Modern Environmental Science and Engineering (ISSN 2333-2581)
November 2018, Volume 4, No. 11, pp. 1110-1116
Doi: 10.15341/mese(2333-2581)/11.04.2018/013
Academic Star Publishing Company, 2018
www.academicstar.us

Proposition and Evaluation of a Territorial Homogeneity Index

Marcos Aurélio Santos da Silva[1], Joel A. de Oliveira[2], Leonardo N. Matos[2], and Márcia H. G. Dompieri[3]

1. Embrapa Tabuleiros Costeiros, Brazilian Agricultural Research Corporation, Brazil

2. Federal University of Sergipe, Department of Computer Science, Brazil

3. Embrapa Territorial, Brazilian Agricultural Research Corporation, Brazil

Abstract: The stratification of municipalities in homogeneous areas facilitates the planning and management of space, especially in large areas, since it allows greater coherence in the adoption of public policies and economic incentives. This paper proposes a homogeneity index to evaluate the degree of similarity of territories previously determined by regional public policies. The index works in conjunction with clustering techniques such as TerraSOM and k-means. The proposed index has been applied in the evaluation of 27 Territories of Identity, state of Bahia, created in 2007 by the Superintendence of Economic and Social Studies of Bahia. The data were composed of a set of 45 socioeconomic and land use variables. The analytical evaluation of the proposed index showed that it was simple to implement and interpret, can be applied with any clustering technique, and can be used in different contexts. The results has shown that to the Territories of Identity results have shown similar results for both clustering techniques, but in the presence of outliers, the homogeneity index was less stable when calculated by the k-means technique. There is not statistically significant spatial autocorrelation for the homogeneity index.

Key words: spatial analysis, spatial dependence, k-means, TerraSOM, regional planning

1. Introduction

The adoption of development policies based on the territorial approach has been the main governmental strategy in order to promote the economic development, decrease social poverty and promote environmental conservation [1, 2]. Noteworthy in this scenario, are the national program for sustainable development of rural areas (Pronat) and the citizenship territorial program (PTC) of the Minister of Agrarian Development, the planning territories from the state of Sergipe and the identity territories from the state of Bahia. On these cases, each territory corresponds to a group of neighboring counties which share similar socioeconomic characteristics according to an objective criterion, as the level of the economic activity, or subjective, as the degree of belonging or identity.

Although the homogeneity is a central concept of policies with a territorial approach in Brazil, yet it has not been defined a method to do it in already defined territories. In general, the use of homogeneity measures is applied before the definition of spatial agglomeration, such as, in processes of agroecological, environmental or climatic risks zoning [3]. In case of already defined territories, it's necessary to check if the supposed homogeneity persists, so that it is possible to monitor and proceed to a realignment of the territorial development policies, when necessary. One strategy would be to choose a set of multidimensional random variables and then evaluate, to each territory, the level of variation within and between territories. However, this strategy assumes that the territories had been

Corresponding author: Marcos Aurélio Santos da Silva, Ph.D. Candidate in Artificial Intelligence, MSc. in Applied Computing; research areas/interests: geocomputation and computational social science. E-mail: marcos.santos-silva@embrapa.br.

corrected defined. Another strategy is to cluster the multivariate dataset for the spatial units and after verifying if there is a spatial dependence inside each territory.

On this work this second strategy had been taken into account, and a territorial homogeneity index (THI) has been defined. The THI is based on the level of aggregation among the spatial units (counties, in the case of public policies mentioned above), and can be determined by using any clustering algorithm. In this work we evaluated the THI using two clustering techniques, TerraSOM and k-means. It's been decided to evaluate the implementation of TerraSOM due to its positive use on spatial analysis [4-6], and the k-means method because it is widely used. The THI has been evaluated in the comparison of the 26 identity territories of the state of Bahia, created in 2007 by the Superintendence of Economic and Social Studies (SEI), attached to the Secretariat of Planning of the State of Bahia (SEPLAN).

