

# Feature engineering vs. extraction: clustering Brazilian municipalities through spatial panel agricultural data via autoencoders

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

<sup>1</sup>Embrapa Coastal Tablelands, Av. Beira Mar, 3250, 49025-040, Aracaju, SE, Brazil

<sup>2</sup>Department of Computer Science – UFS, 49100-000, São Cristóvão, SE, Brazil

<sup>3</sup>Embrapa Territorial, Av. Soldado Passarinho, 303, 13070-115, Campinas, SP, Brazil

<sup>4</sup>Department of Economics – UFS, 49100-000, São Cristóvão, SE, Brazil

marcos.santos-silva@embrapa.br

**Abstract.** *This article compares the clustering of Brazilian municipalities according to their agricultural diversity using two approaches, one based on feature engineering and the other based on feature extraction using Deep Learning based on autoencoders and cluster analysis based on k-means and Self-Organizing Maps. The analyzes were conducted from panel data referring to IBGE's annual estimates of Brazilian agricultural production between 1999 and 2018. Different structures of simple stacked undercomplete autoencoders were analyzed, varying the number of layers and neurons in each of them, including the latent layer. The asymmetric exponential linear loss function was also evaluated to cope with the sparse data. The results show that in comparison with the ground truth adopted, the autoencoder model combined with the k-means presented a superior result than the clustering of the raw data from the k-means, demonstrating the ability of simple autoencoders to represent from their latent layer important features of the data. Although the general accuracy is low, the results are promising, considering that we evaluated the most simple strategy for Deep Clustering.*

## 1. Introduction

Brazilian agricultural production presents different levels of heterogeneity at a landscape scale [Sales and Rodrigues 2019, Sambuichi et al. 2016, Schneider and Cassol 2014]. Due to climate, water resources, geographic constraints, and historical socioeconomic processes, each region specializes in a set of agricultural activities, from high technological agribusiness to familiar smallholders [Teixeira and Ribeiro 2020]. Moreover, we can observe this heterogeneity at regional and intra-regional scales.

As observed by [Tisdell et al. 2019], this agricultural production variation implies different impacts on sustainability in rural space, from native vegetation change to employment rates. It suggests that any territorial public policy should consider incentives to promote agricultural diversification, focusing on diminishing the environmental impact, creating more rural job opportunities, and increasing economic stability at a landscape level.

Designing territorial public policies demands the identification of regional particularities, and we can achieve this by clustering the regions according to their agricultural production similarities. [Silva et al. 2022] combined featuring engineering and clustering analysis to divide the Brazilian municipalities into eight agricultural production diversity trends groups, linking each group to a different level of native vegetation change. Their findings showed evidence that lowly diversified regions tend to present a low degree of sustainability, also observed by [Fatch et al. 2021].

[Silva et al. 2022] used the IBGE's estimates of annual agricultural production from 1999 to 2018 and calculated 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), so eight variables for 20 years and 5570 municipalities. Then, they applied a shallow learning algorithm based on the Self-Organizing Map Artificial Neural Network combined with the k-means to cluster the spatial panel data.

The raw spatial panel data used by [Silva et al. 2022] comprises 196 variables for 20 years. The use of a feature engineering method reduced from 196 to 8 variables and diminished some data challenges as a massive presence of zeros and lots of values near zero. So, the research questions are: What if we performed a clustering analysis directly over the raw panel data? How could a Deep Learning (DL) feature extraction substitute the feature engineering used by [Silva et al. 2022]?

This paper applied a Deep Learning strategy based on autoencoders to feature extract on the raw Brazilian agricultural spatial panel data to cluster the data, considering the clustering obtained by [Silva et al. 2022] as the ground truth. We used two classical clustering algorithms over the extracted features, k-means and Self-Organizing Maps associated with the k-means.

The paper is organized as follows: section 2 presents a brief review on Deep Clustering with autoencoders; section 3 discuss about 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**

Deep learning is a consolidated field in industry and is responsible for a significant transformation in data analysis, mainly in image, video, and text processing. Nevertheless, there are many research challenges, such as in clustering using the DL technique, known as Deep Clustering [LeCun et al. 2015].

