Computers, Environment and Urban Systems 36 (2012) 218–232

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Exploratory geospatial data analysis using the GeoSOM suite

Roberto Henriques [a,*], Fernando Bacao [a], Victor Lobo [a,b]

[a] ISEGI, Universidade Nova de Lisboa Campolide, 1070-312 Lisboa, Portugal
[b] Portuguese Naval Academy, Alfeite, 2810-001 Almada, Portugal

ARTICLE INFO

Article history:
Received 4 August 2010
Received in revised form 14 November 2011
Accepted 15 November 2011
Available online 26 December 2011

Keywords:
Geographic knowledge discovery
Spatial clustering
Self-Organizing Maps
GeoSOM

ABSTRACT

Clustering constitutes one of the most popular and important tasks in data analysis. This is true for any type of data, and geographic data is no exception. In fact, in geographic knowledge discovery the aim is, more often than not, to explore and let spatial patterns surface rather than develop predictive models. The size and dimensionality of the existing and future databases stress the need for efficient and robust clustering algorithms. This need has been successfully addressed in the context of general-purpose knowledge discovery. Geographic knowledge discovery, nonetheless can still benefit from better tools, especially if these tools are able to integrate geographic information and aspatial variables in order to assist the geographic analyst's objectives and needs. Typically, the objectives are related with finding spatial patterns based on the interaction between location and aspatial variables. When performing cluster-based analysis of geographic data, user interaction is essential to understand and explore the emerging patterns, and the lack of appropriate tools for this task hinders a lot of otherwise very good work.

In this paper, we present the GeoSOM suite as a tool designed to bridge the gap between clustering and the typical geographic information science objectives and needs. The GeoSOM suite implements the GeoSOM algorithm, which changes the traditional Self-Organizing Map algorithm to explicitly take into account geographic information. We present a case study, based on census data from Lisbon, exploring the GeoSOM suite features and exemplifying its use in the context of exploratory data analysis.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Advances in database technologies and in data collecting devices originated a huge growth in the amount of spatial data available. Processing these amounts of data requires powerful data mining tools, which form the core of the spatial data mining field. Spatial data mining can be defined as the discovery of interesting relationships, spatial patterns and characteristics that may exist in spatial databases (e.g. Miller & Han, 2001).

One of the most used data mining techniques is clustering. Clustering is a well-established scientific field (Fisher, 1936; Kaufman & Rousseeuw, 1990) allowing the partition of data into groups of similar objects. These objects are usually represented as a vector of measurements or a point in a multidimensional space (Jain, Murty, & Flynn, 1999). Spatial clustering (Han, 2005) is the partition of spatial objects into groups so that objects within a cluster are as similar as possible. Due to spatial dependency, an intrinsic characteristic of spatial data explained by the 1st law of geography (Tobler, 1970), clusters are expected to be grouped in space. Tobler's first law (TFL) states that "*everything is related to everything else, but near things are more related than distant things*". Although Tobler himself (Tobler, 2004) recognizes the first part of TFL is not always true (Sui, 2004), correlation is likely to be higher at short distances.

In spite of TFL we often see clusters produced from spatial datasets which are not spatially contiguous. Some of the known causes are: (1) the aggregation and the scale of data (Openshaw, 1984); (2) the spatial heterogeneity (Anselin, 1988); and (3) the multivariate nature of the clustering.

The problems raised by the aggregation and the scale of data are known as the modifiable areal unit problem (MAUP) (Openshaw, 1984). The problem is that spatial phenomena are normally continuous, but have to be aggregated to obtain a manageable discrete description. The exact outline of the area over which the description is obtained will influence critically the perception of the phenomena. Differences in scale will have a similar effect since they also imply a change in the outline.

Spatial heterogeneity is the property that makes each place on Earth unique due to its specific attributes (Anselin, 1988). This variation implies that standards and design decisions successfully adopted in one region cannot always be generalized and applied in other regions (Goodchild, 2008). This uniqueness of each place makes spatial clustering an even more complex task.

* Corresponding author. Address: Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal. Tel.: +35 1213870413; fax: +35 1213872140.

E-mail addresses: roberto@isegi.unl.pt (R. Henriques), bacao@isegi.unl.pt (F. Bacao), vlobo@isegi.unl.pt (V. Lobo).

0198-9715/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2011.11.003

The third problem with spatial clustering is that different variables (in a multidimensional problem) may have different levels of spatial autocorrelation, and thus the global spatial autocorrelation depends a lot on the relative importance given to each of them. Even in the case when all variables share a similar global spatial autocorrelation (O'Sullivan & Unwin, 2002), it is usually space dependent, and thus the local patterns of this dependency can be very different.

Nevertheless, many applications require spatially contiguous clusters that contain regions as homogeneous as possible (within each cluster), separated from each other by discrete boundaries. Same examples of these applications are image segmentation (Awad, Chehdi, & Nasri, 2007), creation of areas for precision farming (Fleming, Heermann, & Westfall, 2004), estuarine management areas (Bação, Caeiro, Painho, Goovaerts, & Costa, 2005) and zone design problems (Bação, Lobo, & Painho, 2005a; Cockings & Martin, 2005; Openshaw, 1977).

Several methods are available for spatial clustering (Guha, Rastogi, & Shim, 1998; Hu & Sung, 2005; Ng & Jiawei, 2002; Sander, Ester, Kriegel, & Xu, 1998; Sheikholeslami, Chatterjee, & Zhang, 1998). For a more detailed survey on available methods, the reader is referred to (Han, Kamber, & Tung, 2001).

However, many of these methods are not aware of spatial dependence and spatial heterogeneity, assuming that space coordinates are just two (or three) more variables. These methods are based on general-purpose clustering methods which have limited capabilities in recognizing spatial patterns that include neighbors (Guo, Peuquet, & Gahegan, 2003).

GeoSOM, proposed in (Bação, Lobo, & Painho, 2005b; Bação, Lobo, & Painho, 2008), is an extension of Self-Organizing Maps (SOM). It is specially oriented towards spatial data mining. As one of the most known unsupervised artificial neural networks, SOM has been successfully applied to a wide array of spatial data (Bação et al., 2008). GeoSOM, while implementing SOM, recognizes the special inter-relation of spatial dimensions and the importance of this sub-space in the Geographer's analyses. GeoSOM takes into account Tobler's first Law, searching for clusters within certain (but adaptable) geographic boundaries instead of global clusters produced by standard SOM.

This paper extends and consolidates (Bação et al., 2008) in two major ways. First, a tool called GeoSOM suite is presented, integrating features of Artificial Intelligence-based clustering with features of Geographic Information Systems (GISs). This tool implements the standard SOM and the GeoSOM algorithm with a few improvements providing a friendly and ready to use environment for spatial data exploration. Some of the improvements on the GeoSOM are: (1) a tool for cluster outline on a graphical representation of the SOM; (2) auxiliary tools to help on that outline, such as hierarchical clustering; (3) inclusion of parallel coordinate plots (Inselberg, 1985); (4) visualization of the mapping of input data combined with the defined clusters; and (5) possibility of viewing multiple SOM, trained with the same data but different parameters, at the same time. GeoSOM suite enables the user to interact with data and combine multiple clustering solutions, thus gathering knowledge about data and the clusters produced. By providing this exploratory environment GeoSOM Suite fulfils a gap pointed out by Spielman and Thill (2008) in which the connection between the SOM and GIS is usually difficult to achieve, requiring, most of the time, scripting and considerable labor.

Second, this paper assesses GeoSOM suite using Lisbon's census dataset, showing that it is a useful exploratory spatial data analysis (ESDA) and clustering tool.

The paper is organized as follows: Section 2 presents prior work relevant for this paper. Section 3 reviews the SOM and GeoSOM methods. In Section 4, two datasets, used to exemplify this tool, are presented. Section 5 presents GeoSOM suite in detail, and Section 6 demonstrates a case study using Lisbon Metropolitan Area (LMA) 2001 census dataset. Finally, Section 7 concludes the paper and discusses future work.

2. Related work

According to (Guo & Gahegan, 2006), when analyzing geo-referenced data, there are three ways to combine spatial and non-spatial variables. These are: (1) embed the spatial information as *normal* variables (and for that they proposed encoding and ordering spatial data in a particular way); (2) create new data mining algorithms that take into account both types of characteristics, treating spatial variables in a special way; or (3) use multiple views to visually link patterns across different spaces (spatial and non-spatial).

Several tools combining exploratory spatial data analysis and data mining methods have been proposed. One of the oldest tools is GeoMiner (Han, Koperski, & Stefanovic, 1997), which is based on a relational data mining system known as DbMiner (Han, Cai, & Cercone, 1993). GeoMiner proposed a new language (geographic mining query language) to define characteristic rules, comparison rules and association rules. Another characteristic of this system is the integration of data mining, data warehousing technologies and geographic information systems, presenting various outputs, such as maps, tables and charts.

(Maceachren, Wachowicz, Edsall, Haug, & Masters, 1999) proposed the GKConstruck, allowing the integration of knowledge discovery in databases (KDD) and geographic visualization (GVis), with spatiotemporal environmental data. The authors proposed a prototype capable of presenting three dynamically linked representation forms: the geographic map, 3D scatter plots and parallel coordinate plots. These three linked windows allow spatial data exploration through dynamic brushing, focusing and color manipulation.

Another tool for spatial data analysis and visualization is GeoVista Studio (Takatsuka & Gahegan, 2002). In this tool, the user is able to build his own exploratory methods by visual programming. Dynamically linked visual representations such as maps, scatter plots and parallel coordinate visualizations are used for exploration and analysis.

Anselin proposed the GeoDA tool (Anselin, Syabri, & Kho, 2006), including histograms, box plots, scatter plots, choropleth maps, global and local indicators of spatial association (LISA) (Anselin, 1993) and spatial regression. This tool also makes use of dynamically linked windows, combining maps with statistical plotting.

In a recent paper, Mu (Mu & Wang, 2008) proposed a scale-space clustering method for spatial data. This method produces several clustering sets for different scales just like in hierarchical clustering. At the top of the hierarchy there is only one cluster, and at the base the number of clusters is equal to the number of data objects. The method starts by calculating aggregation scores based on the characteristics of each object and its neighbors. These scores allow the creation of directional links, which enables the definition of local minima and maxima: local minima are objects with all directional links pointing towards other objects while local maxima are objects with all directional links pointing towards itself. In the next phase, the method groups objects iteratively, from local minima to local maxima, according to the directional links. This method has, amongst others, the advantage of producing clusters that are always spatially contiguous.

Self-Organizing Maps (SOM) have been used more and more in geospatial problems, and a good overview of these is presented in (Agarwal & Skupin, 2008). Openshaw was one of the first well-known geographers to point out the applicability of SOM in geography, namely for clustering (Openshaw & Wymer, 1995). Other

geospatial clustering applications of SOM include (Céréghino, Santoul, Compin, & Mastrorillo, 2005; Koua & Kraak, 2004; Spielman & Thill, 2008). In (Allouche & Moulin, 2005) and (Sester, 2005) SOMs are used for cartographic generalization. SOMs have also been used as supervised classification tools for geospatial problems, for example in (Mather, Tso, & Koch, 1998; Merenyi, Jain, & Villmann, 2007; Wan & Fraser, 1993). In (Spielman & Thill, 2008) SOM are used for data reduction, allowing the detection of spatial patterns in a socio-demographic analysis of New York City census data. Similar uses include the analysis of airline passenger flows (Yan & Thill, 2008) and linguistic variations (Thill, W.A. Kretzschmar Jr., & X. Yao, 2008).

