UNSUPERVISED MACHINE LEARNING FOR REGIONAL SOCIOECONOMIC SIMILARITY: COMPARING MINAS GERAIS MUNICIPALITIES AND BRAZILIAN FEDERATIVE UNITS

Celestia Olimpio Almeida Nunes[1] - celestiaolimpio@gmail.com
Reginaldo Morais de Macedo[1] - reginaldo.macedo@unimontes.br
Gustavo Gomes Nobre[1] - gustavognobre@gmail.com
Lúcio Júnio Benfica Rosa[1] - luciojuniobenfica@hotmail.com
Charles Soares Cardoso[1] - charlesprofessor.info@gmail.com
[1]Universidade Estadual de Montes Claros, UNIMONTES - Montes Claros, MG, Brazil

Abstract. *Regional socioeconomic analysis is fundamental for territorial planning and public policy, yet identifying similarities among geographic areas requires multivariate approaches. This study proposes an unsupervised machine learning framework to identify socioeconomic patterns among the municipalities of Minas Gerais and the Brazilian federative units using official indicators from the Brazilian Institute of Geography and Statistics (IBGE). Six variables related to health, education, income, employment and demographics were subjected to median imputation, Yeo–Johnson transformation and feature standardisation. K-means clustering was applied, with the number of clusters determined using the silhouette, Calinski–Harabasz, Davies–Bouldin and elbow criteria. Principal Component Analysis supported visualisation, while Ward hierarchical clustering assessed partition robustness. The methodology identified two municipal clusters and three federative-unit clusters. Municipalities exhibited partially overlapping socioeconomic profiles, whereas the federative units formed more clearly separated groups. Agreement between K-means and Ward clustering was high for the federative units but modest for municipalities, indicating greater uncertainty at the local scale. The framework provides an exploratory decision-support instrument for identifying comparable territories, supporting policy benchmarking, differentiated programme design and evidence-based public administration.*

Keywords: *cluster analysis, territorial typology, multivariate analysis, decision support, regional development*

1. INTRODUCTION

Regional socioeconomic analysis supports territorial planning by identifying similarities and disparities among geographic areas, enabling more effective public policies and cooperation between territories (Mouronte-López & Savall Ceres, 2024; Batomunkuev et al., 2024). Because development encompasses health, education, employment and economic conditions, multivariate approaches are better suited to characterise territorial differences than isolated indicators.

The availability of official socioeconomic data has encouraged the use of unsupervised machine learning in regional analysis. Previous studies have applied clustering to agricultural development, socioeconomic similarity, poverty analysis and regional zoning, demonstrating its usefulness for planning and decision-making (Shestakov & Lovchikova, 2023; Batomunkuev et al., 2024; Mouronte-López & Savall Ceres, 2024; Setyadi et al., 2025; Li et al., 2026). Nevertheless, these studies generally employ indicator sets and procedures tailored to specific domains.

This study applies a unified clustering framework to identify socioeconomic similarity patterns among the municipalities of Minas Gerais and the Brazilian federative units using harmonised IBGE indicators. The resulting typologies are intended as exploratory decision-support resources through which public institutions may identify comparable territories, establish policy benchmarks and differentiate interventions according to shared socioeconomic conditions.

2. METHODOLOGY

The workflow comprised data collection, preprocessing, clustering, validation and interpretation. It was implemented in Python using reproducible data-analysis and machine-learning libraries.

2.1 Data Collection

Two datasets were analysed independently: the 853 municipalities of Minas Gerais and the 27 Brazilian federative units. Six indicators were considered: infant mortality, GDP per capita, school enrolment from 6 to 14 years, average monthly salary of formal workers, population density and log-transformed population size.

For both the federative units and Minas Gerais municipalities, infant mortality, GDP per capita and average salary refer to 2023; school enrolment and population to 2022; and territorial area to 2025. Population density was calculated from population and area. Together, the variables represent demographic, educational, economic, labour and health dimensions commonly adopted in regional studies (Mouronte-López & Savall Ceres, 2024; Batomunkuev et al., 2024).

2.2 Data Preprocessing

Numerical formatting was standardised and missing values were treated before clustering. Infant mortality was unavailable for 294 of the 853 municipalities and was therefore imputed using the municipal median. All variables were subsequently transformed using the Yeo–Johnson method and standardised to zero mean and unit variance, reducing skewness and preventing variables with larger numerical magnitudes from dominating Euclidean distances.

