Autoencoder-based feature extraction of spatial panel data for Brazilian agricultural heterogeneity cluster analysis

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

[1] Department of Computer Science – UFS, 49100-000, São Cristóvão, SE, Brazil

[2]Embrapa Coastal Tablelands, Av. Beira Mar, 3250, 49025-040, Aracaju, SE, Brazil

[3]Embrapa Territorial, Av. Soldado Passarinho, 303, 13070-115, Campinas, SP, Brazil

[4]Department of Economics – UFS, 49100-000, São Cristóvão, SE, Brazil

flavioemanuel859@gmail.com, marcos.santos-silva@embrapa.br

leonardo@dcomp.ufs.br, marcia.dompieri@embrapa.br, fabiromoura@gmail.com

Abstract. *Brazilian agricultural production presents a high degree of spatial diversity, which challenges designing territorial public policies to promote sustainable development. This article proposes a new approach to cluster Brazilian municipalities according to their agricultural production. It combines a feature extraction mechanism using Deep Learning based on Autoencoders and clustering based on k-means and Self-Organizing Maps. We clustered the panel data from IBGE's annual estimates of Brazilian agricultural production between 1999 and 2018. The results show that in comparison with the ground truth adopted, the autoencoder model combined with the Self-Organizing Maps and the k-means algorithm presented a better result than clustering the raw data using k-means. It demonstrated the ability of simple stacked autoencoders to reduce the dimensionality and create a new space of features in their latent layer where the data can be analyzed and clustered.*

1. Introduction

At the landscape level, Brazilian agricultural output exhibits various degrees of variation [Sales and Rodrigues 2019]. Each region specializes in some agricultural activities, from high-tech agribusiness to family farming [Teixeira and Ribeiro 2020], due to climate, water resources, geographic limitations, and past socioeconomic processes. Furthermore, this heterogeneity is visible at local and regional scales.

Grouping municipalities according to their shared agricultural production makes it possible to identify the regional particularities that public authorities must consider when creating territorial public policies. [Silva et al. 2022] divided the Brazilian municipalities into eight agricultural production diversity trend groups, each associated with a distinct degree of native vegetation alteration, using clustering analysis and feature engineering. Their work 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 animal), totaling eight variables for 20 years and 5570 municipalities. The raw spatial panel data

they used comprises 196 variables for 20 years. Adopting a feature engineering strategy decreased the number of variables from 196 to 8 and eliminated various data issues like a massive number of zeros and values close to zero. 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. Their research provided compelling evidence for the claim made by [Fatch et al. 2021] that low-diversified regions typically exhibit a low degree of sustainability.

This paper used a Deep Learning strategy based on autoencoders to extract features from raw Brazilian agricultural spatial panel data to cluster the municipalities. The clustering obtained by [Silva et al. 2022] serves as the ground truth. Over the data projected on the new feature space, we used two classical clustering algorithms, k-means and Self-Organizing Maps associated with the k-means. 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 clustering using a deep learning technique known as deep clustering [LeCun et al. 2015]. Therefore, the research questions are: What would happen if we directly ran a clustering analysis on the unprocessed panel data? How could we replace the feature engineering employed by [Silva et al. 2022] with a Deep Learning feature extraction?

There are few works on tabular panel data, as in [Falissard et al. 2018], and 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].

We organized this paper as follows: section 2 discuss the dataset and the proposed approach to feature extraction and spatial panel data clustering; section 3 shows the results and discussion; and section 4 unveil the conclusions.

2. Data and methods

2.1. Spatial panel data

The dataset comprises 196 variables of IBGE's annual estimates for all Brazilian municipalities [IBGE 2021]. These variables correspond to eight groups: herd population, animal production value, planted temporary crops, silviculture, aquaculture, vegetal extractivism, and temporary and permanent crop production value. A detailed data description can be found in [Silva et al. 2022].

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) (Fig. 1). After that, we linearly normalized the data according to the min-max algorithm transforming all variables into the interval $[0, 1]$.

