Metodología de ayuda a la decisión mediante SIG e Inteligencia Artificial: *aplicación en la caracterización demográfica de Andalucía a partir de su residencia*

Methodology of Decision Support through GIS and Artificial Intelligence: *Implementation for Demographic Characterization of Andalusia based on Dwelling*

Autores:
Francisco Javier Abarca-Alvarez
fcoabarca@ugr.es
Francisco Sergio Campos-Sánchez
scampos@ugr.es
Rafael Reinoso-Bellido
rafaelreinoso@ugr.es

Departamento de Urbanística y Ordenación del Territorio
Universidad de Granada

España

Recibido: 16 Abr 2017
Aceptado: 25 Jun 2017

Resumen

Los Sistemas de Información Geográfica (SIG) han sido ampliamente utilizados para el almacenamiento y gestión de la información territorial, mostrándose especialmente útiles para el análisis y para la verificación de hipótesis previamente formuladas y con componentes espaciales relevantes. Existen metodologías heurísticas que en contextos como los actuales, de sobre-abundancia de datos, permiten evidenciar sus coherencias, sin requerir necesariamente hipótesis o formulaciones previas para generar conocimiento. Se propone el uso combinado de (i) técnicas procedentes de la Inteligencia Artificial, como son las Redes Neuronales Artificiales (ANN) del tipo Mapa Auto-organizado (SOM), que han demostrado ser muy eficaces y robustas clasificando y caracterizando perfiles en los datos; integradas con (ii) técnicas de *Machine Learning* como son los árboles de decisión, singularmente funcionales en la creación de modelos predictivos e interpretables para formular hipótesis explicativas de los perfiles anteriores a partir de otras variables diferenciadas. La investigación plantea combinar SIG, SOM y árboles de decisión para la construcción de modelos explicativos de los perfiles demográficos y sociales de Andalucía, a partir de datos de bajo coste sobre la dimensión residencial. Se verifica la viabilidad de tales modelos predictivos y su alto valor para la comprensión y para la toma de decisiones sobre tales territorios.

Palabras clave: árbol de decisión SIG, DSS, mapa auto-organizado.

Abstract:

Geographic Information Systems (GIS) have been widely used for the storage and management of territorial information, being especially useful for the analysis and verification of previously formulated hypotheses and coexisting with relevant spatial components. There are heuristic methodologies that, in contexts such as the present one, of data over-abundance, allow showing their coherence, not necessarily requiring hypotheses or previous formulations to generate knowledge. The combined use of (i) Artificial Intelligence techniques such as the Artificial Neural Network (ANN), namely the Self-Organized Maps (SOM), is proposed. They are very effective and robust by classifying and characterizing profiles in the data. They interact with (ii) machine learning techniques such as decision trees, which are singularly functional in the creation of predictive and interpretable models, with the intention of formulating explanatory hypotheses of the previous profiles, working with other different variables. The research proposes the combination of GIS, SOM and decision trees for the construction of explanatory models of the demographic and social profiles of Andalusia, based on low cost data on the residential dimension. The feasibility of such predictive models and their great value for understanding and as decision support on such territories are evaluated satisfactorily.

Keywords: GIS, decision tree, DSS, self-organizing map, SOM.

Estoa N°11 / Vol. 6 / Julio – Diciembre 2017
ISSN: 1390-7263 e-ISSN: 1390-9274
DOI: 10.18537/est.v006.n011.a03

1. Introducción

Son muy numerosas las experiencias en las que los Sistemas de Información Geográfica (SIG)[1] han sido usados por los gobiernos, investigadores y por las empresas como herramienta de decisión en las que alcanza cierta repercusión e influencia la dimensión espacial (Jarupathirun & Zahedi, 2005, p. 151). Los SIG se desarrollaron originalmente a finales de la década de 1960, incorporados en ese momento en muy pocos departamentos de urbanismo por el elevado coste, siendo en los años ochenta cuando se experimentó un incremento notable en su implantación, tanto en Europa como en EE.UU. (Yeh, 2005). Será a partir de 1990 hasta nuestros días, cuando desarrolle su mayor apogeo y uso por gran número de urbanistas como sistema capaz de integrar datos de diversas fuentes para proporcionar información necesaria para la toma de decisiones en la planificación urbana (Yeh, 2005). Algunos de estos sistemas con los que se integran los SIG son los que se han venido a llamar Sistemas de Ayuda a la Decisión,[2] especialmente en aquellos entornos en los que el análisis de las implicaciones espaciales o territoriales es singularmente relevante. Estos sistemas emergen en el ámbito del Urbanismo en la década de 1990, cuando las metodologías de la planificación pasan de dar por hecho el que el urbanista debe realizar los diseños y los planes para la gente, a una situación en las que ambos –ciudadanos y urbanistas– se convierten en actores importantes de la planificación (Ayedi, 1998).

Los Sistemas de Ayuda a la Decisión son instrumentos o "vehículos" que se han verificado eficaces para la incorporación e integración de realidades y problemas complejos así como para el apoyo de determinadas decisiones; se estima, tradicionalmente, crucial una perfecta definición de lo que se desarrolla y el porqué del mismo (Keen, 1987), evitándose en lo posible la improvisación, la indeterminación y sin prestar especial consideración a la flexibilidad y resiliencia del propio sistema.

Sin embargo la realidad informática en la que se apoyan estos sistemas está cambiando muy rápidamente en los últimos años, con una proliferación de datos e información espacial de acceso libre, intensificándose la idea de campo de investigación propio en torno a los SIG. Este campo ha sido nombrado como GIScience[3] (Goodchild, 2010), asumiendo un papel principal dentro de las Ciencias Sociales para el análisis y comprensión de la información a distintas escalas, pero con una naturaleza eminentemente espacial (Juanes Notario, 2014). Este campo de investigación se apoya en la idea de una nueva Geografía Cuantitativa (Buzai, 2007) de orden básicamente espacial y tendente hacia la gestión y planificación territorial, enfocada en mejorar la calidad de vida de la población.

Según Gustavo D. Buzai (2015), los SIG han pasado del énfasis en la S (Sig) por los problemas computacionales (décadas 1960-1970), pasando a dedicarse a la I (sIg) por el interés en la información (décadas 1980-1990), para finalmente a partir de 2000 enfocarse en la G (siG), por una necesidad de interpretación geográfica, materializándose la "sociedad de la información geográfica" y abriéndose una nueva etapa para la historia de la Geografía (Buzai, 2015). En este contexto los Sistemas de Ayuda a la Decisión se sitúan en el límite de la nueva Geografía Cuantitativa con la Geografía Humana, en cuanto que opera en el campo de las preguntas de la realidad y próximo al campo de las políticas, de sus lenguajes y de sus decisiones.

En este contexto de la nueva Geografía Cuantitativa, la granularidad de los datos geográficos[4] se está elevando al extremo, llegando a una auténtica "n-dimensionalidad" de los datos (André Skupin & Agarwal, 2008).

De este modo Pragya Agarwal y André Skupin han remarcado que el análisis estadístico tradicional se está enfocando en las problemáticas de la autocorrelación espacial, quedando otros múltiples ámbitos totalmente por explorar. Algunos de esos espacios están siendo abordados por enfoques emergentes como son la Inteligencia Artificial o las redes neuronales artificiales, el aprendizaje automático (*Machine Learning*) o de forma específica por la Geo-computación.

Estas nuevas técnicas y enfoques están propiciando un cambio de paradigma en los

[1] A pesar de la enorme difusión en textos en español del acrónimo inglés GIS, se prefiere mantener el uso de SIG en el texto.
[2] En la profusa bibliografía anglosajona sobre los Sistemas de Ayuda a la Decisión el término es acuñado como *Decision Support System* y con acrónimo DSS.
[3] El término acuñado en inglés es *GIScience*.
[4] Se entiende por granularidad de los datos el nivel de detalle y su definición a distintas escalas. En la actualidad se está elevando exponencialmente, tanto a nivel de su geometría espacial (más propiedades y un tamaño más fino de celda), como a nivel de sus atributos (incrementándose los mismos y a su vez la medida de sus valores).

Sistemas de Ayuda a la Decisión, considerándose que en la actualidad pueden ser útiles para la comprensión de la realidad, detección de sus problemas y, en definitiva, la formulación de nuevas hipótesis y no solo como instrumento para verificar aquellas previamente establecidas.

Como objetivo principal la investigación se propone evaluar la viabilidad y el interés de la construcción de modelos que usualmente conllevarían un alto coste aparejado, usando en su lugar información menos costosa, interpretándose mediante técnicas de Inteligencia Artificial y *Machine Learning* apoyados en información SIG. De forma específica para su validación se aplicará en la caracterización poblacional de la región española de Andalucía a partir de información residencial en la que reside la misma.

Para alcanzar tal objetivo la investigación se enmarca en los paradigmas de los Sistemas de Ayuda a la Decisión orientados al conocimiento y orientados a los modelos[5] (Power, Sharda, & Burstein, 2015). El primer paradigma se enfoca en la construcción de un sistema de descubrimiento de conocimiento basado en bases de datos institucionales sobre las cualidades demográficas y sociales de Andalucía (Fase de modelado 1 de la metodología); y, en el segundo paradigma, (Fase de modelado 2 de la metodología) se enfatiza el acceso, la manipulación y creación de un modelo cuantitativo de la realidad social orientado a proporcionar apoyo a la decisión, elaborado a partir de la realidad residencial de los territorios en estudio. Tal y como describe (Power et al., 2015) los Sistemas de Ayuda a la Decisión usan datos y parámetros proporcionados por los agentes de decisión para ayudarles a analizar una situación, aunque no tienen por qué ser datos masivos.

En nuestro caso el modelo sí se construye con datos masivos (tanto demográficos como residenciales), pero puede ser usado para la toma de decisiones con una información muy limitada, incluso escasa.

La investigación, tal y como se ha avanzado, consiste en la construcción de dos modelos para el cual se deben seguir dos fases metodológicas fundamentales:

La Fase de modelado 1 consiste en la construcción de un Sistema de Descubrimiento de Conocimiento a partir de Información y Bases de Datos.[6] En origen estos sistemas no se pensaron como una disciplina autónoma, sino más bien como un método de inteligencia para decisiones a nivel productivo y medioambiental (Longbing Cao, 2009), para pasar recientemente a conformarse como ciencia con identidad propia (*Data Science*).

Esta fase de Conocimiento basado en Datos, se materializa en la investigación mediante técnicas de lo que se vienen a conocer como Mapas Auto-organizados,[7] en adelante SOM. Fueron propuestos inicialmente por Teuvo Kohonen (Kohonen, 1990, 1998; Ritter & Kohonen, 1989). La metodología SOM es una técnica de descubrimiento de conocimiento o de minería de datos consistente en una red neuronal artificial.[8] Procede del campo de conocimiento de la Inteligencia Artificial, habiéndose mostrado muy eficaz y robusta en numerosas disciplinas, presentando diversas capacidades entre las que podemos destacar inicialmente dos: (i) es capaz de mostrar y visualizar la información de partida de forma clara y ordenada; (ii) permite clasificar y, por tanto, etiquetar los sujetos en estudio en clases que no requieren su definición, caracterización o etiquetado nominativo previo (aprendizaje no supervisado).

Frente a otras metodologías de descubrimiento de patrones, como por ejemplo el análisis clúster, la metodología SOM, tiene la ventaja de (i) permitir visualizar un gran conjunto de datos estadísticos (Kaski & Kohonen, 1996), (ii) mostrar las relaciones topológicas de similitud o de diferencia entre los sujetos en estudio, (iii) ser interpretables gráficamente y (iv) constituir por sí mismo un sistema de conocimiento de ayuda a la decisión para el análisis y visualización de indicadores estadísticos (Kaski & Kohonen, 1996).

Como resultado de esta Fase de modelado 1 obtenemos por un lado el etiquetado a modo de clases o perfiles de los diferentes fragmentos de territorio andaluz estudiado, atendiendo al análisis multi-variable de los atributos demográficos y sociales estudiados. Por otro lado, obtenemos un sistema de análisis e interpretación de las clases obtenidas, facilitado por la metodología SOM y materializado en cartografías temáticas de los diferentes atributos incluidos en la red neuronal y en diferentes tablas y datos estadísticos que permiten conocer las características diferenciadoras de cada perfil.

La Fase de modelado 2 trata de materializar, mediante un proceso de *Machine Learning*, una serie de reglas que permitan predecir las clases o perfiles que se determinaron en la Fase de modelado 1 a partir de atributos sobre la realidad residencial de los territorios en estudio. Estas variables residenciales no se tuvieron en consideración en la creación y constitución de la red neuronal SOM ni, en consecuencia, pudieron afectar o correlacionarse en la definición de los perfiles obtenidos

[5] Se suele considerar que existen cinco tipos de Sistemas de Ayuda a la Decisión (Power, Sharda, & Burstein, 2015): (i) orientados a la comunicación; (ii) orientado a los datos; (iii) orientados a los documentos; (iv) orientados al conocimiento; y (v) orientados a los modelos.

[6] Los Sistemas de Descubrimiento de Conocimiento a partir de Información y Bases de Datos se suelen nombrar en la bibliografía anglosajona como *Knowledge Discovery in Databases* y con el acrónimo KDD.

