

UNA CLASIFICACIÓN SOCIOECONÓMICA DE LAS REGIONES EUROPEAS MEDIANTE MAPAS DE KOHONEN.

Esteban Alfaro Cortés Matías Gámez Martínez Noelia García Rubio
Área de Estadística. Departamento de Economía y Empresa.
Universidad de Castilla-La Mancha.
Plaza de la Universidad, s/n. 02071 Albacete. (Fax: 967599220)
e-mail: {Esteban.Alfaro, Matias.Gamez, Noelia.Garcia}@uclm.es

RESUMEN.

El objetivo de esta investigación consiste en la aplicación de modelos neuronales al estudio de las características socioeconómicas de las regiones europeas.

Los modelos neuronales de Kohonen, también conocidos como Mapas de Rasgos Auto-organizados o SOFM¹, constituyen un tipo de redes neuronales cuya principal característica es el uso de aprendizaje no supervisado para tareas de agrupación. La utilización de estos modelos se considera especialmente adecuada cuando se trata de reconocer clusters, a priori desconocidos, dentro de un conjunto de datos, es decir, en aquellos casos en los que no se dispone de una variable de salida objetivo susceptible de ser utilizada durante el proceso de entrenamiento de la red.

Después de mostrar el mecanismo básico de operación del algoritmo de entrenamiento de los Mapas de Kohonen, éste será aplicado al análisis socioeconómico de las regiones europeas para agruparlas de acuerdo a patrones de comportamiento distintivos. Los datos utilizados en esta aplicación se han obtenido de Regio, el banco de datos regional de Eurostat. De esta fuente se ha seleccionado un conjunto de indicadores que, de alguna forma, describen la situación económica y social de las regiones de la Unión Europea, consideradas al nivel de NUTS 2.

PALABRAS CLAVE: Mapas auto-organizados de Kohonen, análisis socio-económico, regiones europeas.

¹ SOFM son las siglas de Self-organizing Feature Maps, nombre en inglés de estos modelos.

1. INTRODUCCIÓN.

Dado que los países europeos se encuentran inmersos en un proceso creciente de integración económica, los análisis económicos regionales se hacen necesarios para diferentes propósitos. Uno de ellos es el intento de definir una clasificación de las unidades territoriales de acuerdo a patrones de comportamiento distintivos y, por otra parte, la no menos importante tarea de analizar las fuentes de las diferencias encontradas en esos patrones de comportamiento.

En este sentido, las redes neuronales artificiales y, en particular, los Mapas auto-organizados de Kohonen, constituyen una interesante herramienta alternativa a los métodos estadísticos más tradicionales como por ejemplo el Análisis Cluster.

Los mapas auto-organizados se componen de dos capas de neuronas. La capa de entrada, denominada también capa presináptica, consta de tantas neuronas como el número de componentes de los vectores de entrada (n). La segunda capa, denominada capa de salida o de competición, consta de m neuronas dispuestas normalmente en una superficie bidimensional. Estas neuronas se encuentran completamente conectadas a las neuronas de la capa de entrada mediante pesos sinápticos. El objetivo es obtener un mapa topológico de tal forma que la localización espacial de las neuronas en la capa de salida reproduzca la estructura de correlación de las señales de entrada. En otras palabras, el proceso de entrenamiento intenta conseguir la especialización de una neurona o de un grupo de neuronas vecinas en un determinado patrón de entrada de manera que neuronas vecinas respondan de forma similar para cada patrón de entrada.

Una vez presentado un patrón de entrada a la red, comienza el proceso de competición, que implica la determinación de la denominada *neurona ganadora*, es decir, aquella cuyo vector de pesos es más similar a la entrada presentada. Entonces esa neurona es actualizada con objeto de parecerse aún más al patrón de entrada para el cual se ha proclamado ganadora.

La principal novedad de los SOFM en relación a otros modelos de aprendizaje competitivo es que un grupo de neuronas alrededor de la ganadora también tendrán la oportunidad de actualizar sus vectores de pesos. La consecuencia es un cambio más suave en los vectores de pesos de las neuronas de una vecindad, es decir, se produce una ordenación espacial de las neuronas en la capa de salida.

Como se mencionó anteriormente, estos efectos serán analizados a través de una aplicación en la cual se pretende realizar un agrupamiento de las regiones de la Unión Europea. El objetivo es agrupar las regiones en clusters de manera que éstos presenten tanta homogeneidad

interna como sea posible, así como el estudio de los patrones de comportamiento de cada grupo de regiones.

Los datos utilizados en esta aplicación se han obtenido de Regio, el banco de datos regional de Eurostat. De esta fuente se ha seleccionado un conjunto de indicadores que, de alguna forma, describen la situación económica y sociales de las regiones de la Unión Europea, consideradas al nivel de NUTS 2.

2. MAPAS AUTO-ORGANIZADOS.

Este modelo fue desarrollado por el físico finlandés Teuvo Kohonen en 1982 . Él mismo lo define como [Kohonen, 1997]: *el resultado de un proceso de regresión no paramétrica que se utiliza principalmente para representar datos multidimensionales relacionados de forma no lineal, a menudo en un espacio bidimensional, y realizar clasificación no supervisada y agrupamiento.*

Los mapas auto-organizados se componen generalmente de dos capas. La capa presináptica o capa de entrada consta de un número n de neuronas igual al número de componentes en los vectores de entrada. La capa postsináptica o capa de salida, también denominada capa de competición, consta de m neuronas dispuestas generalmente en una superficie bidimensional que puede tener forma rectangular o hexagonal. En general, cada neurona de la capa de competición está conectada con todas las neuronas de la capa de entrada a través del vector de pesos sinápticos o vector de referencia² ω_i .

$$\omega_i = [\omega_{i1}, \dots, \omega_{ij}, \dots, \omega_{in}]^T \in \mathfrak{R}^n \quad (2.1)$$

donde ω_{ij} denota el peso de la conexión entre la unidad i de la capa postsináptica y la unidad j en la capa presináptica

² Dado que, como veremos más adelante, el vector de pesos es una representación de los casos de entrenamiento asignados a cada cluster, parece más conveniente la denominación de vectores de referencia que la general de vector de pesos sinápticos, más adecuada en redes como el Perceptrón multicapa.

