sólo se incrementó ligeramente, con excepción del caso de clasificación departamental, en el cual el error topológico aumentó un 72 % aproximadamente, debido a que la disposición geográfica de los departamentos de Colombia es bien irregular y una reducción en el número de neuronas puede afectar de gran manera la precisión de la clasificación.

En cuanto a representación gráfica, en la figura 4 se pueden ver las matrices de distancias del SOM obtenido con los ejemplos de visualización de relieve.

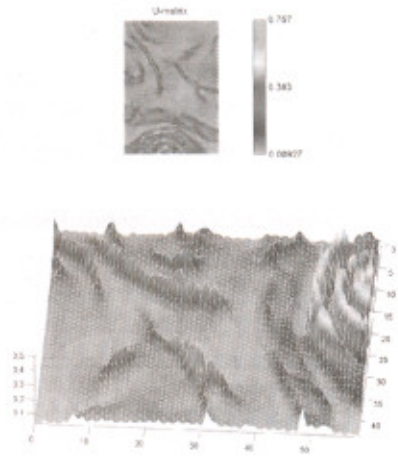

Figura 4. Matrices de distancia obtenidas mediante la aplicación de SOM al primer conjunto de datos para análisis de visualización.

En la figura 5 se observan los vectores prototipo calculados por la red neuronal, donde cada neurona está unida a su vecina por medio de una línea, lo cual genera una malla tridimensional, que simula la forma real del terreno montañoso.

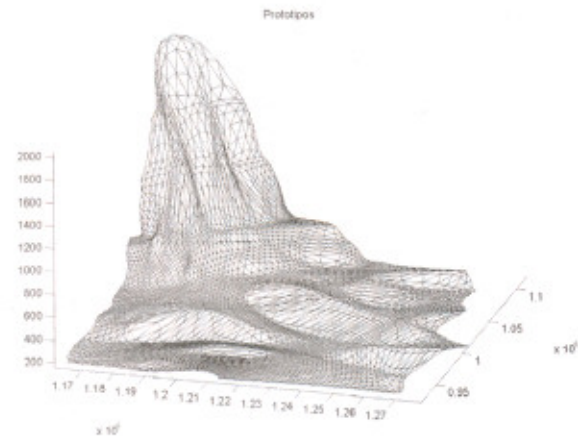

Figura 5. Vectores prototipo obtenidos mediante la aplicación de SOM al primer conjunto de entrenamiento.

Tabla 1. Datos correspondientes a la aplicación de SOM.

Mapas autoorganizativos	Visualización				Clasificación			
	Relieve		Población		Climas		Departamentos	
Ejemplo	x, y ,z		x, y, población		x, y, clima		x, y, departamento	
Número de ejemplos	23368		1090		1090		1090	
Tamaño del mapa	[69, 44]	[34,22]	[29, 23]	[14, 12]	[31, 22]	[15, 11]	[31, 22]	[15, 11]
Error de cuantización	0.0323	0.0786	0.107	0.2028	0.0949	0.2244	0.0988	0.2676
Error topológico	0.0595	0.0775	0.031	0.0312	0.0330	0.0385	0.0284	0.0413

Si se realiza una comparación con los datos reales, y basados en los errores reportados, podemos concluir que el SOM representa con exactitud aceptable la forma real del terreno, obviamente estableciendo un tamaño de mapa adecuado para la cantidad de ejemplos introducidos.

Para el caso de visualización de densidad poblacional, en las figuras 6 y 7 se pueden apreciar las matrices de distancia obtenidas y los vectores prototipo respectivamente.

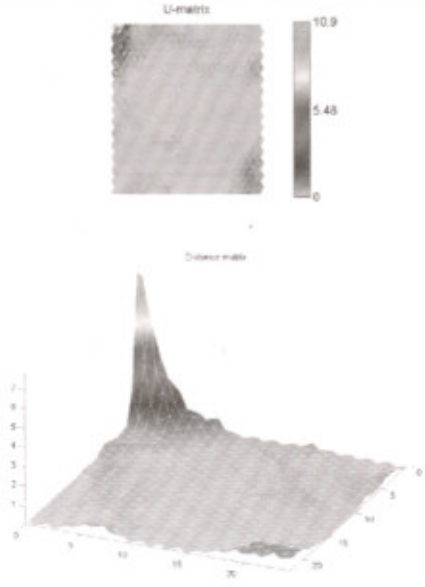

Figura 6. Matrices de distancia obtenidas mediante la aplicación de SOM para visualización de densidad poblacional.

