Evaluating a Self-Organizing Map approach to cluster a Brazilian agricultural diversity spatial panel data

Marcos A. S da Silva[1], **Leonardo N. Matos**[2], **Flávio E. de O. Santos**[2], **Fábio R. de Moura**[3], **Márcia H. G. Dompieri**[4]

[1]Embrapa Tabuleiros Costeiros
Av. Beira Mar, 3250, 49.025-370 – Aracaju – SE – Brazil.

[2]Departmento de Computação – Universidade Federal de Sergipe (UFS)
Cidade Universitária, São Cristóvão – SE – Brazil.

[3]Departmento de Economia – Universidade Federal de Sergipe (UFS)
Cidade Universitária, São Cristóvão – SE – Brazil.

[4]Embrapa Territorial
Av. Sd. Passarinho, 303 - Jardim Chapadão, 13070-115, Campinas – SP, Brazil.

{marcos.santos-silva,marcia.dompieri}@embrapa.br

leonardo.matos@dcomp.ufs.br,{fabiromoura,flavioemanuel859}@gmail.com

Abstract. *Brazil has a substantial diversity of agricultural activities that pose challenges to developing public policies for the sector. Characterizing the evolution of this diversity in time and space is of paramount importance both for agribusiness and for small producers. Based on annual estimates of agricultural production, it is possible to group Brazilian municipalities according to their trajectory over the years. In this work, these estimates were used to calculate the annual agricultural diversity per municipality-category, and cluster them according to the proposed method based on the Self-Organizing Map that is usually more suitable for large datasets with non-convex structure. Results showed that the proposed method is suitable for the Brazilian agricultural spatial panel data and presented better results when compared with the k-means and Genelarized Linear Mixture Model method.*

1. Introduction

The Brazilian agricultural activities show a huge spatial diversity due to economic and historical processes, and this constrains the public policy design to support either smallholder farmers and the agribusiness [Schneider and Cassol 2014]. As stated by [Teixeira and Ribeiro 2020, Sambuichi et al. 2016] the Brazilian rural diversity impacts food security and resilience of agricultural systems, mainly on the small rural business. Moreover, knowing this spatial diversity can be valuable in identifying new agribusiness trends and unveiling regional and local heterogeneities.

One way to investigate agricultural diversity is to use indices to measure diversity from data about the production or cultivated area [Dessie et al. 2019]. In Brazil, there are some studies about family farming as in [Sambuichi et al. 2016], and [Teixeira and Ribeiro 2020]. The latter used the Simpson's diversity index on Pronaf Aptitude Statement (DAP, Declaração de Aptidão ao Pronaf) data to classify all the

Brazilian family farmers into very diverse, diverse, poorly diverse, and not diversified. [Teixeira and Ribeiro 2020] also applied Simpson's index to classify the Minas Gerais municipalities according to the mean of their agricultural quantity production using IBGE estimates between 2014 and 2018. Despite these studies, there is a lack of work about general spatio-temporal agricultural Brazilian diversity.

In this work, a diversity index based on Shannon's entropy index and a proposed machine learning clustering algorithm has been applied to identify spatiotemporal patterns of Brazilian agricultural activities. We used annual values estimated by the IBGE between 1999 and 2018 related to temporary and permanent cultivated crops, animal population (including dairy animals), aquaculture, vegetal extractvism, and silviculture [IBGE 2020].

The proposed method is based on the Self-Organizing Map (SOM), which aims to order the data into a low dimensional grid for clustering, and visual data exploration [Kohonen 2013]. In this paper, the location of each observation (municipality) will not be considered in the clustering process as proposed by [Skupin and Hagelman 2005].

The adopted strategy will consider each observation-year an entry to the SOM and observe what trajectory is generated on the neural map by chronologically linking each one as proposed by [Chen et al. 2018, Ling and Delmelle 2016, Augustijn and Zurita-Milla 2013, Wang et al. 2013]. The spatial patterns will be verified after the clustering process, simply mapping the result into the Brazil map and checking for global and local spatial dependences according to [Qi et al. 2019, Chen et al. 2018, Ling and Delmelle 2016, Wang et al. 2013].

