Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions

Fernando Bação[1], Victor Lobo[1,2], and Marco Painho[1]

[1]Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa, Campus de Campolide 1070-312 Lisboa, Portugal
[2]Academia Naval, Alfeite, 2810-001 Almada, Portugal
{bacao, vlobo, painho}@isegi.unl.pt

Abstract. Regionalization and uniform/homogeneous region building constitutes one of the most longstanding concerns of geographers. In this paper we explore the Geo-Self-Organizing Map (Geo-SOM) as a tool to develop homogeneous regions and perform geographic pattern detection. The Geo-SOM presents several advantages over other available methods. The possibility of "what-if" analysis, coupled with powerful visualization tools and the accommodation of spatial constraints, constitute some of the most relevant features of the Geo-SOM. In this paper we show the opportunities made available by this tool and explore different features which allow rich exploratory spatial data analysis.

1 Introduction

Small area census data constitute a major data source in Geographic Information Science (GISc). The advent of digital census boundaries and the consequent assembly of Geographic Information Systems (GISs) and census data made available huge databases of small administrative geographical features characterized by high dimensional vectors of socio-demo-economic information.

This fact created opportunities for developing an improved understanding of a number of socioeconomic phenomena that are at the heart of GISc. Nevertheless, it also shaped new challenges and raised unexpected difficulties for the analysis of multivariate spatially referenced data. Today, the availability of methods able to perform sensible reduction on huge amounts of high dimensional data, is a central issue in science generically and GISc is no exception.

The urgency of transforming into information the massive digital databases that result from decennial census operations has motivated work in a number of research areas. Geodemographic typologies (Openshaw and Wymer 1995; Openshaw *et al.* 1995; Birkin and Clarke 1998; Feng and Flowerdew 1998), identification of deprived areas (Fahmy *et al.* 2002), and social services provision (Birkin *et al.* 1999) constitute a few examples of subjects where private and public organizations can benefit from techniques that can isolate important trends and patterns from such large datasets, although many more can benefit (i.e., research on suburbanization, residential segregation, immigrant settlement patterns, rural depopulation, etc.).

The challenge is to take advantage of new computationally intensive tools made available in research areas like knowledge discovery (Fayyad *et al*. 1996; Han and Kamber 2000; Miller and Han 2001), which are particularly adapted to process large quantities of data. Nevertheless, it is fundamental to introduce some kind of spatial reasoning into these methods (Openshaw *et al*. 1995), as spatial data comprises special characteristics (Anselin 1990) and spatial analysis should not ignore some important paradigms (e.g., the 1st Law of Geography (Tobler 1970)).

In this paper we put forward a new tool, based on the Self-Organizing Map (Kohonen 1982; Kohonen 2001), for the development of homogeneous regions and spatial pattern detection. Here the zone design problem is approached as a tool for discovery and spatial data exploration and reduction. The idea is to provide a tool that enables the user to interact and explore spatial data, emphasizing the fuzzy nature of most classifications, allowing for "what-if" analysis and providing rich visualization context for exploratory analysis.

The improvements pursued here lie on the capability of introducing more geographical knowledge within the classification process. The option in this work is to emphasize the importance of the geo prefix in small area analysis. In this paper we develop the Geo-SOM which consists of variants of the original Self-Organizing Map (SOM), and is particularly adapted to process and deal with specific features of spatial data, such as geographic location.

2 Discovery and Exploration through Zone Design

In the context of GISc the zone design problem constitutes a paradox as it can be seen both as a problem and an opportunity. On one hand the detrimental effects of arbitrary zoning in administrative and statistical area reporting have plagued geographic-based research (Openshaw 1984; Fotheringham and Wong 1991; Amrhein 1995). On the other hand, computational and algorithmic evolution created the opportunity to transform the problem into a valuable exploratory tool in spatial analysis (Openshaw 1984; Wise *et al*. 1997; Guo *et al*. 2003).

Haggett *et al*. (1977) propose three different types of regions which they classified as uniform regions; nodal regions; and planning or programming regions. Usually, planning or programming regions are developed with a specific and well-defined purpose in mind. They tend to result from explicit needs of institutions that have to manage territorial dispersed activities, providing a management perspective on zone design. Examples of this are the problem of electoral districting (Horn 1995; George *et al*. 1997; Mehrotra *et al*. 1998; Macmillan and Pierce 1994), sales territories (Leischmann and Paraschis 1998), police reporting areas (Sarac *et al*. 1999), and census output Areas (Martin 1997; Martin 1998). Uniform and nodal regions can be viewed as having a fairly exploratory nature, as they usually assist research and discovery processes. Although it is difficult to produce a clear-cut differentiation between zone design as a management tool and zone design as a discovery tool, some distinctions can be made.

Typically, zone design as a management tool assumes restrictive constraints on the geographic configuration of the resulting regions. Thus, the contiguity constraint is

usually present and compactness is generally considered a desirable characteristic. The rationale behind the construction of the algorithms is to produce highly efficient procedures that make use of computational resources and improved optimization techniques to eventually arriving at global optimum solutions. An optimality criterion is previously defined and needs to be encoded into the optimization procedure.

In the case of zone design as a discovery and exploratory tool the aim is more focused on evaluating different possibilities, probably using fewer and less restrictive constraints on the geographic configuration of the resulting zonal systems. Here the main objective is to, "let data speak for themselves" (Gould 1981). This way, strict geographic constraints, such as contiguity, and the need to define in advance the number of regions can be interpreted as restrictive factors that might obscure the identification of interesting patterns. As a discovery tool, regionalization can be seen as a pre-processing method, which is used to generate the basic units for subsequent analysis (Haining *et al.* 1994). Additionally it can be used to detect particular areas with specific characteristics.