2. Material and Methods

2.1 Data

The dataset consists of 45 variables about socioeconomic and land use which present aspects related to indicators of the Atlas of Human Development from PNUD, of "Bolsa Família" Program and agricultural census from IBGE, divided into seven groups: Atlas of Human Development, "Bolsa Família" Program, producer condition, land use, number of cattle, number of goats and number of sheep. It has been analyzed two datasets: the first with all counties and identity territories, which means, 417 counties and 27 territories; and the second without the outliers, which means, observations with values, to certain variables, five standard deviations from the average. To the second set of data, it's been analyzed 375 counties and 26 territories, the Metropolitan Territory of Salvador was excluded because it contained too many outliers.

2.2 Self-Organizing Maps and the TerraSOM Algorithm

The Self-Organizing Map (SOM) is a type of artificial neural network with non-supervised learning which presents the property of neatly mapping the data in the grid of artificial neurons. This mapping preserves the statistical properties of the input data and allows this grid to be exploited for data visualization and clustering. The learning process has three steps: the competitive, when each input vector is associated to a neuron, the BMU (Best Matching Unit); the cooperative, when the neighborhood between the neurons is defined; and the adaptive, when the weight vectors are updated in the direction of the input vector in the inverse intensity of the distance between the neuron and the BMU of each input vector [7, 8].

The TerraSOM algorithm cluster the geospatial data from SOM in three steps [6]. In the first, the neural network is trained using a geospatial dataset. In the second, the trained neural network is automatically partitioned into groups using the Costa-Netto algorithm [9]. As each county will be associated with a single artificial neuron, it is known that the neural network partition determines the dataset partition. In the third step the partitioning is evaluated by two clustering validity indexes, the Davies-Bouldin which evaluates the distance between the centroids of each group [10] and the Composed Density between and within clusters (CDbw) which evaluates the intra and inter-group density through the reference vectors of the input data [11, 12]. The use of validation indexes allows you to choose the neural network parameterization that generates the best results.

The TerraSOM method tests have been conducted to 80 different neural networks architectures. All of them had a two-dimensional hexagonal grid, and they were trained with the sequential learning technique. However, they differed in the number of lines and columns, the initial radius of the update process of the code vectors and in the maximum number of interactions.

2.3 The Territorial Homogeneity Index (THI)

One way of assessing the homogeneity of a territory already constituted from a set of random variables is by applying a clustering algorithm (with or without restriction of spatial contiguity) on the dataset and to compare it with the previous defined territories (e.g., the Territories of Identity). If all spatial objects (e.g., counties), in a territory, belongs to the same cluster we can say that this territory is likely to be homogeneous. Otherwise, if all spatial objects are distributed in several clusters, it means that this territory may be heterogeneous. Therefore, different data cluster algorithms may generate dissimilar results.

To compare these results, it has been idealized a Territorial Homogeneity Index (THI) which takes into consideration: the number c of clusters generated by the clustering algorithm; the area $S_{\max}$, that represents the area covered by the dominant group in the territory i; and the number m_i of groups found in each territory. Then, by the partitioning a territory composed of n geographic objects (counties) in c groups by any clustering algorithm (in this paper, k-means and TerraSOM), it is defined the THI to each-territory T_i as follows:

Let $A = \{S_{i1}, S_{i2}, \ldots, S_{ij}\}$ the set of the summation of areas related to T_i associated to a set of $j = m_i$ groups, where $S_i = \sum_{j=1}^{m_i} S_{ij}$ is the summation of all elements of A and $S_{max,i}$ is the maximum value of this set. The THI for each T_i is given by:

