A Self-Organizing Map Clustering Approach to Support Territorial Zoning[*]

Marcos A. S. da Silva[1][0000−0002−5367−2869], Pedro V. de A. Barreto[1,2][0009−0007−8816−1325], Leonardo N. Matos[2][0000−0002−6302−3299], Gastão F. Miranda Júnior[3][0000−0002−0967−6141], Márcia H. G. Dompieri[4][0000−0001−7689−1602], Fábio R. de Moura[5][0000−0002−6532−110X], Fabrícia K. S. Resende[2][0000−0001−8010−6304], Paulo Novais[6][0000−0002−3549−0754], and Pedro Oliveira[7][0000−0001−7143−5413]

[1] Embrapa Coastal Tablelands, 49025-370, Aracaju, SE, Brazil
marcos.santos-silva@embrapa.br

[2] Dept. of Computer Science, Federal University of Sergipe, São Cristóvão, SE, Brazil
{pedro.araujo,leonardo}@dcomp.ufs.br, fabricia_resende@academico.ufs.br

[3] Dept. of Mathematics, Federal University of Sergipe, São Cristóvão, SE, Brazil
gastao@mat.ufs.br

[4] Embrapa Territorial, 13070-115, Campinas, SP, Brazil
marcia.dompieri@embrapa.br

[5] Dept. of Economics, Federal University of Sergipe, São Cristóvão, SE, Brazil
fabiromoura@gmail.com

[6] Department of Computing, Minho University, Braga, Portugal
pjon@di.uminho.pt

[7] ALGORITMI Centre/LASI, Minho University, Braga, Portugal
pedro.jose.oliveira@algoritmi.uminho.pt

Abstract. This work aims to evaluate three strategies for analyzing clusters of ordinal categorical data (thematic maps) to support the territorial zoning of the Alto Taquari basin, MS/MT. We evaluated a model-based method, another based on the segmentation of the multi-way contingency table, and the last one based on the transformation of ordinal data into intervals and subsequent analysis of clusters from a proposed method of segmentation of the Self-Organizing Map after the neural network training process. The results showed the adequacy of the methods based on the Self-Organizen Map and the segmentation of the contingency table, as these techniques generated unimodal clusters with distinguishable groups.

Keywords: Alto Taquari basin · Ordinal categorical data · Spatial Analysis.

[*] Supported by National Council for Scientific and Technological Development – CNPq, Brazil, and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project 2022.06822.PTDC. The work of Pedro Oliveira was also supported by the doctoral Grant PRT/BD/154311/2022 financed by the Portuguese Foundation for Science and Technology (FCT), and with funds from European Union, under MIT Portugal Program.

1 Introduction

The elaboration of public policies aimed at territorial development is becoming increasingly complex due to the various factors we must consider and mutually influence each other, such as climate, political choices, landscape, economic systems, migration, etc. [6]. One of the first steps in elaborating regional intervention policies is the analysis of homogeneous zones or zoning, which aims to identify areas with similar characteristics and, consequently, which should be submitted to different intervention regimes [16].

Territorial zoning can have different purposes (ecological, economic, risk prevention) and is conducted with the support of some spatial data management and analysis systems. Given the massive availability of spatial data, the complexity of the zoning process, and the urgency that specific applications demand, unsupervised zoning can play an essential role in elaborating territorial public policies [16].

Computer-assisted territorial zoning is elaborated from several superimposed layers of information. We consider one region a homogeneous zone if it presents similar characteristics for all the layers. We can define this homogeneity by choosing a similarity criterion (e.g., Euclidean distance) and a process for determining homogeneous areas (e.g., clustering algorithm).

This work investigates different forms to support automatic territorial zoning based on clustering thematic information layers with ordinal classes (ordinal categorical data). We evaluated three approaches to clustering ordinal data (thematic maps) to support the Alto Taquari basin territorial zoning. The first was based on the segmentation of the multi-way contingency table [7], the second applied a model-based clustering for ordinal data [3], and the last one conducted the multivariate ordinal data transforming it into numerical and used a proposed clustering method based on the segmentation of the Kohonen's Self-Organizing Map Artificial Neural Network [10].

The section 2 briefly reviews territorial zoning with Self-Organizing Maps and clustering ordinal categorical data. The section 3 unveils the case study, the *Alto Taquari* basin territorial zoning, and presents the three evaluated approaches. Section 4 shows the results and discussion, and the last section is dedicated to the conclusions.

2 Related work

2.1 Zoning with Self-Organizing Maps

The capacity of Self-Organizing Maps to preserve the statistical properties of the data, including the proximity between observations, and its ability for quantization, topological ordering, and visualization impacted its application in several spatial analysis tasks, such as regionalization and zoning [12, 1]. Recently, we can highlight several uses in zoning for ecosystem services [14, 11, 21], ecological-economic zoning [15], environmental zoning [18] and determination of adaptive zones [4].