The performance of the proposed method was compared with two other approaches to cluster spatial panel data, k-means adapted to panel data [Genolini et al. 2015] and a Generalized Linear Mixture Model (GLMM) using Markov Chain Monte Carlo (MCMC) to estimate parameters and cluster [Komárek and Komárková 2014, Komárek and Komárková 2013].

This paper is organized as follows: Section 2 presents the spatial panel data, the proposed clustering method based on Self-Organizing Map, and other methods and quality clustering measures. Section 3 shows the results and discussion, and section 4 is dedicated to conclusions.

2. Material and methods

2.1. Spatial panel data - Brazilian agricultural diversity

The diversity of Brazilian agriculture has been evaluated based on the analysis of raw data from eight categories from IBGE annual estimates for the period 1999 to 2018: animal population, including dairy animals (DIV.EFETIVO); the planted area with temporary crops (DIV.PLANT.T), value of production of animal origin (DIV.VL.PRODANI), temporary (DIV.VL.T) and permanent (DIV.VL.P) crops, aquaculture (DIV.AQU.VL), vegetal extraction (DIV.EXTV.VL) and forestry (DIV.SILV.VL) [IBGE 2020].

Then, the panel data are composed of eight diversity indexes for each of the 5570 municipalities for 20 years, from 1999 to 2018, so it comprises 111400 observations. Table 1 presents a statistical summary for them. The aquaculture production value index (DIV.AQU.VL) showed a high coefficient of variation, the animal population, including dairy animals (DIV.EFETIVO), and planted areas with temporary crops (DIV.PLANT.T) indexes showed the highest averages with the smallest coefficients of variation.

Shannon's entropy [Shannon 1948] has been chosen because it is invariant to the number of possible elements in each category. Thus, it is possible to compare the diversity indices of different categories based on entropy (Equation 1).

$$Diversity_{ltp} = -\sum_{i=1}^{m} \left[\frac{X_{ltpi}}{\sum_{j=1}^{m} X_{ltpj}} log_m \left(\frac{X_{ltpi}}{\sum_{j=1}^{m} X_{ltpj}} \right) \right] \tag{1}$$

where t represents the year of reference, l the category, p the municipality, m the number of raw variables used for each category and X_{ltpi} the value of the ith raw variable for the year t, category l and municipality p. The diversity index values vary from zero (without diversity) to one (highest diversity level).

Table 1. Statistical summary for all eight diversity indexes for all year. Source: elaborated by the authors.

VarName	SD	Mean	CV	Median	Max	m
DIV.EFETIVO	0,18	0,48	37,50%	0,52	0,87	11
DIV.PLANT.T	0,074	0,32	23,13%	0,33	0,51	33
DIV.VL.T	0,13	0,28	46,43%	0,3	0,68	33
DIV.VL.P	0,16	0,2	80,00%	0,19	0,72	38
DIV.VL.PRODANI	0,18	0,26	69,23%	0,26	0,88	6
DIV.AQU.VL	0,085	0,026	326,92%	0	0,66	24
DIV.EXTV.VL	0,1	0,088	113,64%	0,041	0,49	44
DIV.SILV.VL	0,14	0,084	166,67%	0	0,75	15

2.2. Self-Organizing Maps, k-means and model based clustering algorithms

The Kohonen Self-Organizing Map is an ANN with two layers (Kohonen, 2001): the input I layer and the output U layer. The input of the lattice corresponds to a vector in d-dimensional space in $\mathbb{R}^d$, represented by $x_i, i = 1, ..., n$, where n represents the number of observations. Each output layer neuron j has a codevector w, also in space $\mathbb{R}^d$.

