Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogra
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Evaluating Autoencoders as a Dimensionality Reduction Mechanism to Support Clustering Brazilian Agricultural Diversity

Avaliando Autoencoders como um Mecanismo de Redução de Dimensionalidade para Apoiar a Análise de Agrupamentos da Diversidade Agrícola Brasileira

Marcos Aurélio Santos da Silva [1], Leonardo Nogueira Matos [2], Gastão Florêncio Miranda Júnior [3], Flávio Emanuel de Oliveira Santos [4], Márcia Helena Galina Dompieri [5], Fábio Rodrigues de Moura [6] and Fabrícia Karollyne Santos Resende [7]

[1] Embrapa Coastal Tablelands, Aracaju, Brazil. marcos.santos-silva@embrapa.br.
ORCID: https://orcid.org/0000-0002-5367-2869
[2] Dept. of Computing, Federal University of Sergipe, São Cristóvão, Brazil. leonardo@dcomp.ufs.br.
ORCID: https://orcid.org/0000-0002-6302-3299
[3] Dept. of Mathematics, Federal University of Sergipe, São Cristóvão, Brasil. gastao@mat.ufs.br.
ORCID: https://orcid.org/0000-0002-0967-6141
[4] Dept. of Computing, Federal University of Sergipe, São Cristóvão, Brazil. flavioemanuel859@gmail.com.
ORCID: https://orcid.org/0000-0002-7041-5581
[5] Embrapa Territorial, Campinas, Brazil. marcia.dompieri@embrapa.br.
ORCID: https://orcid.org/0000-0001-7689-1602
[6] Dept. of Economics, Federal University of Sergipe, São Cristóvão, Brazil. fabiromoura@gmail.com.
ORCID: https://orcid.org/0000-0002-6532-110X
[7] Dept. of Computing, Federal University of Sergipe, São Cristóvão, Brazil. fabricia.resende@outlook.com.
ORCID: https://orcid.org/0000-0001-8010-6304

Recebido: 03.2023 | Aceito: 08.2023

Abstract: Brazilian agricultural production presents high spatial diversity, challenging the conception of public policies. This article proposes an approach for grouping Brazilian municipalities according to their agricultural production. We combine a feature extraction using *autoencoders* and clustering based on k-means and Self-Organizing Maps. We used panel data from IBGE's annual estimates of the production value of permanent and temporary crops, animal products, aquaculture, plant extractivism, forestry, planted areas, and herd population between 1999 and 2018. We analyzed different structures of simple stacked and incomplete *autoencoders*, varying the number of layers and neurons in each, and evaluated the asymmetric exponential linear loss function to handle the sparse data. We applied the Isomap, Kernel PCA, Truncated SVD, and MDS dimensionality reduction methods for comparative purposes. Results showed that the autoencoders could extract characteristics from the transformed raw data to allow the clustering of municipalities to reveal regional and even intra-regional patterns. The autoencoders improved comparative performance as the intrinsic dimensionality increased.

Keywords: Self-Organizing Map. Sparse data. Spatial Analysis.

Resumo: A produção agrícola brasileira apresenta elevada diversidade espacial, o que desafia a concepção de políticas públicas. Este artigo propõe uma abordagem de agrupamento dos municípios brasileiros segundo sua produção agrícola. Combinamos extração de características utilizando *autoencoders* e clusterização baseada em k-médias e Mapas Auto Organizáveis. Utilizamos os dados em painel, entre 1999 e 2018, das estimativas anuais do IBGE sobre valor da produção de culturas permanentes, temporárias, produtos de origem animal, aquicultura, extrativismo vegetal, silvicultura, área plantada e efetivo de animais. Analisamos diferentes estruturas de *autoencoders* simples empilhados e incompletos, variando o número de camadas e neurônios em cada uma delas, e avaliamos a função de perda linear exponencial assimétrica para lidar com os dados esparsos. Comparamos os autoencoders com os métodos de redução de dimensionalidade Isomap, Kernel PCA, Truncated SVD e MDS. Os resultados mostraram que os autoencoders conseguiram extrair características dos dados brutos de forma a permitir a clusterização dos municípios revelasse padrões regionais e intra-regionais. Os autoencoders melhoram o desempenho comparativo à medida que a dimensionalidade intrínseca aumenta.

Palavras Chave: Análise Espacial. Dados esparsos. Mapa Auto-Organizável.

1 INTRODUCTION

Studies have shown that agricultural production's diversity affects sustainability at farm and regional scales (FATCH et al., 2021; SALES; RODRIGUES, 2019). At the farm level, diversity contributes to income stability, increased food security, and greater resilience, mainly related to climate variations (DONFOUET et al., 2017). On a regional scale, the diversification of production systems especially impacts the conservation of natural resources such as soil, water, and native vegetation (TEIXEIRA; RIBEIRO, 2020). Characterizing the diversity of agricultural production is a necessary stepping toward elaborating territorial public policies that encourage incorporating new agricultural elements in rural socioeconomic systems. It is also crucial for the private sector to use this information for commercial purposes.

In Brazil, there are a small number of quantitative studies on its agricultural diversity (SILVA et al., 2022b; PIEDRA-BONILLA; BRAGA; BRAGA, 2020; CALDEIRA; PARRÉ, 2020; SAMBUICHI et al., 2016). They emphasize approaches that use quantitative indices of diversity. Sambuichi et al. (2016) studied the diversity of family farming based on microdata from the Declaration of Agricultural Aptitude (*Declaração de Aptidão Agrícola* - DAP). Piedra-Bonilla, Braga and Braga (2020) analyzed the diversity of municipal agricultural production, correlating it to the average size of establishments per municipality. Silva et al. (2022b) related diversity to the native vegetation change throughout the national territory. Caldeira and Parré (2020) investigated the agricultural diversity in the Cerrado biome, correlating it with rural development. Finally, Teixeira and Ribeiro (2020) established a positive correlation between agricultural diversity and the conservation of forest fragments in the Brazilian state of Minas Gerais.

These studies were based on a feature engineering approach, where raw agricultural production data are transformed to a diversity index based on Shannon entropy or the Simpson index (SHANNON, 1948; SIMPSON, 1949). These indices are calculated from the amounts or values of agricultural production and vary between 0 (absence of diversity) and 1 (maximum diversity). Diversity indices were categorized according to diversification classes to divide municipalities according to their diversity similarities (PIEDRA-BONILLA; BRAGA; BRAGA, 2020; CALDEIRA; PARRÉ, 2020; TEIXEIRA; RIBEIRO, 2020; SAMBUICHI et al., 2016). Silva et al. (2022b) opted for cluster analysis, using Machine Learning techniques, of the diversity indices for the 20 years considered (1999-2018), defining eight groups with different diversity trajectories.

Considering studies of agricultural diversity at the municipal level, we observe that Brazil presents regional and intra-regional differences, a trend towards a decrease in the diversity of agricultural production, and an inverse correlation between farm size and diversity. For example, the South has a higher level of diversity, much of which comes from family farming (SAMBUICHI et al., 2016). The Midwest has had the lowest diversification since 2013 due to agricultural intensification based on export crops such as soybeans, corn, and cotton (PIEDRA-BONILLA; BRAGA; BRAGA, 2020).

Silva et al. (2022b) used the IBGE's annual agricultural production estimates for the years 1999 to 2018 to determine a diversity index based on Shannon's entropy for each category (animal herd, planted area with temporary crops, the production value for temporary and permanent crops, aquaculture, silviculture, vegetal extractivism, and products of animal origin), totaling eight variables for 20 years and 5570 municipalities. They then used a **Shallow Learning** technique to cluster the spatial panel data based on the Self-Organizing Map Artificial Neural Network in conjunction with k-means. The raw spatial panel data used by Silva et al. (2022b) comprises 197 variables for 20 years and presents high dimensionality when we treat this panel data in the wide format ($20 \times 197 = 3940$ variables) and have a huge number of zeros and values close to zero.

Based on these previous works, we focused our study on a **Deep Learning** technique based on autoencoders to extract relevant information from raw panel data about Brazilian agricultural production that also reduces its dimensionality. Our main objective was to cluster Brazilian municipalities by extracting relevant information in lower dimensions and then applying a clustering algorithm. We evaluated the replacement of the feature engineering process (transformation of raw data into diversity indices) by automatic feature extraction from raw data using dimensionality reduction techniques. We used the same raw database used by Silva et al. (2022b) as it is the study that covers a great amount number of variables (197), has good temporal coverage (1999-2018), and focus on municipal clustering according to its diversity trends. The clustering obtained by

Silva et al. (2022b) serves as our main reference. Therefore, the research question is: Can this strategy based on feature extraction achieve results compatible with the studies based on feature engineering that transformed raw data into diversity indices?

The paper is organized as follows: Section 1 presents the general problem of using autoencoders as a dimensionality reduction mechanism to cluster sparse data related to Brazilian agriculture production. Section 2 presents a brief review of quantitative agricultural diversity measure in Brazil and on deep clustering with autoencoders; section 3 discuss the dataset and the proposed approach to feature extraction and spatial panel data clustering; section 4 shows the results and discussion; and section 5 unveil the conclusions.

2 RELATED WORK

2.1 Agricultural production diversity in Brazil

Quantitative analysis of the diversity of Brazilian agricultural production is recent (SILVA et al., 2022b; PIEDRA-BONILLA; BRAGA; BRAGA, 2020; CALDEIRA; PARRÉ, 2020; TEIXEIRA; RIBEIRO, 2020; SAMBUICHI et al., 2016). We can divide it into analyses at the rural establishment level (SAMBUICHI et al., 2016) and the aggregate municipal level based on annual estimates by the IBGE (SILVA et al., 2022b; PIEDRA-BONILLA; BRAGA; BRAGA, 2020; TEIXEIRA; RIBEIRO, 2020; CALDEIRA; PARRÉ, 2020). In most cases, the authors calculated the diversity indices from the raw data, and subsequently, they categorized municipalities or rural production units according to specific diversity ranges. Only Silva et al. (2022b) clustered the cities based on the time series (1999-2018) of the diversity dividing the variables into eight categories.