[7] Los Mapas Auto-organizados se conocen como *Self-Organizing Maps* y su acrónimo SOM. En ciertas ocasiones se pueden encontrar nombrados como *Self-Organizing Feature Maps* (SOFM).

[8] En inglés A*rtificial Neural Network* y muy habitualmente por su acrónimo ANN.

por aquella. Se realiza una aproximación al problema de aprendizaje mediante el paradigma "divide y vencerás" que, al realizarse sobre un conjunto de instancias independientes, conduce naturalmente a un estilo de representación llamado "árbol de decisión" (Witten, Frank, & Hall, 2011, p. 64). En cada nodo del árbol interviene un atributo en particular, comparándose normalmente cada instancia del atributo con el valor de una constante, generándose normalmente dos ramas atendiendo a las instancias que cumplen o no tal regla. El árbol supone una representación asequible para interpretar y usar en la predicción de la realidad demográfica y social de un territorio y, en consecuencia, útil para la toma de decisiones sobre el mismo; usa para ello, una información limitada y generalmente de fácil y económica obtención sobre la realidad residencial del lugar en estudio.

Asimismo al evaluar el "valor" de los perfiles alcanzados en la Fase 1, en su caracterización espacial mediante SIG, se verifica la utilidad de la metodología propuesta.

2. Estado del arte

En este apartado se describirán las principales experiencias y referentes que conforman el estado del arte de los dos ámbitos de conocimiento vinculados con la propuesta metodológica realizada, enfocándose especialmente en las disciplinas vinculadas al urbanismo, la planificación urbana y territorial, así como a la geografía humana y urbana: (i) experiencias de descubrimiento de conocimiento mediante aprendizaje no supervisado con Mapas Auto-organizados y (ii) experiencias de construcción mediante aprendizaje supervisado con árboles de decisión, subrayando singularmente aquellas que se combinan con los resultados obtenidos a partir de metodologías SOM.

2.1 Descubrimiento de conocimiento y clasificación mediante aprendizaje no supervisado. Los Mapas Auto-organizados (SOM)

Las Redes Neuronales Artificiales –a las que pertenecen los Mapas Auto-organizados– son un campo clásicamente integrado en la Neurociencia o Ciencias de la Inteligencia Artificial. Son una categoría de los métodos del *Machine Learning* que han sido ampliamente usados recientemente en problemas de predicción, clasificación y reconocimiento de patrones (Kauko, 2005). Como principales aplicaciones prácticas de las Redes Neuronales destacan (Kohonen, 1995, p. 219): i) la monitorización y control de la instrumentación industrial, ii) aplicaciones médicas como el diagnóstico, prótesis y modelado y iii) la distribución de recursos de redes de telecomunicaciones.

Nosotros usaremos las Redes Neuronales tipo SOM que permiten a partir de información desordenada, analizar y crear perfiles, proporcionando patrones visuales y formando un paisaje del fenómeno descrito por los datos (Kohonen, 1995). Si bien el algoritmo SOM se creó inicialmente para la visualización de relaciones no lineares de datos multidimensionales, rápidamente se evidenciaron útiles para la visualización de relaciones abstractas, como por ejemplo roles contextuales (Kohonen, 1995, p. 219). Los SOM se mostrarían extremadamente útiles en múltiples campos del conocimiento y las ciencias aplicadas. Según el autor de los SOM, Teuvo Kohonen (1982), como aplicaciones de carácter general se pueden destacar multitud de aplicaciones (Kohonen, 1995).[9]

Se puede observar que desde el origen de los SOM en 1982 hasta el año 1995, son prácticamente inexistentes las experiencias con la metodología SOM en campos como las Ciencias Sociales, Geografía, Urbanismo y Ordenación del Territorio y, en general, cualquier investigación que precisara SIG, a pesar de que como se describió en la Introducción tales sistemas de información iniciaran su andadura varias décadas antes. Será necesario que el propio Kohonen demuestre las capacidades de los SOM en el campo de la Geografía Humana (Kaski & Kohonen, 1996), para que estas disciplinas se aproximen, lentamente, a esta metodología. En este trabajo seminal, Samuel Kaski y Teuvo Kohonen utilizan el método SOM para representar un conjunto de datos complejos sobre la distribución de la riqueza y la pobreza en el mundo, de una manera en la que pueden ser mostradas y analizadas las similitudes y diferencias entre los diversos países. Será a partir de este trabajo y especialmente en los últimos 10-15 años cuando se experimenta cierta eclosión en el uso de la metodología SOM, todavía muy distante del sobresaliente uso en otros ámbitos.

A continuación se observarán investigaciones con los SOM vinculadas a las disciplinas próximas al SIG como pueden ser la Urbanística, la Geografía, las Ciencias Sociales, etc., anotando si se enfocan en algunas de los principales campos de experiencia y capacidad de los SOM a la hora de manejar la información. Esto es la

[9] Según Teuvo Kohonen, hasta 1995 se pueden encontrar multitud de aplicaciones a los SOM: (i) el pre-procesado de patrones ópticos, análisis de imágenes, visión artificial, reconocimiento de caracteres, y aplicaciones médicas con procesado de imágenes (ii) el pre-procesado acústico, estudios acústicos y musicales, y procesado de señales y medida de radar, (iii) proceso y monitorización máquina, medidas industriales, control de procesos y otras medidas de la realidad, (iv) diagnóstico de voz, (v) transcripción de voz continua, reconocimiento y análisis del habla, y problemas de Lingüística e Inteligencia Artificial, (vi) análisis de texturas, (vii) mapas contextuales, (viii) organización de grandes ficheros de documentos y procesado de datos (análisis exploratorio de datos financieros, etc.), (ix) control de brazos robóticos, (x) telecomunicaciones, (xi) el SOM como un estimador, (xii) problemas químicos (clasificación o extracción de características de los cromosomas, etc.), (xiii) problemas físicos (visualización de espectros o clasificación de seísmos, etc.), (xiv) problemas matemáticos (optimización, etc.), (xv) diseño de circuitos, o xvi) investigación neuropsicológica.

Representación, la Clasificación, la Caracterización y la Toma de decisiones:

- En el campo de la interpretación de imágenes vinculadas a la geografía o SIG podemos destacar ciertos trabajos con foco en la clasificación, como por ejemplo el uso del SOM para el análisis y clasificación de imágenes de satélite multiespectrales (con más de 200 bandas diferentes) de la superficie de la Tierra y de otros planetas (Villmann, Merényi, & Hammer, 2003), o la aplicación del SOM para la clasificación de suelos y minerales usando imágenes de radio espectral y SIG, permitiendo transformar las cartografías de un volcán (Tayebi, Hashemi Tangestani, & Vincent, 2014).

- En el campo del análisis de la movilidad destaca una investigación en la que se usó el SOM para el análisis de interacciones espaciales, singularmente enfocada hacia la comprensión gráfica y obtención de patrones en las estructuras de transporte aéreo de EE.UU. (Yan & Thill, 2009).

- Por otro lado podemos destacar en el campo de la gestión de catástrofes el uso del SOM como herramienta para clasificar y reconocer patrones en bases de datos sobre epidemiología vinculadas con información espacial para ser representada en un sistema SIG (Zhang, Shi, & Zhang, 2009), o para la toma de decisiones rápida mediante SIG y SOM a partir de la extracción de información precisa de los deslizamientos tras un terremoto (Lin, 2008).

- En el campo del estudio del medioambiente, de la salud y de la calidad de vida podemos encontrar trabajos enfocados en la representación mediante SOM de la distribución del riesgo ecológico por contaminación (Faggiano, de Zwart, García-Berthou, Lek, & Gevrey, 2010); trabajos orientados hacia la clasificación: uso del SOM para la localización de patrones de contaminación por pesticidas en la cuenca del río Asour-Garonne en Francia (Faggiano et al., 2010), para determinar el modelo de localización de plantas para el tratamiento de residuos de la madera (Gomes, Ribeiro, & Lobo, 2007); podemos asimismo observar el uso del SOM y del SIG para la clasificación de la salud comunitaria a partir de variables de las condiciones ambientales (Basara & Yuan, 2008), o la creación de un modelo que evalúa el nivel de calidad ambiental de los suelos, caracterizándolos a partir de concentraciones de elementos mediante SOM y visualizándolo con SIG (C. Yang, Guo, Wu, Zhou, & Yue, 2014). Finalmente podemos encontrar un trabajo enfocado en la toma de decisiones mediante el análisis empírico multidimensional y espacio-temporal de las tendencias de la calidad de vida de los barrios de Charlotte (EE.UU.) mediante la representación gráfica de los SOM (Delmelle, Thill, Furuseth, & Ludden, 2012).

- Con vinculación al ámbito de los estudios demográficos, Ciencias Sociales y Geografía describirá Takatsuka (2001, p. 24) que "El Self-Organizing Map es uno de las más modernas herramientas que los investigadores han encontrado útiles en el análisis de bases de datos multivariables tales como los datos atmosféricos y demográficos". En estos ámbitos del conocimiento, con una vertiente más social y demográfica destacan ciertas investigaciones con un uso del SOM enfocado en la representación de datos, como por ejemplo la visualización conjuntamente con SIG de los cambios demográficos de los condados de Texas (EE.UU.) a lo largo del tiempo mediante SOM (Andre Skupin & Hagelman, 2005) de patrones espacio-temporales de variables geográficas de EE.UU (Guo, Chen, MacEachren, & Liao, 2006), o el uso del SOM para la realización de una representación holística alternativa y complementaria a la representación espacial propia de los SIG, en la que se da simultanea información de 69 atributos censales de EEUU, con información sobre el clima, topografía, suelo, geología, usos de suelo y población (André Skupin & Esperbé, 2011). Con un enfoque hacia el SOM como clasificador podemos encontrar el uso simultaneado con SIG de una variante de los SOM (fuzzy) para creación de regiones demográficas homogéneas a partir de datos del censo del municipio de Atenas (Hatzichristos, 2004), el uso de SOM para caracterizar barrios mediante el etiquetado de secciones censales de Nueva York a partir de 79 atributos geo-demográficos (Spielman & Thill, 2008), o el uso de los SOM como una metodología para el *Data-mining* urbano en la que se realiza una clasificación no supervisada de datos geoespaciales de comunidades alemanas en cuanto a población, migración, impuestos, residencia, empleo y transporte (Behnisch & Ultsch, 2009).

- Por último en el campo más próximo a la investigación que se propone, se pueden encontrar ciertos ejemplos relevantes del uso directo de las metodologías SOM en el ámbito del Urbanismo, la Planificación y Ordenación del Territorio, como por ejemplo: Investigaciones enfocadas en las capacidades de obtener conocimiento mediante la representación de las dinámicas temporales de la ciudad de Harrisburg en EE.UU. (Takatsuka, 2001), o la representación semántica y caracterización de barrios ejemplares de la historia reciente del urbanismo (AUTOR 1 & Osuna Pérez, 2013).

En el marco urbanístico han sido frecuentes los análisis multicriterio a modo de análisis multicapa (Feng & Xu, 1999) con capacidad para la clasificación. Utilizando SOM encontraremos asimismo algunos ejemplos como la identificación y caracterización de *urban sprawl* de Milán (Diappi, Bolchim, & Buscema, 2004), el análisis mediante SOM del mercado residencial a partir de variables de precios, cualidades y características de las viviendas, densidad, habitantes, etc., de Finlandia, Hungría y de Países Bajos como método inductivo de descubrimiento de similitudes y diferencias entre ellos (Kauko, 2005). También podemos destacar la utilización del SOM como clasificador para la detección de edificios mediante tecnología lídar e imágenes multiespectrales y atributos auxiliares (Salah, Trinder, & Shaker, 2009), la clasificación de los tejidos urbanos a partir de indicadores morfológicos relacionadas con la huella de las edificaciones (Hamaina, Leduc, & Moreau, 2012), o la clasificación taxonométrica de las inmigraciones turísticas de los tejidos urbanos de Andalucía a partir de

las cualidades de sus asentamientos AUTOR 1, AUTOR 2, & Osuna-Perez, 2015). Finalmente podemos destacar un enfoque hacia la capacidad de los SOM para facilitar la toma de decisiones en diversos trabajos, como por ejemplo en la integración del SIG y de las técnicas SOM, nombrada específicamente SOFM (*Self-organizing Feature Maps*) para la creación de un modelo difuso para la clasificación de suelos en la provincia China de Zhejiang, con la intención de ser aplicado como parte de un análisis de apoyo a la decisión (H. Yang, Hu, qi Deng, Tian, & Li, 2004), en la caracterización de tejidos urbanos del centro histórico de Santa Fe (España), para la creación de una ordenanza urbanística (AUTOR 1 & Fernandez-Avidad, 2010), o en la combinación de metodologías de aprendizaje supervisado y no supervisado tipo SOM orientado a la investigación del mercado nocturno callejero de Taiwan, usando para ello información espacial SIG (Wu & Hsiao, 2015).

