Tracking the Connection between Brazilian Agricultural Diversity and Native Vegetation Change by a Machine Learning Approach

Marcos Aurélio Santos da Silva, Leonardo Nogueira Matos, Flávio Emanuel de Oliveira Santos, Márcia Helena Galina Dompieri, and Fábio Rodrigues de Moura

***Abstract*—In Brazil, agribusiness has a considerable role in the country's GDP. Because of this, the State needs territorial planning to minimize the impacts on natural resources, especially in the Pantanal and Amazon biomes, where agribusiness has expanded. The lower the agricultural diversification, the lower the pattern of land use homogeneity, generally associated with agribusiness, especially when it occupies large areas with more technological productive units. This paper investigates the relationship between spatial diversification patterns and the dynamics of native vegetation in Brazil. We propose a feature engineering and clustering approach for 5570 Brazilian municipalities between 1999 and 2018. It was based on the unsupervised artificial neural network Self-Organizing Map (SOM) to divide the municipalities into homogeneous groups of agricultural products diversity trends. The results were compared with the change in vegetation area using data from the national land use-mapping project called Mapbiomas. The analysis allowed the identification of three different regimes of modification in native vegetation, particularly related to municipalities in Brazil's Midwest and North regions, indicating substantial changes in the Cerrado and Amazon biomes.**

***Index Terms*—Clustering, Self-Organizing Maps, Shannon's entropy, Spatial panel data, Sustainability.**

I. Introduction

Brazilian agriculture has shown steady growth in productivity and production due to the expansion of the cultivated area. For instance, temporary crops grew 65% between 1999 and 2018, and soybean 166% in the same period [1]. This trend reinforces the agricultural sector as one of the most important for the Brazilian economy and, at the same time, the economic activity that most affects the Brazilian land use and change, mainly the sustainability in terms of native vegetation variation [2]. There is a general trend towards the specialization of agriculture due to scale gains, established market networks, and increased international demand for commodities. Consequently, extensive agriculture puts pressure on the environment, especially natural vegetation, soil, and water resources.

Marcos Silva is a researcher at Embrapa Coastal Tablelands, e-mail:marcos.santos-silva@embrapa.br.

Leonardo Matos and Flávio Santos are with Computer Science Department, Federal University of Sergipe (UFS) e-mail:leonardo@dcomp.ufs.br, flavioemanuel859@gmail.com

Márcia Dompieri is a researcher at Embrapa Territorial e-mail:marcia.dompieri@embrapa.br

Fabio Moura is with Economics Department, Federal University of Sergipe (UFS) e-mail:fabiromoura@gmail.com

Studies have demonstrated that a diversified agricultural production system is fundamental for farmers and society. Diversity makes smallholder systems more adaptive and resilient at the local scale, improves eating habits, increases family income, and promotes the overall reduction of poverty and unemployment rates [3]–[7]. After analyzing family farming data, Sambuiche et al. (2016) noted that the lower the income, the greater the importance of diversity for food security [5].

At the landscape level, the preponderance of monocultures impacts the conservation of biodiversity and natural resources [8]. Tisdell et al. (2019) demonstrated that low-diversification agriculture threatens the agricultural systems themselves [9]. Teixeira and Ribeiro (2020) established a positive correlation between the diversity of municipal agricultural production and vegetation conservation (forest fragments) in Minas Gerais, Brazil [10]. Sambuichi et al. (2016) recommended developing public policies to promote agricultural diversity on a regional scale, including incentives for conservation practices and landscape diversification [5].

However, Brazilian agricultural activities present a huge spatial diversity due to economic and historical processes, challenging territorial public policies design [11]. Thus, finding hidden patterns in spatial data about production diversity becomes essential to support effective public incentives to promote sustainable agriculture considering regional particularities.

This study investigates the relationship between spatial agricultural production diversification patterns and the suppression or increase of native vegetation. We construct eight diversity indicators based on Shannon entropy [12] for 5570 Brazilian municipalities using agricultural IBGE's annual estimates [13] between 1999 and 2018. The indicators cover the themes of herd population, planted area with temporary crops, production value of permanent and temporary crops, animal production, plant extraction, forestry, and aquaculture. This paper proposes a featuring engineering and a clustering analysis approach based on the unsupervised artificial neural network Self-Organizing Map (SOM) of Kohonen [14] to divide the municipalities into homogeneous groups of agricultural production diversity trends. We compared these clusters with the change in vegetation area per municipality using the data from the Brazilian national project called Mapbiomas.

This paper is organized as follows: section II presents a short review of agricultural diversity measures and the use of the SOM to cluster spatial panel data. Section III describes

the dataset, the featuring engineering (construction of the agricultural diversity indices), and the proposed clustering approach. Section IV shows the results and discussions, and the section V is dedicated to the conclusions.

II. Related work

A. Agricultural Diversity Measure

The agricultural diversity has been assessed at local [3], [4] and landscape [8], [10] levels using Simpson [15] and Herfindahl-Hirschmann [16] indices and their variants [3].

In Sambuichi et al. (2016), the authors applied the Simpson's diversity index on Pronaf Aptitude Statement (*Declaração de Aptidão ao Pronaf*–DAP) data to classify Brazilian family farmers into very diverse, diverse, poorly diverse (specialized farm), and not diversified (very specialized farm) [5]. The results suggest that three regions concentrated the specialized farms, Center-West (63%), North (56%), and Southeast (52%), and the Northeast and South regions concentrated the diversified smallholders with almost 60% of diversified and very diversified for both regions. In Dessie et al. (2019), the authors showed that diversity is crucial for smallholder farmers and that they need incentives and technical assistance to address the difficulties. The authors have chosen a modified version of the Herfindahl-Hirschmann index to evaluate the on-farm diversity [4].

The Simpson's index was also applied by [10] to classify the Minas Gerais municipalities by the mean of their cultivated area with temporary and permanent crops between 2014 and 2018 using IBGE's annual estimates. The authors defined two township profiles: the very productive municipalities with low diversity and the conservationists with a high degree of diversity, linking diversity and conservation concepts.

There is a lack of studies about Brazilian agricultural diversity at a landscape level that consider all range of activities such as aquaculture, silviculture, vegetal extractivism, herd population, animal origin products, and permanent and temporary crops. In general, the literature shows a trend for agricultural specialization with low levels of diversification for well-established farmers and even small ones [3]–[6]. However, this leads to environmental pressure due to misuse of natural resources and deforestation. Studies also show that it is necessary to design public policies to promote agricultural diversification, mainly on small farms, and diminish vegetation loss.

B. Spatiotemporal Clustering with Self-Organizing Map

Self-Organizing Map (SOM) is a vector quantization machine learning algorithm used to order multivariate data into a low dimensional grid that can be used for data projection, compression, and clustering.

As proposed by [17], there are at least three main strategies to cope with spatial panel data: data containing time-series observations of spatial units. First, we can use one neural network for each year and analyze the temporal patterns independently [18]. Second, we can transform spatial panel data into a wide one and use only one neural network to observe the temporal pattern [17], [19]. Third, we may consider each observation-year as one input vector and observe what trajectory is generated on the neural grid by chronologically linking each observation-year on the neural map [20]–[23].