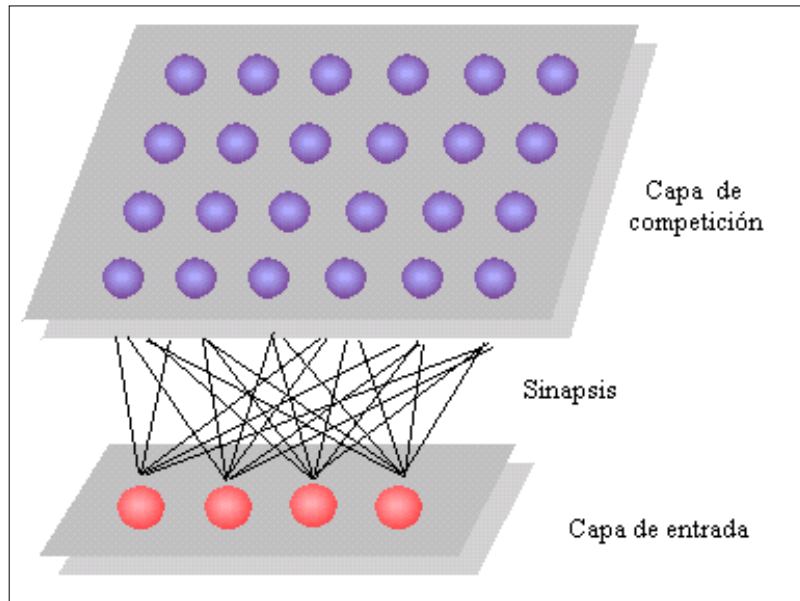


Figura 1. Arquitectura de un mapa auto-organizado

Además de los ya señalados, los mapas de Kohonen constan de los siguientes elementos:

! Un mecanismo que permita calcular el grado de ajuste entre cada neurona en la capa de competición y un vector \mathbf{x} definido como

$$\mathbf{x} = [x_1, \dots, x_j, \dots, x_n]^T \in \mathfrak{R}^n \quad (2.2)$$

Como criterio de ajuste se puede utilizar tanto una medida de similitud como una de disimilitud. En el último caso, la salida de cada neurona en la capa de competición es la distancia entre dos puntos en un espacio n -dimensional, representados por el vector de datos de entrada y el vector de referencia de la neurona. Una medida utilizada en gran parte de las aplicaciones prácticas es la distancia euclídea. En el caso de la neurona i , tenemos:

$$d_i = \|\mathbf{x} - \boldsymbol{\omega}_i\| \quad (2.3)$$

! Después de calcular el grado de ajuste para todas las unidades de salida, necesitamos un mecanismo que los compare y permita elegir la neurona de mejor ajuste, denominada neurona ganadora $g(\mathbf{x})$ para el vector de entrada \mathbf{x} . Es importante señalar que este mecanismo debe ser seleccionado de acuerdo a la métrica del criterio de ajuste. Obviamente, si se utiliza la distancia euclídea como criterio, la neurona ganadora será aquella que cumpla:

$$g(\mathbf{x}) = \arg \min_i \|\mathbf{x} - \boldsymbol{\omega}_i\| \quad i = 1, 2, \dots, m. \quad (2.4)$$

! En tercer lugar, y para conseguir un mapa topológicamente ordenado, es necesario un sistema de interacción local entre las neuronas de la capa de competición que determine el rango excitatorio e inhibitorio de la neurona ganadora. Este mecanismo se denomina *función de*

vecindad y se denota por V_{gi} .

Esta función constituye la principal novedad que los mapas de Kohonen aportan al aprendizaje competitivo y proporciona las siguientes propiedades a los mapas:

- En primer lugar, neuronas pertenecientes a una misma vecindad responden de manera similar tras la presentación de un patrón de entrada a la red.

- En segundo lugar, los vectores de referencia correspondientes a neuronas en vecindades próximas cambian suavemente.

! Finalmente, se necesita un proceso adaptativo que permita actualizar los vectores de referencia correspondientes a la neurona ganadora y sus vecinas. El propósito es conseguir que esas neuronas aprendan “algo” del vector de entrada presentado a la red de manera que cuando ese mismo vector o uno muy similar sea presentado de nuevo a la red, el grado de ajuste sea mayor que el actual.

Como ocurría con el mecanismo de selección de la neurona ganadora, la regla de actualización también debe ser compatible con el criterio de ajuste.

La ecuación de la regla de actualización correspondiente al criterio de la distancia euclídea es³:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t)V_{gi}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)] \quad (2.5)$$

donde $t = 0, 1, 2, \dots$ indica el tiempo discreto o el número de iteración actual. $\eta(t)$ es el parámetro de tasa de aprendizaje ($0 < \eta(t) < 1$), que generalmente se especifica como una función decreciente de t . Finalmente, $V_{gi}(t)$ actúa como una función de vecindad dependiente de la distancia entre la neurona i de la capa de competición y la neurona ganadora g , así como de t ⁴.

Como se ha mencionado anteriormente, la función de vecindad es equivalente a un sistema de interacciones que define cuáles son las neuronas vecinas de la ganadora y, por lo tanto, determina qué vectores de referencia serán actualizados.

Esta función se puede definir de muchas formas distintas, siendo la función escalón la elección más común y sencilla:

$$V_{gi}(t) = \begin{cases} 1 & \text{si } i \in N_g(t) \\ 0 & \text{si } i \notin N_g(t) \end{cases} \quad (2.6)$$

donde $N_g(t)$ representa a conjunto de neuronas vecinas alrededor de la ganadora g dependiendo

³ En Kohonen (1997) se deriva esta regla de manera general partiendo de la definición de una función de error que depende de la distancia entre cada vector de entrada y el vector de referencia de la neurona ganadora correspondiente. Esta función de error es optimizada mediante el método del gradiente descendente. Posteriormente este procedimiento general se particulariza para las medidas de distancia euclídea y de Minkowski.

de un parámetro $R(t)$ denominado radio o tamaño de la vecindad que normalmente se define como una función monótona decreciente en el tiempo.

Siguiendo esta formulación tradicional de la función de vecindad, la regla de actualización queda como:

$$\omega_i(t+1) = \begin{cases} \omega_i(t) + \eta(t)[\mathbf{x}(t) - \omega_i(t)] & \text{si } i \in N_g(t) \\ \omega_i(t) & \text{si } i \notin N_g(t) \end{cases} \quad (2.7)$$

Merece la pena señalar que se pueden diferenciar dos etapas fundamentales en el proceso de entrenamiento además de la inicialización de los vectores de referencia. La primera de ellas, denominada *fase de ordenación global*, consiste en aproximadamente 1000 iteraciones. En esta etapa tiene lugar la ordenación topológica de los vectores de referencia. La segunda etapa, generalmente denominada *fase de ajuste fino*, es mucho más larga que la primera y tiene como objetivo conseguir la convergencia de los vectores de referencia a unos valores asintóticos que representan la imagen de la función de densidad de los datos de entrenamiento $p(x)$. El número de iteraciones en esta fase es crucial para la precisión estadística de la representación final.