2.2 Construcción de modelos predictivos mediante aprendizaje supervisado. Árboles de decisión

Los árboles de decisión son técnicas y métodos de aprendizaje de modelos comprensibles y proposicionales, entendiendo que son (i) modelos en cuanto que construyen una hipótesis o representación de la regularidad de los datos; (ii) comprensibles por expresar de manera simbólica, en forma de un conjunto de condiciones; y, (iii) proposicionales al establecerse en su construcción reglas "atributo-valor" en las que las condiciones se expresan sobre el valor de un único atributo (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, p. 281).

Son muy numerosas las variantes de algoritmos de árboles de decisión. Destacaremos únicamente algunos de ellos. Como antecedentes históricos de los árboles de decisión más usados en la actualidad podemos encontrar los algoritmos CHAID,[10] CART,[11] ID3 o C4.5.[12] Entre los algoritmos que generan árboles de decisión destacaremos por último los métodos de particionado recursivo que se han vuelto en los últimos años muy populares y ampliamente usados para la regresión no paramétrica y para la clasificación en muchos campos científicos (Strobl, Malley, & Tutz, 2010).

A continuación se muestran algunos de los escasos ejemplos de investigaciones que implementan el mismo concepto metodológico que se usa en este artículo: Implementación de un modelo híbrido: sobre el SOM se aplica un árbol de decisión; es decir, el árbol de decisión utiliza la clasificación proporcionada por el SOM como información a predecir.

Como ejemplos de este enfoque de modelo híbrido o dual en los que se combinan técnicas no supervisadas (SOM) y minería de datos supervisada (árboles de decisión) podemos encontrar, algunos trabajos metodológicos como Astudillo & John Oommen (2011); Astudillo & Oommen (2013) o relacionados con la Biología y la Medicina, como por ejemplo un estudio para la reducción del coste del diagnóstico de tiroides (Kinaci & Yucebas, 2015), el análisis de componentes principales para la identificación de bacterias (Simmuteit, Schleif, Villmann, & Kostrzewa, 2009), o un trabajo de minería en datos biológicos (Z. R. Yang & Chou, 2003). En el campo de la Ingeniería podemos encontrar el análisis de la carga eléctrica dinámica (Voumvoulakis, Gavoyiannis, & Hatziargyriou, 2006), o la selección de variables para agrupar muestras de viario (Gómez-Carracedo et al., 2010). Con vinculación a la Economía y empresa hallamos un análisis de la cartera de clientes (Yao, Holmbom, Eklund, & Back, 2010), o el descubrimiento de las preferencias en el comercio de acciones (Tsai, Lin, & Wang, 2009), todos ellos enfocados en este planteamiento de SOM + árboles de decisión. Finalmente con cierta similitud en el enfoque con nuestra investigación podemos encontrar la selección de propiedades en el análisis de datos censales mediante SOM y árboles de decisión (Shanmuganathan & Li, 2016).

3. Materiales y Métodos

3.1. Objetivos de partida

Como objetivo principal de la investigación se propone evaluar la viabilidad y el interés de una alternativa para la construcción de modelos territoriales sobre dimensiones que requieren de datos de alto coste. Para elaborar tal modelo se propone con un menor número de datos, identificar perfiles y conectarlos con información de otras dimensiones que conlleven un menor coste aparejado.

A modo de validación, en nuestro caso concreto, la metodología se va a aplicar para crear un modelo que explique la realidad demográfica de Andalucía, a partir de la realidad residencial, dimensión esta última que está compuesta por datos habitualmente de más fácil acceso y menor coste.

[10] CHAID significa detección automática de interacciones mediante Chi-cuadrado del inglés *Chi-squared Automatic interaction Detection*. Es original de Kass (1980) basado en el test de significancia de Bonferroni y caracterizado por su facilidad de interpretación gráfica y por no ser paramétrico. CHAID es una técnica multivariante en la que hay una única variable a explicar y varias explicativas, en las que se identifican distintas categorías para servir de división en cada rama, seleccionándose para cada una de ellas "la variable que discrimina más y las clases que combinadas proporcionan la mayor discriminación en la variable dependiente objeto de análisis", es decir "detecta las interacciones que más discriminan" (Luque Martínez, 2000, pp. 379-380).

[11] CART significa árboles de clasificación y regresión, del inglés *Classification and Regression Trees*, fue propuesto por Breiman, Friedman, Olshen y Stone (1984).

[12] ID3 dio lugar posteriormente al algoritmo C4.5, siendo ambos muy populares por su enfoque no paramétrico y por su interpretabilidad que los caracteriza (Quinlan, 1986).

3.2. Metodología de investigación

Para la óptima comprensión y para la obtención de los mejores resultados, se siguen las fases definidas por Mark S. Silver (2008): (i) información y funciones de procesado, (ii) conjuntos de datos, (iii) modelos y (iv) representaciones visuales.

(i) Información y funciones de procesado:

La información utilizada en la investigación procede en su totalidad de la información pública del Censo de Población del año 2001 de Andalucía, obtenidos por el Instituto de Estadística y Cartografía de Andalucía (IECA). Se evitó la utilización de fuentes más actualizadas como es el Censo de 2011, por tratarse éste de datos que en gran medida procede de interpolaciones a partir de una reducida muestra de estudio. En cualquier caso se considera que este periodo a causa de la crisis inmobiliaria, no ha producido grandes transformaciones sobre los datos. Sobre esta información se ha realizado una intensa preparación de datos, consistente fundamentalmente en la integración y limpieza de datos, transformación de atributos mediante la creación de indicadores agregados que tienen la peculiaridad de resumir de forma objetiva y compacta las principales cualidades demográficas de mejor modo que los datos originales. Debido a la robustez de los SOM, no es necesario realizar su tipificación o normalización[13] previamente a la agregación e incorporación al modelo.

-Instancias: La unidad de territorio sobre la que se obtienen los datos, es la Sección Censal, alcanzándose la totalidad de las 5381 secciones censales[14] de Andalucía, formando parte del estudio la totalidad de la superficie y de población censada en la región andaluza. Por tanto, no se ha realizado ningún muestreo sino que se ha manejado la totalidad de la población en estudio.

-Atributos: Los atributos que se ha usado en la Fase de modelado 1 constan de diversas variables de la dimensión demográfica, social, laboral, de los equipamientos y de los servicios a los que se tienen acceso en cada sección censal (primera columna de la Tabla 1). Los atributos que se han usado en la Fase de modelado 2 están compuestos de variables de la dimensión residencial (primera columna de la Tabla 2). En la Tabla 1 y Tabla 2 se muestran asimismo ciertos datos de la estadística descriptiva de los mismos, como la media y la desviación estándar (columnas 2 y 3 de las Tablas 1 y 2).

(ii) Conjuntos de datos:

Se opera inicialmente con dos bases de datos desconectadas: una propia de la Fase de modelado 1, con dimensión principalmente demográfica y otra propia para la Fase de modelado 2, basada en la dimensión residencial. El funcionamiento será fundamentalmente independiente, conectándose únicamente tras el Modelado 1 para evaluar cómo se ajustan los perfiles demográficos obtenidos a los datos residenciales (columnas de perfiles de la Tabla 2. Obsérvese que los perfiles han sido obtenidos únicamente a partir de los atributos de la Tabla 1). Finalmente en la Fase de modelado 2 se integrarán en la construcción del árbol de decisión los perfiles obtenidos en la Fase de modelado 1 con los datos residenciales.

(iii) Modelos:

Debemos distinguir entre dos fases de modelado:

Fase de modelado 1: Modelo de clasificación y conocimiento; entre sus objetivos tenemos por un lado, la clasificación y etiquetado de los datos de la dimensión demográfica, social, laboral, de los equipamientos y de los servicios; y, por otro, la generación de conocimiento de la realidad andaluza de tales dimensiones, de los perfiles creados y de los perfiles creados en contraste con la dimensión residencial (que no constituyó parte del modelo creado de clasificación). En esta fase se usará una red neuronal artificial y específicamente un SOM. Esta metodología permite clasificar (propiamente segmentar) sin atribuir a priori una etiqueta con definiciones y significados previamente atribuidos, siendo un modo para reducir la enorme complejidad de los datos (Spielman & Thill, 2008).

El estado del arte demuestra que es muy frecuente el uso de los SOM como metodología de reducción y clasificación (Hamaina et al., 2012) y también para el etiquetado de entidades (Salah et al., 2009). Comparado con otros métodos de reducción de dimensiones como el PCA (análisis de componentes principales) o el MDS (escalado multidimensional), la capacidad de los SOM de preservar la topología de los datos hace que éste tenga un uso más eficiente del espacio disponible en la representación del mapa, con la consecuencia de una mayor distorsión en las distancias relativas (André Skupin & Agarwal, 2008, p. 7). Por otro lado el SOM presenta ventajas muy notables frente a otras técnicas o métodos. Los SOM son relativamente insensibles a los valores perdidos, tolerando a la vez datos con una distribución no normal (Zhang et al., 2009); esto le permite prescindir de verificaciones de difícil cumplimiento, haciéndolo válido para cualquier distribución de datos. Por otro lado como método de clusterización el SOM es más robusto que por ejemplo el K-means, aunque en su contra requiere mayor tiempo de computación (Bação, Lobo, & Painho, 2005; Gomes et al., 2007).

[13] Según el Demartines & Blayo (1992) la normalización no es necesaria cuando el espacio de entrada tiene una dimensión superior a 12, es decir más de 12 variables en estudio.

[14] Sección censal es la partición de los términos municipales del Censo en España. Suelen estar definidos mediante límites fácilmente reconocibles y habitualmente tienen un tamaño de población entre 1000 y 2500 residentes, salvo que el municipio completo tenga una población menor.