3. Outline of SOM

Teuvo Kohonen proposed the Self-Organizing Maps (SOM) in the beginning of the 1980s (Kohonen, 1982). The SOM is usually used for mapping high-dimensional data into one, two, or three-dimensional feature maps. The basic idea of an SOM is to map the data patterns onto an n-dimensional grid of units or neurons. That grid forms what is known as the output space, as opposed to the input space that is the original space of the data patterns. This mapping tries to preserve topological relations, *i.e.* patterns that are close in the input space will be mapped to units that are close in the output space, and *vice versa*. The output space will usually be two-dimensional, and most of the implementations of SOM use a rectangular grid of units. To provide even distances between the units in the output space, hexagonal grids are sometimes used (Kohonen, 2001). Each unit, being an input layer unit, has as many weights as the input patterns, and can thus be regarded as a vector in the same space of the patterns. When training an SOM with a given input pattern, the distance between that pattern and every unit in the network is calculated. While several distance metrics can be used (Kohonen, 2001; Sneath & Sokal, 1973), Euclidean distance is the most common, Then the algorithm selects the unit that is closest as the winning unit (also known as best matching unit-BMU), and that pattern is mapped onto that unit. In a successfully trained SOM patterns that are close in the input space will be mapped to units that are close (or the same) in the output space. Thus, SOM is 'topology preserving' in the sense that (as far as possible) neighborhoods are preserved through the mapping process.

The basic SOM learning algorithm may be described as follows:

Let
$\boldsymbol{X}$ be the set of n training patterns $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$
$\boldsymbol{W}$ be a $p \times q$ grid of units $\boldsymbol{w_{ij}}$ where i and j are their coordinates on that grid
α be the learning rate, assuming values in $]0,1[$, initialized to a given initial learning rate
r be the radius of the neighborhood function $h(\boldsymbol{w_{ij}}, \boldsymbol{w_{mn}}, r)$, initialized to a given initial radius
1 Repeat
2 For $m = 1$ to n
3 For all $\boldsymbol{w_{ij}} \in \boldsymbol{W}$,
4 Calculate $d_{ij} = ||\boldsymbol{x_m} - \boldsymbol{w_{ij}}||$
5 Select the unit that minimizes d_{ij} as the winner $\boldsymbol{w_{winner}}$
6 Update each unit $\boldsymbol{w_{ij}} \in \boldsymbol{W}$: $\boldsymbol{w_{ij}} = \boldsymbol{w_{ij}} + \alpha\, h(\boldsymbol{w_{winner}}, \boldsymbol{w_{ij}}, r)||\boldsymbol{x_m} - \boldsymbol{w_{ij}}||$
7 Decrease the value of α and r
8 Until α reaches 0

The learning rate α, sometimes referred to as η, varies in $[0,1]$ and must converge to 0 to guarantee convergence and stability in the training process. The decrease of this parameter to 0 is usually done linearly, but any other function may be used. The radius, usually denoted by r, indicates the size of the neighborhood around the winner unit in which units will be updated. This parameter is relevant in defining the topology of the SOM, deeply affecting the output space unfolding.

The neighborhood function h, sometimes referred to as Λ or N_c, assumes values in $[0,1]$, and is a function of the position of two units (a winner unit, and another unit), and radius, r. It is large for units that are close in the output space, and small (or 0) for far-away units.

Table 1
Squareville uniformly distributed attributes.

Variable	$35 < x < 65$	$0 < x < 34$ and $66 < x < 100$
Average salary	[0, 100]	[900, 1000]
Number of children	[0, 3]	[1, 5]
Education level	[4, 18]	[0, 9]
Number of residents	[2, 5]	[2, 7]
Number of rooms	[1, 5]	[0, 3]

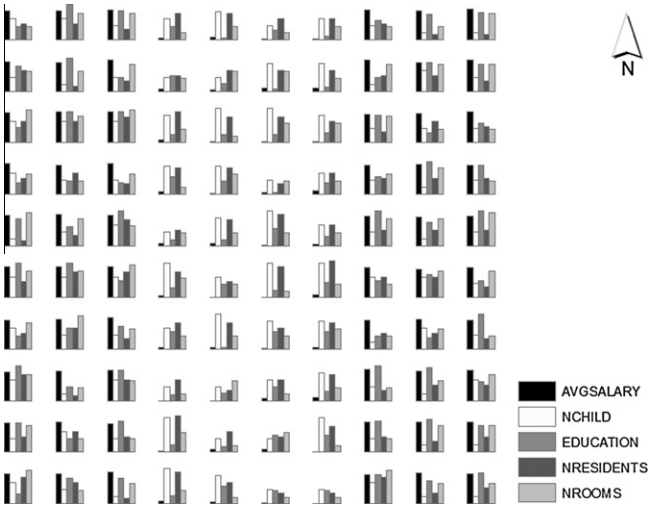

Fig. 1. Squareville house's spatial distribution (the bar chart position represent the geographic coordinates while its bars represent the non-spatial variables).

3.1. GeoSOM

GeoSOM is an adaptation of SOM to consider the spatial nature of data. In GeoSOM, the search for the best matching unit (BMU) has two phases. The first phase settles the geographical neighborhood where it is admissible to search for the BMU, and the second phase performs the final search using the other components. A parameter k controls the search neighborhood defined in the output space. The purpose of this k parameter, and the choice of its value for a given problem is discussed in detail in (Bação, Lobo, & Painho, 2004). The general idea is that instead of defining a fixed geographical neighborhood radius where clustering is admissible, that neighborhood is indirectly defined by fixing a neighborhood in the output space. In areas where data density is high, a given k-radius in the output space will represent a rather small geographic neighborhood, meaning that we will only allow clustering of data that are quite close by. On the contrary, in areas with low data density the same given k-radius in the output space will lead to large geographical neighborhoods, meaning that we will allow clustering of geographically more distant data. Using $k = 0$ will necessarily select as BMU the unit geographically closer. The same result may be obtained by training a standard SOM with only the geographical locations, and then

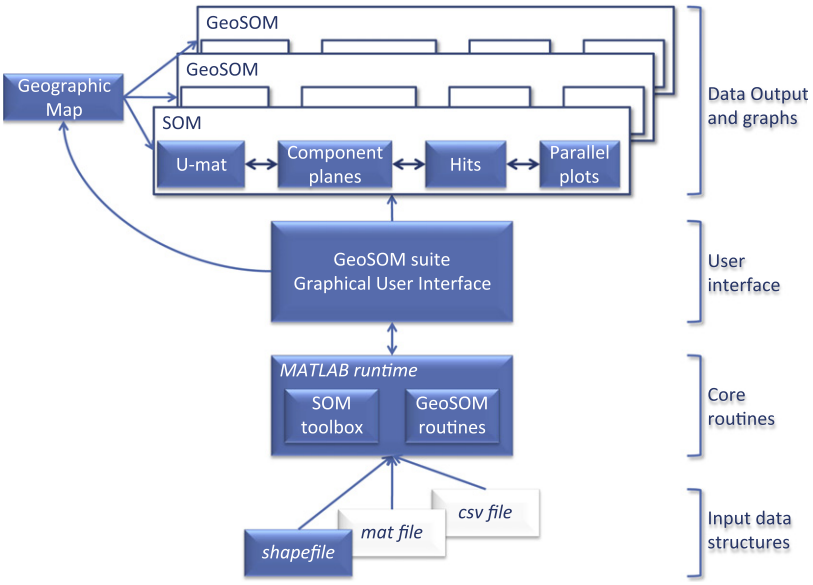

Fig. 2. GeoSOM suite architecture.

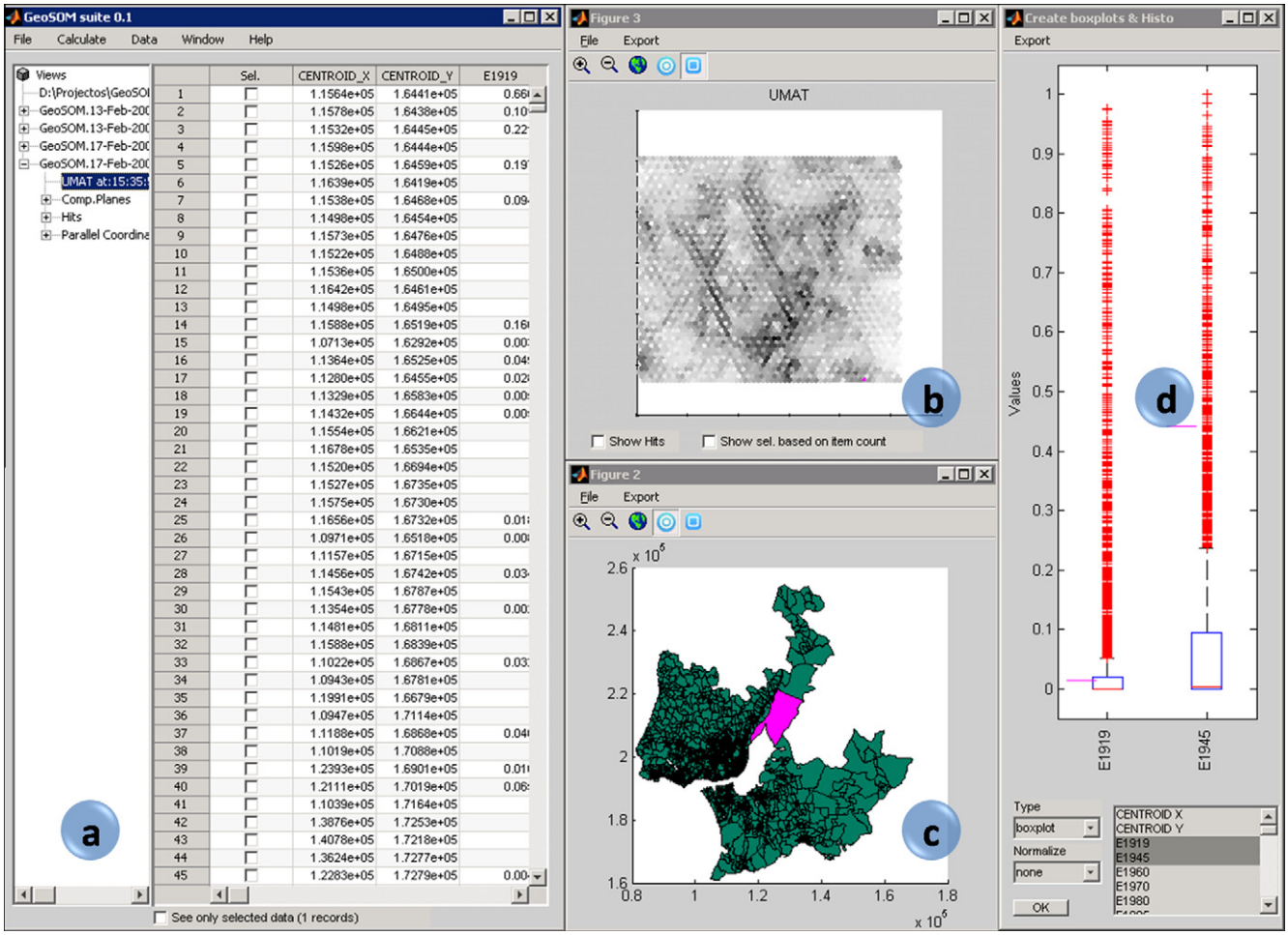

Fig. 3. GeoSOM suite window. From the left to the right, top to bottom: GeoSOM suite main window (a) with a tree-list of available analysis, and the full dataset with all attributes; U-matrix (b) obtained using census data; geographic map (c) of Lisbon Metropolitan Area; and a boxplot (d) showing the distribution of two variables.

using each unit as a low pass filter (*i.e.* a sort of average) of the non-geographic features. As k (the geographic tolerance) increases, the unit locations will no longer be quasi-proportional to the locations of the training patterns, and the "equivalent filter" functions of the units will become more and more skewed, eventually ceasing to be useful as models.

