A Neural Qualitative Approach for Automatic Territorial Zoning

R. J. S. Maciel[1], M. A. Santos da Silva[*2], L. N. Matos[3] and M. H. G. Dompieri[2]

[1]Brazilian Agricultural Research Corporation, Postal Code 13083-886, Campinas, Brazil
[2]Brazilian Agricultural Research Corporation, Postal Code 49040-025, Aracaju, Brazil
[3]Department of Computer Science, Federal University of Sergipe, Postal Code 41000-000, São Cristóvão, Brazil
*Email: marcos.santos-silva@embrapa.br

Abstract

This article presents the application of the Self-Organizing Maps (SOM) as an exploratory tool for automatic territorial zoning by combining the handle of categorical data and the other for automatic clustering. The SOM online learning algorithm had been chosen to treat categorical data by using the dot product method and the Sorense-Dice binary similarity coefficient. To automatically perform a spatial clustering, an adaptation of the automatic clustering Costa-Netto algorithm had been also proposed. The correspondence analysis had been used to examine the profiles of each homogeneous zones. To explore the approach it has been performed the territorial zoning of the Alto Taquari River Basin, Brazil, using as input data a set of thematic maps. The results indicate the applicability of the approach to perform the exploratory territorial zoning.

Keywords: Self-organizing maps, Exploratory spatial analysis, Similarity coefficients, Correspondence Analysis, Alto Taquari River Basin - Brazil.

1. Introduction

Considering the increase in demand for territorial zoning based on the disponibility of geospatial data this study investigates the adaptation of the Self-Organizing Map (SOM) neural network (Kohonen 2001) for territorial zoning by means of a heuristic and automatic process for clustering categorical spatial data. The proposed method had been demonstrated for the territorial zoning of the Alto Taquari River Basin, Brazil, taking as input five thematic maps of the region. The results were compared with those obtained by (Silva & Santos 2011), who used the method of hierarchical cluster analysis of ward and multiple correspondence analysis.

2. The proposed approach

In this paper the SOM online learning algorithm adapted to treat categorical data (Lourenço et al. 2004; Henriques et al. 2012) and it had been used in pair with the SOM's graph partition segmentation, the Costa-Netto algorithm (Costa & Netto 2003; Santos da Silva et al. 2010) to, automatically, cluster thematic maps from the Alto Taquari River Basin, Mato Grosso do Sul, Brazil. The Sorense-Dice binary similarity measure[1] had been tested in the experiments. The Costa-Netto algorithm is defined as follows.

[1] When comparing two binary vectors **x** and **y**, a = number of times that $x_i = 1$ and $y_i = 1$; b = number of times that $x_i = 0$ and $y_i = 1$; c = number of times that $x_i = 1$, the Sorense-Dice measure will be equal to $\frac{2a}{2a+b+c}$. But, when considering the feature vectors generated from K thematic maps of large areas, the number of positive

Algorithm 1 Costa Netto Algorithm
Require: m — the number of neurons
Require: n — the number of input patterns
Require: $G = \{V,A\}$ — graph based on the neural network, where V is the neuron set and A is the set of arcs
Require: $H(\imath)$ — activity function of neuron i, $i \in \{1, \ldots, m\}$
Require: ω — $0.1 \leq \omega < 0.6$
1: $H_{min} \leftarrow \omega(n/m)$ — minimum allowed value for $H(\imath)$
2: **for** pair of adjacent neurons i and j **do**
3: Compute $d(\mathbf{w}_i ; \mathbf{w}_j)$ — the distance between the weights of neurons I and j
4: Compute $\bar{d}_{i,j}$ — mean distance between the i's and j's neighbors
5: Compute c_i — centroid of each neuron
6: **for all** pair of adjacent neurons i and j **do**
7: **if** $(d(\mathbf{w}_i ; \mathbf{w}_j) > 2\bar{d}_{i,j})$ **or**
8: $((H(\imath) < H_{min}/2$ **or** $H(j) < H_{min}/2)$ **and** $(H(i) = 0$ **or** $H(j) = 0))$ **or**
9: $(\mathbf{c}_i > 2d(\mathbf{w}_i ; \mathbf{w}_j)$ or $\mathbf{c}_j > 2d(\mathbf{w}_i ; \mathbf{w}_j))$ **then**
10: Remove arc$(\imath, j)$ from A
11: A distinct label is assigned to each set of connected neurons to identify the group.

