Representación y Clasificación de Datos Geoespaciales Usando Redes Neuronales

Marly Esther De Moya Amarís
Instituto Geográfico Agustín Codazzi, Subdirección de Geografía y Cartografía
Bogotá, Colombia
me_demoya@igac.gov.co
y
Luis Fernando Niño Vásquez, Ph.D.
Universidad Nacional de Colombia, Laboratorio de Sistemas Inteligentes,
Bogotá, Colombia
lfninov@unal.edu.co

Abstract

Approximately 80% of all the existing information in the world corresponds to geo-referenced information, this creates an increasing necessity to have tools more flexible, precise and easy to use to perform visualization, exploration and classification of great volumes of geospatial data. Additionally it is necessary to achieve shorter times to process this kind of information. In this work, different techniques to visualize and classify geo-referenced data using two types of neuronal networks: Kohonen's maps (SOM) and the Neural Gas method (NG) are compared. In visualization, SOM showed a better performance than NG, while in classification NG performed better than SOM.

Keywords: Neural Networks, Self Organizing Map (SOM), Kohonen, Neural Gas, Geographic Information System (GIS).

Resumen

Aproximadamente el 80% de toda la información existente en el mundo corresponde a información geo-referenciada. Esto crea una creciente necesidad de disponer de herramientas más flexibles, precisas y fáciles de usar para la visualización, exploración y clasificación de grades volúmenes de datos Geoespaciales. Adicionalmente es necesario lograr menores tiempos de procesamiento para este tipo de información. En este trabajo se comparan diferentes técnicas para presentar y clasificar datos geo-referenciados utilizando dos tipos de redes neuronales: mapas auto-organizativos de Kohonen (SOM) y el método gas neuronal (GN). Para los casos de visualización, SOM mostró un mejor desempeño que GN, dándose el caso contrario para los ejemplos de clasificación.

Palabras Clave: Redes Neuronales, Mapas auto-organizativos (SOM), Kohonen, Gas Neuronal, Sistemas de Información Geográfica (GIS)

1 INTRODUCCIÓN

Debido a la creciente cantidad de información georreferenciada, se tienen que almacenar, procesar y analizar volúmenes de información cada día más grandes. Con el fin de manejar esta enorme cantidad de datos complejos y multidimensionales, se ha incrementado la demanda de técnicas más sofisticadas de análisis, incluyendo clasificación y visualización de información geoespacial. Tradicionalmente, los sistemas de información geográfica (GIS) han sido ampliamente utilizados para analizar y visualizar datos geoespaciales. En los últimos años, una nueva generación de sistemas de información geográfica han surgido y han extendido sus funcionalidades, generando mapas dinámicamente, logrando grandes capacidades de análisis exploratorio visual de los datos [1], [3], [5], [9]. A pesar de este gran desarrollo, estos sistemas tienen capacidades limitadas para visualizar la interacción de los atributos en mapas con apenas unas pocas dimensiones, de aquí, que dependencias complejas multivariadas sean fácilmente pasadas por alto, teniendo como consecuencia que mucho conocimiento valioso oculto en los datos probablemente nunca sea descubierto o presente mucha dificultad de ser hallado.

En esta investigación se estudia y se demuestra el aporte que las Redes Neuronales Artificiales pueden proveer en este campo, ya que éstas presentan grandes ventajas que pueden ser aplicadas al estudio de datos geoespaciales, tales como sus capacidades de clasificación, reorganización topológica, reducción de datos y visualización. El estudio se centró específicamente en un tipo de redes neuronales conocido como mapas auto-organizativos, básicamente se trabaja con

dos clases de mapas: los mapas de Kohonen (SOM) [7] y el método gas neuronal (GN) [8]. Estas técnicas presentan grandes ventajas para analizar conjuntos de datos masivamente complejos y relacionados espacialmente. El desarrollo de ambientes de visualización basados en mapas auto-organizativos pretende descubrir estructuras y patrones en complejos conjuntos de datos espaciales y proveer reducción de datos y representaciones gráficas que puedan soportar el procesamiento, el análisis, la comprensión y la construcción de conocimiento. La capacidad de estas redes se mostró usando datos del censo de Colombia del año 1993, utilizando tanto el método SOM como el GN para su clasificación, representación y visualización, se estudió cada método en detalle y después se realizó un análisis comparativo de los resultados obtenidos con cada método. Un estudio de las mismas características se realizó para la visualización de modelos digitales del terreno.

El resto de este artículo está organizado como sigue: en la sección 2 se presenta una descripción de los mapas auto-organizativos de Kohonen y del método Gas Neuronal, para posteriormente, en la sección 3, presentar el marco experimental, describiendo 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, a fin de 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 finales

2 MAPAS AUTO-ORGANIZATIVOS Y GAS NEURONAL

Los mapas auto-organizativos y el método gas neuronal corresponden a ejemplos de redes competitivas, en las cuales se define algún tipo de competición entre unidades con el fin de conseguir que una de ellas quede activada. 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, lo cual genera el refuerzo de dichas conexiones y de sus vecinas, y debilitando el de las demás unidades perdedoras. Al final se consigue que cada unidad responda frente a determinados patrones de entrada, y las neuronas vecinas se activarán de manera que los pesos aferentes de esa unidad converjan en el centro del grupo de patrones con características similares. Para este caso de estudio, se analizarán estos dos tipos de redes neuronales competitivas: los mapas auto-organizativos de Kohonen y el método Gas Neuronal.