Except [18], the other applications determine homogeneous zones from numerical data with low dimensionality and combine the use of ANN SOM with a clustering algorithm such as k-means [14] or hierarchical agglomerative [15, 4, 21]. In general, neural networks with few (up to 10^2) neurons predominate in this type of study, and it is possible to conduct studies with very small neural maps, such as a SOM 5×5 [21, 4]. [18] proposed an approach for determining homogeneous zones from thematic maps (categorical data), transforming them to binary data, and adapting the ANN SOM to perform the clustering by segmenting the neural map without the aid of other clustering algorithms.

The literature shows that the SOM is scalable and adequate for non-linear data, enabling applying it to massive data analysis problems. They generally use Euclidean distance as the similarity measure between the numerical input vectors. Still, we can adapt this artificial neural network to tackle different datasets, such as networks, contingency tables, binary data, and graphs [10]. Liu et al. [11] compared the SOM against other automatic zoning methods, showing the unsupervised algorithm's robustness. Still, there is a lack of comparative studies (SOM vs. other) for zoning, zoning from ordinal categorical data (e.g., thematic maps), and clustering using SOM without the support of statistical clustering algorithms.

2.2 Clustering ordinal categorical data

Ordinal categorical data is nominal categorical with a sense of order between each modality whose difference is not directly interpretable. It is common to obtain this information through surveys or other qualitative methods. Clustering this data type implies choosing a similarity measure considering the ordinal character or a distinction process between the probability distribution functions. However, there are cases where it is possible to convert ordinal to numerical, which allows using conventional clustering algorithms (e.g., k-means) by adopting a convenient dissimilarity measure (e.g., Euclidean distance) [2].

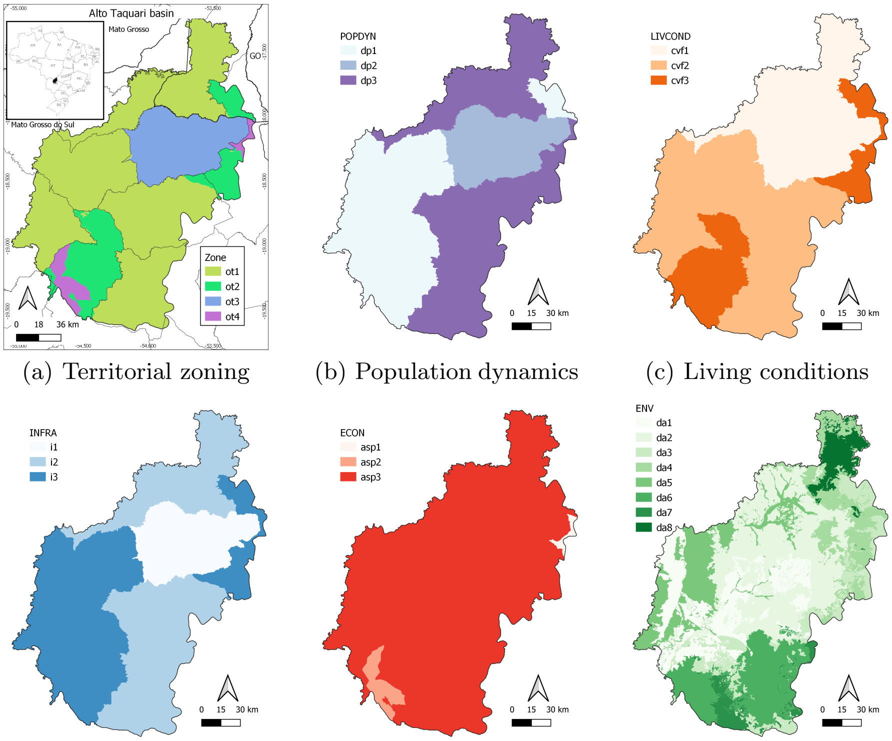
According to [13], in some situations, it is necessary to group the observations associated with ordinal data considering them as such, to allow greater interpretability of dissimilarities (e.g., Goodman & Kruskla γ coefficient), consistency during analysis, and universality of the clustering method (e.g., OrdCIAn-H hierarchical clustering) allowing the treatment of different scales.

This type of clustering can be non-parametric as proposed by [7], which does not use a distance metric between observations. Instead, the authors propose clustering the contingency table cells. Giordan & Diana [7] showed that this approach generates good results for data with low dimensionality compared to the FANNY and PAM methods [9].

Biernacki & Jacques [3] proposed a parametric approach based on modeling the data as a probability distribution function generated by a binary search algorithm. This approach has a solid mathematical foundation. It seeks to model the data to represent important characteristics for clustering, such as the presence of a single mode for each variable per cluster, decreasing probabilities for the other modalities around the mode, and the ability to distinguish between the models. However, the algorithm has a high computational cost as the number of classes increases.

Another way to cluster ordinal data is to transform it into interval numeric and apply a clustering algorithm. There are other ordinal data clustering algorithms, such as ROCK [8]. Still, in this study, we will limit ourselves to evaluating three different and possibly complementary clustering approaches: model-based [3], contingency table [7], and based on the transformation of ordinal data into numeric and subsequent use of the Self-Organizing Map as a clustering algorithm.

3 Material and Methods



(a) Territorial zoning (b) Population dynamics (c) Living conditions

(d) Infrastructure (e) Economic aspects (f) Environmental dimension

Fig. 1. Territorial zoning of the *Alto Taquari* basin (a) obtained by [16] 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).