2.3 Clustering and Validation

Regional similarity was investigated using K-means with k-means++ initialisation, 10 initialisations and a fixed random seed of 42. Candidate values of $k = 2, \ldots, 10$ were evaluated for municipalities and $k = 2, \ldots, 6$ for federative units. The final number of clusters was selected through consensus among the silhouette, Calinski–Harabasz and Davies–Bouldin indices, with inertia used as complementary evidence. Principal Component Analysis (PCA) was employed solely for visualisation.

K-means partitions were compared with agglomerative hierarchical clustering using Ward linkage and the Adjusted Rand Index (ARI). Cluster profiles were interpreted through the means of the original indicators. A cross-scale consistency test was also performed by aggregating the municipal indicators into a synthetic representation of Minas Gerais and classifying it with the federative-unit model.

3. RESULTS AND DISCUSSION

Before clustering, Pearson correlation matrices were computed to evaluate the relationships among the indicators (Figure 1).

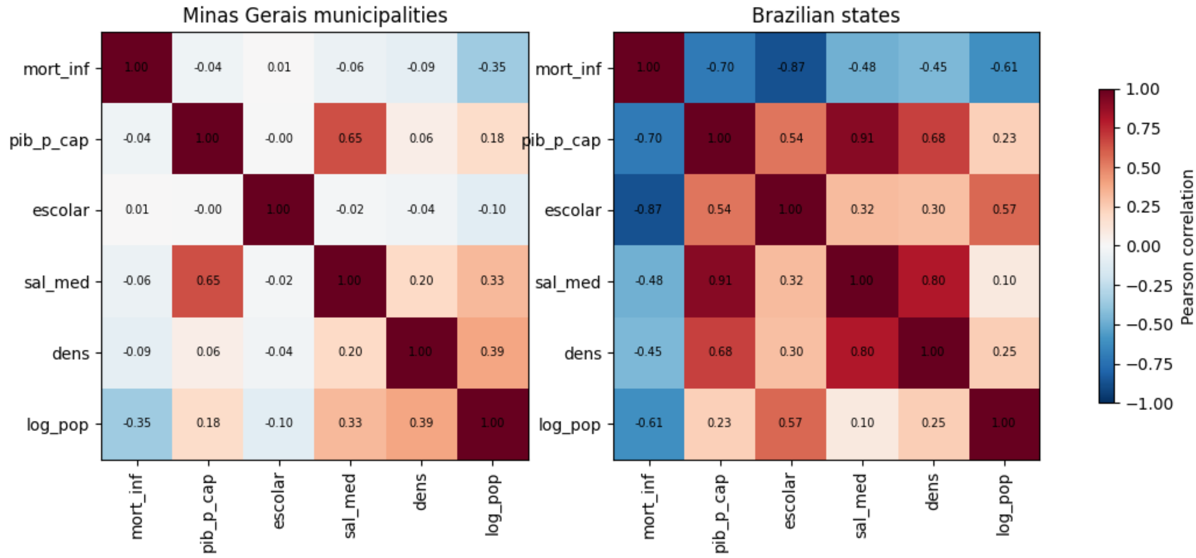

Figure 1: Pearson correlation matrices for municipalities of Minas Gerais and Brazilian federative units.

Municipal indicators exhibited generally weak pairwise correlations, suggesting that they provide complementary socioeconomic information. Stronger relationships were observed among federative units, indicating that income, labour, education, health and demographic dimensions vary more systematically together at this scale.

The number of clusters was evaluated using four internal criteria (Figure 2).

For municipalities, the silhouette and Calinski–Harabasz indices favoured two clusters, whereas Davies–Bouldin favoured seven. For the federative units, silhouette and Calinski–Harabasz favoured three clusters, while Davies–Bouldin favoured four. Following the majority criterion, two municipal clusters and three federative-unit clusters were adopted.