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

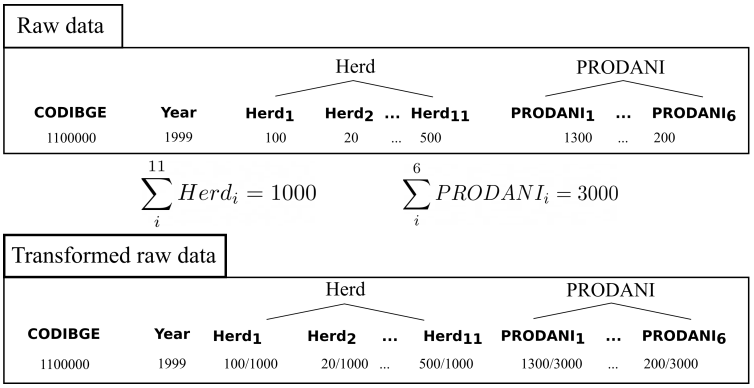
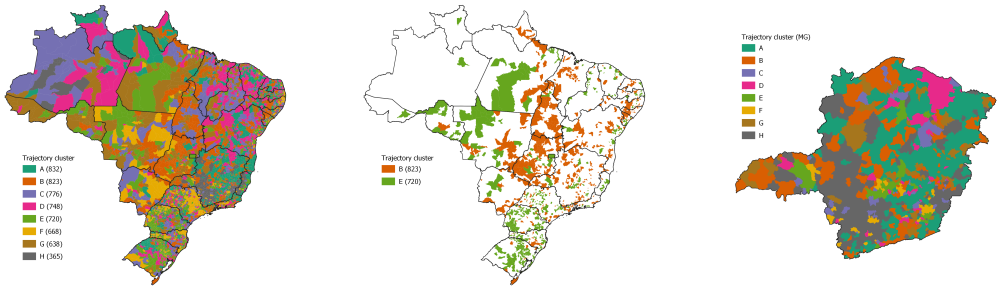


Figure 1. We converted each variable to its unit rate according to the category (e.g., herd or prodani) it belongs. Source: elaborated by the authors.



(a) Trajectory cluster **(b)** Regional patterns **(c)** Intra-regional pattern

Figure 2. a) Clustered municipalities based on the trajectory of Shannon's diversity indices onto a Self-Organizing Map by [Silva et al. 2022]. b) Mapping only two trajectory clusters to highlight regional patterns of clusters 'B' and 'E'. c) The same clustering for the Minas Gerais state, showing intra-regional patterns. Source: elaborated by the authors.

91.49%, respectively. They are structural zeros because most municipalities produce a limited amount of agricultural products, so no available data were set to zero. This unbalanced data challenges the learning process of artificial neural networks that are induced 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.

2.2. Proposed approach

The proposed approach comprises two steps. In the first step, we defined a set of autoencoder models (section 2.2.1), and performed a parameterization of an asymmetric loss function to cope with the data sparsity (section 2.2.2). The next step is responsible for clustering the encoded data onto a new and low dimensional feature space for each autoencoder model (section 2.2.3).

Table 1. Encoder layers structure (number of neurons per hidden layer). Source: elaborated by the authors.

ID	Latent*	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**

*Number of neurons on the latent layer.

2.2.1. 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 (Fig. 3). 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.