En relación a los parámetros de entrenamiento, Kohonen [1997] afirma que la forma en que $\eta(t)$ y $R(t)$ decrecen no es importante, de manera que una elección bastante frecuente en la práctica es hacer que decrezcan linealmente.

Resumimos a continuación el algoritmo completo:

-Paso 1: *Inicialización*. La práctica más común es elegir aleatoriamente valores pequeños para los vectores de referencia iniciales.

- Paso 2: *Muestreo*. Extracción de un patrón de entrada del conjunto de entrenamiento y presentación del mismo a la red.

- Paso 3: *Búsqueda del mejor ajuste*. Cálculo del ajuste entre el vector de entrada y cada uno de los vectores de referencia y comparación de los mismos para seleccionar la neurona ganadora.

- Paso 4: *Actualización*. Ajuste de los vectores de referencia de la neurona ganadora y sus vecinas. En este paso debe tenerse en cuenta el ajuste del radio de la función de vecindad así como del parámetro de tasa de aprendizaje.

- Paso 5: *Repetición*. Vuelta al paso 2 y repetición de los pasos siguientes hasta que no se aprecien cambios significativos en el mapa resultante o hasta que se alcance el número máximo de iteraciones, en caso de que se haya fijado.

⁴ Una condición necesaria para la convergencia del proceso es que $V_{gi}(t) \rightarrow 0$ cuando $t \rightarrow \infty$.

3. APLICACIÓN DE MAPAS DE KOHONEN AL AGRUPAMIENTO DE LAS REGIONES EUROPEAS.

Como se ha señalado en el primer punto de este trabajo, el objetivo de esta investigación es conseguir un agrupamiento de las regiones de la Unión Europea atendiendo a algunas características socioeconómicas.

Los datos utilizados en esta aplicación representan los indicadores regionales más significativos al nivel 2 de la Nomenclatura de Unidades Territoriales para Estadística (NUTS 2) de Eurostat. El conjunto de datos se ha obtenido a partir de la edición de 1999 del Anuario Estadístico Regional elaborado por Eurostat. Dicha publicación contiene una selección de estadísticas comparables que representan la situación económica y social de las regiones de la Unión Europea.

En la tabla 1 del apéndice se muestra la correspondencia entre las divisiones administrativas nacionales y las NUTS a niveles 1, 2, y 3 en cada país de la Unión Europea. Por otra parte, en la tabla 2 se proporciona una breve descripción de cada indicador.

Antes de utilizar el conjunto de datos es conveniente, como en toda técnica estadística, llevar a cabo un preprocesamiento de los datos de entrada, así como de los de salida si existen.

En primer lugar, es muy frecuente encontrar datos *missing*. En este sentido, la matriz de datos utilizada en esta aplicación no es una excepción. Una solución común para este problema es rellenar los datos *missing* mediante la media de los valores para cada variable. Sin embargo, en este caso parece más apropiado reemplazar el caso perdido por el valor de la variable en el nivel NUT inmediatamente superior, si es posible. En otras palabras, si encontramos un dato perdido al nivel NUTS 2, podemos buscar el valor de la variable correspondiente en el nivel NUTS 1 en el que el caso perdido es agrupado. Si encontramos que éste también es un caso perdido, habrá que recurrir al valor de esa variable para el conjunto del país y en el peor de los casos será necesario recurrir a la media de la variable⁵.

Debe ser puesto de manifiesto que los autores de este trabajo habrían querido utilizar algún otro indicador considerado como importante. Sin embargo, debido a la gran cantidad de datos *missing* que presentaban hubo que restringir la aplicación a los “principales indicadores” de la base de datos Regio omitiendo la tasa de variación de la población de 1976 a 1986 por la misma razón. Otro de los principales indicadores (área en km²) fue eliminado por considerar que no contenía información ni sociológica ni económica relevante, siendo más importante la consideración de la densidad de población.

⁵ Esta situación sólo se ha dado en algunas variables en el caso de Irlanda.

Después de completar la matriz de datos, el siguiente paso es la normalización de las variables. Este proceso se vuelve crucial especialmente cuando, como ocurre en esta aplicación, las variables de entrada están medidas en escalas muy distintas. Si una de las variables de entrada tiene un rango de valores mucho mayor que las demás, los valores de las distancias euclídeas calculadas se encontrarán completamente dominadas por esa variable.

Hay muchas opciones para llevar a cabo el proceso de normalización. En este caso los autores eligieron transformar las variables linealmente para tener media cero y desviación típica unitaria.⁶

Una vez que la matriz de datos tiene una forma apropiada para trabajar con ella, el siguiente paso consiste en el diseño de la arquitectura de la red neuronal. Para ello, se han realizado numerosas pruebas⁷, modificando el número de elementos en la capa de competición, los parámetros de aprendizaje así como los vectores iniciales de pesos.

Concretamente se analizaron SOFMs de orden 4x4, 7x7, 10x10 y 14x14, que tienen 16, 49, 100 y 196 nodos de salida respectivamente. Se muestran aquí los resultados para los mapas 4x4 y 14x14. En el primer mapa cada nodo de salida se toma como un cluster, de manera que en total hay 16. Por otra parte, los clusters en el mapa 14x14 se forman mediante grupos de nodos vecinos. Como mostraremos más adelante, los resultados de ambos mapas son consistentes en el sentido de que una consecución de mapas de diferente tamaño puede considerarse como un dendrograma en el Análisis Cluster tradicional.

Comenzamos analizando el mapa auto-organizado 4x4, pues proporciona un agrupamiento suficiente y es más fácil de explicar que los restantes mapas.

Este mapa ha sido entrenado en dos fases. La primera de ellas, consistente en 1000 iteraciones, comienza con un parámetro de tasa de aprendizaje de 0,9 para decrecer hasta 0,01 y un radio de vecindad que varía desde 3 hasta 1.

Durante la segunda etapa de entrenamiento, más larga que la primera (10.000 iteraciones), la tasa de aprendizaje cambia desde 0,01 hasta 0,001 y el radio de vecindad desde 1 hasta 0.