At the end of the process, only connected nodes representing different groups shall remain. The algorithm had been adapted for categorical data and it used the Sorense-Dice binary distance to measure the proximity among code vectors. The code has been implemented in the SOMCode Project (Santos da Silva 2004).

The proposed approach for automatic territorial zoning through the classification of categorical data using SOM can be summarized as follows (Figure 1): 1) A routine for extracting features from thematic maps is applied to the thematic maps[2]; 2) The adapted Self-Organizing map is trained using this binary data as input; 3) The adapted Costa-Netto algorithm is applied to automatic cluster the neural network; 4) The classified thematic map is generated using the labeled SOM; 5) Interpretation of the profiles of each territorial zone using Correspondence Analysis is performed (Greenacre 1984; Benzécri 1992).

co-occurrence a varies from 0 to K at most, and $b=c= K - a$, thus, in this situation Sorense-Dice will be equal to a/K.

[2] Let's consider a set of K thematic maps represented by a set of N raster points, each map has m_k classes, if $M = \sum_{k=1}^{K} m_k$ is the total number of classes, each input vector being represented by $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots x_{iM}]^T$, $i = 1..N$, where $x_{ij} \in \{0,1\}$, and x_{ij} assumes 1 if the class is present, 0 otherwise.

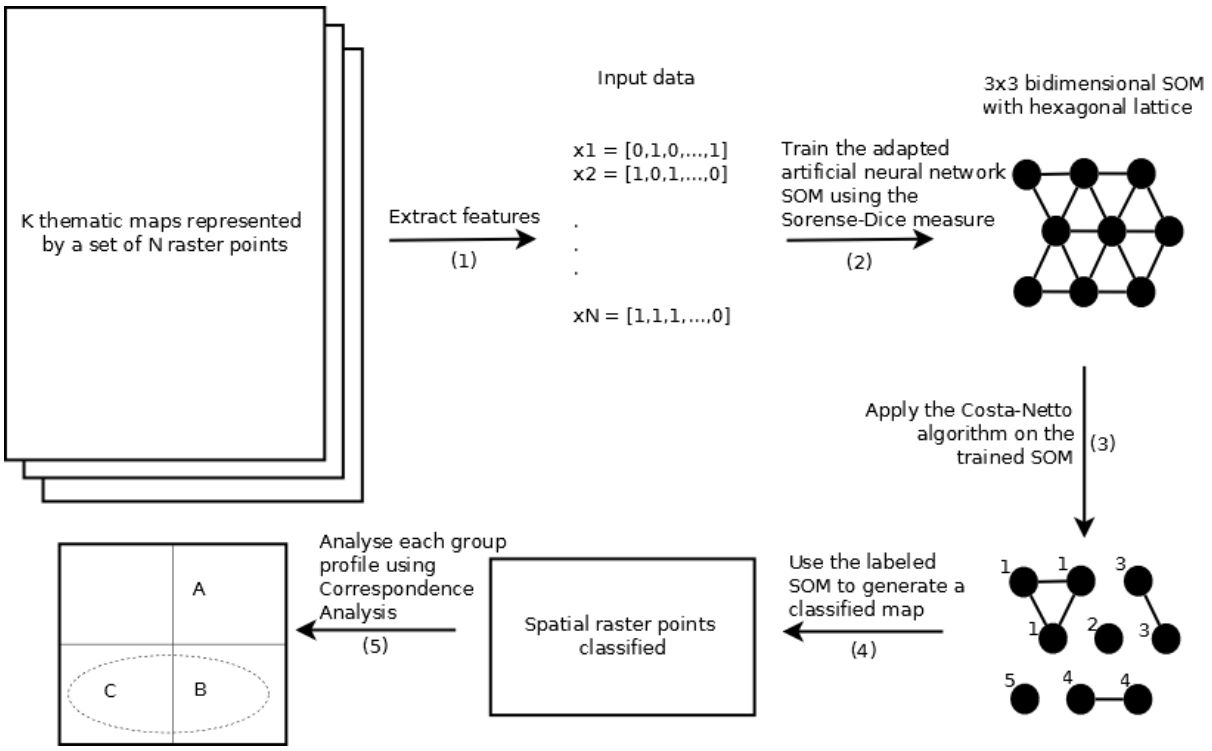


Figure 1. Scheme of the proposed method for investigation of territorial zoning using thematic maps.

3. The Case Study: Territorial Zoning of the Alto Taquari River Basin

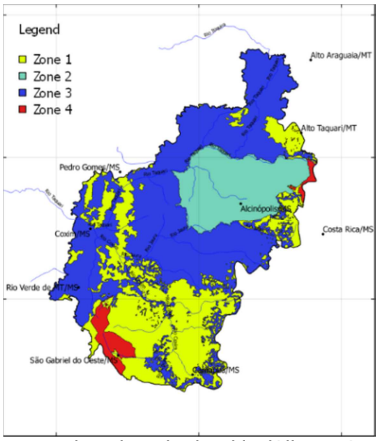
