VISUALIZATION OF GEOSPATIAL DATA BY COMPONENT PLANES AND U-MATRIX

Marcos Aurélio Santos da Silva[1], Antônio Miguel Vieira Monteiro[2] and José Simeão Medeiros[2]

[1]Embrapa Tabuleiros Costeiros - Laboratory of Applied Geothecnologies Av. Beira Mar, 3250,49025-040, Aracaju, SE, Brazil.; [2]National Institute for Space Research - INPE, Caixa Postal 515, 12227-010, São José dos Campos, SP, Brazil.

Abstract: This paper shows an application of two visualization algorithms of multivariate data, U-matrix and Component Planes, in a matter of exploratory analysis of geospatial data. These algorithms were applied in the investigation of urban social exclusion/inclusion in the city of São José dos Campos - SP, Brasil.

Key words: Self-organizing Map; multivariate analysis; spatial analysis; visualization.

1. INTRODUCTION

Modern data acquisition techniques are offering tremendous opportunities that result in more geospatial data to be handled. Analyzing these data becomes a difficult task due to their complexity and hidden patterns. The complexity of the attribute space in such complex datasets does not always allows traditional deductive and statistically based approaches to analyse that data. Like many other techniques, Artificial Neural Network (ANN) is an emerging solution for pattern recognition. Among the ANN models, Self-Organizing Maps (SOM) is seen as a good technique for exploratory analysis of data (Kohonen, 2001). In this paper we explored a SOM, an unsupervised ANN, and their visualization algorithms, U-matrix (Ultsch, 1993) and Component Planes (Kohonen, 2001), for an exploratory analysis of geospatial data. These algorithms are visualization tools developed to work closer to SOM algorithm.

SOM has being applied successfully in a variety of problems of exploratory analysis of multivariated data (Kohonen, 2001), however, few are the works related to the analysis of geospatial data (Winter and Hewitson, 1994; Open-shaw and Turton, 1996; Babu, 1997; Kaski and Kohonen, 1996; Foody, 1999; Cereghino et al., 2001; Park et al., 2003). Urban geospatial problem is also a unexplored theme (Franzini et al., 2001). The main goal here is to find out how the dataset is distributed, how each variable correlates with each other and if there is some spatial correlation between the feature and physical spaces in a exploratory manner of urban geospatial data (Bailey and Gatrell, 1995).

Section 2 explains about the SOM and visualization related algorithms, U-matrix and Component Planes. Section 3 presents our case study, that is mapping urban social exclusion/inclusion in the city of São José dos Campos, São Paulo, Brazil. Finally, Section 4 shows our results and discussion and Section 5 our conclusions.

2. SELF-ORGANIZING MAPS

Kohonen Self-Organizing Map is a competitive artificial neural network structured in two layers (Kohonen, 2001), see Fig. 1. The first one represents the input data, x_k, the second one is a neuron's grid, usually bidimensional, full connected. Each neuron has one codevector associated, w_j.

The main goal of the SOM algorithm is to approximate the input dataset preserving structural local proximities to statistical properties among them. Therefore, it means that SOM acts as a data compressor and feature extractor. A learning process achieves this.

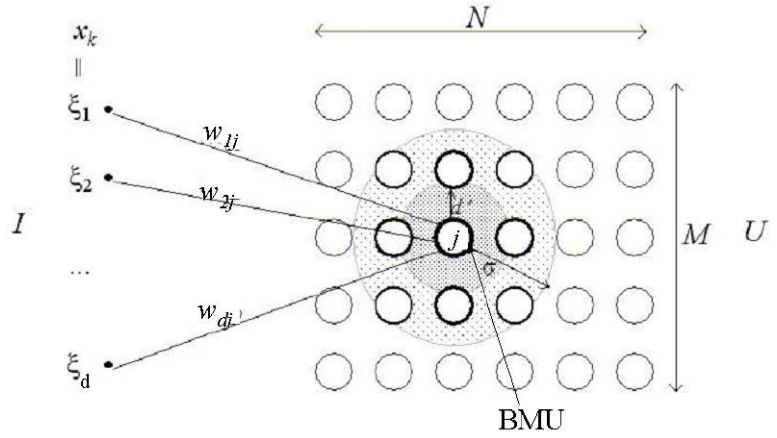

Figure 1. This picture illustrates an architecture of a two dimensional (NxM) SOM with a x_k input vector and w_j codevector.

