DOS SOLUCIONES EMPLEANDO ESTRATEGIA MIXTA DE ESCALAMIENTO ULTIDIMENSIONAL Y REDES NEURONALES DE KOHONEN

E. Miret[1]
Dpto. Ecuaciones Diferenciales, Facultad de Matemática y Computación, Universidad de La Habana, Cuba
F. García-Lagos y G. Joya
Dpto. Tecnología Electrónica, ETSI Telecomunicación, Universidad de Málaga, España

RESUMEN
En este trabajo se realiza un estudio comparativo de dos técnicas de visualización de Datos Multidimensionales: el Escalamiento Multidimensional [1] y los Mapas Auto-Organizativos de Kohonen [3]. Se analiza la posibilidad de empleo de una técnica mixta que permite optimizar la solución de un problema concreto aprovechando los aspectos positivos de ambos métodos. Por último, se verifican en un ejemplo real las posibilidades y desventajas de una y otra técnica aplicando la técnica mixta para explicar las soluciones.

ABSTRACT
This paper makes a comparative study of two techniques about visualizing multidimensional data: Multidimensional Scaling [1] and Self-Organizing Maps [3]).The possibility of employing a combined technique is analyzed aiming to optimize a concrete problem. At last, the possibilities and disadvantages of both techniques are verified in a real case applying the combined technique to explain the solutions.

Key words: Multidimensional Scaling, dissimilarities, Strain, STRESS.

MSC: 90B80

1. INTRODUCCIÓN

Existen variados métodos con el fin de obtener una representación plana o espacial de las posiciones de cierto conjunto de objetos en estudio. Dos de estas técnicas son el Escalamiento Multidimensional (MDS, del inglés, *Multidimensional Scaling*) y los Mapas Auto-Organizativos de Kohonen (SOM, del inglés, *Self-Organizing Map*).

Aunque ambas técnicas persiguen un objetivo común: la proyección en un espacio de baja dimensión de un conjunto de datos de alta dimensión preservando las relaciones topológicas originales, las estrategias seguidas por cada método, así como el tratamiento de la información, son muy diferentes. Así, podemos decir que los SOM son especialmente útiles para el tratamiento de vectores de datos de componentes reales, presentando mayores problemas de aplicación cuando se dispone de vectores de componentes cualitativas. Por otra parte, presentan un algoritmo simple en el que la carga computacional depende más de la dimensión de los vectores que del número de estos. Por su parte, el Escalamiento Multidimensional (EM) resulta especialmente eficiente para el tratamiento de variables cualitativas o donde no se dispone de una descripción vectorial de los elementos a clasificar, sino de información sobre ciertas diferencias o semejanzas entre ellos. La carga computacional del método es más dependiente del número de objetos estudiados que de la dimensión de estos.

Del análisis anterior, podría pensarse en tratar los problemas, que por la información de partida resultan difícilmente abordables mediante Mapas Auto-Organizativos. Por ejemplo, en el Problema de Viajante de Comercio con frecuencia se dispone de las distancias por carretera entre las ciudades, estas distancias son, obviamente, no euclídeas. Como a una distancia no euclídea se le puede asociar una distancia euclídea que sea lo más parecida a la primera mediante el EM, en una primera etapa de la búsqueda de la solución de este problema se puede aplicar EM, con el objetivo de generar un vector de componentes reales que conserve las relaciones de proximidad entre los vectores originales. Para estos efectos, es necesario elegir una función adecuada para medir las proximidades entre todos los pares de objetos en estudio que se almacenan en una matriz de disimilitudes, la cual resulta el punto de partida de dicho procedimiento.

E-mail: [1]elina@matcom.uh.cu, elinamiret@yahoo.com

A la salida, se obtiene una matriz de nxk ($2 \leq k \leq p < n$). Estos nuevos vectores de k componentes que identifican a cada objeto, permitirán una correcta y eficiente aplicación de las técnicas de Kohonen.