The SOM training algorithm consists of three phases. In the first phase, *competitive*, the output layer neurons compete with each other, according to some criterion, in this case, the Euclidean distance, to find a single winner, also called a BMU (Best Match Unit). In the second, *cooperative* phase, the neighborhood of this neuron is defined. In the last phase, *adaptive*, the codevectors of the winning neuron and its neighborhood are updated according to the Equation 2.

$$w_{ij}(t + 1) = w_{ij}(t) + \alpha(t)h(t)[x_{ik}(t) - w_{ij}(t)] \tag{2}$$

where $\alpha(t)$ is the learning rate function, and $h(t)$ is the neighborhood function centered on the winning neuron (BMU).

4. Nonetheless, in this case, a trajectory for a municipality p will be expressed as a matrix $Traj^p_{ij}$ where i denotes an index for each year (1999-2018) and j refers to each

diversity index $(DIV.EFETIVO, ..., DIV.SILV.VL)$. The quality clustering measures Calinski-Harabatz and Davies-Bouldin will support the choice of the best number of clusters. The k-means algorithm and quality measures are implemented in the kml R package [Genolini et al. 2015].

The SOM's trajectory clustering strategy was also be compared with a spatial panel data clustering based on a multivariate mixture Generalized Linear Mixed Model (GLMM) and a Bayesian inference based on the Monte Carlo Markov Chain (MCMC) simulation [Komárek and Komárková 2013]. For this paper, it has been used default parameters of the GLMM MCMC algorithm implemented in mixAK R package [Komárek and Komárková 2014].

To compare the results obtained by the three algorithms were used the quality measures Calinski-Harabatz and Davies-Bouldin, replacing the Euclidean distance by the Frobenius distance between the trajectory matrices $Traj^A$ and $Traj^B$ of municipalities A and B, Equation 3. To assess the degree of homogeneity of the clusters, we chose to evaluate the Coefficient of Variation for each cluster-variable considering all years simultaneously.

$$F_{Traj^A,Traj^B} = \sqrt{trace((Traj^A - Traj^B) * (Traj^A - Traj^B)^T)} \tag{3}$$

2.3. Proposed method - spatial panel data clustering using SOM

The method used to group Brazilian municipalities according to the values of the eight diversity indices between 1999 and 2018 comprises seven steps (Figure 1). The first and second steps have been described in section 2.1. All R code and data are available at [da Silva et al. 2021].

2.3.1. Step 3 - Spatial panel data ordering on the Self-Organizing Map (SOM)

The third step consists of spatial panel data ordering onto a two-dimensional SOM with a hexagonal grid, Gaussian neighborhood function, and stochastic machine learning. The number of neurons was determined empirically based on the quantization error and the projections of the observations in the neural grid. In this work, it has been chosen a small size SOM as used by [Augustijn and Zurita-Milla 2013].

2.3.2. Step 4 - Clustering the SOM's code vectors

In this step, the SOM weights were clustered using the k-means method, considering the elbow method and the Silhouette quality index analysis to determine the number of groups. This clustering will help interpret the Component Planes, generated from the SOM weights, by dividing the neural grid into regions with homogeneous characteristics.