Logically, the differences expressed above have a major impact on the specifications of the algorithms developed to deal with zone design. Exploratory zone design algorithms should allow for user interaction, because the objectives are somewhat fuzzy, and an adequate human/system interaction can guide an otherwise "black box" clustering process (Guo *et al.* 2003). The critical analysis of the results, which can be provided by skilled analysts, can be a valuable contribution, and algorithms should provide adequate means for result interpretation and diagnosis. Additionally, the possibility of "what-if" analysis can be a helpful addition in the sense that it allows probing based on prior knowledge and prompt evaluation between different design options. Finally, visualization capabilities of the algorithms should take advantage of GIS technology and, if possible, push forward analysis based on visualization tools.

3 Characteristics of Typologies Based on Small Area Census Data

We briefly review some of the major characteristics of small area census data typologies. These characteristics constitute the motivation for the development of the GeoSOM.

3.1 The fuzziness associated with small area typologies

According to Feng and Flowerdew (1998) there are two different types of fuzziness in typologies developed from small area census data: one related with the attribute space, the other associated with the geographical space. The first kind of fuzziness is due to the "all or nothing nature of the classification assignment" (Openshaw *et al.* 1995), which conceals the fact that enumeration districts (EDs) may be close to more than one neighbourhood type in the attribute space.

The geographical fuzziness comes from the arbitrary nature of the EDs, which serve as the elementary units on which typologies are based, the well known problem of the modifiable areal unit problem (Openshaw 1984). In fact, the nature of these units reflects the operational needs that steer data collection, and for that matter

should not be assumed to have any type of homogeneity in terms of socioeconomic characteristics (Morphet 1993). Although near things are more related than distant things (Tobler 1970), the problem lies with the fact that the scale on which this empirical regularity can be observed does not have to coincide with the scale represented by the EDs. The neighbourhood effects might be expressed at different scales in different areas of the study region and only fortuitously (probably only miraculously) will coincide with the scale denoted by the EDs.

3.2 Resolution and precision issues

Openshaw *et al.* (1995) point out problems related with the variability in the size of EDs, which influence the precision and resolution of data. This variation is not random and often reflects the duality of urban-rural areas. In urban areas, EDs tend to be more densely populated, which results in mixed characteristics but with accurate values. On the other hand, rural EDs tend to be more homogeneous, but also present more extreme results due to their sizes. Conventional classifiers give equal importance to each ED, this will result in a better representation of extreme results and a poor representation of more mixed ED, this way the least reliable results will benefit from a better representation (Openshaw *et al.* 1995).

Another relevant issue that needs to be pointed out is the reliability and statistical significance of relations found at the ED level. For instance, when linking health data and census data it is fundamental that areas have large enough populations to ensure that rates are reliable; additionally, these areas should be homogeneous with respect to relevant socioeconomic attributes (Haining *et al.* 1994).

3.3 The relevance of providing geographical context

A lot of work in this area is related to the field of geodemographics, where the main focus has been on developing highly efficient variance optimizers, and for this reason little if any attention has been given to the geographical context of the data. In fact, it is clear that the inclusion of contiguity restrictions (Openshaw and Wymer 1995) or geographic references (Lobo *et al.* 2004) will increase the variance of the resulting clusters. Nevertheless the question is: will this variance reduction frenzy improve typologies? Probably no one can answer this question. Nevertheless, we argue that it is geographically coherent to provide a geographical framework in small area based typologies. This is especially pertinent in the light of the characterization provided above, which points to all sorts of fuzziness and uncertainty when dealing with small area census data. The geographical framework can allow the identification of regularities, the detection of unusual EDs, and the discovery of boundaries where major shifts in the phenomena under study occur. Within its geographic context, the possibility of sorting out and understanding the fuzziness in data is much higher. Clearly the assumption here is that the studied phenomena happens at a smaller scale than the scale that characterizes the EDs.

4 Self Organizing Maps (SOM)

Although the term "Self-Organizing Map" could be applied to a number of different approaches, we use it as a synonym of Kohonen's Self Organizing Map, or SOM for short. The basic idea of a SOM is to map the data patterns onto an n-dimensional grid of neurons or units. That grid forms what is known as the output space, as opposed to the input space, which is the original space where the data patterns are. 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 SOM algorithm for training a 2-dimensional map may be defined as follows:

Let

X be the set of n training patterns $x_1, x_2, .. x_n$

W be a $p \times q$ grid of units w_{ij} where i and j are their coordinates on that grid

a be the learning rate, assuming values in]0,1[, initialized to a given initial learning rate

r be the radius of the neighborhood function $h(w_{ij}, w_{mn}, r)$, initialized to a given initial radius

1 Repeat

2 For k=1 to n

3 For all $w_{ij} \in W$, calculate $d_{ij} = || x_k - w_{ij} ||$

4 Select the unit that minimizes d_{ij} as the winner w_{winner}

5 Update each unit $w_{ij} \in W$: $w_{ij} = w_{ij} + \alpha h(w_{winner}, w_{ij}, r) ||x_k - w_{ij}||$

6 Decrease the value of a and r

7 Until α reaches 0

To visualize the results of a SOM, we may use U-Matrices (Ultsch and Siemon 1990). This is a representation of a SOM in which distances, in the input space, between neighbouring neurons are represented, usually by a colour code. If distances between neighbouring neurons are small, then these neurons represent a cluster of patterns with similar characteristics. If the neurons are far apart, then they are located in a zone of the input space that has few patterns and can be seen as a separation between clusters. Thus, a visual inspection of the U-Matrices allows the user to identify different clusters of data with variable "similarity resolution."