The learning process can be split into three phases. In the first phase, competitive, each input pattern is presented to all neurons searching for the Best Match Unit (BMU) using Euclidean distance measure. In the second phase, cooperative, a neighborhood relation, among the BMU and the other neurons, is defined by a neighborhood kernel function (h_{ij}). Finally, in the last phase, adaptive, BMU and neighbors codevectors will be updated using some kind of adaptive rule, see Eqs.(1)-(2) (Vesanto, 1999).

$$s_i(t) = \sum_{j}^{n_{V_i}} x_j \tag{1}$$

$$w_i(t+1) = \frac{\sum_{j}^{m} h_{ji}(t) s_j(t)}{\sum_{j}^{m} n_{V_j} h_{ji}(t)} \tag{2}$$

where: s_i represents an input pattern sum for the *ith*-Voronoi region, V_i; and nV_i is the number of samples for the Voronoi dataset of the *ith*-neuron. *h(t)* is the neighborhood kernel function a *t* time; *m* is the number of Voronoi regions.

After the learning process codevectors should approximate, in a non-linear manner, the input data. Besides, SOM preserves the topological structure of

input data, so nearby patterns in the sample dataset are associated with nearby neurons in the SOM grid.

SOM can vary as learning algorithm, grid's topological structure, neighborhood kernel function, initial parameterization etc. For this work we have choose a SOM batch learning because it generates the same result for the same initial parameters and there isn't learning rate (Kohonen, 2001).

We defined and tested a limited set of SOM's configurations varying initial radius of neighborhood kernel function and the size of grid. All networks were bidimensional, has hexagonal grid, gaussian neighborhood kernel function, linear initialization, batch learning and only one learning phase. We used topological and quantization errors as quality metrics. Visual quality analysis was also used.

To proceed our exploratory analysis we used two software, SOM Toolbox and CASAA. SOM Toolbox is a Matlab based package that implements all algorithms needed here, but we only used it to show visual results (Vesanto et al., 1999). Connectionist Approach for Spatial Analysis of Area (CASAA) is a software for SOM simulation created to manipulate geospatial data stored in a TerraLib database format (Câmara et al., 2002). Figure 2 shows the main screen of that system.

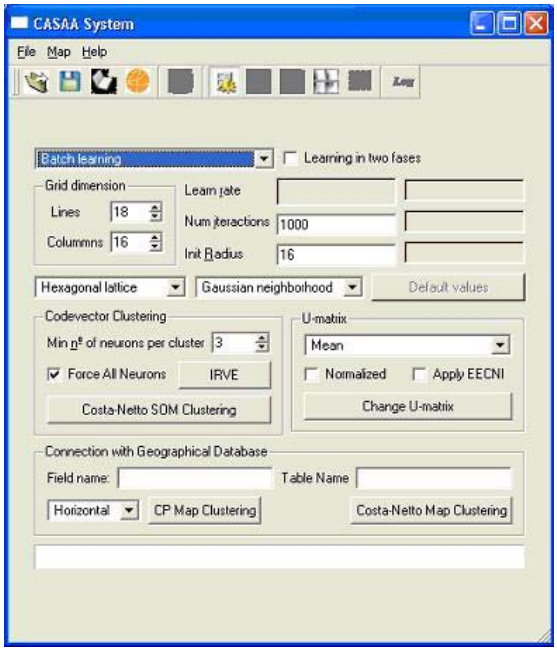

Figure 2. The main screen of the CASAA system.

2.1 U-matrix

The Unified distance matrix (U-matrix) makes the 2D visualization of multi-variate data possible using SOM's codevectors as data source. This is achieved by using topological relations property among neurons after the learning process. This algorithm generates a matrix where each component is a distance measure between two adjacent neurons, therefore we can visualize any multi-variated dataset in a two-dimensional display. Figure 3 shows an representation of an U-matrix calculation for an 3x3 2D hexagonal SOM. By U-matrix we can detect topological relations among neurons and infer about the input data structure.

High values in the U-matrix represent a frontier region between clusters, and low values represent a high degree of similarities among neurons on that region, clusters. This can be a visual task when we use some color schema. Nevertheless, this visual interpretation can be very hard for very short U-matrices because short SOM generates complex U-matrix when we are treating real datasets.