En la figura 7 se observa que el mapa no representa exactamente la distribución geográfica de Colombia (10% de error de cuantización). En la vista tridimensional se puede ver la variación de la densidad poblacional; el pico más alto corresponde a la población de la ciudad de Bogotá, la más alta de Colombia.

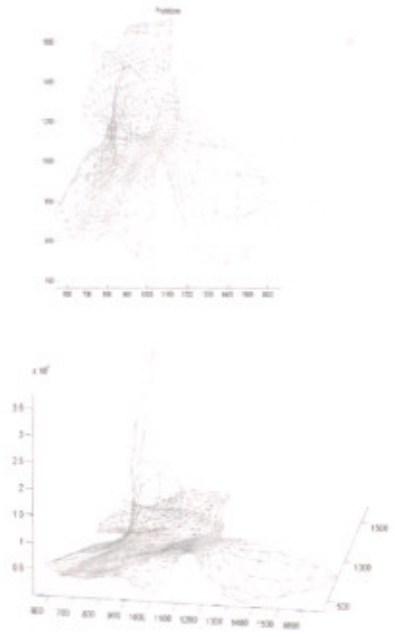

Figura 7. Vectores prototipos obtenidos mediante la aplicación de SOM para visualizar la densidad poblacional.

4.1.2. Desempeño NG

Para el análisis se utilizó la función **neural_gas** del software SOM TOOLBOX para Matlab 5 [4]. La función necesita dos parámetros: *alpha0* y *lambda0* con los significados descritos por la ecuaciones (3) y (4). Cuando el usuario no los indica, estos parámetros (*alpha0* y *lambda0*) son inicializados con sus valores por defecto: *alpha0=0.5* y *lambda0* =n/2. Los valores de $\alpha(k)$ y G(.) de la ecuación (1) declinan en iteraciones sucesivas de acuerdo con las fórmulas (3), (2) y (4). Los valores α_{min} y λ_{min} de las ecuaciones (3) y (4) toman sus valores por defecto: $\alpha_{min}=0.005$ y $\lambda_{min}=0.01$.

En la tabla 2 se muestran los conjuntos de datos elegidos, el número de ejemplos, el número de épocas y el desempeño de la red en cuanto a precisión. Para todos

los conjuntos de ejemplos se tomaron dos números de épocas en relación 1:10. Para el caso de visualización de relieve, debido a que la cantidad de ejemplos es bastante grande, en el primer entrenamiento se tomaron 10 épocas y en el segundo sólo una. Para el conjunto de población, se realizaron dos entrenamientos: el primero con 100 épocas y el segundo con 10.

Como puede verse en la tabla 2, esta relación 1:10 en el número de épocas disminuye el error de cuantización en una relación 2.75:1 aproximadamente. El tiempo de ejecución es directamente proporcional al número de ejemplos por el número de épocas.

En la figura 8 se pueden apreciar los resultados obtenidos al utilizar NG en el primer conjunto de datos para visualización

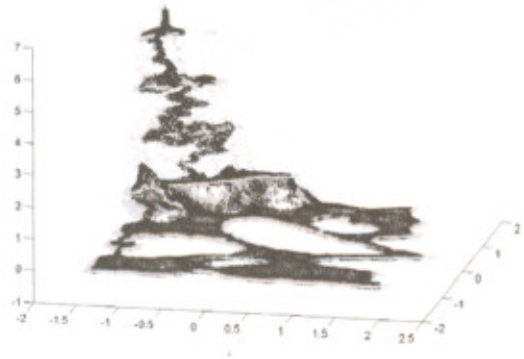

Figura 8. Comparación entre los vectores prototipo (cruces) obtenidos mediante la aplicación de NG a visualización de relieve y los datos originales.

Aquí puede verse con claridad que los vectores prototipo no se distribuyen uniformemente en todo el terreno, y crean *clusters* o agrupamientos que no representan la forma real del relieve.

Los resultados obtenidos para visualización de densidad poblacional se muestra en la figura 9.