3.1 The Alto Taquari basin - MS/MT, Brazil

The Alto Taquari basin is found almost entirely in the northeast part of the MS, bordering the Pantanal, with a small region in the south part of the MT. The region crosses or includes 14 municipalities. According to [16], the delimited area comprises $28,046km^2$, and they defined it from topographic maps on a scale of $1 : 250,000$ and digital images from the Landsat 5 satellite, TM sensor, mainly contour lines and drainage network (Fig. 1(a)).

This region was the subject of a comprehensive study that aimed to identify homogeneous zones to support public policies for the region [16]. The methodology consisted of combining pre-existing information (maps of vocation and environmental fragility) with a map of homogeneous zones for territorial planning, elaborated from categorical data of 37 fundamental indicators on the environment (e.g., geology, soil, climate), economy (e.g., land concentration, types of economic activities) and society (e.g., HDI, energy consumption).

Silva & Santos [16] elaborated the territorial ordering map from intermediate maps on the environmental dimension (ENV) (Fig. 1(f)), economic aspects (ECON) (Fig. 1(e)), infrastructure (INFRA) (Fig. 1(f)), living conditions (LIVCOND) (Fig. 1(c)) and population dynamics (POPDYN) (Fig. 1(b)). The authors generated each map using a clustering method on indicators (categorical data) so that each class of this map represents regions in increasing degrees of homogeneity (ordinal classes). Each of these maps represent ordinal categorical classes (e.g., asp1, asp2, asp3 for Economic aspects map).

The authors transformed the ordinal data into binary and applied the Multiple Correspondence Analysis technique. This data reprojection onto new dimensions allowed using the chi-squared distance to measure dissimilarity between observations. Then, the authors applied the hierarchical agglomerative clustering method to these reprojections. So, here the ordinal data is transformed into binary, and consequently, we lose the ordinal information in the process.

The evaluation of the histograms of the untransformed variables for each of the zones found by [16] shows that they are distinguishable in terms of the mode statistic. However, some histograms are bimodal or do not decay around the mode (Fig. 2).

3.2 Clustering categorical ordinal data

Notation To the next sections considers that for n observations we have J categorical ordinal variables, $J \geq 1$, represented by $Y = Y_1, \ldots, Y_J$, and N_j denotes the number of categories or modalities for Y_j, $j = 1, ...J$.

Clustering based on thresholding the multi-way contingency table The algorithm proposed in [7] is designed explicitly for ordinal categorical data and uses the multi-way contingency table generated from the data set as a starting point. Each cell on this table represents observations with the same characteristics, so they must be part of the same cluster. Besides, as we are dealing with

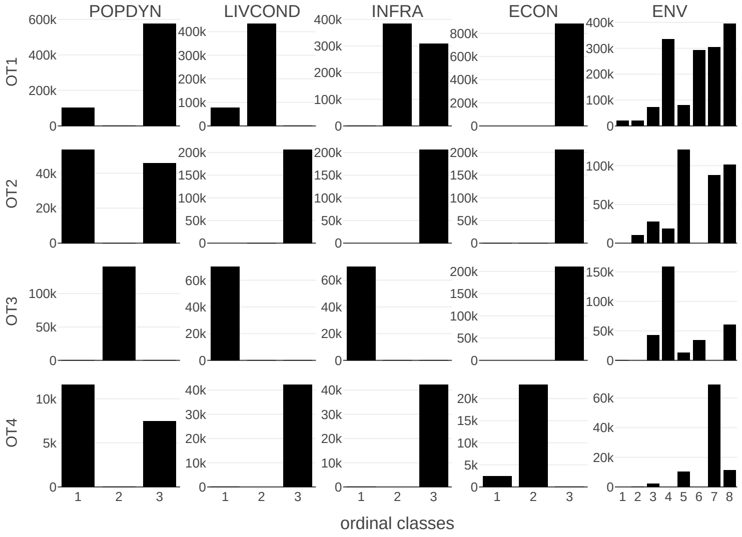


Fig. 2. Histogram for the reference territorial zoning (k=4), Fig. 1(a), according to [16].

ordinal data, neighbor cells may imply some proximity between observations associated with each one. So, the authors use this idea of proximity between cells to propose a clustering algorithm that first considers the cell's density measured as a frequency or proportion and then considers the cell neighborhood to merge groups or associate non-labeled cells into a cluster.