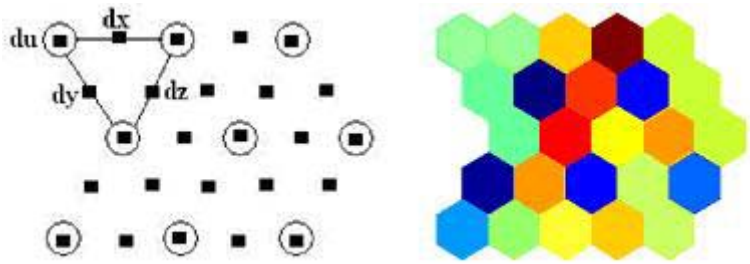

Figure 3. U-matrix generation example for an 3x3 hexagonal SOM.

2.2 Component Planes (CP)

After the learning process we can color each neuron according with each component value in the codevector . Therefore, we will take a colored SOM for each variable (Fig. 4). Trought these Component Planes we can realize emerging patterns of data distribution on SOM's grid (Kohonen, 2001), detect correlations among variables and the contribution of each one to the SOM differentiation only viewing the colored pattern for each Component Plane.

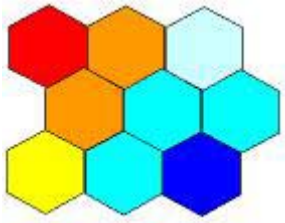

Figure 4. One Component Plane example for an 3x3 hexagonal SOM.

Some works used U-matrix and CP for spatial analysis, Kaski & Kohonen (1996) applied both to classify countries by their socioeconomic global indexes. Winter & Hewitson (1994) used CP to analyze the racial segregation matter in a City South Africa. Franzini et al. (2001) used a short SOM for clustering of urban spatial regions using socioeconomic data.

These applications are different from our approach for some reasons: a) we used a batch SOM learning algorithm, they used an online learning scheme; b) they did not work with intra-urban dataset or high resolution, they work with global values that means low resolution; c) they did not use any automatic segmentation of data using CP, here we proposed an simple but very efficient procedure to do this.

3. CASE STUDY: MAPPING URBAN SOCIAL EXCLUSION/INCLUSION

Our case study is the mapping urban social exclusion/inclusion in the City of São José dos Campos, São Paulo, through composite indexes created by Genovez (2002) using the Brazilian Census Bureau (IBGE) data for each urban census region (Fig. 5). The main goal here is to find out how the dataset is distributed, how each variable correlates with each other and if there is some spatial correlation between the feature and physical spaces (Bailey and Gatrell, 1995).

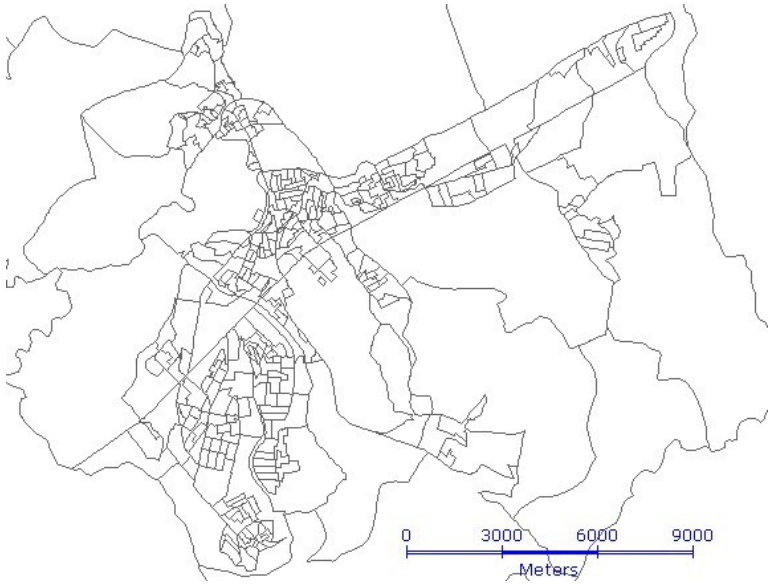

Figure 5. Urban census sectors for São José dos Campos, SP.

For this finding we used 8 socioeconomic indexes. Each one varying from -1, high level of social exclusion, to +1, high level of social inclusion. They are: familiar income (IFH), educational development (ED), educational stimulus (ES), longevity (LONG), environmental quality (EQ), home quality (PQ), concentration of family headed by women (CIWFH) and concentration of family headed by illiterate women (CWFH).