For a thorough review the reader is referred to Kohonen (2001). SOMs have been used in many different areas, and in geographical problems they have been used to perform nonlinear mappings (Skupin 2003), clustering (Openshaw *et al.* 1995; Painho and Bação 2000), and in localization problems (Gomes *et al*. 2004; Hsieh and Tien 2004). Most of these applications focus either on the geographical coordinates or on the other features. However, to address the issues of homogeneous and geographically coherent region design, it is necessary to give special attention to geographical coordinates and at the same time use the non-geographical features for clustering. We identified two distinct ways of doing so (Figure 1). The first consists of including the geographical coordinates in the pattern vector, and giving them increasing importance (Lobo *et al*. 2004). The second is a new SOM architecture, which we named Geo-SOM, and in this paper we discuss its relation with other well known architectures.

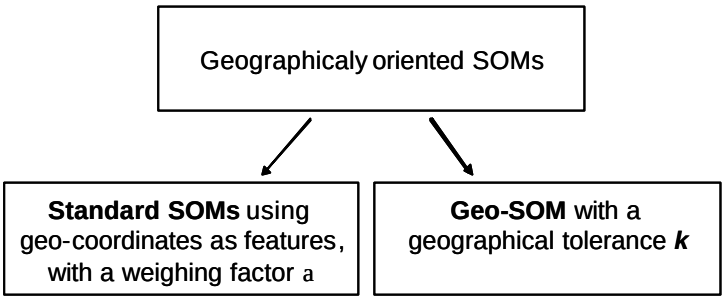

Fig. 1. Types of geographically oriented SOMs

4.1 Geographically oriented SOMs

In the SOM training algorithm, the most important step in establishing which patterns are clustered together is the one where we choose the Best Matching Unit (BMU). In the basic SOM this is done by comparing all components of the input pattern with all components of each unit (and normally calculating the distance between these two vectors). By changing the way the BMU is selected, we can give greater importance to the geographical coordinates.

One way of doing this is simply to include these coordinates in the pattern vector, and scaling them by a parameter a. If $a=0$ the geographical coordinates have no importance whatsoever, whereas if $a=8$, only these are relevant. In this latter case, the BMU will always be the unit geographically closer, and the update process will simply calculate local averages of the other parameters. By observing the way in which these local averages differ (normally with a U-Matrix) we may establish regions with the desired regularity.

The same basic approach of including the geographical coordinates in the data pattern can be used with most other SOM variants, such as Hierarchical SOMs (Ichiki *et al*. 1991; Behme *et al*. 1993) or ASSOM (Kohonen 2001).

4.2 Geo-SOM

Another way of forcing the BMU to be in the geographical vicinity of the input pattern is to explicitly divide the search for the BMU in two phases: first establish a geographical vicinity where it is admissible to search for the BMU, and then perform the final search using the other components. The vicinity where we search for the BMU can be controlled by a parameter k, defined in the output space[1]. If we choose k=0, then the BMU will necessarily be the unit geographically closer. If we allow k to grow up to the size of the map then we will ignore the geographical coordinates altogether.

When k=0, the final locations in the input space of the units will be a quasi-proportional representation of the geographical locations of the training patterns (for a discussion on the proportionality between units and training patterns see (Cottrell *et al*. 1998)), and thus the units will have local averages of the training vectors. Exactly the same final result may be obtained by training a standard SOM with only the geographical locations, and then using each unit as a low pass filter of the non-geographic features. The exact transfer function (or kernel function) of these filters depends on the training parameters of the SOM, and is not relevant for this discussion.

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.

Formally, the Geo-SOM may be described by the following algorithm:

Let

X be the set of n training patterns $\boldsymbol{x_1, x_2, ..x_n}$, each of these having a set of components $\boldsymbol{geo_i}$ and another set $\boldsymbol{ngf_i}$.

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 components $\boldsymbol{wgeo_{ij}}$ and another set $\boldsymbol{wngf_{ij}}$.

α be the learning rate, assuming values in]0,1[, initialized to a given initial learning rate

[1] The geographical tolerance k could be defined in the input space. This would lead to a fixed geographical radius where clustering would be allowed to occur. Choosing k in the output is preferable since it allows a finer resolution in areas with greater pattern density and a coarser resolution in the rest of the space.

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

f be a logical variable that is true if the units are at fixed geographical locations.

1 Repeat

2 For m=1 to n

3 For all $\boldsymbol{w}_{ij} \in W$,

4 Calculate $d_{ij} = ||\boldsymbol{wgeo}_k - \boldsymbol{wgeo}_{ij}||$

5 Select the unit that minimizes d_{ij} as the geo-winner $\boldsymbol{W}_{winnergeo}$

6 Select a set W_{winner} *of* $\boldsymbol{w}_{ij}$ such that the distance in the grid between $\boldsymbol{W}_{winnergeo}$ and $\boldsymbol{w}_{ij}$ is smaller or equal to k.

7 For all $\boldsymbol{w}_{ij} \in W_{winner}$,

8 calculate $d_{ij} = ||\boldsymbol{x}_k - \boldsymbol{w}_{ij}||$

9 Select the unit that minimizes d_{ij} as the winner $\boldsymbol{W}_{winner}$

10 If f is true, then

11 Update each unit $\boldsymbol{w}_{ij} \in W$: $\boldsymbol{wngf}_{ij} = \boldsymbol{wngf}_{ij} + \alpha\, h(\boldsymbol{wngf}_{winner}, \boldsymbol{wngf}_{ij}, r)\, ||\boldsymbol{x}_k - \boldsymbol{w}_{ij}||$

12 Else

13 Update each unit $\boldsymbol{w}_{ij} \in W$: $\boldsymbol{w}_{ij} = \boldsymbol{w}_{ij} + \alpha\, h(\boldsymbol{W}_{winner}, \boldsymbol{w}_{ij}, r)\, ||\boldsymbol{x}_k - \boldsymbol{w}_{ij}||$

14 Decrease the value of α and r

15 Until α reaches 0

4.3 Comparison between Standard SOM and Geo-SOM

In both approaches (Standard and Geo-SOM), geographical coordinates can be the only relevant feature (when a=8 or k=0), or they may be irrelevant (when a=0 or k=maximum size of map), and thus in both limits the approaches produce the same

result (Figure 2). In practical terms it is easier and more efficient to use the Geo-SOM when geographical coordinates are very important, and the standard SOM otherwise.