La tabla 3 muestra el número de regiones en cada grupo. Hay dos clusters con 28 y 23 regiones (los clusters número siete y quince, respectivamente) y 6 clusters con menos de 10 regiones. El resto de clusters tiene entre 10 y 19 regiones.

En la siguiente figura se muestran los clusters obtenidos:

⁶ Véase Bishop [1995], capítulo 8, para un amplio análisis de los métodos de preprocesamiento y extracción de características con Redes Neuronales Artificiales.

⁷ Para realizar el entrenamiento de las redes se ha utilizado el software denominado Trajan Neural Networks.

de4 de8 ded1 ded2 ded3 dee1 dee2 dee3 deg →	nl12 pt13 ukc2 ukd3 ukd4 ukd5 uke1 uke3 uke4 ukf1 ukg2 ukg3 ukl1 ukl2 ukm3 ↘	dk nl11 nl21 nl22 nl23 nl31 nl32 nl33 nl41 nl42 at32 at33 ukd2 uke2 ukf2 ukg1 ukh1 ukh2 ukh3 uki2 ukj1 ukj2 ukj3 ukj4 ukk1 ukm1 ukm2 ukm4 ←	fi16 se01 se02 se04 se06 se07 se08 se0a ←
gr43 pt11 pt12 pt2 pt3 ↗	fr21 fr22 fr23 fr24 fr25 fr3 fr41 fr42 fr43 fr51 fr71 ie fi14 fi15 fi17 ukc1 ukn ↓	be24 de93 def gr22 it31 nl13 nl34 at11 at12 at22 pt15 fi2 se09 ukd1 ukf3 ukk2 ukk3 ukk4 ↖	be1 de3 at13 uki1 ↓
gr41 es11 es12 es13 es21 es23 es41 es42 es52 fr83 it71 it72 it92 →	be31 be32 be33 be34 be35 gr3 gr42 es3 es53 fr26 fr52 fr53 fr61 fr62 fr72 fr81 fr82 it6 fi13 ↑	be21 be22 be23 be25 de72 de73 de91 de92 de94 dea1 dea2 dea3 dea4 dea5 deb1 deb2 deb3 dec es22 es51 it12 it2 it32 ↘	de21 de5 de6 de71 fr1 lu ←
es43 es61 es62 es63 es7 it8 it91 it93 ita itb ↑	gr11 gr12 gr13 gr14 gr21 gr23 gr24 gr25 ↖	es24 fr63 it11 it13 it33 it4 it51 it52 it53 pt14 ↑	de11 de12 de13 de14 de22 de23 de24 de25 de26 de27 at21 at31 at34 ↖

Figura 2. mapa 4x4.

Los nombres completos de las regiones pueden verse en la tabla 4.

El propósito de colorear cada cluster se explicará en la comparación de los mapas 4x4 y 14x14.

En la esquina izquierda de abajo de cada cluster podemos encontrar una flecha cuya dirección representa la distancia más pequeña entre el cluster correspondiente y sus vecinos calculada mediante la distancia euclídea entre los respectivos vectores de referencia. De esta forma, este procedimiento nos permite conocer qué cluster es el más cercano a cada uno.

Describimos ahora brevemente cada cluster diciendo qué regiones contiene y cuáles son sus principales características. Para realizar esta tarea es muy útil analizar los vectores de referencia, que no son otra cosa que una aproximación a los centroides de los casos que son

agrupados en cada cluster. Estos vectores de referencia, para las variables normalizadas, se muestran en la tabla 5.

El cluster 1 contiene 10 regiones, sólo de España e Italia. Este conglomerado presenta altas tasas de desempleo, especialmente femenino, alta tasa de dependencia y también una alta proporción de individuos por debajo de 25 años. Por otra parte, presentan bajas tasas de actividad, también especialmente femenina.

El cluster 2 está formado por 8 regiones, todas ellas griegas. Estas regiones se caracterizan por tener una alta participación de la agricultura en el empleo total y una tasa de mortalidad infantil también alta, así como baja participación del sector servicios tanto en el empleo total como en el producto interior bruto.

El cluster 3 agrupa 10 regiones, principalmente de Italia y también de España, Francia y Portugal. Se caracteriza por bajos valores en la población de menos de 25 años, en la tasa de natalidad y en las tasas de actividad, especialmente masculinas. También presentan una alta proporción de población por encima de los 65 años.

El cluster 4 está formado por 13 regiones de Alemania y Austria. Este cluster presenta altos niveles de participación de la industria en el empleo total, altas tasas de variación anual en la población y un alto producto interior bruto per cápita. Sin embargo, presenta una baja participación de los servicios en el empleo total.

El cluster 5 tiene 13 regiones fundamentalmente de España (8) pero también de Italia (3), Francia y Grecia. Tienen una alta tasa de dependencia y desempleo femenino, así como bajas tasas de actividad, especialmente femenina, y también baja tasa de natalidad.

El cluster 6 contiene 19 regiones, fundamentalmente de Francia (8) y Bélgica (5) y también Grecia, España, Italia y Finlandia. Se caracteriza por un comportamiento bastante estándar en general, sin embargo presenta una participación de los servicios en el empleo total por encima de la media. Por otra parte, la participación de la industria y las tasas de actividad se encuentran por debajo de la media.

El cluster 7 es uno de los grupos más grandes con 23 regiones principalmente de Alemania (14) y también de Bélgica (4), Italia (3) y España (2). Este cluster presenta un alto nivel de participación del sector industrial en el empleo total y también del producto interior bruto per cápita. Sin embargo, presenta una baja proporción de población menor de 25 años.

El cluster 8 tiene sólo 6 regiones, la mayor parte alemanas, una de Francia y Luxemburgo. Este grupo presenta un alto producto interior bruto tanto per cápita como convertido a unidades estándar de poder adquisitivo en relación a la población media. También en este caso se registra

una baja proporción de población menor de 25 años.

El cluster 9 es el segundo más pequeño con sólo 5 regiones, 4 de ellas son de Portugal y una de Grecia. Este grupo se caracteriza por una alta participación de la agricultura en el empleo total, y niveles también relativamente altos de mortalidad infantil y población menor de 25 años. Se caracteriza asimismo por una baja participación del sector servicios en el empleo total y bajos niveles de producto interior bruto y tasa de desempleo masculino.

El cluster 10 está formado por 17 regiones, en su mayoría francesas (11) y también finlandesas (3), del Reino Unido (2) e Irlanda. Este grupo presenta alta proporción de población por debajo de 25 años y también alta tasa de natalidad. Por otra parte, presenta una tasa baja de mortalidad en niños menores de un año.