In a Deep Clustering survey, [Min et al. 2018] identified a myriad of DL model architectures based mainly on autoencoders. Autoencoders are deep artificial neural networks that use an unsupervised learning method to extract features from a dataset by combining an encoder (non-linear mapping function) and a decoder (dataset reconstructor from the representation generated by the encoder). The latent layer, where resides the data representation, can be used to support data clustering while it gathers all necessary information to reconstruct the entire dataset. Nonetheless, this representation on the latent layer does not preserve the original data topology and proximity characteristics.

As stated by [Min et al. 2018], there are at least two ways to perform Deep Clus-

tering using autoencoders. First, use the autoencoder to reduce data dimensionality and apply a clustering algorithm over the encoded data. Second, combine the reconstruction loss with a clustering loss to cluster data while the autoencoder updates parameters in the deep learning process [Charte et al. 2018, Song et al. 2014, Guo et al. 2017]. In fact, the Deep Clustering strategy depends on the dataset properties such as sparsity, size, format (images, sequences, and tabular), and structural complexity, and this will demand different approaches as proposed by [Du et al. 2021] for text and image clustering by a deep multi-view clustering algorithm based on multiple auto-encoder, [Xu et al. 2020] for image clustering using variational autoencoders or [Falissard et al. 2018] for longitudinal tabular data combining recurrent neural network and autoencoders.

From the literature review, we concluded 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 to investigate how autoencoders can map the original tabular panel data to a new latent feature space is following incremental addition of complexity to the model. First, exploring the data clustering directly from the encoded data [Falissard et al. 2018]; second, evaluating the combination of objective and clustering loss functions [Song et al. 2014]; and finally, testing more complex Deep Clustering propositions as in [Du et al. 2021, Xu et al. 2020].

### **3. Data and methods**

#### **3.1. Spatial panel data**

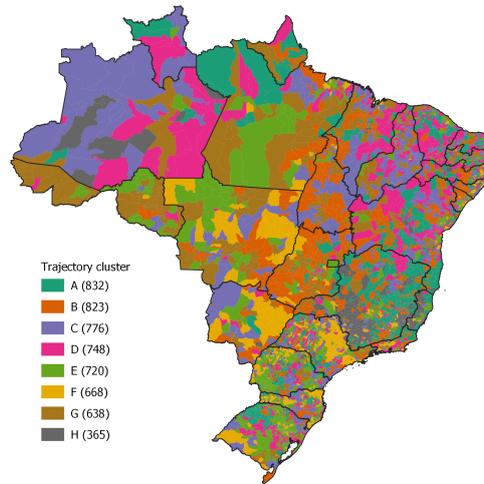
The dataset comprises 196 variables of IBGE’s annual estimates for all Brazilian municipalities [IBGE 2021]. These variables correspond to eight groups: herd population, animal production value, planted temporary crops, silviculture, aquaculture, vegetal extractivisme, and temporary and permanent crop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

The raw data were transformed 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 to. In the end, each variable will correspond to the unit rate of that product for each observation (municipality-year). 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 in the entire dataset are 83.09% and 91.49%, respectively. In fact, 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 will be 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, considering symmetric loss functions or some asymmetric loss functions such as linear and quadratic. Thus, this demanded research for a suitable loss function for a structurally sparse dataset.

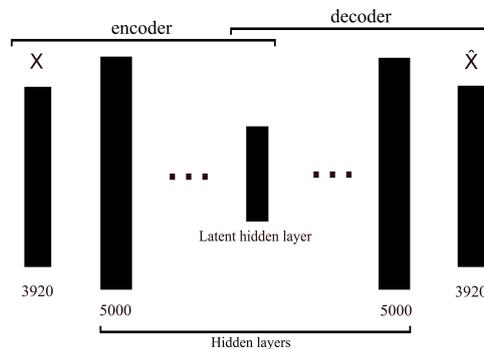
#### **3.2. Autoencoder models**

We have chosen six simple stacked undercomplete autoencoder models with the same optimizer (adam), hidden and output activation functions (relu and sigmoid), fully con-



**Figure 1. Clustered municipalities based on the trajectory of Shannon's diversity indices onto a Self-Organizing Map and considered the ground truth for this paper. Source:[Silva et al. 2022].**

nected, with the same loss function, and varying the number of hidden layers and the size of the latent layer (Fig. 2). We have defined the autoencoder models suitable for sparse data dimensionality reduction according to [Min et al. 2018, Charte et al. 2018].



