SBI Agro 2023

Automatic Segmentation of the Self-organizing Map to Support Territorial Zoning

Pedro V. de A. Barreto[1,2], Marcos A. S. da Silva[1], Leonardo N. Matos[2], Gastão F. Miranda Júnior[3], Márcia H. G. Dompieri[4], Fábio R. de Moura[5], Fabrícia K. S. Resende[2]

[1]Embrapa Coastal Tablelands, 49025-040, Aracaju, SE, Brazil

[2]Dept. of Computer Science, Federal University of Sergipe (UFS), São Cristóvão, SE, Brazil

[3]Dept. of Mathematics, UFS, São Cristóvão, SE, Brazil

[4]Embrapa Territorial, 13070-115, Campinas, SP, Brazil

[5]Dept. of Economics, UFS, São Cristóvão, SE, Brazil

pedro.araujo@dcomp.ufs.br,{marcos.santos-silva,marcia.dompieri}@embrapa.br
lnmatos@ufs.br, gastao@mat.ufs.br, fabiromoura@gmail.com
fabricia.resende@outlook.com

Abstract. *This paper proposes an algorithm for analyzing clusters of thematic maps with ordinal categorical classes to support territorial zoning. The proposed method combines the Self-Organizing Map with graph segmentation techniques for data clustering. The approach was evaluated with synthetic data and applied to the environmental zoning of the Alto Taquari basin, MS/MT. The results showed the ability of the algorithm to separate the data into unimodal differentiable groups.*

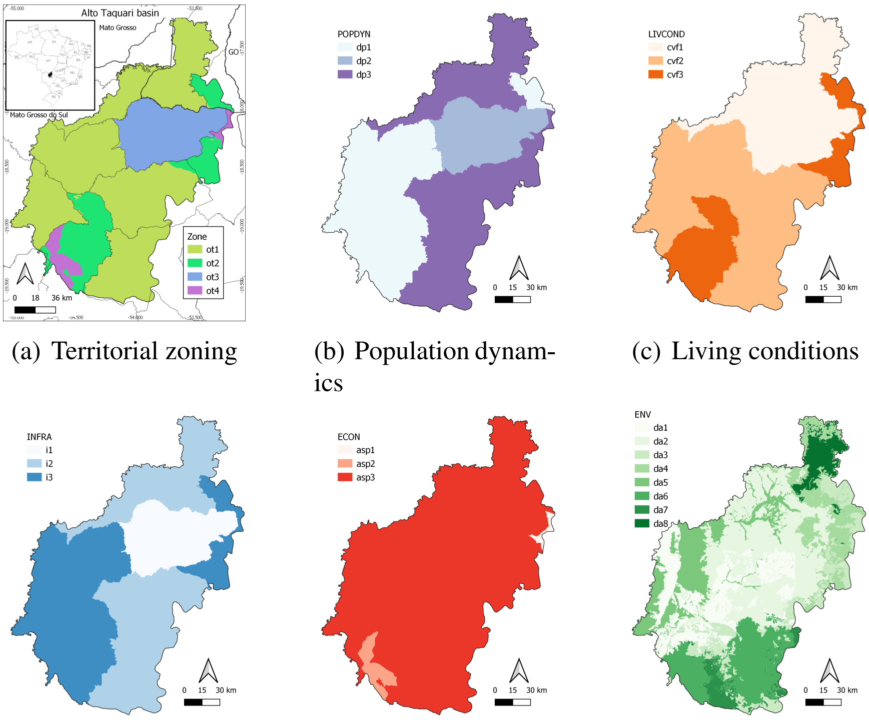
Resumo. *Este artigo propõe um algoritmo para análise de agrupamentos de mapas temáticos com classes categóricas ordinais para suporte ao zoneamento territorial. O método proposto combina o Mapa Auto-Organizável com técnicas de segmentação de grafos para clusterização dos dados. A abordagem foi avaliada com dados sintéticos e aplicada no zoneamento ambiental da bacia do Alto Taquari, MS/MT. Os resultados mostraram a capacidade do algoritmo separar os dados em grupos diferenciáveis unimodais.*

1. Introduction

Public policies concerned with territorial development increasingly use computational resources to achieve greater accuracy and efficiency. Among the reasons for this is a greater complexity deriving from factors such as climate, political choices, landscape, migration, etc. One crucial step in developing an appropriate intervention policy is identifying homogeneous zones or zoning through which similar areas can be targeted with different intervention regimes.

