GEO-SOM AND ITS INTEGRATION WITH GEOGRAPHIC INFORMATION SYSTEMS

Fernando Bação[1], Victor Lobo[1,2] and Marco Painho[1]
[1]ISEGI-New University of Lisbon
[2]Portuguese Naval Academy
Lisbon, Portugal
bacao@isegi.unl.pt, vlobo@isegi.unl.pt, painho@isegi.unl.pt

Abstract – *Clustering geographically referenced data is an important issue in Geographic Information Science. Although the standard SOM can be used in many of these problems, it is useful to have a clustering tool that takes into account the special importance that geographic location has in these problems. In this paper, such a tool, named GEO-SOM, is presented. The differences between the training and mapping algorithms of the standard SOM and GEO-SOM are pointed out, and some simple examples of applications are given. Another important issue in the analysis of geo-referenced data is visualization of results, and integration with well established Geographic Information Systems (GIS). It is shown that GEO-SOM can easily be integrated in such systems, and examples of relevant visualization tools are presented. The fundamental assumption of the GEO-SOM is that some variables (in this case geographical coordinates) are more important, in the sense that they condition any subsequent clustering.*

Key words – Geography, geo-referenced data, spatial data, SOM variants.

1 Introduction

Clustering geographically referenced data, such as census data or remote sensing data, has been an important issue in Geographic Information Science (GIScience) for a long time. With the widespread use of GPS, mobile phones, and other location aware technologies, the amount of geo-referenced data has increased dramatically, and the need for new data reduction and analysis tools has became more urgent than ever.
Self-Organizing Maps (SOMs) [1] have been used in GIScience both for clustering geo-referenced data [2-4] [5] and for the spatialization of various non-geographic datasets [6-10]. The original SOM proposed by Prof. Kohonen does not take into account the particular role that geographic location has in most problems involving the clustering of geo-referenced data. In the original SOM algorithm, all variables are treated equally. When clustering geo-referenced data, spatial location is particularly important, since objects that are geographically far away should not be clustered together, even if they are similar in all other aspects. This is neatly expressed in the 1st Law of Geography [11] "everything is related to everything else, but near things are more related than distant things".

There are many ways of changing the standard SOM so as to give geographic location a relevant role, some of them are reviewed in [12]. One of them is the GEO-SOM that is presented in this paper.

2 GEO-SOM

The basic idea underlying GEO-SOM is that when mapping geo-referenced data to a SOM, only units that have geographic coordinates similar to the datum in question should be considered as candidates for "best matching units" (BMU). In practice this yields results which are constrained foremost by the geographic coordinates and only afterwards by other characteristics or attributes.

To achieve this goal, the search for the BMU is done in two phases. In the first phase, only the geographic locations of the data patterns and units are considered, and thus the "first phase BMU" is the unit that is geographically closer to the data pattern being considered. In the second phase, a variable number of units in the output space vicinity of the first phase BMU are considered as candidates to be the final BMU Figure 1. The actual number of units considered in this phase depends on the neighborhood radius t that we have called geographic tolerance. It must be noted that this geographic tolerance is defined in the output space, *i.e.* in the SOM grid. As a consequence, a given tolerance t corresponds to shorter distances in areas where the geographic density of data is higher, and larger distances where that density is lower. After finding the final BMU the map units are updated according to the standard SOM rule. The choice of t is largely subjective. Because of this different t values should be experimented and the results compared. Basically, t expresses the user's interest in producing local classifications: lower values of t will force the classification of geographic neighboring vectors in closer units.