**Figure 2. General autoencoder structure used in this paper. Source: elaborated by the authors.**

Table 1 shows the number of the hidden layers, defined empirically according to [Charte et al. 2018], to the encoder component of each evaluated autoencoder, including the number of neurons on the latent hidden layer.

### 3.3. Asymmetric loss function

The asymmetric loss function is applicable for situations where the direction of deviation is relevant. For example, [Dress et al. 2018] used an asymmetric loss function to improve the forecast of resale prices of leasing cars where there is a trade-off between overestimation (sell the cars by a lower price, implying that the lessor could have offered higher leasing rates) and underestimation (sell the cars by the highest price, suggesting that the lessor could have offered a lower leasing rate and accumulate more contracts). The authors showed that using these functions reduced costs by about 8%.

**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- <b>500</b>
II	250	3920-5000-3000-2000-1000-500- <b>250</b>
III	100	3920-5000-3000-2000-1000-500-250- <b>100</b>
IV	50	3920-5000-3000-2000-1000-500-250-100- <b>50</b>
V	25	3920-5000-3000-2000-1000-500-250-100-50- <b>25</b>
VI	10	3920-5000-3000-2000-1000-500-250-100-50-25- <b>10</b>

\*Number of neurons on the latent layer.

As stated in section 3.1, the standard mean square error loss function and the linear and quadratic asymmetric 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  $|a| \rightarrow 0$  then the  $LINEX(\hat{x}_i) \rightarrow MSE(\hat{x}_i)$ , 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].

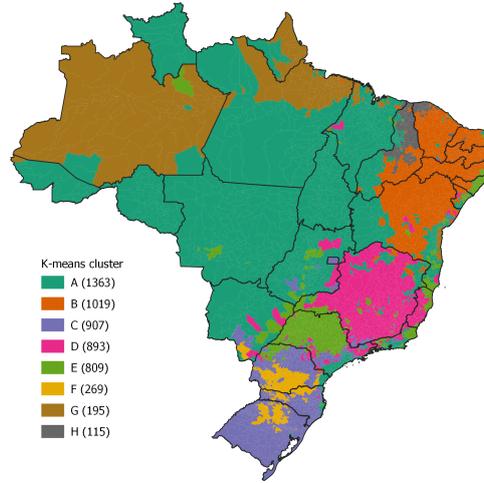
$$LINEX(\hat{x}_i) = \frac{1}{n} \sum_{i=1}^n \frac{2}{a^2} \left( e^{a(x_i - \hat{x}_i)} - a(x_i - \hat{x}_i) - 1 \right) \quad (1)$$

We evaluated  $a \in \{5.0, 10.0, 15.0, 20.0, 25.0\}$  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  $a \geq 5.0$ .

### 3.4. Clustering

To establish a baseline, we clustered the transformed raw data using a k-means algorithm based on joint-trajectories over all 5570 municipalities, 196 variables for 20 years for  $k = 8$  [Genolini et al. 2015].

After defining the  $a$  value for the linex loss function, we performed Deep Learning using the entire transformed raw data for each autoencoder model. We conducted clustering on the encoded data using the k-means, suitable for convex data structures, and the Self-Organizing Map (SOM) [Kohonen 2001] for encoded data reduction and ordering and k-means on the SOM's weights, which is a suitable strategy for non-convex data structures. We have chosen the SOM hyper-parameters as the neural size ( $N \times M$ ), topology, and learning rate according to [Silva et al. 2022, Kohonen 2001]. Generally, the SOM's size is defined empirically and in the function of the data structure, size, and complexity, guided by some SOM's learning quality measure as quantization error



**Figure 3. Geographic mapping of the k-means clustering,  $k = 8$ , applied directly over the transformed raw data using the kml3d function from the kml R package. Source: elaborated by the authors.**

$Eq = \sum_i^n (x_i - BMU_i) / n$ , where  $BMU$  is the neuron associated to the input  $x_i$  after the learning process.

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  $k = 8$ .