2.1 Mapas auto-organizativos de Kohonen

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 una determinada información de entrada a la red. Los pesos de las conexiones se ajustan en función de la neurona de 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. Inicializar 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 números reales.
3. Determinar la neurona vencedora de la capa de salida, la cual será la 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:

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

$e_i^{(k)}$: 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.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)\left[e_i^{(k)} - w_{j^*i}(t)\right] \text{ , donde} \tag{2}$$

$\alpha(t)$: 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, volviendo a presentar todo el juego de patrones de aprendizaje.

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*. Este error corresponde al promedio de la distancia Euclidiana de los vectores de datos a sus representantes más cercanos.
2. *Error Topológico.* Este error indica cual es 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 GN, propuesto por Martinetz [2], se orienta hacia la cuantización del espacio. Este método es más general comparado con los mapas de Kohonen. La mayor diferencia entre los métodos SOM y GN, es que en el modelo GN la colección de neuronas no está 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 se considera una colección de *m* neuronas, cada neurona puede ser imaginada como un punto en el espacio de datos. Las coordenadas de la $i^{ésima}$ neurona será denotada como $w_i = (w_{i1}, ...,w_{ip})$. Al inicio de los cálculos, la colección de neuronas es distribuida aleatoriamente sobre el espacio de datos. En las siguientes iteraciones, las neuronas cambian su posición y se adaptan ellas mismas a la nube de datos. El proceso de adaptación es llamado *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. Asumiendo *h* épocas, el máximo número de iteraciones de ajuste permitidas es K_{max}= n * h. En cada iteración *k* (*k*=0,1,.... K_{max}), se presenta un vector de datos aleatorio **x** al conjunto de neuronas. Para cada vector de datos **x** presentado 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:

$$d(x, w_w) = \min\ d(x, w_i) \quad 1<=i<=m \tag{3}$$

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:

$$w_i = w_i + \alpha(k)G(i,k,x,w,\lambda)(x - w_i) \tag{4}$$

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

$$G(i,k,x,w,\lambda) = \exp\left\{-\frac{d^2(x,w_i)}{2\lambda^2(k)}\right\} \tag{5}$$

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

Analizando la fórmula (4) puede mirarse que los cambios de posición de la *i*-ésima neurona depende de dos factores: α, y el valor de la función G, el cual depende del parámetro λ. Ambos coeficientes α y el parámetro λ depende de *k*, el número de la actual iteración, teniendo el siguiente significado:

α(*k*) : Define el coeficiente de aprendizaje, el cual determina qué tan grande puede ser el cambio de posición. El coeficiente α usualmente decrece con el número de iteraciones: con un valor de inicio α_0, decreciendo gradualmente a un valor final α_{min} alcanzado al final de todas las iteraciones. El valor de decrecimiento de α(*k*) es descrito por la función:

$$\alpha(k) = \alpha_0\left(\frac{\alpha_{min}}{\alpha_0}\right)^{k/k_{max}} \tag{6}$$

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 $k = k_{max}$:

$$\lambda(k) = \lambda_0\left(\frac{\lambda_{min}}{\lambda_0}\right)^{k/k_{max}} \tag{7}$$

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

3 MARCO EXPERIMENTAL

El objetivo general del estudio experimental es realizar un análisis comparativo entre los mapas auto-organizativos de Kohonen (SOM) y el método Gas Neuronal (GN), aplicados a la clasificación y visualización de datos geoespaciales, a fin de evidenciar cual presenta un mejor desempeño en estas tareas. El estudio está dividido en dos tipos de análisis:

- *Clasificación de datos geoespaciales y su correspondiente visualización y representación.* El objetivo de este análisis es realizar un estudio comparativo entre los mapas auto-organizativos de Kohonen y el método gas neuronal, aplicados a la clasificación y representación de la información socioeconómica y demográfica de la población de Colombia, georeferenciada a nivel municipal, correspondiente al censo de 1993. La fuente de datos es el Departamento Nacional de estadística – DANE [2]. Esta muestra incluye los siguientes factores: actividad económica, población por sexo y edad, nivel de educación, ocupación, condición de tenencia de la vivienda, tipo de vivienda, material de las paredes, material del piso, servicios públicos, servicio sanitario, en que lugar cocinan, con qué cocinan. Conformando un total de 118 variables por cada uno de los 1029 municipios.

- *Visualización de datos geoespaciales.* El objetivo es realizar un estudio comparativo entre los mapas auto-organizativos de Kohonen y el método gas neuronal, aplicados a la visualización tridimensional del terreno (generación de modelos digitales del terreno). Para este análisis se seleccionó una muestra de datos tridimensionales de una zona montañosa de Colombia (65535 coordenadas *x, y, z*), véase figura 1. Esta zona corresponde a un área de 38.756 Km^2, con coordenadas: esquina superior derecha: 834.969 mE, 1319476 mN, esquina inferior izquierda: 1'029.969 mE, 1'120.726 mN. La información fue adquirida en el Instituto Geográfico Agustín Codazzi de Colombia. El objetivo es analizar la capacidad de cada uno de los métodos, mapas auto-organizativos de Kohonen y el método gas neuronal, para modelar el terreno, es decir, para presentar visualmente la forma del relieve.

Figura 1: Zona de trabajo

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

3.1 Preprocesamiento de datos

El primer paso es construir los conjuntos de datos y dejarlos en un formato apropiado para su procesamiento. Para facilitar el análisis de las visualizaciones generadas por los mapas auto-organizativos, se dividió el conjunto de datos por departamentos, por tanto se generó un conjunto de datos por cada uno de ellos, con todas las variables generadas en el censo de 1993 para personas, hogares y viviendas (118 variables en total), agregando las correspondientes coordenadas *x* y *y* a cada municipio. Los valores fueron traducidos a porcentajes, ya sea de viviendas, hogares o personas de acuerdo al caso, a excepción de la población total, población de mujeres y población de hombres.