El empleo del EM en la construcción de mapas de ciudades, conocidas las distancias por carretera entre las mismas, es una de las aplicaciones más conocidas de este tipo de técnica del Análisis Exploratorio de Datos. Mardia **et al**. (1979) a partir de las distancias por carretera entre 12 ciudades británicas reconstruye el mapa de Gran Bretaña. También Borg & Groenen (1997) construyen el mapa de las capitales de 10 países de Europa.

El problema del Viajante de Comercio se resuelve satisfactoriamente mediante el empleo de los SOM, pero dicha técnica requiere del conocimiento de las coordenadas en el plano de las ciudades a recorrer por el viajante que no siempre son conocidas.

Nuestro objetivo es involucrar el EM y los SOM para dar solución a la problemática anterior.

2. EL ESCALAMIENTO MULTIDIMENSIONAL Y LOS MAPAS AUTO-ORGANIZATIVOS DE KOHONEN (SOM)

2.1. Acerca del Escalamiento Multidimensional

Teniendo en cuenta que bajo el nombre de Escalamiento Multidimensional se agrupa un número elevado de procedimientos conocidos, sólo abordaremos algunos aspectos de dos técnicas del EM Métrico, a saber: el Escalamiento Clásico y el EM Absoluto [Borg, I. and Groenen, P. (1997)]. Estos dos métodos, si bien, como parte del Escalamiento Métrico (Miret, E. & Linares, G. (2002), [Miret, E., Mederos, M. V. y Linares, G. (2004)) transforman paramétricamente las disimilitudes iniciales para construir la función que mide el ajuste y cuyo óptimo resulta la matriz de configuración de nxk, donde k resulta el número de coordenadas de los n objetos en un subespacio euclídeo de dimensión k = 2 ó 3, existen diferencias notables entre ambas técnicas atendiendo a la función de ajuste usada y al algoritmo utilizado para obtener la solución en cada caso. Así, el Escalamiento Clásico trabaja con la función de ajuste STRAIN, cuyo óptimo se construye mediante un procedimiento algebraico y el EM Absoluto emplea otras funciones como el STRESS y el SSTRESS [Tarazaga & Trosset (1998)] y emplea procedimientos de la optimización numérica para construir sus soluciones.

En ambos casos debe disponerse de una matriz $\Delta = (\delta_{ij})_n$ de disimilitudes, todo lo cual presupone un análisis de las características de los datos de que se dispone. Particularmente, si se tienen datos cualitativos, es necesario proceder eligiendo acertadamente una función de similitud para este tipo de datos (Zhang, B. & Srihary, S. N. (2003)) y luego convertir las similitudes en disimilitudes que se almacenarán en una matriz al iniciar estos procedimientos.

El *Escalamiento Clásico* (Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). Miret, E. & Linares, G. (2002), [Miret, E., Mederos, M. V. & Linares, G. (2004)) puede resumirse en los siguientes pasos:

El procedimiento más utilizado para encontrar la solución $Y_{(k)}$ consiste en:

1. A partir de $\Delta = (\delta_{ij})_n$, construir $A = (a_{ij})_n$, siendo $a_{ij} = -(1\backslash 2)\, d_{ij}^2$.

2. Obtener $B = (b_{ij})_n$, a partir de A mediante la expresión: $B = HAH$, donde $H = I_n - (1\backslash n)\, l_n l_n^t$, siendo l_n el vector columna formado por n unos e I_n la matriz identidad de orden n.