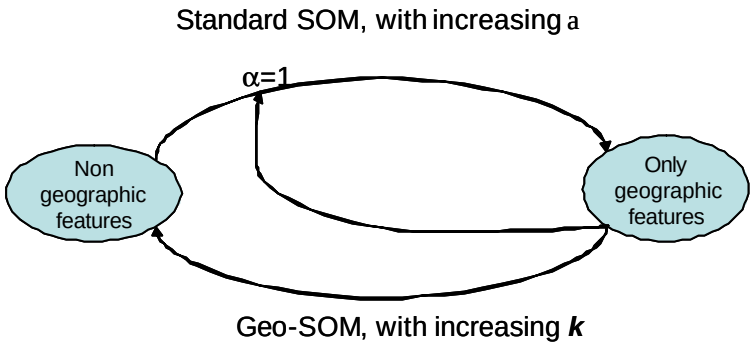

Fig. 2. Different models for standard SOM and Geo-SOM

The Geo-SOM architecture is closely related to the hypermap architecture (Kohonen 1991), the spatio-temporal SOMs (Chandrasekaran and Palaniswami 1995; Chandrasekaran and Liu 1998; Euliano and Principe 1996; Euliano and Principe 1999), and in the kangas map (Kangas 1992) adapted to spatial coordinates (Lobo *et al*. 2004). However, a thorough discussion of these relationships is outside the scope of this paper.

Finally, when using the Geo-SOM, we may include the geographical coordinates in the final search for the BMU, thus obtaining a continuum of models between the pure Geo-SOM and the pure standard SOM. For the sake of simplicity we call all these models Geo-SOM, and in our experimental tests we used the geographical coordinates in the final search for the BMU, with a=1.

5 Some Experimental Results

In order to test the Geo-SOM, we used two datasets. One of them is a very simple artificial problem with only one non-geographic feature. It was used to understand some basic properties of the Geo-SOM. The other dataset refers to EDs of the Lisbon Metropolitan Area and includes 3,968 EDs, which are characterized based on 65 variables. The Geo-SOM was implemented in Matlab® compatible with Somtoolbox (Vesanto *et al*. 2000), and is available at www.isegi.unl.pt/docentes/vlobo/projectos/programas.

5.1 The Artificial Dataset

For this example we used a set of 200 data points evenly spaced on a surface with coordinates $x \in [0,1]$, $y \in [0,2]$. Each point is associated with a single feature z, which is 0 whenever $0.5 < y < 1.5$ and 10 otherwise (Figure 3).

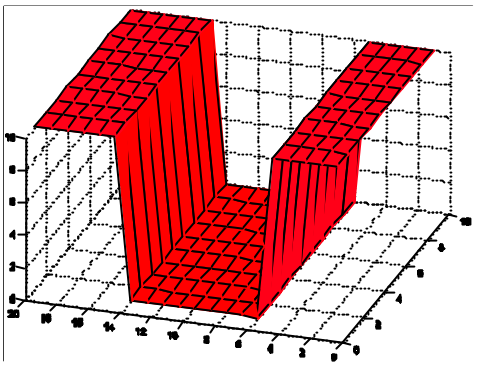

Fig. 3. The artificial dataset

If we cluster the data based on non-geographical features, then we will have two very well defined clusters: one where z=10, another where z=0 (Figure 4). If we consider only geographical coordinates, then we will have no well defined clusters, since the points are evenly spaced. If we consider all three components, we may or may not obtain well defined clusters. If no pre-processing is done, and since in this case the geographical features have a very small scale when compared to the other feature, we will basically obtain only two clusters. If we pre-process the data points to have approximately the same scale in all components, we will obtain rather fuzzy clusters. Depending on the different scalings, we may obtain 1, 2, or 3 clusters, but never clear-cut separations. A Geo-SOM with 0-tolerance will simply calculate local averages, and thus will just smooth the original dataset, and the three clusters will still appear clearly in the U-matrix. The best results are obtained using a Geo-SOM with k=2 (Figure 5). It is interesting to note that a 0-tolerance in the Geo-SOM produces blurred clusters, while relaxing this constraint will allow the clusters to define themselves better without loosing their geographic localization.

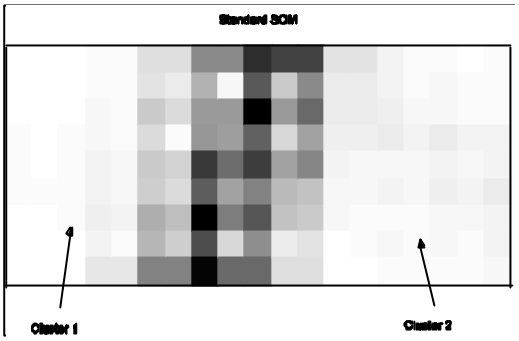

Fig. 4. U-matrix obtained for the artificial dataset with the standard SOM

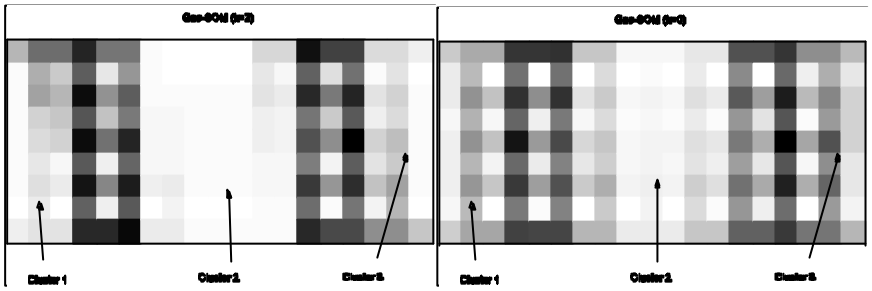

Fig. 5. U-Matrix obtained for the artificial dataset with Geo-SOMs

5.2 Lisbon Dataset

For this dataset we trained SOMs with 20x30 units, and the U-matrices obtained are presented in Figure 6.

