Self-Organizing Map approach to cluster Brazilian agricultural spatiotemporal diversity

Flávio E. de O. Santos[1], Marcos A. S. da Silva[2], Leonardo N. Matos[3], Fábio R. de Moura[4], Márcia H. G. Dompieri[5]

[1]Centro de Ciências Exatas e da Tecnologia – Departamento de Computação – Universidade Federal de Sergipe – Aracaju, SE – Brazil

[2]Embrapa Tabuleiros Costeiros – Aracaju, SE – Brazil.

[3]Centro de Ciências Exatas e da Tecnologia – Departamento de Computação – Universidade Federal de Sergipe – Aracaju, SE – Brazil.

[4]Centro de Ciências Exatas e da Tecnologia – Departamento de Economia – Universidade Federal de Sergipe – Aracaju, SE – Brazil.

[5]Embrapa Territorial – Campinas, SP – Brazil.

{flavioemanuel859,fabiromoura}@gmail.com, leonardo@dcomp.ufs.br, {marcos.santos-silva,marcia.dompieri}@embrapa.br

Abstract. *This work aims to cluster Brazilian municipalities according to their spatiotemporal agricultural diversity pattern. The diversity index has been defined for eight categories and calculated by Shannon's entropy index from annual (1999-2018) IBGE's estimates for agricultural production. The proposed clustering method is based on the Self-Organizing Map, an unsupervised artificial neural network, and comprises visual and automatic steps. The method partitioned the municipalities into eight groups spatially organized in three regions showing different spatiotemporal patterns.*

Resumo. *Este trabalho tem como objetivo agrupar os municípios brasileiros de acordo com seu padrão espaço-temporal de diversidade agrícola. O índice de diversidade foi definido para oito categorias e calculado pela entropia de Shannon a partir das estimativas anuais (1999-2018) do IBGE para a produção agrícola. O método de agrupamento proposto é baseado no Mapa Auto-Organizável, uma rede neural artificial não supervisionada, e compreende etapas visuais e automáticas. O método dividiu os municípios em oito grupos organizados espacialmente em três regiões, mostrando diferentes padrões espaço-temporais.*

1. Introduction

The Brazilian agricultural activities present a high spatial diversity at different scales and dimensions, such as agricultural, industrial, and land distribution. [Sales and Rodrigues 2019, Schneider and Cassol 2014]. The spatiotemporal pattern analysis of the Brazilian agricultural diversity is crucial to design public policies better to support, mainly, small rural activities, but also agribusiness. As stated by [Teixeira and Ribeiro 2020, Dessie et al. 2019, Sambuichi et al. 2016], diversification is crucial in smallholder farmers' resilience and food security, and public policies should support it. Additionally, a diversity characterization can be helpful as criteria to identify new agribusiness hotspots, regional and local heterogeneities (economic clusters), and unveil general trends.

In this work, we applied a diversity index based on Shannon's entropy index and a machine learning clustering approach based on the Self-Organizing Maps to identify spatiotemporal patterns of Brazilian agricultural activities considering annual values estimated by the IBGE between 1999 and 2018 related to temporary and permanent cultivated crops, animal population (including dairy animals), aquaculture, vegetal extraction, and silviculture [IBGE]. Thus, the same spatial object (municipality) will be observed over time (20 years), so the dataset is a balanced spatial panel data.

2. Related Work

Self-Organizing Map (SOM) is a vector quantization machine learning technique used to order multivariate data into a low dimensional grid that can be used for data projection, compression, and clustering. As proposed by [Skupin and Hagelman 2005], there are at least three main strategies to cope with longitudinal data: 1) Use one neural network for each year and analyze the temporal patterns independently [Silva et al. 2010]; 2) Transform longitudinal data into a wide one and use only one neural network to observe the temporal pattern [Qi et al. 2019, Skupin and Hagelman 2005]; 3) Consider each observation-year as one input vector and observe the trajectory generated on the neural grid by chronologically linking each observation-year on the neural map [Chen et al. 2018, Ling and Delmelle 2016, Augustijn and Zurita-Milla 2013, Wang et al. 2013].