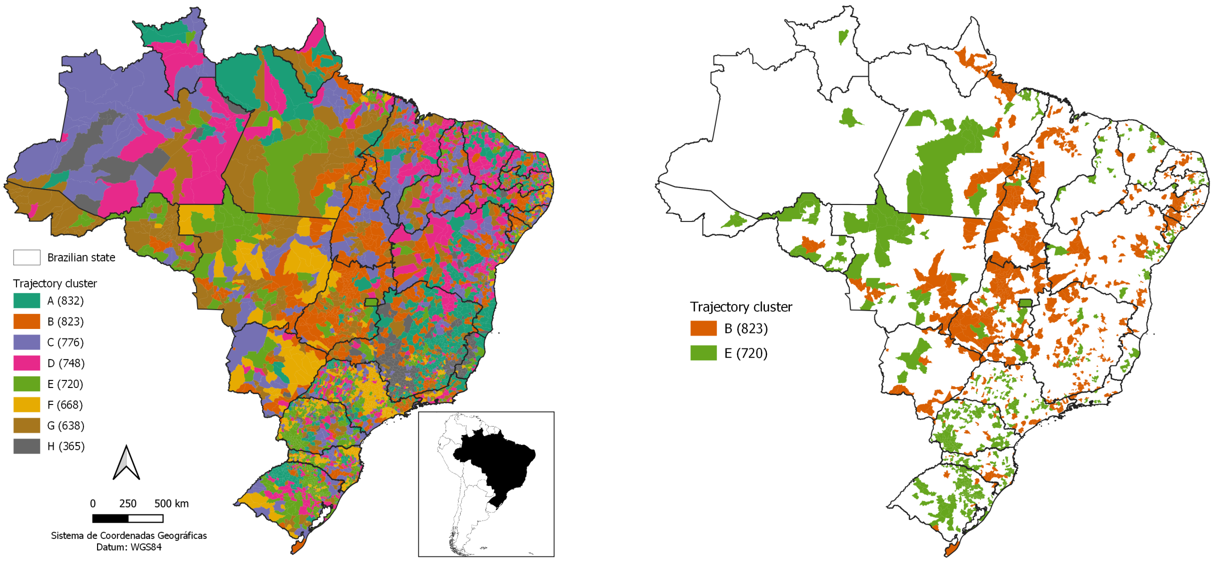
Due to the size of the country and, consequently, its different processes of socioeconomic evolution, significant regional differences can be seen in all national studies (SILVA et al., 2022b; PIEDRA-BONILLA; BRAGA; BRAGA, 2020; SAMBUICHI et al., 2016). Sambuichi et al. (2016) analyzes diversity on a farm scale and shows greater diversity in regions where family farming is more present, such as in the South region. Piedra-Bonilla, Braga and Braga (2020) outlined a scenario based on the temporal analysis (1987-2017) of the IBGE's annual estimates for production values of temporary and permanent crops, plant extractivism, and forestry products of animal origin. In this study, the authors point out that there has been a decline in agricultural diversity from 2013 onwards in the Midwest region due to increased specialization, that the Southeast region is the one with the greatest tendency to reduce agricultural diversity, and that the South region, has the highest level of diversity in agricultural production.

Silva et al. (2022b) proposed the calculation of eight diversity indices from 1999 to 2018, covering the categories of the animal herd, the area planted with temporary crops, and the production value of animal origin, temporary and permanent crops, vegetal extractivism, forestry, and aquaculture. Unlike Piedra-Bonilla, Braga and Braga (2020) and Caldera and Parré (2020)'s approaches, which calculated a single index for all variables from different categories. The authors clustered these indices into eight groups, called trajectory clusters, with some spatial dependence or aggregation at the national scale (Figure 1a). Figure 1b shows a subset (clusters B and E) of these trajectory clusters to highlight how some groups present some degree of spatial dependence (e.g., trajectory cluster B is more concentrated in the Central-West region).

According to Silva et al. (2022b), the Northeast region predominates, municipalities with a high diversity of herd population and low diversity for the animal origin and permanent crop production values (trajectory clusters C and D). In the North and Central-West Brazilian regions, predominate municipalities with a high diversity of herd population and low diversity for the animal origin and permanent and temporary crop production values. In these regions, most municipalities decrease their diversity regarding the animal herd, the area planted with temporary crops, and the animal origin and temporary crop production values (trajectory clusters B and G). In the South region, predominate municipalities with high diversity for the herd population and the animal origin production value (trajectory cluster E). However, there are also municipalities with low diversity in this region. In the Southeast region, predominate municipalities with low diversity for the herd population and high diversity for the animal origin and temporary and permanent crop production values.

Although the national scenario shows specific characteristics of each region, Silva et al. (2022b), Caldeira

Figure 1 – a) Clustered municipalities based on the trajectory of Shannon's diversity indices onto a Self-Organizing Map by (SILVA et al., 2022b). b) Map with only two trajectory clusters to highlight regional patterns of clusters 'B' and 'E'.



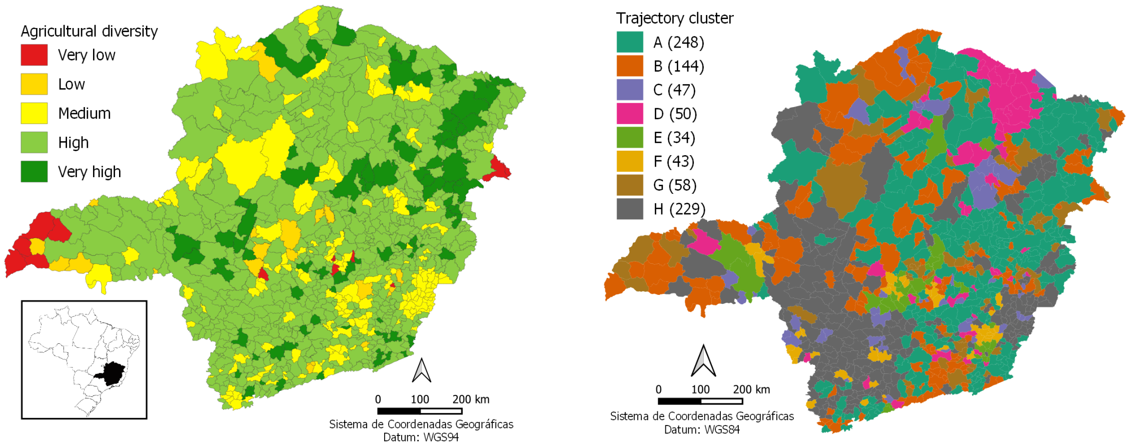
(a) Trajectory cluster

(b) Regional patterns

Elaboration: The authors (2023).

and Parré (2020), and Teixeira and Ribeiro (2020) showed that exists intra-regional distinctions across Brazil, in the Cerrado biome and the state of Minas Gerais, respectively. Teixeira and Ribeiro (2020) have shown that the state of Minas Gerais has high levels of diversity in most municipalities. Still, the region of the Minas Gerais triangle and a municipality in the border region with the state of Espírito Santo stands out with low levels of diversity, and some cities in the northeast part of the state present high levels of diversity for the production value of permanent and temporary crops (Figure 2a). The same result, with a greater level of detail, was found by Silva, who also shows the southern region of Minas (trajectory cluster H) with a tendency towards a decrease in the diversity of the herd population and an increase in diversity for the production value of products of animal origin (Figure 2b).

Figure 2 – a) Agricultural diversity categories according to (TEIXEIRA; RIBEIRO, 2020). b) Clustered Minas Gerais' municipalities based on the trajectory of Shannon's diversity indices onto a Self-Organizing Map by (SILVA et al., 2022b).



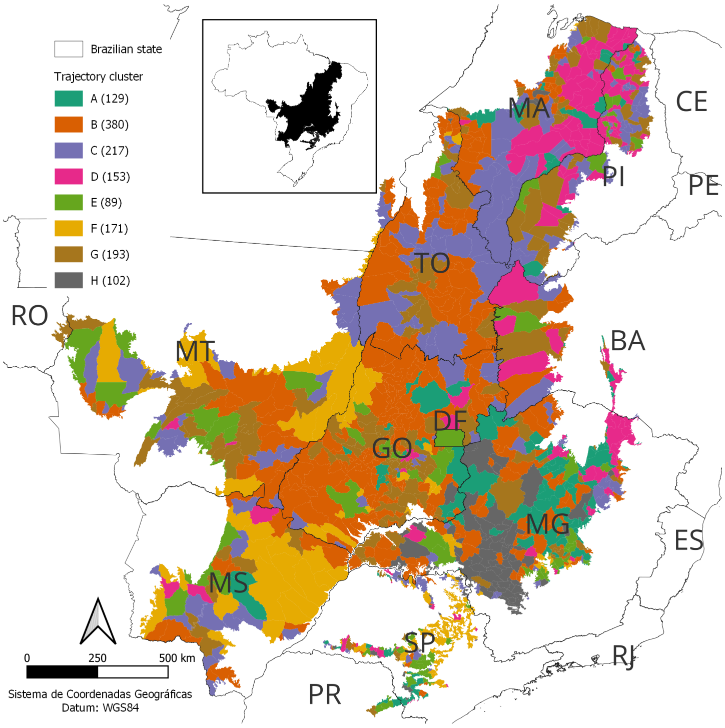
(a) Agricultural diversity categories

(b) Trajectory cluster in the state of Minas Gerais, Brazil

Elaboration: The authors (2023).

Caldeira and Parré (2020) calculated diversity from the Shannon index of municipalities with more than 50% of their territory classified as Cerrado in 2004 from the annual average (2015-2017) of production values for permanent and temporary crops, plant extractivism, forestry, and products of animal origin. The authors concluded that there is spatial dependence when the neighboring municipalities have diversity index values categorized as medium, weak, and strong. According to the authors, there is no spatial dependence when these indexes are at the extremes (super-diversified or super-specialized). In Figure 3, we have the current Cerrado region with its municipalities grouped according to Silva et al. (2022b), which presents results compatible with Caldeira and Parré (2020). According to Silva et al. (2022b), predominate in the Cerrado municipalities with a high diversity of herd population and low diversity of production value of permanent and temporary crops and products of animal origin (trajectory clusters B, C, and D), with the Cerrado biome of Minas Gerais being responsible by the greatest diversification for the production value of temporary and permanent crops and products of animal origin (trajectory cluster A).