Setting k equal to the size of the SOM is equivalent to treat spatial coordinates as any other variable.

Formally, the GeoSOM may be described by the following algorithm:

Let
$\boldsymbol{X}$ be the set of n training patterns $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$, each of these having a set of geospatial components $\boldsymbol{geo_i}$ and another set of non-geospatial components $\boldsymbol{ngf_i}$.
$\boldsymbol{W}$ be a $p \times q$ grid of units $\boldsymbol{w_{ij}}$ where i and j are their coordinates on that grid, and each of these units having a set of geospatial components $\boldsymbol{wgeo_{ij}}$ and another set of non-geospatial components $\boldsymbol{wngf_{ij}}$, $\boldsymbol{w_{ij}} = [\boldsymbol{wgeo_{ij}}\ \boldsymbol{wngf_{ij}}]$.
α be the learning rate, assuming values in]0, 1[, initialized to a given initial learning rate
r be the radius of the neighborhood function $\boldsymbol{h}(\boldsymbol{w_{ij}}, \boldsymbol{w_{mn}}, r)$, initialized to a given initial radius
k be the radius of the geographical BMU that is to be searched
1 Repeat
2 For $m = 1$ to n
3 For all $\boldsymbol{w_{ij}} \in \boldsymbol{W}$,
4 Calculate $d_{ij} = ||\boldsymbol{wgeo_m} - \boldsymbol{wgeo_{ij}}||$
5 Select the unit that minimizes d_{ij} as the geowinner $\boldsymbol{w_{BMUgeo}}$
6 Select a set $\boldsymbol{W_{BMU}}$ *of* $\boldsymbol{w_{ij}}$ such that the distance in the grid between $\boldsymbol{w_{BMUgeo}}$ and $\boldsymbol{w_{ij}}$ is smaller or equal to k.
7 For all $\boldsymbol{w_{ij}} \in \boldsymbol{W_{BMU}}$,
8 Calculate $d_{ij} = ||\boldsymbol{x_m} - \boldsymbol{w_{ij}}||$
9 Select the unit that minimizes d_{ij} as the winner $\boldsymbol{w_{BMU}}$
10 Update each unit $\boldsymbol{w_{ij}} \in \boldsymbol{W}$: $\boldsymbol{w_{ij}} = \boldsymbol{w_{ij}} + \alpha h(\boldsymbol{w_{BMU}}, \boldsymbol{w_{ij}}, r)\ ||\boldsymbol{x_m} - \boldsymbol{w_{ij}}||$
11 Decrease the value of α and r
12 Until α reaches 0

4. Example datasets used in this paper

In this paper, we use two different datasets to illustrate the use of the GeoSOM suite. The first dataset (Squareville) is a fictional example (Lobo, Bação, & Henriques, 2009), consisting of data points with two spatial variables (the x–y coordinates) and five non-spatial variable. The second dataset is taken from the Lisbon Metropolitan Area (LMA) census for 2001, and it is used for a more detailed case study presented at the end of the paper.

4.1. Squareville dataset

Squareville is a fictional dataset benchmark used in spatial clustering problems. Squareville is a small town with square boundaries and 10,000 m^2 of area. Squareville has 100 houses evenly spaced with coordinates $x \in [5, 95]$ and $y \in [5, 95]$. For each house we know five attributes: the average salary of its residents, the number of children, the education level, the number of residents and the number of rooms. Table 1 presents the value intervals used for each variable and within those intervals the variables haven a uniform distribution.

Fig. 1 shows the houses' distribution and the value for each attribute along Squareville.

We could consider a case where only non-spatial attributes are used, *i.e.*, where we perform a traditional, non-spatial, clustering of the data. There is, of course, a continuum between this situation and the case where only spatial attributes exist, *i.e.*, where non-spatial variables are negligible. In this latter case, any clusters arise simply from geographical proximity. In this paper, we do not wish to discuss this problem in detail, but there is no universally optimum way to weigh spatial and non-spatial variables, as we go from pure attribute clustering to spatial clustering. Our objective is to consider situations where the main focus is on non-spatial variables, but these must be considered in their geographical context. We will use this dataset when explaining how to use GeoSOM Suite.

5. GeoSOM suite tool

The GeoSOM suite is implemented in Matlab® and uses the public domain SOM toolbox (Vesanto, Himberg, Alhoniemi, & Parhankangas, 1999). Basically, it consists of a number of Matlab routines (*m-files*). A stand-alone graphical user interface (GUI) was built, allowing non-programming users to evaluate the SOM and GeoSOM algorithms, and explore them with basic GIS tools. The GeoSOM suite is freely available at www.isegi.unl.pt/labnt/geosom. Fig. 2 shows the general GeoSOM suite architecture that consists of: (1) access to spatial and non-spatial data; (2) Matlab runtime components, SOM toolbox and GeoSOM routines; (3) a graphical user interface (GUI) and; (4) the routines that produce the output views. These views consist of geographic maps, U-matrices, component planes, the hit-map plots and parallel coordinate plots, which will be explained later. The GeoSOM suite allows multiple analyses to be shown at the same time. For example, one may use several different SOMs and GeoSOMs on the same dataset, and visually compare the results.

Fig. 3 presents a screen-shot of the GeoSOM suite tool. The main window contains a table of attributes and a tree view pointing to all the views created. The figure also shows three examples of views: a geographic data, a U-matrix and a box plot views.

The GeoSOM suite's main functionalities are: (1) present spatial data; (2) train a self-organized map using the standard SOM or the GeoSOM algorithm; (3) produce several representations (views) and (4) establish dynamic links between windows, allowing an interactive exploration of the data.

5.1. Views

Views are different representations of data allowing the user to analyze it from different perspectives, making interpretation easier. Presently, GeoSOM suite includes the following views:

- Geographic map
- U-matrices
- Component plane plots
- Hit-map plots
- Parallel coordinate plots
- Boxplots and histograms

U-matrices (Ultsch & Siemon, 1990) are calculated by finding the distances in the input space of neighboring units in the output space. The most common way to visualize them is to use a color scheme or a gray scale to represent these distances. In this case, black represents the highest value while white represents the lowest value (Fig. 4e). Low values in the U-matrix (shown as white areas) are an indication that data density is high, thus there is a cluster of data. High values in the U-matrix (shown as dark areas), are an indication that data density is low, thus there is a separation between clusters.

Component planes (Kohonen, 2001) are another SOM representation where each unit gets a color based on the weight of each variable used in the analysis. A component plane exists for each variable showing the units' weights for that variable (Fig. 4b). By observing the component planes one can see how a given variable varies along the map. This may be useful, for example, to understand what characterizes each cluster. By comparing two or more

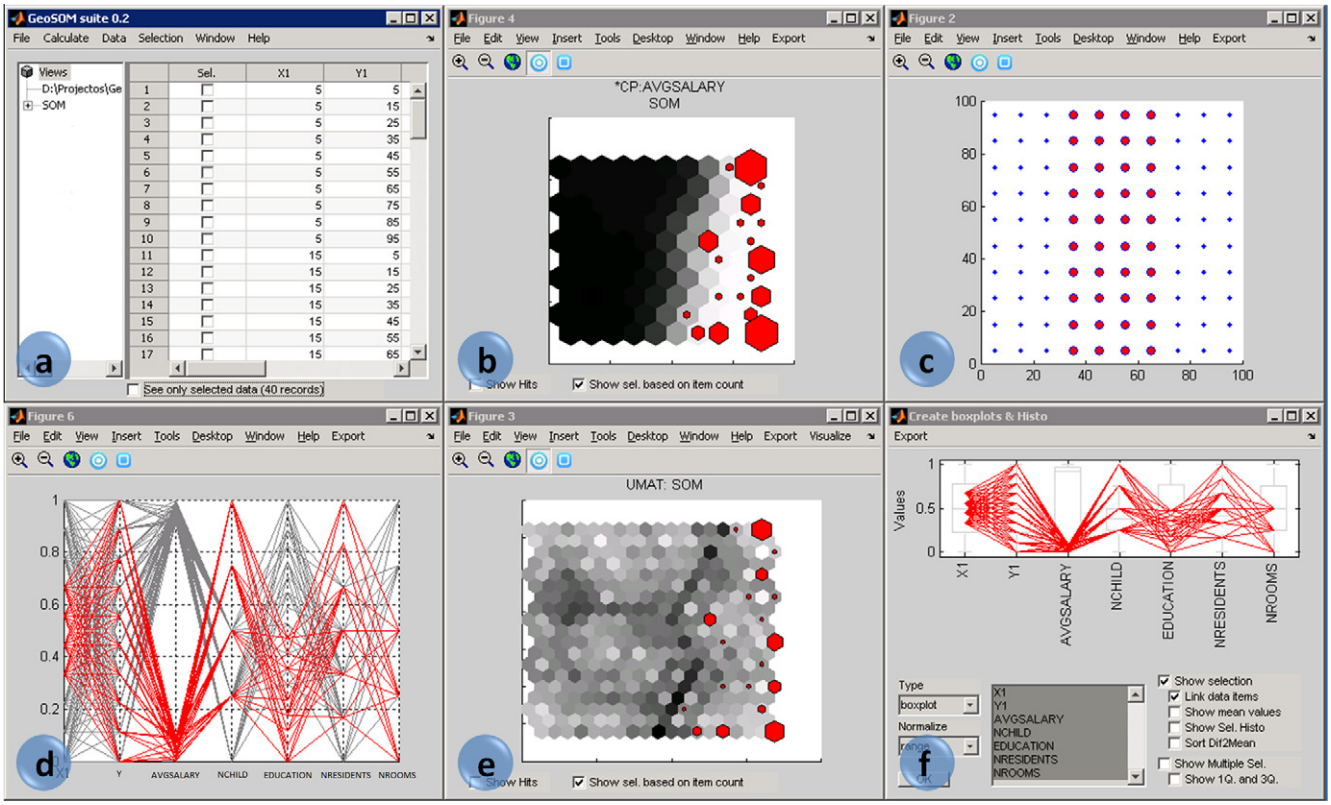

Fig. 4. Dynamically linked views created by GeoSOM suite (selection made in the U-matrix is in red): (a) GeoSOM suite main interface, with a tabular view of the dataset; (b) the average salary component plane; (c)the geographic map ; (d) parallel coordinate plot of all the data; (e) the U-matrix and (f) boxplot view of the seven variables.