The Alto Taquari River Basin (BAT) is adjacent to the Pantanal, in Mato Grosso do Sul, Brazil. This area covers 28,046 km^2. It presents a demographic density of 2.53 inhabitants/km^2, and has extensive livestock farming and agriculture (grains) as the main economic activities (Schiavo et al. 2012). It has been used five thematic maps in the experiments (Table 1), covering five dimensions: environmental, infrastructure, economic aspects, population dynamics and living conditions of the population (Silva & Santos 2011).

Thematic map	**Classes**	**Labels**
Economic aspects	Low, medium and high quality of the economic aspect index	EA1, EA2 and EA3
Environmental dimension	Eight groups organized in the crescent order of homogeneity, from ED1 to ED8	ED1, ED2, ED3, ED4, ED5, ED6, ED7 and ED8
Infrastructure	Low, medium and high quality of the infrastructure index	IS1, IS2 and IS3
Living conditions	Low, medium and high living conditions index	LC1, LC2 and LC3
Population dynamics	Low, medium and high equilibria of the population dynamics	PD1, PD2 and PD3

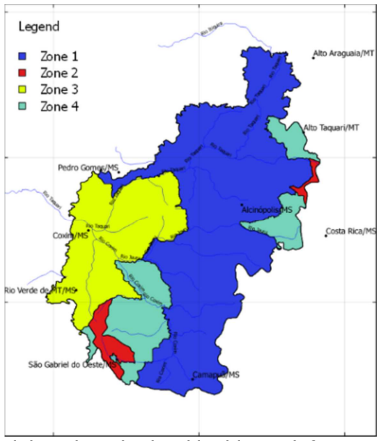
Table 1. Thematic maps of the Alto Taquari River Basin, MS, Brazil

4. Results and Discussion

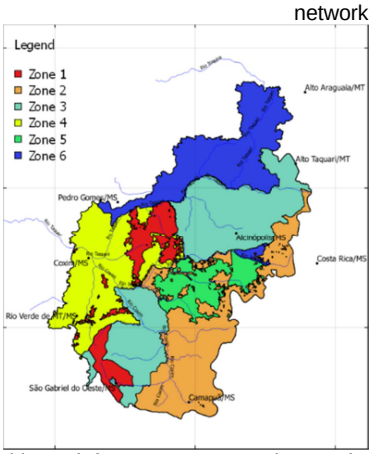
Figure 2(a) presents the result obtained in Silva & Santos (2011), Figures 2(b) and 2(c) show the results obtained using the method proposed in this study. For the SOM map of 1 × 15 with initial radius 3 has been obtained the same number of groups (zones) from Silva & Santos 2011). A different result had been observed in the 17 × 17 SOM map with initial radius 8, in this case, there has been a greater influence on the environmental dimension, which explains the lower similarity in relation to the result obtained in Silva & Santos 2011).