The spatial dimension of the dataset can be explicitly included in the feature vector as proposed by [24]. However, it assumes stationarity of spatial dependence when it is not valid in our case due to the concentration of the municipalities in some Brazilian regions. Furthermore, the spatial proximity matrix between observations can be a constraint, as suggested by [25], [26] as it imposes a constraint appropriate for regionalization purposes but not necessarily for an exploratory one. Then, it is preferable to verify spatial patterns after the clustering process by mapping the cluster into a geographic map and checking for global and local spatial dependencies [19]–[21], [23].

Wang, Biggs, and Skupin (2013) expanded the visual analytic potential of SOM for climate research with conceptual, computational, and visual transformations to find patterns on microwave imagery of snow water equivalent gridded data [23]. A combination of SOM's results, Sammon's projection, and GIS to analyze the dynamics of a spatiotemporal disease (measles outbreaks) diffusion pattern were applied by [22].

Ling and Delmelle (2016) used a spatiotemporal data handling method that clusters trajectories, i.e., changes in coordinates in the SOM distance matrix over time, to automatically cluster and classify urban neighborhoods [21]. Qi et al. (2019) decided to use a combination of the second and third strategies to identify spatiotemporal change patterns of the evolution of land use and change in Beijing from both gridded and aerial data [19].

Hence, considering this short review and our dataset, this study will not explicitly consider the spatial component in the clustering process. We will observe the temporal pattern tracking the trajectories on the neural map by a clustering process as proposed by [17], but incorporating an automatic trajectory clustering as [21], and using a small size neural grid as [22].

III. Data and Methods

A. Raw Data and Diversity Indices

The raw data comprises eight different categories of IBGE's estimates variables, each for all (5570) Brazilian municipalities between 1999 and 2018 [13] as spatial panel data. They are of various types, such as counting (heard population, including dairy animals), area in hectares (planted area with temporary crops), and production value in Brazilian currency (animal origin, temporary and permanent crops, vegetal extractivism, and forestry). Table I shows all categories and a statistical summary for them.

For example, the herd population category has $m = 11$ variables: cattle, buffalo, equine, swine (total), swine (matrices), goats, sheep, poultry(total), poultry(chickens), quail, and dairy animals. For each set of raw data, we have applied a diversity index (Eq. 1) based on Shannon's entropy [12]. This measure have been adopted 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.

TABLE I
STATISTICAL SUMMARY FOR ALL EIGHT DIVERSITY INDEXES. SOURCE: ELABORATED BY THE AUTHORS.

Category	Diversity index	Median / Max	Mean ± sd	m
Herd population + dairy animals	DIV.HERD	0.52 / 0.87	0.480 ± 0.180	11
Animal production value	DIV.VL. PRODANI	0.20 / 0.76	0.220 ± 0.150	6
Temporary crop production value	DIV.VL.T	0.30 / 0.69	0.290 ± 0.130	31
Permanent crop production value	DIV.VL.P	0.20 / 0.73	0.200 ± 0.160	36
Temporary crop planted area	DIV.PLANT.T	0.32 / 0.61	0.300 ± 0.086	31
Aquaculture production value	DIV.AQU.VL	0.00 / 0.66	0.026 ± 0.085	24
vegetal extractivism production value	DIV.EXTV.VL	0.04 / 0.47	0.088 ± 0.100	43
Forestry production value	DIV.SILV.VL	0.14 / 0.75	0.084 ± 0.140	15

$$DIV_l = -\sum_{i=1}^{m}\left[\frac{y_i}{\sum_{j=1}^{m} y_j} log_m\left(\frac{y_i}{\sum_{j=1}^{m} y_j}\right)\right] \quad (1)$$

where m is the number of raw variables for the category l and y_i is the value of the ith raw variable for each year, category, and municipality. The diversity index DIV values vary from zero (without diversity) to one (highest diversity). All unavailable raw data have been replaced with zeros, which means no agricultural diversity is present.

Table I shows that the index DIV.HERD presents the highest mean, median, and max values, DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL presents high levels of Coefficient of Variation. It is a worthing note that the estimates for Aquaculture started only in 2012 and that there are a lot of municipalities without vegetal extractvism and forestry production. Zero is the minimum value for all diversity indices, and they are not normally distributed according to the Kolmogorov-Smirnov test.

Fig. 1 shows a spaghetti graph of all observations with the mean curve (red) to all diversity indexes. All indexes show a slowly decreasing diversity trend (DIV.HERD, DIV.VL.T, DIV.PLANT.T with atypical behavior in 2005 and 2015, DIV.EXTV.VL) or a slowly increasing diversity trend (DIV.VL.PRODANI , DIV.VL.P, DIV.AQU.VL from 2012). The index DIV.SILV.VL shows a cyclical behavior of seven years with some trends to increase diversity.

We calculated the diversity index for the categories harvested area for temporary and permanent crops and cultivated area for permanent crops. Still, they presented a high correlation with other variables, and we excluded them from the study.

All raw data, diversity indices, and metadata are available in [27].

B. Spatial Panel Data Visualization and Clustering using Self-Organizing Maps

Despite the wide application of Self-Organizing Maps to the clustering task, its use on spatial panel data is still scarce. In fact, for each dataset, it is necessary to define or refine a

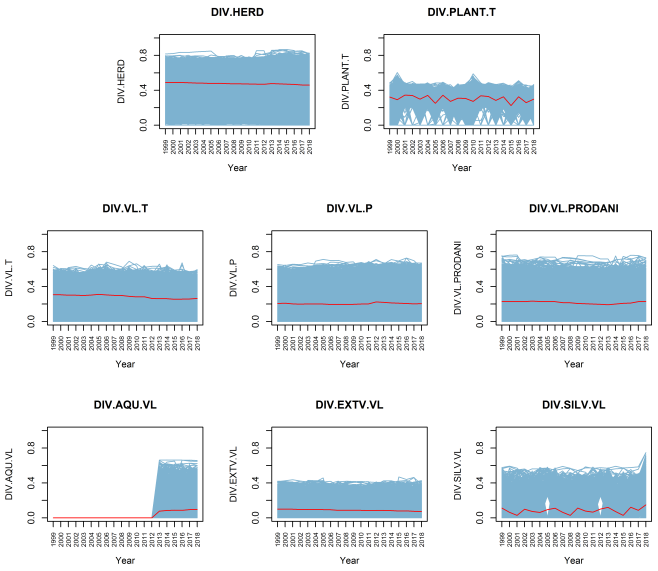


Fig. 1. Spaghetti graph for all diversity indexes highlighting the mean along the years. Source: elaborated by the authors.

method to proceed with the clustering process. Therefore, this work proposes an approach to cluster the spatial panel data for the Brazilian agricultural diversity indices, and it comprises seven steps (Fig. 2).

1) Steps 1 and 2 - Feature engineering: Steps 1 and 2 consists of obtaining the attributes that will represent the data. These steps were detailed in Sec. III-A.

