Spatial Clustering with SOM and GeoSOM

Case study of Lisbon's Metropolitan Area

Roberto Henriques, Fernando Bacao
Instituto Superior de Estatística e
Gestão de Informação
ISEGIUNL
Lisbon, Portugal
e-mail:{roberto,bacao}@isegi.unl.pt

Victor Lobo
Portuguese Naval Academy
Lisbon, Portugal
e-mail: vlobo@isegi.unl.pt

Abstract—**Clustering constitutes one of the most popular and important tasks in data analysis. The size and dimensionality of the existing geospatial databases stress the need for efficient and robust spatial clustering algorithms. In this paper we present the GeoSOM suite as a spatial clustering tool. GeoSOM suite implements the GeoSOM algorithm, which allows to perform clustering using both spatial and aspatial components from geospatial datasets. Finally, a case study is presented, based on census data from Lisbon, exploring the GeoSOM suite features and exemplifying its use in the context of exploratory data analysis.**

Spatial clustering; Exploratory data analysis; GeoSOM;

I. INTRODUCTION

The size and dimensionality of modern geospatial databases poses new challenges to translate their content into information and knowledge.

Data mining tools, or more specifically spatial data mining tools, constitute one of the most promising set of tools that are available to face this flood of data. Spatial data mining can be defined as the discovery of interesting relationships, spatial patterns and characteristics that may exist implicitly in spatial databases.

One data mining technique used to tackle the problems posed by the size and dimensionality of the "new" spatial databases is spatial clustering.

Clustering is the partition of data into groups of similar objects [1]. Spatial clustering is a type of clustering in which the objects to partition are spatially contextualized. This fact adds new complexity to an already difficult task, specially due to the size and dimensionality of datasets. This new complexity results from the new perspective that spatial context gives to the clusters. Thus, when performing spatial clustering we try to achieve a trade-off between spatial and aspatial components. In order words, what we are trying to achieve are groups of spatial objects similar in aspatial attributes and, as close as possible in geographic space.

This trade-off between the two subspaces (aspatial and spatial) can however be difficult to achieve, since usually geospace is characterized by two or three dimensions while aspatial subspace can range from few dimensions to a huge number of dimensions. Yet, these two subspaces are, usually, highly correlated. This is shown by spatial dependency and formulated in the 1st law of geography [2], which states that "*everything is related to everything else, but near things are more related than distant things*".

Thus, it would be expected that by clustering objects only using their aspatial dimensions, the clusters should present well-defined spatial regions.

However, most of the times, clusters produced from spatial data is spatially apart. Factors such as the aggregation and the scale of data, the spatial heterogeneity and the fact clustering only uses a subset of variables for each spatial object are possible causes to this problem.

GeoSOM, which is an extension of self-organizing maps (SOM) proposed in [3] is specially suited to spatial data mining. GeoSOM implements Tobler's First Law [4], looking for local clusters on geographic space instead of the global clusters produced by standard SOMs.

This paper extends and consolidates [3] in two major ways.

First a tool called GeoSOM suite includes GeoSOM algorithm and SOM, providing a friendly and ready to use tool. GeoSOM suite, abilities the user to interact and combine multiple solutions gathering knowledge about data and clusters produced.

Second, GeoSOM suite is assessed with Lisbon's census dataset, providing an efficient exploratory spatial data analysis (ESDA) tool. The paper is organized as follows: section 2 reviews GeoSOM algorithm. Section 3 presents GeoSOM suite. Section 4 presents some experiments with spatial data, as well as the experimental results. Section 5 concludes the paper and discusses future work.

II. SOM AND GEOSOM

Teuvo Kohonen proposed Self-organizing maps (SOM) in the beginning of the 1980s [5]. The SOM is usually used for mapping high-dimensional data into one, two, or three-dimensional feature maps, which are grids of units or neurons. The grid forms what is known as the output space, as opposed to the input space that is the original space of the data patterns. The 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 SOM implementations use a rectangular grid of units. To provide even distances between neighbour units in the output space, hexagonal grids are sometimes used [6]. 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.

The basic idea during training is that each datum is compared to all units. The most similar one, known as BMU (Best Matching Unit) is said to map that datum. The BMU and its neighbours in the grid are then updated so as to be closer to that particular datum. After a sufficiently large number of iterations, the units will be a representation of the dataset in the sense, that they will follow the same density distribution.