(a) Environmental zoning obtained in (Silva & Santos 2011)



(b) Territorial zoning obtained in this work for a 1×15 neural network map with neighborhood radius 3



(c) Territorial zoning obtained in this work for a 17 × 17 neural network map with neighborhood radius 8.

Figure 2: Comparison of results obtained in this work with the result achieved in (Silva & Santos 2011).

It had been applied the Correspondence Analysis over these three territorial partitions. Table 2 shows the mass, the chi-square distance to the data center and the inertia explained by each generated zone. Then, from the zones created by Silva & Santos (2011) only two explains almost all inertia (ZTVila2 and ZTVila3). The one-dimensional SOM distributed the inertia almost equally among the four zones, and the two-dimensional SOM well-distributed the inertia among five zones, only the ZTSOMbi5 represents a small fraction of the total inertia. It is worth to note that the cumulative inertia of the two main dimensions of the two-dimensional SOM explains only 71.0% of the total variation, while the Vila strategy and the one-dimensional SOM explained, each one, more than 84.0%. This suggests that the two-dimensional SOM well separated the data.

The two-dimensional SOM created zones with well distributed masses and more equidistant to the center of the data cloud. The territorial zoning elaborated by Silva & Santos (2011) generated one group (ZTVila1) which represents more than 74% of the total mass but are closer to the center and corresponds to a small fraction of the total inertia. The same occurs to the zone ZTSOMUni1 which concentrates a great portion of the mass, and are close to the barycenter of the data.

Zone/group	Mass	ChiDist	Inertia
ZTVila1	0.743479	0.199826	0.029688
ZTVila2	0.113763	1.010926	0.116262
ZTVila3	0.115641	1.291045	0.192750
ZTVila4	0.027118	1.919004	0.099863
ZTSOMUni1	0.581934	0.603867	0.212205
ZTSOMUni2	0.037206	2.469866	0.226968
ZTSOMUni3	0.226756	1.043839	0.247074
ZTSOMUni4	0.154103	1.231564	0.233736
ZTSOMBi1	0.084398	1.206881	0.122931
ZTSOMBi2	0.216021	0.880525	0.167487
ZTSOMBi3	0.275548	0.889580	0.218056
ZTSOMBi4	0.173119	1.157951	0.232127
ZTSOMBi5	0.073440	1.105970	0.089829
ZTSOMBi6	0.177473	1.078160	0.206300

Table 2: The mass, the chi-distance to the data center and the inertia calculated by the Correspondence Analysis method for each territorial zoning set of groups.

The graphical Correspondence Analysis (Figure 3) had been performed on the best data partition, the two-dimensional SOM territorial zoning. It has been considered the first two principal dimensions, which explains 70.95% of the total inertia. Analyzing the first axis, which explains 41.60% of the total variation and considering the four zones and the fifteen classes which most contributed to this axis, it is observed an opposition among the zones ZTSOMbi2/ZTSOMbi6 and ZTSOMbi3/ZTSOMbi4. The first two zones are associated with high and medium environmental homogeneity, high equilibria of population dynamics and medium quality of infrastructure index. The second group is associated with low and medium environmental homogeneity, low and medium population dynamics, low and high quality of infrastructure index and high living conditions index. The analysis of the second axis (Figure 3), which explains 29.35% of the total inertia, shows an opposition between the zones ZTSOMbi3 and ZTSOMbi4, the first is associated to medium equilibria of the population dynamics, low quality of the infrastructure index and low and high living conditions. The ZTSOMbi4 zone is associated with medium living conditions and medium environmental homogeneity.