2) Step 3 - Spatial panel data ordering on the Self-Organizing Map (SOM): The third step consists of spatial panel data ordering using the unsupervised artificial neural network called Kohonen Self-Organizing Map [14]. It projects a d-dimensional dataset, represented by $\mathbf{x}_i, i = 1, ..., n$ where n represents the number of observations, into a two-dimensional $M \times N$ grid composed of a discrete number $m = M * N$ of artificial neurons by a stochastic machine learning process. Each neuron j has a weight vector $\mathbf{w}_j$ associated to it, also in space $\mathbb{R}^{\mathrm{d}}$ [14]. The neurons are arranged as a hexagonal grid because it increases the number of neighbors, improving the machine learning quality and generating good maps, as stated by [14, p. 159].

The number m of neurons depends on the volume and complexity of the data and is determined empirically from the combined observation of quantization error (Eq. 3) and the analysis of the projections of the observations on the neural grid after the learning process. The learning algorithm keeps the original data topology, so as m grows more detailed becomes the data partitioning [14]. Thus, the SOM's learning processes guarantee a robust data ordering that generates consistent data partitioning for different neural network sizes. In this work, we evaluated a set of medium size SOMs as used by [22]: 10x15, 20x25, 25x30, 30x35, 50x60 and 80x100. We compared the quantization error and the data projection by clustering with k-means (see step 4) and observing the Component Planes (see step 5).

The SOM iterative machine learning process consists of

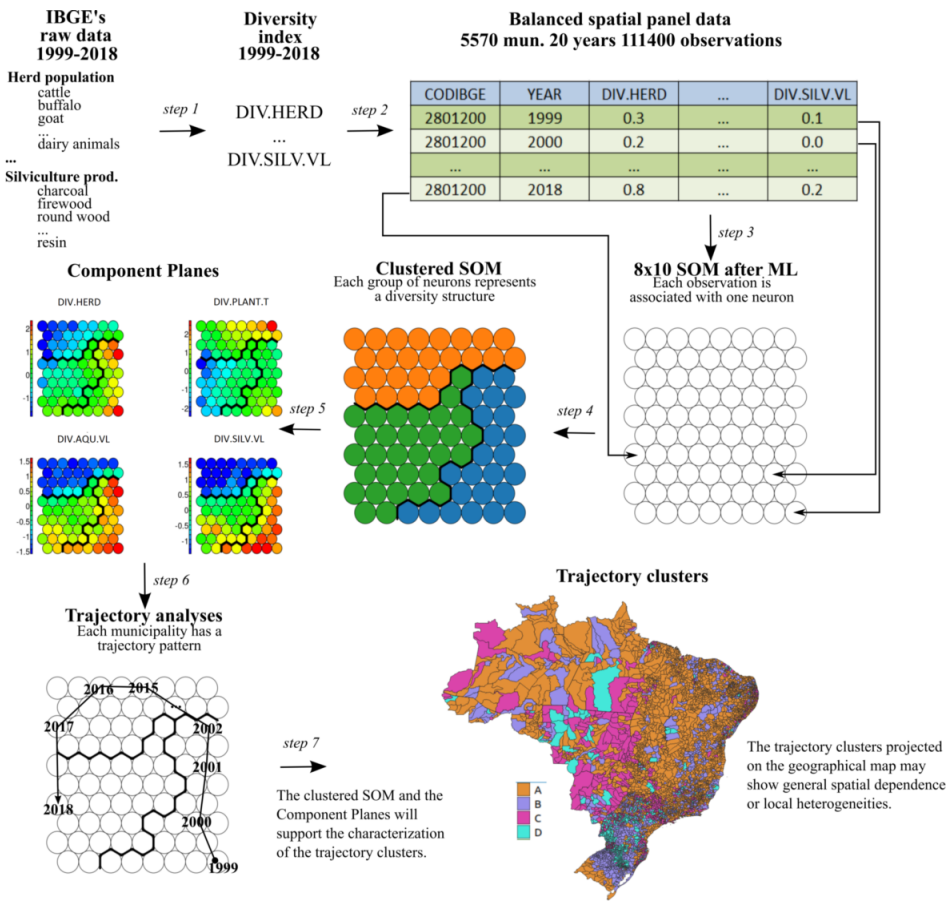


Fig. 2. Featuring engineering and clustering analysis of spatial panel data clustering using the Self-Organizing Map. Source: elaborated by the authors.

three phases. In the first phase, *competitive*, each input vector $\mathbf{x}_i$ searches for the nearest neuron, according to the Euclidean distance, and this is considered the Best Match Unit (BMU). In the second phase, *cooperative*, the BMU and its neighbors are defined according to the Gaussian function. In the last phase, *adaptive*, the weights of the BMU and its neighborhood are adjusted according to the Eq. 2 to approximate them to the input vector $\mathbf{x}_i$. The algorithm iterates T times, previously defined and generally greater than n, and the weights are linearly initialized according to [14, p. 142].

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)h(t)\|\mathbf{x}_i(t) - \mathbf{w}_j(t)\|_2 \qquad (2)$$

where $\alpha(t)$ is the learning rate function at time t and $h(t)$ is the Gaussian neighborhood function centered on the winning neuron (BMU). The quantization error, Eq. 3, can measure the quality of the mapping process, Eq. 3, which represents the mean of all distances between each input vector $\mathbf{x}_i$ and the weights of its BMU $\mathbf{w}_{i,BMU}$.

$$E_q = \frac{\sum_i^n \|\mathbf{x}_i, \mathbf{w}_{i,BMU}\|_2}{n} \qquad (3)$$

The machine learning process was conducted using the SOM PAK software [28], while the computation of the other steps were accomplished utilizing the Kohonen R package [29].

3) Step 4 – Clustering the SOM's weights: In this step, we have clustered the SOM weights using the k-means method, with support of the elbow method, and the Silhouette quality index analysis to help define the number k of groups. After clustering, a comparison of means for each diversity is conducted to verify if they are statistically distinct between clusters using the Kruskal-Wallis non-parametric test, considering a significance level of 0.05. This clustering will also help the interpretation of the Component Planes generated from the SOM weights by dividing the neural grid into regions with homogeneous characteristics and facilitating the visual inspection.

4) Step 5 – Component Planes analysis: To determine how each component of the feature vector $\mathbf{x}_i$ was organized in the trained map, a coloring method based on the values of each component is used. For a given $j-th$ component of the SOM's weights, an image is generated with dimensions equal to those of the map $M \times N$, where each pixel will correspond to the value of the j component at the position (a, b) on the neural map using a divergent palette pattern (dark blue represents minimum values, dark red maximum 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 [17], [19], [22].

5) Step 6 – Trajectory analysis: 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 [21]. A trajectory for a municipality p can be expressed as a matrix $Traj_{ij}^{p}$ where each row corresponds to a coordination (a, b) on the neural grid. Hence, to cluster all trajectories, it has been applied a k-means algorithm using the Eq. 4 as a matrix distance measure [30]. The number c clusters will be defined with the support of the Davies-Bouldin and Calinski-Harabatz quality indices, also implemented in [30].