3. Extraer los k valores singulares estrictamente positivos y más grandes $\lambda_1 \geq ... \geq \lambda_k$ de la descomposición espectral de $B = V\Sigma V^t$ y sus correspondientes vectores singulares normalizados. Si se denota por $V_{(k)} = (V_1,...,V_k)$ a la matriz de las k primeras columnas de V y $\Sigma_k^{1/2} = \mathrm{diag}(\sqrt{\lambda_1},...,\lambda_k)$, entonces:

$$Y_{(k)} = B_{(k)}^{1/2} = V_{(k)}\, \Sigma_k^{1/2} = (Y^{(1)},...,Y^{(n)})^t.$$

La solución algebraica ofrecida por estos pasos, es el óptimo de la función STRAIN de Carroll & Chang (1972) [Tarazaga & Trosset (1998)]. El STRAIN puede expresarse de la forma siguiente:

$$STRAIN = \left\| B - B_{(k)} \right\|^2 = tr(B - B_{(k)})^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

Donde B y $B_{(k)}$ son las matrices definidas en los pasos antes descritos y $\sigma_{k+1} \geq ... \geq \sigma_n$ son los n-k valores singulares de B más pequeños.

El EM Absoluto [Tarazaga & Trosset (1998)] es un procedimiento métrico que suele tomar como función de ajuste el STRESS que está dado por:

$$S = STRESS = 2 \sum_{i<j} W_{ij} [d_{ij}(X) - d_{ij}]^2$$

El *EM Absoluto* es un caso particular del *Interval MDS o Escalamiento Lineal* [Borg, I. and Groenen, P. (1997)], que se resume en los siguientes pasos:

Dada la matriz de disimilitudes $\Delta = (\delta_{ij})_n$ y una matriz de configuración inicial X_0 (obtenida de aplicar Escalamiento Clásico a Δ o generada aleatoriamente).

- Calcular la función $\hat{\delta}_{ij} = f(\delta_{ij}) = a + b\delta_{ij}$, siendo a y b los estimadores habituales de la regresión .

- Minimizar S respecto a X, empleando optimización numérica y tomando como punto de inicio del algoritmo de búsqueda del óptimo a $X = X_0$.

Sea S_1 la evaluación de la función STRESS en la nueva configuración $X = X_1$, entonces:

Analizar ciertas condiciones de S_1 (alcanzado en $X = X_1$), con respecto a una constante prefijada α:

Si $S_1 < \alpha$, entonces la configuración final es: $X_f = X_1$.

Si $S_1 \geq \alpha$, se pasa al próximo paso.

(2) Actualización de X (X = X1).

Se minimiza S respecto a X, obteniéndose S_2 y la configuración X_2.

Si $S_1 - S_2 < \alpha$, entonces la configuración final es: $X_f = X_2$.

Si $S_1 - S_2 \geq \alpha$, se pasa al próximo paso.

(k) Actualización de X ($X = X_{k-1}$).

Se minimiza S respecto a X, obteniéndose S_k y la configuración X_k.

Si $S_{k-1} - S_k < \alpha$, entonces la configuración final es: $X_f = X_k$.

Si $S_{k-1} - S_k \geq \alpha$, se pasa al próximo paso, hasta una cantidad determinada de pasos.

El *Escalamiento Lineal* se llama Escalamiento de Razón (*Ratio MDS*) si la función de las disimilitudes es de la forma $f(\delta_{ij}) = b\delta_{ij}$ y el caso más simple es el EM Absoluto, cuando $f(\delta_{ij}) = \delta_{ij}$.

2.2. Acerca de los Mapas Auto-Organizativos de Kohonen

Dentro de las Redes Neuronales Artificiales, los Mapas Auto-Organizativos de Kohonen (SOM del inglés *Self Organizing Maps*, (Haykin, S. (1994), Martín del Brío, B. & Sanz Molina, A) se caracterizan por su capacidad para clasificar un conjunto complejo de patrones de manera no supervisada, extrayendo criterios de clasificación no obvios ni expresados de manera explícita. Esta clasificación se lleva a cabo mediante la distribución de un espacio de entrada VI de alta dimensión en un espacio de salida VO de menor dimensión, preservándose las relaciones topológicas existentes entre los patrones de entrada. El espacio de salida

viene constituido por un conjunto de neuronas ordenadas en un plano o una línea, en el cual se define una función de vecindad. Cada una de estas neuronas tiene asignado un vector de pesos de igual dimensión que los vectores del espacio original.