The spatial dimension of the data set can be explicitly included in the feature vector as proposed by [Hagenauer and Helbich 2013], but it assumes stationarity of spatial dependence when it is not valid in our case due to the concentration of the cities in some regions. The spatial proximity matrix among observations can be a constrain as proposed by [Luo et al. 2021, Teixeira et al. 2019]. It imposes a restriction suitable for regionalization purposes but not for an exploratory one. Then, it is preferable to verify spatial patterns after the clustering process simply mapping the cluster into a geographic map and checking for global and local spatial dependencies [Qi et al. 2019, Chen et al. 2018, Ling and Delmelle 2016, Wang et al. 2013].

The research question guides the interpretation of the trajectories on the neural map. [Wang et al. 2013] expands the visual-analytic potential of SOM for climate research with a series of conceptual, computational, and visual transformations to find patterns on microwave imagery of snow water equivalent gridded data. [Augustijn and Zurita-Milla 2013] used a combination of SOM's results, Sammon's projection, and GIS to analyze the dynamics of a spatiotemporal disease (measles outbreaks) diffusion pattern.

[Ling and Delmelle 2016] have used a clustering method over the trajectories' coordinates based on matrix distance to cluster and classify urban neighborhoods automatically. Notwithstanding, [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.

Hence, considering this short review and our dataset, this study will not consider the spatial component explicitly, and the temporal pattern we observed the trajectories on the neural map by a clustering process as proposed by [Skupin and Hagelman 2005], but incorporating an automatic trajectory clustering as [Ling and Delmelle 2016], and using a small size neural grid as [Augustijn and Zurita-Milla 2013]. We have used The SOM's code vectors clustering to support the interpretation of the trajectory patterns, and mapping it into a geographic map will support our analysis of spatial dependencies and heterogeneities.

3. Material and Methods

3.1. Diversity Index

We have chosen Shannon's entropy [Shannon 1948] because it is invariant to the number of possible elements in each category. Thus, it is possible to compare the diversity indices (Equation 1) of different categories based on entropy. 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); planted area with temporary crops, value of production of animal origin, temporary and permanent crops, aquaculture, vegetal extraction and forestry [IBGE].

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

Where *t* represents the year of reference, *l* the category, *p* the municipality, *m* the number of raw variables used for each category, and Xltpi the value of the *i*th 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).

Then, the longitudinal spatial data comprises eight diversity indexes for each of the 5570 municipalities for 20 years, from 1999 to 2018, so it comprises 111400 observations. In general, all indexes show a slowly decreasing diversity trend (DIV.EFETIVO, DIV.VL.T, DIV.PLANT.T) with atypical behavior in 2005 and 2015 for DIV.EXTV.VL or a slowly increasing diversity trend for DIV.VL.PRODANI, DIV.VL.P, DIV.AQU.VL. The index DIV.SILV.VL shows a cyclical behavior of seven years with a clear trend to increase diversity.

3.2. Longitudinal data vsiualization and clustering using Self-Organizing Maps

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). Section 3.1 describes the first and second steps.

3.2.1 Step 3 – Spatial panel data ordering on the 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. We defined the number of neurons based on the data size and complexity. We have used the quantization error and quality of the data projection on the neural map. In this work, we used a small size SOM as [Augustijn and Zurita-Milla 2013].

The Kohonen Self- Organizing Map is an ANN with two layers (Kohonen, 2001): the input *I* and the output *U* layers. 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 vicinity are adjusted according to Equation 2.

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

Where *a(t)* is the learning rate function, and *h(t)* is the neighborhood function centered on the winning neuron (BMU).