$$THI_i = \frac{S_{max,i}}{S_i + \frac{(m_i - 1)}{c}(S_i - S_{max,i})},$$

The THI_i assumes the maximum value when $S_{\max,i}$ equals to S_i, and $m_i = 1$, and when the territory is associated to only one group. The THI assumes the minimum value when m_i equals c and $S_{\max,i}$ gets closer to S_i/m_i, thus, when all the groups are represented in the territory and the bigger group gets closer to S_i/m_i. Scaling THI_i to the range [0,1] we have the value THI'_i, given by:

$$THI'_i = \frac{x_i z - c}{z - c}, z = c^2 + (c-1)^2$$

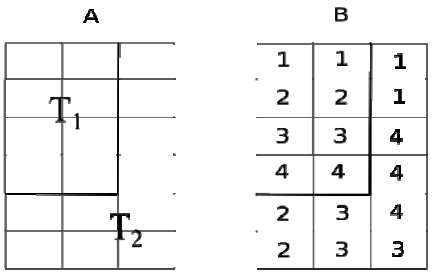


Fig. 1 Hypothetical geometric area with 18 unitary areas sub-areas divided in two (A) and partitioned in four groups (c = 4).

The Fig. 1 shows un hypothetical example to demonstrate the numerical calculation of the proposed index. Then, for the territory T_1, A = {2,2,2,2}, $m_1 = 4$, $S_{\max,1} = 2$, $S_i = 8$, $c = 4$, and the territorial homogeneity index will be equal to $THI_1 = 2/(8 + (4\text{-}1)*(8\text{-}2)/4) = 0.16$. Then, scaling THI_1, we have $z = 4^2+(4\text{-}1)^2$, $THI'_1 = (0{,}16*25\text{-}4)/25\text{-}4 = 0$, so the territory one has a very low THI.

As each algorithm generates more than one partition of data we have considered the average of THI'_i from all valid considered partitions as the homogeneity index. The spatial autocorrelation of the indicator was evaluated by the Moran's I index and c of Geary [13, 14].

It is observed that the proposed homogeneity index is independent of the algorithm used to data partitioning as well as the type of multivariate data set used for the analysis of homogeneity.

2.4 Software

The R language, version 3.2.1, was used with packages *maptools* version 0.8, *spdep* version 0.5 to the manipulation of geographic data and the calculation of the spatial autocorrelation index [15]. The maps had been made by the program *QGis* version 2.6.0.

The clustering analysis *k*-means has been done based on the existent implementation package *stat* of the R language. The execution of the TerraSOM method was implemented through a call to the library SOMCode, implemented in C++ language and packed as a R script.

3. Results and Discussion

3.1 Clustering Analysis

It is observed in the graphics A and B on the Fig. 2 that, as the number of clusters increase, the vectorial quantization error decreases in the two sets of data (with or without outliers). In fact, the number of groups, *c*, depends largely on the total number of neurons, *m*, so *c* and *m* are positively directly proportional. The graphics C and D on the Fig. 2 shows the partition validation indexes results for Davies-Bouldin (DB) and CDbw.