Figure 3 – Trajectory cluster in the Cerrado biome by Silva et al. (2022b).



Elaboration: The authors (2023).

The diversity of agricultural production can generate significant impacts on the loss of native vegetation, as pointed out Silva et al. (2022b), Teixeira and Ribeiro (2020), and on rural development according to Caldeira and Parré (2020), just as the inherent characteristics of rural production unit impact it (PIEDRA-BONILLA; BRAGA; BRAGA, 2020; SAMBUICHI et al., 2016). Silva et al. (2022b) and Teixeira and Ribeiro (2020) have shown a positive correlation between the diversity of agricultural production and the loss of native vegetation. Caldeira and Parré (2020) showed that the gains from agricultural specialization (low diversification) do not necessarily convert into positive rural development. In general, all studies showed that there are specialization trends, possibly due to the direct and indirect costs of adding new agricultural processes both at the farm level and at the municipal scale due to issues of access to markets, logistics, cost of adding new agricultural methods, and edaphoclimatic conditions.

We used these studies as references to compare the results of the approach proposed in this work. Thus, the outcome of the analysis of clusters based on the dimensionality reducers of the raw data might show regional and intra-regional distinctions throughout the national territory.

2.2 Feature extraction with Deep Learning

Deep learning is a consolidating field in the industry, responsible for a significant transformation of data analytics, primarily in image, video, and text processing. Still, there are many research challenges, such as using a deep learning technique known as deep clustering over tabular data (LECUN; BENGIO; HINTON, 2015).

In a detailed clustering investigation, Min et al. (2018) identified various Deep Learning model architectures, mainly based on autoencoders. Autoencoders are deep artificial neural networks that use an unsupervised learning method to extract features from a dataset by combining an encoder (a nonlinear mapping function) and a decoder (a dataset reconstructor from the encoder-generated representation). They have a latent layer that stores the data representation to help cluster the data while gathering all the information needed to reconstruct the entire dataset. Nevertheless, this latent layer representation does not preserve the original data topology and neighborhood properties.

According to Min et al. (2018), there are at least two methods for performing Deep Clustering with autoencoders. First, use the autoencoder to reduce the dimensionality of the data before applying a clustering algorithm to the encoded data. Second, while the deep learning process updated the autoencoder parameters, combine the reconstruction loss with a clustering loss to cluster data (CHARTE et al., 2018; SONG et al., 2014; GUO et al., 2017). The dataset properties, such as sparsity, size, format (images, sequences, and tabular), and structural complexity, determine the more appropriate Deep Clustering strategy. For example, Du et al. (2021) used a deep multi-view clustering algorithm based on multiple auto-encoders for text and image clustering, Xu et al. (2020) applied a variational autoencoder for image clustering using, and Falissard et al. (2018) combined a recurrent neural network and autoencoders for longitudinal tabular data.

We concluded from the literature review that there are few works on tabular panel data, as in Falissard et al. (2018), and that a clustering analysis with autoencoders implies an empirical data-driven process. Then, the most appropriate strategy for investigating how autoencoders can map the original tabular panel data to a new latent feature space is incrementally adding complexity to the model, exploring data clustering directly from encoded data (FALISSARD et al., 2018), evaluating the combination of objective and clustering loss functions (SONG et al., 2014), and finally testing more complex Deep Clustering propositions (DU et al., 2021; XU et al., 2020).

3 DATA AND METHODS

3.1 Spatial panel data

The dataset comprises 197 variables of IBGE's annual estimates for all Brazilian municipalities (IBGE, 2021). These variables correspond to eight categories: herd population, animal origin production value, planted temporary crops, silviculture, aquaculture, vegetal extractivism, and temporary and permanent crop production value. The data and its description can be found in Silva et al. (2022a).

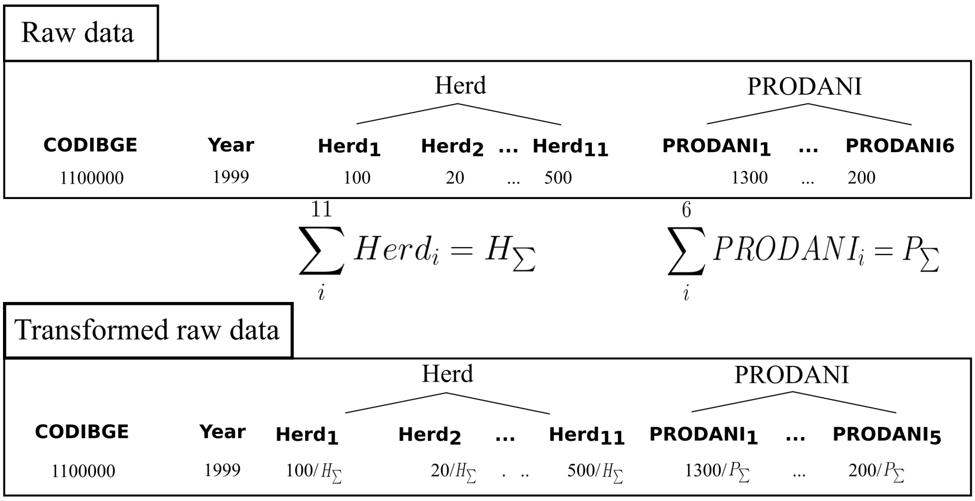
We transformed the raw data as follows: a) each variable is associated with only one category; b) for each observation (municipality-year), we calculate the sum for each category; c) each variable is updated by dividing its value by the sum of the category it belongs. In the end, each variable will correspond to the unit rate of that product for each observation (municipality-year) (Figure 4). After that, we linearly normalized the data according to the min-max algorithm transforming all variables into the interval $[0, 1]$. We set "not available"data to zero.

The main characteristic of this dataset is the considerable presence of zeros. The mean and median percentages of zeros per variable in the entire dataset are 83.09% and 91.49%, respectively. They are structural zeros because most municipalities produce a limited amount of agricultural products, so we set them to zero. This unbalanced data challenges the learning process of artificial neural networks by inducing them to learn the zeros instead of the rest of the patterns. Initial tests showed that any autoencoder structure models used in this research converged for symmetric and some asymmetric (linear and quadratic) loss functions. Thus, this demanded investigating a suitable loss function for a very sparse dataset.

3.2 Proposed approach

The study intends to evaluate the Autoencoder artificial neural network (ANN) as a dimensionality reducer for subsequent clustering of Brazilian municipalities based on this reprojected data Z in a new feature

Figure 4 – Example of raw data and its transformation. We converted each variable to its unit rate according to the category (e.g., herd population, Herd, or animal origin production value, PRODANI) it belongs.



Raw data

		Herd				PRODANI		
CODIBGE	Year	Herd1	Herd2	...	Herd11	PRODANI1	...	PRODANI6
1100000	1999	100	20	...	500	1300	...	200

$$\sum_{i}^{11} Herd_i = H_{\sum} \qquad \sum_{i}^{6} PRODANI_i = P_{\sum}$$

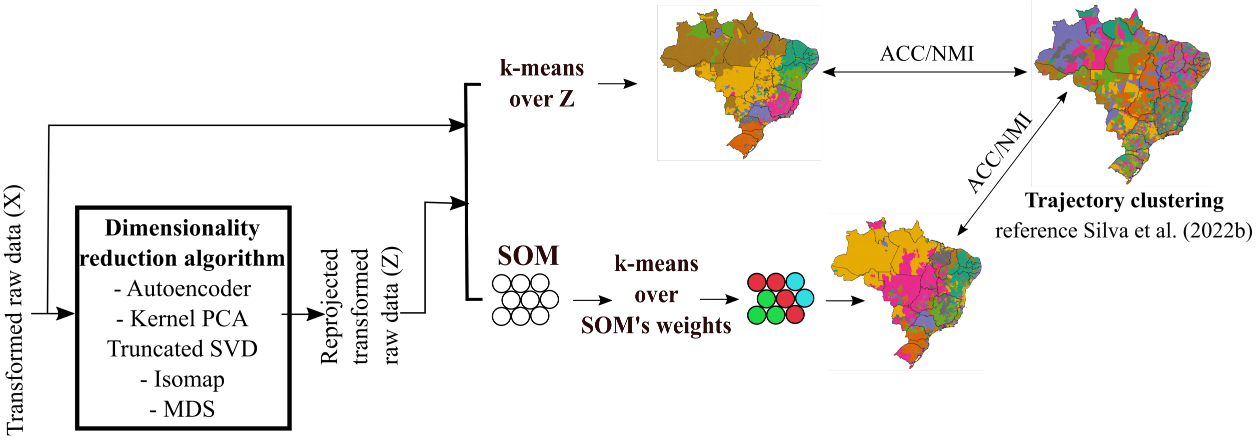
Transformed raw data

		Herd				PRODANI		
CODIBGE	Year	Herd1	Herd2	...	Herd11	PRODANI1	...	PRODANI5
1100000	1999	$100/H_{\sum}$	$20/H_{\sum}$	. ..	$500/H_{\sum}$	$1300/P_{\sum}$	...	$200/P_{\sum}$

Elaboration: The authors (2023).

space. We applied two clustering approaches, the k-means method recommended for convex data with simple structure and a combination of Z data ordering with Self-Organizing Maps (SOM) associated with k-means clustering over the neural network weights. This last method is more appropriate for non-convex data with more complex structures. We compared the clustering results with the result obtained by Silva et al. (2022b), so it is necessary to maintain the number of clusters found by these authors, $k = 8$, in all cases. We compared the clusters and our reference, the trajectory clusters, using the accuracy (ACC) and the Normalized Mutal Information (NMI) measures. The aim is not to obtain a clustering identical to that of Silva et al. (2022b) but to compare the effects of different dimensionality reduction methods based on a standard reference (Figure 5).

Figure 5 – The clustering approach of the transformed raw data (X) uses two strategies: directly clustering X and reprojecting it based on dimensionality reduction. We adopted two clustering methods (k-means and Self-Organizing Mapa (SOM) as a data ordering method associated with the k-means algorithm. Finally, we used the ACC and NMI measures to compare the results with our reference work (SILVA et al., 2022b).