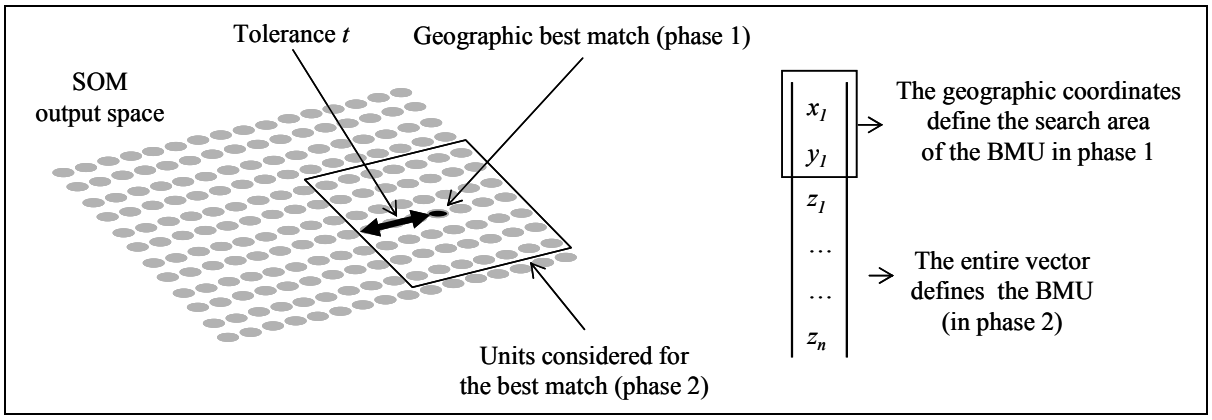

Figure 1- Geo-SOM architecture, showing the unit that is selected amongst all units in phase (1), using only geographical coordinates, and the units that are considered as candidates for BMU in phase (2).

This approach has similarities with the Hypermap approach [13], where only part of the input features are used to find the best match, and with the Kangas architecture [14] where only a small number of neighbors (in the output space) of the previous winner are searched. A combination of these two ideas leads to the spatio-temporal feature map – STFM – [15], where a "spatial gating function" is used, together with a similar temporal gating function, to select the next winner unit.

Formally, the GEO-SOM training algorithm may be described as follows:

```
Let
        X be the set of n training patterns x1,x2,..xn, each of these having a
           set of components geo_i and another set ngf_i.
        W be a p×q grid of units wij where i and j are their coordinates on
           that grid, and each of these units having a set of components wgeo_ij
           and another set 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 h(wij,wmn,r), initialized
           to a given initial radius
        t be a radius surrounding geographical BMU where the final BMU is to be
           searched
1 Repeat
2    For m=1 to n
3      For all wij∈W,
4        Calculate dij = ||wgeok - wgeoij||
5        Select the unit that minimizes dij as the geo-winner wwinnergeo
6        Select a set W_winner of wij such that the distance in the grid between
           wwinnergeo and wij is smaller or equal to t.
7        For all wij∈W_winner, calculate dij = ||xk - wij||
8          Select the unit that minimizes dij as the winner wwinner
9          Update each unit wij∈W: wij = wij +α h(wwinner,wij,r)||xk – wij||
10   Decrease the value of α and r
11 Until α reaches 0
```

To simplify notation, we indicate in parenthesis the t value used when building a given Geo-SOM, *i.e.*, Geo-SOM(0) refers to a Geo-SOM with geographic tolerance t=0, Geo-SOM(1) to t=1, *etc*.

A very simple dataset (presented in Figure 2) will help understand differences between a standard SOM and a Geo-SOM. In this case 200 data points were generated with spatial coordinates uniformly distributed ($x \in [0,1]$, $y \in [0,2]$). Without loss of generality, we associated a single non-geographic feature z, which is 0 whenever $0.5<y<1.5$ and 10 otherwise. If we use a standard SOM and U-Matrix to cluster this data, we will obtain two clusters. These are very well defined in the U-Matrix (left side of Figure 3), one corresponding to points where z=10, another to points where z=0. By pre-processing the data we may give greater importance to spatial coordinates, but since these are uniformly distributed, we will cease to have well defined clusters, as discussed in [16].

