ISSN: 2317-0840

Similarity assessment between hierarchical and partitioned clustering: study with fuel prices

Wylliam Eduardo Alves Silva[1†], Iêda Maria de Siqueira Bezerra[1] , Mayara Macedo da Mata[2]

[1]*Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife - Pernambuco, Brasil.*
[2]*Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, Programa de Pós Graduação em Química, Campina Grande - Paraíba, Brasil.*

Abstract: *This study aimed to compare the similarity between clusters generated by hierarchical and partitioning methods (k-means) in the analysis of fuel prices, using data on regular gasoline and ethanol in the municipality of Campina Grande – PB, Brazil, in 2019. Cluster analysis, a multivariate statistical technique, was applied to classify fuel stations into homogeneous groups based on price proximity. The analyses were performed using R software, demonstrating the feasibility of applying these techniques to real-world data. In the non-hierarchical method (k-means), the number of groups was determined using the Elbow method, with significant differences identified in the mean prices between clusters. For the hierarchical method, we employed Euclidean distance and complete linkage, the resulting clusters exhibited structures similar to those obtained with k-means. The comparison between both approaches revealed consistency in group formation and cluster structural similarity, indicating that the methods produced convergent results for fuel station price segmentation, this demonstrates the reliability of the employed techniques. In conclusion, cluster analysis proves to be an effective tool for fuel market studies, both hierarchical and partitioning methods, while distinct in their approaches, generated coherent groupings in this context.*

Keywords: *Cluster analysis; Hierarchical and partitioning methods; Fuel prices; k-means; Euclidean distance.*

Introduction

Fuel stations are facilities that sell petroleum products and biofuels - including gasoline, diesel, ethanol, and natural gas - regulated by Brazil's National Agency of Petroleum, Natural Gas and Biofuels (ANP) (PETRÓLEO, G. N. E. B. AGÊNCIA NACIONAL do, 2017). Fuel price fluctuations are heavily influenced by the volatility of the international oil market, which has significant effects on economic performance and household consumption (Dallal et al., 2024). In particular, gasoline and ethanol, widely used fuels in Brazil, have exhibited interconnected price trajectories, influenced by factors such as production costs, taxation, and local market competition.

In this context, multivariate statistical techniques, particularly cluster analysis, serve as valuable tools for identifying fuel price similarity patterns across different retail locations. This approach, classified as unsupervised machine learning, groups observations based on their similarities, maximizing within-cluster homogeneity while maximizing between-cluster heterogeneity (Silva, 2021). This technique has proven versatile across multiple disciplines, particularly in economic and management sciences, by enabling the organization of large datasets into groups with similar characteristics, thereby facilitating the analysis and interpretation of relevant patterns. (Ferreira et al., 2020).

[†]Autor correspondente: wylliameduardo99@gmail.com
Received: 16/04/2025. Revised: 12/05/2025. Accepted:19/05/2025.

Among available clustering techniques, hierarchical and partitioning methods stand out as the most prominent, differing primarily in their group formation strategies during the analysis process (Pereira, 2023). While hierarchical methods group data based on similarity or dissimilarity measures, producing a dendrogram structure, partitioning methods like k-means require predefined cluster numbers and use an iterative process to optimize element allocation (Oliveira et al., 2022). The choice between these techniques depends on the study objectives and data characteristics, but comparing their effectiveness in generating consistent clusters is essential for result validation.

This study evaluates the similarity between clusters generated by hierarchical and partitioning methods in analyzing regular gasoline and ethanol prices in Campina Grande, PB, Brazil. To this end, cluster analysis was applied to identify stations with similar pricing patterns, comparing the consistency of groups generated by each method.

Materials and methods

The data used in this study were collected in person by the Municipal Fund for the Defense of Diffuse Rights (PROCON) in Campina Grande, PB, Brazil, and consist of 2019 fuel price records. The dataset comprises 57 fuel stations distributed across the city, containing the following variables: station ID, brand, neighborhood, regular gasoline price, and ethanol price. Following data collection, a thorough verification process was conducted to identify and address missing values or storage inconsistencies, ensuring data quality prior to analysis.

All statistical procedures, calculations, and visualizations were performed using R software (R Core Team, 2024), using specialized packages for data analysis. Among the main packages employed were: cluster (for cluster analysis) (Maechler et al., 2025), factoextra (for multivariate data visualization) (Kassambara; Mundt, 2020), ClustOfVar (for variable clustering) (Chavent et al., 2025) e ggdendro (for dendrogram creation) (de Vries; Ripley, 2024).

Cluster analysis

Cluster analysis is a set of multivariate techniques designed to group objects based on their shared characteristics (Hair et al., 2009). Objects within each group tend to be similar to one another, yet distinct from objects in other groups, (Malhotra, 2019).

