The Self-Organizing Map and it's variants as tools for geodemographical data analysis: the case of Lisbon's Metropolitan Area

Victor Lobo[123], Fernando Bação[1], Marco Painho[1]
[1]ISEGI – UNL, Lisbon, Portugal
[2]Portuguese Naval Academy, Lisbon, Portugal
[3]Contact author: vlobo@isegi.unl.pt

SUMMARY

In this paper we explore the advantages of using Self-Organized Maps (SOM) when analyzing geodemographic data. The standard SOM algorithm is presented, together with a few variants. The strengths and weaknesses of those different variants are shown, and their use in this type of problems is discussed. A practical application of these techniques is given, analyzing geodemographic data from Lisbon's metropolitan area.

KEYWORDS: *Geodemographic typologies, Self-Organizing Map, Spatial-kangas map*

INTRODUCTION

The convergence of Geographical Information Systems (GIS) and census data made available an immense volume of digital geo-referenced data. This fact created opportunities for developing an improved understanding of a number of socio-economic phenomena that are at the heart of human geography. Nevertheless, it also shaped new challenges and raised unexpected difficulties on the analysis of multivariate data spatially referenced. 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 GIScience 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 constitute one of the ways by which private and public organizations can benefit from massive reduction of census data. Although geodemographics may use a variety of different data sources, census data remains the backbone of these typologies. Geodemographics can be seen as a "*snapshot of the most important population types within a given locality*" (Birkin and Clarke 1998).

A number of problems related with geodemographic typologies have been described in the relevant literature (Birkin and Clarke 1998; Feng and Flowerdew 1998; Harris 1998; Openshaw, Blake *et al.* 1995; Openshaw and Wymer 1994). Among these there are problems that are inescapable, but there are also difficulties that can be mitigated as long as proper approaches and tools are available. It has been suggested that neurocomputing paradigms (Openshaw, Blake *et al.* 1995; Openshaw and Wymer 1994) may provide the power and flexibility to improve the overall quality of this kind of classifications. Fuzzy clustering methods have also been suggested as a step forward in the sophistication of geodemographics (Feng and Flowerdew 1998).
Although limitations and debilities can be found in the available methodologies used to develop geodemographic typologies, they have proved to be useful tools for a number of areas (Goss 1995). For this reason, it is important to develop efforts in order to improve the quality and reliability of these classifications.

(Openshaw, Blake *et al.* 1995) argue that the quality of geodemographic typologies is directly related with three major factors. "*First, the classification algorithm that is used; second, the manner and extent to which knowledge about socio-spatial structure is used and is represented in the* *classification; and, third, the sensitivity of the technology to the geographical realities of the spatial census data classification problem*". The combination of better classification algorithms and explicit focus on the issues of geographic representation will certainly reap sizeable rewards. The improvements pursued here lie on the capability of introducing more geographical knowledge within the classification process, which leads to geodemographic typologies. The option in this work is to emphasize the importance of the *geo* prefix in geodemographics. It is considered here that the *geo* component has been neglected in spite of being the true strength behind geodemographics.

The objective of this paper is to evaluate the interest of using neural classifiers in order to analyze geodemographic data. More specifically we try to explore the potential and flexibility of the Self-Organizing Map to develop approaches that can improve the existing classification methods in geodemographic classifications. In this paper we present 4 different ways of using the Self-Organizing Map and it's variants, which constitute promising approaches to deal with the problems of incorporating the geographic reasoning within geodemographic typologies. In the next section a brief introduction to the main problems of geodemographics is presented, followed by a presentation of the data used in the practical example. An explanation of the workings of SOM is given next, along with a more detailed explanation of the variants used. After that we present the main results of the application of the different variants to the Lisbon metropolitan area data. Finally, some conclusions are drawn.

PROBLEMS AND DIFFICULTIES IN DEVELOPING GEODEMOGRAPHIC TYPOLOGIES

One of the most important characteristics of gedemographics is the fuzziness associated on one hand with the data use to develop the typologies and on the other hand with the typologies themselves as they result from a "*crisp*" classification process (Feng and Flowerdew 1998). In this context the term fuzziness is employed to underline the uncertainties that arise from imprecision and ambiguity that characterizes the end product resulting from the classification process of small area census data.