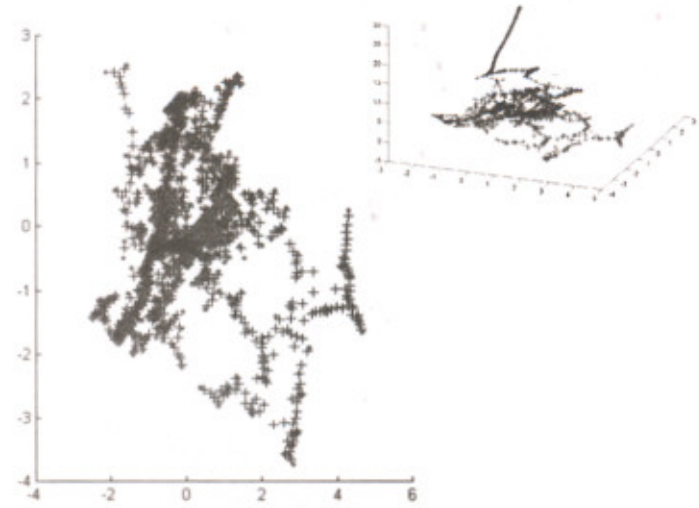

Figura 9. Vectores prototipo obtenidos (cruces) mediante la aplicación de NG para visualización de densidad poblacional.

Aquí puede apreciarse que los vectores prototipo se distribuyen de manera adecuada sobre el espacio geográfico de Colombia.

4.2 Clasificación de datos geoespaciales

A continuación se presenta un análisis de los desempeños obtenidos con SOM y NG para clasificación de información georreferenciada.

4.2.1 Desempeño SOM

En las figuras 10 y 11 se encuentran las matrices de distancias y los vectores prototipo obtenidos en el proceso de clasificación climática. Aquí se observa que los

Tabla 2. Datos correspondientes a la aplicación de NG.

NEURAL GAS	Visualización				Clasificación			
	Relieve		Población		Climas		Departamentos	
Ejemplo	x, y ,z		x, y, población		x, y, clima		x, y, departamento	
Número de ejemplos	23368	1090	1090	1090				
Número de épocas	10	1	100	10	100	10	100	10
Error de cuantización	0.0347	0.0902	0.0420	0.1048	0.0269	0.0740	0.0312	0.0844

vectores prototipo se concentran en las regiones donde los municipios se encuentran más cercanos, mientras que los más alejados son totalmente ignorados por el SOM. Esto se ve reflejado con claridad en los altos porcentajes de error de cuantización generados (10.7 y 22.4%. Véase tabla 1).

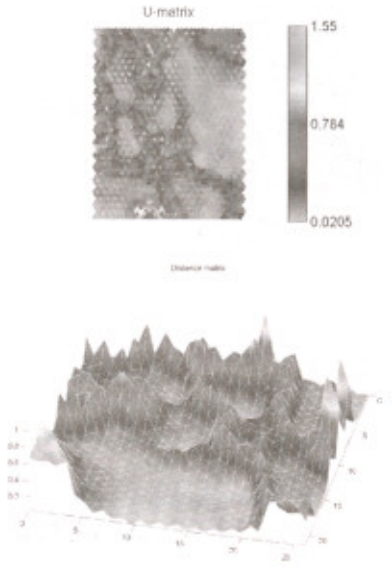

Figura 10. Matrices de distancia obtenidas mediante la aplicación de SOM para clasificación de zonas climáticas.

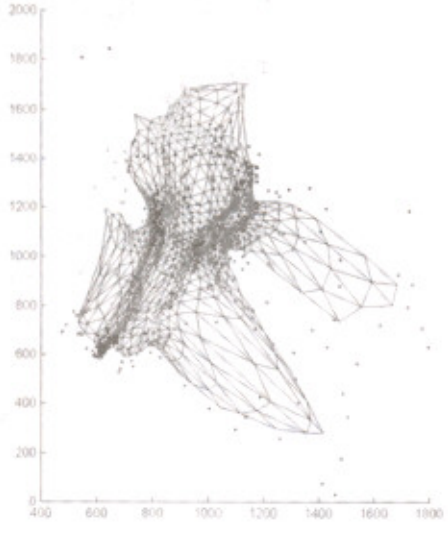

Figura 11. Comparación entre los prototipos (líneas) generados y los datos reales (puntos) para clasificación climática con SOM.