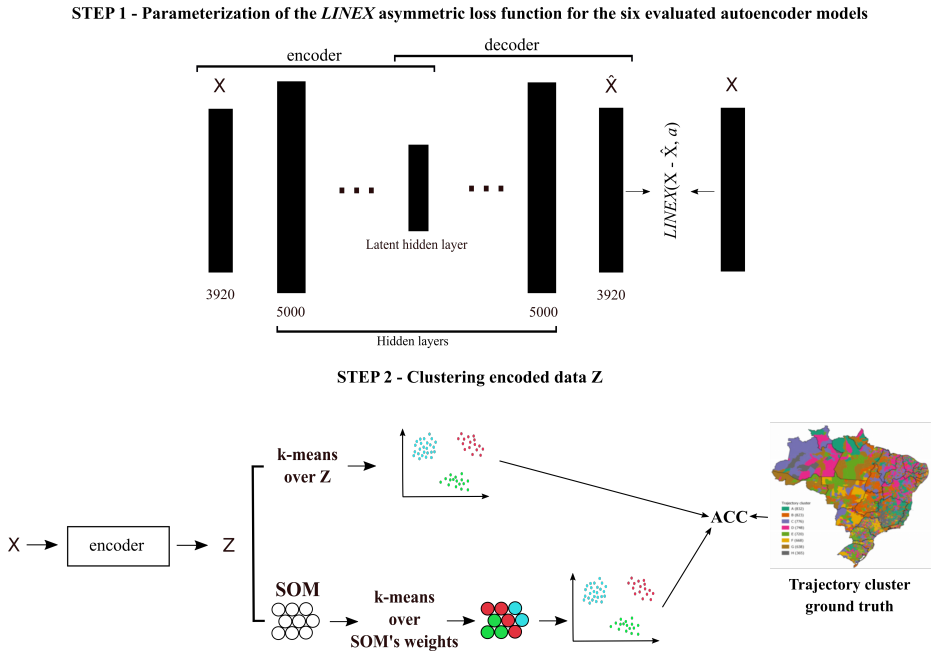


Figure 3. General autoencoder structure used in this paper. Source: elaborated by the authors.

2.2.2. Asymmetric loss function

Asymmetric loss functions are suitable for cases where bias is relevant or a considerable imbalance in data representation, as in our case (very sparse data), or fault detection [Dress et al. 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 et al. 2020] proposed an asymmetric function based on [Huber 1964] that is quadratic when the error is small but is like mean absolute error when the error is larger than a threshold.

As stated in section 2.1, 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 of the zero [Khatun and 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 et al. 2022, Khatun and 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 $a \geq 5.0$.

2.2.3. Clustering encoded data

To establish a baseline, we clustered the transformed raw data using a k-means algorithm based on joint-trajectories [Genolini et al. 2015] over all 5570 municipalities, 196 variables for 20 years, for $k = 8$ to be able to compare with the results obtained by [Silva et al. 2022].

After defining the a value for the LINEX loss function, we performed Deep Learning using the entire dataset for each autoencoder model. We conducted clustering on the encoded data using the k-means and the Self-Organizing Map (SOM) [Kohonen 2001] for encoded data reduction and k-means on the SOM's weights. To compare these two methods, we used four clustering validity indices: Silhouette [Rousseeuw 1987], Davies-Bouldin [Davies and Bouldin 1979] using centroids and medoids, and CDbw [Halkidi and Vazirgiannis 2008]. We conducted all clustering considering $k = 8$.

The six deep clustering and the k-means were compared using an accuracy (ACC) measure for the eight clusters (Eq. 2), where b_i represents the ground truth for the municipality i, c_i the cluster obtained by the evaluated clustering method and m a mapping function based on the Hungarian method [Kuhn 1955] to match k_i and c_i. A greater ACC means a good match between our clustering and the one obtained by [Silva et al. 2022].

$$ACC = max_m \frac{\sum_{i=1}^{n} \mathbf{1}(b_i = m(c_i))}{n} \quad (2)$$

To check for spatial dependence and regional and intra-regional distinction, we mapped the clustering into the Brazilian municipal geographical map.

2.3. Software

We modeled the deep neural networks using Python and Keras framework, clustered using R packages and used SOMPAK for the Self-Organizing Map processing. The geographical maps were generated by QGIS version 3.6.

3. Results and discussion

3.1. $LINEX$ parameterization

Fig. 4 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 the MSE loss function and for different values for the a parameter for the LINEX loss function. 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. For both training and test datasets, the LINEX loss function with $a = 7$ presents the best result for all autoencoder models. In fact, the performance decreases for $a < 7$ and $a > 7$.

3.2. Clustering encoded data Z

After defining a value for the LINEX loss function parameter a, we presented all datasets to the deep learning process for each evaluated autoencoder model. After that, we encoded the data reducing its dimensionality to the latent layer dimension. Then, we proceeded to the clustering using two strategies: a) applying k-means over the encoded data; b) using the Self-Organizing Map as an encoded data ordering and reduction and k-means over the SOM's weights after an unsupervised shallow learning process. We evaluated three SOM sizes (8×6, 10×15, and 20×15), and we chose the big one because it presented the best quantization error.