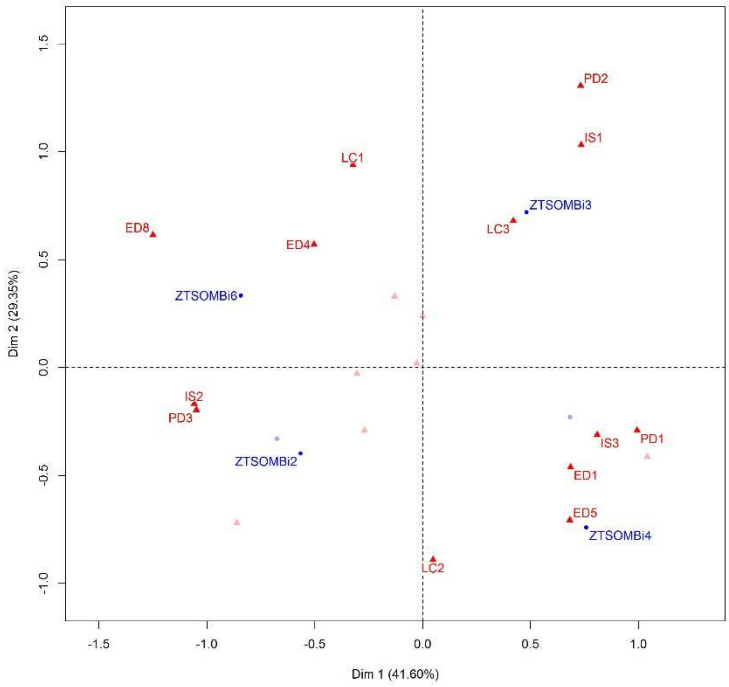


Figure 3: Correspondence analysis graph for the territorial zoning created by the two-dimensional SOM. It has been considered the zones and classes which most contributed to the two main axes.

5. Conclusions

The two-dimensional Self-Organizing map, associated with an automatic segmentation strategy (Costa-Netto algorithm), well separated the spatial binary patterns as showed the correspondence analysis of the group's profiles. The Sorense-Dice measure of binary similarity used in this context had been reduced to the calculation of the percentage of matched classes between two feature vectors (*a/K*). Future works may improve this calculation to take into account the differences among thematic maps with a different number of classes because if one thematic map has many classes the probability of two locations to present the same class is less than another thematic map with fewer classes. Another improvement should be the adaptation of the SOM to treat ordinary data.

6. References

Benzécri, J.P., 1992. *Correspondence analysis handbook*, New York: Dekker.

Costa, J.A.. & Netto, A.M.L., 2003. Segmentação do SOM baseada em particionamento de grafos. In *VI Congresso Brasileiro de Redes Neurais*. São Paulo: SBRN, pp. 451–456.

Greenacre, M.J., 1984. *Theory and applications of correspondence analysis*, London: Academic Press.

Henriques, R., Bação, F. & Lobo, V., 2012. Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, 36, pp.218–232.

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

Lourenço, F., Lobo, V. & Bação, F., 2004. Binary-based similarity measures for categorical data and their application in self-organizing maps. In *Jornadas de classificação e análise de dados*. Lisboa, p. 11.

Santos da Silva, M.A., 2004. SOMCode project. Available at: http://www.cpatc.embrapa.br/somcode/index.htm.

Santos da Silva, M.A. et al., 2010. Using self-organizing maps for rural territorial typology. In *Computational Methods for Agricultural Research: Advances and Applications*. pp. 107–126. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84900643299&partnerID=MN8TOARS.

Schiavo, J.A. et al., 2012. Characterization and classification of soils in the taquari river basin-pantanal region, state of mato grosso do sul, Brazil. *Revista Brasileira de Ciência do Solo*, 36(3), pp.697–708.

Silva, J. d. S.V. d. & Santos, R.F. d., 2011. *Estratégia metodológica para zoneamento ambiental: a experiência aplicada na Bacia Hidrográfica do Alto Rio Taquari*, Campinas: Embrapa Informática Agropecuária.