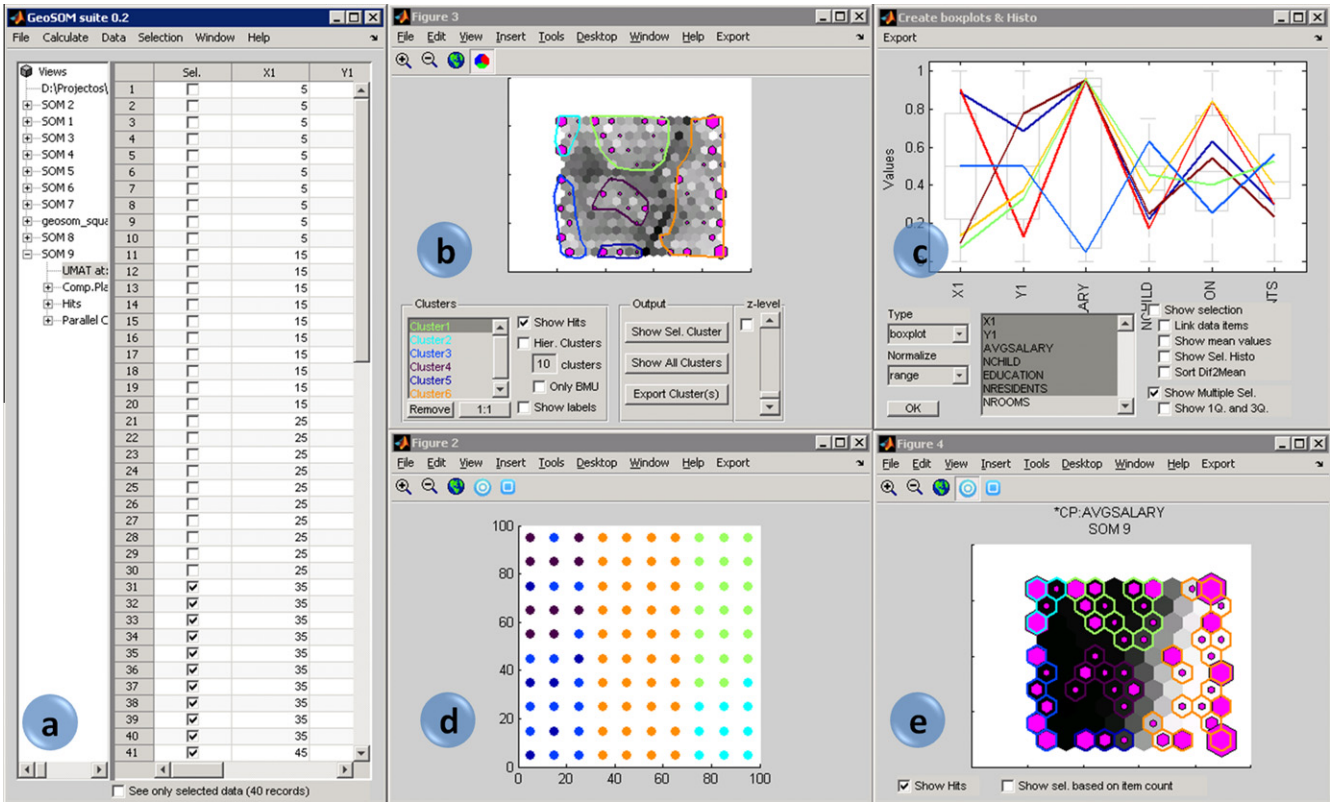

Fig. 5. Defining clusters from a standard SOM trained with Squareville data (a). In the figure, six clusters are delimited by the user on top of the U-matrix (b) produced from the SOM. The boxplot (c), the geographic map (d) and the average salary plane (e) are also presented showing the clusters.

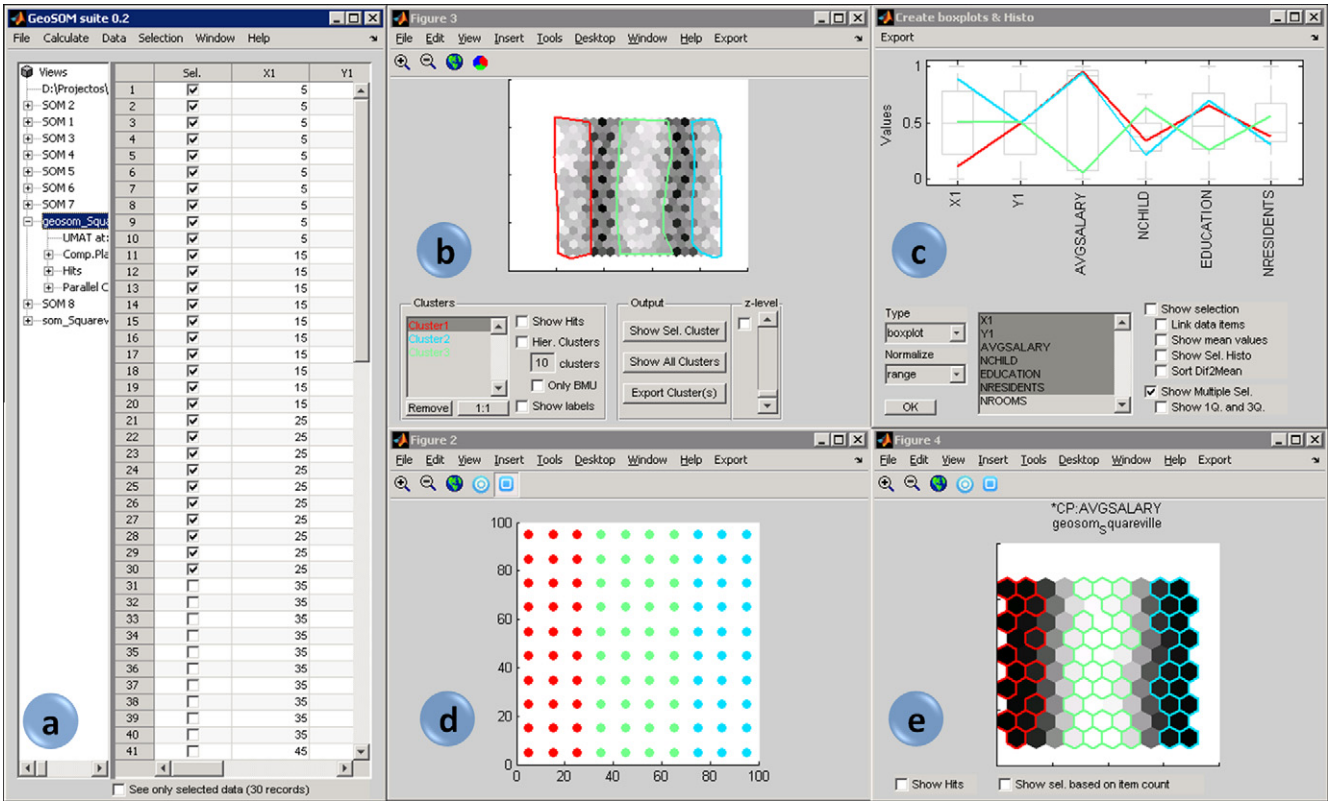

Fig. 6. Defining clusters from GeoSOM method trained with Squareville data (a). In the figure, three clusters (represented by red, green and blue) are delimited by the user on top of the U-matrix (b) produced from the SOM. The boxplot (c), the geographic map (d) and the average salary plane (e) are also presented showing the clusters.

component planes, one can visually identify correlations between variables, both globally and at a local scale.

Another possible view in the GeoSOM suite is the hit-map plot (Kohonen, 2001). This representation is usually superimposed on the U-matrix or on the component planes, and gives information about the number of data items represented by each unit, *i.e.* data items with the same BMU (Fig. 4b and e, red[1] hexagons). It can be used to see how a certain set of data points are mapped on the SOM, gaining more information about the clustering structure.

A parallel coordinate plot (Inselberg, 1985) is a data analysis technique for plotting multivariate data. This technique starts by drawing a set of parallel axis, one for each variable. A line connecting a given value on each variable axis will then map each data item (Fig. 4d). This allows us to visually compare multivariate data vectors.

Other possible representations in GeoSOM suite are the boxplots (also known as box plots, or box-and-whisker diagrams) and histograms (Fig. 4f). Boxplots are 2-dimensional graphics displaying several statistics for each variable (the smallest observation of a given variable, the lower quartile, the median, the upper quartile, the largest observation and the outliers). Besides the boxplot it is possible to plot the histogram, which is a graphical display of tabulated frequencies, shown as bars. Fig. 4 shows some views created by GeoSOM suite using the Squareville dataset. In this example, we trained an SOM with 10×10 units. Also in this figure, it is possible to notice the dynamically linked property of the views allowing the brushing of data items through different views.

The use of dynamically linked windows promotes interaction with the data, allowing users to analyze data from different perspectives. Observing the U-matrix (Fig. 4e) the first thing that emerges is that on the right hand-side we have a lighter area separated from the rest by a vertical dark region. This means that the units in this area form a cluster. Observing the component plane (Fig. 4b) we can confirm that the right hand-side cluster corresponds to low income houses. A closer inspection of the left hand-side area on the U-matrix would lead to it being divided into a upper part corresponding to houses from the East side and a lower part corresponding to houses on the West side. However this more detailed distinction is not very clear from the U-matrix. Selecting all units belonging to one of the clusters we are implicitly selecting a subset of the original data, and that data will be highlighted in all other views. Thus, when analyzing the salary component plane it is possible to find that the units selected on the U-matrix correspond to the lowest value of average salary. This is reinforced by inspecting the parallel coordinate plot, where low average salary houses are selected. Finally, in the geographic map, it is possible to view the spatial distribution of the houses with a lower average salary.

5.2. Clustering in GeoSOM suite

Clustering with SOM can be made using "*k*-means" SOM (Bação, Lobo, & Painho, 2005c) or "emergent" SOM (Ultsch, 2005). The distinction between these two approaches, which vary only in the number of units used, is not very common amongst the geospatial community, but is detailed in (Behnisch & Ultsch, 2008; Ultsch, 2005). In a "*k*-means" SOM, each unit is a cluster centroid (thus the process is similar to *k*-means clustering, hence the name), while in emergent SOM each cluster is composed of a large number of units, and identified by the borders on a U-Matrix. GeoSOM suite allows the users to use both methods for clustering. While "*k*-means" clustering does not require any special tool (each unit is

[1] For interpretation of color in Figs. 2–9 and 11–16, the reader is referred to the web version of this article.