According to the authors, the position of a cell C in a multi-way contingency table is given by $(C_1, \dots, C_J)$, such that $C_j = 1, \dots, N_j$, then the *neighborhood* of C is the set of cells whose coordinates are $(I_1, \dots, I_J) \in Int(1) \times \cdots \times Int(J) \setminus (C_1, \dots, C_J)$ where

$$Int(j) = \begin{cases} \{1, 2\}, & \text{if } C_j = 1 \\ \{N_j - 1, N_j\}, & \text{if } C_j = N_j \\ \{C_j - 1, C_j, C_j + 1\}, & \text{otherwise} \end{cases}$$

From this, we can describe the following algorithm [7, p. 1318]:

Model-based ordinal clustering As stated by [3], mixture models have become a well-established method for clustering due to their mathematical background for parameter estimation and model selection, capacity to generalize some geometric methods, and successful use in real-world situations. A key issue in this approach is defining an appropriate probability distribution function for ordinal data. It can be based on cumulative probabilities, on the constraining of a multinomial model to tackle the ordinality, on assuming that the ordinal data

Algorithm 1 Multi-way contingency table clustering

```
Require: $P$ — Multi-way normalized contingency table containing the proportions $p$
    for each cell
Require: $\lambda \in [0, 1]$ — Threshold for determining initial clusters
Find the cells with high proportion $p$, $p \geq \lambda$
Assign neighboring cells with high proportion $p$ to the same cluster
while there are no cells with $0 < p < \lambda$ in the neighboring of labeled ones do
    if the cell is not labeled and $0 < p < \lambda$ and has only one labeled neighbor then
        this cell assumes the label of the neighbor
    else
        this cell assumes a noise label
    end if
end while
Non-labeled cells will be labeled as noise
```

are the discretization of a continuous latent variable, or on the artificial construction of a model that exhibits some desired properties such as the presence of a unique mode, a decrease of the probabilities from each side of this mode, the possibility to achieve a uniform or a Dirac distribution [3].

The proposition of [3] is based on the last strategy, and it assumes that the ordinal data-generating process results from a Binary Ordinal Search (BOS) algorithm that only uses ordinal information. This probabilistic model (Eq. 1) has two parameters, one related to the position (the mode of the distribution, μ) and the other associated with precision (prominence of the mode, π), that can be estimated by a Maximum Likelihood approach using an Expectation-Maximization algorithm. Then, the model is extended to perform a multivariate ordinal clustering based on these unimodal and univariate distributions (Eq. 2). We used the implementation of this method available at the *ordinalClust* R package, version 1.3.5.

$$p(Y_j, \mu, \pi) = \sum_{e_{N_j-1},\ldots,e_1} \prod_{i=1}^{N_j-1} p(e_{i+1}|e_i; \mu; \pi)p(e_1) \tag{1}$$

where Y_j is the jth-variable, N_j is the number of modalities in this variable, e_i represents an interval $e_i = \{b_i^-, ..., b_i^+\} \subset \{1, ..., N_j\}$ in the ith iteration of a binary search.

$$p(\mathbf{Y}|w_k = 1; \mu_{\mathbf{k}}, \pi_{\mathbf{k}}) = \prod_{h=1}^{J} p(Y_h; \mu_k^h; \pi_k^h) \tag{2}$$

where k represents the cluster, w_k is a variable, such that $w_k = 1$ if the observation belongs to cluster k, and $w_k = 0$ otherwise, $\mu_{\mathbf{k}} = (\mu_k^1, ..., \mu_k^J)$ and $\pi_{\mathbf{k}} = (\pi_k^1, ..., \pi_k^J)$.

Proposed method based on Self-Organizing Map The proposed method transforms ordinal data into interval numerical data as its starting point so that the distance between two subsequent classes is the same for all variables. 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 $\mathbf{Y}^{'}$ has five components varying their values between 1/8 and 8/8.

The standard Self-Organizing Map defined by Teuvo Kohonen is an artificial neural network with unsupervised machine learning. The artificial neurons are represented by weight vectors, w, with the same dimension as the input data. They are organized in a two-dimensional grid, $N \times M$, with rectangular or hexagonal lattice that defines the neighborhood between neurons. The sequential or stochastic Machine Learning mechanism is iterative and can be divided into three phases. In the first phase, competitive, the input data are randomly presented to the neural network, and the neuron closest to the input vector according to the Euclidean distance is considered the Best Match Unit (BMU). In the second phase, cooperative, the neighboring neurons of the BMU are defined, which will also be updated in the third phase, adaptive, where each weight vector w of the BMU and its neighbors is updated according to the Eq. 3.

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(t)h(t)(\mathbf{Y}_i^{'} - \mathbf{w}(t)) \tag{3}$$

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 $\mathbf{Y}_i^{'}$ 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 SOM feature allows clustering algorithms (e.g., k-means) on the neural network's weights as an indirect way to partition the input data [19].

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, level of activation of neurons (number of input vectors associated with it), and data density between neurons. Costa & Netto [5] proposed a graph-based SOM partitioning model that uses all this information and automatically determines the number of clusters, which Silva et al. [18] successfully applied. This proposal has as a limiting factor, there are three hyperparameters that we must adjust to each data set. Silva & Costa [17] also proposed a graph-based method for segmenting the SOM, but using the density between neurons exclusively as a segmentation method using 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 interpreting the SOM as an undirected graph, which uses all the information available after the machine learning process without needing hyperparameter adjustment. It is only necessary to define the desired number k of clusters (Algorithm 2). We implemented it in Python version 3.8 using the Minisom version 2.3.0 as the SOM solution [20].

Algorithm 2 SOM-based proposal

```
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 connect nodes in $T$
```

Fig. 3 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 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).