According to (Feng and Flowerdew 1998) there are two different types of fuzziness in geodemographics. One related with the attribute space, the other one associated with the geographical space. The first kind of fuzziness is due to the "*all or nothing nature of the classification assignment*" (Openshaw, Blake *et al.* 1995), which conceals the fact that enumerations districts (EDs) may be close to more than one neighbourhood type in the attribute space. A specific ED can have similarities with more than one neighbourhood type, this is a probable circumstance especially if we take into account that we are dealing with multivariate classifications. An ED can resemblance different neighbourhood types depending on the subset of variables considered. This fact is even more problematic if the sub-optimal nature of the classifier algorithms usually used (K-means algorithm) is taken into account (Bradley and Fayyad 1998). As a result EDs which differ by small amounts in the classification space can be assigned to very different clusters (Feng and Flowerdew 1998).

The geographical fuzziness comes from the arbitrary nature of the EDs which serve as the elementary units on which geodemographic classifications are based, the well known problem of the modifiable areal unit problem (Openshaw 1984). In fact, the nature of these units reflect the operational needs that guide data collection, and for that matter should not be assumed to have any type of homogeneity in terms of socio-economic characteristics (Feng and Flowerdew 1998; Morphet 1993). The lines that separate EDs have no meaning in terms of neighbourhood, they can isolate true neighbourhoods but more probably just separate the same neighbourhood into two or more EDs. The main assumption that supports geodemographic classifications is the 1º law of Geography (Tobler 1970), which in this specific case can be rephrased as "*people who live near to each other tend to share behavioral characteristics despite other differences*" (Feng and Flowerdew 1998). The problem lies on 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 neighborhood effects might be expressed at different scales in different areas of the study region and only fortuitously will coincide with the scale denoted by the EDs.

Associated with the issues of representation there is also the problem of the ecological fallacy (Freedman 1999; Freedman, Klein *et al.* 1998; King 1997), which results from the assumption that the characteristics observed at aggregated levels apply equally well to individual observations. Aggregate area descriptors represent relatively crude averages of the population, and selecting one particular classification cluster for one specific ED, may result in the overlook of other important household groups that may also be part of the ED (Birkin and Clarke 1998). The false sense of homogeneity constitutes one important danger in geodemographic classifications, and "*crisp*" clustering algorithms only makes it worst. In spite of some efforts the ecological fallacy can only be tackled if finer resolution elementary units are available for study.

(Openshaw, Blake *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 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 to present more extreme results, due to its size. Conventional classifiers give identical 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, Blake *et al.* 1995). This fact also emphasizes the need to use classification algorithms which are robust to outliers, unfortunately this is not the case of the K-means algorithm, probably the most used algorithm in geodemographic classifications.

LISBON'S METROPOLITAN AREA DATA

For this study we used data form the Portuguese Institute of Statistics, referring to Lisbon's Metropolitan Area. This data is at the Enumeration District (ED) level (*Figure 1*), and refers to 65 socio-demographic variables. These variables describe ED based on 6 main topics: information about buildings, families, households, age structure, education levels, and economic activities. Additionally, we introduced two explicitly geographic variables, representing by the *x* and *y* coordinates of the centroids of the EDs. These centroids were calculated in ArcGis and represent the geometric centroids of the shapes of the EDs.

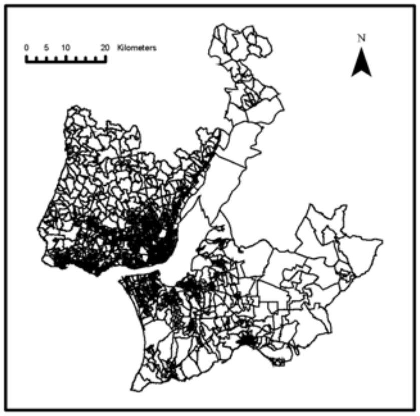

Figure 1 - Map of Lisbon's metropolitan area with its Enumeration Districts.

All variables used were made to be invariant to size, by calculating ratios whenever necessary. Further more all were normalized to be in the [0, 1] interval. In the particular case of the *x*, *y* coordinates, and in order to preserve the original form of the region, the *y* coordinate was allowed to have a different range so as to keep proportionality with the *x* coordinate.