4. RESULTS AND DISCUSSION

Looking at the topological and quantization error graphics (Fig. 6) we realized that the quantization curve declines smoothly for an *y* asymptotic, so how big is the SOM how bit will be the quantization error, that means that more neurons represents better input pattern. In the topological graphic we can see an random behavior after 15th SOM's network configuration, so we cannot make any strong conclusion, only that for very short SOM we will have high values for the topological metric. These results were also reached by others authors (Kohonen, 2001). Therefore, we avoid short SOMs for U-matrix and Component Planes analysis.

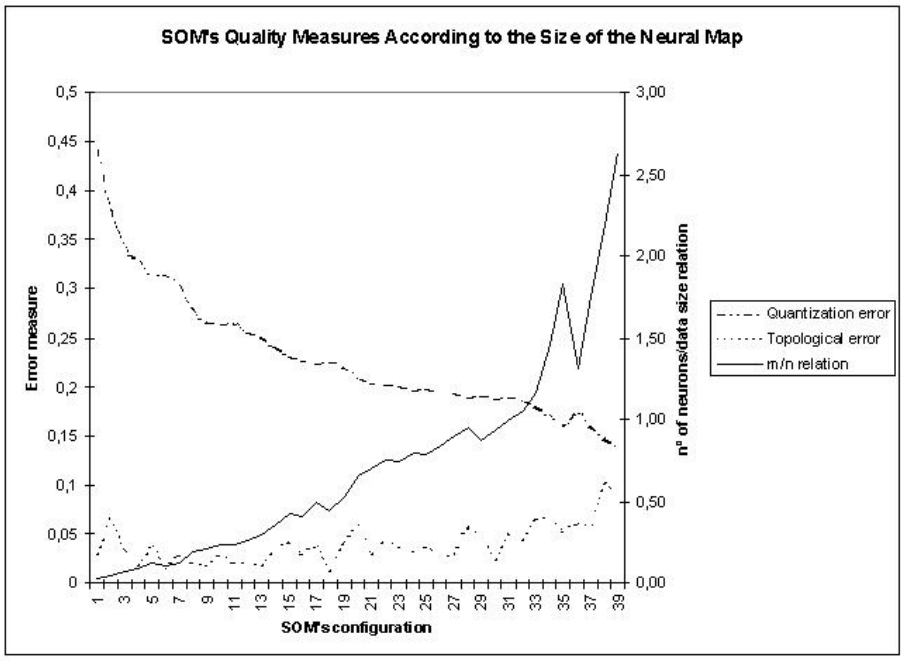

Figure 6. Error graphics for topologial and quantization error metrics.

4.1 General structure of dataset and outliers

A well formation of an U-matrix depends of the quality convergence learning, that depends of the dataset structure, SOM grid size and the initial radius of the neighborhood kernel function. Figure 7 shows that with a short U-matrix, 9x9, we cannot see significant differences thought colored patterns because there is too few neurons to absolve the complexity of dataset. The same figure shows also that to a big U-matrix, 99x59, happens the opposite, there are too many neurons to map the data complexity, so we can see many short clusters, that represents an overtrained neural network. This visual analysis confirms the graphical analysis of the quality metric errors and shows to us that very big neural networks does not fit well, considering our case. Consequently, we've chose an intermediate 20x15 SOM to proceed our analysis.

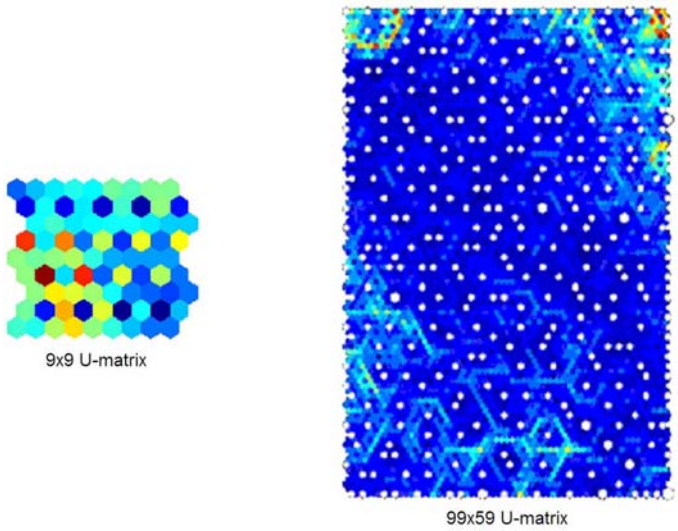

Figure 7. These figures show that for our case neither short nor big SOM does not fits well.