Elaboration: The authors (2023).

We modeled the deep neural networks using Python version 3.7.13 and keras framework version 2.11.0, clustered using R packages, and used SOMPAK Kohonen et al. (1996) for the Self-Organizing Map processing. We generated the geographical maps using QGIS version 3.6.

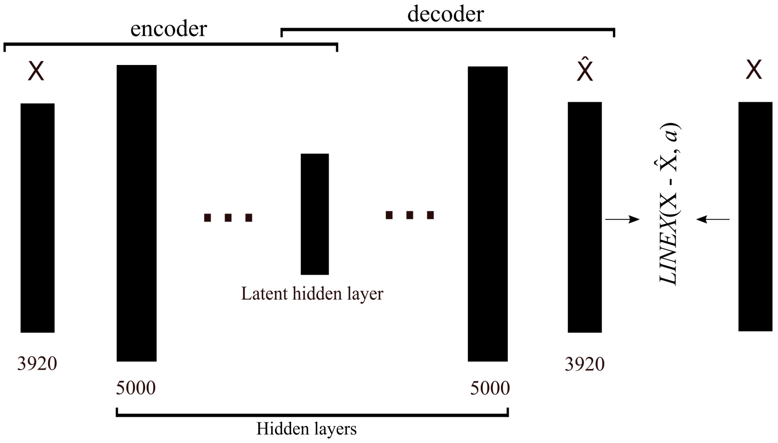
3.2.1 DIMENSIONALITY REDUCTION STRATEGIES

We applied all dimensionality reduction algorithms to the transformed raw data in a wide format (3940 variables) normalized between zero and one, evaluating six values $(10, 25, 50, 100, 250, and 500)$ for the number of reduced components ($ncomp$). Autoencoders do not preserve the original data's distance or probability distributions. Therefore it is an instrument for discovering new features in the data. We compared the results obtained with autoencoders with other dimensionality reduction strategies such as Multidimensional Scaling (MDS) (KRUSKAL, 1964), Isomap (TENENBAUM; SILVA; LANGFORD, 2000), Truncated Singular Vector Decomposition (SVD) (HALKO; MARTINSSON; TROPP, 2009), and Kernel Principal Component Analysis (KPCA) (MÜLLER et al., 2001; SCHÖLKOPF; SMOLA; MÜLLER, 1998). The Truncated SVD method is linear and tuned to handle sparse data. The MDS, Isomap, and KPCA methods are appropriate when the intrinsic dimension is low, and the Isomap maintain local distances from the original data.

3.2.2 AUTOENCODER MODELS

We have chosen six simple stacked undercomplete autoencoder models with the same optimizer (adam), hidden and output activation functions (relu and sigmoid), fully connected, with the same loss function, and varying the number of hidden layers and the size of the latent layer (Figure 6). Table 1 shows the number of the hidden layers to the encoder component of each evaluated autoencoder, including the number of neurons on the latent hidden layer.

Figure 6 – General autoencoder stacked structure used in this study.



Elaboration: The authors (2023).

Table 1 – Encoder layers structure (number of neurons per hidden layer).

ID	Number of neurons on the latent layer	Encoder layers structure
I	500	3920-5000-3000-2000-1000-**500**
II	250	3920-5000-3000-2000-1000-500-**250**
III	100	3920-5000-3000-2000-1000-500-250-**100**
IV	50	3920-5000-3000-2000-1000-500-250-100-**50**
V	25	3920-5000-3000-2000-1000-500-250-100-50-**25**
VI	10	3920-5000-3000-2000-1000-500-250-100-50-25-**10**

Elaboration: The authors (2023).

3.2.3 ASYMMETRIC LOSS FUNCTION

Asymmetric loss functions are suitable for cases where bias is relevant or there is a considerable imbalance in data representation, as in our case (very sparse data), or fault detection (GUPTA; HAZARIKA; BERLIN, 2020; DRESS; LESSMANN; METTENHEIM, 2018). The most frequently used loss functions for regression are linear and quadratic and have their asymmetric variants (BERK, 2011). The most relevant characteristics of an asymmetric function are its capability to cope with different error situations and directions. Gupta, Hazarika and Berlin (2020) proposed an asymmetric function based on Huber (1964) that is quadratic when the error is small but is like mean absolute error (mse) when the error is larger than a threshold.

As stated in section 3.1, the standard mean square error loss function and the linear and quadratic asymmetric variant functions failed to guide the learning process of the evaluated autoencoders to reach a convergence curve. Thus, we assessed the linear exponential (LINEX) loss function that rises exponentially on one side of the zero and almost linearly on the other side (KHATUN; MATIN, 2020; VARIAN, 1975).

The LINEX loss function is given by the Eq. (1), where $\hat{x_i}$ represents the model-based forecast of actual x_i for case i, and $a \neq 0$ is a constant that determines the degree of asymmetry. The direction of the asymmetry can be defined by the signal of a or by change the subtraction $(x_i - \hat{x_i})$ by $(\hat{x_i} - x_i)$. For $|a| \to 0$ then the $LINEX(\hat{x_i}) \to MSE(\hat{x_i})$, so the $LINEX$ loss function could be thought of as an asymmetric generalization of the mean squared error loss function (MOHAMMED; ALSHANBARI; EL-BAGOURY, 2022; KHATUN; MATIN, 2020; VARIAN, 1975).

$$LINEX(\hat{x_i}) = \frac{1}{n}\sum_{i=1}^{n}\frac{2}{a^2}\left(e^{a(x_i - \hat{x_i})} - a(x_i - \hat{x_i}) - 1\right) \tag{1}$$

We evaluated $a \in \{5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 11.0, 12.0\}$ using cross-validation with hold-out method splitting the data into training (80%) and test (20%) datasets. We observed that the autoencoders demonstrated some convergence for $a \geq 5.0$.

3.2.4 CLUSTERING REPROJECTED DATA Z

For cluster analysis, we used two simple techniques. First, we apply the k-means algorithm to the reprojected data Z. In the second, we apply a data ordering algorithm based on the Unsupervised Artificial Neural Network (ANN) Self-Organizing Map (SOM) that associates to each input data a vector of weights, and, in general, the number of SOM weight vectors is defined as less than the total number of observations (5570 municipalities). It, associated with the data ordering property of the SOM, helps partition the input data from the clustering of the weight vectors of the SOM (SILVA et al., 2022b; KOHONEN, T., 2013; KOHONEN, Teuvo, 2001). The main hyperparameter of ANN SOM is its size. This work uses a two-dimensional SOM with hexagonal topology, sequential learning, Gaussian neighborhood function, and initial learning rate equal to 0.5, with dimension 20 × 15 (300 weight vectors). As our objective was to reduce the data volume while preserving its topology for the clustering task, we opted for a SOM network with the number of weight vectors (300) much smaller than the number of observations.

To establish a baseline, we clustered the transformed raw data (X) using a k-means algorithm based on joint-trajectories proposed by Genolini et al. (2015) over all 5570 municipalities, 197 variables for 20 years in the wide format (3940 variables) for $k = 8$ to be able to compare with the results obtained by Silva et al. (2022b). We also clustered the raw data, ordering the X data with SOM and applying the k-means algorithm to its weights.

For each grouping, we calculated three validity indices: Silhouette (ROUSSEEUW, 1987), Davies-Bouldin (DAVIES; BOULDIN, 1979), and CDbw (HALKIDI; VAZIRGIANNIS, 2008). We conducted all clustering considering $k = 8$ to compare with our reference. We used the Hungarian method proposed by Kuhn (1955), which defines a function $m(c_i)$, to match the cluster obtained by the evaluated clustering method for each observation i, c_i, and our referential cluster, y_i, obtained by Silva et al. (2022b). After this, we calculated two clustering quality measures, an unsupervised equivalent of classification accuracy (ACC), Eq. (2), and the Normalized Mutual Information (NMI) based on entropy, Eq. (3). The former is a unitary rate that represents the amount of corrected labeled observations and varies from zero (wrong clustering) to one (perfect clustering) (VINH; EPPS; BAILEY, 2010). A greater NMI means a good match and ranges from zero to one.

$$ACC = \frac{Number\ of\ observations\ where\ y_i = m(c_i)}{n} \tag{2}$$

where n is the number of observation, in our study $n = 5570$.

$$NMI(Y, C) = \frac{2I(Y, C)}{H(Y) + H(C)} \tag{3}$$

where Y is the true labels vector, C the cluster labels vector, $I(Y, C)$ the mutual information that is the entropy of class labels within each cluster, and $H(Y)$ and $H(C)$ the entropy value for Y and C.

We mapped the clustering into the Brazilian municipal geographical map to check for spatial dependence and regional and intra-regional distinction.

4 RESULTS AND DISCUSSION

4.1 $LINEX$ parameterization and data encoding

Figure 7 shows the mean loss for the train and test dataset for 30 runs (50 epochs each), randomly changing train and test data. We observed these values considering different values for the a parameter for the LINEX loss function and an MSE loss function for comparison. For all loss functions and datasets (train and test), the MSE is almost a horizontal line denoting that the autoencoders did not converge with this loss function. The LINEX loss function with $a = 7$ presents the best result for all autoencoder models for training and test datasets. The performance decreases for $a < 7$ and $a > 7$.

After defining the parameter a, $a = 7$, for the LINEX function, we proceeded with Deep Learning of the six autoencoder models defined in Table 1. We observed that the greater the dimension in the latent layer, the greater the proportion of components with all their values equal to zero. Only one component was zeroed for the autoencoder with dimension 10 in the latent layer. For the autoencoder with dimension 500 in the latent layer, 40% of the components were vectors of zeros. These zeroed vectors were disregarded in the subsequent step, clustering.

4.2 Clustering encoded data Z

4.2.1 VALIDITY CLUSTERING INDICES

Figure 8 shows the result of the cluster validation indices for all dimensionality reduction strategies. We normalized the values between 0 (worst partition) and 1 (best partition) for all indexes. In all cases, the best partition for the Davies-Bouldin and Cdbw indices was for the number of components ($ncomp$) equal to 10. For the Silhouette index, we have the same, except for the clusters using the MDS ($ncomp = 25$) associated with the SOM plus k-means clustering and from the Truncated SVD (ncomp=500) associated with k-means.