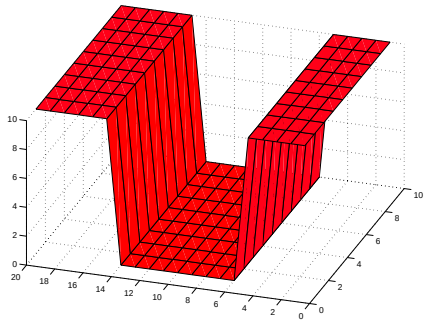

Figure 2 – Simple example of data with spatial coordinates in ([0,1],[0,2]), and a non-spatial attribute z with values 0 and 10.

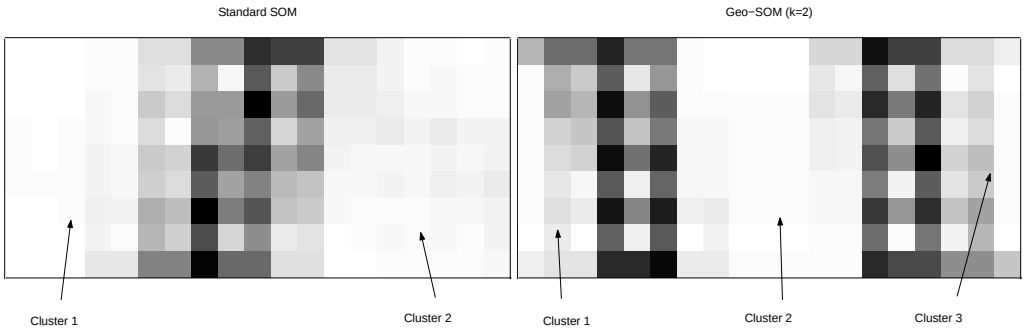

Figure 3 - U-matrices obtained with a standard SOM (left) and GEO-SOM (right).

If we use a Geo-SOM, then points with similar *z* will only be clustered together if they are spatially close. Using a geographic tolerance *t*=2, for example, we obtain the U-Matrix presented in the right side of Figure 3, where we can clearly identify the 3 clusters.

3 Visualization environment

To illustrate the integration between the Geo-SOM results and GIS we used a dataset of mainland Portugal. The objective is to explore the tools that the user can employ in order to discover knowledge through the use of the Geo-SOM. We emphasize the exploration environment and the visualization possibilities as these are closely connected with the exploratory nature of the task.

The dataset used has a high dimensionality, containing 65 variables which characterize each one of the 250 counties of mainland Portugal. For each of these areal units we calculated the geometric centroid and included the *x,y* geographic coordinates in the dataset. We ran 2 Geo-SOM using different geographic tolerance parameters (0 and 1) with 50 units (5*10).

For visualization purposes, the Geo-SOM (implemented in Matlab®) produces 3 data files with the following names and contents:

Patterns file – Contains one line for each county. The first *n* columns contain the *n* variables that characterize the county (direct copy of the input datafile), followed by number (id) of the unit to which they are mapped in the Geo-SOM.

Units File – Contains one line for each unit. The first *n* columns contain the *n* variables that characterize each unit. The last column contains the average quantization for the Thiessen (or Voronoi) polygon defined by the unit's geographical coordinates. It must be stressed that this is not the quantization error as measured by the standard SOM. To calculate the standard quantization error, we measure the difference (considering all variables) between the data patterns and the units to which they are mapped (also considering all variables). To calculate the quantization error for the Thiessen polygons, we measure the difference (considering all variables) between the data patterns and the units geographically closer.

U-Mat file – U-Matrix of the Geo-SOM, in which the distances in the output space between the neighboring units are expressed.

The exploration environment was developed based on ArcView®, using multiple dynamically linked windows and the possibility of linking the files presented above. The final result is an environment where the user can probe the information available in different windows and build *what-if* scenarios.