Artificial intelligence techniques offer new ways of automatically getting insight from data. In the artificial neural network class of models, the Self-Organizing Map (SOM) excels at preserving statistical properties from the dataset and allows visualization of high-dimensional input patterns. SOM's recent use in territorial zoning applications is increasing, such as in [Silva et al. 2018] and [Sadeck et al. 2022].

In this paper, we investigate a new approach for the clustering problem, proposing an algorithm that extensively uses the SOM properties and performs automatic segmentation upon the trained map. To test this approach's usefulness in support of territorial zoning application, we study its performance on the *Alto Taquari* basin zoning based on thematic maps, along with other standard datasets used for clustering benchmarks.



(a) Territorial zoning

(b) Population dynamics

(c) Living conditions

(d) Infrastructure

(e) Economic aspects

(f) Environment

Figure 1. Territorial zoning of the *Alto Taquari* basin (a) obtained by [Silva and Santos 2011] by applying a hierarchical agglomerative clustering over five ordinal data transformed into a binary: the population dynamics (b), the living conditions (c), the infrastructure (d), the economic aspects (e) and the environmental dimension (f). Source: elaborated by the authors.

2. Material and Methods

2.1. The Alto Taquari basin - MS/MT, Brazil

The Alto Taquari basin is located in northeastern MS, Brazil. [Silva and Santos 2011] elaborated an territorial zoning of this region to subsidize public policies for the region

(Fig. 1(a)). It was elaborated from intermediate maps on the environment (Fig. 1(f)), economic aspects (Fig. 1(e)), infrastructure (Fig. 1(d)), living conditions (Fig. 1(c)) and population dynamics (Fig. 1(b)). Each of these maps represent ordinal categorical classes (e.g., asp1, asp2, asp3 for Economic aspects map). We transformed the ordinal data into numerical sequences (e.g., $1,\ 2,\ and\ 3$ for the ECON map) to present the data to the SOM. After, we divided these values by the highest value of all maps (eight from the ENV map). Then, the input vector $\mathbf{x}$ has five components varying their values between $1/8$ and $8/8$.

2.1.1. Proposed method based on Self-Organizing Map

A standard Self-Organizing Map [Kohonen 2001] is an artificial neural network where the neurons are organized in a $N \times M$ hexagonal or rectangular grid, where there is a weight vector associated with each one. The iterative learning mechanism consists of approximating these weights in the direction of the input data. For a fixed number of iterations and all input vectors, we randomly pick an input vector and search for the closer neuron using the Euclidean distance, also known as the Best-Matching Unit (BMU). After, we define the neighborhood of its BMU. Then, we update the BMU and the neighboring neurons' weights according to Eq. 1.

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(t)h(t)(\mathbf{x}_i – \mathbf{w}(t)) \tag{1}$$

where t represents the iteration, $\mathbf{w}(t)$ is the neuron weight vector in the iteration t, $\alpha(t)$ is a small value representing the learning rate, $h(t)$ is a neighborhood function, and $\mathbf{x}_i$ an input data vector taken randomly.

At the end of the iterations, each input data will be associated with a single neuron, which can represent more than one input vector. SOM weights preserve the data's topology, meaning that neighboring neurons can represent nearby input vectors. This feature of SOM allows the use of clustering algorithms (e.g., k-means) on the neural network's weights as an indirect way to partition the input data [Silva et al. 2022].