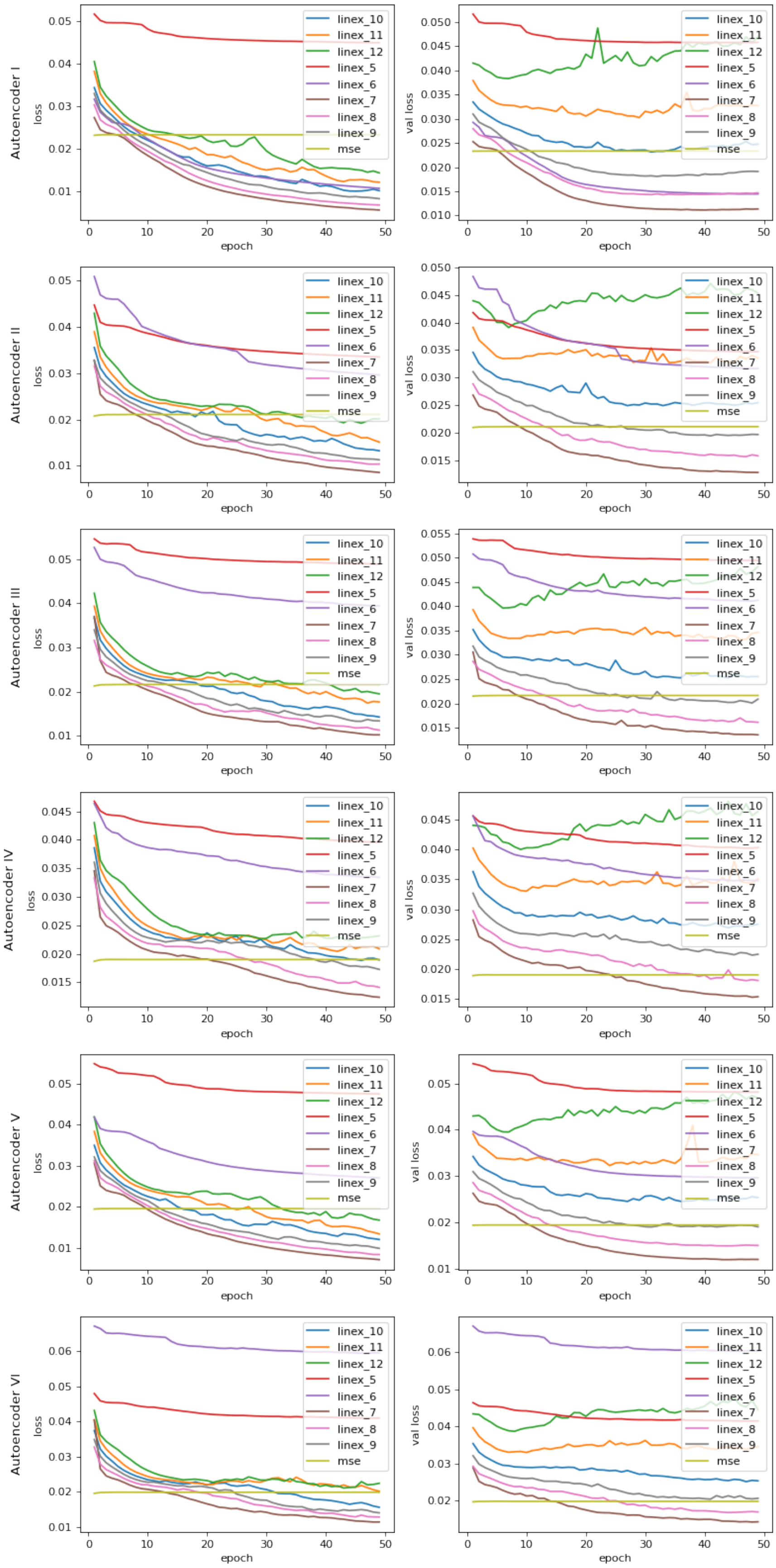
When one observes the differences in the clustering methods for the same dimensionality reduction strategy, there is almost the same behavior for the clustering validation indices for Isomap, some differences for KPCA, and substantial distinctions for MDS, Truncated SVD, and Autoencoders. The role of SOM in the process is to order the data in the new feature space to facilitate clustering with the k-means method by approximating nearby vectors. Then, if the reprojected data do not preserve its topology, the SOM will reinforce it. So, the greater the topology distinction between original and reprojected data, the more significant the differences between the results obtained by k-means and SOM plus k-means strategies will be.

Autoencoders are the only dimensionality reduction strategy whose index curves do not monotonically decay as the number of components grows for both clustering methods. This fact suggests that there is an improvement in pattern separation as the number of components in the latent layer grows. The autoencoder could be more appropriate for dimensionality reduction by extracting relevant features considering high values for the number of components.

4.2.2 ACC AND NMI CLUSTERING QUALITY MEASURES

We obtained the highest value for the ACC combining the Isomap ($ncomp = 100$) and the k-means algorithm (Figure 9a) and the second greater value combining the Truncated SVD ($ncomp = 25$) with SOM and k-means (Figure 9c). Significant differences exist between the ACC statistic values for the two clustering methods, and the Truncated SVD algorithm generates the best results in most cases for SOM plus k-means clustering. For the case of clustering only with k-means, the Isomap and KPCA methods dominate. The ACC for the autoencoder strategy shows low values for almost all numbers of components ($ncomp$) in both clustering strategies.

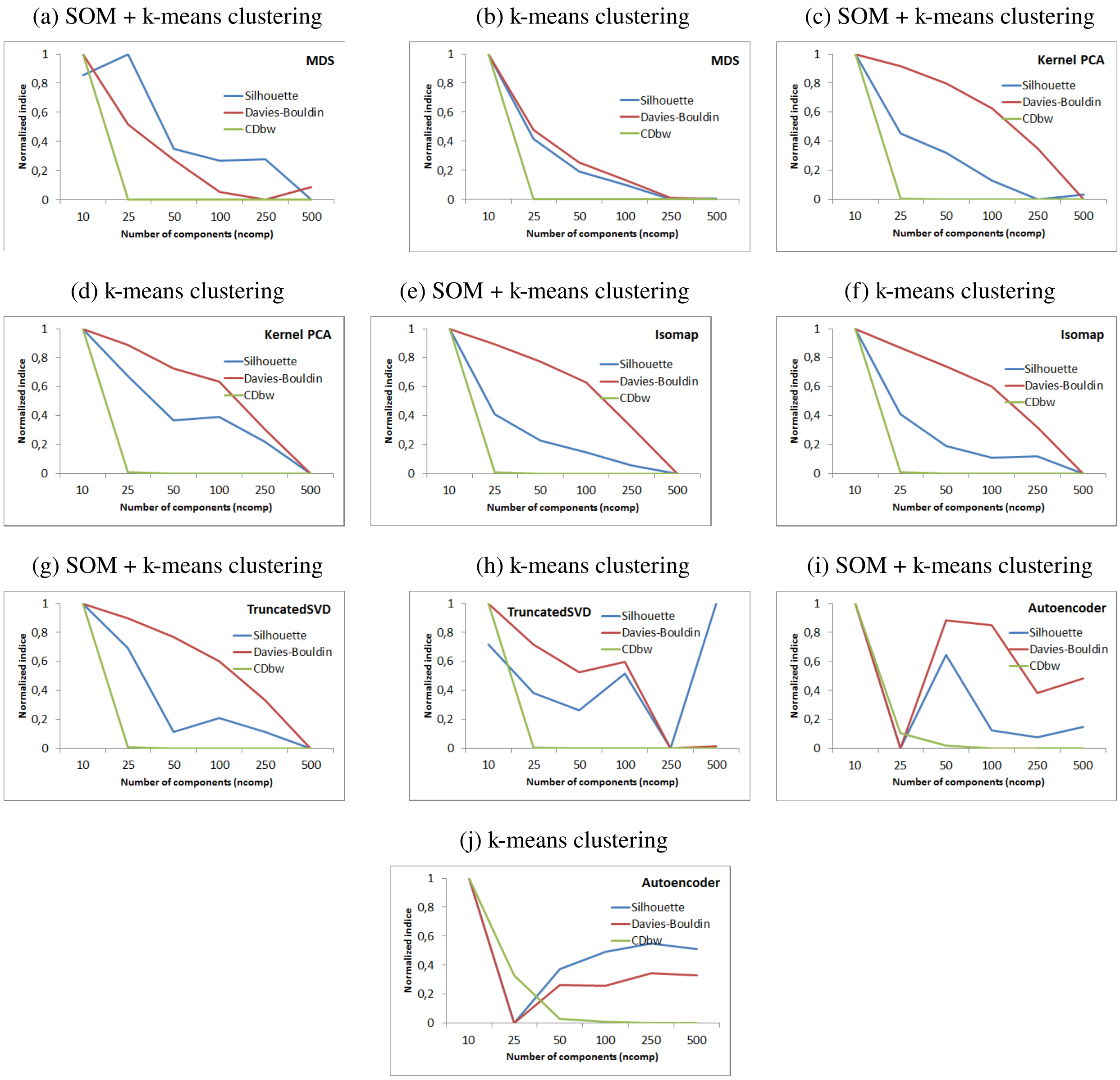
Figure 7 – Mean loss values for the train (first column) and test data for validation (second column) considering different loss functions (MSE and LINEX for different a values, linex_a) considering fifty runs for each autoencoder model.



Elaboration: The authors (2023).

The NMI index is more balanced as it evaluates entropy per cluster. For this indicator, the dimensionality reduction strategy with the best performance was with the MDS algorithm for all component values when we

Figure 8 – Normalized clustering validity indices for all dimensionality reduction strategies considering two clustering strategies (SOM + k-means and k-means) and varying the number of components (*ncomp*) generated by the dimensionality reduction strategy.