The typical visualization setting includes a window with a component plane superimposed on the U-Matrix, linked with a second window which displays the geographical map (in this case of mainland Portugal) and a graph or database window where the selected elements are

displayed. Figure 4 shows such an environment. On the left side of the picture the orange colored map consists of the U-Matrix for the Geo-SOM (0) with 50 units. The green squares superimposed on the U-Matrix represent a component plane, in this case the dimension of the square represents the GDP per capita (Gross Domestic Product) for each unit. The user may select one or more units in this window, and the corresponding data is highlighted in the other windows. On the right side of the figure the map of Portugal is presented, along with a database window where the elements selected in the U-Matrix are represented.

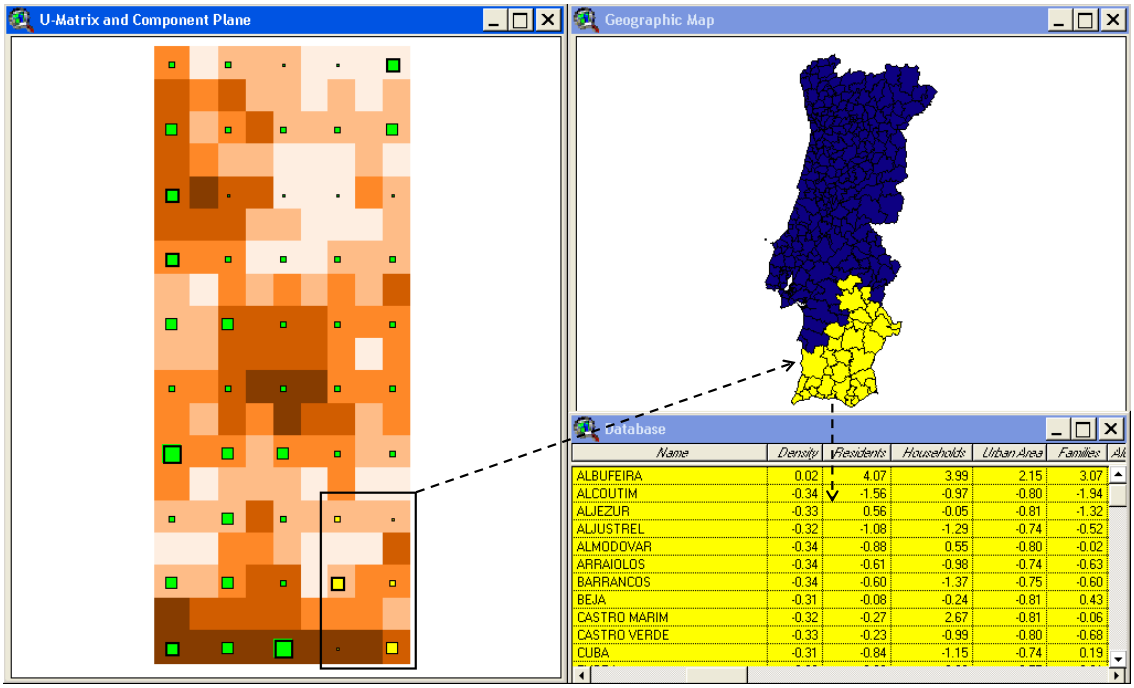

Figure 4 - The exploration environment developed to support the Geo-SOM

Another visualization setting is the Thiessen polygons window, shown in Figure 5. To obtain these maps it was necessary to geocode the units and define the Thiessen polygons for the set of 50 units. As shown in the figure one can produce a series of maps depicting the different variables used to develop the classification. In the particular case of the Geo-SOM (0) the counties that are contained within each of the Thiessen polygons are exactly the same that are classified in each unit. This concept can be quite useful when building homogenous areas. The idea is to use the Thiessen polygon mapping in order to group the areas of small variations and isolating areas of high variation. Additionally, the Thiessen polygons map can be used as a components plane where the different variables can be represented increasing the information context available for the user (as can be seen in the center and right maps of Figure 5).