Tras el proceso de aprendizaje, un vector de entrada $\vec{x} = (x_1,...,x_n)$ activará la neurona i del espacio de salida cuyo vector de pesos $\vec{w}_i = (w_{i1},...,w_{in})$ tenga menor distancia al vector $\vec{x}$. De esta manera, el vector $\vec{w}_i$ podrá ser considerado como el prototipo de la región del espacio de entrada cuyos vectores activan a la neurona i. Finalmente, dos vectores de entrada similares según la relación definida en VI, activarán la misma neurona o dos neuronas cercanas en el espacio de salida.

En este trabajo nosotros aplicamos una SOM unidimensional de 2n neuronas donde n es el número de ciudades. La regla de adaptación de pesos utilizada aparece en la ecuación:

$$\vec{w}_i(t+1) = \vec{w}_i(t) + l_r h_v(\vec{x} - \vec{w}_i(t)) \tag{1}$$

donde $\vec{x}$ es un vector de entrada, $\vec{w}_i$ es el vector de pesos de la neurona i, y h_v, la función de vecindad, determina el incremento del peso de cada neurona como una función de su proximidad a la neurona ganadora para el patrón $\vec{x}$. En nuestro caso, el área de vecindad viene dada por un cuadrado centrado en la neurona ganadora, cuyo lado decrece hasta cero a lo largo del entrenamiento. l_r es la razón dinámica de aprendizaje, la cual evoluciona a lo largo del proceso de entrenamiento según la ecuación:

$$l_r(t) = \frac{l_{r0}}{(1 + \frac{c \cdot t}{nn})} \tag{2}$$

siendo l_{r0} la razón de aprendizaje inicial (0.1, en nuestro caso), c una constante (0.2), t la iteración actual y nn el número de neuronas de la red.

El paradigma SOM se ha mostrado cómo una potente herramienta en el análisis exploratorio de datos, tanto por su capacidad de clasificación no supervisada como por su capacidad para sacar a la luz criterios de clasificación y relaciones desconocidas entre los datos. Más recientemente, se ha resaltado su utilidad como predictor de la evolución de un determinado sistema. Su ventaja principal es que la información suministrada es de carácter visual y puede ser procesada de directamente cualquier observador. Desde un punto de vista del interés social, su aplicación en el área de la socio-economía (p.ej., análisis de la distribución de la población en categorías laborales en función de otros factores como nivel cultural, estatus social, etc.) o en el área de la salud (p.ej., análisis de la distribución de personas afectadas por una determinada enfermedad en función de factores como sexo, estado civil, orientación sexual, edad, color de piel, etc.) este paradigma puede suministrar una información en algunos casos revolucionaria.

2.2. El problema del Viajante de Comercio (Travelling Saleman Problem)

Este problema consiste en encontrar el orden en el cual un viajante de comercio debería visitar n ciudades sólo una vez, de modo que la distancia recorrida sea mínima (Haykin, S. (1994), Martín del Brío, B. & Sanz Molina, A. (2001)).

Una consideración a tener en cuenta es que el viajante debe regresar a la ciudad de inicio.

Los SOM resuelven satisfactoriamente este problema siempre que se disponga de las coordenadas de las n ciudades, pero suele ocurrir que en la práctica no se disponga de las mismas, debido a que sólo se conocen las distancias por carretera entre las ciudades que, además, no son siempre distancias euclidianas debido a las deformaciones a causa del relieve (ríos, lagos, montañas, etc). Para estos casos no se puede aplicar dicho método.