SELF ORGANIZING MAPS (SOM)

Although the term "*Self-Organizing Map*" could be applied to a number of different approaches, we shall use it as a synonym of Kohonen's Self Organizing Map, or SOM for short. These maps are also referred to as "*Kohonen Neural Networks*" (Fu 1994), "*Self Organizing Feature Maps-SOFM*", "*Topology preserving feature maps*" (Kohonen 1995), or some variant of these names. Kohonen describes SOM as a "*visualization and analysis tool for high dimensional data*", but they have been used for clustering (Vesanto and Alhoniemi 2000), dimensionality reduction, classification, sampling, vector quantization, and data-mining (Kohonen 2001).

The basic idea of a SOM is to map the data patterns onto a *n*-dimensional grid of neurons or units. That grid forms what is known as the *output space*, as opposed to the *input space* that 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 output space will usually be 2-dimensional, and most of the implementations of SOM use a rectangular grid of units. So as to provide even distances between the units in the output space, hexagonal grids are sometimes used (Kohonen, Hynninen *et al.* 1995). Single-dimensional SOMs are common (eg. for solving the traveling salesman problem), and some authors have used 3-dimensional SOMs. Using higher dimensional SOMs, although posing no theoretical obstacle, is rare, since it is not possible to easily visualize the output space.

Each unit, being an input layer unit, has as many weights or coefficients as the input patterns, and can thus be regarded as a vector in the same space as the patterns. When we train or use a SOM with a given input pattern, we calculate the distance between that pattern and every unit in the network. We then select the unit that is closest as the winning unit, and say that the pattern is mapped onto that unit. If the SOM has been trained successfully, then patterns that are close in the input space will be mapped to neurons that are close (or the same) in the output space, and vice-versa. Thus, SOM is "*topology preserving*" in the sense that (as far as possible) neighborhoods are preserved through the mapping process.

Before training, the neurons may be initialized randomly. During the first part of training, they are "*spread out*", and pulled towards the general area (in the input space) where they will stay. This is usually called the unfolding phase of training. After this phase, the general shape of the network in the input space is defined, and we can then proceed to the fine tuning phase, where we will match the neurons as far as possible to the input patterns, thus decreasing the quantization error.

The Basic Learning Algorithm

The basic SOM learning algorithm may be described as follows.

Let $\boldsymbol{w}_{ij}$ be the weight vector associated with unit positioned at column *i* row *j*. Let $\boldsymbol{x}_k$ be the vector associated with pattern *k*. Let d_{ij} be the distance between weight vector $\boldsymbol{w}_{ij}$ and a given pattern. Let *h* be a neighborhood function described below. Let α be the learning rate also described below.
For each input pattern:

a) Calculate the distance between the pattern and all units of the SOM ($d_{ij} = \| x_k - w_{ij} \|$)
This is what we call the calculation phase.

b) Select the nearest unit as winner w_{winner} ($w_{ij} : d_{ij} = \min(d_{mn})$).
This is what we call the voting phase.

c) Update each unit of the SOM according to the update function

$$w_{ij} = w_{ij} + \alpha\, h(w_{winner},w_{ij}) \, \| x_k - w_{ij} \|$$

Where α is the learning rate, and *h* is a neighborhood function.

This is what we call the updating phase.

d) Repeat the steps a) to c), and update the learning parameters, until a certain stopping criterion is met.

This algorithm can be applied to a SOM with any dimension. The learning rate α, sometimes referred to as η, must converge to 0 so as to guarantee convergence and stability for the SOM. The decrease from the initial value of this parameter to 0 is usually done linearly, but any function may be used. The update of these two parameters may also be done after each training pattern is processed or after the whole training set is processed.

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. It is large for units that are close in the output space, and small (or 0) for units far away. Usually, it is a function that has a maximum at the center, monotonically decreases up to a radius *r* (sometimes called the neighborhood radius) and is zero from there onwards. For the sake of simplicity, this radius is sometimes omitted as an explicit parameter. The two most common neighborhood functions are the bell-shaped (Gaussian-like) and the square (or bubble), in both cases, we force $r \rightarrow 0$ during training to guarantee convergence and stability.

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 neighboring neurons are represented, usually by a color code. If distances between neighboring 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. Distances are usually coded as gray shades, but with the software used for this paper short distances are coded as blue, and large ones as red.