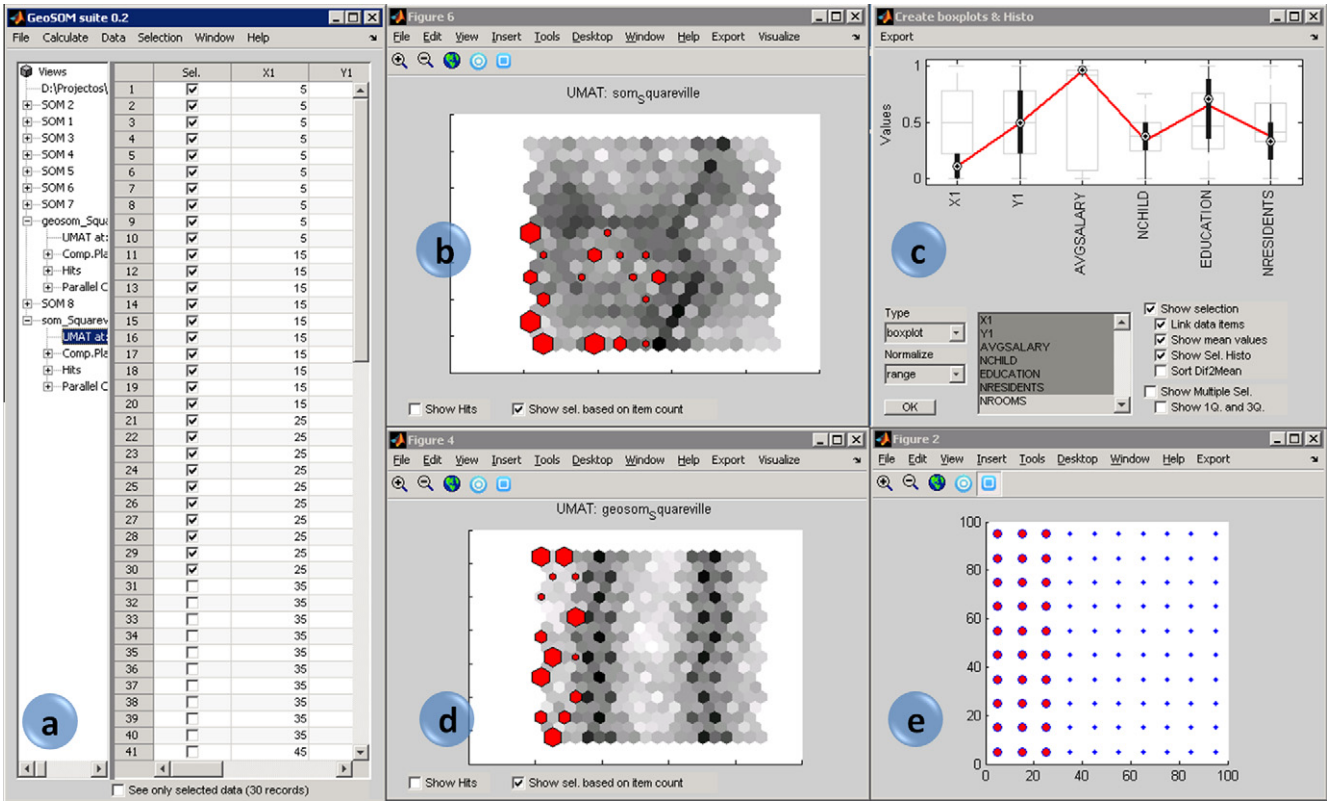

Fig. 7. Comparison between SOM and GeoSOM clustering. GeoSOM has the capability of detecting spatial contiguous clusters. The selection in red shows one region with high average salary in the west part of the map. (a) Main GeoSOM window; (b) U-matrix produced from a standard SOM; (c) boxplot; (d) U-Matrix produced from GeoSOM; and (e) geographic map.

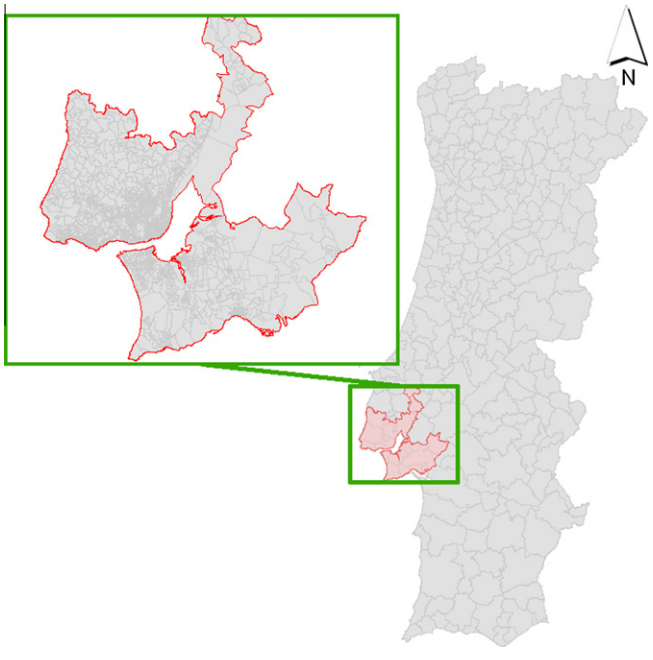

Fig. 8. Lisbon metropolitan area enumeration districts.

one cluster), to use "emergent" SOM clustering GeoSOM suite allows the user to delineate the clusters on top of U-matrices. To help users outline clusters, two extra tools are available in GeoSOM suite: a hierarchical clustering of units and what we call a *z-level tool*. In the first case, we cluster the units based on distance and position (on the U-matrix) using a hierarchical algorithm (single-linkage), and label each unit on the U-matrix. The *z*-level tool is a simple query on the U-matrix that highlights all units that are below a certain threshold. If we consider the U-matrix in three dimensions (assuming as third dimension the distance between units) high-density areas correspond to valleys while low-density areas correspond to mountains. Thus, clusters match valley zones while mountains are cluster frontiers. After defining clusters on top U-matrices, it is possible to see them on all open views (Fig. 5).

Fig. 5 shows a possible cluster configuration using the Squareville dataset. From the clusters' outline that is manually drawn on top of the U-matrix it is possible to analyze the clusters on the salary component plane, on the geographic map and on the parallel coordinate plot.

5.3. Clustering spatial data

As shown before, the standard SOM algorithm allows the detection of several clusters in the Squareville dataset. However, there is no unquestionable cluster arrangement and solutions using two, three or six clusters are possible.

A visual inspection of the variables distribution in the geographic map will suggest us three clusters. However, because in this example SOM is giving the same weight to each variable, spatial variables have proportionally less weight than the non-spatial variables (spatial variables are the *x* and *y* coordinates versus five non-spatial variables). Thus, in the U-matrix presented in Fig. 5, SOM is capturing the differences between the non-spatial variables which makes the clustering structure less clear.

At this point, a reasonable doubt can arise: if SOM does not detect the three clear-cut clusters due to the low weight of spatial variables, why not increase their weight? The answer to this

Table 2
Variables used in the cluster analysis of LMA census.

Category	Variable name	Description
Age of the buildings	E1945	% of buildings built before 1945
	E1970	% of buildings built between 1946 and 1970
	E1980	% of buildings built between 1971 and 1980
	E1990	% of buildings built between 1981 and 1990
	E2001	% of buildings built between 1991 and 2001
Age of residents	Id0_13	% of residents with age bellow 13
	Id14_19	% of residents with age between 14 and 19
	Id_19_24	% of residents with age between 20 and 24
	Id_25_64	% of residents with age between 25 and 64
	Id_65	% of residents with more than 65 years of age
Residents' level of education	Ens0	% of resident with no formal education
	EnsBas1	% of resident with 4 years of education
	EnsBas23	% of resident with 6–8 years of education
	EnsSec	% of resident with 12 years of education
	EnsSup	% of resident with higher education
Residents attending school	EstBas1	% of students attending years 1–4
	EstBas2	% of students attending years 5–6
	EstBas3	% of students attending years 7–9
	EstSec	% of students attending years 10–12
	EstSup	% of students attending University
Sector of economy	Sect1	% of residents working in the primary sector
	Sect2	% of residents working in the secondary sector
	Sect3	% of residents working in the tertiary sector
	PensRef	% of retired residents

question is that by increasing the importance of the spatial variables we will be decreasing the importance of the other variables, and thus blur the distinction between clusters defined by those variables. The clear distinction between clusters will fade away as the importance of the variables that define them decreases. In the limit, if only spatial variables are used, no clusters whatsoever will emerge since in this case spatial variables vary uniformly. It is not easy to find a point along this process where the clusters that arise are spatially continuous, but defined by non-spatial variables.

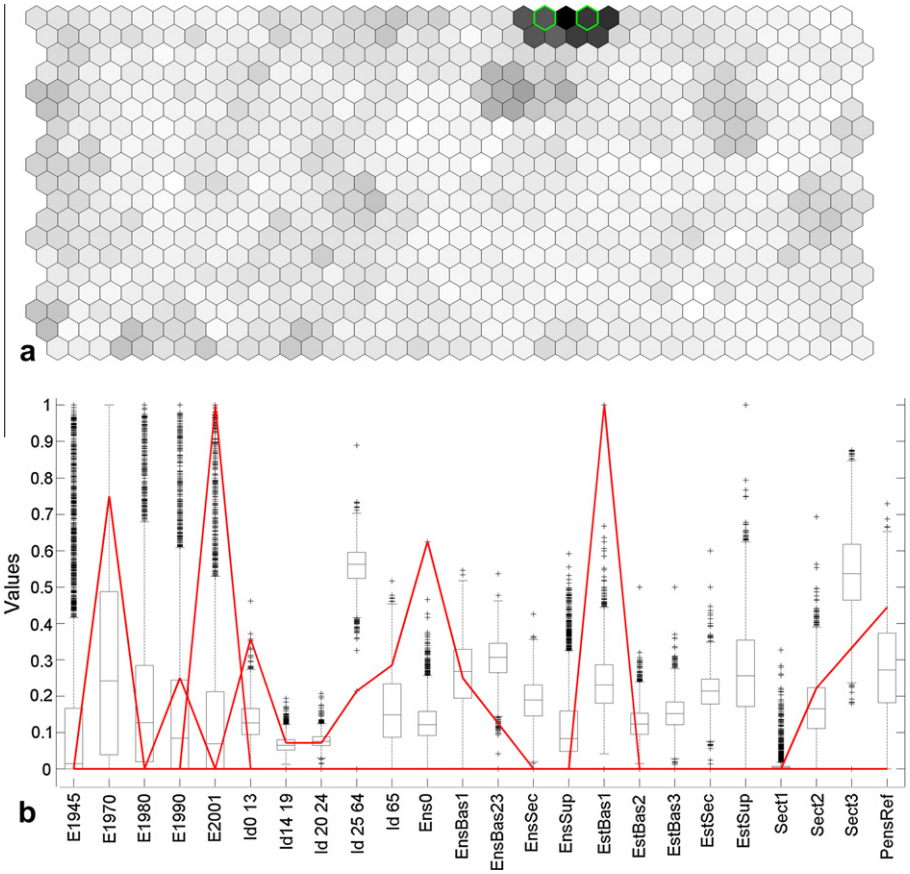

Fig. 9. U-matrix (a) for Lisbon Metropolitan Area SOM and box plot (b) showing the outliers (red features both in U-matrix and in the boxplot).