2.3. Colaboración entre Escalamiento Multidimensional y SOM

Atendiendo a las posibilidades de las técnicas de EM y los SOM como procedimientos de clasificación de conjuntos de datos multidimensionales, pueden considerarse estrategias combinadas para resolver el Problema del Viajante de Comercio cuando no se disponga de las coordenadas de las ciudades.

Si se dispone de una medida δ_{ij} de disimilitud (o distancia no euclidiana) entre todos los pares de ciudades i y j, entonces luego de almacenar en una matriz $\Delta = (\delta_{ij})_n$ todas las disimilitudes, considerar las siguientes estrategias:

Clásico-SOM

Aplicar Escalamiento Clásico a la matriz de disimilitudes Δ para obtener como solución una matriz configuración Y de tamaño nxk, eligiéndose el subespacio euclidiano de dimensión k = 2 en el cual se representan las filas de Y, empleando el STRAIN para medir la pérdida de información, identificándose cada fila con las correspondientes coordenadas de las n ciudades en el plano R2. Seguidamente, aplicar el algoritmo SOM para encontrar el recorrido óptimo.

Clásico-EM Absoluto-SOM

Aplicar Escalamiento Clásico a la matriz de disimilitudes Δ para obtener como solución una matriz de configuración Yo, en un subespacio euclidiano de dimensión k = 2 en el cual las filas de Y_0 son las coordenadas de las n ciudades, bajo el criterio de optimalidad STRAIN. Seguidamente, empleando la configuración Yo como punto de partida de otro criterio de optimización (el STRESS) del Escalamiento Métrico Absoluto, encontrar una configuración final Y_f que minimice el STRESS, resultando las filas de Y_f nuevas coordenadas de las n ciudades en R^2, que mejoran las iniciales. A continuación, utilizar Y_f como punto de partida para aplicar el algoritmo SOM y obtener finalmente, la trayectoria óptima.

Nótese cómo las configuraciones ofrecidas por las estrategias anteriores resuelven las dificultades para aplicar el algoritmo SOM a las n ciudades cuando son desconocidas las coordenadas de las mismas, todo lo cual permite establecer una colaboración entre el EM y el algoritmo SOM. Por supuesto que, estas son dos de las tantas variantes del EM que pueden resolver las dificultades para iniciar el trabajo del algoritmo SOM en casos para los que no era posible. Nuestra elección se debe a la disponibilidad de programas de dichos métodos, todo lo cual permite desarrollar este trabajo en ejemplos concretos.

3. DOS SOLUCIONES DEL PROBLEMA DEL VIAJANTE DE COMERCIO

Se experimentó con las estrategias combinadas del EM y los SOM en dos ejemplos. La primera estrategia Clásico-SOM fue aplicada a 21 ciudades de España (Ver resultados en Figuras 1, 2 y 3, ANEXOS 1 y 2.) y la segunda Clásico-EM Absoluto-SOM) fue aplicada a las 15 capitales de las provincias de Cuba. (Ver resultados en Figuras 4 y 5, ANEXO 3).

Se pudo apreciar cómo quedaban superadas las dificultades del algoritmo SOM al problema TSP, todo lo cual permite establecer una colaboración entre el EM y los SOM. Estas son dos de las tantas variantes del EM que pueden resolver las dificultades de los SOM. La aplicación de técnicas mixtas EM-SOM en el problema TSP no aparece reportada en la literatura, por lo que resulta una técnica novedosa para resolver el mismo.

4. CONCLUSIONES

Las técnicas de EM y los SOM resultan herramientas útiles en el Análisis Exploratorio como técnicas aisladas. Sin embargo, su empleo combinado no es de amplio conocimiento.

Las estrategias combinadas Clásico-SOM y Clásico-EM Absoluto-SOM permiten analizar cómo el EM es una vía para ampliar la gama de problemas a resolver por los SOM en particular, aquellos relacionados con el Problema del Viajante de Comercio para los que no se conocen las coordenadas de las ciudades. La información necesaria para iniciar el proceso del algoritmo SOM es suministrada por la solución del EM.