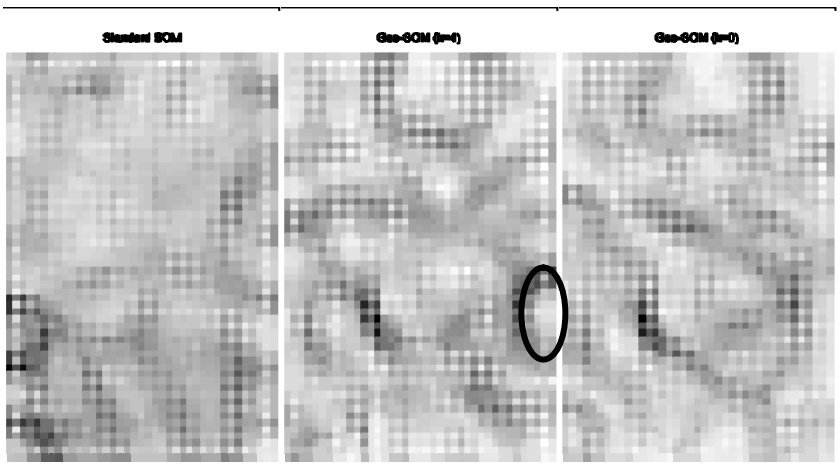

Fig. 6. U-matrices obtained for the Lisbon dataset. In the center figure (Geo-SOM with k=4) one of the clearly separated clusters, which we later use as example, is marked with an ellipse

The connection between the U-Matrices and the geographic map is a key issue in using SOM-based methods. The objective is to provide an interactive exploration environment that interconnects both spaces. At this time, this interaction is provided by ArcView, where the U-Matrix is geocoded, and linked to the geographic map. This way the selection of a unit on the U-Matrix automatically highlights the geographical areas that are classified in that specific unit. Through this mechanism, one can analyse the U-Matrix, define clusters in the data and, by selecting them in the U-Matrix, automatically get a "picture" of their geographic location.

A rough analysis of these U-matrices allows us to see that the standard SOM clusters most units in a single cluster (top half of the map), and separates, although not very clearly, a few clusters in the bottom half. An analysis of the EDs mapped onto these clusters shows, as expected, that while these EDs do in fact have similar characteristics, they are not geographically close. The Geo-SOMs lead to a very different clustering.

We would like to emphasize the exploration possibilities provided by the Geo-SOM. After training the Geo-SOM the units were georeferenced in Lisbon's map. The first aspect that is important to highlight is related with the geographic distribution of the units of the Geo-SOM. As can be seen in Figure 7 there is an important difference between the distribution of the units in the standard SOM and the Geo-SOM. In the Geo-SOM (with k=0 and 4) the units are geographically spread, mimicking the density of the centroids of the EDs. This way, the more densely populated areas will receive more classifying resources (i.e., more units), but sparsely populated areas will also receive some resources. It is quite clear from the analysis of Figure 7 that the unfolding of the Geo-SOM is ruled by the spatial distribution of the EDs. As would be expected the standard SOM distributes its units in the centre of the region, as it is there that the global variance can be minimized.

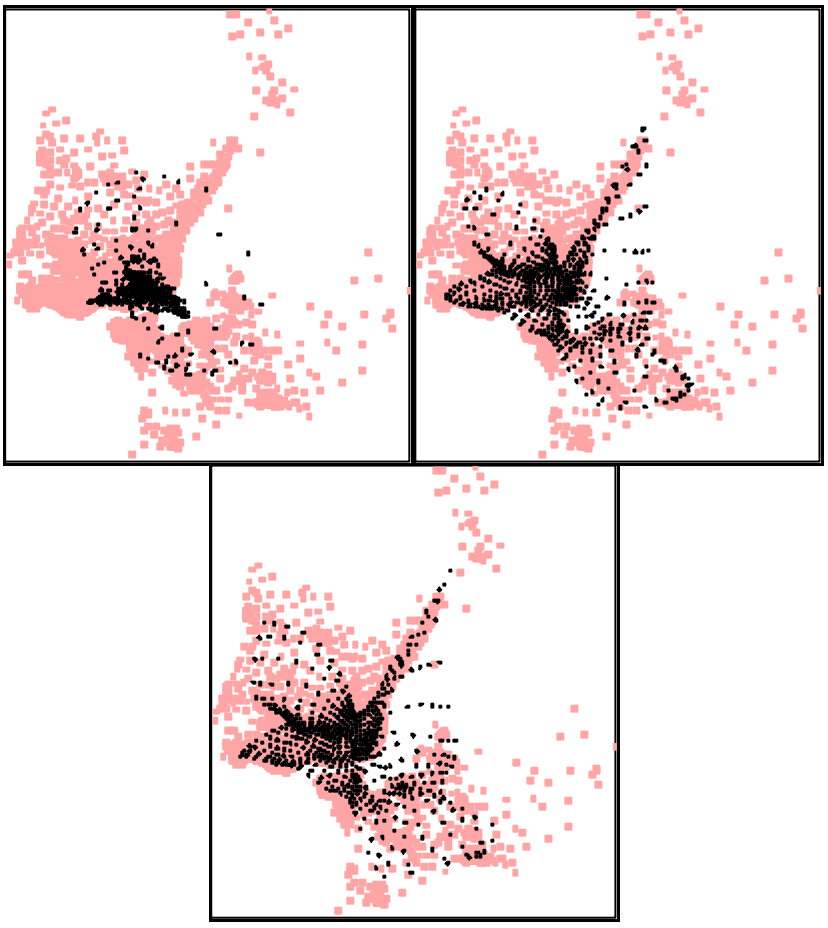

Fig. 7. Geographic distribution of Geo-SOM units for the Standard SOM (upper left), Geo-SOM with k=4 (upper right) and Geo-SOM with k=0 (units are shown as points and ED's centroids as squares)