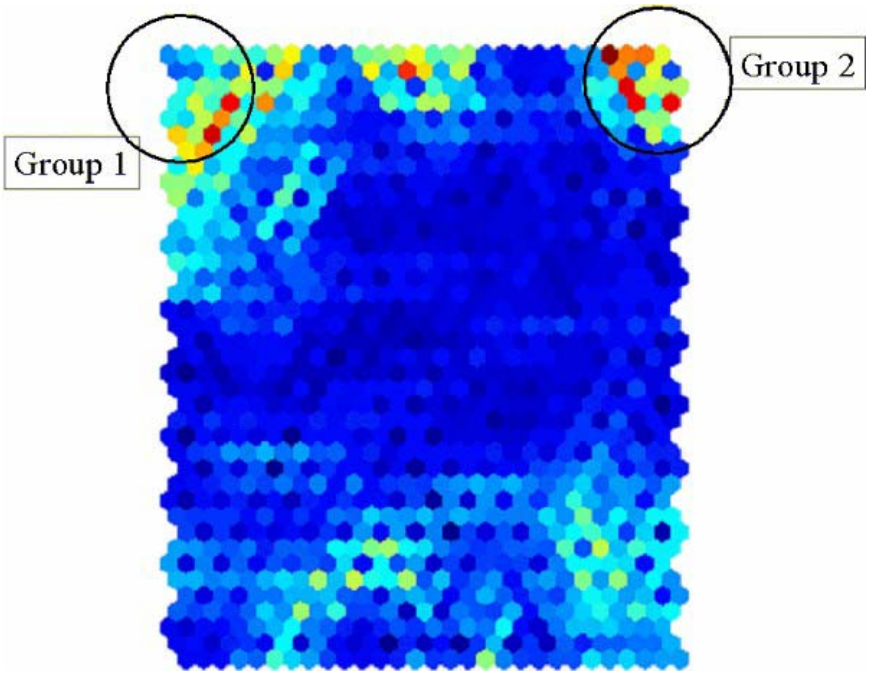

Figure 8. U-matrix for an 20x15 hexagonal SOM grid.

Figure 8 shows an U-matrix for an 20x15 SOM network. Looking at that picture we identify three patterns. First, there are two short clusters areas on the two top corners of picture. This suggests that these areas could be associated with census sectors outliers. Second, there is a big homogeneous area on the middle of the U-matrix. Here we can make two inferences, or the dataset has a high degree of similarity or the U-matrix could not separate the dataset in a proper manner. Finally, on the bottom of the U-matrix we register some differentiation, but not clear enough to be a cluster.

Mapping census areas associated to neurons located on the two top corners of U-matrix we identify, visually, that we have a set of census sectors outliers, Fig. 9. In fact, these areas have differences when compared with the others. Some of them has high exclusion indexes values but is located in a high inclusion area. Another ones is located in a high exclusion area but has high inclusion values for some indexes.