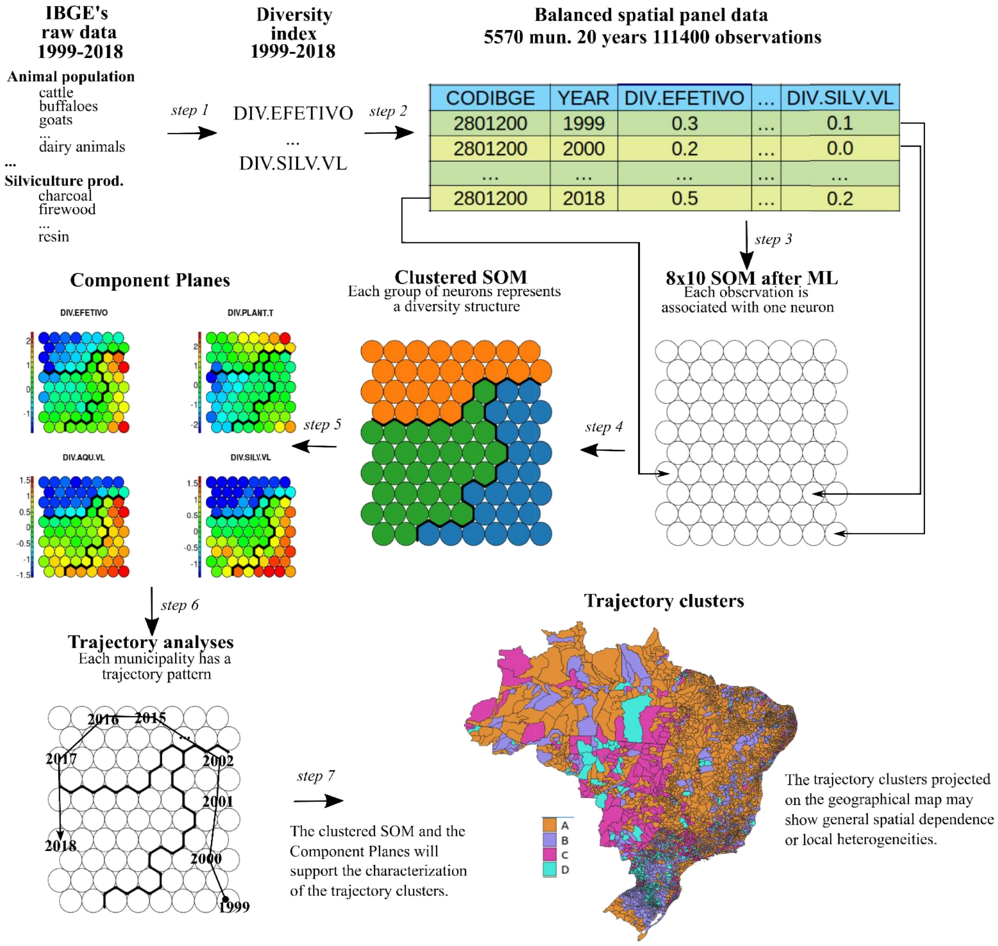

Figure 1. The proposed method of spatial panel data clustering is based on trajectory analysis onto the neural map. Source: elaborated by the authors.

2.3.3. Step 5 - Component Planes (CP) analysis

To check the pattern of each variable on the neural map a coloring method based on the values of each component is used - the Componente Planes. For a given $j - th$ component of the SOM's code vectors, an image $f(x, y)$ is generated with dimensions equal to the map MxN. Each pixel will correspond to the value of the j component at the position (x, y) using a divergent palette pattern (dark blue represents maximum values, dark red minimum values, and shades of green and yellow for intermediate values). Thus, Component Planes can be used to check for correlation between variables, visual clustering, and, in this paper, to explain each region on the clustered neural grid generated in the precedent step as proposed by [Qi et al. 2019, Augustijn and Zurita-Milla 2013, Skupin and Hagelman 2005]. Due to the limited number of pages, this article will not cover the visual interpretation of CP.

2.3.4. Step 6 - SOM's Trajectory clustering

In the sixth step, the trajectory generated by chronologically linking each observation-year on the neural grid can be visually analyzed for each municipality or applying a clustering algorithm as proposed by [Ling and Delmelle 2016]. A trajectory for a municipality p can be expressed as a matrix $Traj_{ij}^{p}$ where each row corresponds to a coordination (x, y) on the neural grid. Hence, to cluster all trajectories, it has been applied a k-means algorithm using the matrix distance defined in the Equation 4 [Genolini et al. 2015]. The Davies-Bouldin quality index has measured the quality of the trajectory clustering, also implemented in [Genolini et al. 2015].

$$Dist(Traj^1, Traj^2) = \sqrt{\sum_i \sum_j (Traj_{ij}^1 - Traj_{ij}^2)^2} \tag{4}$$

2.3.5. Step 7 - Projection on the geographic map

In the last step, the clusters will be mapped on the geographic map to observe spatial dependence and spatial heterogeneities as proposed by [Qi et al. 2019, Ling and Delmelle 2016]. That is, whether the distribution of groups follows any regional or local spatial pattern.