Once the units are geocoded, the next step consists on defining to which unit is each ED centroid associated. This was done by generating Thiessen polygons based on the units and assigning each ED centroid to the nearest unit. Figure 8 shows the assigning process. In this particular case Lisbon's downtown area is depicted. As it can be seen several Geo-SOM units are placed in the river, due to the fact that it separates two high-density areas. This is natural as the Geo-SOM produces a surface for the whole of the study area.

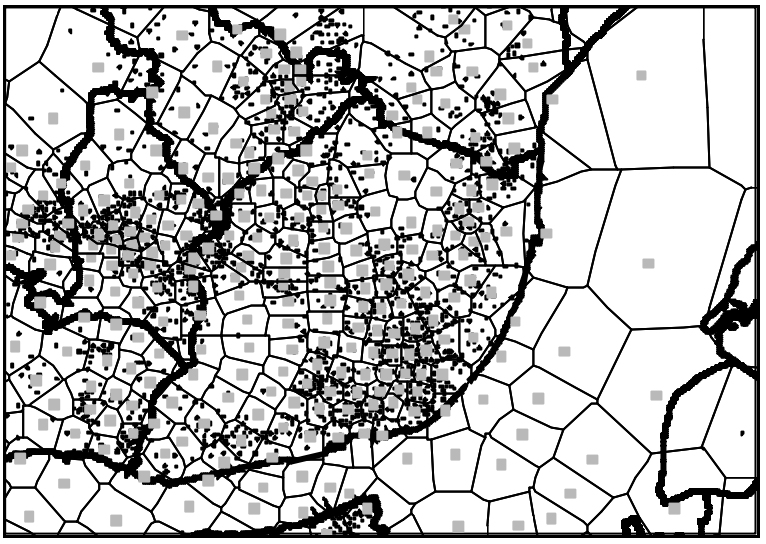

Fig. 8. Geographic distribution of Geo-SOM units and the EDs that are mapped to them in the center of Lisbon (units are shown as squares and EDs centroids as points)

Building homogeneous regions based on the Geo-SOM can be done in a number of ways. One of them is to define thresholds based on which homogeneous regions will be built. In Figure 9, three different thresholds are tested and as the threshold grows the same happens to the number of homogeneous regions.

Fig. 9. Identification of homogeneous regions (darker areas represent homogeneous regions). The leftmost figure was obtained using a low threshold (forcing very homogeneous regions), and the other two were obtained using increasing thresholds

This type of visualization is in effect a visualization of the U-Matrix in the geographic sub-space of the input space, as opposed to the traditional visualizations of that matrix in the output space (Figures 4-6). The mapping of clusters identified in the

traditional visualization of the U-Matrix to this geographical representation, although not linear, is quite simple. As an example, the cluster detected in the U-Matrix of Figure 6 can be readily identified in Lisbon's geographical map (Figure 10).

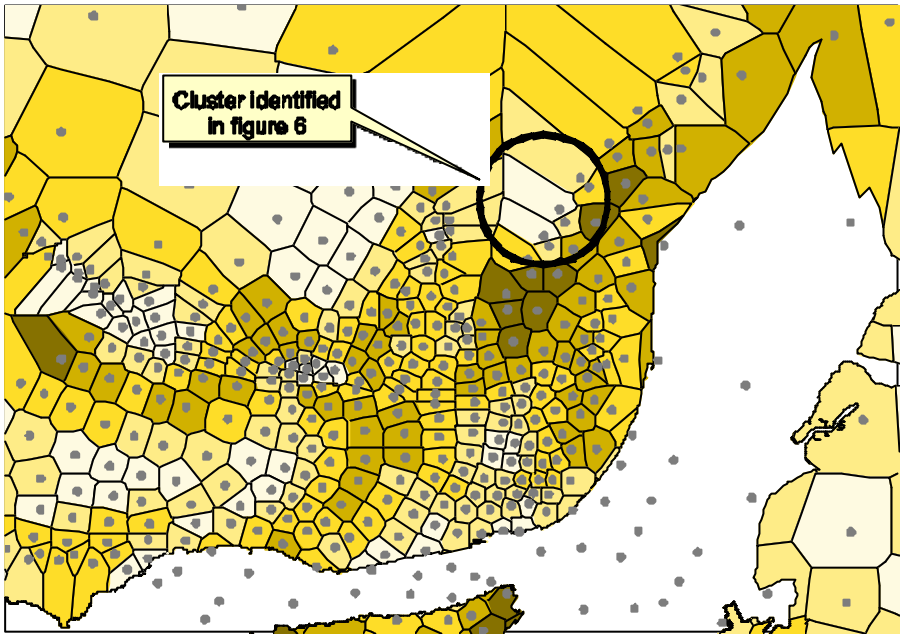

Fig. 10. Map identification of the cluster detected in figure 6 U-Matrix, here with a set of five similar units (light shade) near a well defined "boundary" of very dissimilar units (dark shade)

Similarly, if a surface is built based on the similarities of each unit and its neighbours, an elevation model can be developed (Figure 11). A progressive flooding of this surface will indicate which areas should be aggregated first. Additionally, the ridges indicate areas of change, where EDs present important dissimilarities. For instance, areas, such as the Lisbon International Airport and Monsanto (a big green area with no housing within the city limits), constitute obvious transition areas, which are well depicted by the Geo-SOM.