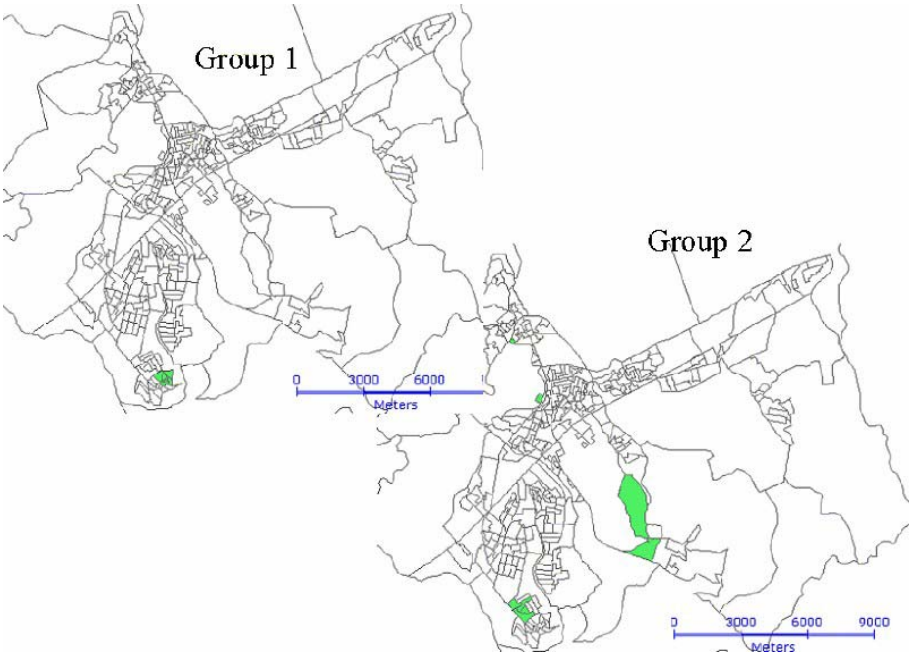

Figure 9. Census sectors highlighted represents outliers detected by the U-matrix visualization.

4.2 Component analysis and Spatial distribution of phenomena

If the U-matrix vary, significantly, for different size and neighbor radius the CP presents an opposite behavior. Although short SOM hide something

from us, in general, all CP shows a colored pattern that represents how variables is distributed in the SOM grid, Figure 10 shows two CPs for a short and a big CP SOM grid. Therefore, we've chose the same 20x15 hexagonal SOM to proceed our component analysis.

Figure 11 shows an U-matrix and eight CP, one for each studied variable. High values is associated with red color and low values with blue colors. High values, near +1, also mean high level of urban social inclusion, and low values, near -1, mean high degree of urban social exclusion. Looking at Component Planes we can identify some visible patterns among components. First, IFH and ED have a very similar color pattern, this suggests that they should be strongly correlated. Second, LONG and EQ presents a very homogeneous colored pattern, so these variables could have a low contribution for differentiation among dataset.

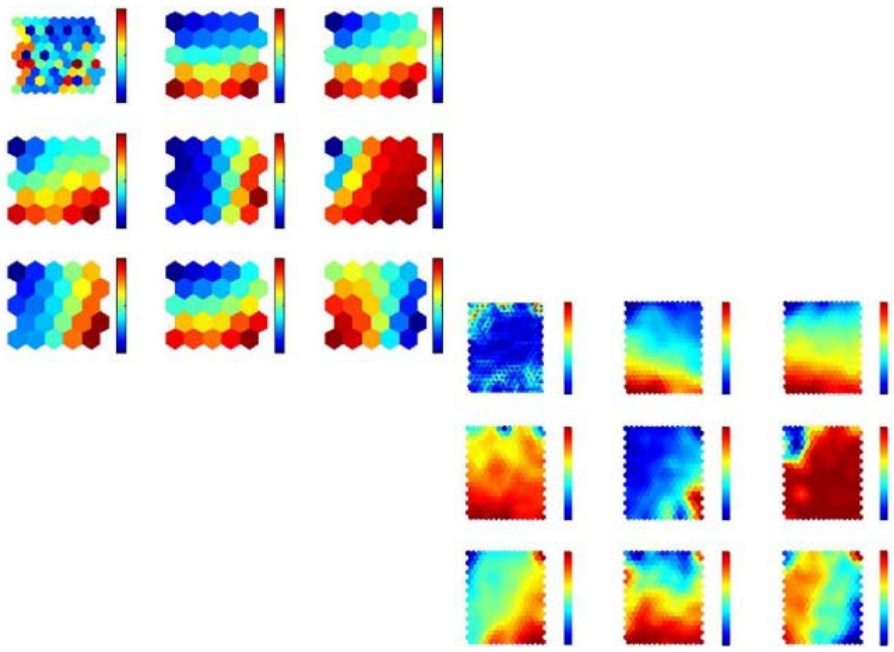

Figure 10. These figures shows that we can find the same color pattern for short or big Self-Organizing Maps.

Component Planes presents the same color pattern distribution for all CP, every one has high values on bottom and low values on top, with some exceptions. Therefore, we have a social exclusion-inclusion direction in SOM grid, and it is vertical (Fig. 12). Labeling all neurons starting on top to the bottom (vertical direction) and mapping this labeling to the map of census sectors we will generate a colored map illustrated by Fig. 13a.