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

6) Step 7 – Projection on the geographic map: In this step, we will map the clusters on the geographic map to observe spatial dependence and spatial heterogeneities as proposed by [19], [21] and verify if the distribution of groups follows any regional or local spatial pattern.

This approach was compared by [31] with a k-means for panel data [30], and a model clustering algorithm based on Generalized Linear Mixture Model [32]. The proposed approach showed better results considering the Coefficient of Variation for each variable per cluster as a degree of cluster homogeneity. In addition, using SOM allows data exploration and cluster explanation by visual inspection tools such as Component Planes and trajectory analysis onto the neural map.

C. Evaluating the Vegetation Change

To analyze the relationship between agricultural diversity and variation in native vegetation, we will focus on municipality land-use change based on the "Mapbiomas Collection 6" Land Use and Change database [33]. This project monitors the municipality's land use and changes through time (1985-2020), classifying each spatial unit into five general classes: forest (class 1), non-forest natural formation (class 2), farming (class 3), non-vegetated area (class 4) and water (class 5). This data is published aggregated by area (municipality), and the total area for each class is expressed in hectares.

For each municipality i we can define a variable rt_veg_i that represents the rate of vegetation (classes 1 and 2) per municipality considering all mapped area (Eq. 5). The vegetation rate variation, var_veg_i between 2000 and 2019, also a 20 years interval, is given by the Eq. 6. The lower the var_veg_i, the greater the negative impact of the land use, including very specialized and diverse farming.

$$rt\_veg_{i,year} = \frac{total\_class_{1,i,year} + total\_class_{2,i,year}}{total\_all\_classes_{i,year}} \quad (5)$$

$$var\_veg_i = \frac{rt\_veg_{i,2019} - rt\_veg_{i,2000}}{rt\_veg_{i,2000}} \quad (6)$$

After the trajectory clustering of municipalities by the proposed method, a Kruskal-Wallis means test will be applied to verify if each trajectory cluster c represents a statistically different profile according to the var_veg variable.

IV. Results and Discussion

A. Data Ordering by Machine Learning (step 3)

The neural input corresponds to a vector with eight diversity values (DIV.HERD, ..., DIV.SILV.VL) for each municipality and year, and it comprises 111440 observations that were not normalized. We tested six SOM neural networks, varying the SOM's dimensions ($M \times N$) and observing the quantization error (E_q): 10×15 ($Eq = 0.20$), 20×25 ($Eq = 0.17$), 25×30 ($Eq = 0.16$), 30×35 ($Eq = 0.16$), 50×60 ($Eq = 0.15$) and 80×100 ($Eq = 0.14$). The SOM's training was conducted using the sequential strategy, with 10^7 iterations.

We observed that there were not a noticeable drop of quantization error when we increased the size of the neural network. We also applied the k-means to the neural network weights combined with the elbow method and silhouette quality index. We found $k = 5$ for five of them, suggesting the robustness of the SOM that projected the input data into the neural grid preserving its original topology and changing only the "resolution" of this mapping. Then, considering it and the high computational cost of big neural networks, we eliminated the neural network candidates 50×60 and 8×100. Observing the quantization error, we eliminate the candidates 10×15 and 20×25. So, as the neural networks 25×30 and 30×35 present the same quantization error and weights partition ($k = 5$), we have chosen the smallest one, the 25×30 2D-SOM.

B. Homogeneous Regions on the Neural Map (Steps 4 and 5)

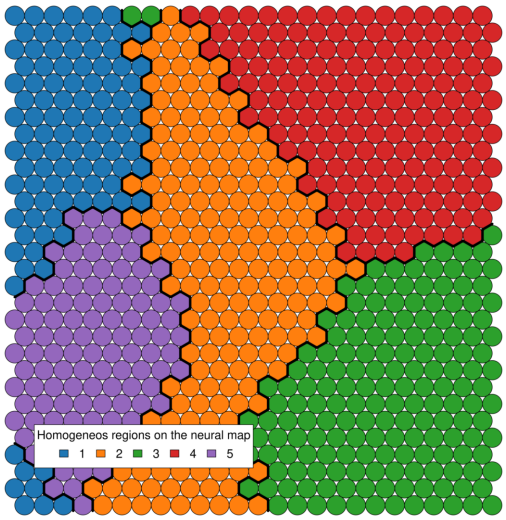
The 25×30 SOM was segmented using the k-means algorithm, the elbow curve, and the Silhouette index to define k equal to five. These homogeneous regions onto the neural map, Fig. 3a, were characterized by the interpretation of the Component Planes, Fig. 3b, and Table II where the mean and median for each variable can be found, as well as the means comparison test.

The Component Planes show that high diversity for aquaculture (DIV.AQU.VL), vegetal extractivism (DIV.EXT.VL), and silviculture (DIV.SILV.VL) are mutually exclusive. It suggests that these activities compete with each other or demand very specified and different edaphoclimatic and landscape conditions.

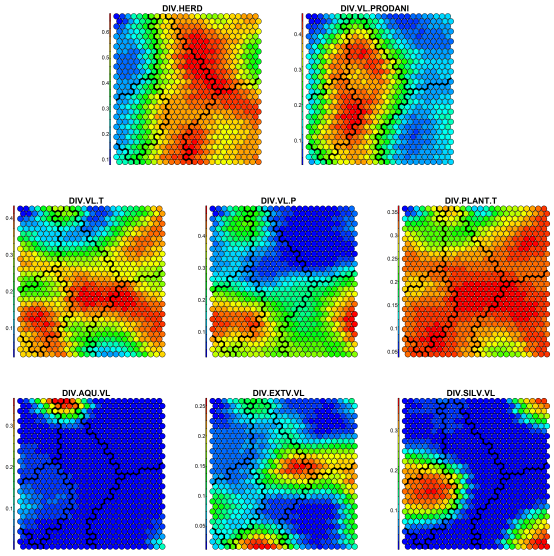
Comparing the means of each group from Table II and the general means from Table I, it is possible to describe a group profile based on mean values greater or less than the values of Table I. The homogeneous region one on the neural map is associated with low diversity for DIV.HERD, DIV.VL.T and DIV.PLANT.T; region 2 with high diversity for DIV.HERD and low diversity for DIV.VL.PRODANI and DIV.VL.P; region 3 with high diversity for DIV.VL.PRODANI, DIV.VL.T and DIV.VL.P, and low diversity for DIV.HERD; region 4 with high diversity for DIV.HERD, DIV.VL.P, and low diversity for DIV.VL.PRODANI; and region 5 with high diversity for DIV.HERD and DIV.VL.PRODANI.

C. Municipal Clustering by SOM's Trajectory Analysis (Step 6)

As described in section III, there will be a trajectory onto the neural map for each municipality mapped by coordinates



(a) Segmented 25x30 SOM's neural map by applying the k-means algorithm on the neural weights ($k = 5$)



(b) A Component Plane for each diversity index made from the neural map weights

Fig. 3. Segmented neural map and Component Planes for the 25x30 SOM. Source: elaborated by the authors.

TABLE II
MEAN AND STANDARD DEVIATION FOR EACH DIVERSITY INDEX AND SOM'S HOMOGENEITY REGION. SOURCE: ELABORATED BY THE AUTHORS.