Se dispone de la información de un total de 1029 municipios, en la figura 2 se puede apreciar la distribución geográfica de los mismos.

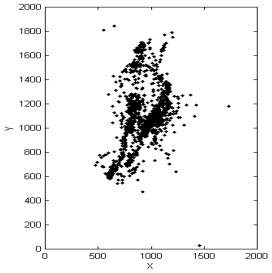

Figura 2. Distribución de municipios

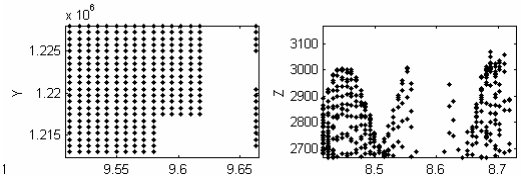

Figura 3. Acercamiento de la malla de puntos

Para el caso del modelo digital del terreno se generaron 16 archivos, 14 con 4096 puntos y 2 con 4095 puntos, con coordenadas *x, y* y *z*, por tanto, en total se dispone de una malla de 65535 puntos. En la figura 3 se muestra un acercamiento al conjunto de puntos.

Una vez los conjuntos de datos han sido generados y almacenados en su correspondientes formato, se procede a la normalización. Debido a que los algoritmos SOM y GN 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 la organización de los resultados completamente. Por esta razón, los componentes de los datos deben ser normalizados. El proceso de normalización llevado a cabo es por rango, es decir, se toma el valor máximo y mínimo de cada conjunto de datos y se realiza una transformación lineal la cual escala los valores entre [0, 1].

4 RESULTADOS EXPERIMENTALES

A continuación se describen los resultados obtenidos con los experimentos. Para analizar el comportamiento del GN y del SOM se aplicaron diferentes parámetros sobre cada conjunto de datos, a fin de evaluar el comportamiento y determinar los valores óptimos de desempeño.

4.1 Clasificación, representación y visualización de los datos del censo

Para realizar el entrenamiento de las redes se seleccionó una muestra de 10 departamentos, de un total de 30, tal como se ilustra en la tabla 4.1. En esta muestra se incluyen departamentos representativos de cada una de las regiones geográficas de Colombia.

REGION	DEPARTAMENTO	CARACTERÍSTICA
NORTE	ATLÁNTICO	Puerto aéreo, marítimo y fluvial
	BOLIVAR	Turismo
	ANTIOQUIA	Mayor número de municipios, comercio, agricultura, industria
PACIFICO	CHOCO	Pobreza
	VALLE	Puerto marítimo, agricultura, industria, comercio
CENTRO	BOYACA	Agricultura, artesanías
	CUNDINAMARCA	Capital de Colombia, agricultura
ZONA CAFETERA	CALDAS	Agricultura, comercio
LLANOS ORIENTALES	META	Agricultura
AMAZONIA	CAQUETA	Zona selvática de Colombia

Tabla 1: Departamentos seleccionados

Cada departamento cuenta con información de las 116 variables mencionadas en la sección 3 y con sus respectivas coordenadas planas *x* y *y,* las cuales permiten ubicar geográficamente cada municipio. Debido a que se dispone de un total de 116 variables por cada uno de los diez departamentos, a fin de facilitar el proceso de análisis, se seleccionó un conjunto de 53 de éstas, con el objeto de realizar un estudio del nivel de socioeconómico de la población. Este análisis pretende descubrir las interrelaciones existentes entre las diferentes variables, tales como su nivel de educación, la actividad económica, la ocupación, la condición de tenencia de vivienda, el material del piso, material de las paredes y los servicios públicos.

Se realizó el entrenamiento de una red neuronal por cada uno de los diez departamentos seleccionados, por los métodos de Kohonen y de gas neuronal. La red recibe como entrada las coordenadas geográficas de cada uno de los municipios que conforman el respectivo departamento y el conjunto de las 53 variables.

4.1.1 Desempeño con SOM

El proceso de entrenamiento de los SOM para los diez departamentos escogidos generó los resultados relacionados en la tabla 2. Aquí puede observarse, el tamaño del SOM seleccionado para cada departamento tiene un número de neuronas mayor que el número de municipios (ejemplos de entrenamiento), por esta razón los errores topográficos y de cuantización son bastante pequeños. Esto puede generar sobreentrenamiento en la red, pero debido a que el objetivo del análisis es utilizar las ventajas del despliegue gráfico de los mapas auto-organizativos de Kohonen, una red grande brinda muchas ventajas para el agrupamiento (clusters) y el análisis visual.

Departamento	No. de Munic.	SOM		GN	
		No. de Neuronas	Error de Cuantización	No. de Neuronas	Error de Cuantización
Antioquia	124	1944	0.0260	1240	0.0071
Atlántico	23	384	0.0020	230	4.8348e-004
Bolívar	32	448	0.0104	320	0.0011
Boyacá	123	2016	0.0152	1230	0.0105
Caldas	25	400	0.0026	250	6.8099e-004
Caquetá	15	180	0.0082	150	1.3502e-004
Chocó	19	216	0.0196	190	3.0326e-004
Cundinamarca	115	1980	0.0115	1150	0.0074
Meta	29	448	0.0066	290	0.0011
Valle	42	560	0.0272	420	0.0019

Tabla 2: Cuadro comparativo entrenamientos SOM y GN – Departamentos

Para un análisis más detallado, se seleccionó el departamento de Caldas, por tener buenas características de agrupamiento, facilitando esto la explicación de los mecanismos de análisis que brindan los SOM. El departamento de Caldas cuenta con 25 municipios distribuidos geográficamente tal y como se muestra en la figura 4, se seleccionó un tamaño para la red neuronal de 20 x 20 neuronas, generando un mapa de 400 neuronas (figura 5).