Una técnica combinada Clásico-SOM fue considerada por Miret **et al**. (2004) para emplear el algoritmo SOM en problemas de clasificación para los que no se conoce la naturaleza topológica de los vectores que identifican a los objetos como elementos de un espacio vectorial de métrica desconocida. Se resuelve un ejemplo cuyas variables iniciales son cualitativas.

REFERENCIAS

BORG, I. and P. GROENEN (1997): **Modern multidimensional scaling**. Springer-Verlag. New York, Inc.

COTTRELL, M. and P. LETREMY (1995): "Classification et analyse des correspondences au moyen de L'Algorithme de Kohonen: Application a l'étude de données économiques". **Prépublication du Samos** (42), University of Paris I, Francia.

HAYKIN, S. (1994): "Neural Networks. A comprehensive Foundation", Macmillan College Publishing Company. New York

KEARSLEY, A.; R.A. TAPIA and M.W. TROSSET (1998): "The solution of the metric SSTRESS and STRESS problems in Multidimensional Scaling using Newton's method". Computational Statistic, 13, 369-396.

MARDIA, K.V.; J.T. KENT and J.M. BIBBY (1979): "Multivariate Analysis". Academic Press, Inc., London.

MARTÍN del BRÍO, B. and A. SANZ MOLINA (2001): "Redes Neuronales y Sistemas Borrosos". RA-MA Editorial. Madrid.

MIRET, E. **et al**. (2004): "Escalamiento Multidimensional y Mapas Auto-Organizados de Kohonen. Análisis Comparativo." Ponencia presentada en el CLAIO, Ciudad Habana, Cuba.

MIRET, E. and G. LINARES (2002): "Extensión de las coordenadas canónicas de Rao a algunas técnicas del Escalamiento Multidimensional". **Revista Ciencias Matemáticas** 20, .34-43,

MIRET, E.; M.V. MEDEROS and G. LINARES (2004): "Un enfoque unificado para técnicas de representación euclidiana." Revista Ciencias Matemáticas. (Presentado para publicar).

TARAZAGA and TROSSET (1998): "An Approximate Solution to the Metric SSTRESS Problem in Multidimensional Scaling." http://citeseer.ist.psu.edu/tarazaga98approximate.html.

ZHANG, B. and S.N. SRIHARY (2003): "Properties of Binary Dissimilarity Measures." **Cedar Publications**. http://www.cedar.buffalo.edu/papers/pubs2000.html**.**

A N E X O S

ANEXO 1

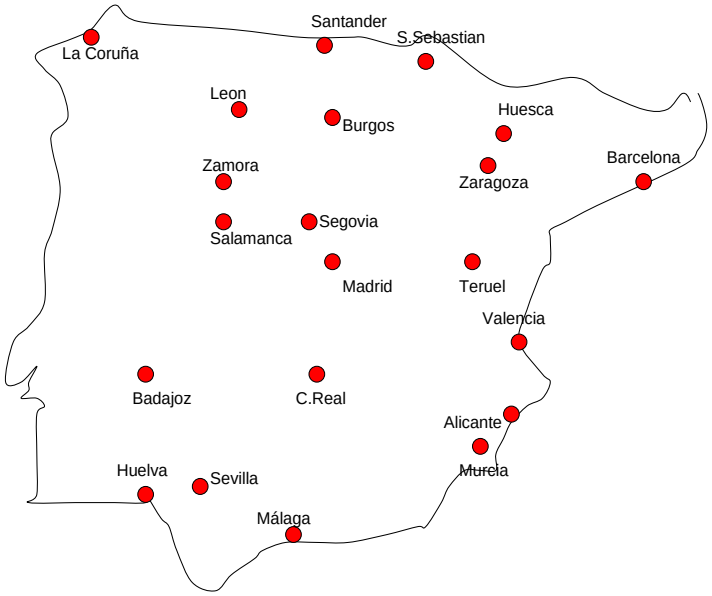

Figura 1. Mapa real de España.