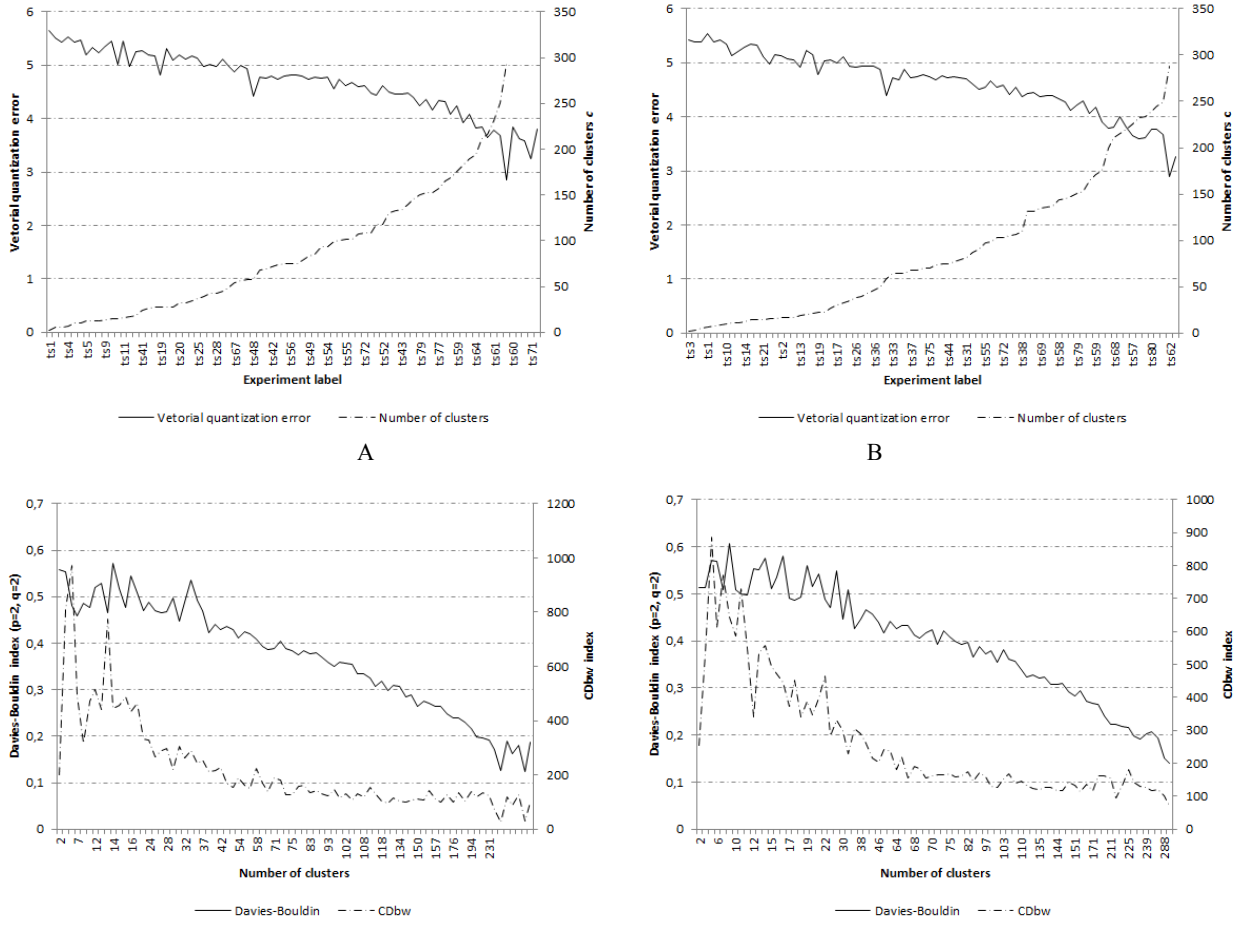


Fig. 2 Graphics of the vectorial quantization errors *Eq* and number of clustering *c* to experiments with (A) and without (B) outliers. Graphics of the clustering validity indexes (Davies-Bouldin and CDbw) results to experiments with (C) and without (D) outliers.

The Davies-Bouldin index identified as the best experiments those that returned a cluster number above 64 for the complete data set, and above 45 for the data set without no outliers, in that interval the value of the Davies-Bouldin index remained below 0.2. On the other hand, the CDbw index indicated the best experiments (indexes with high values) those with the lowest number of clusters. Therefore, for the subsequent analysis, the networks that presented intermediate values for the Davies-Bouldin and CDbw indexes were considered in the evaluation of the territorial homogeneity index.

For the test with the *k*-means method, 22 clusters were evaluated, with *k* varying between 10 and 31.

It could be observed by means of the Fig. 3 that both techniques had similar results for the Territorial Homogeneity Indexes when considering the two datasets, however with greater variation for the indexes

calculated for the *k*-means clustering. Although not perfectly matching, the *k*-means and the TerraSOM methods classified the Territories fairly well.