The clustering validity indices results for both k-means and SOM+kmeans strategies suggest that we achieved the best solution with the lowest number of neurons in the latent layer. The Davies-Bouldin presents the same behavior using centroids and medoids as centrality references. Hence, they decay until the smallest values for the autoencoder VI with 10 neurons on the latent layer. The Silhouette index increases while the number of neurons on the latent layer decreases for both clustering methods. The CDbw validity index presents a more erratic behavior for the k-means clustering and a more smooth curve for the SOM 20×15 + k-means approach. All this suggests that the autoencoder would achieve the best data partition with fewer neurons on the latent layer. Still, an inspection of the geographic projection and the accuracy measure shows that the autoencoder with more neurons on the latent layer clustered the Brazilian municipalities more appropriately.

Fig. 5 shows the clustering accuracy when compared with the ground truth (Fig. 2a) for all autoencoder models and clustering algorithms. We observe that clustering using

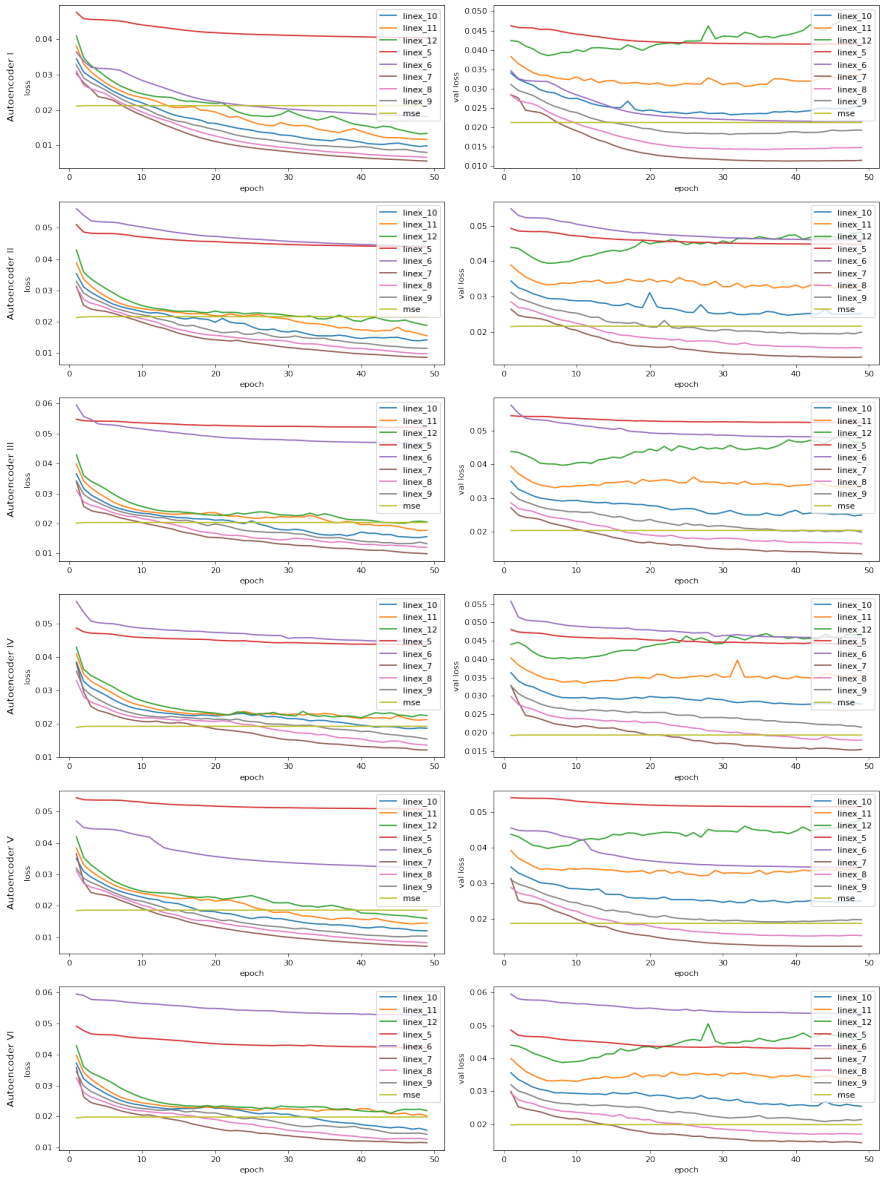