El cluster 11 contiene 18 regiones de multitud de países. Al Reino Unido pertenecen 5, 3 a Austria, 2 a Alemania y Holanda, y, finalmente 1 a Bélgica, Grecia, Italia, Portugal, Finlandia y Suecia. Estas regiones se caracterizan por un comportamiento estándar, una alta proporción de población mayor de 65 años y bajas tasas de desempleo, tanto masculino como femenino.

El cluster 12 es el más pequeño de los grupos con sólo 4 regiones, cada una de ellas de un país (Bélgica, Alemania, Austria y el Reino Unido). Este conglomerado se caracteriza por una alta densidad de población, una importante participación de los servicios en el empleo total y también alto producto interior bruto.

El cluster 13 está formado por 9 regiones, todas ellas alemanas. Estas regiones presentan bajas tasas de natalidad, de dependencia y producto interior bruto per capita por debajo de la media. Por otra parte, las tasas de desempleo son altas, especialmente las masculinas.

El cluster 14 contiene 15 regiones, principalmente del Reino Unido (13) y también de Holanda y Portugal. Este cluster presenta un comportamiento estándar excepto por las relativamente altas tasas de mortalidad infantil y natalidad. Registra así mismo un bajo desempleo femenino y escasa importancia de la agricultura en el empleo total.

El cluster 15 es el mayor de todos los grupos. Contiene 28 regiones, mayoritariamente del Reino Unido (16), Holanda (9) y también de Austria (2) y Dinamarca. Se caracteriza por presentar tasas de actividad relativamente altas, tanto masculina como femenina y también por bajas tasas de desempleo, masculino, femenino y total.

El cluster 16 está formado por 8 regiones, 7 de ellas corresponden a Suecia y una a Finlandia. Presenta altas tasas de actividad, especialmente femenina, gran participación del sector servicios en el empleo total y altos niveles de producto interior bruto per cápita. Asimismo, presenta bajas tasas de mortalidad infantil y de dependencia.

Analizamos a continuación el mapa 14x14 mostrado en la figura 5. En este mapa la capa de competición consta de 196 nodos, casi tantos como casos de entrada. Por esta razón, no es extraño que después del entrenamiento de la red, muchos de los aquellos nodos se encuentren vacíos, informando a menudo de posibles límites de separación entre clusters.

Una vez coloreados los nodos formado los clusters obtenidos para el mapa 4x4, es interesante señalar los siguientes hechos:

- Regiones que habían formado un cluster en el mapa más pequeño, presentan localizaciones próximas en el mapa mayor con muy pocas excepciones.

- Se mantiene la estructura de vecindad, es decir, clusters vecinos en el mapa 4x4 lo son también en el 14x14, existiendo en este último un cierto solapamiento entre algunos de los grupos más cercanos del mapa más pequeño.

En algunas ocasiones y con objeto de analizar las fronteras entre clusters se puede recurrir a una útil herramienta que es un mapa de niveles de colores. Este método, propuesto por Kraaijveld et al., consiste en representar las distancias relativas entre vectores de referencia vecinos mediante grados en una escala de color. De esta forma, una distancia promedio pequeña se representa mediante un color claro, mientras que tonalidades oscuras representarían grandes distancias. A continuación y mediante la superposición de este mapa sobre el mapa auto-organizado es posible visualizar la separación de clusters. Las figuras 3 y 4 muestran estos mapas⁸

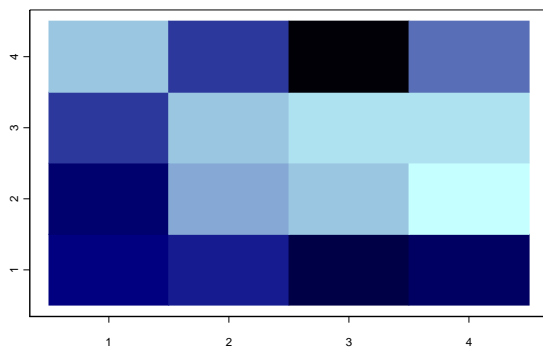


Figura 3. Mapa de niveles de colores para el SOFM 4x4.

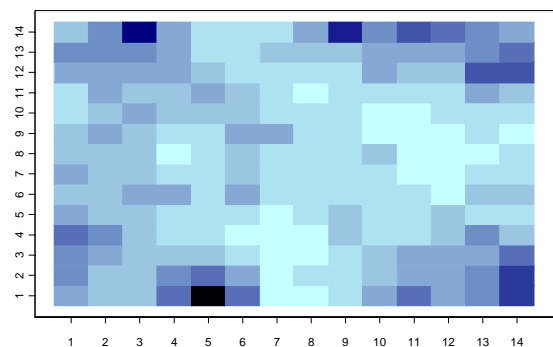


Figura 4. Mapa de niveles de colores para el SOFM 14x14.

⁸ El cálculo de las distancias y su representación mediante mapas de niveles de colores se ha realizado utilizando el paquete S-plus.

be1 uki1	de3	fr1		nl31 ukj1	nl32 nl33	dk at32	nl34	nl12	at34		pt12	gr43	gr11 gr14
							nl11	nl42 at33	nl21 nl41				gr23
at13		fi16	se01		ukh2	ukm1		nl22		at31	pt11		gr12
de5 de6			se08	se02	ukg1 ukj3	ukk1		ukd2	ukf2				pt2 pt3
Lu		se07	se04 se06		uke2	ukh3	ukh1 ukj4 ukm2		ukf1 ukg2	uke4	uki2		nl23
											ukd3 ukg3		
de21 de71		se0a			ukj2	ukk4	ukm4		ukd4 uke1		ukd5		ie ukn
de11 de12 de25		se09 ukd1	nl13	def fi2		ukf3 ukk2		pt13	uk12	uke3 uk11	ukc2 ukm3		ukc1
	de24			de93		at11 at12		pt15 ukk3				fr71	fr42
de14 de22 de23 de27	de13 de26	dea4 deb3	de94 dea3		deb1 deb2	at22		gr22		be24	fr41	fr43	fr24 fr51
										be31 be34			
		de72	de73 de92	dea2	it12	it31		gr42		es53		fr21 fr25	fi14 fi17
at21		de91	dea1 dec		be21 be23 be25				be35	it6	be33		fr22 fr23 fi15
		it2 it32	dea5		be22		fr53 fr62 fr72	fr61	fr81 fr82	be32	gr3		fr3
it33 it4	it11 it53		es22		es23 it71		fr26 fr52		fi13		es3	es62 es7	es63
it51	it52	es24	gr13 gr24		it72			es51		de4 de8		itb	it8
it13	fr63	pt14	gr25	gr21	gr41 es12 es41	fr83 it92	es13 es42	es21 es52		ded1 ded2 ded3 dee1 dee2 dee3 deg		es43 es61	it91 it93 ita
				es11									

Figura 5. mapa 14x14.