3.2 The calculated THI

Figs. 4 and 5 show the spatial distribution of homogeneity indices for Identity Territories with the two data sets, complete data set (map A) and no atypical data (map B). Choropleth maps show the indexes through the clear breaks, it is confirmed by them that there is greater variation of the indices for the partitioning of the data by the *k*-means method and better separation of the data for both cases when the atypical data is eliminated. The calculation of the indices *I* of Moran and *c* of Geary for the homogeneity indices did not confirm spatial autocorrelation.

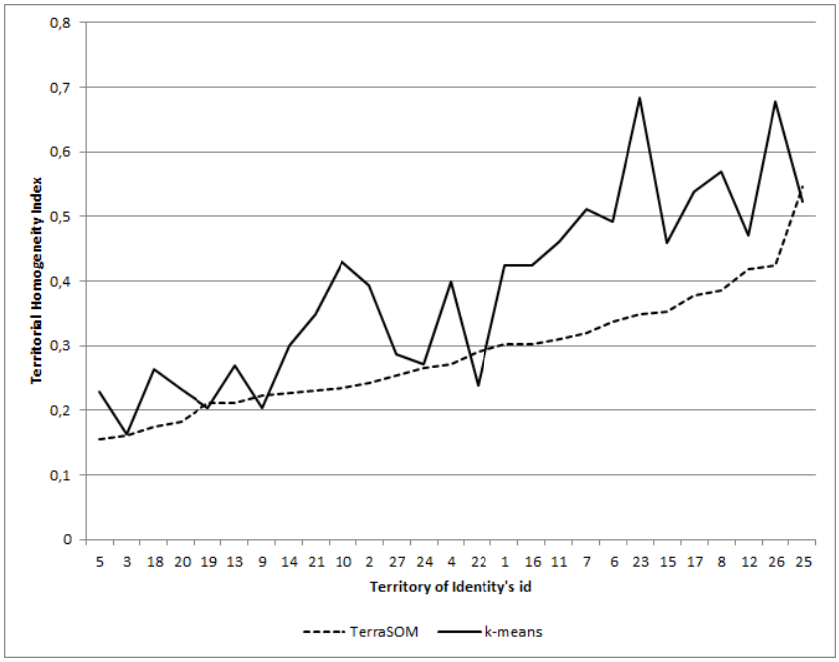
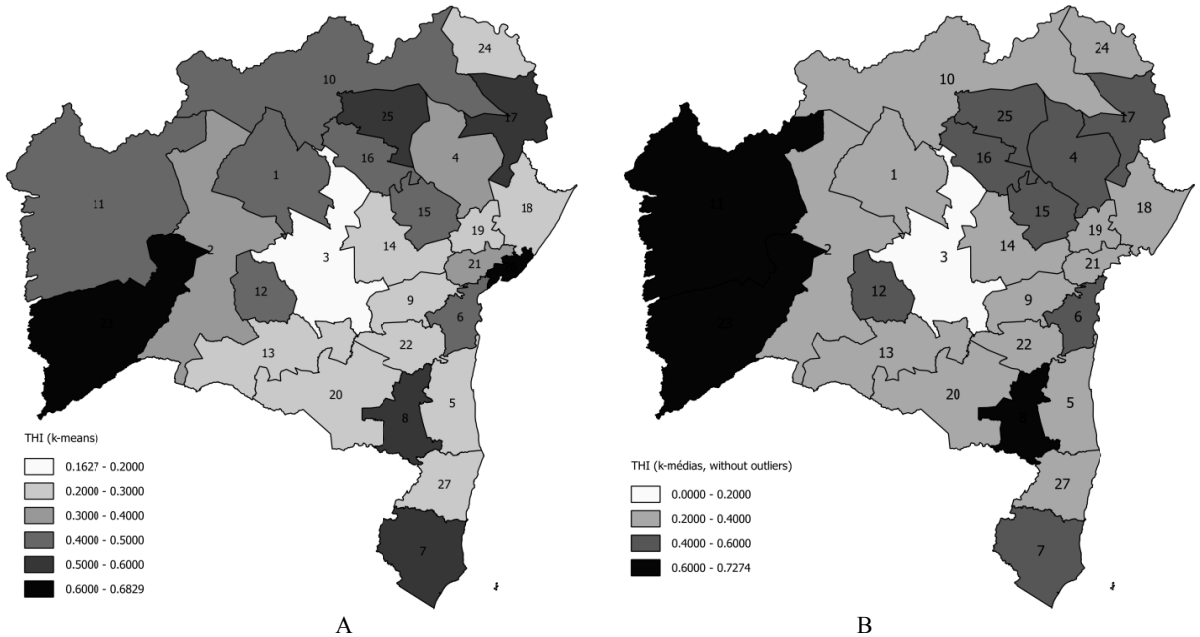
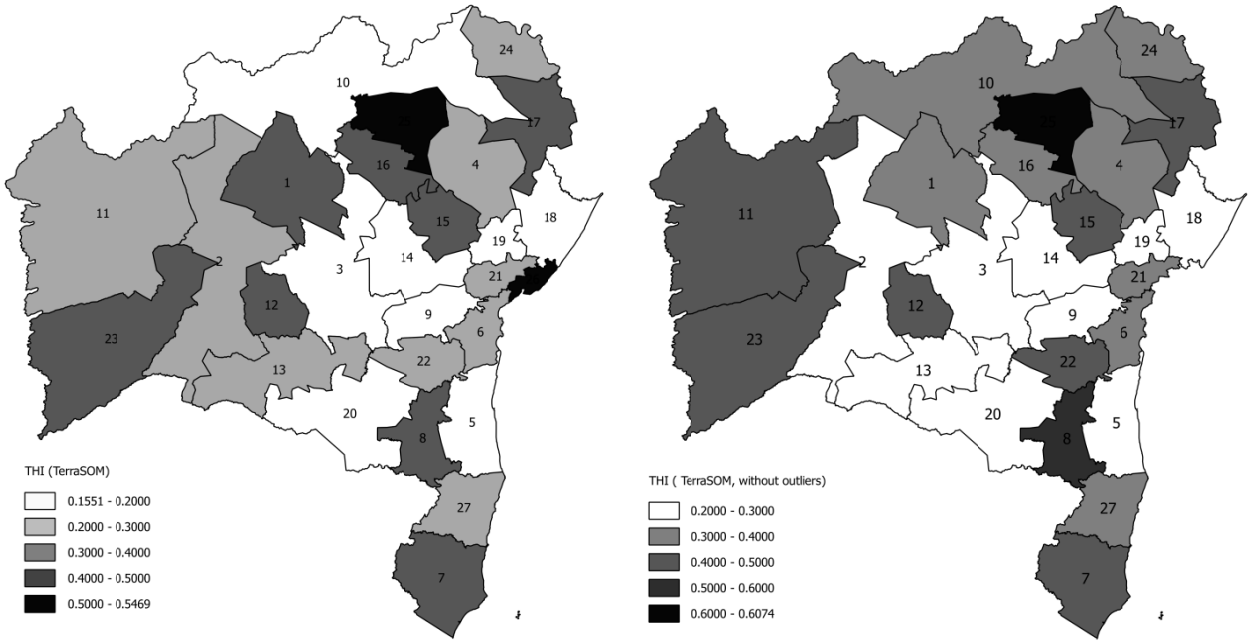


Fig. 3 The average curves of the *THI'* by the Territory of Identity using TerraSOM and k-means clustering algorithms.



A B

Fig. 4 Choropleth maps to the *THI*' calculated from k-means clustering algorithm.



A B

Fig. 5 Choropleth maps to the *THI*' calculated from TerraSOM clustering algorithm.