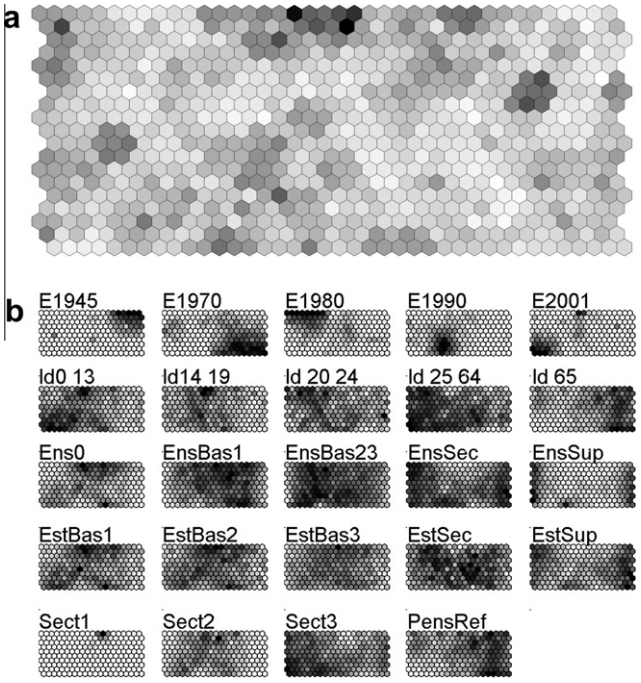

Fig. 10. U-matrix (a) and component planes (b) for Lisbon Metropolitan Area dataset after exclusion of the outliers. The top row of component planes refers to the age of the building. The next row refers to the age of the residents, the third one the student status, the forth the achieved education levels, and the last the employment sector.

To include spatial data and ensure the clusters' contiguity we apply the GeoSOM algorithm (Fig. 6).

In this case, three clusters are clearly detected (green, blue and red) matching the two spatially apart regions with high average salary and the middle one with a small average salary.

5.4. Combining multiple clustering solutions in GeoSOM suite

Another analysis possible with GeoSOM suite is to train several SOM or GeoSOM using the same dataset. This possibility has different applications. First, it allows the comparison between the "*k*-means" and "emergent" SOM clustering methods using the same dataset. Therefore, the user can compare the clusters produced using a predefined number of clusters with those obtained by searching for "natural" clusters.

This feature also allows a sensitivity analysis of SOM and GeoSOM by comparing results using different input parameters. Comparisons between SOM and GeoSOM algorithms in clustering can also be performed. Using multiple clustering options at the same time will give the user a better insight on the nature of the data. Using the two examples shown before (Figs. 5 and 6), it is possible to compare the SOM and GeoSOM results (Fig. 7), and thus understand the geographical separation of the high income cluster.

Finally, the user might choose to make several SOM's using different subsets of variables. This can be thought of as building different thematic classifications. For the census dataset that we will analyze later, for example, one may separately use building characteristics, family characteristics, or unemployment characteristics, which can, in the end, be evaluated together.

Fig. 7 shows the comparison between the SOM and the GeoSOM U-matrices. The difference in clarity between the SOM U-matrix and GeoSOM U-matrix (Fig. 8b and d) is striking. In both we can identify three clusters but they are far more pronounced in the GeoSOM than in the SOM. Selecting one cluster in the GeoSOM U-matrix (an area with a high salary), it is possible to analyze the cluster distribution on the previously trained SOM.

6. Case study: Lisbon's census

To evaluate GeoSOM suite with real data we performed a cluster analysis using Lisbon's Metropolitan Area (LMA) 2001 census, obtained from the Portuguese statistical institute (Statistics Portugal). The data is aggregated by 3978 enumeration districts (ED) (*secções estatisticas* in Portuguese) and describes buildings, families, households, age, education levels and economic activities using more than 65 variables. An ESRI™ shapefile with the ED's spatial outline and attributes is used as starting point in the GeoSOM suite. One may also use other file type, such as *.csv* or *.mat* files, but in that case the geographical map will not be produced. Fig. 8 presents a map with LMA delineation and relative location within Portugal.

Table 2 describes the variables used in the cluster analysis.

We started by training a 20×10 SOM using all enumeration districts (ED). Fig. 9 shows the U-matrix produced and a boxplot for all the variables. Next, we identified outliers by searching for high values in the U-matrix. As can be seen, in this case, the U-matrix has a very dark area (corresponding to very low density of data) at the top. The data that is mapped to this area are clearly

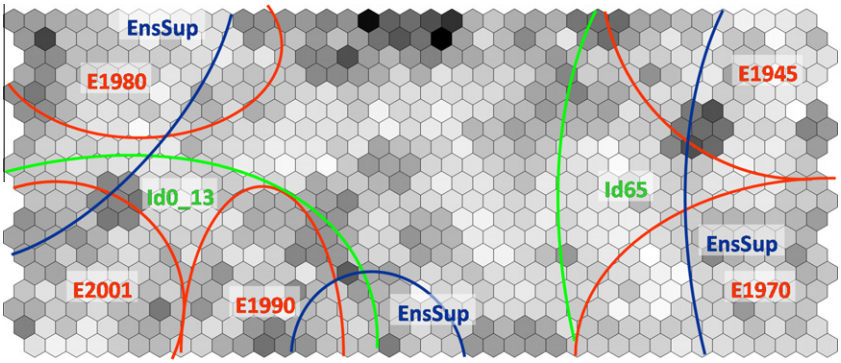

Fig. 11. U-matrix with outlines of some component planes hotspots. Areas of the component planes that have high values are shown with colors (one for each thematic group of variables) on top of the U-matrix. There are two areas where age (in green) plays a predominant role: on the right there is an area with many people over 65, and on the lower left an area with infants (under 13 years of age). There are three areas where education level (in blue) plays a predominant role: on the extreme right, upper left, and middle bottom, there are many people with tertiary education. Finally there are 5 areas (in red) where buildings have a well-defined age structure: in the top right there are many old buildings (built before 1945), in the bottom right buildings built before the 70s, in the top left, buildings of the 70s, in the middle-left bottom the 80s, and in the bottom left the 90s.

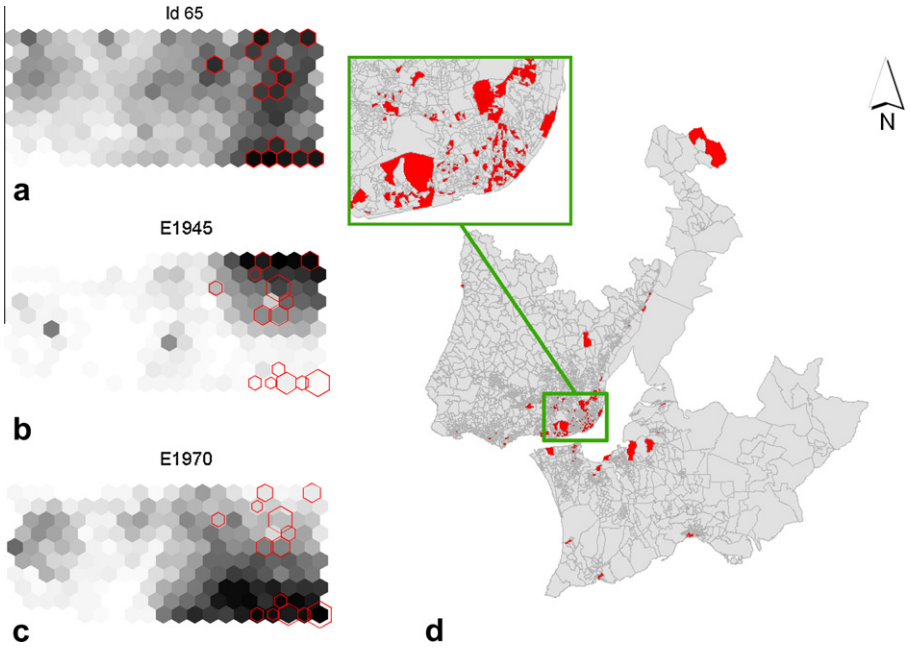

Fig. 12. Component planes for the variables Id65 (a), E1945 (b) and E1970 (c) and Lisbon map (d) showing the selection of the units with higher percentage of oldest people.

outliers, and the values of the variables that characterize them are superimposed on the boxplot, as red lines. As we can see in the boxplot these outliers (represented by the red lines) refer to ED where most variables are null or deserted areas where a single house can skew the results significantly. These ED were removed from the original dataset, so that a better understanding of the remaining ED can be achieved. This process of removing outliers before proceeding with the analysis is quite common, since the