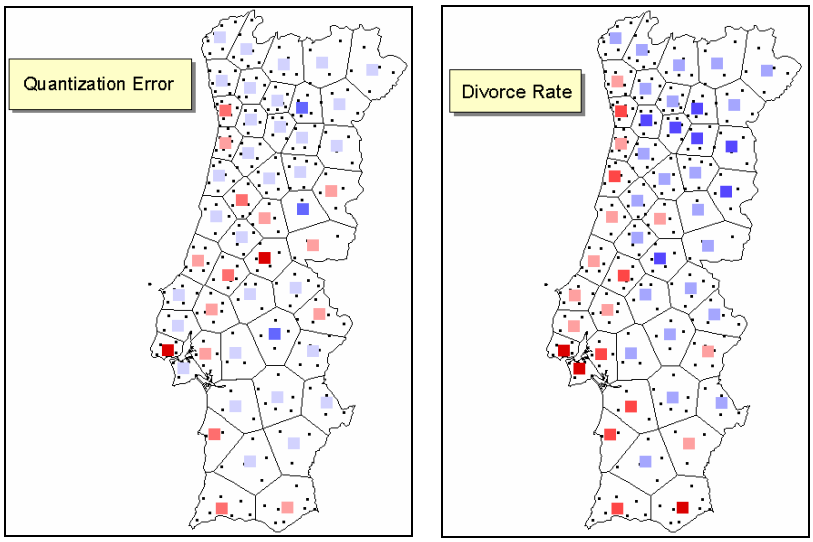

Figure 5 – Using the Geo-SOM (0) to build Thiessen polygons and mapping different variables

Effect of the *t* parameter: comparing Geo-SOM(0) and Geo-SOM (1)

A Geo-SOM (0), by choosing the BMU solely with the geographic coordinates, is calculating local averages of the remaining features. The locations where those averages are calculated, and consequently the Thiessen polygons obtained, depend on the geographic density of the available data patterns. The use of the Geo-SOM (1) is more complex as the values of the different units do not involve calculations solely based on the geographically closest neighbors. The workings of the Geo-SOM (1) (and higher geographic tolerances) can be described as "averages of similar counties" in the sense that within a geographic tolerance the Geo-SOM will try to group similar counties. This can be viewed as the possibility of lessening the geographic constraint providing the Geo-SOM with the possibility of clustering counties with similar profiles and which are located in the same general area. In this case the results are not contiguous regions but sets of areas with similar characteristics that are relatively close in geographic terms.

In Figure 6 we compare the membership of a specific county (Braga, shown in red) in three different SOMs: a standard SOM (the *x,y* coordinates of the counties centroids were added to the 65 attribute variables), a Geo-SOM (0) and a Geo-SOM (1). In all three classifications Braga is grouped with different counties. In the standard SOM Braga, which is a district capital, is grouped in a cluster which contains most of the Porto Metropolitan Area, as well as two other district capitals (Viseu and Leiria). Both Viseu and Leiria are located far away from Braga. In the Geo-SOM (0), Braga is grouped in a geographically contiguous set which includes coastal counties north of Oporto Metropolitan Area. Finally, in Geo-SOM (1) only two other counties are grouped with Braga. The contiguous county, Guimarães, can be seen as a twin city as they share a number of administrative services and a university campus. Viana do Castelo, like Braga, is also a district capital. This example shows some fundamental differences between the workings of the different SOM variants. The standard SOM clusters with a strong influence of the attribute variables. In the Geo-SOM (0), on the other hand, attribute variables are less relevant and geographic location becomes central. Finally, in Geo-SOM (1) a compromise between attributes and geographic location is achieved. It is probably useless to argue the superiority of any of these variants, as the combination of the three analysis produces an improved understanding of the problem. Nevertheless, we argue that from

a GIScience perspective it is sensible to use space as a determining factor in the outcome of clustering.