We compared the six deep clustering and the k-means using a cluster accuracy (ACC) measure (Eq. 2), where  $k_i$  represents the ground truth for the municipality  $i$ ,  $c_i$  is the cluster obtained by the evaluated clustering method and  $m$  is a mapping function based on the Hungarian method [Kuhn 1955] to match the  $k_i$  and  $c_i$ .

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

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

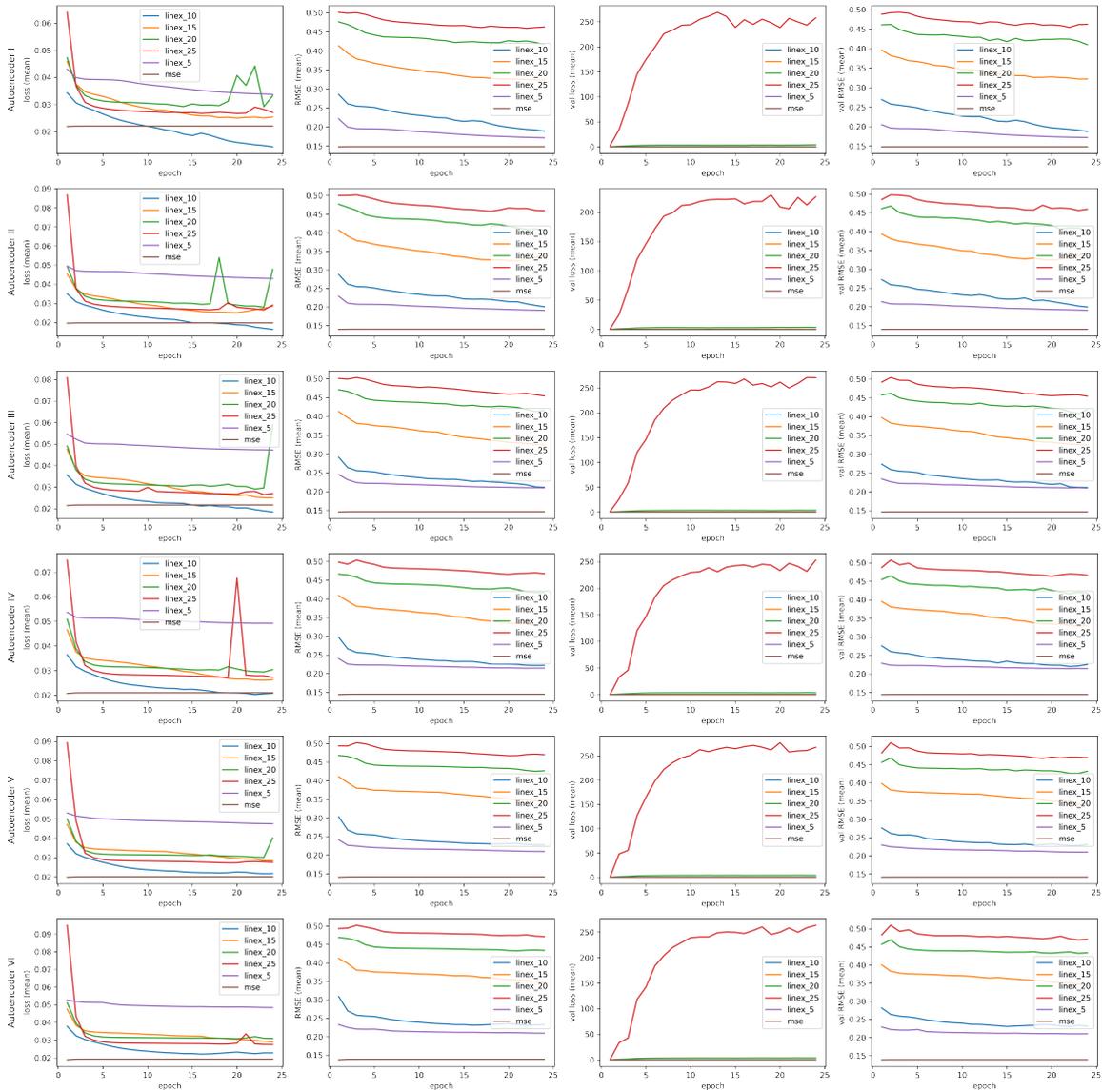
### 3.5. Software

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

## 4. Results and discussion

The k-means clustering over the transformed raw 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. Nonetheless, the k-means algorithm did not identify intra-regional patterns in these regions and put in one group almost all municipalities of the Center-West and North regions (Fig. 3).