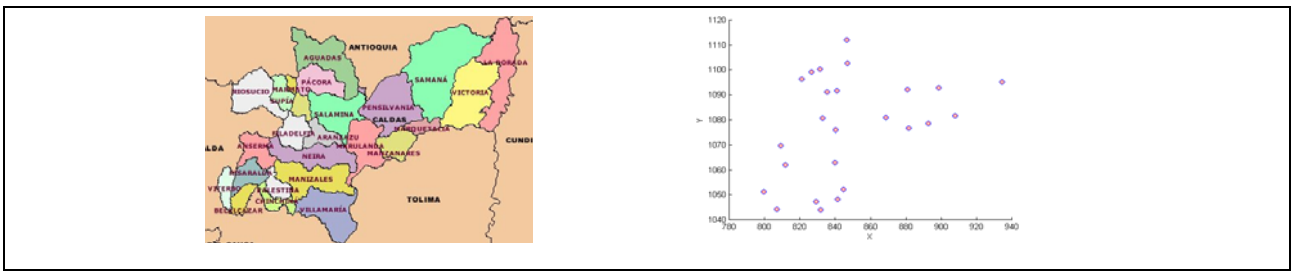

Figura 4: Ubicación geográfica de los municipios del Departamento de Caldas.

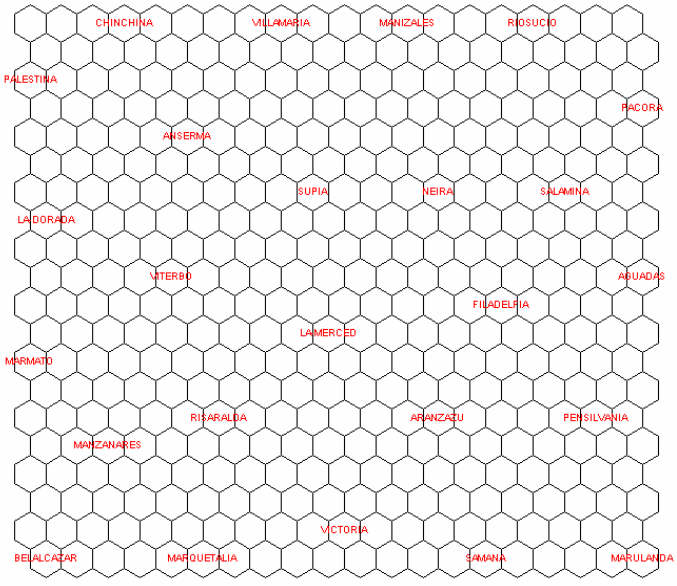

Figura 5: SOM generado para el departamento de Caldas

En la figura 5, se pueden apreciar cada uno de los municipios y demás vectores prototipo creados (neuronas o unidades de mapa), en el espacio de proyección del SOM que corresponde a la malla de la red neuronal. Sin embargo, los datos de entrada y los vectores formados también pueden ser visualizados en el espacio geográfico correspondiente (ver figura 6). Las cruces negras corresponden a cada uno de los vectores prototipo generados durante el entrenamiento y cada uno de los puntos en otros tonos corresponden a los datos de entrenamiento de la red. Esta red generó un error de cuantización de 0.0026 y un error topográfico de 0.

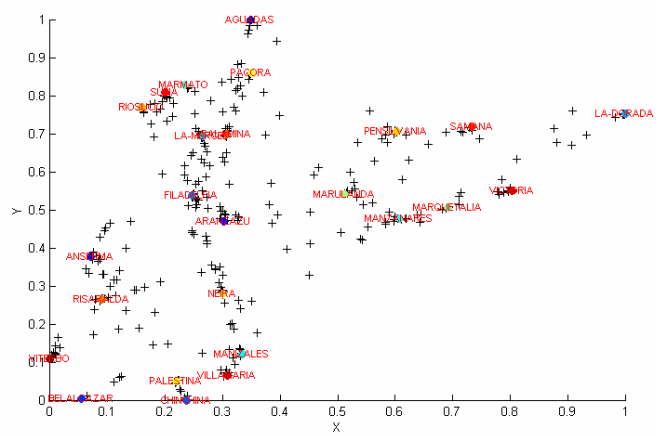

Figura 6: Datos de entrada y vectores prototipo en el espacio geográfico de Caldas

Los SOM ofrecen grandes ventajas de visualización como son las matrices de distancias y las componentes planas. Cada componente plana muestra los valores de una variable en cada unidad de mapa usando la misma codificación de color descrita para la matriz de distancia. Esto da la posibilidad de examinar visualmente cada celda (correspondiente a cada unidad de mapa). En la figura 7 se pueden apreciar las componentes planas y la matriz de distancias de las variables correspondiente al material en que están construidas las casas de Caldas.