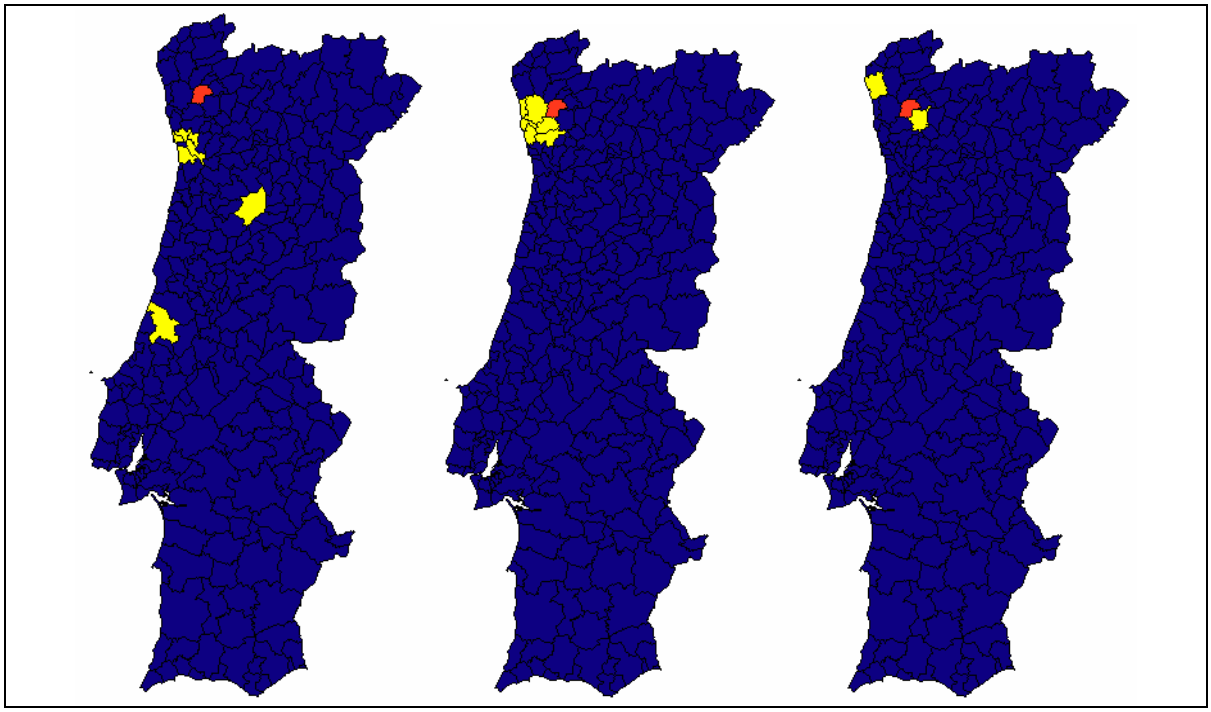

Figure 6 - Comparison of the areas clustered together with Braga using three different SOM variants: a Geo-enforced SOM (right), a Geo-SOM (0) (center) and a Geo-SOM (1) (left)

4 Conclusions and future work

The Geo-SOM can be thought of as a method which projects multidimensional data into the geographic space. The amount of geographic error and quantization error is controlled by the *t* parameter. Thus, as *t* increases the geographic error also increases and the quantization error decreases. The user must experiment with different *t* values in order to strike an acceptable trade-off. The fundamental aspect of the Geo-SOM is its ability to cluster in similar areas of the output space, geographic features (in our case counties) which are similar in terms of attributes, but more importantly are located in the same geographic area. This implies that the output space of the Geo-SOM has a geographic expression and can be geographically mapped. The output space is draped over the geographic space and the units are arranged in such a way that the quantization error is minimized constrained by the value of parameter *t* (geographic tolerance). The fundamental assumption of the Geo-SOM is that in spatial analysis space should take the centre stage and attribute variables should be analyzed within their spatial context.