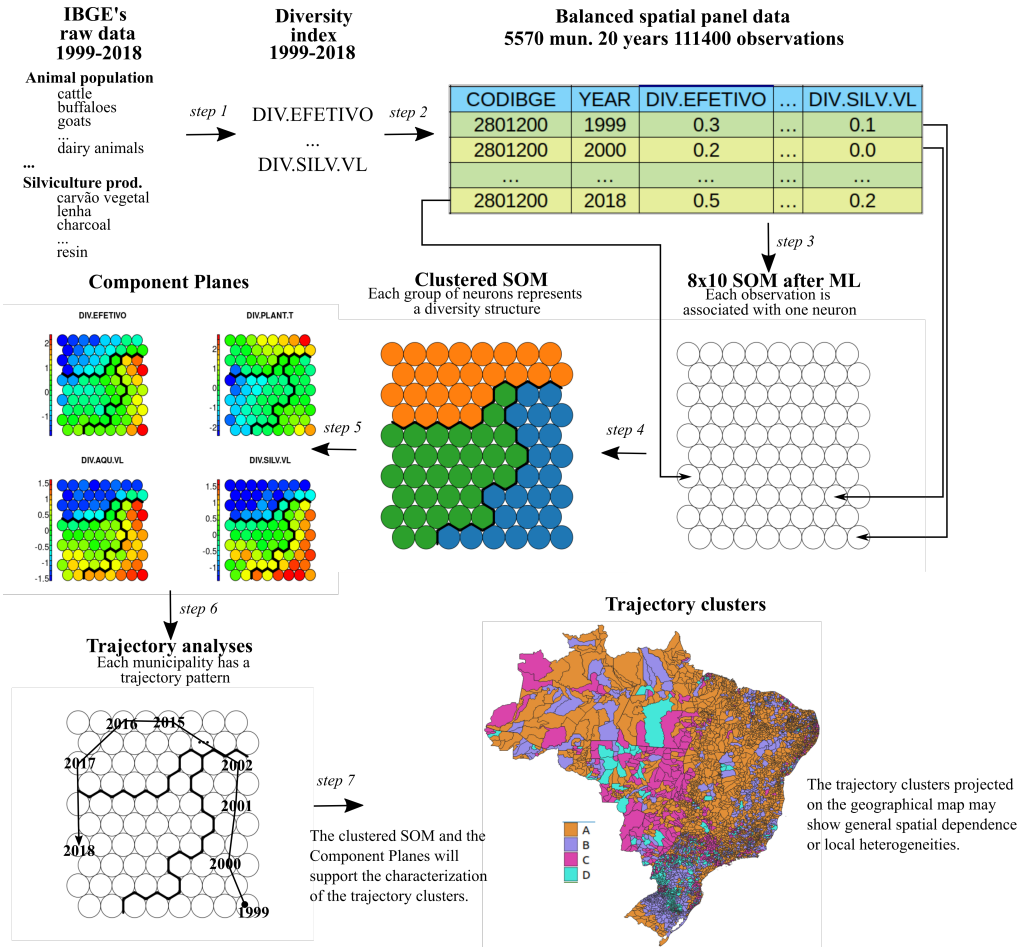

Figure 1. Methodology of spatial longitudinal data clustering based on temporal trajectory on the neural map. Source: elaborated by the authors.

3.2.2 Step 4 – Clustering the SOM's code vectors

In this step, the SOM weights are clustered using the k-means method, considering 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.

3.2.3 Step 5 – Component Plane analysis

The Component Planes unveil the patterns on the neural grid after the machine learning process by applying a coloring method based on the values of each component on the weight W matrix. So, for a given *jth* component of the SOM's code vectors, an image *f(x, y)* is generated with dimensions equal to those of the Map *MxN*, where each pixel will correspond to the value of the *j* component at the position *(x, y)* using a divergent palette pattern. Where dark blue represents maximum values, dark red minimum values, and shades 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].