Diversity index mean* (median)	SOM's homogeneous regions k (# of observations)				
Number of obs. per region	1 (15922)	2 (26939)	3 (13595)	4 (28229)	5 (26715)
DIV.HERD	0.20a (0.20)	0.54b (0.55)	0.32c (0.31)	0.57d (0.58)	0.56e (0.56)
DIV.VL. PRODANI	0.18a (0.15)	0.12b (0.11)	0.34c (0.35)	0.13d (0.12)	0.37e (0.36)
DIV.VL.T	0.22a (0.23)	0.26b (0.28)	0.34c (0.35)	0.31d (0.32)	0.30e (0.32)
DIV.VL.P	0.17a (0.18)	0.05b (0.00)	0.34c (0.35)	0.31d (0.35)	0.20e (0.19)
DIV.PLANT.T	0.26a (0.28)	0.29b (0.31)	0.32c (0.34)	0.32d (0.33)	0.32e (0.33)
DIV.AQU.VL	0.04a (0.0)	0.02b (0.0)	0.04ab (0.0)	0.02c (0.0)	0.03d (0.0)
DIV.EXTV.VL	0.14a (0.0)	0.08b (0.03)	0.07c (0.0)	0.08b (0.03)	0.14d (0.14)
DIV.SILV.VL	0.07a (0.0)	0.06b (0.0)	0.25c (0.25)	0.10d (0.0)	0.02e (0.0)

*For each diversity index, non statistically different cluster means by Kruskal-Wallis test, with a confidence level of 95%, are indicated with the same letter.

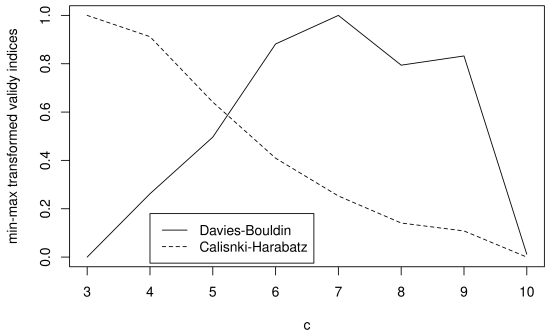


Fig. 4. Transformed validity indices into $[0, 1]$ interval, zero meaning bad partition and one the best one considering the evaluated number of clusters. Source: elaborated by the authors.

(a, b) of the chronological path from 1999 to 2018.

To define the c number of trajectory clusters, we used the Calinski-Harabatz and Davies-Bouldin validity indices, Fig. 4, and observed the number of municipalities by clusters searching for a partition that explicitly differentiates the Brazilian regions. The monotonically decreasing curve of the Calinski-Harabatz index while c increases denotes that the clustering quality also increases with c. The Davies-Bouldin curve shows that the candidate c should be between six and nine. Then, we have chosen that $c = 8$ divided the municipalities in a way to unveil the intra-regional and inter-regional heterogeneities and also partitioned the dataset into a balanced set of groups (see the second column of Table III).

To represent these eight clusters on the neural map, we have plotted the mean and the median trajectory for each one (Fig. 5). There are six mean trajectory clusters that are completely located at one of the homogeneous regions on the neural map; cluster A is located at region 5, cluster B at region 4, clusters C and D at region 2, cluster E at region 5, cluster F at region 1. Moreover, two clusters (G and H) present some displacement onto the neural map. The same occurs to the median trajectory, but with greater displacements for the clusters G and H. The first group of clusters represents municipalities that do not tend to change their diversity profiles between 1999-2018. The second group suggests that some municipalities migrate from

one diversity profile to another.

The trajectory **cluster A** is associated with region three on the neural map and represents municipalities with herd population diversity below the global mean, but with animal, temporary, and permanents crops production values above the global mean. The trajectory **cluster B**, associated with the neural region 4, presents and maintains through time a high diversity of herd population and production value for temporary and permanent crops, associated with a low diversity for animal production value.

The trajectory **clusters C and D** share the same region two on the neural map and represent a set of municipalities with similar characteristics as herd population diversity above the global mean and animal and permanent crops production values below the global mean. The trajectory **cluster E** is associated with region five on the neural map and gathers municipalities with high diversity for herd population and animal production value.

The trajectory **cluster F**, associated with region one on the neural map, represents the lowest levels of diversity for herd population, planted area with temporary crops, and animal and temporary crop production values. The trajectory **cluster G** represents a displacement from the SOM's homogeneous region three toward region 1. It implies that these municipalities are decreasing their diversity towards the characteristics of cluster F. The trajectory **cluster H** also presents a displacement from the region 4 to 3, suggesting that these municipalities are decreasing their herd population diversity, increasing the animal production value diversity, but keeping the level of diversity for temporary and permanent crops production values.

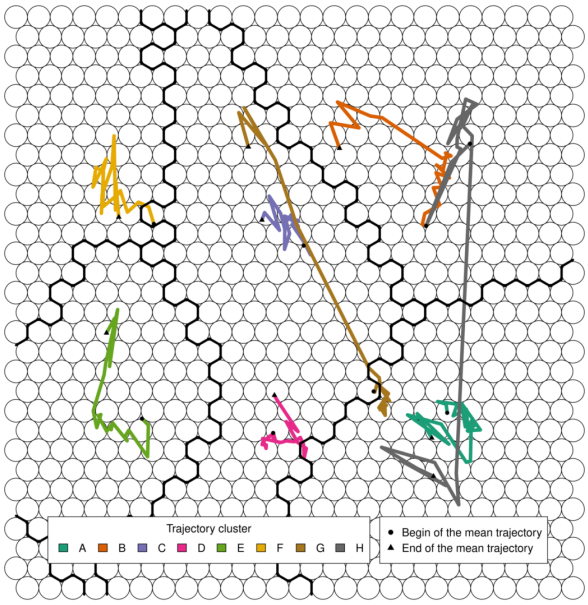
D. Geographic Mapping of the Trajectory Clusters (Step 7)

Fig. 6 shows the spatial distribution of the trajectory clusters by state and region. Table III indicates that, in the North region, it predominates municipalities associated with trajectory cluster G, 25.33%, representing a trend in changing their diversity profile toward less diversity. The cluster B is the most frequent in the Center-West region, 40.47%; in the Northeast, we observe the predominance of cluster D, 24.19%; in the South region, the trajectory cluster E, 34.09%; and in the Southeast is the trajectory cluster A. As observed by [5], there is a regional agricultural diversity pattern, but remarkable intra-regional differences exist.