SOM VARIANTS

A very large number of SOM variants have been proposed, and reviews of some of these can be found in (Kangas, Kohonen *et al.* 1990; Kaski 1997; Kohonen 2001). Some of these are just parameterizations or minor adjustments to the basic SOM algorithm, while others differ quite a lot and do not have the same mapping and visualization properties of the SOM. We shall now review some of the major variants, and present ways in which they may be applied to geo-referenced problems.

One simple way of experimenting with SOM, producing quasi-variants and testing spatial effects is through the use of pre-processing. Our first experiment, which is called here Geo-enforced quasi-variant, was based on weighting the geographic coordinates in order to make them as important as the whole other 65 variables. This way each coordinate was multiplied by 32.5, thus attributing the same importance to the two geographical variables as the entire set of socio/demographic variables.

Introducing the first law of Geography (Tobler 1970) in the training of a SOM would suggest that when seeking the best match in a SOM for a certain data pattern, only the neurons geographically closer should be searched. This approach has similarities with the Hypermap approach (Kohonen 1991), where only part of the input features are used to find the best match, and with the Kangas architecture (Kangas 1992) 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 – (Chandrasekaran and Palaniswami 1995), where a "*spatial gating function*" is used, together with a similar temporal gating function, to select the next winner unit. More generally the idea of selecting only a subset of neurons as candidates for winning unit can lead to what we call a spatial-kangas map (see *Figure 2*). In this architecture the selection of the best match unit, or winning neuron, is done in two steps. First, a best match is searched using only the geographical coordinates. Only the units in the output space vicinity of this first best match are then compared to the complete input pattern, to select the final best match. The units are then updated according to the standard rule. This Spatial-kangas map forces units close in the output space to be close in the input space too, thus creating clusters of areas that will be geographically close.

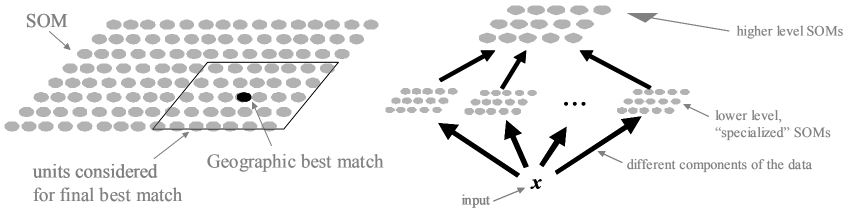

Figure 2 – Topology of a spatial-kangas map on the left and the hierarchical SOM on the right.

Hierarchical SOMs have been used in many contexts (Guimarães and Urfer 2000; Kohonen 2001). The main idea here is that instead of a single SOM that receives as inputs all the features of the data, a number of different SOMs are used, each receiving a certain part of the features. The output of these low level SOMs is then fed to a top level SOM (see *Figure 2*). This allows each lower level SOM to specialize on a particular aspect of the data, providing a potentially more intuitive interpretation of the results. Hierarchical SOMs have been used to deal with different time scales in temporal series (Behme, Brandt *et al.* 1993), or with different sources of data (Guimarães and Urfer 2000). In geo-referenced data, it makes sense to use the geographical coordinates as direct inputs to the top level SOM together with the "*summarized*" information provided by the lower level SOMs.

RESULTS

One of the major problems facing researchers in the field of geodemographics is related with the evaluation of the quality of different classifications. It is not clear what key indicators can be used to objectively evaluate the final product. Comparisons of different classifications can be made at two different levels, qualitative and quantitative. In the first case end users or experts should be called upon to judge the quality of the classification. This is not a simple process and is beyond the scope of this research. Nevertheless we provide description of major characteristics of the classifications performed and evaluate them based on our knowledge of the study region. Quantitative comparisons "*are more difficult because it is not clear what the performance measure should be.*" (Openshaw, Blake and Wymer 1995). Here we use the quantization error and the topological error in order to derive conclusions on the different SOM variants used. These measures can help understanding the effects of the different variants in the classification of the data and their impact on the complexity of the output space provided by the SOM. The quantization error is a measure of the quality of the representation of each ED, representing the distance, in the input space between each ED and the neuron to which it is mapped. This way the sum of the quantization errors of all the EDs gives a notion of the quality of the representation of the map produced. The topological error provides a measure of the complexity of the output space, by measuring the average number of times the second closest neighbor to a given ED is not mapped to a neighbor of the first.