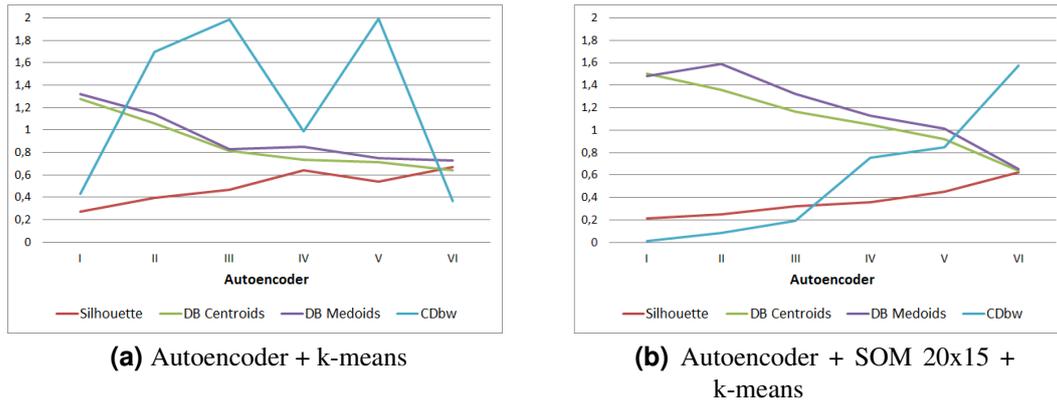
Fig. 4 shows the mean loss and RMSE for the train and test dataset for 30 runs (25 epochs each), randomly changing train and test data. We observed these values considering the MSE loss function and for five different values for the  $a$  (5, 10, 15, 20, and 25) parameter in the linex loss function. 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 the training dataset, the linex loss function with  $a = 10$  presents the best result for all autoencoder models. For the test dataset, the linex loss function with  $a \geq 20$  shows a terrible prediction performance. The RMSE quality measure shows that the linex loss function with  $a \in \{5, 10\}$  presents the best results. Based on these results, we proceed to the deep clustering using the linex loss function with  $a = 10$ .



**Figure 4. Mean values of learning metrics (loss and rmse) for train, first two columns, and test, last two columns, data considering six loss functions (mse and linear exponential for five different  $a$  values) considering thirty runs for each autoencoder structure. Source: elaborated by the authors.**

After defining the loss function parameter, 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) only k-means over the encoded data; b) and 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 ( $8 \times 6$ ,  $10 \times 15$ , and  $20 \times 15$ ), and we chose the big one because it presented the best quantization errors.

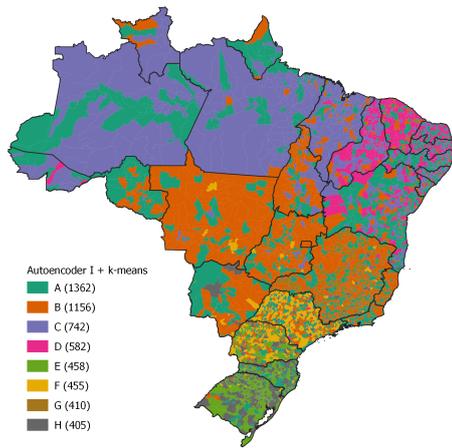
Fig. 5 shows the results for the clustering validity indices for both k-means and SOM  $20 \times 15$  + kmeans strategies. The Davies-Bouldin presents the same behavior using centroids and medoids as centrality references. Hence, they decay until the smallest values for the autoencoder VI with 10 neurons 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  $20 \times 15$  + k-means method. All this suggests that the autoencoder would achieve the best data partition with fewer neurons on the latent layer.