Por último, se ha realizado un análisis de la varianza para estudiar si existen diferencias significativas de los vectores medias entre los 16 grupos. A continuación se muestra un resumen de los resultados para el análisis de la varianza, tanto en el caso multivariante (MANOVA) como en el univariante (ANOVA). Podemos observar que las diferencias entre los grupos son significativas para todas las variables conjunta e individualmente.

Contrastes multivariados (MANOVA)

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Signif
GRUPO	Traza de Pillai	7,710	9,714	288,000	3008,000	,000
	Lambda de Wilks	,000	38,172	288,000	2120,912	,000
	Traza de Hotelling	18031,423	10713,984	288,000	2738,000	,000
	Raíz mayor de Roy	18001,465	188015,301 ^a	18,000	188,000	,000

a El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación.

Pruebas de los efectos para cada variable (ANOVA)

Fuente	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo	POP_DENS	149333147,090	16	9333321,693	30,630	,000
	POP96_86	453,419	16	28,339	1,694	,051
	POP25_96	195946,011	16	12246,626	3184,474	,000
	POP65_96	53149,459	16	3321,841	1211,580	,000
	BIRTH_96	245,582	16	15,349	753,865	,000
	INF_MO96	61,759	16	3,860	389,946	,000
	ACT_T_98	652065,320	16	40754,083	5442,931	,000
	ACT_M_98	902656,399	16	56416,025	8462,913	,000
	ACT_F_98	456845,843	16	28552,865	1982,499	,000
	DEP_RT98	619,464	16	38,716	507,804	,000
	UN_T_98	24483,694	16	1530,231	277,754	,000
	UN_M_98	17114,328	16	1069,646	187,461	,000
	UN_F_98	40164,164	16	2510,260	269,141	,000
	EMP_AGR	14848,571	16	928,036	82,807	,000
	EMP_IND	174614,711	16	10913,419	435,013	,000
	EMP_SER	875101,956	16	54693,872	1991,966	,000
	GDPECUHA	2020702,938	16	126293,934	484,546	,000
GDPPPSHA	1965179,759	16	122823,735	624,374	,000	

4. CONCLUSIONES.

Para finalizar, merece la pena resaltar los siguientes puntos de este trabajo. En primer lugar, dado que el número de casos perdidos en el conjunto original de variables era elevado, se propuso como solución rellenar la matriz de datos acudiendo al nivel inmediatamente superior (NUT 1 o NUT 0). En este sentido hay que señalar que sería deseable disponer de estadísticas más completas que permitieran obtener mejores resultados en el análisis de diferentes aspectos relacionados con la Unión Europea.

En segundo lugar, creemos interesante destacar la gran utilidad de este tipo de redes neuronales no supervisadas, así como la posibilidad de considerar una consecución de mapas de

diferente tamaño como un procedimiento de cluster jerárquico.

Finalmente, debe remarcarse que aunque en este trabajo se han utilizado los mapas de Kohonen con un propósito únicamente descriptivo, eso no significa que no puedan ser utilizados como técnica de predicción. Por ejemplo, podría resultar interesante tomar observaciones relativas a los indicadores utilizados en este trabajo para otras regiones que aún no pertenecen a la Unión Europea pero que lo harán en breve. Estos casos pueden ser ejecutados en la red de manera que conozcamos cuál es el vector de referencia más próximo y, por tanto, a qué cluster pertenecerían o con qué regiones de la Unión Europea quedarían agrupados.

5. APÉNDICE.

TABLA 1. CORRESPONDENCIA ENTRE NIVELES NUTS Y DIVISIONES ADMINISTRATIVAS NACIONALES.				
		NUTS 1	NUTS 2	NUTS 3
Belgique/België	(be)	Regions	Provinces	Arrondissements
Danmark	(dk)	---	---	Amter
Deutschland	(de)	Länder	Regierungsbezirke	Kreise
Ellada	(gr)	NUTS 2 groupings	Entwicklungsregionen	Nomoi
España	(es)	NUTS 2 groupings	Comunidades Autónomas + Ceuta y Melilla	Provincias + Ceuta y Melilla
France	(fr)	ZEAT + DOM	Régions + DOM	Départements + DOM
Ireland	(ie)	---	Regions	Regional Authority Regions
Italia	(it)	NUTS 2 groupings	Regioni	Provincia
Luxembourg	(lu)	---	---	---
Nederland	(nl)	Landsdelen	Provincies	COROP - Regio's
Österreich	(at)	Gruppen von Bundesländern	Bundesländern	Gruppen von Politischen Bezirken
Portugal	(pt)	NUTS 2 groupings	Comissoes de coordenação regional + Regioes autónomas	Grouping of Concelhos
Suomi/Finland	(fi)	Manner-Suomi / Ahvenanmaa	Suuralueet	Maakunnat
Sverige	(se)	---	Riksomraden	Län
United Kingdom	(uk)	Government Office Regions	NUTS 3 groupings	Counties, local authority regions

TABLA 2. DESCRIPCIÓN DE LOS PRINCIPALES INDICADORES.			
pop_dens	Densidad de Población.	dep_rt98	Tasa de dependencia obtenida como cociente de inactivos sobre población activa.
pop96_86	Tasa de variación anual de la población.	un_t_98	Tasa de desempleo total.
pop25_96	Población menor de 25 años (%).	un_m_98	Tasa de desempleo masculino.
pop65_96	Población mayor de 65 años (%).	un_f_98	Tasa de desempleo femenino.
birth_96	Tasa de natalidad.	emp_agr	Participación de la agricultura en el empleo total.
inf_mo96	Tasa de mortalidad antes del año sobre nacidos vivos.	emp_ind	Participación de la industria en el empleo total.
act_t_98	Tasa de actividad total.	emp_ser	Participación de los servicios en el empleo total.
act_m_98	Tasa de actividad masculina.	gdpecuha	Producto interior bruto per cápita.
act_f_98	Tasa de actividad femenina.	gdpppsha	PIB convertido a unidades estándar de poder adquisitivo en relación a población media y expresado en índices (Media de la UE=100)