3.3 Clustering assessing

Considering that we will evaluate three very different forms of analysis of ordinal data groupings, we chose to assess the distinction between clusters based on analyzing the distribution of ordinal classes by variable and group. From the histograms, we evaluated the type of distribution per group (unimodal, bimodal, multimodal) and the distribution format (whether it decays around the mode or not).

4 Results and Discussion

We performed the model-based clustering for different values of the number of groups. However, the algorithm only converged for values equal to or less than four. For analysis, we have chosen to evaluate the result for $k = 4$ that generated the groups illustrated in the map in Fig. 4(a). The analysis of the histograms by each of the four clusters for the five ordinal variables considered showed that, although in most cases, we have unimodal distributions, it was not possible to

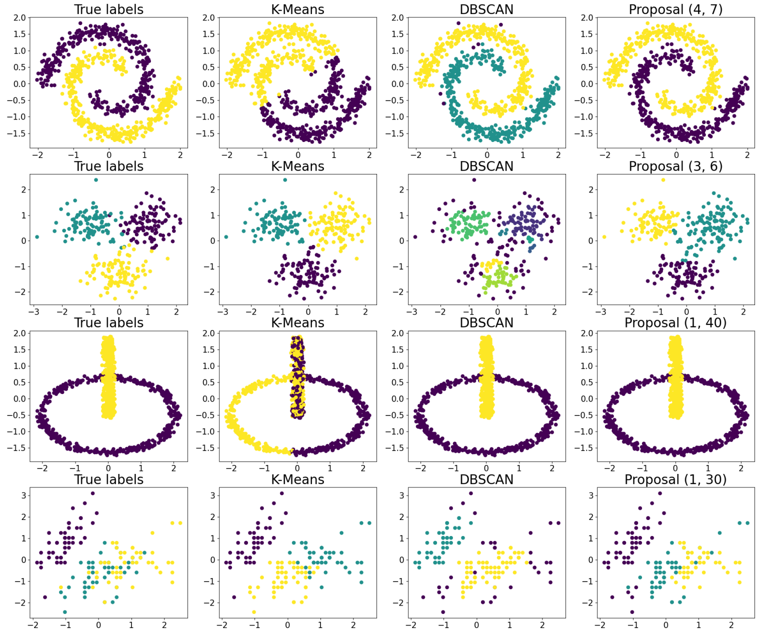


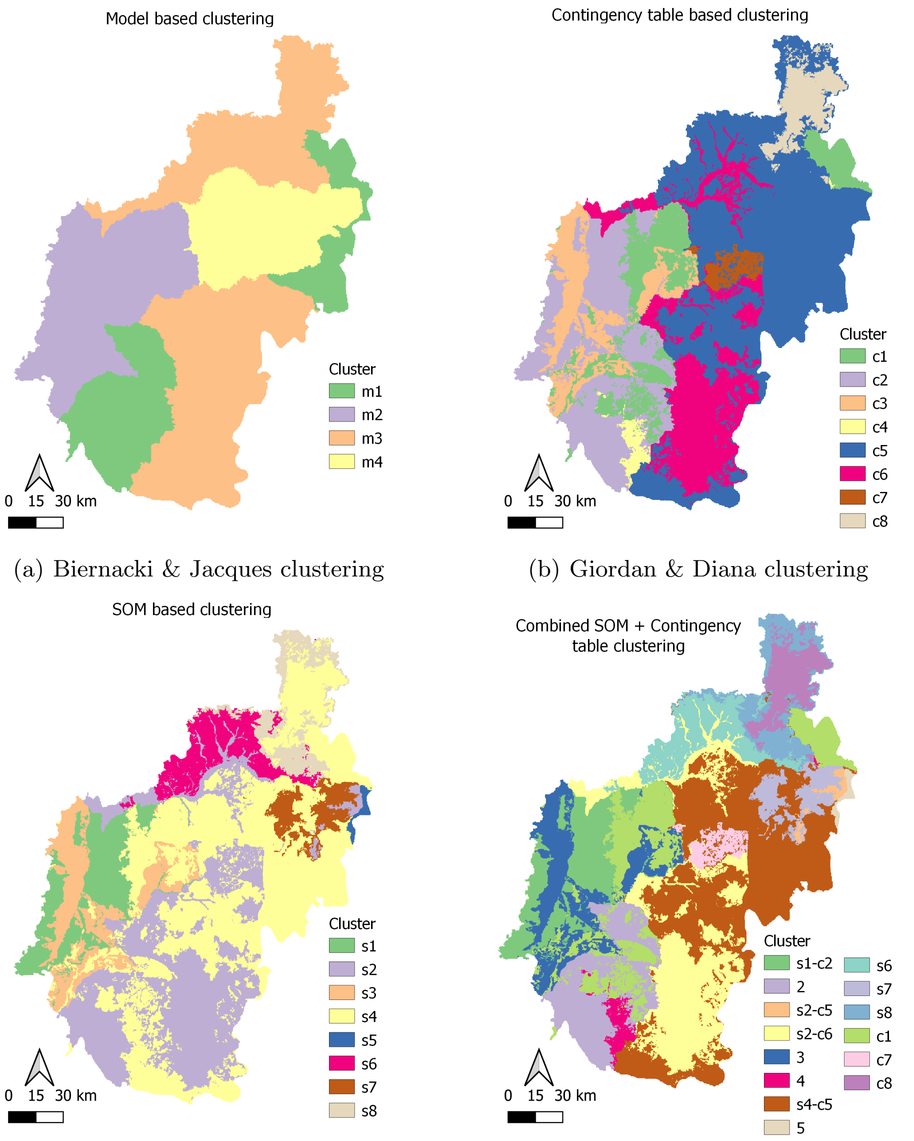
Fig. 3. In the first column on the left, we have the four labeled datasets (spiral, gaussian, chainlink, and iris data sets) used to evaluate the proposed method (last column), compared with the k-means methods in the second column and DBSCAN in the third column.