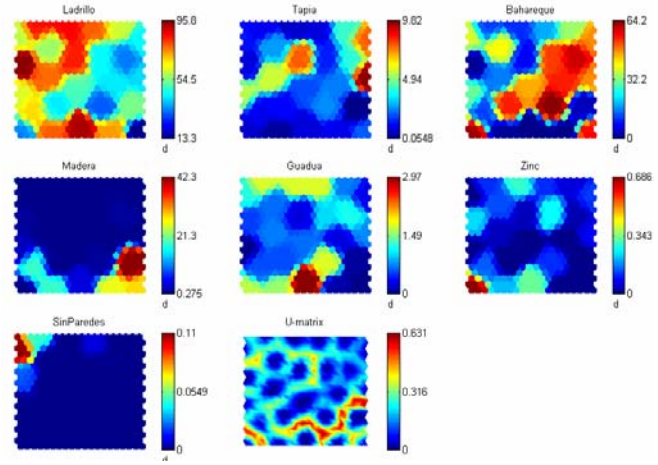

Figura 7: Material de las casas del departamento de Caldas

Aquí puede apreciarse que en los municipios de Manizales, Chinchiná, Villamaría, Supia, Viterbo, La Dorada, Marquetalia, Victoria y Samaná, se visualiza un gran porcentaje de casas construidas en ladrillo, con porcentajes que oscilan alrededor del 90%, aunque un alto porcentaje de población aún tiene casas construidas con bahareque, tal como es el caso de los municipios de Belalcazar, Marulanda, Risaralda, Aranzú, La Merced, Filadelfia, Aguadas, Neira, Salamina y Pácora, en los cuales las casas de bahareque oscilan entre el 50 y el 64%. Si se forma una matriz de distancia (U-matrix en Figura 7) con estas variables, se visualizan básicamente dos grupos, el formado por las casas de ladrillo y el formado por las casas de bahareque y un pequeño tercer grupo formado por las casas de madera (42.3%) del municipio de Pensilvania.

Una pregunta interesante es en qué parte de un SOM se encuentra localizado un dato de entrada específico, y cuanta precisión tiene esa localización. La forma más sencilla de resolver esta pregunta es encontrando la unidad del mapa que más concuerde con el vector del dato que se está buscando (BMU). Como ejemplo se van a localizar en el SOM las BMU de dos municipios, Manizales y Belalcazar (indicadas por etiquetas en el mapa), luego se van a calcular los errores relativos de cuantización, el error de cuantización entre estos ejemplos y todas las unidades del mapa. Esto puede visualizarse en la figura 9.

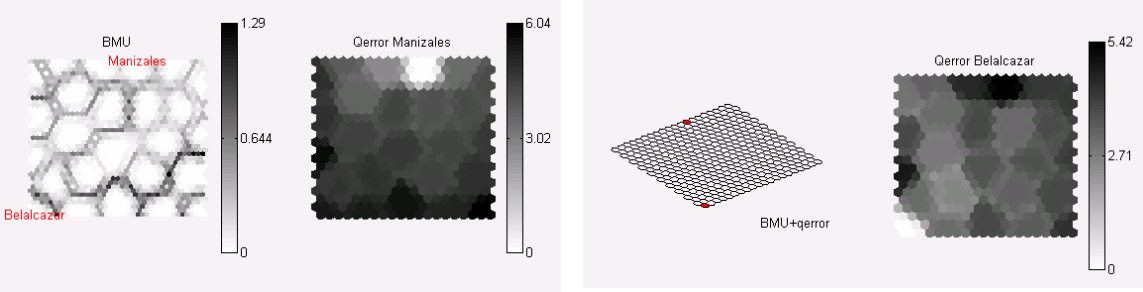

Figura 9: Localización de datos en el SOM y su precisión

El error de cuantización calculado para el municipio de Manizales, es decir la distancia hasta su BMU es de 0.0019, a su segunda BMU es de 0.007 y a su peor unidad WMU es de 6.0446. Los mismos errores se calcularon para el municipio de Belalcazar, obteniéndose 0.0011, 0.0077 y 5.4205 respectivamente.

4.1.2 Desempeño con GN

Los parámetros para el entrenamiento fueron inicializados con los valores: *alpha=0.5* y *lambda* =n/2. Los valores de $\alpha(k)$ y G(.) de la ecuación (1) declinan en iteraciones sucesivas de acuerdo a las fórmulas (3), (2) y (4). Los valores α_{min} y λ_{min} de las ecuaciones (3) y (4) toman sus valores por defecto: α_{min}=0.005 y λ_{min}=0.01.

El entrenamiento se realizó sobre los mismos diez departamentos trabajados con SOM. Los datos de entrenamiento y los errores generados se encuentran relacionados en la tabla 2. El número de neuronas seleccionadas para cada departamento es el número de ejemplos de entrenamiento (número de municipios) multiplicado por un factor de diez. El número de épocas de entrenamiento es de 250. Estos valores fueron escogidos debido a que ofrecen un menor error de cuantización. Debido a que el método Gas Neuronal no trabaja comportamiento topológico, solo se maneja el error de cuantización.

Para un estudio más detallado, al igual que en el caso de SOM, se seleccionó el departamento de Caldas, a fin de poder comparar los dos métodos para el mismo conjunto de datos. La red generada para el método GN consta de 250 neuronas, a diferencia del SOM, esta red no está conformada por filas y columnas, ni tiene una topología específica, simplemente es una red constituida por 250 neuronas que pueden viajar libremente por el espacio de datos. Cuando la red es entrenada, se generan una serie de vectores prototipo, los cuales son representantes de los datos reales, es decir, los datos de entrenamiento (los 25 municipios). En la figura 10 se pueden apreciar cada uno de los municipios y demás vectores prototipo creados (neuronas). Esta visualización está referenciada en el espacio libre de los datos, es decir en el espacio geográfico, esto debido a que el método Gas Neuronal trabaja sobre el espacio de los datos de entrenamiento.