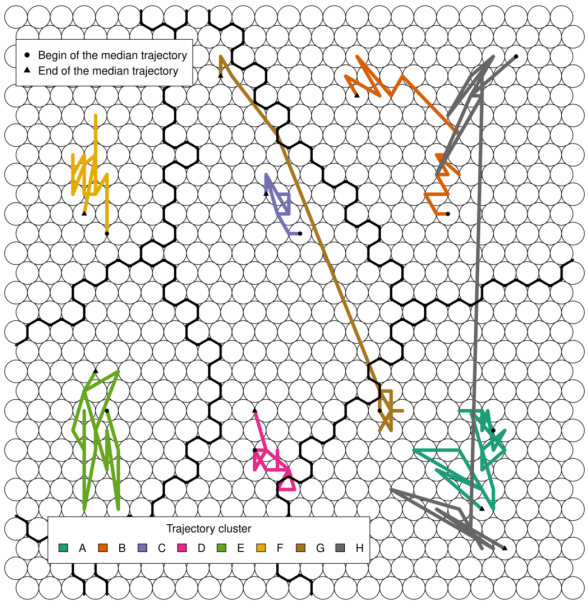
For example, the predominance of the trajectory cluster A in the Southeast region is mainly due to the Minas Gerais state, which presents a remarkable diversity in the agricultural pattern as observed by [10]. The trajectory cluster B predominates in the Center-West region, but it concentrates in two states, Tocantins and Goiás. The Amazon and Roraima states present a particular spatial pattern in the trajectory cluster distribution compared with the entire North region.

E. Linking Trajectory Clusters and Vegetation Change

Fig. 7a shows the spatial distribution for the change in the rate of vegetation (var_veg) between 2000 and 2019. The red region on the map denotes an intense vegetation loss, observed mainly in the Center-West and North states and some areas of



(a) Representation on the neural map of each trajectory cluster mean



(b) Representation on the neural map of each trajectory cluster median

Fig. 5. Representation on the neural map of each trajectory cluster mean and median. Source: elaborated by the authors.

Northeast, Minas Gerais state, and South region. The green regions denote relative vegetation growth in some states from South to Northeast.

The last column of Table III shows the average change in the vegetation rate for each trajectory cluster. It shows that the trajectory clusters B and G present the lowest (and statistically similar according to the Kruskal-Wallis test) means for the variation in vegetation rate. The trajectory clusters A, F, and

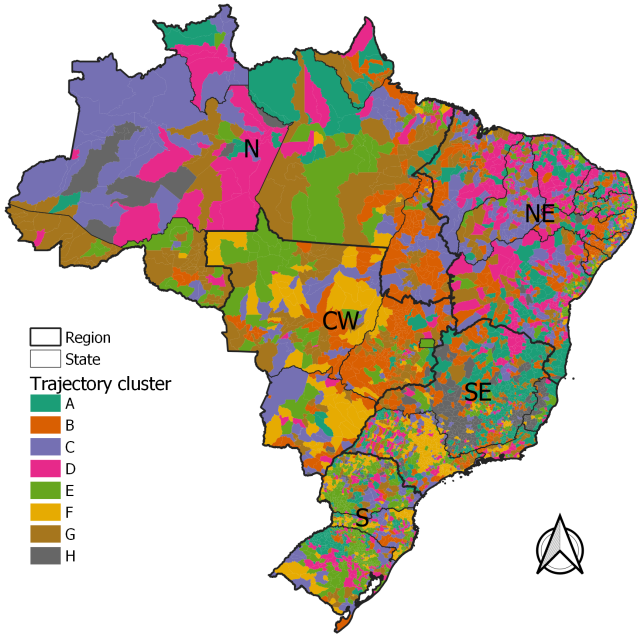


Fig. 6. Spatial projection of the trajectory clusters. Source: elaborated by the authors.

TABLE III
DISTRIBUTION OF THE TRAJECTORY CLUSTERS (c) ON BRAZIL'S REGION AND THE MEAN FOR THE VAR_VEG VARIABLE FOR EACH ONE. SOURCE: ELABORATED BY THE AUTHORS.

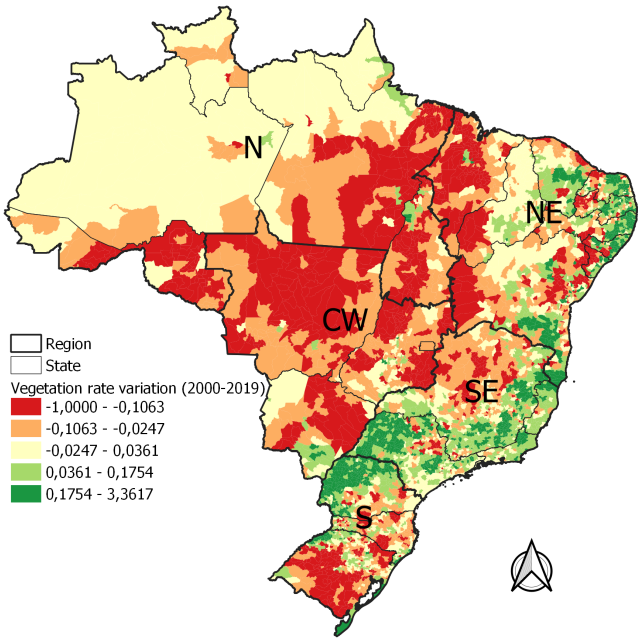
c	#*	Percentage per region (%)					var_veg
		CW	N	NE	S	SE	(mean)**
A	832	3.21	5.78	15.33	13.52	21.28	0.0811a
B	823	40.47	24.22	14.49	3.78	13.19	-0.0158b
C	776	10.49	20.00	21.29	10.33	7.91	0.0355c
D	748	3.00	9.33	24.19	10.92	7.67	0.0399cd
E	720	8.35	9.11	6.13	34.09	7.43	0.0575 d
F	668	13.92	4.22	4.79	18.05	16.97	0.1030a
G	638	19.49	25.33	12.49	7.05	7.49	0.0036b
H	365	1.07	2.00	1.28	2.27	18.05	0.0810a

$*$ Number of municipalities per cluster.
$**$ Non statistically different cluster means by Kruskal-Wallis test, with a confidence level of 95%, are indicated with the same letter.

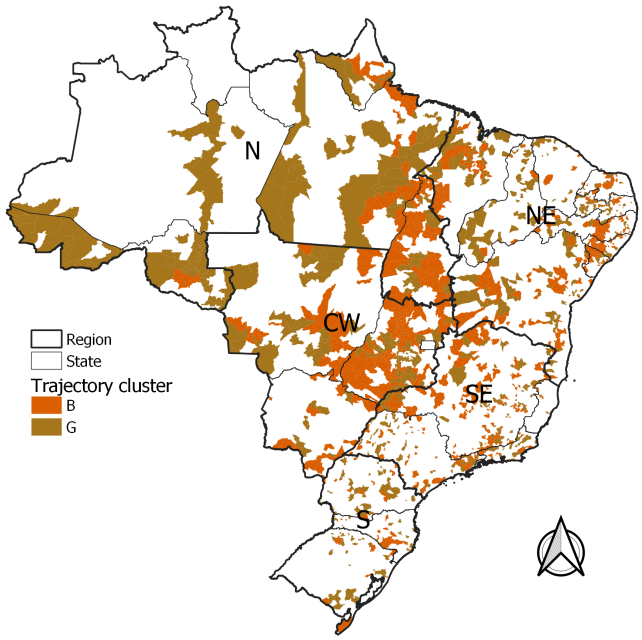
H, present the highest means for the var_veg variable and are statistically similar. The trajectory clusters C and D, and D and E present intermediary values for the change in vegetation rate between 2000 and 2019 and are statistically similar. Therefore, the eight trajectory clusters have three different regimes of vegetation variation.