(a) SOM + k-means clustering

(b) k-means clustering

(c) SOM + k-means clustering

(d) k-means clustering

(e) SOM + k-means clustering

(f) k-means clustering

(g) SOM + k-means clustering

(h) k-means clustering

(i) SOM + k-means clustering

(j) k-means clustering

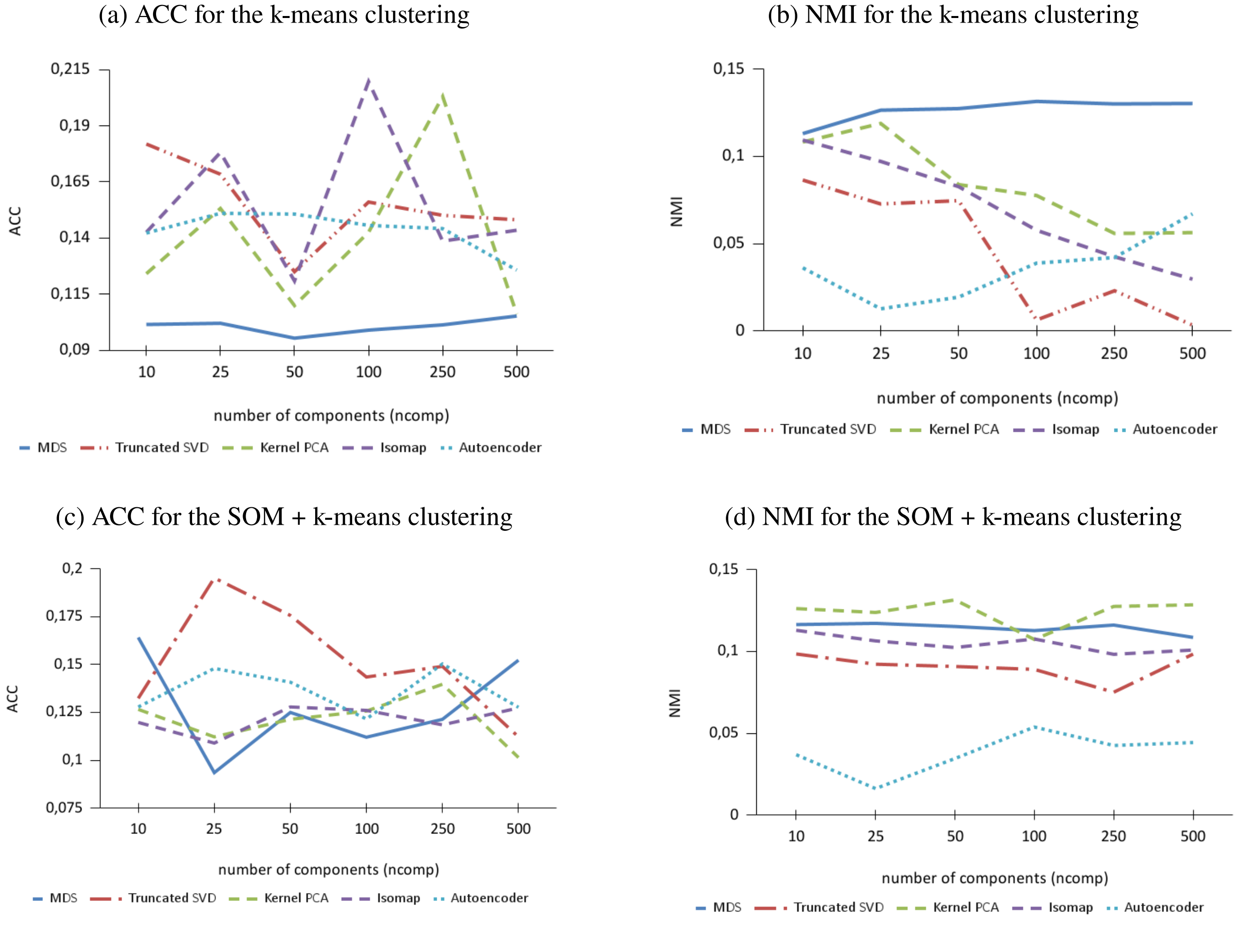
Elaboration: The authors (2023).

consider clustering with k-means (Figure 9b). In this case, the approaches with Isomap, TruncatedSVD, and Kernel PCA presented decreasing performances as *ncomp* increases, while with autoencoders, the opposite occurs.

When we analysed the NMI metric for clustering with SOM + k-means, we found that, except for the autoencoders that showed a slight increase as *ncomp* increased, the values remained almost constant (Figure 9d). As the SOM neural network sorts the Z reprojected data before clustering the SOm weights by k-means, it may have increased the clustering performance by leveling up the clustering effectiveness compared to the benchmark.

The geographic mapping of the clusters can provide additional information on whether the combined methods of dimensionality reduction and clustering managed to capture regional and intra-regional distinctions from the national agricultural production data.

Figure 9 – ACC and NMI measures for all dimensionality reduction strategies for the two clustering methods (k-means and SOM+k-means.



(a) ACC for the k-means clustering

(b) NMI for the k-means clustering

(c) ACC for the SOM + k-means clustering

(d) NMI for the SOM + k-means clustering

Elaboration: The authors (2023).

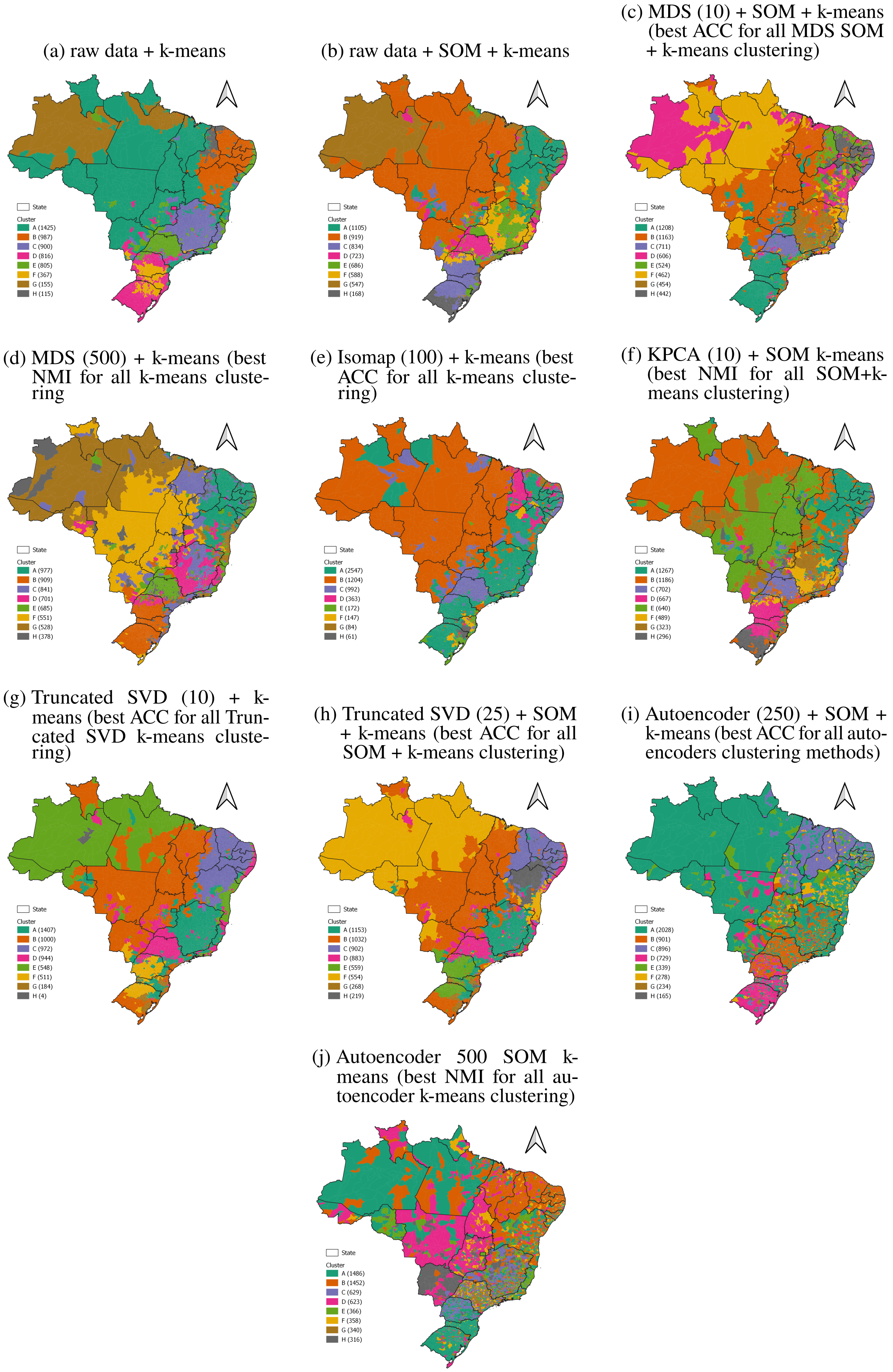
4.2.3 REGIONAL SPATIAL DISTINCTIONS

To interpret the spatial distribution of the clusters found with the different dimensionality reduction methods, we selected those with the highest value for the ACC and NMI measures and others that could reveal some spatial pattern (Figure 10). We also plotted the clustering maps obtained by applying the two methods to the raw data (Figure 10a-b).

These maps show that distinctions between the five Brazilian regions are observed in all cases, with greater or lesser detail. However, it is observed that the highest value for the ACC measure (Isomap, $ncomp = 100$) does not indicate a better characterization of the regions, as they associate Minas Gerais and the South region in the same group (Figure 10e) and according to Silva et al. (2022b), Piedra-Bonilla, Braga and Braga (2020), and Sambuichi et al. (2016), this is not true.

The clustering obtained from the MDS does not vary much between the extremes of the number of components (Figures 10c-d). The Autoencoders' clusterings improve in distinguishing regional patterns as the number of components grows from 250 (Figure 10i) to 500 (Figure 10j). For all other dimensionality reduction strategies, the greater the number of components, the smaller the regional distinction, a fact that we can not infer from the ACC and NMI measures.