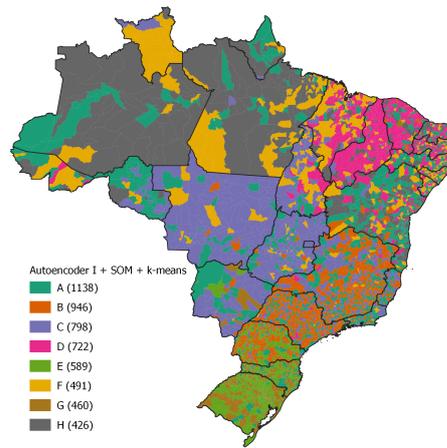
**Figure 5. Clustering validity indices for the six autoencoder structures considering the two clustering strategies (autoencoder + k-means and autoencoder + SOM + k-means). Source: elaborated by the authors.**

Fig. 6 shows the geographic mapping for the k-means and SOM+k-means clustering for the autoencoder models I, II, and VI. Despite the validity indices, only the autoencoder I with 500 neurons on the latent layer extracted enough features to allow the clustering algorithms to unveil the general regional and intra-regional distinctions (Fig. 6 a and b). In general, as the number of neurons in the latent layer decreases, the ability of clustering algorithms to identify regional and intra-regional patterns also decreases, as we can observe in the autoencoder model VI (Fig. 6 e and f).

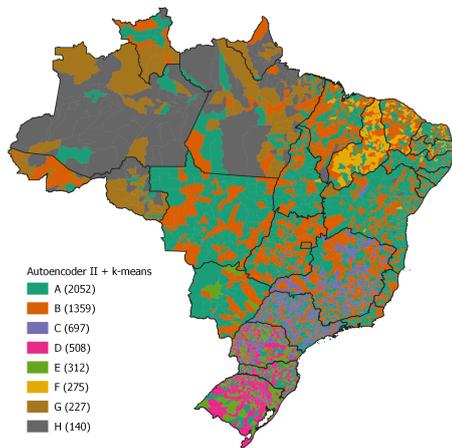
The intra-regional distinction is hard to achieve and verify. Still, taking the trajectory clustering as the ground truth, we can check how close to it the deep learning arrived by observing the global clustering accuracy and by cluster. Table 2 shows the results for all autoencoder models and clustering algorithms, including the clustering obtained by applying the k-means over the transformed raw data. The autoencoder I + k-means got the best global accuracy.



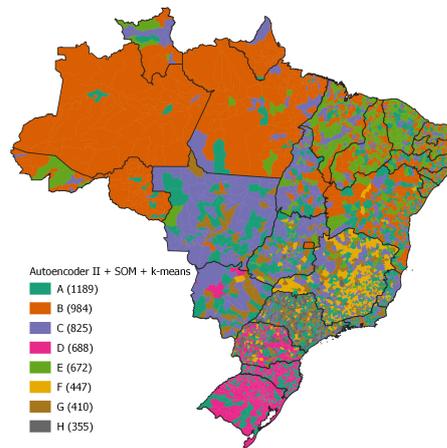
(a) Autoencoder I + k-means



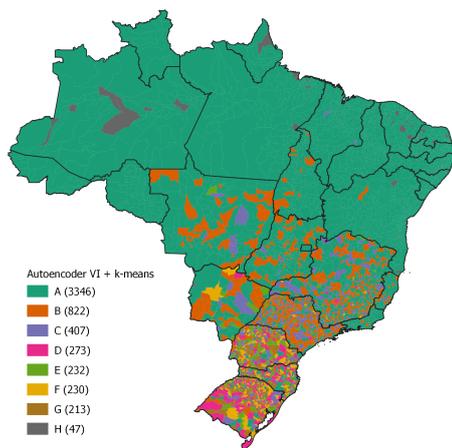
(b) Autoencoder I + SOM + k-means



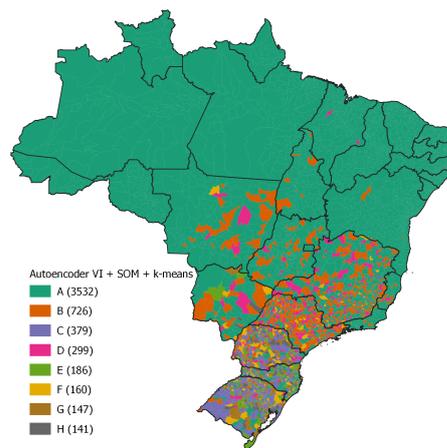
(c) Autoencoder II + k-means



(d) Autoencoder II + SOM + k-means



(e) Autoencoder VI + k-means



(f) Autoencoder VI + SOM + k-means

**Figure 6. Geographic mapping of deep clustering for the autoencoder structures I, II, and VI. In parentheses is the number of observations per cluster. Source: elaborated by the authors.**