3.2.4 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 [Ling and Delmelle 2016]. A trajectory for a municipality *p* can be expressed as a matrix $Traj_{ij}^{p}$ where each row corresponds to coordination *(x,y)* on the neural grid. Hence, to cluster all trajectories, it has been applied a k-means algorithm using the Frobenius Norm (p=2) (Equation 3) as a matrix distance measure [Genolini et al. 2015]. The Davies-Bouldin quality index has measured the quality of the trajectory clustering, also implemented in [Genolini et al. 2015].

$$Dist(Traj^1, Traj^2) = \sqrt{\sum_i \sum_j (Traj_{ij}^1 - Traj_{ij}^2)^2} \quad (3)$$

3.2.4 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 heterogeneity 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.

4. Results and Discussion

For the trajectory analysis, a SOM's Artificial Neural Network was defined with 25 columns and 30 lines, a non-toroidal hexagonal grid, Gaussian neighborhood function, and stochastic machine learning (online) processed in 100,000 iterations with a monotonically decreasing learning rate starting from 0.05.

As described in section 3.2, after the machine learning process, the SOM was partitioned by the k-means method, using the elbow method and the Silhouette quality indicator to define the number of groups. Then, the Component Planes were analyzed to characterize each of the partitioned regions.

As shown in Figure 2, the 25x30 RNA was partitioned using the k-means method into six regions according to the elbow method and the Silhouette indicator. These regions can be characterized from the observation of the Component Planes (Figure 3), where it is possible to visualize the ordered distribution of each variable

represented by the elements of the SOM codevetor.

From the Component Planes (Figure 3) it can be seen that the aquaculture diversity index has fewer neurons dedicated to high values (in region 6 of the neural grid), which denotes the indication of outliers. The high values of the standard deviation for this index (Table 1) confirm this hypothesis. The same is observed for the diversity index of the production value of vegetal extraction. Thus, these indices will not be used to characterize the six regions of the neural grid.

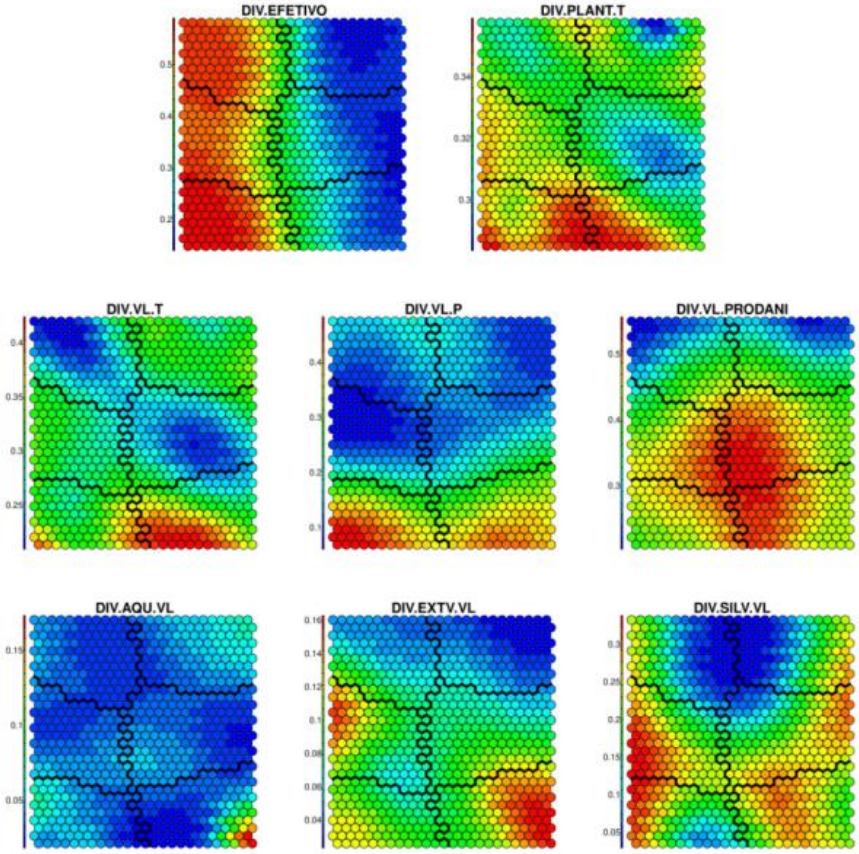

Figure 3. The eight Component Planes. One for each studied variable. The color of the neural grid follows a divergent pattern with high values mapped with red, low values with blue and intermediate values with colors varying shades of green and yellow. Source: elaborated by the authors.