Analyzing that map we concluded that the most inclusion's sectors, with high values, are located in center of the map, and the most of exclusion's sectors are located on peripherical areas. So, there are a general direction of the urban social exclusion/inclusion spatial distribution on sectors map, and it is center-to-peripherical zone. This result was also achieved using multivariated statistics (Genovez, 2002), see Fig. 13b.

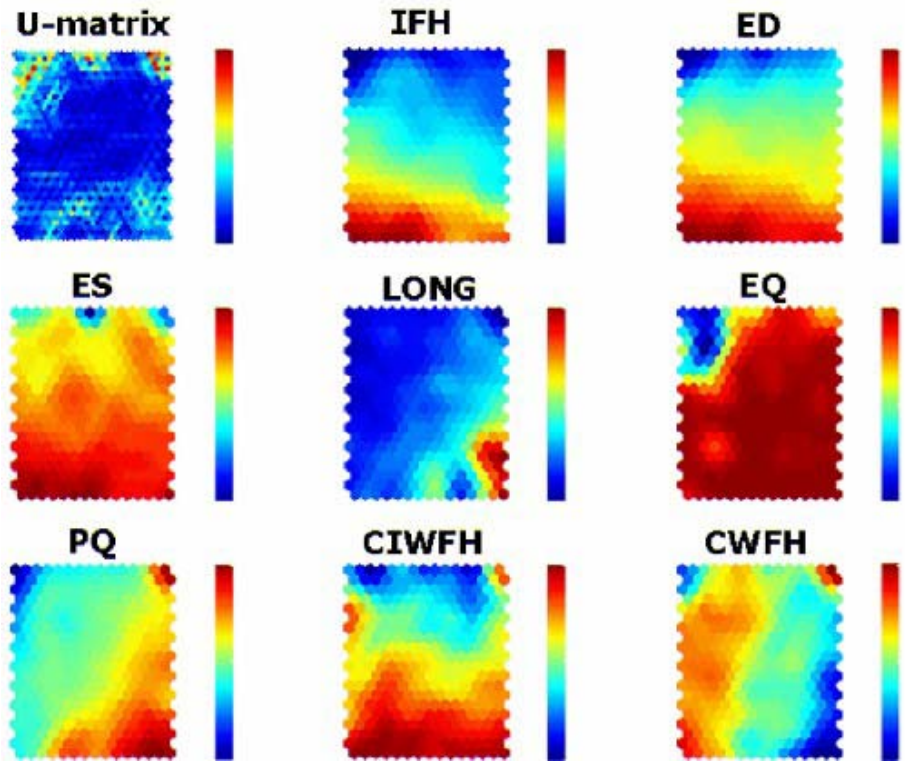

Figure 11. Component Planes for each variable using an 20x15 hexagonal SOM.

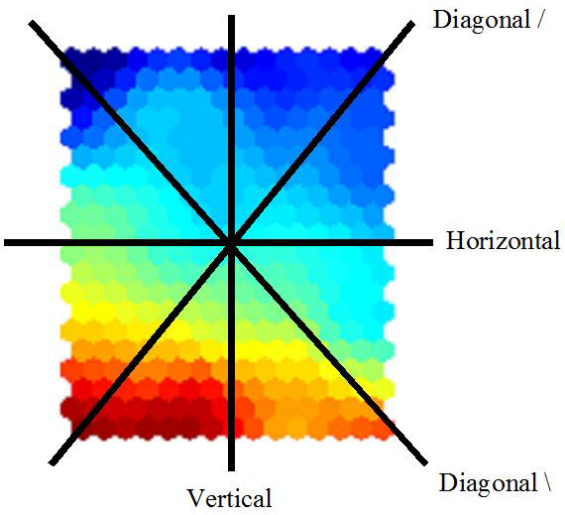

Figure 12. Possible directions of data distribution in SOM's grid.