In our experiments we used four different types of SOM:

A) Standard SOM.
B) Geo-enforced SOM
C) Spatial-kangas SOM
D) Hierarchical SOM

In all our experiments we used the SomToolbox for Matlab (Vesanto, Himberg *et al.* 2000), and trained SOMs with 30x20 units. The maps were initialized using the "*Randinit*" routine, with a rectangular grid, a “*sheet*” topology, “*bubble*” neighborhoods, and Euclidean distances. Two sequential training phases were used, with initial learning rates of 0.3 and 0.05, and initial neighborhood radius of 20 and 6. Both parameters where linearly reduced to 0. The initial neighborhood radius of the second phase was chosen after observing the U-Matrices obtained with the first phase, and corresponded to slightly more than the radius of the clusters observed in that phase. Ten training epochs were used in the first phase, and 20 in the second.

Using the standard SOM, the topology error in the final phase, as calculated by the SomToolbox, was around 6%, which indicates a fairly good unfolding, considering we are mapping a 67 dimensional space onto 2 dimensions. The U-matrix obtained is presented in Figure 3. In that figure, some clusters are quite evident and indicated with black circles.

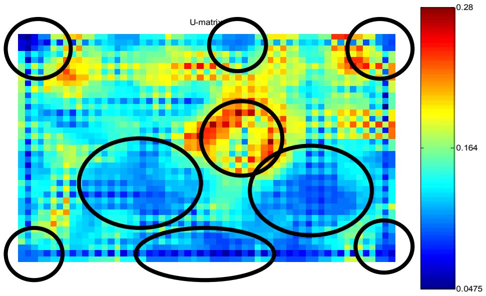

Figure 3 - U-matrix of a standard SOM of Lisbon's geodemographic data. Clusters are pointed out by black ellipses.

The quite pronounced boarders of the cluster located around (19, 9) has lead us to refine the characterization of this cluster of 60 EDs. A close look at the areas belonging to that cluster, shown in Figure 4, reveals areas with a percentage of rented houses quite higher than average, as well as buildings with more than 5 floors, and a high ratio of population between 14 and 19 years old. Personal knowledge of the areas in question indicates that these are areas with a high density of poor families relocated in “*social housing buildings*”. The most impressive aspect of this cluster is the precision with which it identifies the relocated “*social housing buildings*” in Lisbon metropolitan area. Similar analysis could be made to the areas belonging to the other clusters, although space constraints make it impossible here.

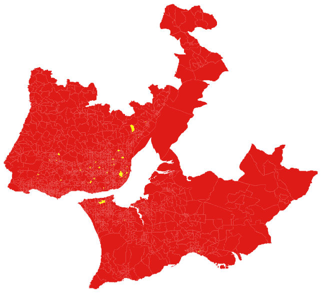

Figure 4 - Geographic representation of one of the clusters identified by the standard SOM. This cluster refers to poor families relocated in “*social housing buildings*”.

SOMs may also be used to find “*directions of maximum change*”. This concept is related to the concept of principal components (e.g.Jolliffe 1986) and principal curves (Hastie and Stuetzle 1989).

While in these latter cases a parametric curve is found along which the variance of data is maximized, when using a SOM the two axis form an arbitrarily non-parametric line that maximizes the separation amongst the data. Thus, by analyzing the map coordinates of the neurons to which each area is mapped, we can estimate which factors differentiate the most the areas. If we plot Lisbon's areas with colors proportional to the position where they are mapped on the SOM, we obtain the results shown in Figure 5. Some clear patterns emerge from those figures. The first coordinate increases with the degree of recent growth and of wealth, while the second one can be interpreted as an accessibility index. It is particularly interesting to note the clearness of major railroad and motorway axis around Lisbon.