3. Results and Discussion

3.1. K-means

The k-means algorithm was applied to all eight agricultural diversity indices using the Davies-Bouldin and Calinski & Harabatz 2 validation indices as a reference for the definition of the number of clusters (Figure 2). These algorithms divided the municipalities into eight clusters, see Figure 3, where there are significant differences between the five regions of Brazil, with emphasis on the NE region where municipalities associated with cluster A predominate.

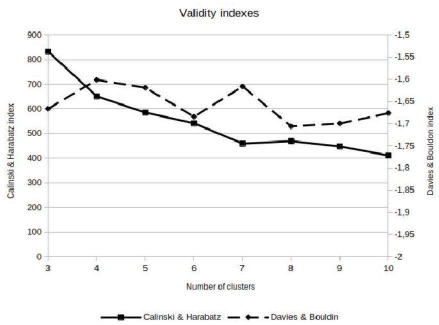

Figure 2. Quality measures for all teste number of clusters for the k-means algorithm. Source: elaborated by the authors.

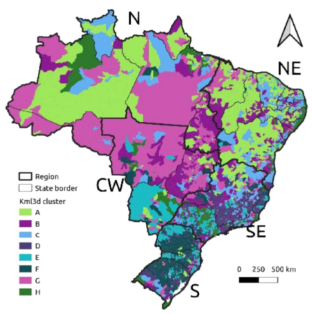

Figure 3. Brazilian municipalities' agricultural diversity clustering by the k-means algorithm. Source: elaborated by the authors.

Table 2 shows the coefficients of variation for all variables and the number N of observations associated with each group generated by the k-means algorithm. Note a balanced distribution of the number of observations per group, strong CV for the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL and greater homogeneity for the variable DIV.PLANT.T.

Table 2. Coefficient of Variation (%) for all diversity indices for each cluster (N is the number of observations per group) generated by the k-means algorithm. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI	DIV.AQU.VL	DIV.EXTV.VL	DIV.SILV.VL
A	892	20,52	15,71	36,21	69.91	64.64	375.18	36.57	570.49
B	885	25.47	19.81	43.77	143.58	98.87	546.11	120.62	456.92
C	791	21.41	15.92	34.45	41.85	63.15	544.83	131.49	315.67
D	706	18.36	17.79	37.98	73.92	51.48	432.71	202.55	68.54
E	651	57.02	17.93	39.61	68.87	39.63	305.08	194.96	92.48
F	610	40.99	15.01	26.40	39.88	28.47	186.46	74.72	60.39
G	554	28.61	17.90	36.33	72.71	74.93	172.02	88.04	413.40
H	481	56.79	51.62	118.74	99.80	92.22	533.11	276.39	290.47
CV Mean		33.65	21.46	46.69	76.31	64.17	386.94	140.67	283.54
CV SD		0.141	0.108	0.260	0.293	0.215	1.362	0.696	1.709

3.2. GLMM MCMC

The GLMM MCMC algorithm was applied to six diversity indices and the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL were eliminated from the analysis due to their peculiar characteristics, such as extremely skewed distribution that prevented convergence of the algorithm. The variables DIV.VL.P and DIV.VL.PRODANI were transformed from the cubic root function to approximate their density curves to the Gaussian distribution.

The GLMM MCMC algorithm generated the spatial cluster shown in Figure 4. It can be seen from the map that the majority (2982) of the municipalities was associated with group D, which is predominant in all five regions of Brazil.

Table 3 shows the coefficients of variation for all variables and the number N of observations associated with each group generated by the GLMM MCMC algorithm. There is a very balanced distribution of the number of observations per group, with low variation of CV for all variables, with the variable DIV.PLANT.T showing greater homogeneity.