establish distinctions between the clusters through the analysis of the respective modes (Fig. 5).

We evaluated different values for the threshold for the clustering based on the contingency table, which varied between 0.007 and 0.082 in intervals of 0.005. The number of clusters generated ranged between eight for the lowest threshold and four for the highest value, and there was consistency between the different clusters, that is, without random variation between the clusters. Thus, we chose to analyze the partition with the largest number of clusters, eight, seeking to identify as many distinctions as possible, as seen in Fig. 4(b).

Fig. 6 shows the histograms for each of the eight analyzed clusters, where we observe unimodal histograms with a clear distinction between the clusters, except for cluster 1 and the variable *ENV*, and for the variable *ECON* where no difference is observed between the groups as also shown in the model-based clustering histograms.

We evaluated different hyperparameters related to the ANN size, initial radius (sigma), and lattice of the neural grid (hexagonal or rectangular) for clustering the ordinal data transformed to numeric from the proposed post-training ANN SOM segmentation algorithm. To help choose the best result, we calculated the Davies-Bouldin clustering validation index, indicating the 6×5 hexagonal



(a) Biernacki & Jacques clustering

(b) Giordan & Diana clustering

(c) Proposed clustering

(d) Merge of the results from (b) and (c)

Fig. 4. Maps in raster format obtained from the clustering of the five ordinal variables (POPDYN, LIVCOND, INFRA, ECON, and ENV) from the method proposed by Biernacki & Jacques (a), Giordan & Diana (b), and the method proposed in this article (c). In map (d) we have a merge of maps (b) and (c) in order to highlight their total coincidences (clusters 2, 3, 4, and 5), complementarities (clusters that represent regions differentiated by one and not by the other algorithm, clusters $s6$, $s7$, $s8$, $c1$, $c7$ and $c8$) and partial matches (when there are partial matches between them, clusters $s1 - c2$, $s2 - c5$, $s2 - c6$, and $s4 - c5$)

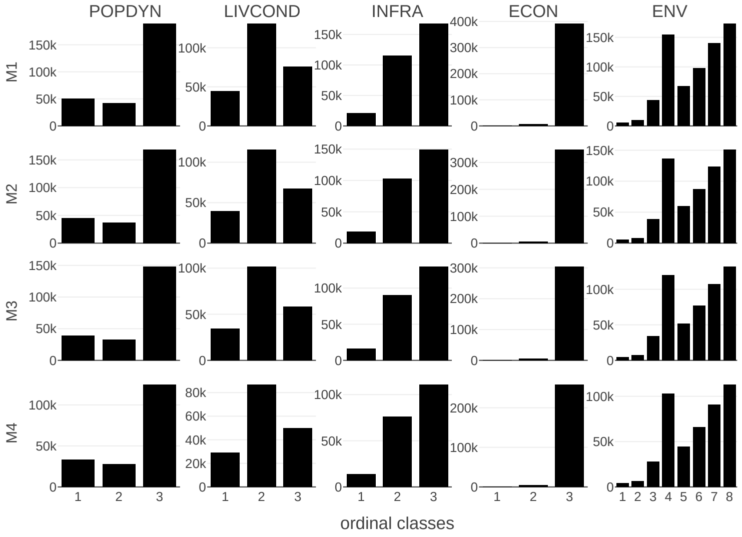


Fig. 5. Histogram for the model-based clustering method over the five ordinal categorical variables (POPDYN, LIVCOND, INFRA, ECON, and ENV) proposed by Biernacki & Jacques considering $k = 4$.

SOM ANN with sigma equal to 1.0 and partitioned it into six groups. After analyzing this 6×5 hexagonal ANN SOM for other values of k, we decided to analyze this same network partitioned into eight groups, resulting in the map in Fig. 4(c).

We observed that clustering based on the contingency table and the proposed method generated coincident partitions regarding spatial location, even if not entirely, and each strategy identified spatially distinct clusters. The analysis of the histograms from Fig. 7 for the clustering according to the proposed method shows that, as well as the method based on the contingency table, there is a substantial distinction between the clusters in terms of unimodal distributions, with emphasis on the variable *ECON* where the algorithm was able to distinguish, unlike the other techniques. In this clustering, the *ENV* variable shows better-defined distributions when compared to those generated by the contingency table method.