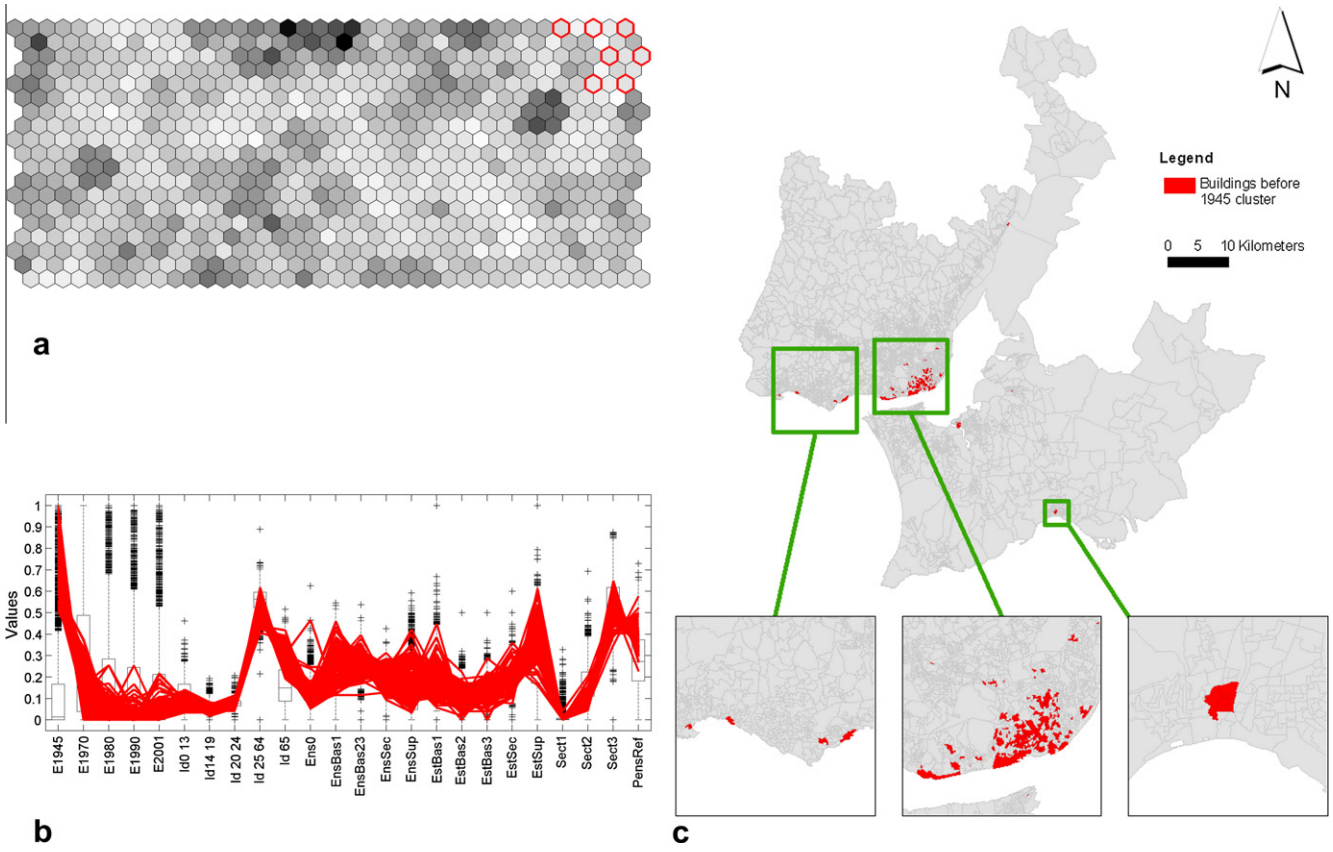

Fig. 13. U-matrix (a) and parallel coordinate plot (b) and Lisbon Metropolitan Area map (c) with the highest percentage of buildings built before 1945' enumeration districts in red. From the previous figure we conclude that the "old buildings" cluster formed by SOM is spatially distributed along Lisbon Metropolitan Area, matching the oldest town centers (Lisbon, Cascais, Oeiras Almada and Setubal).

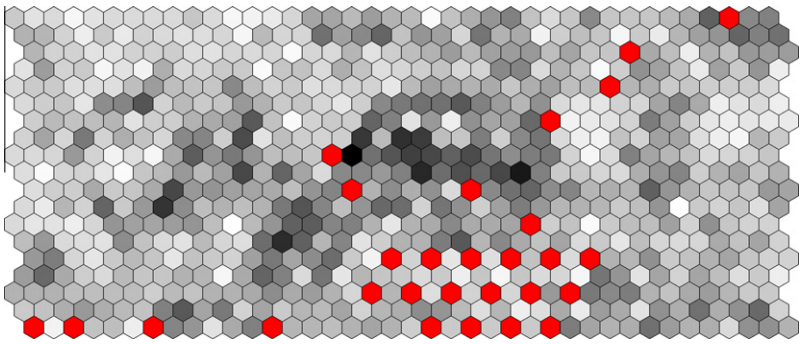

Fig. 14. U-matrix obtained with GeoSOM for Lisbon Metropolitan Area dataset after exclusion of the outliers. The original cluster of old buildings detected by the standard SOM is mapped to the red units.

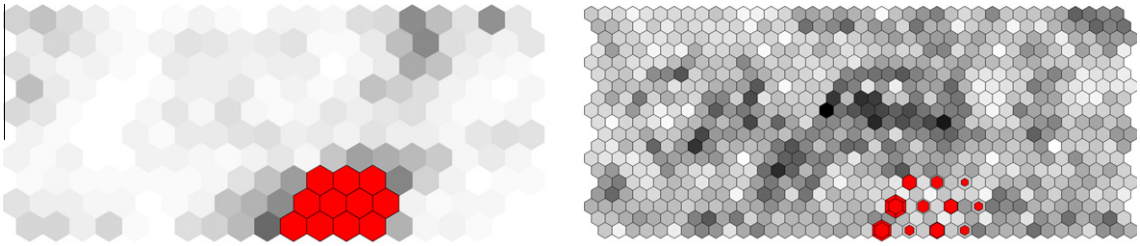

Fig. 15. Oldest buildings cluster selected on the ED1945 component plane (a) and on the U-matrix (b).

outliers will usually "squash" the rest of the data, thus hiding the general structure of the dataset.

After removing those outliers a new SOM was trained with the remaining ED. Again, for this training set, a 20 × 10 SOM was used. Fig. 10 shows the U-matrix produced and the component planes for all the variables used.

A detailed cross analysis between the U-matrix and the component planes reveals that the age of the buildings is the factor that better defines clear clusters, and thus may be dubbed more "*clusterable*". This means that these variables present natural clusters that are easily detected by SOM. Concerning the age of residents it is possible to conclude that young people (<13 years old) are more associated with "newer" areas (buildings built after the 90s) while older people (>65 years old) are found in areas with buildings built before 1970. As expected the number of students is highly related with the age of residents. Fig. 11 shows the U-matrix emphasizing units with high values for residents' age, buildings age and education related variables.

Fig. 12 shows the component planes for the variables Id65 (residents with more than 65 years of age), E1945 and E1970 (buildings built before 1940 and between 1940 and 1970, respectively). Selecting the units with higher values for the variable Id65 it is possible, due to dynamically linked windows of GeoSOM suite, to verify the corresponding selection on the E1945 and E1970 component planes. On the two last component planes the size of the red hexagons represent the number of ED selected. It is possible to conclude that ED with eldest people are highly related with those with buildings built before 1945 or before 1970. The spatial distribution of these ED is also shown in Fig. 12. The ED with higher percentages of older people are located mostly in the center of Lisbon.

In the following figure (Fig. 13) the cluster representing the highest percentage of buildings built before 1945 is selected on the U-matrix. In the same figure a box plot characterizing this cluster and its spatial distribution is shown.

From the previous figure we conclude that the "old buildings" cluster formed by SOM is spatially distributed along Lisbon Metropolitan Area, matching the oldest town centers (Lisbon, Cascais, Oeiras Almada and Setubal).

However, depending on the final objective, the creation of spatially homogeneous clusters can be an important goal in the cluster analysis. If, for example, we wanted to decide where an historical buildings visitor center should be located, we would like to select a cluster of old buildings that are spatially close to each other. The clusters are thus formed not only by the age dimension (age of the buildings) but also by their spatial location. To obtain these spatially homogenous clusters we applied the GeoSOM algorithm. A 20 × 10 GeoSOM was trained and the U-matrix produced is shown in Fig. 14.

On this U-matrix we selected (red hexagons) the "old buildings" that belong to the cluster detected by the standard SOM. The highlighted GeoSOM units corroborate the fact that this cluster is not spatially contiguous. In other words, since the GeoSOM U-matrix represents the distances between its units, and this distance takes into account the attributes and geographic distances, spatial homogeneous clusters are represented by units close to each other in the U-matrix. Fig. 15 shows the ED1945 component plane for the GeoSOM where high value units are selected in red. Corresponding features are also selected on the U-matrix produced from GeoSOM.

In Fig. 16, clusters were delineated on top of the U-matrix produced from the GeoSOM. From this partition, ED were colored on the map according to the respective cluster. A parallel coordinate plot is also shown, characterizing the units belonging to each cluster.

Analysing Lisbon Metropolitan Area 2001 census we may conclude that:

- Young people (<13 years old) are associated with "newer" areas (buildings built after the 90s), while older people (>65 years) are found in areas with buildings built before 1970.
- ED with eldest people are located throughout all LMA, but with special focus near Lisbon's center and centers of old villages such as Sintra, Cascais, Oeiras, Almada and Setubal.

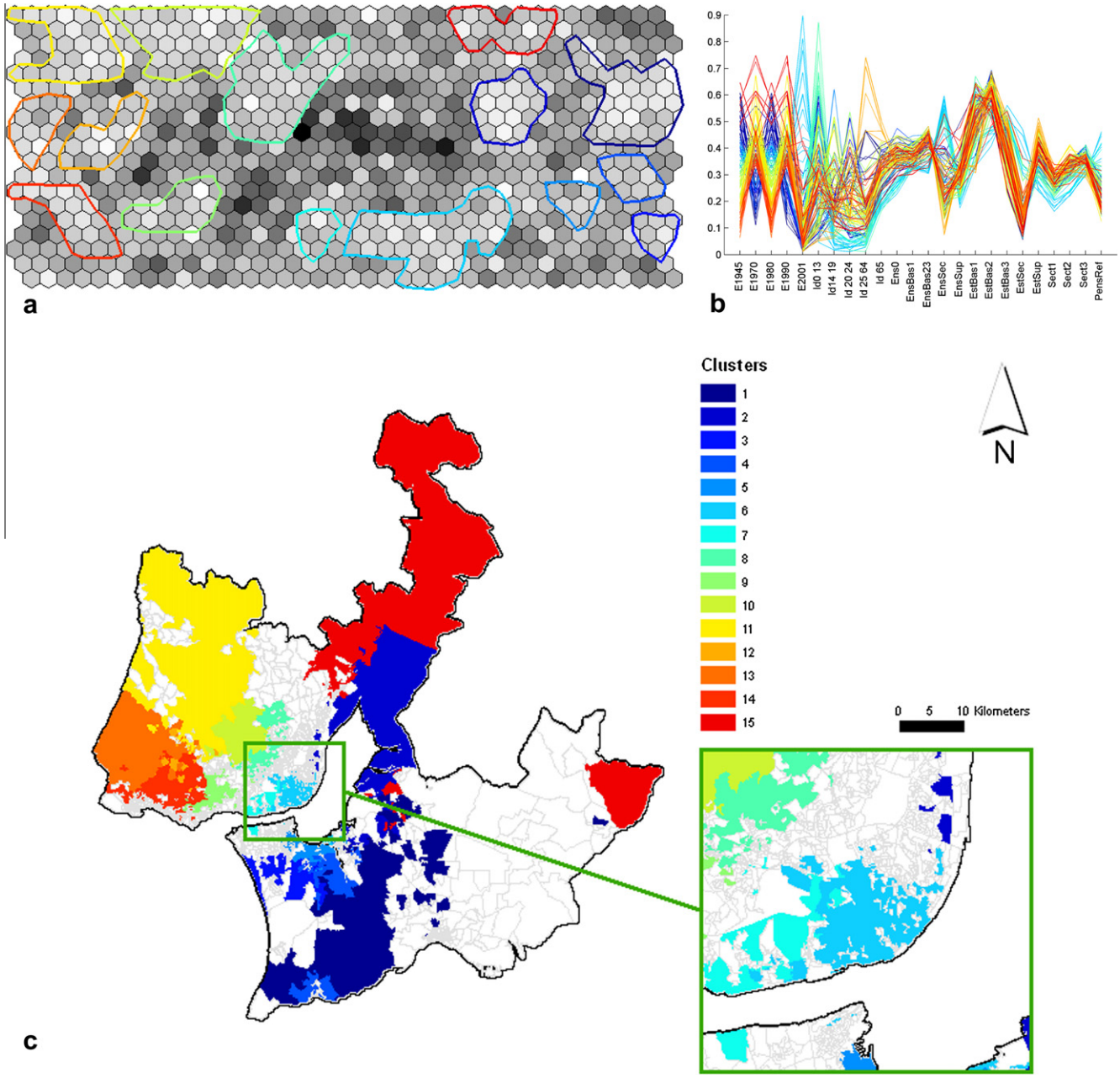

Fig. 16. Clusters created for Lisbon Metropolitan Area presented in the: (a) U-matrix; (b) parallel coordinate plot of units and (c)Lisbon Metropolitan Area map.