TABLA 3. FRECUENCIAS GANADORAS.							
Cluster	Regiones	Cluster	Regiones	Cluster	Regiones	Cluster	Regiones
1	10	5	13	9	5	13	9
2	8	6	19	10	17	14	15
3	10	7	23	11	18	15	28
4	13	8	6	12	4	16	8

TABLA 4. CÓDIGOS DE LAS NUTS 2.

be1	Région capitale/Brussels gewest	Bruxelles- hoofdstad	dea1	Dusseldorf	es23	La Rioja
be21	Antwerpen		dea2	Köln	es24	Aragón
be22	Limburg (B)		dea3	Münster	es3	Comunidad de Madrid
be23	Oost-Vlaanderen		dea4	Detmold	es41	Castilla y León
be24	Vlaams Brabant		dea5	Arnsberg	es42	Castilla-la Mancha
be25	West-Vlaanderen		deb1	Koblenz	es43	Extremadura
be31	Brabant Wallon		deb2	Trier	es51	Cataluña
be32	Hainaut		deb3	Rheinhessen-Pfalz	es52	Comunidad Valenciana
be33	Liège		dec	Saarland	es53	Baleares
be34	Luxembourg (B)		ded1	Chemnitz	es61	Andalucía
be35	Namur		ded2	Dresden	es62	Murcia
dk	Denmark		ded3	Leipzig	es63	Ceuta y Melilla (ES)
de11	Stuttgart		dee1	Dessau	es7	Canarias (ES)
de12	Karlsruhe		dee2	Halle	fr1	Île de France
de13	Freiburg		dee3	Magdeburg	fr21	Champagne-Ardenne
de14	Tübingen		def	Schleswig-Holstein	fr22	Picardie
de21	Oberbayern		deg	Thüringen	fr23	Haute-Normandie
de22	Niederbayern		gr11	Anatoliki Makedonia, Thraki	fr24	Centre
de23	Oberpfalz		gr12	Kentriki Makedonia	fr25	Basse-Normandie
de24	Oberfranken		gr13	Dytiki Makedonia	fr26	Bourgogne
de25	Mittelfranken		gr14	Thessalia	fr3	Nord - Pas-de-Calais
de26	Unterfranken		gr21	Ipeiros	fr41	Lorraine
de27	Schwaben		gr22	Ionia Nisia	fr42	Alsace
de3	Berlin		gr23	Dytiki Ellada	fr43	Franche-Comté
de4	Brandenburg		gr24	Stereia Ellada	fr51	Pays de la Loire
de5	Bremen		gr25	Peloponnisos	fr52	Bretagne
de6	Hamburg		gr3	Attiki	fr53	Poitou-Charentes
de71	Darmstadt		gr41	Voreio Aigaio	fr61	Aquitaine
de72	Gießen		gr42	Notio Aigaio	fr62	Midi-Pyrénées
de73	Kassel		gr43	Kriti	fr63	Limousin
de8	Mecklenburg-Vorpommern		es11	Galicia	fr71	Rhône-Alpes
de91	Braunschweig		es12	Principado de Asturias	fr72	Auvergne
de92	Hannover		es13	Cantabria	fr81	Languedoc-Roussillon
de93	Lüneburg		es21	Pais Vasco	fr82	Provence-Alpes-Côte d'Azur
de94	Weser-Ems		es22	Comunidad Foral de Navarra	fr83	Corse

TABLA 4. CÓDIGOS DE LAS NUTS 2 (CONTINUACIÓN).

ie	Ireland	at11	Burgenland	ukd3	Greater Manchester
it11	Piemonte	at12	Niederösterreich	ukd4	Lancashire
it12	Valle d'Aosta	at13	Wien	ukd5	Merseyside
it13	Liguria	at21	Kärnten	uke1	East Riding and North Lincolnshire
it2	Lombardia	at22	Steiermark	uke2	North Yorkshire
it31	Trentino-Alto Adige	at31	Oberösterreich	uke3	South Yorkshire
it32	Veneto	at32	Salzburg	uke4	West Yorkshire
it33	Friuli-Venezia Giulia	at33	Tirol	ukf1	Derbyshire and Nottinghamshire
it4	Emilia-Romagna	at34	Vorarlberg	ukf2	Leicestershire, Rutland and Northants
it51	Toscana	pt11	Norte	ukf3	Lincolnshire
it52	Umbria	pt12	Centro (P)	ukg1	Herefordshire, Worcestershire and Warks
it53	Marche	pt13	Lisboa e Vale do Tejo	ukg2	Shropshire and Staffordshire
it6	Lazio	pt14	Alentejo	ukg3	West Midlands
it71	Abruzzo	pt15	Algarve	ukh1	East Anglia
it72	Molise	pt2	Açores (PT)	ukh2	Bedfordshire, Hertfordshire
it8	Campania	pt3	Madeira (PT)	ukh3	Essex
it91	Puglia	fi13	Itä-Suomi	uki1	Inner London
it92	Basilicata	fi14	Väli-Suomi	uki2	Outer London
it93	Calabria	fi15	Pohjois-Suomi	ukj1	Berkshire, Bucks and Oxfordshire
ita	Sicilia	fi16	Uusimaa (suuralue)	ukj2	Surrey, East and West Sussex
itb	Sardegna	fi17	Etelä-Suomi	ukj3	Hampshire and Isle of Wight
lu	Luxembourg	fi2	Åland	ukj4	Kent
nl11	Groningen	se01	Stockholm	ukk1	Gloucestershire, Wiltshire and North Somerset
nl12	Friesland	se02	Östra Mellansverige	ukk2	Dorset and Somerset
nl13	Drenthe	se04	Sydsverige	ukk3	Cornwall and Isles of Scilly
nl21	Overijssel	se06	Norra Mellansverige	ukk4	Devon
nl22	Gelderland	se07	Mellersta Norrland	ukl1	West Wales and The Valleys
nl23	Flevoland	se08	Övre Norrland	ukl2	East Wales
nl31	Utrecht	se09	Småland med öarna	ukm1	North Eastern Scotland
nl32	Noord-Holland	se0a	Västsverige	ukm2	Eastern Scotland
nl33	Zuid-Holland	ukc1	Tees Valley and Durham	ukm3	South Western Scotland
nl34	Zeeland	ukc2	Northumberland, Tyne and Wear	ukm4	Highlands and Islands
nl41	Noord-Brabant	ukd1	Cumbria	ukn	Northern Ireland
nl42	Limburg (NL)	ukd2	Cheshire		