	Muestra completa		Perfil 1				Perfil 2				Perfil 3				Perfil 4				Perfil 5			
N. samples (%)	5381	1.00	2054		38.17%		1659		30.83%		1002		18.62%		550		10.22%		115		2.14%	
	Mean	(SD)	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES
2.1_EquipSaludPor1000Hab	1.39	6.16	2.08	9.10	***		0.72	1.30	***		0.85	1.78	***		1.56	2.45	ns		2.75	14.22	ns	
2.1_EquipEducacionPor1000Hab	1.14	2.36	1.18	3.04	ns		0.91	1.26	***		0.91	1.33	***		2.16	3.20	***	+	1.04	1.60	ns	
2.1_EquipBienestarPor1000Hab	0.83	1.68	0.87	1.42	ns		0.62	0.93	***		0.55	0.83	***		1.85	3.77	***	++	0.78	1.86	ns	
2.1_EquipCulDepPor1000Hab	0.72	1.47	0.63	1.27	**		0.56	0.93	***		0.48	0.82	***		1.94	3.07	***	+++	0.66	0.92	ns	
2.1_EquipamientosPor1000Hab	4.08	8.31	4.75	11.71	**		2.80	2.69	***		2.79	3.24	***		7.51	7.25	***	+	5.23	15.50	ns	
2.2_%NoAguaCorriente	0.91	2.84	0.55	1.96	***		0.69	1.44	***		0.66	1.80	***		3.16	6.47	***	++	1.90	4.59	*	+
2.2_%Gas	22.77	33.76	30.00	34.45	***	+	18.94	34.23	***		21.61	32.51	ns		12.31	28.29	***	-	9.12	18.27	***	-
2.2_%Telefono	86.96	17.48	94.65	9.13	***	+	83.08	17.59	***	-	90.51	12.48	***	+	63.45	23.69	***	---	87.41	16.92	ns	
2.5_%PocaLimpiezaCalles	35.10	19.82	43.75	16.23	***	+	28.69	20.49	***	-	35.95	15.78	ns		21.23	22.00	***	--	31.76	16.79	*	
2.6_%Delincuencia	25.32	23.67	42.70	23.28	***	++	11.27	13.90	***	--	24.03	17.43	*		4.94	10.62	***	---	26.54	17.59	ns	
3.1_Poblacion	1367	518.00	1272	390.50	***		1359	433.20	ns		1772	539.30	***	++	900.20	410.50	***	---	1904	934.60	***	+++
3.1_EdadMediaPoblacion	38.01	4.41	38.91	3.98	***	+	37.91	3.30	ns		33.87	3.37	***	---	42.74	4.02	***	+++	36.83	4.52	**	-
3.1_%Nacimientos	11.32	3.21	9.98	2.87	***	-	11.74	2.27	***		14.32	3.21	***	+++	9.54	2.78	***	--	11.46	2.44	ns	
3.2_PersonasPorEdificio	14.46	22.91	27.80	30.16	***	++	4.31	5.34	***	-	11.39	16.08	***		2.43	3.79	***	--	6.81	9.87	***	-
3.2_%Hogares1Adulto	18.80	7.55	18.86	7.15	ns		17.51	5.92	***		14.99	5.68	***	--	27.87	7.37	***	+++	26.51	10.23	***	+++
3.2_%Hogares1adultoYMenor	1.83	1.12	2.06	1.12	***	+	1.28	0.73	***	-	2.44	1.08	***	++	1.28	0.92	***	-	3.08	1.45	***	+++
3.2_%Hogares2adulto	41.24	6.75	37.47	5.22	***	--	42.88	5.47	***	+	46.53	7.70	***	++	41.01	5.11	ns		40.02	5.94	*	
3.2_%Hogares3adulto	18.41	3.36	19.20	3.34	***	+	19.17	2.87	***	+	16.94	3.04	***	-	16.53	3.13	***	--	14.92	4.10	***	---
3.2_%Hogares4adulto	19.71	6.64	22.41	6.50	***	+	19.16	6.29	***		19.10	5.21	***		13.32	4.70	***	---	15.47	6.71	***	--
3.2_Hogares	449.20	167.50	415.00	117.20	***	-	442.30	136.10	*		560.60	173.20	***	++	340.20	145.70	***	--	709.20	430.60	***	+++
3.2_HabitantesPorHogar	3.04	0.36	3.06	0.36	***		3.07	0.31	***		3.17	0.26	***	+	2.62	0.31	***	---	2.79	0.51	***	--
3.2_RatioEdifViviendasPorHogar	0.74	0.56	0.33	0.37	***	---	0.99	0.34	***	+	0.79	0.59	*		1.42	0.48	***	+++	0.98	0.75	**	+
3.3_%PoblacionArraigada	80.11	9.49	83.99	4.69	***	+	82.84	4.40	***	+	68.84	12.65	***	---	82.19	5.00	***	+	59.39	9.81	***	---
3.3_%PoblacionInmigranteProvincial	3.87	5.81	1.83	1.68	***	-	2.55	2.18	***	-	9.92	10.46	***	+++	3.83	3.52	ns		6.89	4.98	***	++
3.3_%PoblacionInmigranteRegional	1.35	1.28	1.35	1.06	ns		0.80	0.78	***	-	2.26	1.63	***	++	0.97	1.09	***	-	3.21	1.72	***	+++
3.3_%PoblacionInmigranteNacional	1.58	1.33	1.35	0.90	***		1.11	0.83	***	-	2.44	1.67	***	++	1.89	1.59	***	+	3.67	2.51	***	+++
3.3_%PoblacionInmigranteExtranjero	1.32	2.79	0.96	1.02	***		0.56	0.97	***	-	1.85	2.37	***		1.13	1.95	*		14.96	8.23	***	+++
3.3_%España	97.92	4.58	98.46	1.65	***		99.20	1.59	***	+	96.98	4.12	***	-	98.58	2.68	***		74.94	12.25	***	---
3.3_%EuropaUE	0.77	2.97	0.33	0.63	***		0.30	1.22	***		1.23	2.77	***		0.82	2.21	ns		11.14	13.37	***	+++
3.3_%EuropaNoUE	0.18	0.56	0.10	0.23	***		0.07	0.19	***		0.32	0.72	***	+	0.12	0.38	***		2.18	1.92	***	+++
3.3_%AmericaDelNorte	0.05	0.18	0.05	0.12	ns		0.01	0.06	***	-	0.08	0.17	***		0.02	0.08	***		0.50	0.86	***	+++
3.3_%AmericaCentral	0.04	0.09	0.05	0.11	***		0.01	0.05	***	-	0.05	0.08	***		0.01	0.05	***	-	0.13	0.20	***	+++
3.3_%AmericaDelSur	0.41	0.83	0.46	0.62	**		0.16	0.34	***	-	0.65	1.15	***	+	0.20	0.47	***	-	2.33	2.38	***	+++
3.3_%Asia	0.08	0.34	0.09	0.23	ns		0.02	0.08	***		0.13	0.39	***		0.02	0.12	***		0.88	1.47	***	+++
3.3_%Africa	0.54	2.01	0.46	0.86	***		0.23	0.53	***		0.56	1.09	ns		0.23	1.05	***		7.87	10.08	***	+++
3.3_%Oceania	0.00	0.02	0.00	0.02	***		0.00	0.03	ns		0.00	0.01	ns		0.00	0.01	***		0.04	0.08	***	+++
3.3_%Apatridas	0.00	0.01	0.00	0.00	***		0.00	0.01	ns		0.00	0.01	ns		0.00	0.00	***		0.00	0.00	ns	
3.4_%TrabajaEnProvincia	6.43	5.48	4.19	2.59	***	-	5.87	3.99	***		11.62	8.31	***	+++	6.81	4.43	*		7.40	5.68	ns	
3.4_%TrabajaEnRegion	0.87	1.00	0.71	0.53	***		1.19	1.49	***	+	0.66	0.51	***	-	0.98	0.95	**		0.30	0.61	***	--
3.4_%TrabajaEnEspaña	0.47	0.57	0.38	0.30	***		0.49	0.48	ns		0.35	0.32	***	-	1.00	1.26	***	+++	0.23	0.35	***	-
3.4_%TrabajaOtroPais	0.12	0.29	0.14	0.32	***		0.08	0.18	***		0.10	0.13	***		0.16	0.53	ns		0.23	0.29	***	+
4.1_%Ocupados	33.19	6.51	32.88	4.91	**		31.54	5.42	***	-	38.53	5.05	***	+++	27.38	7.11	***	---	43.92	9.84	***	+++
4.1_%Parados	10.48	5.11	10.46	3.91	ns		11.34	5.77	***		8.18	3.14	***	-	12.95	7.53	***	+	6.64	3.10	***	--
4.1_%Inactivos	38.02	7.24	40.05	5.75	***	+	38.02	5.96	ns		31.36	6.12	***	---	43.94	7.25	***	+++	31.39	10.14	***	---
4.2_LocalComercialPor1000Hab	26.90	98.90	32.00	66.40	***		20.60	27.90	***		20.80	24.90	***		19.40	20.90	***		113.80	594.90	ns	
4.2_OficinaYServiciosPor1000Hab	10.70	44.60	16.70	69.20	***		5.60	12.60	***		7.30	15.50	***		8.10	12.20	***		22.00	33.90	***	+
4.2_Industrial1000Hab	3.07	12.11	1.80	5.71	***		4.97	19.72	***		2.30	5.41	***		3.63	7.94	ns		2.47	4.66	ns	
4.2_LocalAgrario1000Hab	0.65	6.71	0.17	1.32	***		0.59	2.89	ns		0.36	1.78	***		3.14	19.84	**	+	0.80	3.83	ns	
4.2_LocalInactivoPor1000Hab	13.37	20.22	14.65	22.42	**		10.03	13.89	***		12.21	16.34	*		19.63	29.47	***	+	18.73	21.97	*	+
4.3_EstadoConstruccionesD	0.40	0.23	0.42	0.28	**		0.38	0.18	***		0.33	0.20	***	-	0.50	0.21	***	+	0.38	0.19	ns	
4.4_IDH_COMBINADO-	0.00	0.00	0.00	0.00	ns		0.00	0.00	***	+	0.00	0.00	***		0.00	0.00	***	+	-0.02	0.02	***	---
4.4_IDH_COMBINADO+	0.00	0.00	0.00	0.00	***		0.00	0.00	***		0.00	0.00	***		0.00	0.00	ns		0.00	0.00	***	+++
4.4_IDH_COMBINADO	0.00	0.00	0.00	0.00	ns		0.00	0.00	***	+	0.00	0.00	***		0.00	0.00	***	+	-0.02	0.02	***	---
4.5_%AgriculturaGanaderia	4.16	6.06	1.15	1.90	***	-	6.75	6.14	***	+	3.28	5.29	***		7.28	5.74	***	++	13.36	17.83	***	+++
4.5_%Pesca	0.14	0.59	0.26	0.91	***		0.05	0.18	***		0.11	0.30	**		0.03	0.12	***		0.13	0.28	ns	
4.5_%Industria	3.81	2.46	3.32	1.29	***	-	4.79	3.48	***	+	3.98	1.80	**		2.81	2.18	***	-	1.70	0.85	***	---
4.5_%Construccion	4.47	2.44	3.37	1.92	***	-	5.33	2.42	***	+	5.44	2.73	***	+	4.35	2.08	ns		3.85	1.47	***	-
4.5_%Servicios	20.61	8.12	24.78	6.15	***	++	14.61	5.07	***	--	25.73	7.31	***	++	12.91	4.83	***	---	24.87	9.99	***	++
4.5_%Paro	24.09	11.26	24.20	8.99	ns		26.09	11.69	***		17.59	6.66	***	--	31.61	16.40	***	++	13.80	6.87	***	---
4.6_%ViviendaEnPropiedad	82.28	13.30	78.99	17.03	***	-	84.19	9.60	***		86.71	8.67	***	+	82.73	10.10	ns		72.85	10.91	***	--
4.6_%ViviendasEnAlquiler	9.39	11.29	14.25	15.15	***	+	5.55	5.98	***	-	7.22	5.89	***		5.05	5.16	***	-	17.98	8.82	***	++
4.6_%ViviendaNoAlquilerNoPropiedad	8.32	8.69	6.76	9.25	***		10.26	8.20	***	+	6.07	5.83	***	-	12.23	9.62	***	+	9.17	8.95	ns	
7_%Analfabetismo	4.63	3.77	3.41	3.20	***	-	6.59	3.96	***	++	2.85	2.17	***	-	6.81	3.83	***	++	3.32	3.20	***	-

Mean: Media
SD: Desviación Típica
conf: Confianza p-valor: ns: p>0.05; *: p<=0.05; ** p<=0.01; ***: p<=0.001
ES: Effect Size. Tamaño del Efecto. +++: Grande positivo; ++:Mediano positivo; +:Pequeño positivo; ---: Grande negativo; --:Mediano negativo; -:Pequeño negativo

Tabla 1: Variables de dimensión demográfica, social, laboral de los equipamientos y de los servicios. Características estadísticas de la totalidad de los datos y de los Modelo SOM de 5 Perfiles. Se marcan las variables con elevado Tamaño del efecto en la constitución del Perfil
Fuente: Elaboración propia

El agrupamiento, clusterización o clasificación de los sujetos se realiza mediante un análisis adicional Ward-clúster sobre el mapa (Figuras 1 y 2). Se generan perfiles o prototipos mediante modelado de patrones y tendencias en la información (Weiss & Indurkhya, 1998). La elección del número de perfiles puede ser guiada por diversos métodos o elegida por la capacidad de interpretación que tenga el investigador sobre los mismos.[15]

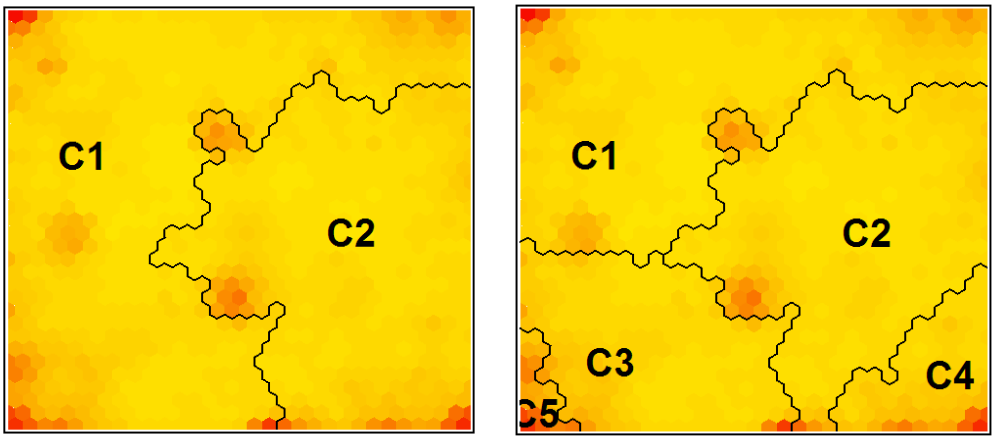
Para facilitar la comprensión de los perfiles obtenidos, se puede caracterizar cada clúster determinado por el

[15] Un buen punto de partida para la definición del número de perfiles puede ser el uso de la regla de Sturges ($c=1+\log_2 N$; N=tamaño de la muestra), habitualmente empleada para la definición del número de clases al elaborar histogramas. En otros casos el número de perfiles usualmente se va ajustando dependiendo del objetivo del investigador y de la interpretación de los resultados que se van obteniendo, deteniendo el proceso cuando la nueva división no aporte conocimiento extra al usuario.