Algorithm 1 Proposed SOM segmentation based on graph partition

```
Require: $G = (V, E)$ — Graph of the trained SOM
Require: $H$ — Neurons' activity level data
Require: $D$ — Distance matrix between weights
Require: $k$ — The number of desired clusters
$T \leftarrow$ minimum spanning tree of $G$ using $D$ as edges' weights
for each edge $(u, v) \in T$ do
    $cost(u, v) \leftarrow DBI(u, v)$
end for
Prune the $k - 1$ edges in $T$ with lesser costs
Assign a cluster label to each set of connected nodes in $T$
```

However, it is possible to segment the SOM without the aid of traditional clustering algorithms using neural network internal information such as distance and neighborhood between the SOM weights, the number of input vectors associated with it, and data

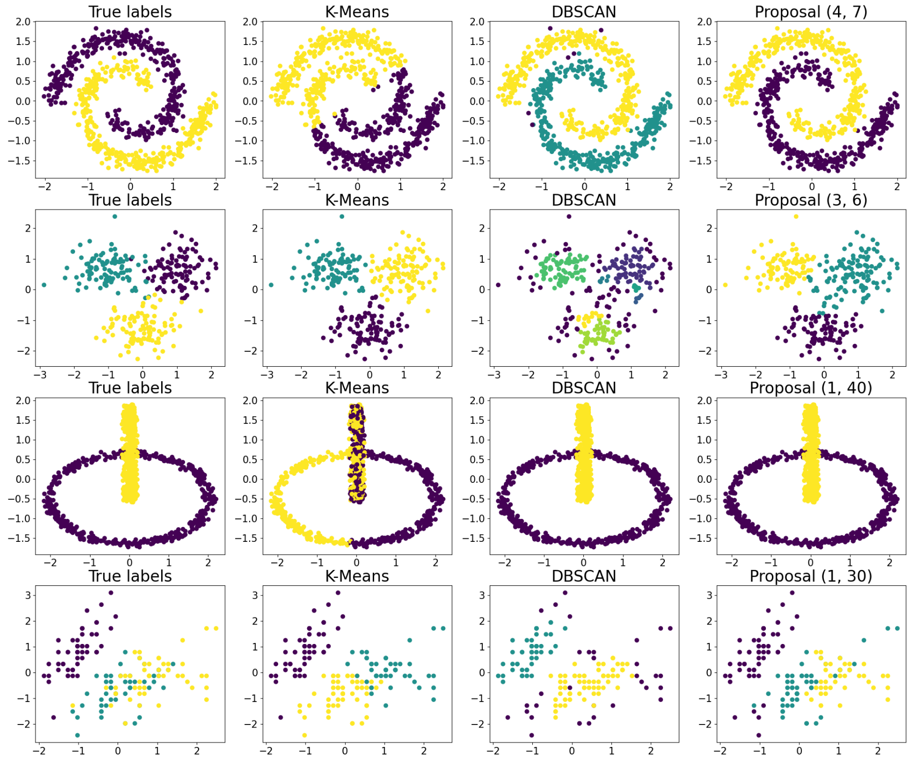
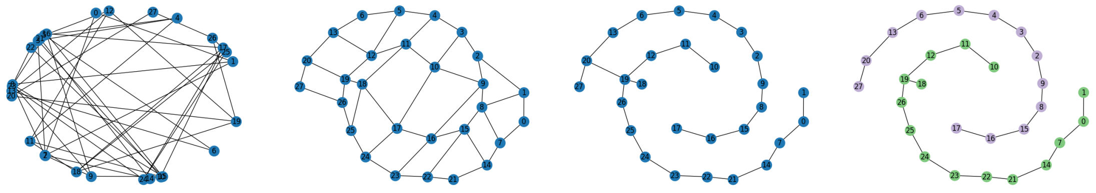