3.3. Proposed method

Cluster analysis with the proposed method was performed with a 25x30 two-dimensional SOM, non-toroidal hexagonal, the neighborhood defined by a Gaussian function, and

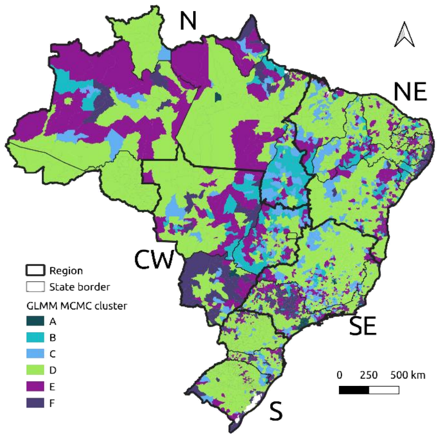

Figure 4. Brazilian municipalities' agricultural diversity clustering by the GLMM MCMC algorithm. Source: elaborated by the authors.

sequential (online) stochastic machine learning with 100,000 iterations with the learning rate monotonically decreasing starting from 0.05.

Following the proposed method, in the next step, the SOM was segmented into six homogeneous regions on the neural grid that, together with the Component Plans generated in step five, helps in the characterization of the groups that will be generated at the end of the process (Figure 5). Due to space limitations in the article, we will not address the characterization of the neural grid groups and the clusters performed in the last step of the process. The effective clustering of the municipalities was performed by clustering the trajectories of each observation in the neural grid using the k-means algorithm, dividing the 5570 municipalities into eight groups.

Figure 6 shows the average trajectories for each group. It is observed that groups A, C, H, and D represent trajectories that do not shift much considering the edges created by the homogeneous regions of the neural grid generated in step 4 of the proposed method. It denotes that the major trend of municipalities associated with these groups is not to change their diversity indices between 1999 and 2018. On the contrary, groups B, E, F, and G represent the average of municipalities whose trajectory in the grid tends to move between the six homogeneous regions of the neural grid. It implies that there are municipalities with trends towards changes in the profile of agricultural diversity and that there are at least four types of trends towards changes.

The distribution of clusters on the map (Figure 7) shows that the predominant group A is mainly concentrated in the NE, MG, and part of the states of TO and GO. Cluster C predominates in the Southern region and B in regions N and CW in the southern region. Cluster B represents a set of municipalities with a tendency to decrease in diversity.

3.4. Comparing the methods

The analysis of the coefficients of variation shows that the proposed method partitioned the municipalities to ensure greater homogeneity than the k-means method, es-

Table 3. Coefficient of Variation (%) for all diversity indexes for each cluster (N is the number of observations per group) generated by the GLMM MCMC algorithm. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI
A	80	37.38	24.60	47.89	81.13	69.58
B	456	37.12	22.84	45.48	77.84	69.41
C	599	37.31	22.15	45.52	78.11	67.93
D	2982	37.08	22.97	45.77	77.19	68.67
E	984	36.97	22.89	46.21	78.40	68.73
F	469	37.91	24.25	47.47	79.07	69.16
Mean		37.29	23.28	46.39	78.62	68.91
SD		0.003	0.009	0.010	0.013	0.005

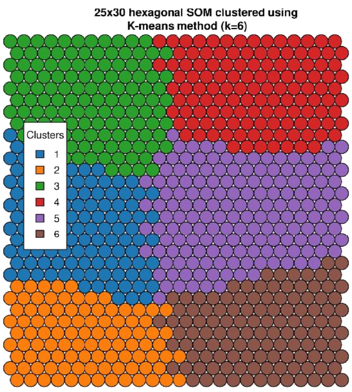

Figure 5. Clustered SOM's code vectors using the k-means method. Source: elaborated by the authors.