The cluster G, which presents a trend to decrease diversity, confirms our hypothesis that less diversity leads to loss of vegetation. Cluster B shows the opposite: a group with municipalities with high diversity also impacts the vegetation loss, mainly in the Cerrado biome, suggesting that more studies must be conducted to identify the source of vegetation impact. Maybe these municipalities show a good landscape-level diversity but not an intra-farm one.

Fig. 7b shows only the trajectory clusters B and G in order to aid the visualization of their spatial distribution. It is clear that cluster B is mostly located in Goiás and Tocantins states, and cluster G in Pará, Rondônia, and Acre states.



(a) Spatial distribution of the vegetation rate variation using quantile (equal count)



(b) Spatial distribution of trajectory clusters B and G which are associated with the lowest values means for var_veg variable

Fig. 7. Spatial comparison between the spatial distribution of the vegetation variation and the spatial distribution of the two trajectory clusters with the lowest mean for the var_veg. Source: elaborated by the authors.

Some regions with high vegetation loss are not related to the cluster trajectories B and G. For instance, a significant portion of the Mato Grosso do Sul state presents a high negative vegetation rate, and these municipalities are more related to the trajectory cluster F, which offers low levels of agricultural diversity and, at the same time, a positive and highest mean

for vegetation variation between 2000 and 2019. The same occurs in the Pampa biome, but we have a more heterogeneous landscape here. Consequently, the municipalities are associated with different trajectory clusters such as B, C, D, E, F, and G. Again, making a direct link between vegetation loss and trajectory clusters is impossible. It could be explained by the Brazilian agricultural heterogeneity and the impact of global parameters on the interpretation of local/regional trajectory profiles. It suggests that we must conduct a regional or a more local analysis to explore the link between vegetation variation and trajectory clusters.

F. Benefits for Sustainability

The feature engineering to measure agricultural diversity can be applied at different scales. From the local level using, for instance, farm production information as proposed in [5] to another type of landscape-level using spatiotemporal satellite data to identify crop diversity as presented by [34], instead of using IBGE's data estimates aggregated by a spatial unity (municipality). It can offer very different perspectives on agricultural diversity for different purposes.

The proposed approach combines visualization techniques such as Component Planes inspection and the observation of the trajectory of each municipality onto the neural map through time. It offers an interactive way to explore agricultural production diversity, so we can use it for designing a monitor or alert system to follow land use and change associated with different diversity trends.

The territorial public policy design can take advantage of the proposed approach by using it to choose regional targets to invest in municipality production diversity or to support small farmers to improve food security, and family income as observed by [3]–[7].

G. Other Applications

The featuring engineering and trajectory clustering approach can be applied to investigate other types of diversity (e.g., industrial) and to explore any panel or longitudinal numerical data with a small number of variables, associated or not with a geographical object (e.g., census tract, district).

V. Conclusions

The proposed approach showed to be an effective clustering strategy for agricultural diversity spatial panel data, but it can be adapted to any panel data framework.

The clustering procedure partitioned the Brazilian municipalities into eight homogeneous trajectory clusters. These clusters can be split into two groups, those (six) where municipalities do not tend to change their agricultural diversity profile and those (two) with a clear trend to change. In general, trajectory clusters presented spatial dependence when projected to the geographic map.

Calculating the vegetation rate variation between 2000 and 2019 for each municipality allowed us to verify that there are three levels of vegetation rate variation for the eight trajectory clusters. The loss of vegetation is associated with clusters B and G; minimal vegetation gain is related to the trajectory clusters C, D, and E; small vegetation gain for clusters A, F, and H.

The trajectory clusters B and G, associated with loss of vegetation between 2000 and 2019, are concentrated in Center-West and North regions, covering the Cerrados and Amazon Forest biomes. Cluster G groups municipalities with a tendency to decrease agricultural diversity and confirms our initial hypothesis that less diversity means significant vegetation loss. Cluster B is associated with significant vegetation loss but presents many diversity indices above the global mean. Then, it does not confirm our initial hypothesis and suggests more work.

Future work includes a sensitivity analysis of the proposed method for different SOM's weights clustering methods; applying clustering methods based on Deep Learning to explore the complexity of the spatial panel data; and consider including economic, social, and biodiversity dimensions in analyzing the environmental impact.

Acknowledgments

This paper was carried out with the support of the Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe (FAPITEC) through public notice nº 06/2021 FAPITEC/SE/FUNTEC.

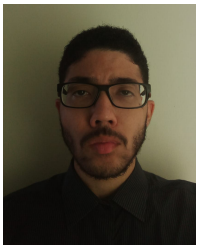
References

[1] IBGE, "Sistema IBGE de recuperação automática: tabela 1612 - culturas temporárias." Available at https://sidra.ibge.gov.br (2022/06/01), 2022.
[2] C. Klink and R. Machado, "Conservation of the Brazilian Cerrado," *Conserv. Biol.*, vol. 19, pp. 707–713, 2005.
[3] P. Fatch, C. Masangano, T. Hilger, I. Jordan, I. Mambo, J. Francesca, M. Kamoto, A. Kalimbira, and E.-A. Nuppenau, "Holistic agricultural diversity index as a measure of agricultural diversity: A cross-sectional study of smallholder farmers in Lilongwe district of Malawi," *Agricultural Systems*, vol. 187, p. 102991, 2021.
[4] A. Dessie, T. Abate, T. Mekie, and Y. Liyew, "Crop diversification analysis on red pepper dominated smallholder farming system: evidence from Northwest Ethiopia," *Ecological Processes*, vol. 8, no. 50, 2019.
[5] R. Sambuichi, E. Galindo, R. Pereira, M. Constantino, and M. Rabetti, "Diversidade da produção nos estabelecimentos da agricultura familiar no Brasil: uma análise econométrica baseada no cadastro da declaração de aptidão ao PRONAF (DAP)," tech. rep., Brasília: Rio de Janeiro, 2016.
[6] L. Pellegrini and L. Tasciotti, "Crop diversification, dietary diversity and agricultural income: empirical evidence from eight developing countries," *Canadian Journal of Development Studies*, vol. 35, no. 2, pp. 211–227, 2014.
[7] S. Schneider and A. Cassol, "Diversidade e heterogeneidade da agricultura familiar no Brasil e algumas implicações para políticas públicas," *Cadernos de Ciência & Tecnologia*, vol. 31, no. 2, pp. 227–263, 2014.
[8] M. Ogorevc and R. Slabe-Erker, "Assessment of the European Common Agricultural Policy and landscape changes: an example from Slovenia," *Agricultural economics (Praha)*, vol. 64, no. 11, pp. 489–498, 2018.
[9] C. Tisdell, M. Alauddin, M. Sarker, and M. Kabir, "Agricultural diversity and sustainability: general features and Bangladeshi illustrations," *Sustainability*, vol. 11, pp. 6004–6015, 2019.
[10] M. Teixeira and S. Ribeiro, "Agricultura e paisagens sustentáveis: a diversidade produtiva do setor agrícola de Minas Gerais, Brasil," *Sustainability in Debate*, vol. 11, no. 2, pp. 29–41, 2020.
[11] C. Sales and R. Rodrigues, "Espaço rural brasileiro: diversificação e peculiaridades," *Revista Espinhaço*, vol. 8, no. 1, pp. 54–65, 2019.
[12] E. Shannon, "Mathematical theory of communication," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1948.
[13] IBGE, "Tabelas 74, 94, 289, 291, 1612, 1613, 3939 e 3940: sistema IBGE de recuperação automática." Available at https://sidra.ibge.gov.br (2021/06/15), 2021.