TABLA 5. VECTORES DE REFERENCIA.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
POP_DENS	0,1991	-0,3676	-0,2835	-0,2229	-0,3250	-0,2097	-0,0809	0,6112	-0,2629	-0,2979	-0,2911	5,5366	-0,2409	0,4016	0,0504	-0,3384
POP96_86	-0,1001	0,0252	-0,2230	1,1500	-0,1917	-0,0535	-0,0272	0,0171	-0,1628	-0,0826	-0,0351	0,2660	-0,3844	-0,1338	-0,0050	-0,0622
POP25_96	1,5579	0,1555	-1,7568	-0,4110	-0,1634	0,1917	-0,8248	-0,9787	1,9070	1,2323	-0,3403	-0,6903	-0,7982	0,4631	0,2613	0,0916
POP65_96	-1,1160	0,2492	2,1704	-0,4281	0,8826	0,2640	0,1043	-0,3886	-0,8248	-0,7699	0,5367	-0,0553	-0,2896	-0,0973	-0,5780	0,5642
BIRTH_96	0,2342	-1,0077	-1,3624	0,1694	-1,0246	0,1012	-0,2651	0,3968	0,5447	1,2495	0,0273	0,2121	-2,1163	0,6038	0,6238	-0,0651
INF_MO96	1,1061	1,5449	-0,0647	-0,4530	-0,0026	-0,4382	-0,2841	-0,4062	2,0572	-0,8007	-0,5278	-0,2840	0,2850	0,8865	0,2907	-1,1987
ACT_T_98	-1,4918	-0,9127	-1,1775	0,5286	-1,6965	-0,6647	-0,4480	0,2684	0,6399	0,0619	0,3343	0,4118	0,7490	0,3769	1,1946	1,5941
ACT_M_98	-0,8063	-0,5984	-1,3542	0,7734	-1,5022	-0,9182	-0,3141	0,2679	1,0978	-0,3335	0,3837	0,2580	0,1597	0,2694	1,3741	0,9032
ACT_F_98	-1,7368	-1,0451	-0,9520	0,3452	-1,6685	-0,4443	-0,4860	0,2495	0,3630	0,2858	0,2687	0,5091	1,0402	0,4021	0,9675	1,8198
DEP_RT98	2,3089	0,8379	0,5457	-0,4980	1,7573	0,2995	0,1891	-0,3678	-0,2208	-0,2344	-0,3573	-0,7670	-1,1023	-0,3487	-0,8013	-0,9107
UN_T_98	2,5862	0,2150	-0,2409	-0,7029	1,1526	0,4505	-0,3013	-0,3911	-0,9476	0,2862	-0,8553	0,3469	1,7391	-0,4347	-0,9164	0,0137
UN_M_98	2,3184	-0,3352	-0,6382	-0,6341	0,7168	0,3814	-0,2665	-0,1340	-1,1825	0,3673	-0,8254	0,8062	1,8110	-0,0748	-0,8650	0,6058
UN_F_98	2,7170	0,6964	0,0937	-0,6932	1,4732	0,4249	-0,3059	-0,5554	-0,6623	0,0937	-0,7878	-0,0491	1,3357	-0,6886	-0,8455	-0,4312
EMP_AGR	0,5805	3,4378	0,0371	-0,3066	0,6919	-0,1139	-0,4604	-0,6602	2,2211	-0,0886	0,1037	-0,8903	-0,3692	-0,6053	-0,4937	-0,4564
EMP_IND	-1,0028	-0,6405	0,6563	1,4484	0,3348	-0,7091	0,9223	-0,4395	0,2106	0,3321	-0,3715	-1,5913	0,7781	0,1263	-0,6257	-0,5895
EMP_SER	0,4210	-2,0807	-0,5302	-0,9242	-0,7708	0,7109	-0,3685	0,8995	-1,8378	-0,1830	0,1924	2,0377	-0,3192	0,3338	0,7479	0,8690
GDPECUHA	-1,1536	-1,4881	-0,0962	0,9275	-0,8909	-0,1864	0,6348	2,8308	-1,5969	0,0554	-0,1372	2,1878	-0,6758	-0,5930	0,1487	0,9312
GDPPPSHA	-1,1077	-1,3445	0,3466	0,5709	-0,7446	-0,2475	0,5423	2,7149	-1,3410	-0,1562	-0,1465	2,7254	-1,2305	-0,2903	0,3738	0,3274

6. REFERENCIAS.

- BISHOP, C.M.**(1995): “*Neural Networks for Pattern Recognition*”. Ed. Clarendon Press, Oxford.
- HAYKIN, S.** (1994): “*Neural Networks. A Comprehensive Foundation*”. Ed. Prentice Hall, New Jersey.
- HILERA, J.R. & MARTÍNEZ. V.J.** (1995): “*Redes Neuronales Artificiales. Fundamentos, modelos y aplicaciones*”. Ed. RA-MA.
- KOHONEN, T.** (1982): “Self-organized formation of topologically correct feature maps”. *Biological Cybernetics* 43:59 - 69.
- KOHONEN, T.** (1989): “*Self-Organization and Associative Memory*”. 3th ed. Ed. Springer-Verlag.
- KOHONEN, T.** (1990): “The Self-Organizing Map”. *Proceedings of the IEEE* 78, 1464-1480.
- KOHONEN, T.**(1997): *Self-Organizing Maps* (Springer, Berlin, Heidelberg 1995). 2nd ed. 1997.
- KRAAIJVELD, M. A. ; MAO, J. & JAIN, A. K.:** *In Proc. IJICPR, Int. Conf. On Pattern Recognition* (IEEE Comput. Soc. Press, los Alamitos, CA 1992).
- MARTÍN DEL BRÍO, B. & SANZ MOLINA, A.** (1997): “*Redes Neuronales y Sistemas Borrosos*”. Ed. RA-MA.
- PERALTA, M. J., RUA, A., FERNÁNDEZ, L. Y BORRAS, F.** (2000): “*Tipología socioeconómica de las regiones europeas. Comparativa estadística*”. Instituto de Estadística de la Comunidad de Madrid. Colección: Metodología e Infraestructura estadística. Madrid.
- REFENES, P.** (editor) (1995): “*Neural Networks in the Capital Markets*”. Ed. Wiley.
- REGIONS: ANNUAIRE STATISTIQUE 2000.** Office des Publications Officielles des Communautés Européennes, 2000. Luxembourg.