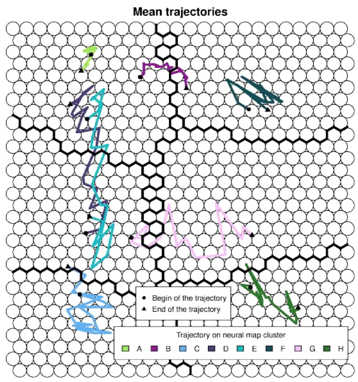

Figure 6. Illustration of the average trajectory of each group (A-H) defined from the trajectory clustering. Source: elaborated by the authors.

pecially when observing the mean and standard deviation for the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL. Although many observations (2015) were associated with cluster A, there was a balance in the distribution in terms of the number of observations per group.

Although the model-based clustering method has generated the groups with greater homogeneity, considering the CV per variable, the imbalance in the distribution of observations per group and the need to exclude three variables showed that this might not be the most suitable method for the evaluated data set.

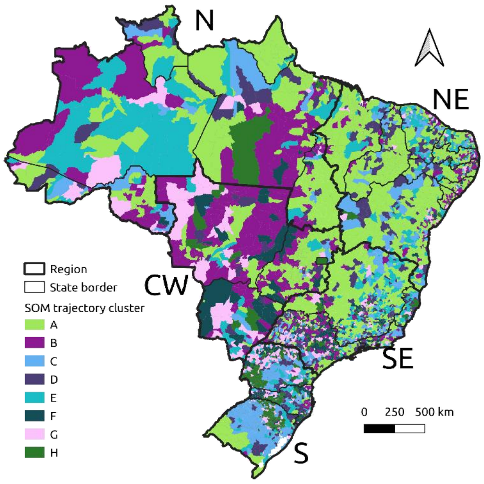

Figure 7. Brazilian municipalities' agricultural diversity clustering by the proposed method. Source: elaborated by the authors.

Table 4. Coefficient of Variation (%) for all diversity indices for each cluster (N is the number of observations per group) generated by the proposed method. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI	DIV.AQU.VL	DIV.EXTV.VL	DIV.SILV.VL
A	2105	17.81	22.31	49.16	91.17	85.00	395.24	111.29	286.92
B	675	36.50	28.05	53.41	82.56	70.73	329.42	132.75	325.54
C	574	16.16	16.11	32.89	33.73	32.91	241.81	88.64	96.86
D	564	16.18	18.00	38.33	72.39	46.53	357.23	112.33	141.46
E	510	18.46	17.30	35.50	61.28	47.60	332.59	102.24	154.56
F	454	66.76	38.32	61.73	94.66	69.36	333.99	209.30	124.15
G	361	36.75	16.55	35.57	55.03	32.98	273.69	133.60	101.26
H	327	49.92	18.66	37.12	35.00	27.26	212.39	92.57	69.97
Mean		32.32	21.91	42.96	65.73	51.55	309.55	122.84	162.59
SD		0.165	0.068	0.092	0.209	0.186	0.542	0.341	0.821

The method based on the k-means algorithm is relatively easy to apply and understand. However, it has the weakness of being more appropriate for data with a convex structure, a premise challenging to be true for large data sets. The k-means algorithm tends to present good values for data partitioning quality measures developed for convex data, such as the Calinsk-Harabask and Davies-Bouldin indices (Table 5).

The proposed clustering method presented groups with less variance per variable

Table 5. Comparing the three algorithms using clustering quality measures (best results highlighted). Source: elaborated by the authors.

Quality measure	Proposed method	GLMM MCMC	k-means
Calinski-Harabatz	195.03	401.12	567,82
Davies-Bouldin	1.05	0.60	0.84

per cluster when compared to the k-means algorithm, which denotes a greater capacity to capture the complexity of the dataset with a non-convex structure. The characterization of the homogeneous groups in the neural grid generated in step 4, the visual analysis of the Component Plans generated in step 5, and the interpretation of the trajectories of each municipality in the neural grid add essential explanatory elements during the clustering process.