Para la clasificación departamental, figuras 12 y 13, puede verse un comportamiento muy similar al caso anterior. Los vectores prototipo se agrupan en las regiones con más concentración de municipios, y son ignorados en la clasificación los municipios más alejados.

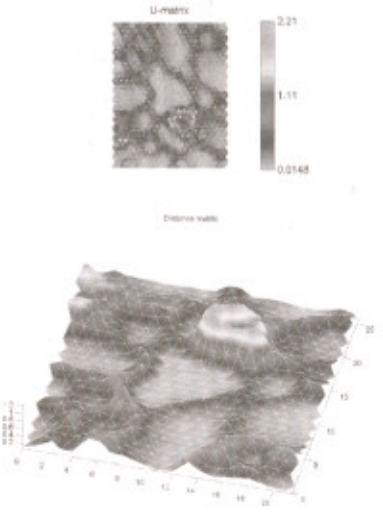

Figura 12. Matrices de distancia obtenidas mediante la aplicación de SOM para análisis de clasificación departamental.

Aquí los errores de cuantización tienden a ser más altos debido a que la disposición geográfica de los departamentos de Colombia es bastante irregular.

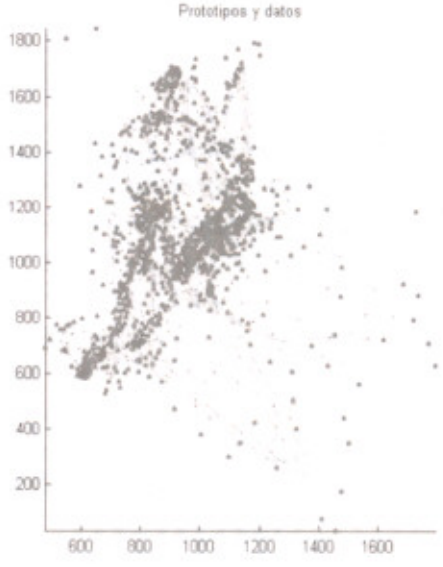

Figura 13. Comparación entre los prototipos generados (líneas) y los datos reales (puntos) para clasificación departamental con SOM.

Es de notar que el mejor desempeño de los SOM se presentó en la visualización de relieve, debido a que la cantidad de ejemplos presentados es mucho más grande y mucho más uniforme, geográficamente hablando. Además, la estructura de malla de los mapas autoorganizativos se acopla bastante bien a la representación gráfica del relieve.

Para los casos de visualización de densidad poblacional y clasificación de climas y departamentos, si se desea reducir el error, es necesario aumentar el tamaño del mapa autoorganizativo con una relación directamente proporcional al tiempo de ejecución.

4.2.2 Desempeño NG

En la tabla 2 se muestran los conjuntos de datos elegidos, el número de ejemplos, el número de épocas y el desempeño de la red en relación a precisión. Para todos los conjuntos de ejemplos se tomaron dos números de épocas en relación 1:10. Para los conjuntos de datos de climas y departamentos se realizaron dos entrenamientos: el primero con 100 épocas y el segundo con 10.

Como puede verse en la tabla 2, la relación 1:10 en el número de épocas disminuye el error de cuantización en una relación 2.75:1 aproximadamente. El tiempo de ejecución es directamente proporcional al número de ejemplos por el número de épocas.

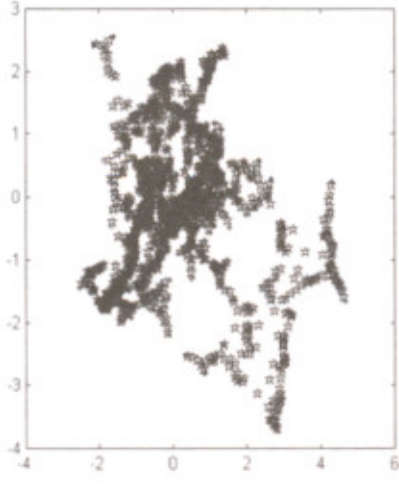

Figura 14. Prototipos obtenidos (estrellas) al aplicar NG para análisis de clasificación climática.

Los resultados obtenidos se muestran en las figuras 14 y 15. Como puede apreciarse, la distribución de los vectores prototipo, con relación a los datos reales, es bastante precisa y abarca todo el espacio geográfico.