Figure 4. Mean loss values for train and test data considering different loss functions (MSE and LINEX for different a values) considering fifty runs for each autoencoder model. Source: elaborated by the authors.

SOM + k-means strategy performed better in most situations, suggesting that clusters may have a non-convex structure.

Figs. 6 (a) and (c) show the geographic mapping for the k-means partition over the raw data X and for the best clustering considering the ACC as a metric. The k-means clustering over the raw transformed data showed a solid regional spatial dependence, highlighting a clear distinction between the semi-arid region, the states of Amazon, Minas Gerais, and São Paulo, and the Brazilian South (Fig. 6a). The partition using autoencoder+SOM+k-means also identified regional patterns, separating Minas Gerais from the rest of the country. The Center-West was split into many groups, with Mato Grosso do Sul showing more similarity with the Paraná state.

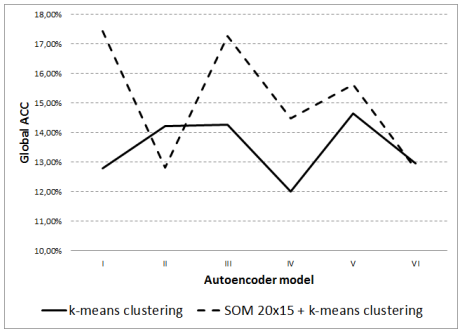
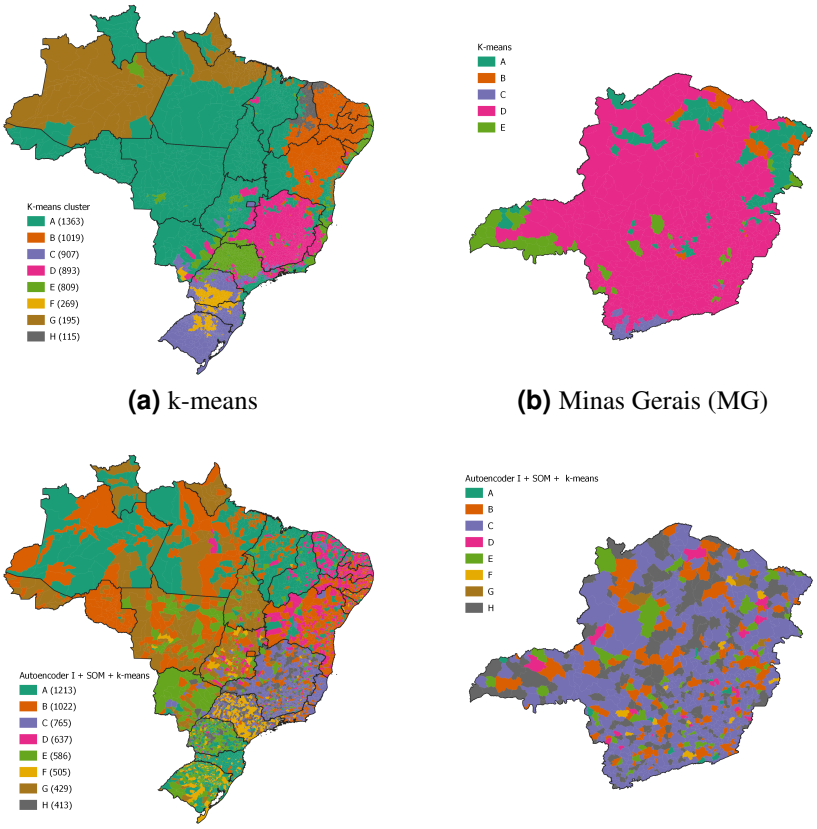


Figure 5. Clustering accuracy (ACC) considering the trajectory cluster from [Silva et al. 2022] as the ground truth. Source: elaborated by the authors.