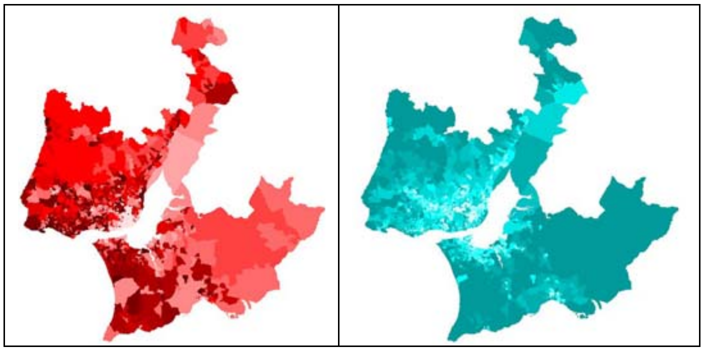

Figure 5 - Color coded coordinates of the geographical areas on the SOM. The left figure represents the coordinates along the first axis, while the right one the coordinates along the second one.

The results provided by the Geo-enforced quasi-variant differ from the standard SOM essentially in the fact that the output space produced is more complex as can be seen by observing that the topological error increases significantly. Interestingly, the quantization error does not increase, in fact it decreases, which means that the representation of the ED by the neurons doesn't deteriorate. When compared with the standard SOM it is clear that the quantization error in the Geo-enforced has a very different distribution (Figure 6), presenting an explicit geographical pattern unlike the standard SOM which presents a random distribution of the quantization errors. It is clear from the analysis of Figure 6 that the extra importance given to the geographic coordinates in the Geo-enforced quasi-variant, increases the quantization error in the periphery of the study area. Also, EDs with large areas will tend to be less well represented due to the fact that they are far from any other ED. The complexity of the output space, indicated by the increase of the topological error, of the Geo-enforced quasi-variant can be attributed to EDs with similar profiles but distant is geographic space. The output space, while unfolding, is affected by the tension that derives from the fact that some EDs which present similar profiles should be represented neighbors in the output space but are distant in geographic space.

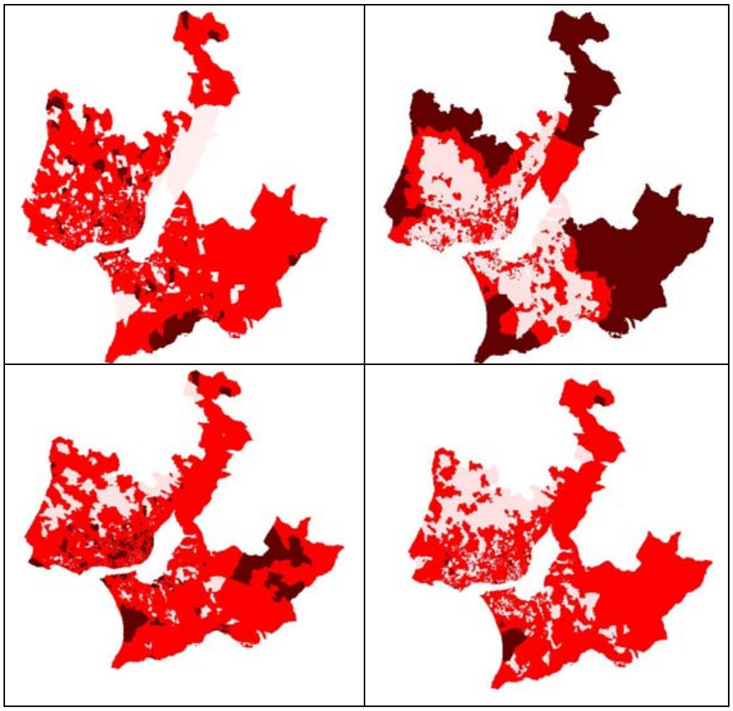

Figure 6 - Quantization errors for each ED. The top left refers to the standard SOM, top right the geo-enforced SOM, bottom left the spatial-kangas, and bottom right the hierarchical SOM.

In the case of the results referring to the Spatial-kangas map the quantization error is similar to the Standard SOM but once again, as in the case of the Geo-enforced quasi-variant, the output space is much more complex. In fact, the Spatial-kangas map is even more restrictive than the Geo-enforced quasi-variant, forcing EDs which are close in geographical space to be represented in near neurons of the output space. In a sense the areas with highest values of quantization error, in the Spatial-kangas map can seen as outliers in their particular neighborhood. It is possible to look at the classification produced by the Spatial-kangas map as incorporating *n* local classifications, instead of just one global classification.