When clustering using SOMs we can follow two major methodologies. In the first method, each unit is a cluster centroid, so the number of required clusters should define the size of the SOM. Because this is similar to k-means clustering, this is usually referred as “k-means” SOM.

The second method uses a much larger SOM with much more units than the required number of clusters. The SOM units are in fact mapping the input space into a 2-dimensional space. We may visualize the density distribution of the units in the input space (and by proxy of the original data) mapped onto the 2-dimensional output space. The best way to analyse the density of the input space in the output space is through the use of U-matrices [7], often referred to as simply “u-mats”. This approach is usually known as “emergent” SOMs [8].

The GeoSOM constitutes a variation of the original SOM algorithm, and it was devised to explicitly 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 neighbourhood where it is admissible to search for the BMU, and the second phase performs the final search using the other multidimensional components. The search neighbourhood is controlled by a parameter k, defined in the output space. Using k=0 the algorithm will necessarily select as BMU the closest geographical unit. On the other hand, setting k equal to the size of the map (the SOM), space is ignored, and the algorithm will perform just like a regular SOM. The results obtained with k=0, will be similar to the training of a standard SOM with only the geographical locations (x,y coordinates), and then using each unit as a low pass filter for 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. For a detailed explanation of the workings of the GeoSOM the reader is referred to [9].

III. GEOSOM SUITE

The GeoSOM suite implementation is based on Matlab® and SOM toolbox [10]. A stand-alone graphic user interface (GUI) was built, allowing non-programmer users to evaluate and use the SOM and the GeoSOM algorithms. The GeoSOM suite is freely available at [11]. Fig. 1 shows the general GeoSOM suite architecture, which consists of (1) a GUI, (2) MATLAB runtime components including SOM toolbox and GeoSOM routines, (3) access to spatial and non-spatial data and (4) production of outputs. These outputs consist of geographic maps, U-matrices, component planes, hit maps, and parallel coordinate plots. The GeoSOM suite also allows the use of multiple analysis tools (*e.g.*, the possibility of using several SOM and GeoSOM) on the same dataset.

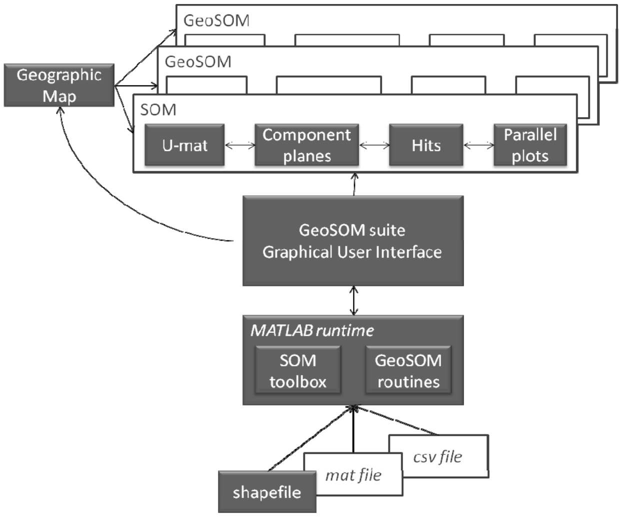
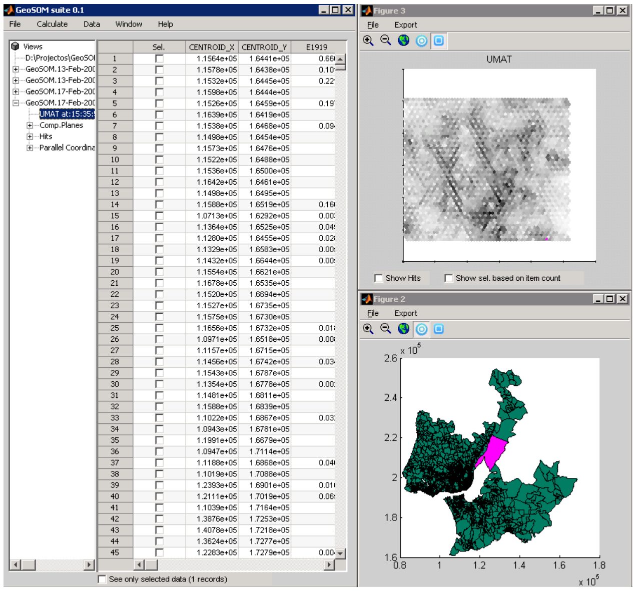


Figure 1. GeoSOM suite architecture.