	Muestra completa		Perfil 1				Perfil 2				Perfil 3				Perfil 4				Perfil 5			
N. samples (%)	5381		2054		38.17%		1659		30.83%		1002		18.62%		550		10.22%		115		2.14%	
Atribute	Mean	(SD)	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES	Mean	(SD)	conf	ES
2.3_%Garaje	19.03	18.79	13.09	16.53	***	-	18.98	15.17	ns		31.06	23.56	***	++	16.70	14.07	***		32.22	21.09	***	++
2.4_%NoAccesible	72.81	30.40	63.24	31.24	***	-	82.44	27.06	***	+	68.44	30.27	***		87.94	22.67	***	+	70.36	24.28	ns	
4.3_EdadMediaConstruccionesM	1963	20.00	1962	17.60	*		1963	17.60	ns		1976	12.60	***	++	1943	28.10	***	---	1975	11.30	***	++
4.3_EdadMediaConstruccionesD	23.38	14.79	20.07	14.76	***	-	26.07	13.76	***		19.03	11.14	***	-	37.01	14.78	***	+++	16.61	9.81	***	-
4.3_EstadoConstruccionesM	1.16	0.21	1.21	0.27	***	+	1.12	0.13	***		1.10	0.14	***	-	1.23	0.20	***	+	1.13	0.14	*	
5.1_%RuidosExteriores	32.53	18.35	44.88	13.88	***	++	24.35	16.28	***	-	31.99	13.62	ns		13.15	15.83	***	---	27.39	13.03	***	-
5.1_%Contaminacion	19.47	14.83	26.29	14.18	***	+	15.70	14.39	***	-	18.38	12.24	**		8.11	11.66	***	--	15.80	8.90	***	-
6.1_AlturaMediaConstruccionesM	2.68	1.60	3.73	1.82	***	++	1.81	0.55	***	--	2.53	1.46	**		1.70	0.51	***	--	2.27	1.10	***	-
6.1_AlturaMediaConstruccionesD	5.46	5.27	9.12	5.73	***	++	2.34	1.59	***	--	5.10	4.90	*		1.94	1.34	***	--	5.12	4.38	ns	
6.2_EdifViviendas	327.50	292.70	134.30	157.80	***	--	434.80	197.60	***	+	428.40	345.90	***	+	460.10	225.30	***	+	715.70	751.60	***	+++
6.2_%Unifamiliares	61.10	33.88	34.12	31.99	***	--	80.82	16.42	***	++	68.43	29.68	***	+	87.48	13.05	***	++	68.38	24.39	**	+
6.2_%UnifamiliaresAgrupadas	21.66	21.86	33.10	26.30	***	++	13.76	13.16	***	-	18.80	17.97	***		8.52	8.70	***	--	19.22	13.65	ns	
6.2_%UnifamiliaresConLocales	17.04	24.09	32.40	29.64	***	++	5.35	6.83	***	-	12.64	18.87	***		3.94	6.38	***	--	12.30	17.27	**	
6.2_%Plurifamiliares	0.20	0.80	0.37	1.19	***	+	0.07	0.23	***		0.13	0.53	***		0.06	0.20	***		0.10	0.39	**	
6.2_%EdifLocalesConAlgunaVivienda	0.80	2.59	1.44	3.87	***	+	0.33	0.84	***		0.51	1.21	***		0.24	0.74	***	-	1.45	2.73	*	+
6.2_%EdifLocales	8.93	19.87	10.76	24.17	***		7.20	13.11	***		9.06	23.73	ns		6.95	7.79	***		9.70	15.41	ns	
6.2_%EdifAlojamientos	0.08	1.05	0.02	0.37	***		0.04	0.47	***		0.05	0.78	ns		0.22	1.79	ns		1.32	4.83	**	+++
6.3_%MalasComunicaciones	13.95	15.95	12.28	12.17	***		12.06	16.52	***		19.36	17.54	***	+	15.06	20.53	ns		18.45	16.06	**	+
6.4_%PocasZonasVerdes	48.85	25.97	54.96	22.89	***	+	47.33	28.18	*		45.49	21.97	***		37.60	31.06	***	-	44.57	22.70	*	
6.5_%FaltaAseosEnVIviendas	1.32	2.95	1.34	2.88	ns		1.16	2.07	**		0.79	1.42	***		2.20	5.11	***	+	3.48	6.55	***	++

Mean: Media
SD: Desviación Típica
conf: Confianza p-valor: ns: p>0.05; *: p<=0.05; ** p<=0.01; ***: p<=0.001
ES: Effect Size. Tamaño del Efecto. +++: Grande positivo; ++:Mediano positivo; +:Pequeño positivo; ---: Grande negativo; --:Mediano negativo; -:Pequeño negativo

Tabla 2: Variables de dimensión residencial. Características estadísticas de la totalidad de los datos y de los Modelo SOM de 5 Perfiles. Se marcan las variables con elevado Tamaño del efecto en la constitución del Perfil
Fuente: Elaboración propia



Figuras 1 y 2: A la Izquierda: Mapa Auto-organizado para la clasificación en 2 perfiles
A la derecha: Mapa Auto-organizado para la clasificación en 5 perfiles
Fuente: Elaboración propia

análisis SOM con datos estadísticos básicos como son la Media, la Desviación Estándar, el Máximo y el Mínimo (Faggiano et al., 2010), tratando de conseguir principalmente dos resultados adicionales: i) el factor o variable que es más importante para el efecto y ii) el valor de tal factor (Wu & Hsiao, 2015). Para el análisis de los perfiles además de la información estadística que los define, son valiosos los Mapas SOM monovariables (Figuras 3 y 4) porque permiten según la distribución de valores en el mismo, evaluar relaciones y correlaciones entre variables.

Para cumplir con las recomendaciones de la *American Statistical Association* (Wasserstein & Lazar, 2016) para cada variable y perfil, además de la significación estadística[16] se calcula el tamaño del efecto.[17] En las Tablas 1 y 2 se muestran los resultados de la clasificación en 5 perfiles que se describirán en Resultados.

[16] La significación estadística se lleva a cabo mediante la Prueba *T-Student* bilateral (p-valor≤0.05).

[17] Tamaño del Efecto (Effect Size: SE) es una medida de cómo influye en los valores alcanzados en la variable el hecho de estar o no dentro del perfil en cuestión. Se calcula como el cociente de la diferencia de la media entre el grupo experimental (perfil) y la media del grupo control (media de la población) dividido entre la desviación Estándar de la población (Coe & Merino, 2003, p. 149). En las tablas se indican los tamaños del efecto para cada atributo/variable que interviene en la construcción del perfil: +++ efecto positivo grande, ++ efecto positivo medio, + efecto positivo bajo, - efecto negativo bajo, - - efecto negativo medio, - - - efecto negativo grande (Cohen, 1998), obteniéndose una información muy relevante del efecto que tiene las variables en la definición y singularidad de cada perfil.

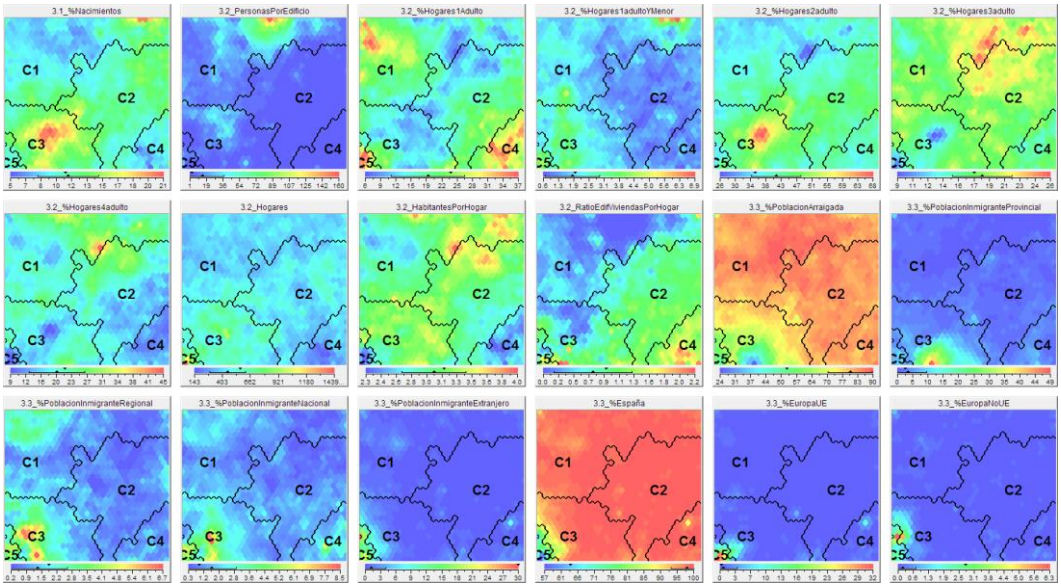
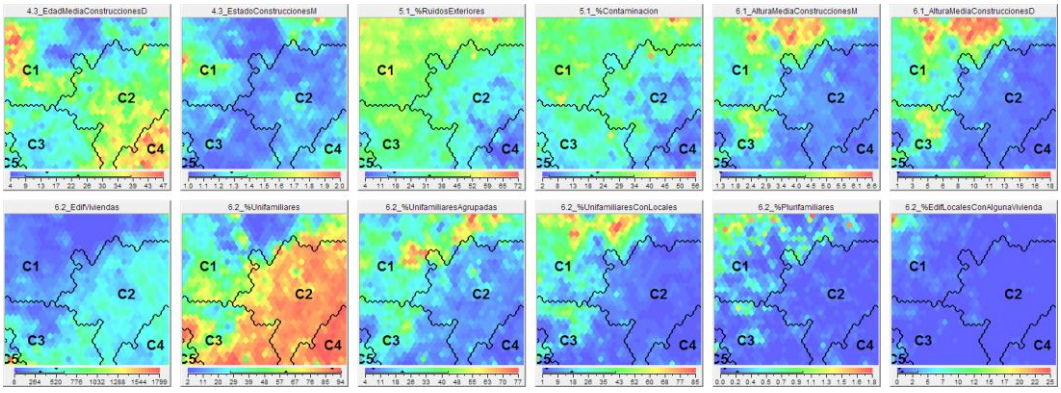


Figura 3: Conjunto de Mapas auto-organizados monotemáticos para la clasificación de 5 perfiles; variables demográficas
Fuente: Elaboración propia



Figuras 4: Conjunto de Mapas auto-organizados monotemáticos para la clasificación de 5 perfiles. Variables residenciales
Fuente: Elaboración propia

Fase de modelado 2: Modelo de predicción. Como objetivo principal tenemos la construcción de un modelo que permita simultáneamente explicar y predecir la vertiente demográfica, social, laboral y de los equipamientos y servicios que quedó caracterizada y etiquetada en perfiles en la Fase de modelado 1 (variable explicada, dependiente o respuesta), a partir de variables de la dimensión residencial (variables explicativas, independientes o predictoras). En esta fase se usará el modelo árbol de decisión y específicamente el árbol de inferencia condicional.

De este modo una vez etiquetadas las secciones censales mediante el SOM en la Fase de modelado 1, se procede a la construcción del modelo de árbol de decisión que permita la predicción basada en reglas, materializándose mediante la representación de condiciones sucesivas que permitirán identificar, con su grado de probabilidad, a qué categoría o tipología de realidad geo-social de las etiquetadas por el SOM pertenece (perfil), atendiendo a las características residenciales (véanse Figuras 5 y 6); para ello, se han usado los árboles de inferencia condicional.[18]

iv) Representaciones visuales:

Una de las cualidades que presentan las cartografías SOM es la capacidad de representación de la información

[18] Se usan los árboles de inferencia condicional que se basan en el test de permutación definidos por Strasser y Weber (1999). Este enfoque utiliza pruebas no paramétricas como criterios de división de las ramas, no precisándose poda de las mismas. Se ha usado el software el paquete "rpart" del software *R-Project* (Hothorn, Hornik, & Zeileis, 2006).

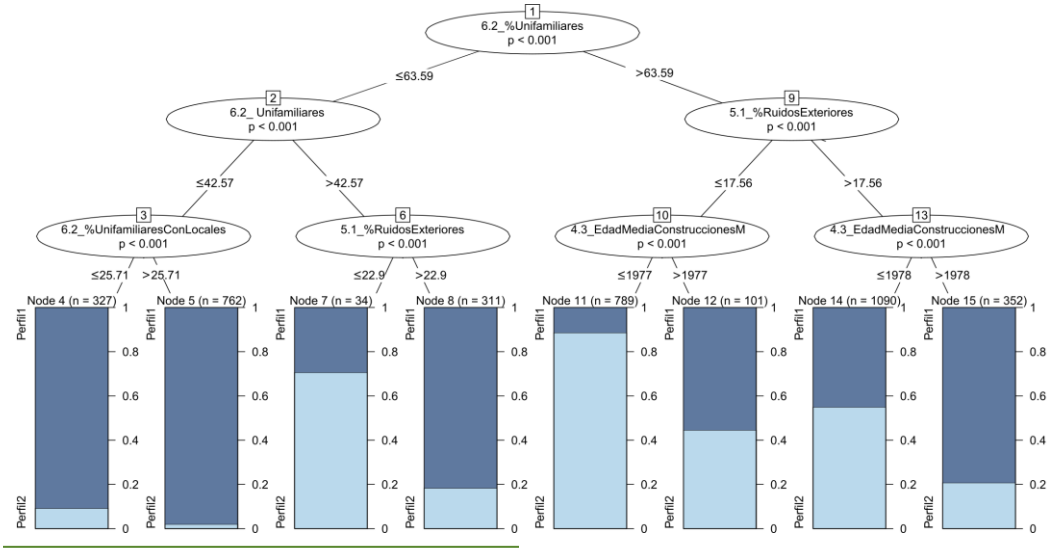
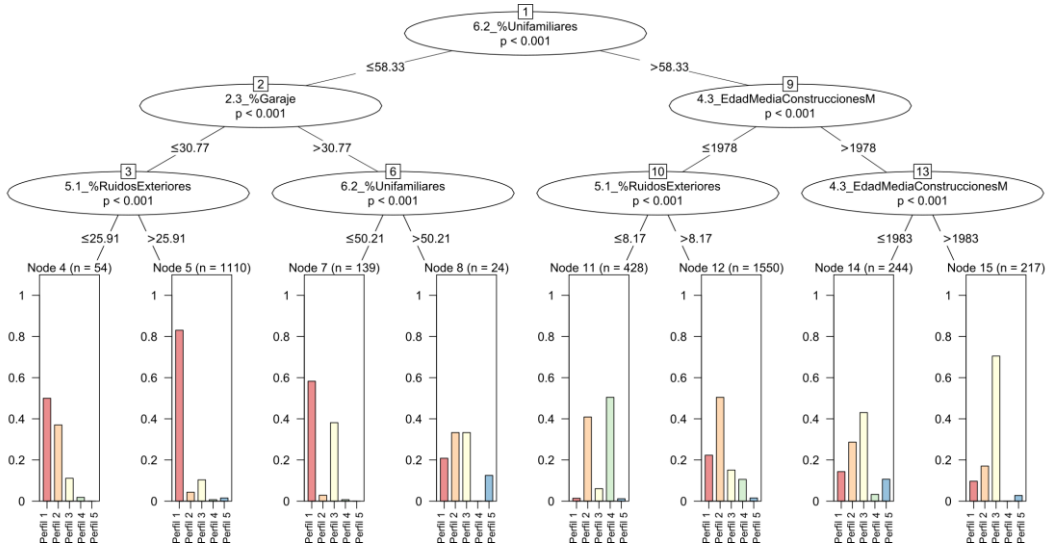


Figura 5: Árbol de la clasificación en 2 perfiles
Fuente: Elaboración propia



Figuras 6: Árbol de la clasificación en 5 perfiles
Fuente: Elaboración propia

resultante de un modo relativamente sencillo de comprender, al mostrar una representación bidimensional de las instancias de partida, con la característica de que cada una de ellas tiene por "vecina" la instancia con cualidades más semejantes. En la misma cartografía se suele representar las agrupaciones de las instancias en los distintos perfiles conformados (Figuras 1 y 2). Esta representación se suele completar con un mapa por cada uno de los atributos o variables que construyeron el mapa SOM (Figuras 3 y 4).