The model-based method was the only method that could have been more successful in partitioning the data to generate distinguishable groups from analyzing the distributions of the modalities. We should conduct further studies before discarding this approach for analyzing thematic maps with ordinal classes. This approach had a higher computational cost, followed by the proposed method, which depends mainly on the number of neurons in the ANN. The solution proposed by Giordan & Diana has the lowest computational cost.

The clustering method from the contingency table is deterministic. It depends solely on the definition of the threshold, which can be defined through trial

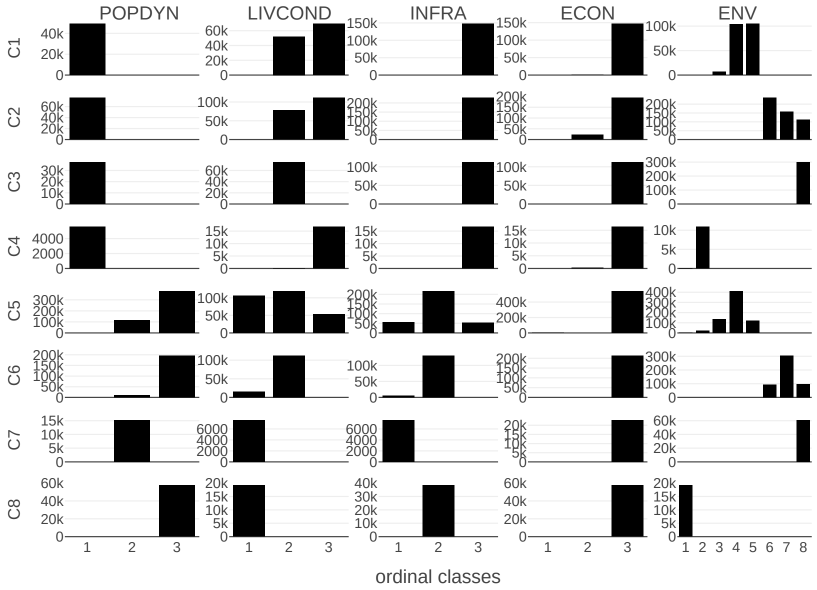


Fig. 6. Histogram for the multi-way contingency table clustering method over the five ordinal categorical variables (POPDYN, LIVCOND, INFRA, ECON, and ENV) proposed by Giordan & Diana considering $k = 8$ and a threshold equal to 0.007.

and error or by some other hyperparameter optimization method. The proposed method can be indicated for data with complex geometry. Still, it requires a more significant effort in choosing its hyperparameters, mainly the size of the neural network, which will determine the level of ability to separate the different intrinsic patterns in the dataset.

The environmental zoning process is complex and requires a multidisciplinary effort to establish the boundaries comprising the homogeneous zones. For the case of the Alto Taquari basin, the present work suggests, for example, the combined analysis of the partitions performed by the proposed method and the one based on the contingency table, as shown in Fig. 4(d). In this way, we could take advantage of the complementarities of each approach to have clusters with more significant distinctions.

The proposed method and the one based on the contingency table could partition the data into groups with characteristics of unimodal distributions, with the other classes decaying around the mode and quite distinguishable from each other. Both methods identified homogeneous areas with reduced area and, even so, distinguishable, as was the case of cluster $c4$ for clustering based on the contingency table and cluster $s5$ generated by the proposed algorithm. There is identification of regions with homogeneity detected by these two methods, emphasizing clusters 2, 3, 4, and 5, Fig. 4(d).

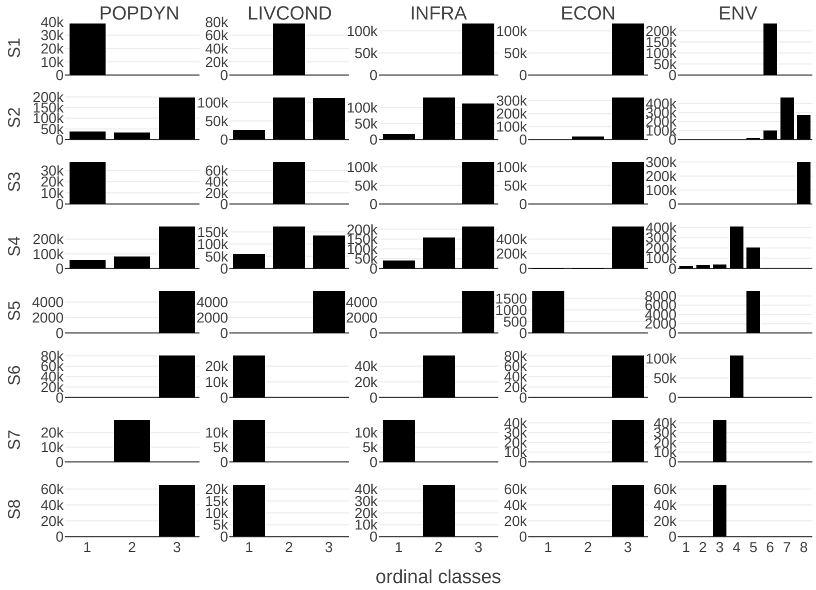


Fig. 7. Histogram for the SOM's segmentation clustering proposed method over the five ordinal categorical variables (POPDYN, LIVCOND, INFRA, ECON, and ENV) considering $k = 8$ and a and a 6×5 neural network.