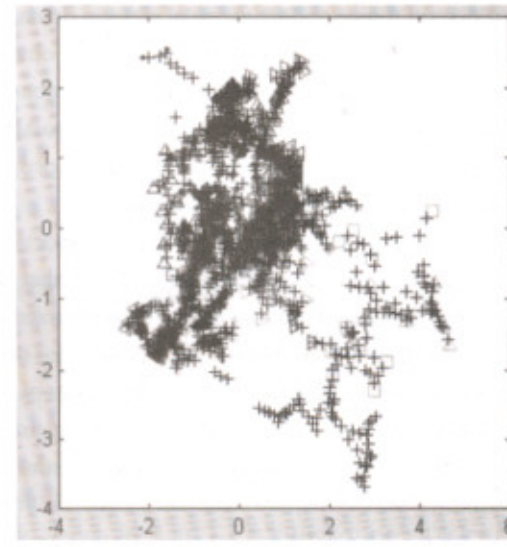

Figura 15. Prototipos obtenidos (cruces) al aplicar NG para análisis de clasificación departamental.

5. Desempeño SOM Vs. NG

En la figura 8 se nota con claridad que el desempeño del NG no es el mejor como método de visualización, ya que la malla generada con los SOM es más ilustrativa (figura 5). Además, el error de cuantización generado tiende a ser más alto que el obtenido mediante SOM.

Para los casos de visualización de densidad poblacional y clasificación climática y departamental (figuras 9, 14 y 15), se puede observar que el comportamiento del NG es mucho más eficiente que el de los SOM. Debido a que las neuronas del método NG viajan libremente en el espacio y a que en cada época aumenta el número de las mismas, ninguno de los municipios fue excluido de la clasificación y el conjunto de vectores prototipo fue distribuido equilibradamente sobre toda la geografía de Colombia, aun cuando ésta es tan irregular. Si recordamos la tabla 2, los errores de cuantización para estos conjuntos de datos son mucho más bajos que los generados mediante SOM.

6. Conclusiones

Los SOM representan un camino para visualizar modelos digitales del terreno, de manera rápida y con una precisión aceptable; sin embargo, no es la herramienta más óptima para realizar clasificaciones de datos geoespaciales. El método NG permite generar vectores prototipo con mejor cobertura en el espacio de datos. Esto se observa en los errores de cuantización obtenidos y en los análisis gráficos realizados. Sin embargo, los SOM tienen una capacidad invaluable para visualizar la topología de los datos en el espacio original, propiedad que no posee el método Gas Neuronal.

Las redes neuronales tienen mucho que aportar al manejo de datos geoespaciales, en el nivel de abstracción y en el de compresión de datos. Esto permite lograr tiempos de ejecución mucho más bajos que los ofrecidos por los métodos tradicionales en los cuales es necesario procesar todo el conjunto masivo de datos. En esta investigación se ha comprobado que constituyen un mecanismo eficiente para visualizar y clasificar información georreferenciada con menores costos en software y hardware.

Referencias

[1] Kohonen, T. Self-Organizing Maps. *Springer Series in Information Sciences*, 30, Berlin 1995.

[2] Martinetz, M.; Berkovich, S.; Schulten, K. 'Neural-gas' network for vector quantization and its application to time series prediction. *IEEE Trans. Neural Networks*, V. 4, 1993, 558–569.

[3] Bartkowiak, A.; Evelpidou, N.; Vassilopoulos, A. Choosing data vectors representing a huge data set: Kohonen's SOMs applied to the Kefallinia erosion data.

[4] Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. *SOM Toolbox for Matlab 5*. Som Toolbox team, Helsinki University of Technology, Finland, Libella Oy, Espoo 2000, 1–54.

[5] Ultsch, A. Self-organizing Neural Networks for Visualization and Classification. In O. Optiz; B. Lausen and R. Klar (Eds.). Information and Classification, Berlin: Springer-Verlag, 1999, 307-313.

[6] Gahegan, M. On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, Vol. 32, No. 2, 2000, 113-119.

[7] Heinke, D.; Hamker, F.H. Comparing neural network benchmarks on growing neural gas, growing cell structure, and fuzzy ARTMAP. *IEEE Trans. on Neural Network*, 1998, pp. 1279–1291.