Estos mapas contribuyen a la comprensión de la distribución de los datos en el mapa SOM. Finalmente como las instancias evaluadas tienen su identidad y forma en el espacio, se representarán mediante un SIG la información de los perfiles determinados en la Fase de modelado 1 (Figuras 7, 8 y 9). Esta retorno al SIG de las instancias una vez clasificadas en clases, ha sido frecuente como por ejemplo en investigaciones médicas en análisis no lineales de múltiples variables en ciertas enfermedades (Basara & Yuan, 2008), en la representación de resultados de la clusterización SOM sobre el riesgo ecológico de contaminación (Faggiano et al., 2010) o aplicándolos experimentalmente a datos procedentes de información socio-demográfica oficial del Área Metropolitana de Lisboa (Bação, Lobo, & Painho, 1995).

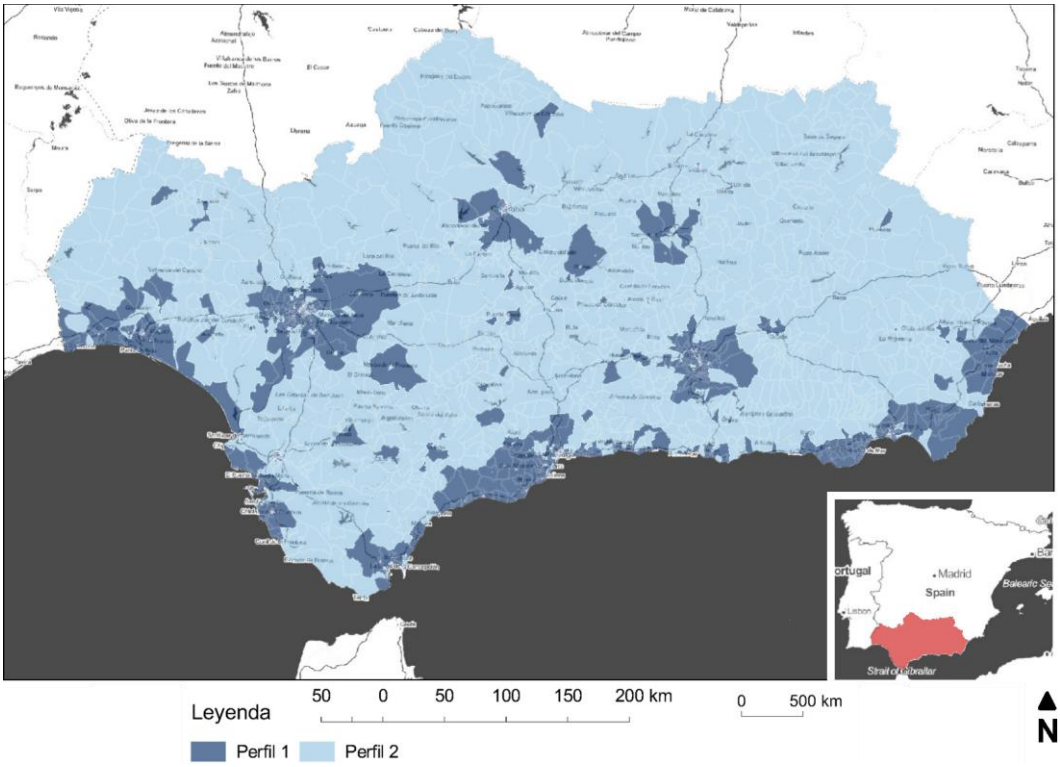
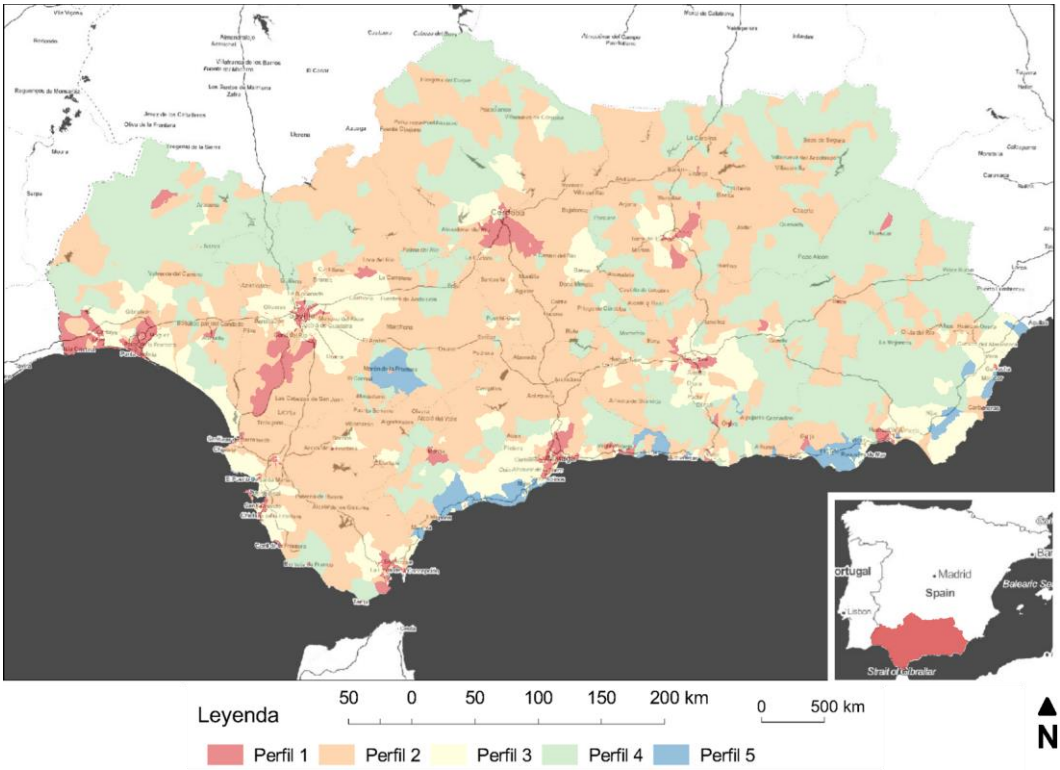


Figura 7: Representación SIG de la clasificación SOM de 2 perfiles para toda Andalucía.
Fuente: Elaboración propia



Figuras 8: Representación SIG de la clasificación SOM de 5 perfiles para toda Andalucía
Fuente: Elaboración propia

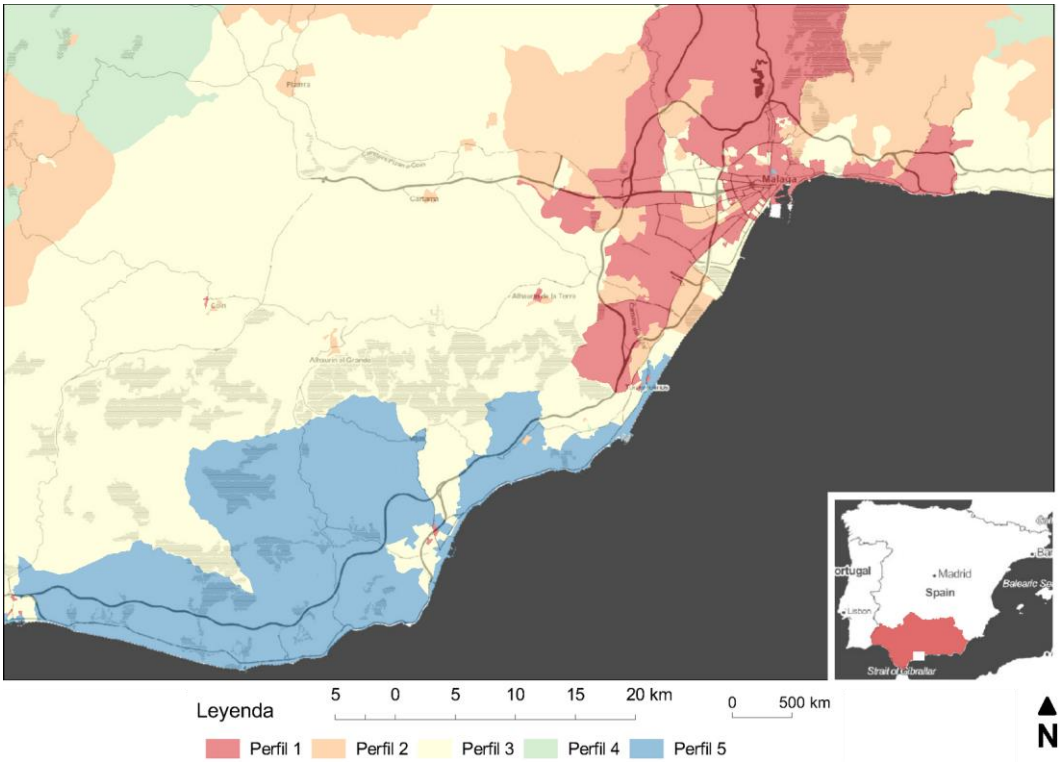


Figura 9: Representación SIG del territorio próximo a la ciudad de Málaga. Clasificación SOM de 5 perfiles.
Fuente: Elaboración propia

4. Resultados

Según la metodología se realizan las tareas propias de los apartados (i) y (ii), obteniéndose dos bases de datos independientes. Se puede observar una síntesis descriptiva de tales variables de partida en las 3 primeras columnas de la Tabla 1 para los datos principalmente demográficos y de la Tabla 2 para la dimensión residencial. Siguiendo con el apartado (iii) de la metodología, se lleva a cabo la Fase de modelado 1 usando para ello 63 variables de la dimensión demográfica, social, laboral, de los equipamientos y de los servicios, obteniéndose los perfiles que caracterizan la realidad demográfica. A continuación en la Fase de modelado 2 se obtiene el árbol que permite "predecir" a qué perfil (demográfico) corresponde cada sección censal a partir exclusivamente de datos de la dimensión residencial.

A continuación se describirán los resultados obtenidos siguiendo las distintas fases de la metodología, realizándose en primer lugar una evaluación previa con el estudio de la estructura demográfica basada en dos únicos perfiles y a continuación la evaluación definitiva mediante la clasificación en cinco perfiles.

4.1. Resultados y evaluación previa (2 perfiles demográficos)

En un primer análisis SOM se determinan un número de perfiles mínimo –concretamente dos– con la intención de realizar una primera evaluación de los resultados. Al trasladarlas dos categorías demográficas al espacio mediante SIG (Figura 7) se observa que los resultados son básicamente los esperados, se identifican espacialmente claramente las "dos Andalucías", una nítidamente rural (Perfil 2) y otra con un protagonismo principal de lo urbano (Perfil 1).

A continuación se pudo evaluar el resultado obtenido con el árbol de decisión de la Figura 5, en el que se observa perfectamente qué variables de la dimensión residencial definen las distintas ramas del modelo de perfiles demográficos. Tal y como podríamos prever las variables que principalmente construyen el modelo demográfico son: la proporción de viviendas unifamiliares, la cantidad de viviendas unifamiliares con locales, la cantidad de quejas por ruidos exteriores y la edad media de las construcciones. Se puede observar cómo todas y cada una de estas variables aportan información relevante y definen lo rural frente a lo urbano en el modelo y en la realidad.