Figs. 6 (b) and (d) show intra-regional distinction in the Minas Gerais (MG) Brazilian state for k-means and autoencoder+SOM+k-means strategies. The k-means algorithm did not identify intra-regional patterns in MG and put in one group almost all municipalities of the Center-West and North regions Fig. 6. The clustering based on autoencoder+SOM+k-means identified more intra-regional patterns as observed in Minas Gerais (MG) in Fig. 6d.



(a) k-means

(b) Minas Gerais (MG)

(c) autoencoder I + SOM + k-means

(d) Minas Gerais (MG)

Figure 6. Geographic mapping of the k-means clustering applied directly over the transformed raw data (a) and MG (b). Geographic mapping of deep clustering for the autoencoder model I + SOM + k-means (c) and MG (d). Source: elaborated by the authors.

4. Conclusions

Performing a clustering analysis over the raw data using a simple k-means algorithm showed that the clusters divide the municipalities showing a huge regional partition but did not identify any intra-regional distinctions. On the other hand, the Deep Clustering proceeded in two steps. Dimensionality reduction by autoencoders and clustering of this new data representation using SOM and k-means improved the general accuracy compared to the k-means strategy over the raw data.

Future work should include evaluating larger latent layers and other nonlinear dimensionality reduction techniques such as Isomap [Tenenbaum et al. 2000], and kernel PCA [Müller et al. 2001], 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

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

Berk, R. (2011). Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, 27(1):107–123.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(PAMI-1):224–227.

Dress, K., Lessmann, S., and Mettenheim, H.-J. (2018). Residual value forecasting using asymmetric cost functions. *International Journal of Forecasting*, 34(4):551–565.

Du, G., Zhou, L., Yang, Y., Lü, K., and Wang, L. (2021). Deep multiple auto-encoder-based multi-view clustering. *Data Science and Engineering*, 6:323–338. 10.1007/s41019-021-00159-z.

Falissard, L., Faghreazzi, G., Howard, N., and Falissard, B. (2018). Deep clustering of longitudinal data. *ArXiv*.

Fatch, P., Masangano, C., Hilger, T., Jordan, I., Mambo, I., Francesca, J., Kamoto, M., Kalimbira, A., and Nuppenau, E.-A. (2021). Holistic agricultural diversity index as a measure of agricultural diversity: A cross-sectional study of smallholder farmers in Lilongwe district of Malawi. *Agricultural Systems*, 187:102991.

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.

Gupta, D., Hazarika, B. B., and Berlin, M. (2020). Robust regularized extreme learning machine with asymmetric huber loss function. *Neural Computing and Applications*, 32:12971–12998.

Halkidi, M. and Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29:773–786.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

IBGE (2021). 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).

Khatun, N. and Matin, M. A. (2020). A study on linex loss function with different estimating methods. *Open Journal of Statistics*, 10:52–63.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Mohammed, M., Alshanbari, H. M., and El-Bagoury, A.-A. H. (2022). Application of the linex loss function with a fundamental derivation of liu estimator. *Computational Intelligence and Neuroscience*, (2307911):–. Artificial Intelligence and Machine Learning-Driven Decision-Making.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 12(2):181–201.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65.

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

Silva, M. A. S. d., Matos, L. N., Santos, F. E. d. O., Dompieri, M. H. G., and Moura, F. R. d. (2022). Tracking the connection between brazilian agricultural diversity and native vegetation change by a machine learning approach. *IEEE Latin America Transactions*, 20(11):2371–2380.

Song, C., Y, Y. H., Liu, F., Wang, Z., and Wang, L. (2014). Deep auto-encoder based clustering. *Intelligent Data Analysis*, 18(6):S65–S76. 10.3233/IDA-140709.

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.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

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

Xu, C., Dai, Y., Lin, R., and Wang, S. (2020). Deep clustering by maximizing mutual information in variational auto-encoder. *Knowledge-Based Systems*, 205(106260). 10.1016/j.knosys.2020.106260.