**Table 2. Percentage of accuracy by cluster when compared with the trajectory clustering by [Silva et al. 2022]. Source: elaborated by the authors.**

Clustering method	ACC by cluster (%)								Global ACC(%)
	A	B	C	D	E	F	G	H	
Autoencoder I + K-means	26	33	25	23	23	17	4	4	20.90
Autoencoder II + K-means	35	32	7	9	13	1	8	1	14.87
Autoencoder III + K-means	35	32	9	10	14	9	3	1	15.91
Autoencoder IV + K-means	73	5	13	6	3	1	1	0	14.76
Autoencoder V + K-means	45	32	8	6	15	4	1	1	16.03
Autoencoder VI + K-means	58	18	5	5	10	6	2	1	15.04
Autoencoder I + SOM+k-means	22	16	10	23	29	5	5	2	15.19
Autoencoder II + SOM+k-means	20	18	13	12	7	7	6	6	11.97
Autoencoder III + SOM+k-means	32	37	22	10	3	7	3	1	16.27
Autoencoder IV + SOM+k-means	58	4	13	9	3	6	4	1	13.91
Autoencoder V + SOM+k-means	35	34	8	18	23	5	1	2	17.54
Autoencoder VI + SOM+k-means	62	15	5	3	9	4	1	1	14.49
K-means	12	18	11	7	13	5	5	0	9.93

## 5. Conclusions

This paper showed how to identify groups in panel data from raw data. We developed an experiment based on IBGE data on agricultural production from 1999 to 2018 in more than five thousand Brazilian municipalities. In a first approach using the k-means algorithm, we observed that when representing municipalities on a map with colors associated with groups, the obtained clustering allowed us to distinguish inter-regional characteristics but not intra-regional ones. This stems from the fact that panel data is particularly challenging due to its high dimensionality and complex spatial distribution, which is not necessarily circular around a centroid. This required a more elaborate approach. We developed a second approach using a two-step process. In the first step, the data was projected into a lower dimensional space using an autoencoder latent space. We clustered the data using the k-means algorithm and self-organizing maps (SOM) in the second step. The maps generated from the clusters obtained by this approach also allowed us to distinguish intra-regional characteristics.

In order to assess the quality of the generated clusters by a less subjective method than simply checking inter-regional differences, the clusters generated by different types of autoencoders were compared with the ground truth obtained trajectory of Shannon's diversity indices onto a Self-Organizing Map. That is, socioeconomic indicators already used by academia served as feature extractors for the raw data, and the clusters generated from them, based on the method in [Silva et al. 2022], served as ground truth. By admitting the existence of ground truth, it was possible to use accuracy as a measure for comparing the different types of clustering. It was then possible to verify which of the different combinations of autoencoders and clustering methods best approximated ground truth. The results showed that the autoencoder-I + k-means combination produced the best alignment. It should be noted that this alignment captures both the similarity of the different municipalities based on Shannon's diversity indices and the temporal dynamics of changes in economic indicators over the observation period. The temporal dynamics

are taken into account due to the methodology used in [Silva et al. 2022] to generate the groups.

Future work should include evaluating larger latent layers, 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

- Charte, D., Charte, F., García, S., Jesus, M. J., and Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 44:78–96.
- 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., Kalimbara, 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.
- Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). Deep clustering with convolutional autoencoders. *Lecture Notes in Computer Science*, (10635):373–382. 10.1007/978-3-319-70096-0\_39.
- Halkidi, M. and Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29:773–786.
- 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.
- Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514. 10.1109/ACCESS.2018.2855437.
- 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.
- 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.
- 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.
- 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.
- Tisdell, C., Alauddin, M., Sarker, M., and Kabir, M. (2019). Agricultural diversity and sustainability: general features and Bangladeshi illustrations. *Sustainability*, 11:6004–6015.
- 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.