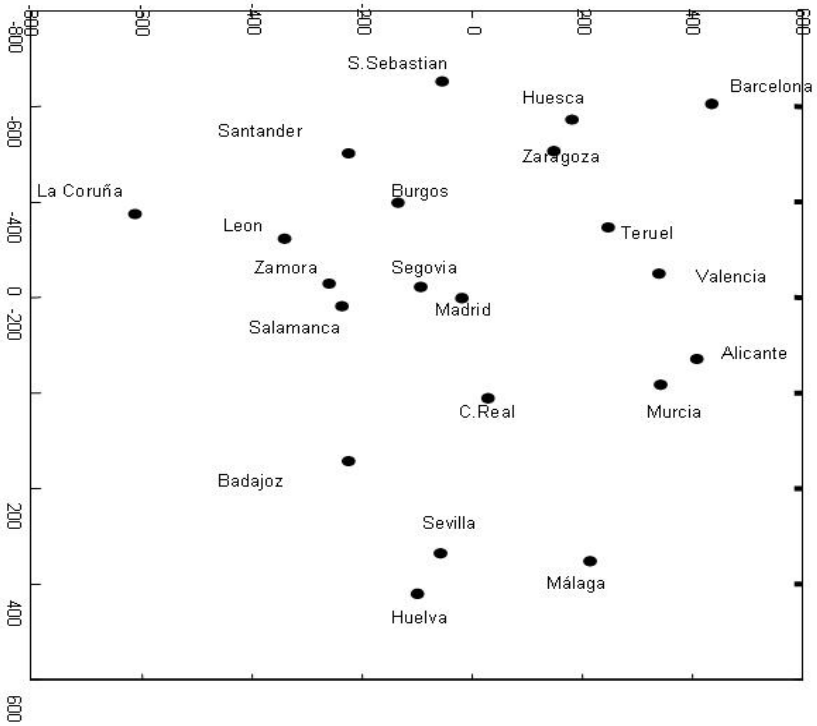

Figura 2. Mapa de las 21 ciudades de España
(Solución del Clásico).

ANEXO 2

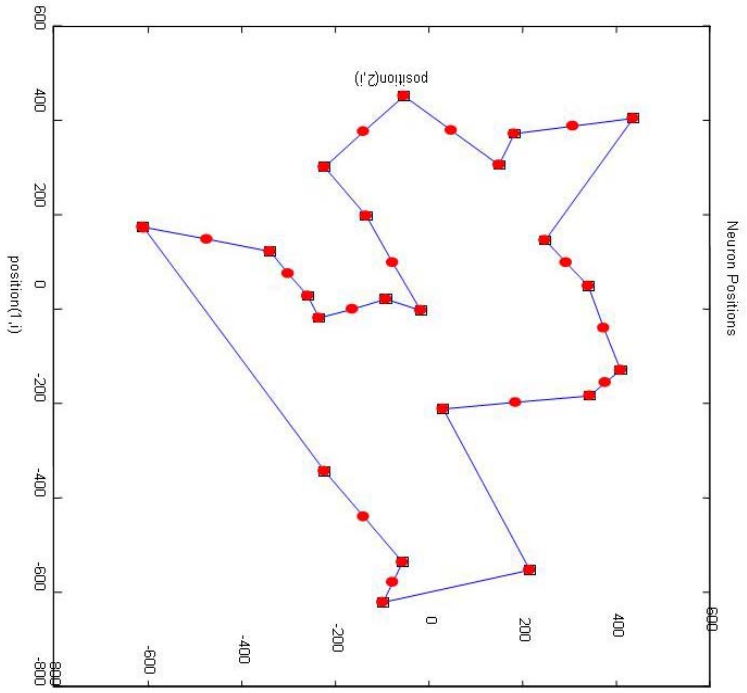

Figura 3. Organización final SOM.
(Problema TSP para las 21 ciudades de España).