Region 1 is characterized by a high diversity of quantities of animal population, including dairy animals; planted area with temporary crops and values of animal production and forestry; and low diversity in the value of production of permanent crops.

Region 2 is characterized by a high diversity of animal population, including dairy animals, planted areas with temporary crops and production values of permanent crops, animal production, and forestry. In fact, what differentiates regions 1 and 2 is the diversity in the production value of permanent crops.

Region 3 is characterized by a high diversity of animal population, including dairy animals, and low diversity of production values for temporary and permanent crops, animal production, and forestry.

Region 4 is characterized by a low diversity of animal population, including dairy animals and values of animal production, permanent crops, and forestry.

Region 5 is characterized by a high diversity in the value of animal and forestry production and low diversity in the animal population, including dairy animals, planted areas with temporary crops, and the production values of temporary and permanent crops.

Region 6 is characterized by a low diversity of animal population, including dairy animals, and high diversity of planted area with temporary crops and values of animal production, temporary and permanent crops, and forestry.

From the partitioning of the SOM into homogeneous regions, it is possible to analyze the most recurrent patterns of trajectories of the municipality in the neural grid for the period 1999 to 2018. Then, as established in section 3.2, the analysis of these trajectories was performed using the matrix distance method.

From the partitioning of the SOM into homogeneous regions, it is possible to analyze the most recurrent patterns of trajectories of the municipality in the neural grid for the period 1999 to 2018. Then, as established in section 3.2, the analysis of these trajectories was performed using the matrix distance method.

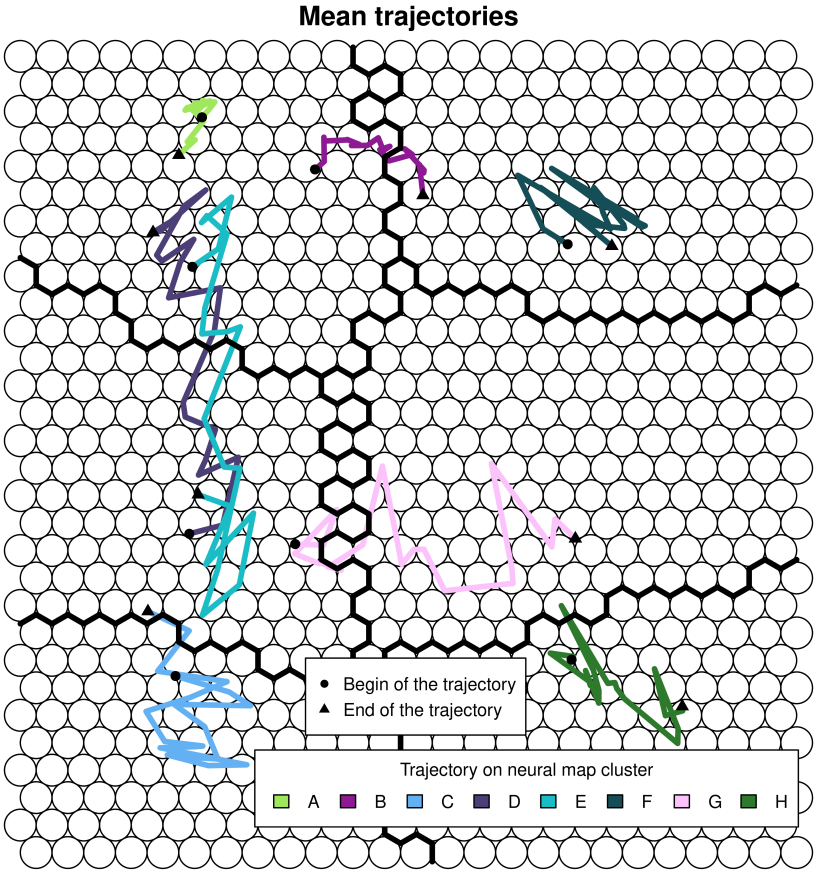

Figure 4. Illustration of the average trajectory of each group (A-H) defined from the trajectory clustering. Source: elaborated by the authors.