4.2. Resultados y evaluación de 5 perfiles demográficos

Si analizamos los 5 perfiles obtenidos del análisis SOM final, comparando por un lado la información estadística que caracteriza a cada uno de ellos (Tabla 1), con la espacialización de los propios perfiles en la región andaluza (Figuras 8 y 9), alcanzamos los siguientes resultados y conclusiones:

-Perfil 1: Estadísticamente se comprueba que las secciones censales contenidas en este perfil presentan, comparados con el resto de perfiles: mayor presencia de delincuencia, mayor número de personas por edificio, mayor dedicación en empleos de servicios y un menor número de viviendas por cada hogar ocupado. Al representar los perfiles espacialmente, mediante SIG, se observa que son coincidentes con las principales áreas urbanas y sus inmediatas conurbaciones a lo largo de toda la región. Tenemos pues un perfil con unas connotaciones urbanas claramente de ciudad consolidada.

-Perfil 2: En el análisis estadístico de este perfil poblacional se observa cierta diversificación laboral, aunque con poca presencia del sector servicios, una población eminentemente española, escasos inmigrantes y un elevado número de analfabetos, no siendo frecuentes los hogares con un único adulto y menores. Espacialmente podríamos decir que este perfil se trata de una población emplazada en un entorno rural, diferenciándose con respecto al otro perfil rural (perfil 4) en que su población es más joven que en aquel, con mayor población activa, con más actividades propias de esa realidad, como por ejemplo mayor dedicación a la construcción o la industria, y con unos hogares con mayor número de habitantes.

-Perfil 3: Estadísticamente destaca por un mayor número de nacimientos, mayor número de inmigrantes de origen provincial y, en menor, regional o nacional que a su vez trabaja en la provincia, con un elevado porcentaje de ocupados laboralmente y menor tasa de paro. Asimismo presenta una edad inferior a la media, con pocos hogares unipersonales y con bajo nivel de arraigo. Espacialmente se localizan como zonas periféricas de las principales ciudades.

-Perfil 4: El análisis estadístico desvela que este perfil poblacional presenta una elevada edad media, gran cantidad de hogares con un único ocupante, abundancia de viviendas vacías y con ciertos problemas como carencia de agua corriente en proporción mayor que el resto. Curiosamente la estadística delata que habitan en asentamientos con buenos ratios de equipamientos culturales y de bienestar por población, probablemente derivado del bajo número de habitantes de tales poblaciones y una aceptable distribución de tales funciones. Espacialmente se observa que corresponden con los emplazamientos rurales más aislados y a mayor distancia de las principales ciudades. Comparando este perfil con el perfil 2, se observa que coincide con una población rural más envejecida que en muchas ocasiones vive sola, en entornos urbanos con poca población, con poca ocupación de las viviendas y con altos índices de analfabetismo, paro e inactividad. Podemos localizar este perfil entre otros ámbitos como en la Hoya de Baza (Granada), en los Campos de Tabernas (Almería), altos de la Sierra de Gádor (Almería) o Sierra de Aracena (Huelva).

-Perfil 5: Destaca por un elevado número de viviendas ocupadas por una persona, en bastantes ocasiones con algún menor a su cargo, alta presencia de inmigrantes procedentes del resto de Andalucía, del resto de España y especialmente extranjeros con el consiguiente bajo arraigo de su población. Presentan una alta tasa de ocupación, bajo paro y baja inactividad, trabajando primordialmente en el sector servicios o en la agricultura. Espacialmente se reconocen y se identifican como áreas urbanizadas bien conocidas por su fuerte y singular presencia de residentes extranjeros, ya sean en enclaves turísticos como Marbella (Málaga), o Almuñécar-Cerro Gordo (Granada) o de fuerte producción agraria intensiva, como la zona de invernaderos del Campo de Dalías (Almería).

A modo de resumen de los perfiles demográficos podemos distinguir la presencia de un perfil eminentemente urbano (perfil 1); dos perfiles suburbanos, entre los que podemos diferenciar un perfil (3) en el que abunda una población joven y activa, con familia, inmigrante de corta distancia (provincial), con vivienda en propiedad y trabajo en la provincia, frente a otro perfil (5) caracterizado fundamentalmente por la abundancia de inmigrantes de larga distancia (regionales, nacionales o extranjeros), muy activos en trabajos vinculados con la agricultura o los servicios y con vivienda sobre todo en alquiler. Finalmente destacan dos perfiles eminentemente rurales, uno en el que se observa cierta vitalidad, juventud y actividad económica (perfil 2) y otro en clara depresión, envejecimiento de su población y recesión (perfil 4).

Los perfiles pueden ser analizados estadísticamente tal y como hemos hecho anteriormente, o de un modo gráfico mediante los Mapas auto-organizados monotemáticos (Figuras 3 y 4). En ellos podemos observar una distribución "semántica" de todas las secciones censales estudiadas, situándose cada una de ellas junto a una semejante, atendiendo a la globalidad de las variables incorporadas al SOM. Asimismo las secciones censales semejantes se agrupan en los distintos perfiles o *clusters* (etiquetados en los gráficos como C1, C2, C5). El interés de esta representación consiste en su capacidad de proporcionar conocimiento de forma heurística, permitiendo descubrir cualidades y relaciones entre las distintas partes del mapa y, en consecuencia, entre los distintos perfiles. Por ejemplo, podemos observar cómo los índices más altos (rojo) de nacimientos, variable *3.1_%Nacimientos*, salvo alguna excepción se sitúan en el Perfil 3 (C3), coincidiendo en gran medida, como era de esperarse, con los hogares que tienen dos adultos, tal y como se puede observar con las zonas marcadas igualmente en rojo en el mapa *3.2_%Hogares2adulto*.

Son numerosas las relaciones que se pueden encontrar entre los mapas auto-organizados monotemáticos, dependiendo del interés o enfoque particular del investigador.

Tras la Fase de modelado 1 y su interpretación, en la Fase de modelado 2 se obtendrá la identificación, o más bien la probabilidad de realizar correctamente la identificación de los perfiles demográficos, a partir de variables exclusivamente de la dimensión residencial. Es decir, es posible –para una sección censal– inferir o predecir la probabilidad de pertenencia a un perfil demográfico a partir de determinadas variables residenciales, como son el porcentaje de viviendas unifamiliares, el porcentaje de viviendas que tienen garaje, la edad media de construcción o el porcentaje de usuarios que se quejan por ruidos (Figura 6).

Se observa de este modo que hay determinadas variables que participan más que otras en la construcción del modelo; y, en el caso de que alguna de ellas no tenga relevancia para el mismo o incluso únicamente aporte ruido, el análisis estadístico tras la red neuronal o el árbol de decisión mostrará tal falta de significación y ausencia de valor, sin influir en ningún momento en la construcción del modelo basado en la red neuronal artificial.

5. Discusión y conclusiones

A partir del análisis de las experiencias bibliográficas de aplicación del modelo de clasificación y conocimiento mediante la metodología SOM y corroborado mediante la propia experiencia llevada a cabo, se puede concluir que la metodología SOM es útil para:

1. Realizar un análisis exploratorio (Spielman & Thill, 2008).
2. Hacer más potentes, robustas y más completas las clasificaciones descriptivas tradicionales (Hamaina et al., 2012).
3. Comprender los patrones de distribución espacial que se dan en un territorio atendiendo a las variables en estudio (Faggiano et al., 2010).
4. Explorar visualmente, validar y evaluar eficazmente gracias a las consistentes propiedades geométricas de los resultados de los SOM (AUTOR 1 & Osuna Pérez, 2013; Yan & Thill, 2009).
5. Analizar eficazmente complejos conjuntos de datos geográficos, específicamente, demográficos (Takatsuka, 2001).
6. Inferir consideraciones espaciales a partir de los grupos taxonométricos hallados (AUTOR 1 et al., 2015; Faggiano et al., 2010).
7. Codificar las clasificaciones resultantes en un SIG consiguiéndose hacerlos más accesibles y comprensibles a una audiencia no familiar con los SOM (Kauko, 2005).
8. Etiquetar la realidad geográfica sin tener que nombrar tales categorías, evitando así las problemáticas inherentes del análisis de factores y de las técnicas geodemográficas (Spielman & Thill, 2008).
9. Superar los retos tradicionales asociados con el estudio de la complejidad de las comunidades ambientales, evidenciándose un gran potencial de la combinación del SOM y del SIG (Basara & Yuan, 2008).
10. Evaluar los efectos de la concurrencia de ciertas variables en estudio (Faggiano et al., 2010).
11. Constituir una potente solución alternativa, en un tiempo caracterizado por las tecnologías de la información y por la proliferación de datos (Hatzichristos, 2004).
12. Ser usado como un sistema de apoyo a la decisión para analizar y visualizar conjuntos de indicadores estadísticos para diversas aplicaciones (Kaski & Kohonen, 1996).

Por su lado la metodología basada en árboles de decisión a partir de una clasificación SOM se considera útil para:

1. Atribuir de forma muy sencilla patrones de comportamiento que pueden ser muy complejos.
2. Predecir de forma eficaz comportamientos de variables que presentan cierto coste o dificultad de evaluación, como son las variables demográficas o sociales, a partir de otras variables con menor complejidad y coste de evaluación, como las variables residenciales.
3. Generar y verificar hipótesis sobre realidades y comportamientos complejos, sin que sea necesaria la participación del usuario para su formulación.
4. Hacer accesible los sistemas de apoyo a la decisión a un público no experto.
5. Identificar variables que se relacionan de forma significativa y su peso o tamaño del efecto en la realidad estudiada.

Como oposición a las anteriores, es necesario tener presente ciertas precauciones y limitaciones en el uso de estas metodologías:

1. Un análisis de la población de una sección censal no es propiamente un análisis de la población y debe extremarse la precaución y limitar la inferencia a la escala de la observación, sin

alcanzar directamente a los individuos (Spielman & Thill, 2008); es decir, no se debe extrapolar a individuos las conclusiones obtenidas del estudio de grupos de individuos.

2. La integración completa entre SOM y SIG es compleja (André Skupin & Hagelman, 2003), quedando limitada a una conexión más o menos manual. Salvo algunos intentos de conexión, aún no se ha implementado una conexión directa "amigable" entre ninguno de los principales softwares SIG y SOM (Takatsuka, 2001).

3. Las metodologías apoyadas en los sistemas basados en el conocimiento no se encuentran desarrolladas para la integración directa en los procesos de desarrollo y planificación urbana y territorial (Behnisch & Ultsch, 2009; Streich, 2005); situación que sugiere, en conjunción con la anterior, que existe una importante brecha tecnológica que puede convertirse en un espacio de desarrollo técnico, tecnológico y de oportunidad de investigación y/o de negocio.

4. La combinación del conocimiento experto con los resultados SOM requieren cierta creatividad (Kauko, 2005), sin ser en absoluto inmediatos ni obvios.

Mediante la investigación aplicada al caso de estudio de la región de Andalucía, se han obtenido unos árboles de decisión basados en una metodología de clasificación no supervisada basada en Mapas auto-organizados que han demostrado ser sencillos de usar, a la vez que útiles y capaces de predecir –con un relativo bajo error– fenómenos demográficos complejos y de relevancia a partir de la realidad residencial. Se puede, por tanto, concluir que existe una conexión y relación entre los fenómenos demográficos y la configuración residencial en Andalucía; por ello, se debe ser cauto y evitar *a priori* un establecimiento causa-efecto entre tales fenómenos, pues requerirían otras pruebas alejadas de los objetivos de esta investigación.