Figure 2. In the first column on the left, we have the four labeled datasets (spiral, gaussian, chainlink, and iris data sets) used to evaluate the proposed method (last column), compared with the k-means methods in the second column and DBSCAN in the third column. Source: elaborated by the authors.

density between neurons. [Costa and Netto 2003] proposed a graph-based SOM partitioning model that uses all this information and automatically determines the number of clusters, which has been successfully applied in some case studies [Silva et al. 2018]. This proposal has as a limiting factor, there are three hyperparameters that must be adjusted to each data set. [Silva and Costa 2011] also proposed a graph-based method for segmenting the SOM, but using only the density between neurons by applying the Davies-Bouldin Validation Index (DBI). In this case, the algorithm automatically detects the number of clusters, we have a single hyperparameter, but we still do not have applications in real situations and do not use all the information available from SOM.

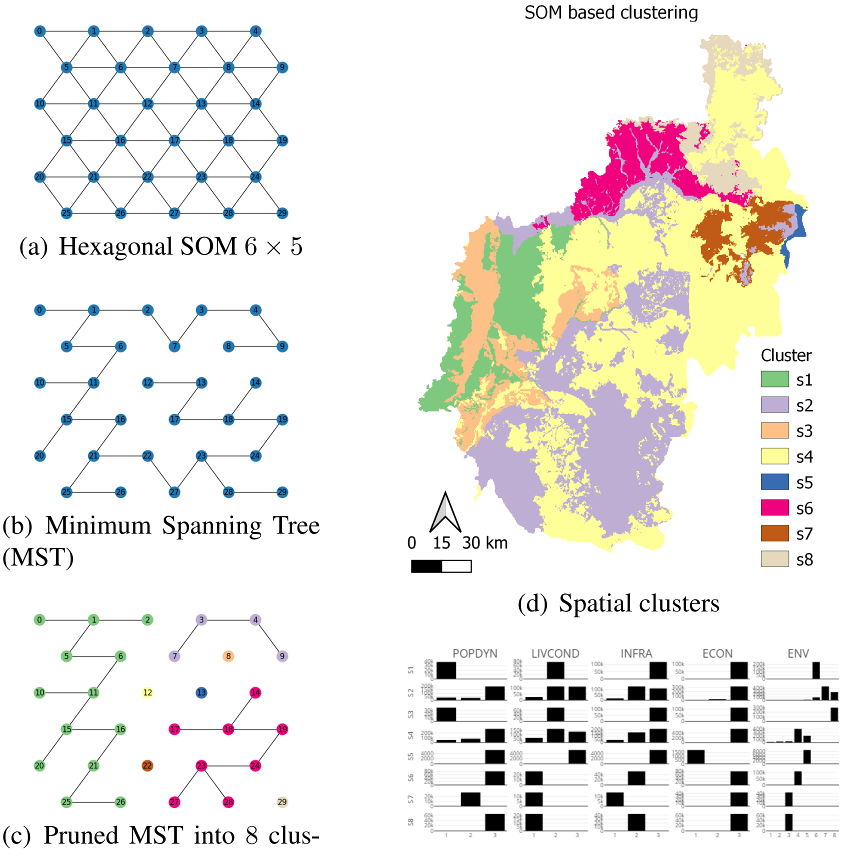
We propose a segmentation algorithm based on the interpretation of the SOM as an undirected graph, which uses all the information available after the machine learning process without the need for hyperparameter adjustment. It is only necessary to define the desired number k of clusters (Algorithm 1). We implemented it in Python version 3.8 using the Minisom version 2.3.0 as the SOM implementation [Vettigli 2018].

Fig. 2 shows the results of clustering some artificially and benchmark labeled data (spiral, gaussian, chainlink, and iris) using the k-means, DBSCAN, and the proposed method. We observed that the proposed method performs well for all four data sets. Different hyperparameters (number of neurons and grid lattice) were evaluated for the ANN SOM, using the accuracy measure ACC to choose the final configuration. The proposed method has a higher computational cost when compared to the k-means and



(a) Untrained SOM (b) Trained SOM (c) MST (d) Pruned MST

Figure 3. Application of the proposed algorithm on the spiral data using a 4×7 SOM and representation of its weights and their neighborhood. Untrained SOM (a), the trained SOM (b), the Minimum Spanning Tree (MST) using the distance between neurons as weights (c), pruning of the MST considering the lesser DBI between neighbor neurons (d). Source: elaborated by the authors.