3.3 Analysis of the THI

The proposed index is simple, it can be calculated from any cluster algorithm, it is applicable in several contexts where it is desired to determine the degree of homogeneity in zones or territories already delimited and it is easy to interpret. However, the correspondence between the calculated and the effective homogeneity should be verified by means of fieldwork to support research projects or design public actions. Another relevant factor is the choice of the variables used in cluster analysis. They may or may not correspond to the criteria used for the creation of territories. In fact, the researcher may want to verify if a territory created from certain factors would also be homogeneous from others.

4. Conclusion

The *THI* for the Bahia's Territories of Identity was more sensitive to the presence of outliers data when calculated using the k-means method when compared to the TerraSOM method.

No statistically significant spatial autocorrelation was found for the *THI* calculated from both the k-means and the TerraSOM algorithms.

The proposed *THI* is easy to apply and interpret, and complementary studies are needed both to analyze its response to other cluster analysis algorithms (e.g., considering spatial neighborhood issues such as SKATER) and to verify its utility in real world problems of homogeneity analysis in pre-defined partitions.

References

[1] S. P. Leite and V. J. Wesz Júnior, Um estudo sobre o financiamento da política de desenvolvimento territorial no meio rural brasileiro, *Rev. Econ. e Sociol. Rural* 50 (2012) (4) 645-666.

[2] R. Boueri and M. A. Costa, *Brasil em desenvolvimento 2013*, Vol. 1, Brasília, DF: IPEA, 2013.

[3] J. D. S. V. D. Silva and R. F. D. Santos, Estratégia metodológica para zoneamento ambiental: a experiência aplicada na Bacia Hidrográfica do Alto Rio Taquari. Campinas: Embrapa Informática Agropecuária, 2011.

[4] M. A. S. da Silva, E. R. de Siqueira, O. A. Teixeira, M. G. L. Manos and A. M. V Monteiro, Using self-organizing maps for rural territorial typology, in: *Computational Methods for Agricultural Research: Advances and Applications*, 2010, pp. 107-126.

[5] M. A. S. da Silva, E. R. de Siqueira and O. A. Teixeira, Abordagem conexionista para análise espacial exploratória de dados socioeconômicos de territórios rurais, *Rev. Econ. e Sociol. Rura* 48 (Dec. 2010) (2) 429-446.

[6] M. A. Santos da Silva, R. J. S. Maciel, L. N. Matos and M. H. Galina, *TerraSOM: Sistema para Análise de Dados Geoespaciais Agregados por Área Baseado na Rede Neural do Tipo Mapa Auto-Organizável de Kohonen*, Embrapa Tabuleiros Costeiros, Aracaju, 2015, p. 38,

[7] T. Kohonen, *Self-Organizing Maps* (3rd ed.), Berlin: Springer, 2001.

[8] T. Kohonen, Essentials of the self-organizing map, *Neural Networks*37 (2013) 52-65.

[9] J. A. Costa and A. M. L. Netto, Segmentação do SOM baseada em particionamento de grafos, in: *VI Congresso Brasileiro de Redes Neurais*, 2003, pp. 451-456.

[10] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. IntellAnal. Mach. Intell* PAMI-1 (1979) (2) 224-227.

[11] S. Wu and T. W. Chow, Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density, *Pattern Recognit.* 37 (2004) (2) 175-188.

[12] M. Halkidi and M. Vazirgiannis, A density-based cluster validity approach using multi-representatives, *Pattern Recognit. Lett.* 29 (2008) 773-786.

[13] A. D. Cliff and J. K. Ord, *Spatial Autocorrelation*, London: Pion, 1973.

[14] A. C. Bailey and T. C. Gatrell, *Interactive Spatial Data Analysis*, Essex: Longman, 1995.

[15] R. S. Bivand, E. J. Pebesma and V. Gómez-Rubio, *Applied Spatial Data Analysis with R*, New York: Springer-Verlag, 2008.