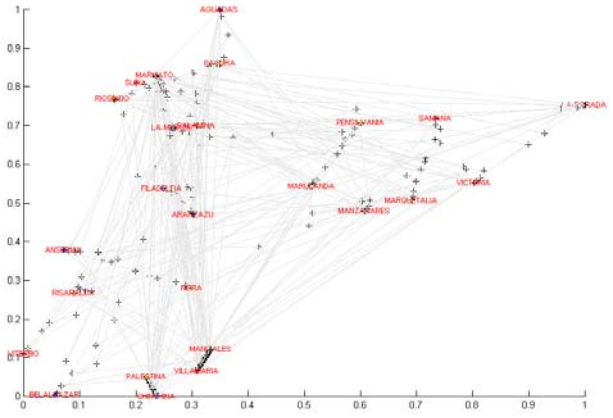

Figura 10: Datos y vectores prototipo en el espacio geográfico de Caldas

Se obtuvo un error de cuantización de 6.8099e-004. En la figura 10 puede observarse que los vectores prototipo tienden a agruparse en los sectores donde están ubicados los datos de entrenamiento y las áreas donde no existen municipios no se crearon neuronas, o se crearon muy pocas.

En un método como el GN que no tiene herramientas de visualización gráfica, las proyecciones constituyen una herramienta de gran utilidad para los procesos de análisis. A fin de facilitar el análisis visual se generó la proyección PCA para el conjunto de vectores prototipo creados durante el proceso de entrenamiento, en la figura 11 puede apreciarse la distribución de los vectores y las etiquetas indicando el nombre del municipio. Al igual que en el estudio realizado con SOM se tomaron específicamente las variables correspondientes al material de la casas. Se pueden distinguir claramente dos grupos, el primero abarcando los municipios de Pensilvania, Samaná, Manzanares, La Dorada, Marquetalia, Manizales, Chinchiná, Villamaría y un poco más alejado el municipio de Victoria. El segundo grupo puede apreciarse en la parte izquierda conformado por los municipios de Marulanda, Aguadas, Pácora, Filadelfia, Risaralda, Supia, Marmato, Viterbo, Riosucio, La Merced, Salamina, Anserma, Aranzazu, Palestina, Neira y en la parte inferior el municipio de Belalcazar. Estos dos grupos corresponden básicamente el grupo de municipios donde predomina el ladrillo y al grupo donde predomina el bahareque respectivamente, llegando de esta forma a la misma conclusión con los análisis realizados con SOM.

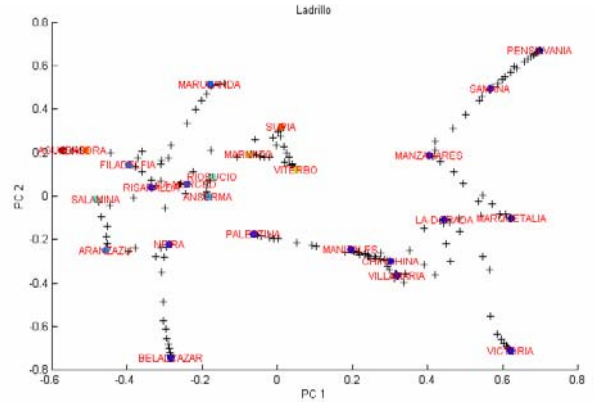

Figura 11: Proyección PCA GN – Material paredes – Caldas

4.2 VISUALIZACIÓN DE MODELOS DIGITALES DEL TERRENO

El objetivo del estudio de visualización de modelos digitales del terreno, es comparar el desempeño que ofrecen los mapas auto-organizativos de Kohonen y el método gas neuronal para representar eficientemente la forma del terreno. Como ya se explicó en la sección de descripción de los datos, se dispone de 65534 puntos, los cuales contienen coordenadas *x, y* y *z*. La red neuronal se alimentará con estos tres datos y su función será reorganizarse topológicamente y generar un modelo digital del terreno correspondiente a los puntos que ha recibo como entrenamiento. A fin de facilitar el análisis y el comportamiento de la red, se dividió el conjunto de datos en 16 regiones, 14 de ellas con 4096 puntos y 2 con 4095. Debido a que el terreno es bastante montañoso, esta división permite un análisis más detallado de cada una de las regiones.

4.2.1 Desempeño con SOM

Para el estudio del modelamiento del terreno por medio de SOM, se trabajó con 16 regiones, generando los errores de cuantización y topológicos detallados en la tabla 3. El número de puntos de cada región es de 4096 con excepción de las regiones 1-4 y 2-4 que tienen 4095 puntos. Para realizar un estudio en detalle, se seleccionó una de las 16 regiones (1-4), la cual es representativa, presentando variaciones en el detalle del relieve y diferentes alturas. En la figura 12 se muestran los vectores prototipo del área seleccionada (malla) vs. los datos reales (puntos).

Región	SOM		GN
	No. de Neuronas	Error de Cuantización	Error de Cuantización
1-1	1312	0.0230	0.0148
1-2	1302	0.0220	0.0145
1-3	1305	0.0195	0.0135
1-4	1312	0.0240	0.0147
2-1	1312	0.0190	0.0134
2-2	1290	0.0176	0.0129
2-3	1287	0.0201	0.0137

2-4	1302	0.0201	0.0135
3-1	1288	0.0163	0.0124
3-2	1290	0.0164	0.0121
3-3	1290	0.0155	0.0118
3-4	1312	0.0150	0.0109
4-1	1290	0.0211	0.0141
4-2	1312	0.0159	0.0121
4-3	1302	0.0153	0.0115
4-4	1032	0.0147	0.0114

Tabla 3: Cuadro comparativo entrenamientos SOM y GN - Modelamiento de terreno

Esta región tiene alturas entre los 185 y los 2600 metros, coordenadas mínimas 835.720 mE y 1.269.977 mN y coordenadas máximas 881.470 mE y 1319477 mN, conformando una región de 45.750 metros por 49.500 metros, es decir 2.264.625.000 m^2, alrededor de 2.265 Km^2. Los datos de entrada corresponden a 4095 puntos equidistantes entre sí, con una separación entre ellos de 750 metros tanto en la coordenada X como en la coordenada Y.