5 Conclusions

We evaluated three approaches with different strategies for clustering ordinal data with many observations ($n = 448000$), low dimensionality ($d = 5$), and a few modalities (four variables with three and one with eight). The model-based clustering proposed by [3] did not achieve the objective of dividing the data into distinguishable groups, suggesting modifications in the algorithm's parameterization or the need to evaluate other model-based solutions for ordinal data. The proposed method and the one based on the contingency table showed good results regarding the distinction of groups. They showed the ability to identify coincident homogeneous regions but also complementary ones. The ordinal data clustering algorithm based on the contingency table is low computational cost, simple to understand, and requires the adjustment of a single parameter. However, we must evaluate its ability to separate ordinal data into distinguishable groups for data with higher dimensionality and complexity. The proposed method also deserves more exhaustive tests considering new datasets (ordinal and numerical) and comparisons with other non-parametric and parametric clustering methods.

References

1. Agarwal, P., Skupin, A. (eds.): Self-Organising Maps: Applications in Geographic Information Science. John Wiley and Sons, Chichester (2008)

2. Agresti, A.: Analysis of Ordinal Categorical Data. Wiley Series in Probability and Statistics. Wiley-Interscience, New York (2010)
3. Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistical Computing **26**, 929–943 (2016)
4. Bustos-Korts, D., Boer, M.P., Layton, J., Gehringer, A., Tang, T., Wehrens, R., Messina, C., de la Vega, A.J., van Eeuwijk, F.A.: Identification of environment types and adaptation zones with Self-Organizing Maps; applications to sunflower multi-environment data in Europe. Theoretical and Applied Genetics **135**, 2059–2082 (2022)
5. Costa, J.A.F., Netto, M.L.A.: Segmentação do SOM baseada em particionamento de grafos. In: VI Congresso Brasileiro de Redes Neurais. pp. 451–456 (2003)
6. Furtado, B.A., Sakowski, P.A.M., Tóvolli, M.H. (eds.): Modeling complex systems for public policies. Institute for Applied Economic Research, Brasília, DF (2015)
7. Giordan, M., Diana, G.: A clustering method for categorical ordinal data. Communications in Statistics—Theory and Methods **40**(7), 1315–1334 (2011)
8. Guha, S., Rastogi, R., Shims, K.: ROCK: A robust clustering algorithm for categorical attributes. Information Systems **25**(5), 345–366 (2000)
9. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. J. Wiley Sons, New York (1990)
10. Kohonen, T.: Self-Organizing Maps. Berlin: Springer (2001)
11. Liu, Y., Li, T., Zhao, W., Wang, S., Fu, B.: Landscape functional zoning at a county level based on ecosystem services bundle: methods, comparison and management indication. Journal of Environmental Management **249**(109315), 1–11 (2019)
12. Nikparvar, B., Thill, J.C.: Machine learning of spatial data. International Journal of Geo-Information **10**(600), 1–32 (2021), doi.org/10.3390/ijgi10090600
13. Podani, J.: Braun-Blanquet's legacy and data analysis in vegetation science. Journal of Vegetation Science **17**, 113–117 (2006)
14. Pérez-Hoyos, A., Martínez, B., García-Haro, F.J., Álvaro Moreno, Gilabert, M.A.: Identification of ecosystem functional types from coarse resolution imagery using a Self-Organizing Map approach: A case study for spain. Remote Sensing **6**, 11391–11419 (2014)
15. Sadeck, L.W.R., de Lima, A.M.M., Adami, M.: Artificial neural network for ecological-economic zoning as a tool for spatial planning. Pesquisa Agropecuária Brasileira **52**(11), 1050–1062 (2022)
16. Silva, J.S.V., Santos, R.F.: Estratégia metodológica para zoneamento ambiental: a experiência aplicada na Bacia Hidrográfica do Rio Taquari. Embrapa Informática Agropecuária, Campinas, SP (2011)
17. Silva, L.A., Costa, J.A.F.: A graph partitioning approach to SOM clustering. In: 12th Int. Conf. on Intelligent Data Engineering and Automated Learning (2011)
18. Silva, M.A.S.d., Maciel, R.J.S., Matos, L.N., Dompieri, M.H.G.: Automatic environmental zoning with Self-Organizing Maps. MESE **4**(9), 872–881 (2018)
19. Silva, M.A.S.d., Matos, L.N., Santos, F.E.d.O., Dompieri, M.H.G., Moura, F.R.d.: Tracking the connection between Brazilian agricultural diversity and native vegetation change by a machine learning approach. IEEE Lat. Am. T. **20**(11), 2371–2380 (2022)
20. Vettigli, G.: Minisom: minimalistic and NumPy-based implementation of the Self Organizing Map (2018), `https://github.com/JustGlowing/minisom/`
21. Yan, Y., Deng, Y., Yang, J., Li, Y., Ye, X., Xu, J., , Ye, Y.: Exploring the applicability of Self-Organizing Maps for ecosystem service zoning of the Guangdong-Hong Kong-Macao greater bay area. SPRS Int. J. Geo-Inf. **11**(481), 1–20 (2022)