Figure 10 – Geographic mapping of the clustering results by using raw data with high dimensionality (3940) and by adopting a dimensionality reduction strategy (number of components in parentheses) followed by the clustering method (k-means or SOM + k-means).


(a) raw data + k-means

(b) raw data + SOM + k-means

(c) MDS (10) + SOM + k-means (best ACC for all MDS SOM + k-means clustering)

(d) MDS (500) + k-means (best NMI for all k-means clustering

(e) Isomap (100) + k-means (best ACC for all k-means clustering)

(f) KPCA (10) + SOM k-means (best NMI for all SOM+k-means clustering)

(g) Truncated SVD (10) + k-means (best ACC for all Truncated SVD k-means clustering)

(h) Truncated SVD (25) + SOM + k-means (best ACC for all SOM + k-means clustering)

(i) Autoencoder (250) + SOM + k-means (best ACC for all autoencoders clustering methods)

(j) Autoencoder 500 SOM k-means (best NMI for all autoencoder k-means clustering)

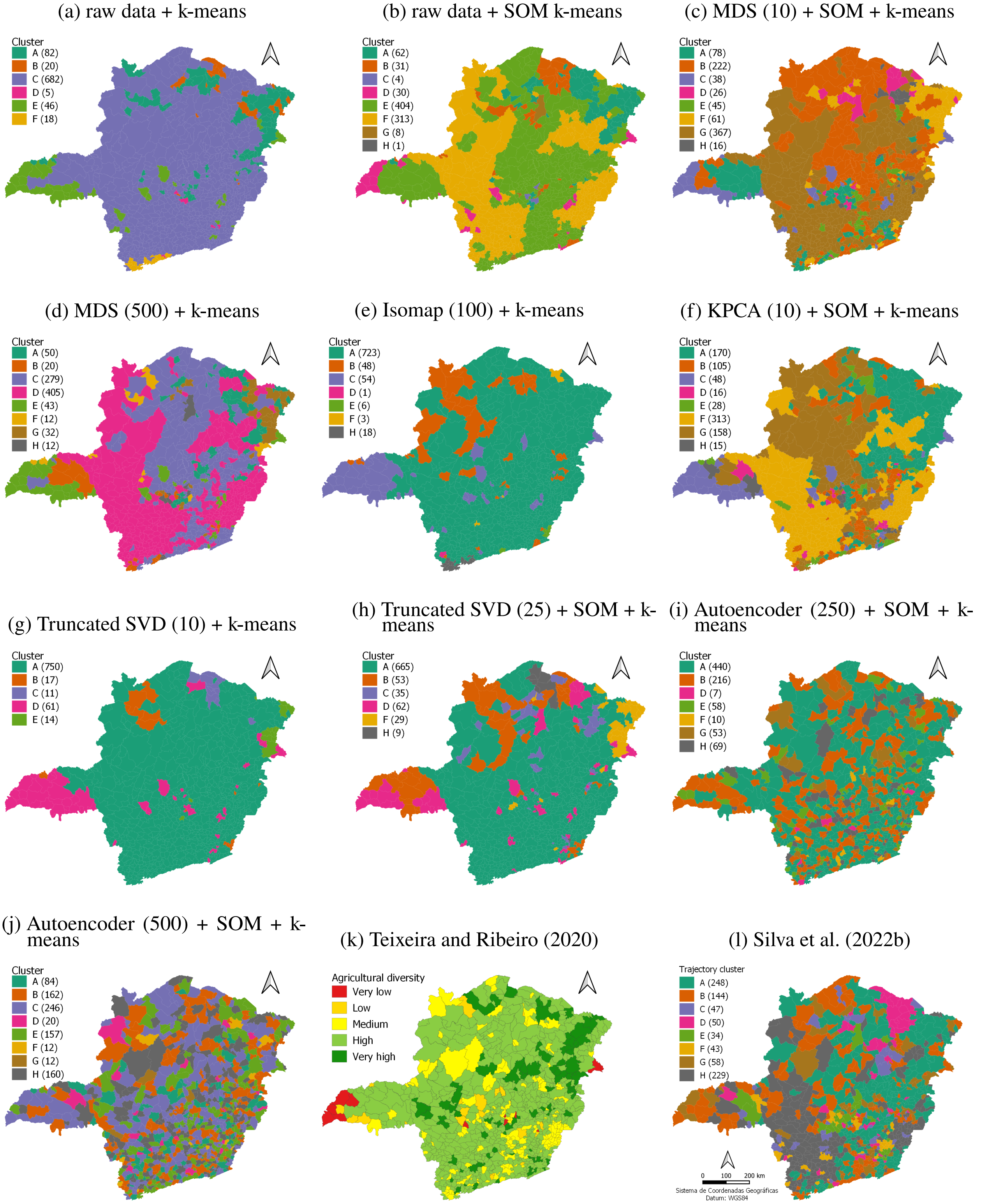
Elaboration: The authors (2023).

DOI: http://dx.doi.org/10.14393/rbcv75n0a-68733

4.2.4 INTRA-REGIONAL SPATIAL DISTINCTIONS IN THE BRAZILIAN STATE OF MINAS GERAIS

In Figure 11, we highlight the Brazilian state of Minas Gerais for all the clusters shown in Figure 10.

Figure 11 – The Brazilian state of Minas Gerais clustering for the raw data with high dimensionality (3940) and adopting a dimensionality reduction strategy (number of components in parentheses) followed by the clustering method (k-means or SOM + k-means). We also have the clustering obtained by Teixeira and Ribeiro (2020) and by Silva et al. (2022b) as a comparison source.



(a) raw data + k-means

(b) raw data + SOM k-means

(c) MDS (10) + SOM + k-means

(d) MDS (500) + k-means

(e) Isomap (100) + k-means

(f) KPCA (10) + SOM + k-means

(g) Truncated SVD (10) + k-means

(h) Truncated SVD (25) + SOM + k-means

(i) Autoencoder (250) + SOM + k-means

(j) Autoencoder (500) + SOM + k-means

(k) Teixeira and Ribeiro (2020)

(l) Silva et al. (2022b)

Elaboration: The authors (2023).

The Figure 11 allow us to verify if the dimensionality reduction applied to the raw data extracted the necessary features for identifying intra-regional patterns like those found by Silva et al. (2022b) and Teixeira & Ribeiro (2020). In all cases, except for autoencoders, intra-regional distinctions compatible with those found by (TEIXEIRA; RIBEIRO, 2020) are identified even though they are labeled differently. That is, the extreme of the Minas Gerais triangle and a municipality bordering the state of Espírito Santo (ES) are pretty different from the other municipalities, and according to Teixeira and Ribeiro (2020), they represent very low diversified cities.

The clustering that combines the SOM with the k-means method applied to the transformed raw data (Figure 11b), with the MDS (Figure 11c), for $ncomp = 10$, and KPCA (Figure 11f), for $ncomp = 10$, identified a group compatible with the trajectory cluster H obtained by Silva et al. (2022b).

The clustering obtained from the dimensionality reduction with the autoencoders showed a very different pattern from the others. The level of regional and intra-regional detail increased as the number of components increased. However, the cluster patterns did not follow those obtained by the other dimensionality reduction methods. In the case of Minas Gerais, it is clear that the autoencoders extracted characteristics that identify differences between its municipalities in an even more balanced way than the others if we observe the number of occurrences per cluster.

Considering the entire Brazilian territory, the results show that, for the conditions established for the experiments, the dimensionality reduction methods Isomap, KPCA, MDS, and Truncated SVD were not able to capture the intra-regional differences obtained by Silva et al. (2022b), even with the increase in the number of components. The results confirm the vocation of these algorithms for situations where the intrinsic dimensionality is low. On the contrary, the autoencoders showed that they improve performance as the number of components increases. Possibly, autoencoders are more suitable when the intrinsic dimensionality is high, despite not having reached the same level of intra-regional differences obtained by Silva et al. (2022b).

5 Conclusions

Identifying patterns through unsupervised methods in the temporal database of annual estimates of agricultural production from the IBGE poses a great challenge, handling data with high sparseness. Although it was not designed to handle this type of data, we were able to approach the problem using an asymmetric loss function, the LINEX function, requiring adjustment of the function parameter.

We observed that the larger the latent layer, the greater the percentage of zeroed components. For sparse data, there is a limit of neurons in the latent layer. It is necessary to fine-tune the autoencoder to find the most suitable number of components for the proposed analysis.

For our case study, the greater the number of neurons in the latent layer (number of components), the better its performance in identifying regional patterns of differentiation and some intra-regional patterns, as in the case of Minas Gerais Brazilian state.

Both for the autoencoders and the other dimensionality reduction methods, we observed the difficulty in characterizing the groups found from the data reprojected into a new feature space. The difficulty is even greater as autoencoders do not preserve raw data's topology or probability distribution. However, this approach can be used as a complementary way of finding spatial or non-spatial patterns in the database.

In conclusion, all dimensionality reduction techniques associated with clustering algorithms effectively identified regional patterns of agricultural diversity from raw data. However, only the method with autoencoders showed potential for identifying intra-regional patterns. Nevertheless, More studies are needed to explore the full capability of feature extraction from sparse data from deep neural networks.

Future work should include exploring other Deep Clustering techniques such as the combination of objective and clustering loss functions as in Song et al. (2014), the use of multi-view clustering as in Du et al. (2021), or using variational autoencoders as in Xu et al. (2020).

Acknowledgments

The authors would like to thank the Foundation to Support Research and Technological Innovation of the State of Sergipe (Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe - FAPITEC/SE) that supported this research through its program of bourses for scientific initiation (Programa de Bolsas em Iniciação Científica - PBIC) notice FAPITEC/SE/FUNTEC Nº 06/2022.

Authors’ Contribution

M.A.S.d.S.: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Validation, Visualization, Writing - initial draft and Writing - revision and editing; L.N.M.:Conceptualization, Formal Analysis, Methodology, Supervision, Writing - revision and editing; G.F.M.J.:Conceptualization, Formal Analysis, Methodology, Writing - revision and editing; F.E.d.O.S.: Software, Visualization, Writing - revision and editing; M.H.G.D.: Conceptualization, Data Curation, Methodology, Resources, Writing - revision and editing ; F.R.d.M.: Conceptualization, Writing - revision and editing; F.K.S.R.: Software, Visualization, Writing - revision and editing.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

BERK, R. Asymmetric loss functions for forecasting in criminal justice settings. **Journal of Quantitative Criminology**, v. 27, n. 1, p. 107–123, 2011.