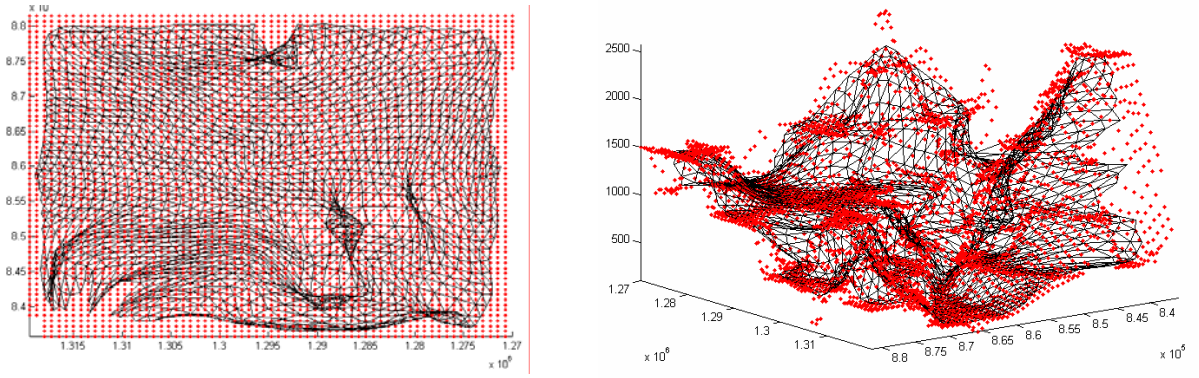

Figura 13: Vectores prototipo Vs. datos – Modelamiento de terreno con SOM

El entrenamiento del SOM permite realizar una gran reducción de la cantidad de datos a ser procesados, de los 4095 datos de entrada, se redujo a 1312 (número de neuronas), es decir, aproximadamente el 68%, esta reducción de datos generó un error de cuantización de 0.0240 y un error topográfico de 0.0256, es decir, se tiene un modelo digital del terreno con un error aproximado de 2.5%. En la figura 13 puede apreciarse la diferencia entre los datos de entrada (puntos) y los vectores prototipos creados (malla).

Como puede observarse, la malla de puntos se ajusta a los datos de entrada, aunque queda un pequeño perímetro de puntos en la parte externa de la región sin representación de vectores prototipo, a fin de obtener un mayor cubrimiento puede elegirse un mapa con mayor número de neuronas. Seleccionando un SOM de 56x44 (2464) neuronas se obtiene un error topográfico de 0.019 y un error de cuantización de 0.027. Se redujo el error de cuantización del espacio al 1.9% y se obtuvo un SOM con mayor cobertura. Comparando la tabla de errores y los SOM generados, podemos verificar que entre más homogéneo es el relieve, el entrenamiento del mapa genera errores de cuantización y topológicos más pequeños, por tanto en zonas con mayor variación de alturas y diversos tipos de relieve es aconsejable utilizar mapas con mayor número de neuronas a fin de obtener mayor precisión a la forma real del terreno y mayor cobertura al espacio de datos.

Puede concluirse que los SOM se convierten en una herramienta efectiva para generar modelos digitales del terreno, la estructura de malla propia de los SOM permite una visualización adecuada de la superficie del terreno, adicionalmente brinda la gran ventaja de reducir la cantidad de datos, modelos de relieve complicado como el usado en el análisis, permitió la reducción de datos en un 68%, con un error del 2.5% y usando más neuronas, se logró una reducción de datos en un 50% con un error del 1.9%, aumentando el tamaño del mapa se puede reducir aún más el porcentaje de error. En modelos más sencillos se puede disminuir aún más la cantidad de datos sin perder el grado de precisión, lo cual permite disminuir los tiempos de procesamiento y el espacio de almacenamiento.

4.2.2 Desempeño con GN

Para el análisis del desempeño del método GN, se trabajó con las mismas 16 regiones entrenadas con SOM, esto a fin de tener parámetros de comparación entre los dos métodos. El número de neuronas escogido para cada región (1300) es en

promedio el mismo utilizado en el entrenamiento con SOM, esto con el objetivo de analizar el comportamiento del GN con la misma cantidad de neuronas. Los errores de cuantización generados se detallan en la tabla 3.

Para el análisis detallado se seleccionó la misma región seleccionada para el estudio de SOM (1-4), esto a fin de tener parámetros de comparación entre el desempeño presentado por los dos métodos. En la figura 14 se muestran los vectores prototipo generados durante el entrenamiento del GN para el área escogida (puntos oscuros). El entrenamiento de esta región con el método GN, permitió la reducción de los datos de entrada en un 69% generando un error de cuantización del 1.4%. Para una reducción de datos tan alto, este porcentaje de error es bastante bajo. Si se desea disminuir el error obtenido pueden agregarse más neuronas a la red. Analizando la representación del relieve obtenida en la figura 14, puede observase que la adaptación al terreno del GN es muy precisa, esto debido a que este método permite que las neuronas viajen libremente por el espacio geográfico.