- SOM and GeoSOM produce different clusters: while SOM groups ED with similar characteristics albeit quite far from each other, GeoSOM groups nearby ED with similar characteristics.
- Using GeoSOM it is possible to create regions with similar attributes and high spatial autocorrelation (corresponding to GeoSOM clusters), but it is also possible to detect regions where the spatial autocorrelation is low (corresponding to areas outside the clearly defined clusters).

7. Conclusions

In this paper, we presented GeoSOM suite as a new and efficient tool for exploratory spatial data analysis (ESDA) and clustering. This tool implements two major methods, the standard SOM (Kohonen, 2001) and the GeoSOM (Bação et al., 2008). The SOM is a well-known algorithm that has proved to be of interest in spatial clustering. The GeoSOM, by explicitly considering spatial autocorrelation, is able to detect spatial homogeneous and heterogeneous areas. These heterogeneous areas are regions where, although spatial attributes are related (data points are close to each other) non-spatial attributes have little correlation.

GeoSOM suite implements several visualization features, all dynamically linked, allowing a strong interaction between user and data, and thus an improved understanding of the data analyzed. It is also possible to compare both methods through several views such as U-matrices, component planes, parallel coordinate plots, *etc.*

Spatial clusters were produced from Lisbon Metropolitan Area 2001 census dataset using the SOM and GeoSOM methods available in GeoSOM suite. Four main conclusions were drawn from this analysis as explained in the previous section.

The main conclusion is that GeoSOM suite not only is easy to use (it has been used extensively by our students), but provides a wide range of powerful tools that enable the user to detect

patterns that are hard to find using other methods. It constitutes a useful environment for exploratory geospatial analysis, even if the particularities of the GeoSOM algorithm are not used. A second important conclusion is that, as suggested in earlier papers, GeoSOM in real world problems does produce clusters that, while defined by non-spatial attributes, are geographically compact.

Future research will focus on evaluating the performance of GeoSOM with factors such as scale, zoning and time. In addition, some research will be done on using GeoSOM to detect clusters in different thematic areas and to produce general clusters based on these "lower level" clusters.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compenvurbsys.2011.11.003.

References

Agarwal, P., & Skupin, A. (2008). *Self-organising maps: Applications in geographic information science*. Wiley.

Allouche, M. K., & Moulin, B. (2005). Amalgamation in cartographic generalization using Kohonen's feature nets. *International Journal of Geographical Information Science, 19*(8), 899–914.

Anselin, L. (1988). *Spatial econometrics: Methods and models (studies in operational regional science)*. Springer.

Anselin, L. (1993). Local indicators of spatial association – LISA. Geographic information systems data (GISDATA) specialist meeting on geographic information systems (GIS) and spatial analysis. Amsterdam, Netherlands: Ohio State Univ Press.

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis, 38*(1), 5–22.

Awad, M., Chehdi, K., & Nasri, A. (2007). Multicomponent image segmentation using a genetic algorithm and artificial neural network. *Geoscience and Remote Sensing Letters, IEEE, 4*(4), 571–575.

Bação, F., Caeiro, S., Painho, M., Goovaerts, P., & Costa, M. (2005). Delineation of estuarine management units: Evaluation of an automatic procedure. In P. Renard, H. Demougeot-Renard, & R. Froidevaux (Eds.), *Geostatistics for environmental applications* (pp. 429–441). Netherlands: Springer-Verlag.

Bação, F., Lobo, V., & Painho, M. (2004). Geo-self-organizing map (Geo-SOM) for building and exploring homogeneous regions. *Geographic Information Science, Proceedings, 3234*, 22–37.

Bação, F., Lobo, V., & Painho, M. (2005a). Applying genetic algorithms to zone design. *Soft Computing, 9*, 341–348.

Bação, F., Lobo, V., & Painho, M. (2005b). The self-organizing map, the Geo-SOM and relevant variants for geosciences. *Computers and Geosciences, 31*, 155–163.

Bação, F., Lobo, V., & Painho, M. (2005c). Self-organizing maps as substitutes for K-means clustering. In V. S. Sunderam, G. v. Albada, P. Sloot, & J. J. Dongarra (Eds.). *Lecture notes in computer science* (Vol. 3516, pp. 476–483). Berlin Heidelberg: Springer-Verlag.

Bação, F., Lobo, V., & Painho, M. (2008). *Applications of different self-organizing map variants to geographical information science problems. Self-organising maps: Applications in geographic information science*. P. Agarwal and A. Skupin, pp. 21–44.

Behnisch, M., & Ultsch, A. (2008). Urban data mining using emergent SOM. In H. Burkhardt, L. Schmidt-Thieme, R. Decker, & C. Preisach (Eds.), *Data analysis, machine learning and applications* (pp. 311–318). Berlin, Heidelberg: Springer.

Céréghino, R., Santoul, F., Compin, A., & Mastrorillo, S. (2005). Using self-organizing maps to investigate spatial patterns of non-native species. *Biological Conservation, 125*(4), 459–465.

Cockings, S., & Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine, 60*(12), 2729–2742.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, VII*(II), 179–188.

Fleming, K. L., Heermann, D. F., & Westfall, D. G. (2004). Evaluating soil color with farmer input and apparent soil electrical conductivity for management zone delineation. *Agronomy Journal, 96*(6), 1581–1587.

Goodchild, M. F. (2008). Geographic information science: The grand challenges. In J. P. Wilson & A. S. Fotheringham (Eds.), *The handbook of geographic information science* (pp. 596–608). Malden, MA: Blackwell.

Guha, S., Rastogi, R., Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on management of data*. Seattle, Washington, United States, ACM.

Guo, D., & Gahegan, M. (2006). Spatial ordering and encoding for geographic data mining and visualization. *Journal of Intelligent Information Systems, 27*(3), 243–266.

Guo, D., Peuquet, D. J., & Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica, 7*(3), 229–253.

Han, J. (2005). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers Inc..

Han, J., Cai, Y., & Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering, 5*(1), 29–40.

Han, J., Kamber, M., & Tung, A. K. H. (2001). Spatial clustering methods in data mining: A survey. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 188–217). London: Taylor and Francis.

Han, J., Koperski, K., & Stefanovic, N. (1997). GeoMiner: A system prototype for spatial data mining. *SIGMOD Record, 26*(2), 553–556.

Hu, T., & Sung, S. (2005). Clustering spatial data with a hybrid EM approach. *Pattern Analysis & Applications, 8*(1), 139–148.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer, 1*(2), 69–91.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley & Sons.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *RecMap: Rectangular Map Approximations, 43*(1), 59–69.

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer.

Koua, E., & Kraak, M.-J. (2004). Geovisualization to support the exploration of large health and demographic survey data. *International Journal of Health Geographics, 3*(1), 12.

Lobo, V., Bação, F., Henriques, R. (2009). GeoSOM repository: SquareVille dataset, 2009. <http://www.isegi.unl.pt/labnt/geosom/georepository/> Retrieved 17.07.09.

Maceachren, A. M., Wachowicz, M., Edsall, R., Haug, D., & Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science, 13*(4), 311–334.

Mather, P. M., Tso, B., & Koch, M. (1998). An evaluation of LAndsat TM spectral data and SAR-derived textural information for lithological discrimination in the Red Sea Hills, Sudan. *International Journal of Remote Sensing, 19*(4), 587–604.

Merenyi, E., Jain, A., & Villmann, T. (2007). Explicit magnification control of self-organizing maps for "Forbidden" data. *IEEE Transactions on Neural Networks, 8*(3), 786–797.

Miller, H., & Han, J. (2001). *Geographic data mining and knowledge discovery*. London, UK: Taylor and Francis.

Mu, L., & Wang, F. (2008). A scale-space clustering method: Mitigating the effect of scale in the analysis of zone-based data. *Annals of the Association of American Geographers, 98*(1), 85–101.

Ng, R. T., & Jiawei, H. (2002). CLARANS: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions, 14*(5), 1003–1016.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, 2*(4), 459–472.

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, England: GeoBooks – CATMOG 38.

Openshaw, S., & Wymer, C. (1995). *Classifying and regionalizing census data. Census users handbook*. Cambrige, UK: GeoInformation International, S. Openshaw, pp. 239–268.

O'Sullivan, D., & Unwin, D. J. (2002). *Geographic information analysis*. John Wiley and Sons.

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery, 2*(2), 169–194.

Sester, M. (2005). Optimization approaches for generalization and data abstraction. *International Journal of Geographical Information Science, 19*(8), 871–897.

Sheikholeslami, G., Chatterjee, S., Zhang, A., (1998). WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24rd international conference on very large data bases*. Morgan: Kaufmann Publishers Inc.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. W.H. Freeman.

Spielman, S. E., & Thill, J.-C. (2008). Social area analysis, data mining, and GIS. Computers. *Environment and Urban Systems, 32*(2), 110–122.

Sui, D. Z. (2004). Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers, 94*(2), 269–277.

Takatsuka, M., & Gahegan, M. (2002). GeoVISTA studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers & Geosciences, 28*(10), 1131–1144.

Thill, J.-C., W.A. Kretzschmar Jr., I. Casas, X. Yao, 2008. Detecting geographic associations in English dialect features in North America within a visual data mining environment integrating self-organizing maps. In P. Agarwal, A. Skupin (Eds.), *Self-organising maps: applications in geographic information science*. pp. 87–106.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography, 46*, 234–240.

Tobler, W. (2004). On the first law of geography: A reply. *Annals of the Association of American Geographers, 94*(2), 304–310.

Ultsch, A. (2005). Clustering with SOM: U*C. WSOM 2005, Paris.

Ultsch, A., Siemon, H.P. (1990). Kohoneńs self-organizing neural networks for exploratory data analysis. In *Proceedings of the international neural network conference, Paris, Kluwer*.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. (1999). Self-organizing map in Matlab: The SOM toolbox. In *Proceedings of the Matlab DSP Conference, Espoo, Finland, Comsol Oy*.

Wan, W., Fraser, D. (1993). M2dSOMAP: clustering and classification of remotely sensed imagery by combining multiple Kohonen self-organizing maps and associative memory. In *Proceedings of 1993 International Joint Conference on Neural Networks, IJCNN '93 Nagoya*. Japan: Supermolecular Science Division Electrotechni.

Yan, J., Thill, J.-C., 2008. Visual exploration of spatial interaction data with self-organizing maps. In A.S. Pragya Agarwal, (Ed.), *Self-organising maps*. pp. 67–85.