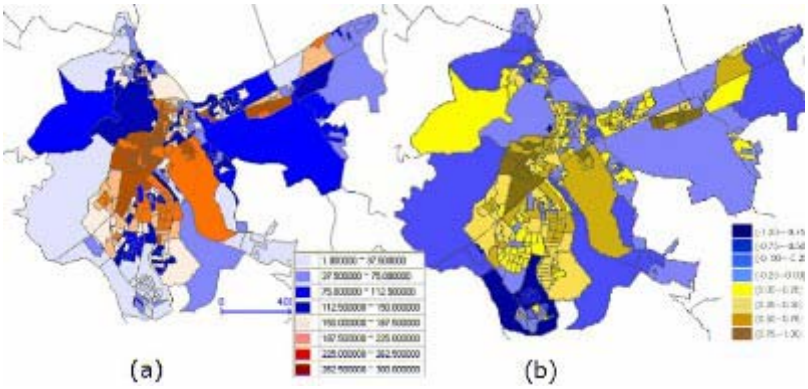

Figure 13. Figure 'a' shows the social inclusion/exclusion map segmentation using neural approach (SOM). Figure 'b' shows the same segmentation using statistical techniques, Iex created by Genovez (2002).

5. CONCLUSIONS

This experiment showed that SOM and related visualization algorithms could be applied as an exploratory tool to investigate an urban matter with good results. In our case we also confirmed previous statistical results reached by Genovez (2002). The quality of U-matrix and CP can be viewed easily and can be measured by topological and quantization errors with some

cautions. CP was also used as spatial analysis tool, mapping the direction distribution on SOM grid onto the census sectors map.

Initial radius of the neighborhood kernel function and the size of the SOM grid affect the final quality of the neural network and, consequently, the U-matrix and CP. Nevertheless, the size influences more than the initial radius. Short and big SOMs weren't good for visual interpretations.

The automatic segmentation of census sectors based on CP patterns helped us to see the spatial distribution of the phenomena and suggests that there are a strong spatial dependence between feature and physical spaces.

Although, more studies must be carried out to evaluate these visualization techniques for other spatial pattern problems, there is not any restriction for U-matrix and CP applications to other kind of spatial problems.

6. REFERENCES

Babu, G. P. (1997). Self-organizing neural networks for spatial data. *Pattern Recognition Letters,* 18:133-142.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis.* Longman.

Cereghino, R., Giraudel, J., and Compin, A. (2001). Spatial analysis of stream invertebrates distribution in the adour-garonne drainage basin (france), using kohonen self organizing maps. *Ecological Modelling,* 146(1-3): 167-180.

Câmara, G., Neves, M., Monteiro, A., Souza, R., Paiva, J. A., and Vinhas, L. (2002). Spring and terralib: Integrating spatial analysis and gis. In for Spatially Integrated Social Science, C., editor, *Specialist Meeting on Spatial Data Analysis Software Tools,* Santa Barbara, CA.

Foody, G. (1999). Applications of the self-organising feature map neural network in community data analysis. *Ecological Modelling,* 120:97-107.

Franzini, L., Bolchi, P., and Diappi, L. (2001). Self Organizing Maps,: A Clustering neural method for urban analysis. In Banos, A., Banos, F., Bolot, J., and Coutherut, C., editors, *Proceeding of the VRecontres de Théo Quant.,* pages 1—15.

Genovez, P. C. (2002). Território e desigualdades: Análise espacial intra-urbana no estudo da dinâmica de exclusão/inclusão social no espaço urbano em são josé dos campos-sp. Master's thesis, INPE.

Kaski, S. and Kohonen, T. (1996). Exploratory Data Analysis by The Self-Organizing Map: Structures of Welfare and Poverty in the World. In Refenes, A.-P. N., Abu-Mostafa, Y., Moody, J., and Weigend, A., editors, *Proceeding of the Third International Conference on Neural Networks in the Capital Markets,* pages 498-507. World Scientific.

Kohonen, T. (2001). *Self-organizing maps.* Springer. Third Edition.

Openshaw, S. and Turton, I. (1996). A parallel kohonen algorithm for the classification of large spatial datasets. *Computers & Geosciences,* 22(9): 1019-1026.

Park, Y.-S., Cereghino, R., Compin, A., and Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling,* 160:265-280.

Ultsch, A. (1993). Knowledge extraction from self-organizing neural networks. In Opitz, O., editor, *Information and Classification.* Springer.

Vesanto, J. (1999). Som based data visualization methods. *Intelligent Data Analysis,* 3:111-126.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (1999). Self-Organizing Map in Matlab: the SOM Toolbox. In *Proceeding of the Matlab DSP Conference,* pages 35^10.

Winter, K. and Hewitson, B. (1994). Self organizing maps - applications to census data. In Hewitson, B. and Crane, R., editors, *Neural nets: applications in geography*. Kluwer.