(a) Hexagonal SOM 6 × 5

(b) Minimum Spanning Tree (MST)

(c) Pruned MST into 8 clusters

(d) Spatial clusters

(e) Histogram of each variable per cluster

Figure 4. (a) Trained hexagonal SOM 6×5. (b) Minimum sapanning Tree generated by the distance between neurons' weights. (c) Pruned MST into 8 clusters considering the data density between neurons using the Davies-Bouldin index. (d) Mapping the clusters into the *Alto Taquari* basin geographical map. (e) Histogram for the SOM's segmentation clustering proposed method over the five ordinal categorical variables (POPDYN, LIVCOND, INFRA, ECON, and ENV). Source: elaborated by the authors.

DBSCAN methods. Still, it manages to be efficient in situations more appropriate for algorithms based on data partitioning, such as k-means (iris and gaussian data sets), and density-based algorithms like DBSCAN (spiral and chainlink data sets). Fig. 3 shows the segmentation process of a 4×7 SOM trained upon the spiral-shaped dataset.

3. Results and Discussion

We evaluated several configurations (topology, size) of the SOM network for different values of k and used the Davies-Bouldin Index (DBI) to define the best combination. The DBI indicated a 6×5 hexagonal SOM ANN with sigma equal to 1.0 and partitioned into six clusters. But, after analyzing this 6×5 hexagonal ANN SOM for other values of k, we decided to analyze this same network partitioned into eight groups (Fig. 4(c)), resulting in the map in Fig. 4(d).

The histograms in Fig 4(e) indicated a good distinction between the cluster, and it means unimodal distributions, with other ordinal classes decaying around it.

4. Conclusions

The proposed method showed promising results, with a fair computational cost while maintaining a simple-to-understand approach that led to good group distinction. For future work, we suggest a more extensive testing of the proposed method, aiming at evaluating its performance with high-dimensional complex data. Also interesting would be to perform a deeper comparison with other SOM-based algorithms. And to investigate how larger networks behave with real data since many of this work's neural maps were small.

Acknowledgments

This work was supported by the National Council for Scientific and Technological Development – CNPq, Brazil, by the Institutional Scientific Initiation Scholarship Program.

References

Costa, J. A. F. and Netto, M. L. A. (2003). Segmentação do SOM baseada em particionamento de grafos. In *VI Congresso Brasileiro de Redes Neurais*, pages 451–456.

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

Sadeck, L. W. R., de Lima, A. M. M., and Adami, M. (2022). Artificial neural network for ecological-economic zoning as a tool for spatial planning. *Pesquisa Agropecuária Brasileira*, 52(11):1050–1062.

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

Silva, L. A. and Costa, J. A. F. (2011). A graph partitioning approach to SOM clustering. In *12th Intl Conf. on Intelligent Data Engineering and Automated Learning - IDEAL*.

Silva, M. A. S. d., Maciel, R. J. S., Matos, L. N., and Dompieri, M. H. G. (2018). Automatic environmental zoning with Self-Organizing Maps. *MESE*, 4(9):872–881.

Silva, M. A. S. d., Matos, L. N., Santos, F. E. d. O., Dompieri, M. H. G., and Moura, F. R. d. (2022). Tracking the connection between Brazilian agricultural diversity and native vegetation change by a machine learning approach. *IEEE Lat Am T*, 20(11):2371–2380.

Vettigli, G. (2018). Minisom: minimalistic and NumPy-based implementation of the Self Organizing Map. Available at: https://github.com/JustGlowing/minisom/.