Fig. 2 presents a screen-shot of the GeoSOM suite. The main interface contains a table of attributes and a tree view pointing to all the views created. Also presented in the figure are two examples of views: a geographic data, and a u-matrix views.

GeoSOM suite’s main functionalities are:

- present and select by point and click geographical data;
- train SOMs and GeoSOMs according to the nature of the data used;
- generate several graphs and tables known as views (U-matrices, Component planes, etc) and;
- link the views dynamically (*i.e.* when one view is clicked data is selected on all views);



	Sel.	CENTROID_X	CENTROID_Y	E1919
1	□	1.1564e+05	1.6441e+05	0.66
2	□	1.1578e+05	1.6438e+05	0.10
3	□	1.1532e+05	1.6445e+05	0.22
4	□	1.1598e+05	1.6444e+05	
5	□	1.1526e+05	1.6459e+05	0.19
6	□	1.1639e+05	1.6419e+05	
7	□	1.1538e+05	1.6468e+05	0.09
8	□	1.1498e+05	1.6454e+05	
9	□	1.1573e+05	1.6476e+05	
10	□	1.1522e+05	1.6488e+05	
11	□	1.1536e+05	1.6500e+05	
12	□	1.1642e+05	1.6461e+05	
13	□	1.1498e+05	1.6495e+05	
14	□	1.1588e+05	1.6519e+05	0.16
15	□	1.0713e+05	1.6292e+05	0.00
16	□	1.1364e+05	1.6525e+05	0.04
17	□	1.1280e+05	1.6455e+05	0.02
18	□	1.1329e+05	1.6583e+05	0.00
19	□	1.1432e+05	1.6644e+05	0.00
20	□	1.1554e+05	1.6621e+05	
21	□	1.1678e+05	1.6535e+05	
22	□	1.1520e+05	1.6694e+05	
23	□	1.1527e+05	1.6735e+05	
24	□	1.1575e+05	1.6730e+05	
25	□	1.1656e+05	1.6732e+05	0.01
26	□	1.0971e+05	1.6518e+05	0.00
27	□	1.1157e+05	1.6715e+05	
28	□	1.1456e+05	1.6742e+05	0.03
29	□	1.1543e+05	1.6787e+05	
30	□	1.1354e+05	1.6778e+05	0.00
31	□	1.1481e+05	1.6811e+05	
32	□	1.1588e+05	1.6839e+05	
33	□	1.1022e+05	1.6867e+05	0.03
34	□	1.0943e+05	1.6781e+05	
35	□	1.1991e+05	1.6679e+05	
36	□	1.0947e+05	1.7114e+05	
37	□	1.1188e+05	1.6868e+05	0.04
38	□	1.1019e+05	1.7088e+05	
39	□	1.2393e+05	1.6901e+05	0.01
40	□	1.2111e+05	1.7019e+05	0.06
41	□	1.1039e+05	1.7164e+05	
42	□	1.3876e+05	1.7253e+05	
43	□	1.4078e+05	1.7218e+05	
44	□	1.3624e+05	1.7277e+05	
45	□	1.2283e+05	1.7279e+05	0.00

□ See only selected data (1 records)

Figure 2. GeoSOM suite interface: from the left to the right, GeoSOM suite main window with non-spatial attributes, U-mat using census data and geographic map from Lisbon Metropolitan Area.

A. *Views in the GeoSOM suite*

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

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

B. *Dynamically linked windows*

This feature allows the selection of elements on a view providing the user with the matching selection on any other open view. Thus, if the user selects some features in a geographic map, the corresponding selection will be presented on the u-mat, the component planes and the parallel coordinate plots. The use of dynamically linked windows enables strong interaction with the data, allowing users to analyze data from different perspectives.

C. *Clustering in GeoSOM suite*

GeoSOM suite allows the users to define clusters on plots of U-matrices. Thus the user can draw the clusters according to his interpretation of the U-matrix, and label input data points accordingly. After defining the clusters on top of U-matrices, it is possible to visualize the clusters on all open views.

D. *Combining multiple clustering solutions with GeoSOM suite*

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

This feature also allows a sensibility analysis of SOM and GeoSOM by comparing results using different input parameters. It also allows the comparison of SOM and GeoSOM algorithms in clustering. Using this possibility will give the user an insight on the geographic nature of data.

Finally, the user might choose to train several SOM's using different subsets of variables. This can be thought of as building different thematic classifications (*e.g.* for the census dataset it can be building characteristics, family characteristics, unemployment characteristics, or other) which can, in the end, be evaluated together.