During this paper we formulated and explained the fundaments of the Geo-SOM, additionally we tried to show that it constitutes a true knowledge discovery tool. It emphasizes the ability to rapidly and efficiently highlight and isolate unusual or unexpected patterns. It provides a spatial context to highly dimensional patterns and this helps revealing subtleties underlying spatial interactions between neighbours. A large number of issues remain to be explored in the Geo-SOM. The effect that the relation between the density of the input patterns (in the geographic space) and the distance between them (in the variable space) has on the distribution

of the units is still an open problem. Another interesting issue to address in future developments is the possibility of using dynamical *t* values. The idea is to specify the *t* parameter according to the specific spatial autocorrelation index of the area of the input pattern.

References

[1] T. Kohonen, Self-Organizing Maps, 3rd ed. Berlin-Heidelberg: Springer, 2001.

[2] S. Openshaw, M. Blake, and C. Wymer, "Using neurocomputing methods to classify Britain's residential areas," in Innovations in GIS, vol. 2, P. Fisher, Ed.: Taylor and Francis, 1995, pp. 97-111.

[3] A. Skupin and R. Hagelman, "Attribute space visualization of demographic change," presented at eleventh ACM international symposium on Advances in geographic information systems, New Orleans, Louisiana, USA, 2003.

[4] B. Jiang and L. Harrie, "Selection of Streets from a Network Using Self-Organizing Maps," Transactions in GIS, vol. 8, pp. 335-350, 2004.

[5] M. Takatsuka, "An Application of the Self-Organizing Map and Interactive 3-D Visualization to Geospacial Data," presented at GeoComputation'01 (6th International Conference on GeoComputation), Brisbane, Australia, 2001.

[6] A. Skupin, "Cartographic Considerations for Map-like Interfaces to Digital Libraries," presented at Workshop on Visual Interfaces to Digital Libraries, Roanoke, Virginia, 2001.

[7] A. Skupin and S. I. Fabrikant, "Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization," Cartography and Geographic Information Science, vol. 30, pp. 95-115, 2003.

[8] L. Girardin, "Mapping the virtual geography of the World Wide Web," presented at Fifth International World Wide Web Conference, Paris, France, 1996.

[9] D. M. Mark, A. Skupin, and B. Smith, "Features, Objects, and Other Things: Ontological Distinctions in the Geographic Domain," in Lecture Notes in Computer Science, vol. 2205. Heidelberg: Springer-Verlag, 2001.

[10] P. Agarwal, "Contested nature of 'place': knowledge mapping for resolving ontological distinctions between geographical concepts," in Lecture Notes in Computer Science, vol. 3234, M. Egenhofer, C. Freksa, and H. Miller, Eds.: Springer-Verlag, Berlin, 2004, pp. 1-21.

[11] W. Tobler, "A Computer Model Simulating Urban Growth in the Detroit Region," Economic Geography, vol. 46, pp. 234-240, 1970.

[12] F. Bação, V. Lobo, and M. Painho, "Geo-Self-Organizing Map (Geo-SOM) for building and exploring homogeneous regions," in Lecture Notes in Computer Science, vol. 3234, M. Egenhofer, C. Freksa, and H. Miller, Eds.: Springer-Verlag, Berlin, 2004, pp. 22-37.

[13] T. Kohonen, "The Hypermap Architecture," in Artificial Neural Networks, vol. 1, T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, Eds.: Elsevier Science Publishers, 1991, pp. 1357-1360.

[14] J. Kangas, "Temporal Knowledge in Locations of Activations in a Self-Organizing Map," in Artificial Neural Networks, vol. 2, J. T. I. Aleksander, Ed.: Elsevier Science Publisher, 1992, pp. 117-120.

[15] V. Chandrasekaran and M. Palaniswami, "Spatio-temporal Feature Maps using Gated Neuronal Architecture," IEEE Transactions on Neural Networks, vol. 6, pp. 1119-1131, 1995.

[16] F. Bação, V. Lobo, and M. Painho, "The Self-Organizing Map, Geo-SOM, and relevant variants for GeoSciences," Computers & Geosciences, vol. 31, pp. 155-163, 2004.