	Quantization Error	Topological Error
Standard SOM	0.70	0.06
Geo-enforced	0.57	0.15
Spatial-Kangas	0.79	0.15
Hierarquical	0.60	0.07

CONCLUSIONS

The first idea we would like to highlight is that Standard SOM algorithm posses a number of important features and exploration tools that can be very useful in the geographical context. The possibility of identifying what can be called “natural clusters” is particularly relevant, as it allows the spotting of areas with particular characteristics, such as deprivation areas. Another interesting possibility provided by the SOM is the comprehensive analysis that can be done using the U-Matrix.

The colored surface can easily induce the understanding of the major clusters present in the data. Additionally, the coordinates of the areas on the SOM can give an idea of the defining components of the dataset. We also showed that there are many ways in which the SOM algorithm and it's variants can be used, and that the particular choice of variant will allows us to concentrate on particular aspects of the data.

Two major contributions of this paper are the adaptation of a hierarchical SOM for geodemographic data, and the development of the spatial-kangas map. We feel that this latter type of SOM can be very useful in problems where we want to impose a coherent geographical ordering of regions. We also wrote the Matlab software to implement these variations, which is available at http://www.isegi.unl.pt/ensino/docentes/fbacao/index.html.

REFERENCES

Behme, H., W. D. Brandt and H. W. Strube (1993). Speech Recognition by Hierarchical Segment Classification. ICANN 93, 416-419, Springer.

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

Bradley, P. and U. Fayyad (1998). Refining initial points for K-means clustering. International Conference on Machine Learning (ICML-98), 91-99.

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

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

Freedman, D. A., S. P. Klein, M. Ostland and M. R. Roberts (1998). "Review of A Solution to the Ecological Inference Problem." Journal of the American Statistical Association 93: 1518-22.

Freedman, D. A., prepared for the (1999). Ecological Inference and the Ecological Fallacy, International Encyclopedia of the Social & Behavioral Sciences.

Fu, L. (1994). Neural Netwrorks in Computer Intelligence, McGraw Hill.

Goss, J. (1995). "We Know Who You Are and We Know Where You Live: The Instrumental Rationality of Geodemographic Systems." Economic Geography 71: 171-98.

Guimarães, G. and W. Urfer (2000). Self-Organizing Maps and its Applications in Sleep Apnea Research and Molecular Genetics, University of Dortmund - Statistics Department.

Harris, R. (1998). Considering (mis-) representation in geodemographics and lifestyles. 3rd International Conference on GeoComputation.

Hastie, T. and W. Stuetzle (1989). "Principal curves." Journal of the American Statistical Association 84(406): 502-516.

Jolliffe, I. T. (1986). Principal component analysis. New York :, Springer-Verlag,.

Kangas, J. (1992). Temporal Knowledge in Locations of Activations in a Self-Organizing Map. Artificial Neural Networks. J. T. : I. Aleksander, Elsevier Science Publisher. 2: 117-120.

Kangas, J. A., T. K. Kohonen and J. T. Laaksonem (1990). "Variants of Self-Organizing Maps." IEEE Transactions on Neural Networks 1(1): 93-99.

Kaski, S. (1997). Data exploration using self-organizing maps. Helsinki, Finland, Helsinki University of Technology.

King, G. (1997). A Solution to the Ecological Inference Problem, Princeton University Press.

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

Kohonen, T. (1995). Self-Organizing Maps, Springer.

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

Kohonen, T., J. Hynninen, J. Kangas and J. Laaksonen (1995). The Self-Organizing Map Program Package, Helsinki University of Technology.

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., Ed. (1984). The Modifiable Areal Unit problem. CATMOG, Geo-abstracts. Norwich.

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

Openshaw, S. and C. Wymer (1994). Classifying and regionalizing census data. Census Users Handbook. S. Openshaw. Cambridge, UK, Geo Information International: 239-270.

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 Neural Networks for Exploratory Data Analysis. Intl. Neural Network Conf. INNC90, Paris, 305-308.

Vesanto, J. and E. Alhoniemi (2000). "Clustering of the Self-Organizing Map." IEEE Transactions on Neural Networks.

Vesanto, J., J. Himberg, E. Alhoniemi and J. Parhankangas (2000). SOM toolbox for Matlab 5. Espoo, HUT.