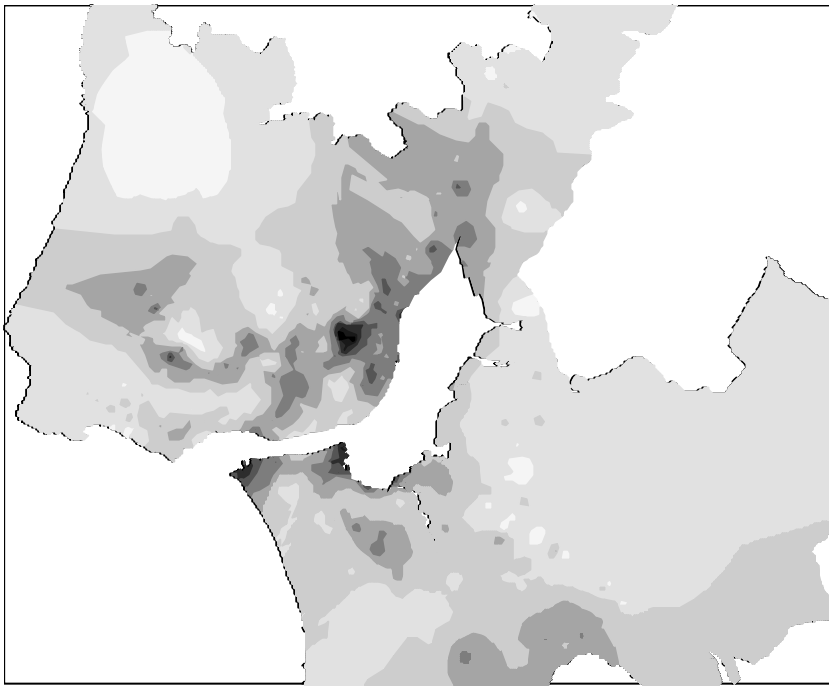

Fig. 11. Elevation map where ridges represent transition areas

6 Conclusions

An overview of different techniques for using SOMs as tools for designing homogeneous geographical regions and pattern detection was presented, and a new SOM-based architecture, named Geo-SOM, was proposed. It was shown, both in an artificial problem and in a real problem, that this new architecture can provide better insights for the region design problem. One of the advantages of using the Geo-SOM is related with its exploratory nature. Various ways of exploring the information provided by the Geo-SOM were explained. Finally, it was shown that this new architecture provides a meaningful clustering of the Lisbon Metropolitan area given a large set of census data. The idea of creating a bridge between the geographic space and the feature space is particularly appealing, as it allows processing features and subsequent visualization with geographical context. The result is a partition of space, which is primarily ruled by the density of geographic occupation and secondly by the similarity of the patterns. Further work needs to be done in two major areas. The first one is related with extending the exploratory tools provided by the Geo-SOM. The second one is rather theoretical and regards the relations between the Geo-SOM and geographical concepts like spatial autocorrelation and spatial heterogeneity.

References

Amrhein, C. G. (1995). Searching for the elusive aggregation effect: evidence from statistical simulation. Environment and Planning A 27: 105-120.

Anselin, L. (1990). What is special about spatial data? Alternative perspectives on spatial data analysis. In D. A. Griffith (ed.) Spatial Statistics, Past, Present and Future. Ann Arbor, MI, Institute of Mathematical Geography: 63-77.

Behme, H., W. D. Brandt and H. W. Strube. (1993). Speech Recognition by Hierarchical Segment Classification. In: S. Gielen and B. Kappen (eds), Proceedings of the International Conference on Artificial Neural Networks, London, UK. 416-419.

Birkin, M. and G. Clarke. (1998). "GIS, geodemographics and spatial modeling in the UK financial service industry." Journal of Housing Research 9: 87-111.

Birkin, M., G. Clarke and M. Clark. (1999). GIS for Business and Service Planning. In M. Goodchild, P. Longley, D. Maguire and D. Rhind (eds.) Geographical Information Systems. Cambridge, Geoinformation.

Chandrasekaran, V. and Z.-Q. Liu. (1998). "Topology Constraint Free Fuzzy Gated Neural Networks for Pattern Recognition." IEEE Transactions on Neural Networks 9(3): 483-502.

Chandrasekaran, V. and M. Palaniswami. (1995). "Spatio-temporal Feature Maps using Gated Neuronal Architecture." IEEE Transactions on Neural Networks 6(5): 1119-1131.

Cottrell, M., J. C. Fort and G. Pagès. (1998). "Theoretrical Aspects of the SOM Algorithm." Neurocomputing, 21: 119-138.

Euliano, N. and J. Principe. (1996). Spatio-temporal self-organizing feature maps. Proceedings of the International Conference on Neural Networks, Washington. 1900-1905.

Euliano, N. and J. Principe. (1999). A Spatio-Temporal Memory Based on SOMs with Activity Diffusion. In E. Oja and S. Kaski (eds.) Kohonen Maps. Amsterdam, Elsevier: 253-266.

Fahmy, E., D. Gordon and S. Cemlyn. (2002). Poverty and Neighbourhood Renewal in West Cornwall. Social Policy Association Annual Conference, Nottingham, UK. URL: http://www.bris.ac.uk/poverty/cornw/02SPA.doc

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. (1996). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Cambridge, MA. ISBN 0-262-56097-6.

Feng, Z. and R. Flowerdew. (1998). Fuzzy geodemographics: a contribution from fuzzy clustering methods. In S. Carver (ed.) Innovations in GIS 5. London, Taylor & Francis: 119-127.

Fotheringham, A. S. and D. W. S. Wong. (1991). "The modifiable areal unit problem in multivariate statistical analysis." Environment and Planning A 23: 1025-1044.