4. Conclusions

The choice to analyze the Coefficients of Variation (CV) by variable and by group seemed to be an appropriate strategy for comparing the algorithms compared to quality measures such as Davies-Bouldin and Calinsky-Harabatz indices that are more appropriate for convex data.

The proposed method presented lower CV with lower dispersions (standard deviation) when compared to the k-means method. The fair performance in terms of the CV of the model-based method ended up being hampered by the concentration of municipalities in a single group, which also compromised their spatial distribution.

The proposed method and the k-means algorithm presented different but compatible results regarding the spatial distribution of the groups in the five regions of Brazil. Both show that there is substantial homogeneity in the NE region, that the states of Pará, Mato Grosso, Goiás, and the Tocantins have similarities and that Mato Grosso do Sul is more similar to the southern states than to the Midwest. The differences between the two strategies stand out more in the Southeast region.

In order to improve the robustness of the proposed method, its application in datasets with distinct structures is mandatory, and the use of non-convex data partition validation measures. Furthermore, an investigation of outliers for each method should also be carried out.

Acknowledgement

This work was carried out with the support of the CAPES and FAPITEC through notice no. 04/2019 PBIC/FAPITEC/SE/FUNTEC/CAPES.

References

Augustijn, E. W. and Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics*, 12(1).

Chen, I.-T., Chang, L.-C., and Chang, F.-J. (2018). Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. *Journal of Hydrology*, 556:131–142.

da Silva, M. A. S., Matos, L. N., and de O. Santos, F. (2021). Data and R scripts - Self-Organizing Map approach to cluster Brazilian agricultural spatiotemporal diversity. `https://github.com/marcossantos-silva/SOMPanelData`.

Dessie, A., Abate, T., Mekie, T., and Liyew, Y. (2019). Crop diversification analysis on red pepper dominated smallholder farming system: evidence from northwest Ethiopia. *Ecological Processes*, 8(50).

Genolini, C., Alacoque, X., Sentenac, M., and Arnaud, C. (2015). Kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4):1–34.

IBGE (2020). Sistema IBGE de recuperação automática. Available at `https://sidra.ibge.gov.br` (2021/06/15).

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

Komárek, A. and Komárková, L. (2013). Clustering for mmultivariate continuos and discrete longitudinal data. *The Annals of Applied Statistics*, 7(1):177–200.

Komárek, A. and Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, 59(12):1–38.

Ling, C. and Delmelle, E. (2016). Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach. *Annals of GIS*, 22(3):173–186.

Qi, J., Liu, H., Liu, X., and Zhang, Y. (2019). Spatiotemporal evolution analysis of time-series land use change using self-organizing map to examine the zoning and scale effects. *Computers, Environment and Urban Systems*, 76:11–23.

Sambuichi, R., Galindo, E., Pereira, R., Constantino, M., and Rabetti, M. (2016). Diversidade da produção nos estabelecimentos da agricultura familiar no Brasil: uma análise econométrica baseada no cadastro da declaração de aptidão ao PRONAF (DAP). Technical report, Brasília: Rio de Janeiro.

Schneider, S. and Cassol, A. (2014). Diversidade e heterogeneidade da agricultura familiar no Brasil e algumas implicações para políticas públicas. *Cadernos de Ciência & Tecnologia*, 31(2):227–263.

Shannon, E. (1948). Mathematical theory of communication. *The Bell System Technical Journal*, 28(4):656–715.

Skupin, A. and Hagelman, R. (2005). Visualizing demographic trajectories with self-organizing maps. *GeoInformatica*, 9(2):159–179.

Teixeira, M. and Ribeiro, S. (2020). Agricultura e paisagens sustentáveis: a diversidade produtiva do setor agrícola de Minas Gerais, Brasil. *Sustainability in Debate*, 11(2):29–41.

Wang, N., Biggs, T., and Skupin, A. (2013). Visualizing gridded time series data with self organizing maps: An application to multi-year snow dynamics in the northern hemisphere. *Computers, Environment and Urban Systems*, 39:107–120.