IV. Lisbon's Metropolitan Area Spatial Clustering

In this section, we evaluate GeoSOM suite through a cluster analysis using Lisbon's Metropolitan Area (LMA) 2001 census data. Variables used in this analysis are those related to the age of the buildings, the population age structure and the education level. Table I presents the variables used in the cluster analysis.

TABLE I. Variables used in the cluster analysis

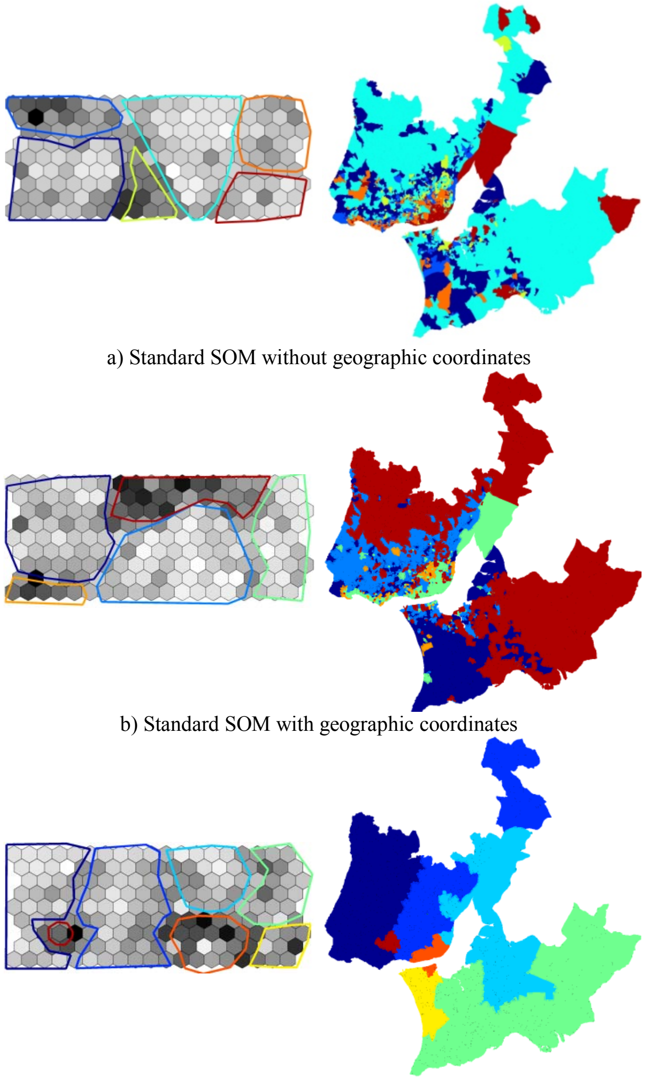
Category	**Variables**	
	Variable	***Description***
Age of the buildings	E1945	% of buildings built before 1945
	E1970	% of buildings built between 1946 and 1970
	E1980	% of buildings built between 1971 and 1980
	E1990	% of buildings built between 1981 and 1990
	E2001	% of buildings built between 1991 and 2001
Residents' age	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' education level	Ens0	% of resident with no formal education
	EnsBas1	% of resident with 4 years of education
	EnsBas23	% of resident with 6 to 8 years of education
	EnsSec	% of resident with 12 years of education
	EnsSup	% of resident with higher education

After removing outliers from the dataset, we start this analysis by applying three different approaches of SOM to the dataset. Thus, we applied:

- a standard SOM using the variables presented in Table I;

- a standard SOM using the variables presented in Table I and the *x* and *y* coordinates of each ED;
- the GeoSOM method;

In all three cases a 10x5 network was used, and the resulting U-matrices are shown in Fig. 3. On top of each U-matrix clusters are drawn (based on visual inspection) and for each cluster a colour code is defined. Clusters are then presented in the geographical space using the same colour code.