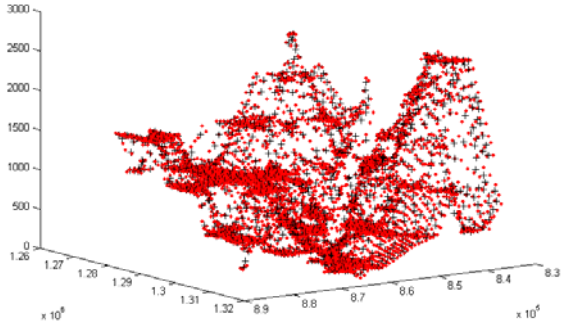

Figura 14: Vectores prototipo para modelamiento de terreno con GN

5 SOM VS. GN

Observando las tablas 2 y 3 puede observarse que el método GN con menor número de neuronas genera errores de cuantización mucho más bajos, esto es debido a que el método GN permite que sus neuronas viajen libremente por el espacio de los datos, esto facilita que las neuronas se distribuyan más eficientemente y con mayor precisión. Los SOM, por el contrario, tiene sus neuronas encasilladas en una malla y las neuronas, a pesar de que pueden moverse por el espacio de datos, tienen que respetar la malla a la que pertenecen, quitándoles esto libertad de movimiento. La gran ventaja de la malla, es que le permite a los SOM tener preservación topológica, los SOM permiten la visualización de cada una de sus neuronas y sus neuronas vecinas y respetan esta topografía, por ello los SOM manejan también el error topográfico. El método GN aunque maneja el concepto de vecindad, solo permite el manejo del error de cuantización.

De igual manera, se pueden ver las grandes ventajas que ofrece la malla de los SOM, permite diagramar las matrices de distancias y las componentes planas y facilitan la visualización de grupos y las correlaciones que existen entre las diferentes componentes planas. Los SOM ofrecen ventajas de visualización que el método GN no permite. Sin embargo, tanto los SOM como el método GN pueden combinarse con otros métodos de proyección a fin de contar con más herramientas visuales para corroborar hipótesis.

6 CONCLUSIONES

Esta investigación presentó las características de los mapas autoo-organizativos de Kohonen (SOM) y del método gas neuronal (GN) para la representación, clasificación y el análisis de datos geoespaciales. El objetivo principal fue explorar las herramientas que ofrecen estos métodos para soportar la extracción de información en grandes conjuntos de datos georreferenciados y la construcción de conocimiento a través de representaciones visuales. Se realizaron experimentos que mostraron el potencial de este tipo de redes neuronales. Del entrenamiento generado para visualización y clasificación de los datos del censo puede observarse, que el método GN con menor número de neuronas genera errores de cuantización mucho más bajos, esto es debido a que el método GN permite que sus neuronas viajen libremente por el espacio de los datos, esto facilita que las neuronas se distribuyan más eficientemente y con mayor precisión. Los SOM, por el contrario, tiene sus neuronas encasilladas en una grilla y las neuronas, a pesar de que pueden moverse por el espacio de datos, tienen que respetar la grilla a la que pertenecen, quitándoles esto libertad de movimiento. Sin embargo, la grilla ofrece grandes ventajas, permite diagramar las matrices de distancias y las

componentes planas y facilita la visualización de clusters y las correlaciones que existen entre las diferentes componentes planas.

Este trabajo mostró que las redes neuronales constituyen una herramienta de gran ayuda para el análisis, representación y visualización de datos geoespaciales, por tanto un trabajo futuro importante sería desarrollar estas aplicaciones para que puedan ser utilizadas desde los software de Sistemas de Información Geográfica existentes y así ayudar a los procesos de análisis y de visualización tradicionales.

Existen diversas aplicaciones en las cuales las redes neuronales y más específicamente los mapas auto-organizativos pueden ser de gran utilidad. Una aplicación interesante es agregar la variable tiempo a los análisis de datos georreferenciados, es decir, realizar análisis espacio-temporales de datos geoespaciales. Los mapas auto-organizativos ofrecen la posibilidad de representar los cambios ocurridos a través de tiempo mediante el trazo de trayectorias. Esto representaría una valiosa herramienta tanto para el estudio de datos geoestadísticos (como en el caso de los datos del censo), como para el estudio de modelos digitales del terreno, ya que se podrían analizar los cambios geográficos sufridos en una determinada zona a través del tiempo. De igual forma, el estudio también puede ser ampliado a información tipo *raster*, tales como imágenes de satélites y fotografías aéreas, para hacer tareas como clasificación, compresión de imágenes, procesos geoestadísticos, entre otros. Otro trabajo futuro importante sería explorar otros métodos de redes neuronales como soporte a diversas aplicaciones SIG.

Referencias

[1] Andrienko, G., Andrienko, N.. "Interactive Maps for Visual Data Exploration", *International Journal of Geographical Information Science* 13(5), 355-374, 1999

[2] Departamento Nacional de Estadística, DANE. Http://www.dane.gov.co/

[3] Dykes, J., "Exploring spatial data exploration with dynamic graphics", *Computers and Geosciences*, 23, 345-370, 1997

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

[5] Gitis V., Dovgyallo A., Osher B., Gergely T., "GeoNet: an information technology for WWW on-line intelligent Geodata analysis", *Abstracts of 4th EC-GIS Workshop*, Hungary, 1998

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

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

[8] 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, 558–569, 1993

[9] Openshaw, S., Turton, I., Macgill, J. and Davy, J., "Putting the Geographical Analysis Machine on the Internet", in Gittings, B. (ed.) *Innovations in GIS 6,* Taylor and Francis, London, 1999

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

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