Como citar este artículo/*How to cite this article*: Abarca-Alvarez, F., Campos-Sánchez, F. & Reinoso-Bellido, R. (2017). Metodología de ayuda a la decisión mediante SIG e Inteligencia Artificial: aplicación en la caracterización demográfica de Andalucía a partir de su residencia. *Estoa, Revista de la Facultad de Arquitectura y Urbanismo de la Universidad de Cuenca*, 6(11), 33-51. doi:10.18537/est.v006.n011.a03

Bibliografía

- Abarca-Alvarez, F., Campos-Sánchez, F. S., & Osuna-Perez, F. (2015). Taxonomía de las inmigraciones turísticas de Andalucía basada en las cualidades de sus asentamientos urbanos. En *Migraciones Contemporáneas, Territorio y Urbanismo*.
- Abarca-Alvarez, F., & Fernandez-Avidad, A. (2010). Generation of downtown planning-ordinances using self organizing maps. En *10th International Conference on Design and Decision Support Systems, DDSS 2010*.
- Abarca-Alvarez, F., & Osuna Pérez, F. (2013). Cartografías semánticas mediante redes neuronales: los mapas auto-organizados (SOM) como representación de patrones y campos. *EGA. Revista de expresión gráfica arquitectónica*, 18(22). http://doi.org/10.4995/ega.2013.1692
- Astudillo, C. A., & John Oommen, B. (2011). Imposing tree-based topologies onto self organizing maps. *Information Sciences*, 181(18), 3798-3815. http://doi.org/10.1016/j.ins.2011.04.038
- Astudillo, C. A., & Oommen, B. J. (2013). On achieving semi-supervised pattern recognition by utilizing tree-based SOMs. *Pattern Recognition*, *46*(1), 293-304. http://doi.org/10.1016/j.patcog.2012.07.006
- Ayedi, B. (1998). The design of spatial decision support systems in urban and regional planning. En Timmermans, H. Decesion Support Systems in Urban Planning. Routledge.
- Bação, F., Lobo, V., & Painho, M. (1995). The Self-Organizing Map and it's variants as tools for geodemographical data analysis: the case of Lisbon's Metropolitan Area. *Computers & Geosciences*, *31*(Goss), 155-163. http://doi.org/10.1016/j.cageo.2004.06.013
- Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. *Computational Science–ICCS 2005*, *3516*, 476-483. http://doi.org/10.1007/11428862_65
- Basara, H. G., & Yuan, M. (2008). Community health assessment using self-organizing maps and geographic information systems. *International journal of health geographics*, 7, 67. http://doi.org/10.1186/1476-072X-7-67
- Behnisch, M., & Ultsch, A. (2009). Urban data-mining: spatiotemporal exploration of multidimensional data. *Building Research & Information*, 37(5-6), 520-532. http://doi.org/10.1080/09613210903189343
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall.
- Buzai, G. D. (2007). Sistemas de Información Geográfica: Aspectos conceptuales desde la teoría de la Geografía. *XI Conferencia Iberoamericana de Sistemas de Información Geográfica (XI CONFIBSIG)*. En Sociedad Iberoamericana de Sistemas de Información Geográfica, Luján, Argentina.
- Buzai, G.D. (2015). Geografía global y Neogeografía. La dimensión espacial en la ciencia y la sociedad. *Polígonos. Revista de Geografía,* 27, 49-60.
- Cao,L. S.; Y. Philip, C. Zhang, & H. Zhang (eds.) (2009). *Data mining for business applications*. New York.
- Coe, R., & Merino, C. (2003). Magnitud del efecto: Una guía para investigadores y usuarios. *Revista de Psicología*, *21*(1), 147-177.
- Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences* (Vol. 2nd Editio). Lawrence Erlbaum Associates, Publishers. http://doi.org/10.1234/12345678
- Delmelle, E. C., Thill, J. C., Furuseth, O., & Ludden, T. (2012). Trajectories of Multidimensional Neighbourhood Quality of Life Change. *Urban Studies*, 50(5), 923-941. http://doi.org/10.1177/0042098012458003
- Demartines, P., & Blayo, F. (1992). Kohonen Self-Organizing Maps : Is the Normalization Necessary? *Complex Systems*, 6(2), 105-123.
- Diappi, L., Bolchim, P., & Buscema, M. (2004). Improved Understanding of Urban Sprawl Using Neural Networks. En J. P. Van-Leeuwen & H. J. P. Timmermans (Eds.), *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning* (pp. 33-49). Politecn Milan, Dept Architecture and Planning, I-20133 Milan, Italy.: Springer.
- Faggiano, L., de Zwart, D., García-Berthou, E., Lek, S., & Gevrey, M. (2010). Patterning ecological risk of pesticide contamination at the river basin scale. *Science of the Total Environment*, 408(11), 2319-2326. http://doi.org/10.1016/j.scitotenv.2010.02.002
- Feng, S., & Xu, L. D. (1999). Decision support for fuzzy comprehensive evaluation of urban development. *Fuzzy Sets and Systems*, *105*(1), 1-12. http://doi.org/10.1016/S0165-0114(97)00229-7
- Goodchild, M. F. (2010). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science, 1,* 3-20. http://doi.org/10.5311/JOSIS.2010.1.2
- Gomes, H., Ribeiro, A. B., & Lobo, V. (2007). Location model for CCA-treated wood waste remediation units using GIS and clustering methods. *Environmental Modelling and Software*, 22(12), 1788-1795. http://doi.org/10.1016/j.envsoft.2007.03.004
- Gómez-Carracedo, M. P., Andrade, J. M., Carrera, G. V. S. M., Aires-de-Sousa, J., Carlosena, A., & Prada, D. (2010). Combining

Kohonen neural networks and variable selection by classification trees to cluster road soil samples. *Chemometrics and Intelligent Laboratory Systems*, 102(1), 20-34. http://doi.org/10.1016/j.chemolab.2010.03.002
- Guo, D., Chen, J., MacEachren, A. M., & Liao, K. (2006). A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6), 1461-1474. http://doi.org/10.1109/TVCG.2006.84
- Hamaina, R., Leduc, T., & Moreau, G. (2012). Towards Urban Fabrics Characterization based on Buildings Footprints. En J. Gensel (Ed.), *Bridging the Geographic Information Sciences* (pp. 231-248). http://doi.org/10.1007/978-3-642-29063-3_13
- Hatzichristos, T. (2004). Delineation of demographic regions with GIS and computational intelligence. *Environment and Planning B: Planning and Design*, 31(1), 39-49. http://doi.org/10.1068/b1296
- Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos*. Pearson Prentice Hall.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(July), 651-674. http://doi.org/10.1198/106186006X133933
- Jarupathirun, S., & Zahedi, F. (2005). GIS as Spatial Decision Support Systems. En J. B. Pick (Ed.), *Geographic information systems in business*. Idea Group Pub.
- Juanes Notario, P. (2014). La Geografía y la Estadística. Dos necesidades para entender Big Data. http://hdl.handle.net/10366/125197
- Kaski, S., & Kohonen, T. (1996). Exploratory Data Analysis By The Self-Organizing Map: Structures Of Welfare And Poverty In The World (1996). *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, 498-507. http://doi.org/10.1.1.53.3954
- Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2), 119-127. http://doi.org/10.2307/2986296
- Kauko, T. (2005). Using the self-organising map to identify regularities across country-specific housing-market contexts. *Environment and Planning B: Planning and Design*, 32(1), 89-110. http://doi.org/10.1068/b3186
- Keen, P. G. W. (1987). Decision support systems: The next decade. *Decision Support Systems*, 3(3), 253-265. http://doi.org/10.1016/0167-9236(87)90180-1
- Kinaci, A. C., & Yucebas, S. C. (2015). Cost Reduction in Thyroid Diagnosis: A Hybrid Model with SOM and C4.5 Decision Trees. En *International Conference on Neural Information Processing* (pp. 440-448). http://doi.org/10.1007/11893257
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59-69. http://doi.org/10.1007/BF00337288
- Kohonen, T. (1990). The Self-Organizing Map. En *Proceeding of the IEEE* (Vol. 78, pp. 1464-1480). http://doi.org/10.1109/5.58325
- Kohonen, T. (1995). *Self-Organizing Maps*. Springer. http://doi.org/10.1007/978-3-642-88163-3
- Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21(1-3), 1-6. http://doi.org/10.1016/S0925-2312(98)00030-7
- Lin, W. (2008). Earthquake-induced landslide hazard monitoring and assessment using SOM and PROMETHEE techniques: A case study at the Chiufenershan area in Central Taiwan. *International Journal of Geographical Information Science*, 22(9), 995-1012. http://doi.org/10.1080/13658810801914458
- Luque Martínez, T. (2000). *Técnicas de análisis de datos en investigación de mercados*. (T. Luque Martínez, Ed.). Madrid: Pirámide.
- Power, D. J., Sharda, R., & Burstein, F. (2015). Decision Support Systems. En C. L. Cooper (Ed.), *Wiley Encyclopedia of Management* (pp. 1-4). Chichester, UK: John Wiley & Sons, Ltd.
- Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81-106. http://doi.org/10.1023/A:1022643204877
- Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61(4), 241-254. http://doi.org/10.1007/BF00203171
- Salah, M., Trinder, J., & Shaker, A. (2009). Evaluation of the self-organizing map classifier for building detection from lidar data and multispectral aerial images. *Journal of Spatial Science*, 54(2), 15-34. http://doi.org/10.1080/14498596.2009.9635176
- Shanmuganathan, S., & Li, Y. (2016). An AI based approach to multiple census data analysis for feature selection. *Journal of Intelligent & Fuzzy Systems*, 31(2), 859-872. http://doi.org/10.3233/JIFS-169017
- Silver, M. S. (2008). On the Design Features of Decision Support Systems : The Role of System Restrictiveness and Decisional Guidance. En F. Burstein & C. W. Holsapple (Eds.), *Handbook on Decision Support Systems 2: Variations* (pp. 261-291). Springer-Verlag Berlin Heidelberg.
- Simmuteit, S., Schleif, F. M., Villmann, T., & Kostrzewa, M. (2009). Hierarchical PCA using tree-som for the identification of bacteria. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5629 LNCS*, 272-280. http://doi.org/10.1007/978-3-642-02397-2_31
- Skupin, A., & Agarwal, P. (2008). Introduction:

What is a Self-Organizing Map? En P. Agarwal & A. Skupin (Eds.), *Self-organising maps : applications in geographic information science* (pp. 1-20). Wiley.
- Skupin, A., & Esperbé, A. (2011). An alternative map of the United States based on an n-dimensional model of geographic space. *Journal of Visual Languages and Computing*, 22(4), 290-304. http://doi.org/10.1016/j.jvlc.2011.03.004
- Skupin, A., & Hagelman, R. (2003). Attribute space visualization of demographic change. *Proceedings of the eleventh ACM international symposium on Advances in geographic information systems - GIS* 2003, 56-62. http://doi.org/10.1145/956676.956684
- Skupin, A., & Hagelman, R. (2005). Visualizing Demographic Trajectories with Self Organizing Maps. *GeoInformatica*, 9(2), 159-179.
- Spielman, S. E., & Thill, J.-C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, *32*(2),110-122. http://doi.org/10.1016/j.compenvurbsys.2007.11.004
- Strasser, H., & Weber, C. (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, 8, 220-250. http://doi.org/10.1007/s10551-011-0925-7
- Streich, B. (2005). *Stadtplanung in der Wissensgesellschaft Ein Handbuch*. VS Verlag für Sozialwissenschaften.
- Strobl, C., Malley, J., & Tutz, G. (2010). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*, 14(4), 323-348. http://doi.org/10.1037/a0016973.An
- Takatsuka, M. (2001). An application of the Self-Organizing Map and interactive 3-D visualization to geospatial data. *Proceedings of the 6th International Conference on GeoComputation*, 24-26.
- Tayebi, M. H., Hashemi Tangestani, M., & Vincent, R. K. (2014). Alteration mineral mapping with ASTER data by integration of coded spectral ratio imaging and SOM neural network model. *Turkish Journal of Earth Sciences*, *23*(6), 627-644. http://doi.org/10.3906/yer-1401-9
- Tsai, C.-F., Lin, Y.-C., & Wang, Y.-T. (2009). Discovering Stock Trading Preferences By Self-Organizing Maps and Decision Trees. *International Journal on Artificial Intelligence Tools*, *18*(4), 603-611. http://doi.org/10.1142/S0218213009000299
- Villmann, T., Merényi, E., & Hammer, B. (2003). Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4), 389-403. http://doi.org/10.1016/S0893-6080(03)00021-2
- Voumvoulakis, E. M., Gavoyiannis, A. E., & Hatziargyriou, N. D. (2006). Dynamic Security Assessment and Load Shedding Schemes Using Self Organized Maps and Decision Trees. En *Hellenic Conference on Artificial Intelligence* (pp. 1-7).
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *1305*(April), 00-00. http://doi.org/10.1080/00031305.2016.1154108
- Weiss, S. M., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*.
- Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. *Data Mining* (Third Edit, Vol. 277). Elsevier. http://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C
- Wu, P. K., & Hsiao, T. C. (2015). Factor Knowledge Mining Using the Techniques of AI Neural Networks and Self-Organizing Map. *International Journal of Distributed Sensor Networks*, *2015*. http://doi.org/10.1155/2015/412418
- Yan, J., & Thill, J.-C. (2009). Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B: Planning and Design*, 36(3), 466-486. http://doi.org/10.1068/b34019
- Yang, C., Guo, R., Wu, Z., Zhou, K., & Yue, Q. (2014). Spatial extraction model for soil environmental quality of anomalous areas in a geographic scale. *Environmental Science and Pollution Research*, *21*(4), 2697-2705. http://doi.org/10.1007/s11356-013-2200-1
- Yang, H., Hu, Y., qi Deng, F., Tian, X., & Li, B. (2004). Fuzzy SOFM-GIS space cluster model and its application analysis. *2004-8th International Conference on Control, Automation, Robotics and Vision-Icarcv 1*, (December), 6-9.
- Yang, Z. R., & Chou, K.-C. (2003). Mining biological data using self-organizing map. *J. Chem. Inf. Comput. Sci.*, *43*(6), 1748-1753.
- Yao, Z., Holmbom, A. H., Eklund, T., & Back, B. (2010). Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis. En *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6171 LNAI, pp. 292-307). http://doi.org/10.1007/978-3-642-14400-4_23
- Yeh, A. G.-O. (2005). Urban planning and GIS. En: Longley, P.A.; Goodchild, M.F.; Maguire, D.J & Rhind, D.W. *Geographical Information Systems: Principles, Techniques, Management and Applications*. En 877-888. Recuperado a partir de http://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/files/ch62.pdf
- Zhang, J., Shi, H., & Zhang, Y. (2009). Self-organizing map methodology and google maps services for geographical epidemiology mapping. *DICTA 2009 - Digital Image Computing: Techniques and Applications*, 229-235. http://doi.org/10.1109/DICTA.2009.46