Table 1 summarises the resulting socioeconomic profiles. Municipal Cluster 0 contained 355 observations and combined lower infant mortality with higher GDP per capita, salaries, density

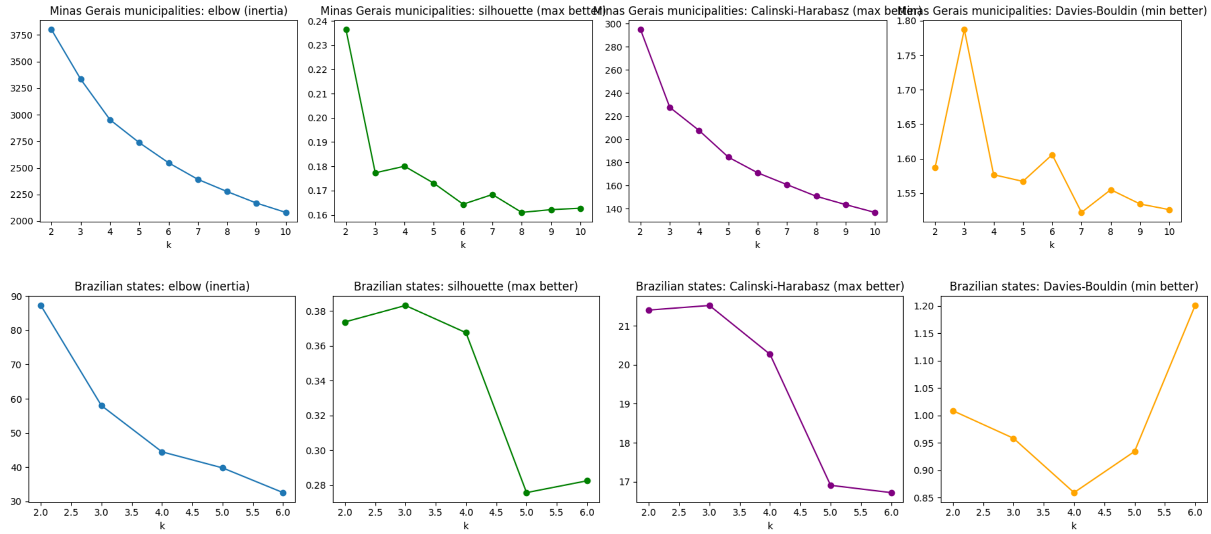

Figure 2: Internal validation indices used to determine the number of clusters.

and population. Municipal Cluster 1 contained 498 generally smaller and less economically advantaged municipalities, although it presented a slightly higher mean school-enrolment rate.

At the federative level, Cluster 1 presented the most favourable average profile, with the lowest infant mortality and the highest GDP per capita, salary, school enrolment, density and population. Cluster 0 comprised eight predominantly less densely populated units, while Cluster 2 combined the lowest income and salary averages with comparatively high infant mortality.

Table 1: Mean socioeconomic profiles of the identified clusters.

Scale and cluster	n	Infant mortality	GDP per capita	Enrolment	Salary	Density	Population
Municipalities 0	355	12.79	50,347.77	98.62	2.05	134.24	48,226
Municipalities 1	498	20.40	23,180.79	99.25	1.68	22.68	6,867
Federative units 0	8	14.86	39,056.63	97.31	2.44	4.93	1,956,125
Federative units 1	9	9.24	60,682.66	98.78	2.96	149.93	13,850,332
Federative units 2	10	13.73	23,597.54	97.71	2.03	56.25	6,276,112

The clusters were visualised through PCA (Figure 3). The first two components explained 58.9% of municipal variance and 82.5% of federative-unit variance.

Municipalities formed partially overlapping groups, indicating gradual transitions rather than sharply separated socioeconomic categories. In contrast, the federative units formed three more clearly differentiated groups. Geographically distant units occurred within the same cluster, confirming that the model captures multivariate socioeconomic similarity rather than geographic proximity.

Agreement between K-means and Ward clustering reached an ARI of 0.282 for municipalities and 0.887 for federative units. The high federative-unit value supports the stability of the three-group partition across algorithms. The lower municipal value indicates that local socioeconomic structure is less sharply defined and that the two-cluster municipal solution should be interpreted as a broad exploratory typology rather than a definitive classification.

The cross-scale consistency test assigned the synthetic representation of Minas Gerais to Federative-unit Cluster 1, matching the classification of the official state observation. This result indicates internal coherence between the municipal aggregation and the national model, although it constitutes only a single-case consistency check.