a) Standard SOM without geographic coordinates

b) Standard SOM with geographic coordinates

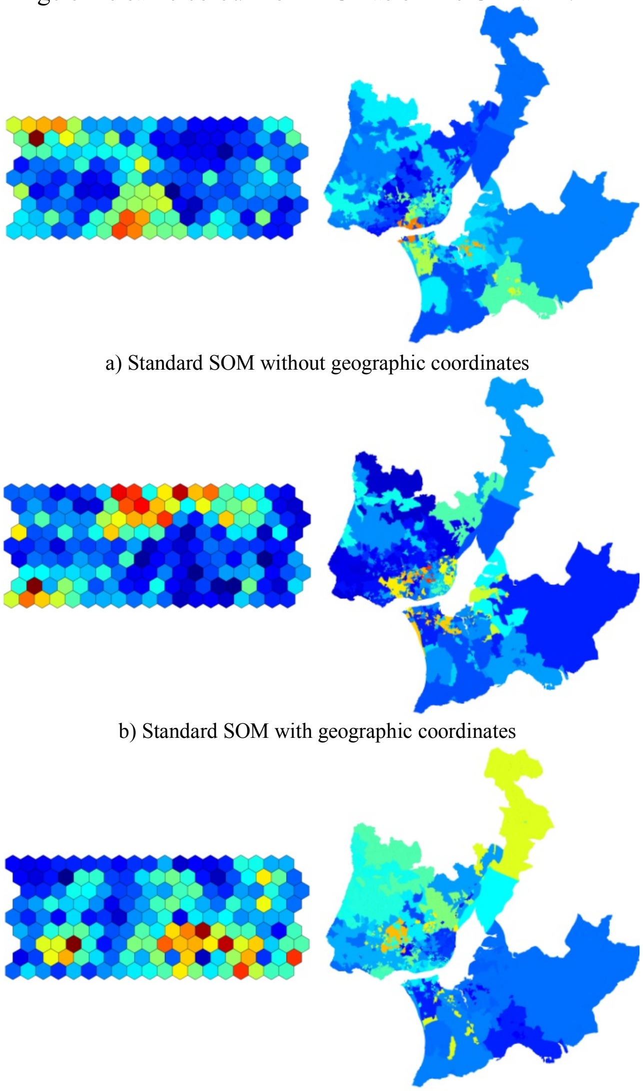
c) GeoSOM

Figure 3. U-matrices and geographical map of Lisbon Metropolitan Area showing the clusters achieved using: a) standard SOM with only aspatial components; b) standard SOM using spatial and asaptial components and; c) GeoSOM approach.

Comparing the clusters produced in the three cases it is possible to conclude that:

- The use of geographical coordinates in the standard SOM, in a high-dimensional problem, does not alter significantly the clusters
- GeoSOM creates continuous geographic clusters, although this restriction makes clusters more heterogeneous.

Another example is shown in Fig. 4 where we present for the same three cases, the U-matrices colour coded between blue and red (blue means short distances and red means longer distances between units). For each U-matrix, Lisbon's Metropolitan Area map is also presented were each ED gets the same colour its BMU has on the U-matrix.



a) Standard SOM without geographic coordinates

b) Standard SOM with geographic coordinates

c) GeoSOM

Figure 4. Representation of the unit's distances on the geographical map using the same colors used in the U-matrices for the following cases: a) standard SOM with only aspatial components; b) standard SOM using spatial and asaptial components and; c) GeoSOM approach.

The graphics presented in Fig. 4 allow the identification of the geographical regions where the units' density is lower and therefore EDs that are outliers.

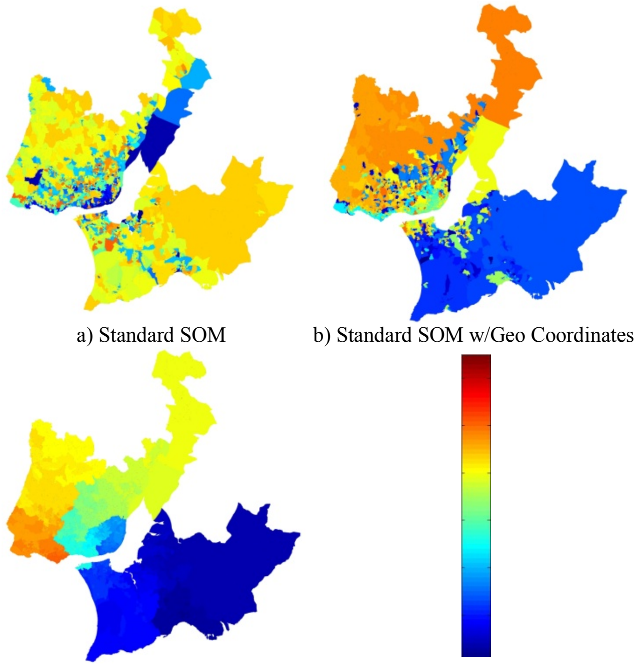
Finally, to understand the relation of SOM, GeoSOM and the inclusion of the space in cluster analysis, we decided to use one-dimensional neural networks in the three assessments. Therefore, two one-dimensional SOMs and one GeoSOM are created with 50 x 1 units.