CALDEIRA, Charly; PARRÉ, José Luiz. Diversificação agropecuária e desenvolvimento rural no bioma Cerrado. **Revista Americana de Empreendedorismo e Inovação**, v. 2, n. 1, p. 344–359, 2020.

CHARTE, David; CHARTE, Francisco; GARCÍA, Salvador; JESUS, María J.del; HERRERA, Francisco. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. **Information Fusion**, v. 44, p. 78–96, 2018. DOI: 10.1016/j.inffus.2017.12.007.

DAVIES, David L.; BOULDIN, Donald W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 2, PAMI-1, p. 224–227, 1979. DOI: 10.1109/TPAMI.1979.4766909.

DONFOUET, Hermann Pythagore Pierre; BARCZAK, Aleksandra; DÉTANG-DESSENDRE, Cécile; MAIGNÉ, Elise. Crop Production and Crop Diversity in France: A Spatial Analysis. **Ecological Economics**, v. 134, p. 29–39, 2017.

DRESS, Korbinian; LESSMANN, Stefan; METTENHEIM, Hans-Jörgvon. Residual value forecasting using asymmetric cost functions. **International Journal of Forecasting**, v. 34, n. 4, p. 551–565, 2018. DOI: 10.1016/j.ijforecast.2018.01.008.

DU, Guowang; ZHOU, Lihua; YANG, Yudi; LÜ, Kevin; WANG, Lizhen. Deep Multiple Auto-Encoder-Based Multi-view Clustering. **Data Science and Engineering**, v. 6, p. 323–338, 2021. DOI: 10.1007/s41019-021-00159-z.

FALISSARD, L.; FAGHREAZZI, G.; HOWARD, N.; FALISSARD, B. Deep clustering of longitudinal data. **ArXiv**, 2018.

FATCH, Paul; MASANGANO, Charles; HILGER, Thomas; JORDAN, Irmgard; MAMBO, Isaac; FRANCESCA, Judith; KAMOTO, Mangani; KALIMBIRA, Alexander; NUPPENAU, Ernst-August. Holistic agricultural diversity index as a measure of agricultural diversity: A cross-sectional study of smallholder farmers in Lilongwe district of Malawi. **Agricultural Systems**, v. 187, p. 102991, 2021.

GENOLINI, Christophe; ALACOQUE, Xavier; SENTENAC, Mariane; ARNAUD, Catherine. kml and kml3d: R Packages to Cluster Longitudinal Data. **Journal of Statistical Software**, v. 65, n. 4, p. 1–34, 2015. DOI: 10.18637/jss.v065.i04.

GUO, X.; LIU, X.; ZHU, E.; YIN, J. Deep Clustering with Convolutional Autoencoders. **Lecture Notes in Computer Science**, n. 10635, p. 373–382, 2017. DOI: 10.1007/978-3-319-70096-0_39.

GUPTA, D.; HAZARIKA, B. B.; BERLIN, M. Robust regularized extreme learning machine with asymmetric huber loss function. **Neural Computing and Applications**, v. 32, p. 12971–12998, 2020.

HALKIDI, M.; VAZIRGIANNIS, M. A density-based cluster validity approach using multi-representatives. **Pattern Recognition Letters**, v. 29, p. 773–786, 2008.

HALKO, Nathan; MARTINSSON, Per-Gunnar; TROPP, Joel A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. arXiv, 2009. DOI: 10.48550/ARXIV.0909.4061.

HUBER, P. J. Robust estimation of a location parameter. **The Annals of Mathematical Statistics**, v. 35, n. 1, p. 73–101, 1964.

IBGE. **Tabelas 74, 94, 289, 291, 1612, 1613, 3939 e 3940: sistema IBGE de Recuperação Automática**. Rio de Janeiro: IBGE, 2021. Available at https://sidra.ibge.gov.br (2021/06/15).

KHATUN, N.; MATIN, M. A. A Study on LINEX Loss Function with Different Estimating Methods. **Open Journal of Statistics**, v. 10, p. 52–63, 2020. DOI: 10.4236/ojs.2020.101004.

KOHONEN, T. Essentials of the self-organizing map. **Neural Networks**, v. 37, p. 52–65, 2013.

KOHONEN, Teuvo. **Self-Organizing Maps**. Berlin: Springer, 2001.

KOHONEN, Teuvo; HYNNINEN, Jussi; KANGAS, Jari; LAAKSONEN, Jorma. **SOM PAK: The Self-Organizing Map Program Package**. A31. Espoo: Finland, 1996.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**, v. 29, n. 1, p. 1–27, 1964.

KUHN, H. W. The Hungarian method for the assignment problem. **Naval Research Logistics Quarterly**, v. 2, n. 1-2, p. 83–97, 1955. DOI: 10.1002/nav.3800020109.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015. DOI: 10.1038/nature14539.

MIN, E.; GUO, X.; LIU, Q.; ZHANG, G.; CUI, J.; LONG, J. A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture. **IEEE Access**, v. 6, p. 39501–39514, 2018. DOI: 10.1109/ACCESS.2018.2855437.

MOHAMMED, M.A.; ALSHANBARI, Huda M.; EL-BAGOURY, Abdal-Aziz H. Application of the LINEX Loss Function with a Fundamental Derivation of Liu Estimator. **Computational Intelligence and Neuroscience**, n. 2307911, p. 1–9, 2022. Artificial Intelligence and Machine Learning-Driven Decision-Making. DOI: 10.1155/2022/2307911.

MÜLLER, Klaus-Robert; MIKA, Sebastian; RÄTSCH, Gunnar; TSUDA, Koji; SCHÖLKOPF, Bernhard. An Introduction to Kernel-Based Learning Algorithms. **IEEE Transactions on Neural Networks**, v. 12, n. 2, p. 181–201, 2001.

PIEDRA-BONILLA, Elena Beatriz; BRAGA, Cícero Augusto S.; BRAGA, Marcelo José. Diversificação agropecuária no Brasil: conceitos e aplicações em nível municipal. **Revista de Agronomia e Agronegócio**, v. 18, n. 2, p. 1–28, 2020.

ROUSSEEUW, Peter J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

SALES, C.M.C.F.; RODRIGUES, R.N. Espaço rural brasileiro: diversificação e peculiaridades. **Revista Espinhaço**, v. 8, n. 1, p. 54–65, 2019. DOI: 10.5281/zenodo.3345145.

SAMBUICHI, R.H.R.; GALINDO, E.P.; PEREIRA, R.M.; CONSTANTINO, M.; RABETTI, M.d.S. **Diversidade da produção nos estabelecimentos da agricultura familiar no Brasil: uma análise econométrica baseada no cadastro da declaração de aptidão ao PRONAF (DAP)**. v. 2202. Brasília: Rio de Janeiro, 2016. (Texto para discussão).

SCHÖLKOPF, Bernhard; SMOLA, Alex; MÜLLER, Klaus-Robert. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. **Neural Computation**, v. 10, n. 5, p. 1299–1319, 1998.

SHANNON, Claude E. A mathematical theory of communication. **The Bell system technical journal**, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SILVA, M. A. S. da; MATOS, Leonardo N.; O. SANTOS, Flávio E. de; DOMPIERI, Márcia H. G.; MOURA, Fábio R. de. **Data and R script - Tracking the Connection Between Brazilian Agricultural Diversity and Native Vegetation Change by a Machine Learning Approach.** São Francisco: Github, 2022. Available at: <https://github.com/marcos-silva-inf/SOMSpatialPanelData>.

SILVA, M. A. S. da; MATOS, Leonardo Nogueira; SANTOS, Flavio Emanuel de Oliveira; DOMPIERI, Marcia Helena Galina; MOURA, Fabio Rodrigues de. Tracking the Connection Between Brazilian Agricultural Diversity and Native Vegetation Change by a Machine Learning Approach. **IEEE Latin America Transactions**, v. 20, n. 11, p. 2371–2380, ago. 2022. Special Issue on Artificial Intelligence for Sustainability. DOI: 10.1109/tla.2022.9904762.

SIMPSON, Edward H. Measurement of diversity. **Nature**, v. 163, n. 4148, p. 688–688, 1949.

SONG, C.; Y, Y Huang; LIU, F.; WANG, Z.; WANG, L. Deep auto-encoder based clustering. **Intelligent Data Analysis**, v. 18, n. 6, s65–s76, 2014. DOI: 10.3233/IDA-140709.

TEIXEIRA, M.L.C.; RIBEIRO, S.M.C. Agricultura e paisagens sustentáveis: a diversidade produtiva do setor agrícola de Minas Gerais, Brasil. **Sustainability in Debate**, v. 11, n. 2, p. 29–41, 2020.

TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, p. 2319–2323, 2000.

VARIAN, H. R. A bayesian approach to real estate assessment. **Studies in Bayesian Econometric and Statistics in Honor of Leonard J. Savage**, v. 5, p. 195–208, 1975.

VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. **Journal of Machine Learning Research**, v. 11, p. 2837–2854, 2010.

XU, Chaoyang; DAI, Yuanfei; LIN, Renjie; WANG, Shiping. Deep clustering by maximizing mutual information in variational auto-encoder. **Knowledge-Based Systems**, v. 205, n. 106260, set. 2020. DOI: 10.1016/j.knosys.2020.106260.

First author biography

Marcos Aurélio Santos da Silva was born in Aracaju, Sergipe, Brazil. He holds a bachelor's degree in Computer Science from the Federal University of Sergipe (UFS), a master's in Applied Computing from the National Institute for Space Research (INPE), and a Ph.D. in Computer Science from the University of Toulouse 1 Capitole (UT1), France. Since 2006, he has been working as a researcher at the Brazilian Agricultural Research Corporation (Embrapa) at the Center for Agricultural Research on Coastal Tablelands, Aracaju, developing studies in social modeling and simulation, and applying artificial neural networks and spatial statistics to territorial typification, regionalization, and zoning.