ANEXO 3

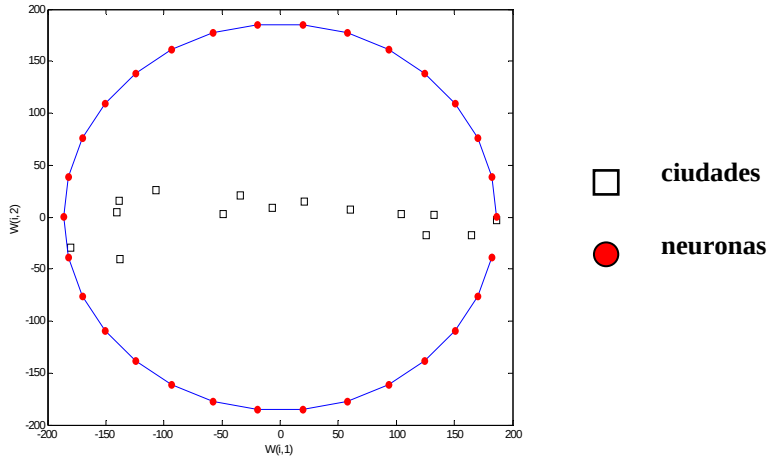

Figura 4. Mapa inicial de los SOM tomando Solución del Clásico-EM Absoluto.
(Problema TSP para las 15 ciudades de Cuba).

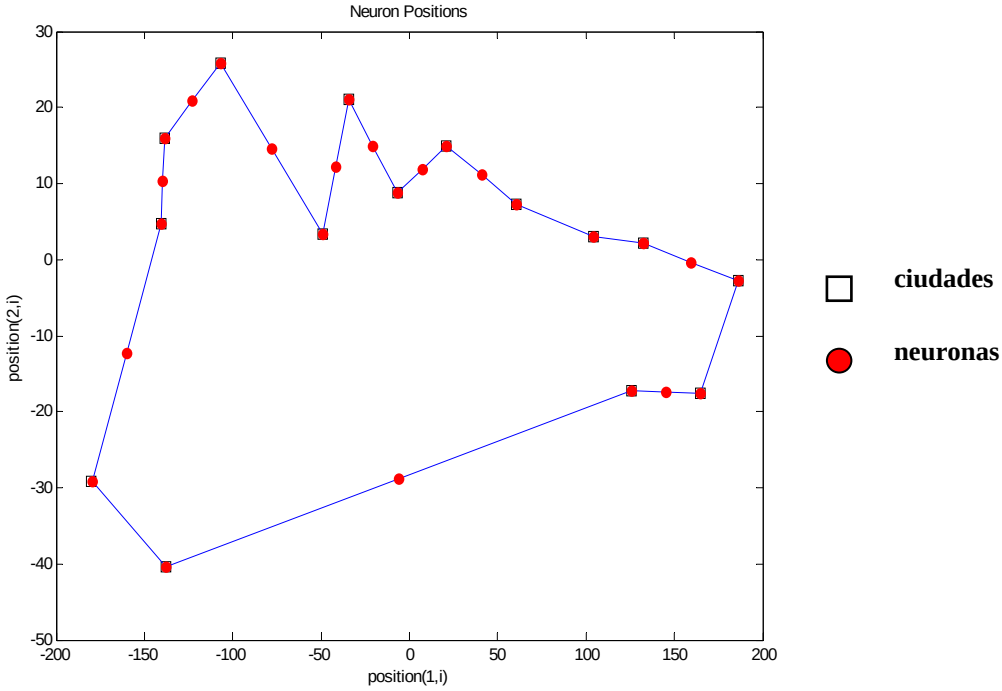

Figura 5. Organización final SOM.
(Problema TSP para las 15 ciudades de Cuba)