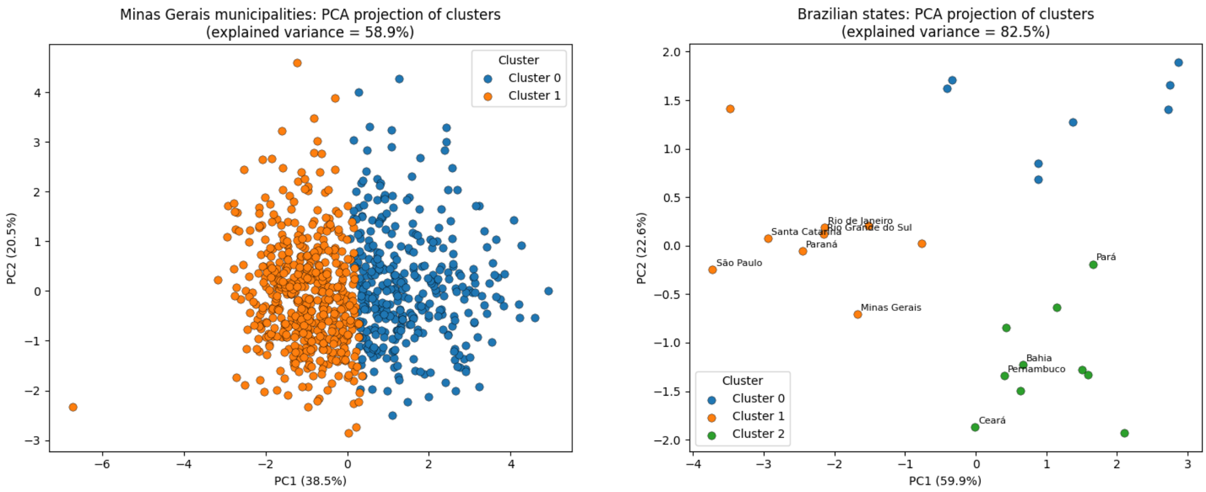

Figure 3: PCA projections of the identified clusters.

From a public-administration perspective, these typologies may support the identification of territories facing comparable conditions, the selection of reference cases for policy benchmarking and the design of differentiated programmes. The clusters should not replace administrative judgement or detailed local diagnosis, but can serve as an exploratory AI-based decision-support instrument for prioritising further analysis.

4. CONCLUSIONS

This study developed an unsupervised machine-learning framework for identifying socioeconomic similarity patterns among the municipalities of Minas Gerais and the Brazilian federative units using official IBGE indicators. The analysis identified two broad municipal profiles and three more clearly separated federative-unit profiles. The municipal partition presented considerable overlap and modest agreement with Ward clustering, whereas the federative-unit classification exhibited clearer separation and high cross-algorithm agreement.

The results demonstrate the potential of clustering as an exploratory tool for public-sector territorial comparison, policy benchmarking and differentiated programme planning. Nevertheless, the classifications should not be interpreted as definitive administrative categories. The state-level analysis contains only 27 observations, while the municipal analysis is sensitive to indicator selection and to the substantial imputation required for infant mortality.

Future research should incorporate additional indicators, temporal observations and alternative clustering methods, as well as evaluate the typologies with public managers and domain specialists. Such extensions could improve the framework's interpretability and its practical integration into evidence-based public administration.

Acknowledgements

Os autores agradecem à FAPEMIG (grant n OET-00243-26).

REFERENCES

Batomunkuev, V.S.; Gomboev, B.O.; Sharaldaev, B.B.; Zhamyanov, D.T.-D.; Tsybikova, A.B.; Badmaev, A.G.; Zangeeva, N.R.; Motoshkina, M.A.; Rygzynov, T.S.; Banzaraktcaev, Z.E.; Lygdenova, A.B.; Alekseev, A.V.; Ulzetueva, A.D.; Badmaeva, A.S. (2024), "Territorial Production and Resource Structures of Asian Russia: Assessment, Typology, and Zoning", *Sustainability*, 16(23), 10518.

Li, Y.; Zheng, H.; Li, F.; Li, L.; Mao, L. (2026), "Interpretable Machine Learning Reveals Regional Heterogeneity and Key Environmental Drivers of BOD_5 in Complex River Networks", *Journal of Environmental Management*, 405, 129664.

Mouronte-López, M.L.; Savall Ceres, J. (2024), "Exploring Socioeconomic Similarity-Inequality: A Regional Perspective", *Humanities and Social Sciences Communications*, 11, 279.

Setyadi, S.; Indriyani, L.; Widiastuti, A. (2025), "The Role of Economic and Social Safety Nets in Extreme Poverty in Indonesia", *Etikonomi*, 24(2), 653–672.

Shestakov, R.B.; Lovchikova, E.I. (2023), "Clustering of Regions Using Basic Agricultural and Economic Criteria", *Ekonomika Regiona (Economy of Regions)*, 19(1), 178–191.