George, J. A., B. W. Lamar and C. Wallace. (1997). "Political District Determination Using Large-Scale Optimization." Socio-Economic Planning Sciences 31(1): 11-28.

Gomes, H., V. Lobo and A. Ribeiro. (2004). Application of Clustering Methods for Optimizing the Location of Treated Wood Remediation Units. XI Jornadas de Classificação e Análise de Dados, April 1-3, Lisbon, Portugal.

Gould, P. (1981). "Let Data Speaking for Themselves." Annals of the Association of American Geographers 71: 166-176.

Guo, D., D. Peuquet and M. Gahegan. (2003). ICEAGE: Interactive clustering and Exploration of Large and High-Dimensional Geodata. GeoInformatica, 7(3): 229-253.

Haggett, P., A. D. Cliff and A. E. Frey. (1977). Locational Analysis in Human Geography. Second Edition. London, Arnold.

Haining, R., S. M. Wise and M. Blake. (1994). "Constructing regions for small area analysis: material deprivation and colorectal cancer." Journal of Public Health Medicine 16: 429-438.

Han, J. and M. Kamber (2000). Data Mining: Concepts and Techniques. New York, Morgan Kaufmann.

Horn, M. E. T. (1995). "Solution Techniques for Large Regional Partitioning Problems." Geographical Analysis 27(3): 230-248.

Hsieh, K.-H. and F. C. Tien. (2004). "Self-organizing feature maps for solving location–allocation problems with rectilinear distances." Computers & Operations Research 31(7): 1017-1031.

Ichiki, H., M. Hagiwara and N. Nakagawa. (1991). Self-Organizing Multi-Layer Semantic Maps. Proceedings of International Conference on Neural Networks, Seattle, WA. 357-360.

Kangas, J. (1992). Temporal Knowledge in Locations of Activations in a Self-Organizing Map. Proceedings of the International Conference on Artificial Neural Networks, Brighton, England. 117-120.

Kohonen, T. (1982). Clustering, Taxonomy, and Topological Maps of Patterns. Proceedings of the 6th International Conference on Pattern Recognition, Munich. 114-128.

Kohonen, T. (1991). The Hypermap Architecture. In T. Kohonen, K. Mäkisara, O. Simula and J. Kangas (eds.) Artificial Neural Networks. 1: Helsinki, Elsevier. 1357-1360.

Kohonen, T. (2001). Self-Organizing Maps, Springer.

Leischmann, B. and J. N. Paraschis. (1998). "Solving a Large Scale Districting Problem: A Case Report." Computers and Operations Research 15(6): 521-533.

Lobo, V., F. Bação and M. Painho. (2004). Regionalization and homogeneous region building using the spatial kangas map. In F. Toppen and P. Prastacos (eds.) 7th AGILE Conference on Geographic Information Science, Heraklion, Greece. 301-313.

Macmillan, W. D. and T. Pierce. (1994). Optimization modelling in a GIS framework: the problem of political redistricting. In S. Fotheringham and P. Rogerson (eds.) Spatial Analysis and GIS. London, Taylor & Francis. 221-246.

Martin, D. (1997). "From enumeration districts to output areas: experiments in the automated design of a census output geography." Population Trends 88: 36-42.

Martin, D. (1998). "Optimizing census geography: the separation of collection and output geographies." International Journal of Geographical Information Science 12(7): 673-685.

Mehrotra, A., E. L. Johnson and G. L. Nemhauser. (1998). "An Optimization Based Heuristic for Political Districting." Management Science 44(8): 1100-1114.

Miller, H. and J. Han. (2001). Geographic Data Mining and Knowledge Discovery. London, UK, Taylor & Francis.

Morphet, C. (1993). "The mapping of small area census data-a consideration of the effects of enumeration district boundaries." Environment and Planning A 25(9): 1267-1277.

Openshaw, S. (1984). The Modifiable Areal Unit problem. CATMOG, Geo-abstracts. Norwich, UK.

Openshaw, S., M. Blake and C. Wymer. (1995). Using neurocomputing methods to classify Britain's residential areas. In P. Fisher (ed.) Innovations in GIS. Taylor & Francis. 2: 97-111.

Openshaw, S. and C. Wymer. (1995). Classifying and regionalizing census data. In S. Openshaw (ed.) Census Users Handbook. Cambrige, UK, GeoInformation International: 239-268.

Painho, M. and F. Bação. (2000). Using Genetic Algorithms in Clustering Problems. Proceedings of the 5th International Conference on GeoComputation. University of Greenwich, UK. (CD-ROM) ISBN 0-9533477-2-9

Sarac, A., R. Batta, J. Bhadury and C. Rump. (1999). "Reconfiguring police reporting districts in the City of Buffalo." OR Insight 12(3): 16-24.

Skupin, A. (2003). A Novel Map Projection Using an Artificial Neural Network. 21st International Cartographic Conference, Durban, South Africa. 1165-1172. (CD-ROM)

Tobler, W. (1970). "A Computer Model Simulating Urban Growth in the Detroit Region." Economic Geography 46: 234-240.

Ultsch, A. and H. P. Siemon. (1990). Kohonen´s Self-Organizing Feature Maps for Exploratory Data Analysis. Proceedings of the International Neural Network Conference, Dordrecht, Netherlands. Kluwer. 305-308.

Vesanto, J., J. Himberg, J. Alhoniemi and J. Parhankangas. (2000). SOM Toolbox for Matlab 5. Espoo, Helsinki University of Technology: 59.

Wise, S. M., R. P. Haining and J. Ma. (1997). Regionalisation tools for the exploratory spatial analysis of health data. In M. Fisher and A. Getis (eds.) Recent Developments in Sp atial Analysis - Spatial Statistics, Behavioural Modelling and Neurocomputing. Berlin, Springer. 83-100.