The application of the matrix distance clustering method, combined with the Davies-Bouldin cluster quality indicator, resulted in the partition of Brazilian municipalities into eight groups representing different trajectory patterns in the neural grid (Figure 4). Four of these (A, C, F, and H) show trajectories with short displacement. That is, they represent the municipalities that did not change their characteristics over the twenty years. Contrarily, there are four other groups (B, D, E,

and G) where it is clearly perceived when confronted with the borders of the six regions in the neural map, representing the displacement of municipalities across neural grid regions. These clusters group the cities that show tendencies to change their agricultural diversity profiles.

From the characteristics of the six regions of the neural map, it is possible to characterize each cluster of Brazilian municipalities according to the diversity indices. Then, the municipalities associated with groups A, C, F and H represent those that did not show changes in their characteristics regarding the diversity of economic activity in rural areas, with group A being more associated with the characteristics of region 1 on the neural grid, the group C to region 2, group F to region 4, and group H to region 6 on the neural map. Group B represents the municipalities that are migrating from area 3 onto the neural grid towards region 4. It indicates changes, above all, in the decrease of diversity of animal population, including dairy animals.

Groups D and E represent the municipalities that showed similar trends but with opposite directions. While group D shows a trend of migration of diversity characteristics from region one onto the neural map to region 3, group E shows the opposite. It denotes changes in the diversity of these municipalities in terms of the value of producing temporary crops, animal production, vegetal extraction, and forestry.

Group G represents the municipalities that move from the neural region 1 to 5. In short, it denot

5. Conclusions

By applying the machine learning method based on the Artificial Neural Network of the Self-Organizing map type, it was possible to group the Brazilian municipalities into eight groups. In four of them, the municipalities showed tendencies to maintain the characteristics of agricultural diversity. The other groups comprise approximately 38% of the municipalities that show trends in changing characteristics.

Globally, it is observed that Brazil could be divided into three large clusters, each one presenting a global pattern and different local regimes of spatial heterogeneities. In region I (All NE states, MG, GO, and TO) are located the municipalities that did not change their diversity profiles between 1999 and 2018. In region II (All N states, MT and MS region) predominate municipalities with a tendency to decrease the diversity of animal population, including dairy animals. In region III (the entire region S, SP, RJ, and ES), there is a good distribution of the eight groups, implying greater agricultural diversity, with a strong presence of local clusters.

Acknowledgments

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (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.

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.

Hagenauer, J. and Helbich, M. (2013). Hierarchical self-organizing maps for clustering spatiotemporal data. International Journal of Geographical Information Science, 27(10):2026–2042.

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

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.

Luo, Z. T., Sang, H., and Mallick, B. (2021). A bayesian contiguous partitioning method for learning clustered latent variables. Journal of Machine Learning Research, 22:1–52.

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.

Sales, C. and Rodrigues, R. (2019). Espac¸o rural brasileiro: diversificação e peculiaridades. Revista Espinhaço, 8(1):54–65.

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.

Silva, M., Siqueira, E., and Teixeira, O. (2010). Abordagem conexionista para análise espacial exploratória de dados socioeconômicos de territórios rurais. Revista de Economia e Sociologia Rural, 48:429–446.

Skupin, A. and Hagelman, R. (2005). Visualizing demographic trajectories with selforganizingmaps. GeoInformatica, 9(2):159–179.

Teixeira, L. V., o, R. M. A., and Loschi, R. H. (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. Journal of Machine Learning Research, 20:1–35.

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.