[14] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer, 2001.
[15] E. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 688, 1949.
[16] A. Hirschmann, "The paternity of an index," *American Economic Review*, vol. 54, no. 761, 1964.
[17] A. Skupin and R. Hagelman, "Visualizing demographic trajectories with self-organizing maps," *GeoInformatica*, vol. 9, no. 2, pp. 159–179, 2005.
[18] M. Silva, E. Siqueira, and O. Teixeira, "Abordagem conexionista para análise espacial exploratória de dados socioeconômicos de Territórios Rurais," *Revista de Economia e Sociologia Rural*, vol. 48, pp. 429–446, 2010.
[19] J. Qi, H. Liu, X. Liu, and Y. Zhang, "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*, vol. 76, pp. 11–23, 2019.
[20] I.-T. Chen, L.-C. Chang, and F.-J. Chang, "Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps," *Journal of Hydrology*, vol. 556, pp. 131–142, 2018.
[21] C. Ling and E. Delmelle, "Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach," *Annals of GIS*, vol. 22, no. 3, pp. 173–186, 2016.
[22] E. W. Augustijn and R. Zurita-Milla, "Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns," *International Journal of Health Geographics*, vol. 12, no. 1, 2013.
[23] N. Wang, T. Biggs, and A. Skupin, "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*, vol. 39, pp. 107–120, 2013.
[24] J. Hagenauer and M. Helbich, "Hierarchical self-organizing maps for clustering spatiotemporal data," *International Journal of Geographical Information Science*, vol. 27, no. 10, pp. 2026–2042, 2013.
[25] Z. T. Luo, H. Sang, and B. Mallick, "A bayesian contiguous partitioning method for learning clustered latent variables," *Journal of Machine Learning Research*, vol. 22, pp. 1–52, 2021.
[26] L. V. Teixeira, R. M. Assunção, and R. H. Loschi, "Bayesian space-time partitioning by sampling and pruning spanning trees," *Journal of Machine Learning Research*, vol. 20, pp. 1–35, 2019.
[27] M. A. S. da Silva, L. N. Matos, F. E. de O. Santos, M. H. G. Dompieri, and F. R. de Moura, "Data and R script - Tracking the connection between Brazilian agricultural diversity and native vegetation change by a Machine Learning approach. Available at https://github.com/marcos-silva-inf/somspatialpaneldata," 2022.
[28] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "SOM PAK: The self-organizing map program package," tech. rep., Espoo: Finland, 1996.
[29] R. Wehrens and J. Kruisselbrink, "Flexible self-organizing maps in kohonen 3.0," *Journal of Statistical Software*, vol. 87, no. 7, pp. 1–18, 2018.
[30] C. Genolini, X. Alacoque, M. Sentenac, and C. Arnaud, "Kml and kml3d: R packages to cluster longitudinal data," *Journal of Statistical Software*, vol. 65, no. 4, pp. 1–34, 2015.
[31] M. A. S. d. Silva, L. N. Matos, F. E. O. Santos, F. R. Moura, and M. H. G. Dompieri, "Evaluating a self-organizing map approach to cluster a Brazilian agricultural diversity spatial panel data," in *Proceedings GEOINFO - Brazilian Symposium on Geoinformatics*, pp. 75–86, São José dos Campos: INPE, 2021.
[32] A. Komárek and L. Komárková, "Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data," *Journal of Statistical Software*, vol. 59, no. 12, pp. 1–38, 2014.
[33] Mapbiomas, "Project Mapbiomas - Collection 6.0 of Brazilian Land Cover & Use map series." Available at http://mapbiomas.org (2022/03/21), 2022. MapBiomas Project - is a multi-institutional initiative to generate annual land cover and use maps using automatic classification processes applied to satellite images.
[34] L. A. Santos, K. Ferreira, M. Picoli, G. Camara, R. Zurita-Milla, and E.-W. Augustijn, "Identifying spatiotemporal patterns in land use and cover samples from satellite image time series," *Remote Sensing*, vol. 13, no. 974, pp. 1–21, 2021.

Marcos A. S. da Silva is a researcher at the Brazilian Agricultura Research Corporation (Embrapa) and received his Ph.D. degree in Computer Science - Artificial Intelligence at the University of Toulouse 1 Capitole, France, and an MSc in Applied Computing at the National Institute for Space Research (INPE), Brazil. He had worked as Executive Coordinator of the Sergipe Geotechnology Network between 2006 and 2008 and is now a member of the editorial board of the Review of Artificial Societies and Social Simulation forum and the International Network on Complex Problems, Thought, and Systems (InComplex). His current work includes Geocomputing, and Computational Social Science applied to the study of socio-territorial and socioecological systems.

Leonardo N. Matos is an Associate Professor at the Department of Computing, Federal University of Sergipe, Brazil. His main area of interest is Machine Learning, particularly Explainable Artificial Intelligence, model compression and applications, including speech recognition and computer vision. He has several papers published in conferences and scientific journals. Some of his recent work with undergraduate students has won awards in events in Brazil. He has taught Machine Learning in the Graduate Program in Computer Science at the Federal University of Sergipe, oriented students and participated as an evaluator of master's dissertation defenses. He is a member of Brazilian Computer Society, collaborator of ISLab group of Minho University in Portugal and reviewer of papers in scientific conferences and journals in the area of Computing.



Flávio E. de O. Santos has been an undergraduate student in computer science at Federal University of Sergipe since April 2019. His main interests involve machine learning, big data and computer vision. He has an award-winning scientific work at an event in Brazil and currently is enrolled in an undergraduate research program regarding the project "Development of Deep Learning-based method for features extraction and spatial panel data clustering regarding Brazilian farming" which is developed at the Brazilian Agricultural Research Corporation.

Márcia H. G. Dompieri has a PhD in Geography from the State University of São Paulo (UNESP) since 2006. She holds a bachelor's degree in Statistics and Geography and since 2012 has been a researcher in the field of Geotechnology at the Brazilian Agricultural Research Corporation (EMBRAPA) in the line of research related to the analysis of the territorial dynamics of Brazilian agriculture.

Fábio R. de Moura received his Ph.D degreee in Economics from the Superior School of Agriculture "Luiz de Queiroz" (ESALQ) of the University of São Paulo (USP), São Paulo, Brazil in 2016. He is currently a professor with the Department of Economics and the Postgraduate Program in Economics, Federal University of Sergipe (UFS), Brazil. His research interests include agricultural economics, applied economics, and education and health economics.