As before, the first SOM uses only the aspatial data, while the second SOM and the GeoSOM combine aspatial and spatial data.

Because we used a grid of units of one dimension, it is easy to define a similarity between these units, *i.e.* the position each unit in the grid represents its similarity with the other units (unit in position 1 is more similar to the unit in position 2 than to the unit in position 3).

Using this metric, from each U-matrix produced, we created for each ED a degree of similarity (by identifying the position in the U-matrix of its best matching unit).

Fig. 4 shows the maps with this EDs' similarity for the three cases used. The colours used in the maps represent how similar EDs are with each other.



a) Standard SOM

b) Standard SOM w/Geo Coordinates

c) GeoSOM

d) Colormap

Figure 5. Similarity EDs maps using a color ramp. Maps created a one dimensional: a) standard SOM with only aspatial components; b) standard SOM using spatial and asaptial components and; c) GeoSOM approach. d) presents the color map used in which similar EDs get similar colors.

In these examples, the one-dimensional SOMs used are imposing a linear ordering to the EDs. This projection from a high-dimensional space to one dimension has to generalise data, and important information can be overlooked. However, this arrangement allows a simple comparison of the EDs which can easily be spatially represented.

V. Conclusions

In this paper we presented the GeoSOM suite, a new and efficient tool for exploratory spatial data analysis (ESDA) and clustering. This tool uses two methods, the SOM [6] and the GeoSOM [3]. The SOM is a well-known algorithm that has proved to be of interest in spatial clustering. The GeoSOM, by explicitly taking into account spatial autocorrelation, has the capability of detecting both spatially homogeneous and spatially heterogeneous areas.

As shown in this paper, the inclusion of geographical coordinates in the SOM, combined with the use of multiple aspatial variables does not produce clusters very different from those obtained without geographical coordinates. This fact results from the dilution of the geographical coordinates' weight in the set of variables.

GeoSOM by assuming the special nature of geospatial data allows the creation of spatially continuous clusters, which can be of great use for urban planning or environmental management.

Acknowledgment

This work was supported in part by the Fundação para a Ciência e a Tecnologia under PhD grant SFRH/BD/30360/2006.

References

[1] Han, J., *Data Mining: Concepts and Techniques*. 2005: Morgan Kaufmann Publishers Inc.

[2] Tobler, W.R., *A Computer Movie Simulating Urban Growth in the Detroit Region*. Economic Geography, 1970. **46**: p. 234-240.

[3] Bação, F., V. Lobo, and M. Painho, *Applications of Different Self-Organizing Map Variants to Geographical Information Science Problems*, in *Self-Organising Maps: Applications in Geographic Information Science*, P. Agarwal and A. Skupin, Editors. 2008. p. 21-44.

[4] Tobler, W., *A continuous transformation useful for districting*. Annals, New York Academy of Sciences, 1973. **219**: p. 215-220.

[5] Kohonen, T., *Self-organizing formation of topologically correct feature maps*. RecMap: rectangular map approximations, 1982. **43 (1)**: p. 59-69.

[6] Kohonen, T., *Self-Organizing Maps*. 3rd edition ed. 2001, Berlin: Springer.

[7] Ultsch, A. and H.P. Siemon. *Kohonen´s self-organizing neural networks for exploratory data analysis*. in *Proceedings of the International Neural Network Conference*. 1990. Paris: Kluwer.

[8] Ultsch, A. and F. Moerchen, *ESOM-Maps: tools for clustering,visualization, and classification with Emergent SOM*. 2005, Dept. of Mathematics and Computer Science, University of Marburg, Germany.

[9] Bação, F., V. Lobo, and M. Painho, *The self-organizing map, the Geo-SOM and relevant variants for geosciences*, in *Computers and Geosciences*. 2005, Elsevier. p. 155-163.

[10] Vesanto, J., *et al. Self-organizing map in Matlab: the SOM Toolbox*. in *Proceedings of the Matlab DSP Conference*. 1999. Espoo, Finland: Comsol Oy.

[11] Lobo, V., F. Bação, and R. Henriques. *GeoSOM suite*. 2009 15-11-2009]; Available from: www.isegi.unl.pt/labnt/geosom.