According to Falqueto e Cezar (2022), the cluster analysis process involves several fundamental steps: defining the sample to be clustered, selecting the most relevant variables to represent the individuals' characteristics, and choosing a clustering method (either agglomerative or partitioning), while ensuring evaluation of the resulting clusters quality and coherence.

In this study, the number of clusters was determined using the Elbow method, which is generally considered a reliable indicator for identifying the appropriate quantity of groupings. The Elbow method represents one of the most traditionally used approaches for estimating the optimal number of clusters in a dataset. While widely adopted for its application simplicity, it remains sensitive to graphical interpretation (Alves et al., 2024).

Similarity and distance measures

According to Paz (2024), there are two primary approaches for quantifying relationships between objects in an analysis: similarity measures, which express the degree of direct correspondence between elements (with higher values indicating greater likeness), and dissimilarity or distance measures, where larger values reflect greater divergence between compared items.

Correlation coefficients

Similarity measures play a crucial role in cluster analysis by enabling the quantification of association degrees between objects. Among these measures, Pearson's correlation coefficient stands out for its ability to capture linear relationships between variables. It is widely applied across diverse fields due to its intuitive interpretation and straightforward implementation (Albuquerque et al., 2022). Although the correlation coefficient is commonly attributed to Karl Pearson, its origins trace back to Francis Galton's studies on regression and correlation, which directly influenced Pearson's work (Chattamvelli, 2024).

Distance measures

Most clustering techniques rely on calculating a measure that quantifies the degree of separation between objects, typically represented by distance functions or specific metrics (Gao et al., 2023). Multiple functions can serve as distance measures in cluster analysis (Buccianti; Gozzi, 2023). Some distance metrics are classified as similarity measures, including: Euclidean distance, Squared Euclidean distance, Manhattan distance, Minkowski distance, and Mahalanobis distance (Crispim et al., 2019).

Hierarchical methods

This procedure organizes data into a hierarchical structure by grouping elements according to their similarities (Ran et al., 2023). Hierarchical methods can be classified into two main types: agglomerative, which build clusters through progressive merging of elements, and divisive, which form clusters via successive splitting (Wang et al., 2023). The agglomerative method encompasses various widely-used techniques in practice, (Abushilah; Abbas, 2023): single linkage, complete linkage, average linkage, median linkage, centroid method, Ward's Method. According to Cabezas et al. (2023), in hierarchical methods, dendrograms (tree diagrams) are commonly used to visualize clustering results. In this representation, the branches represent individual elements, while the root symbolizes the complete dataset grouping.

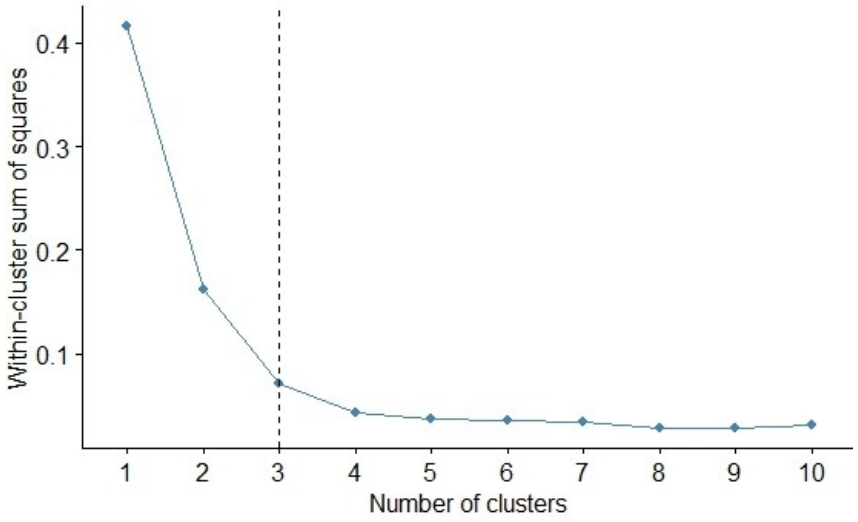
Results and discussion

Initially, cluster analysis techniques were applied, both the non-hierarchical approach (k-means method) and hierarchical (with complete linkage and Euclidean distance), to identify natural groupings among fuel stations based on gasoline and ethanol prices. Following cluster determination by each method, a detailed descriptive analysis was conducted of prices within each formed group.

Non-hierarchical method

Determining the number of clusters using the Elbow method involves analysis of a scree plot, where the location of one (elbow) is normally considered indicative of the ideal number of clusters.

Figure 1: Determining the optimal number of clusters using the Elbow method (Scree Plot)

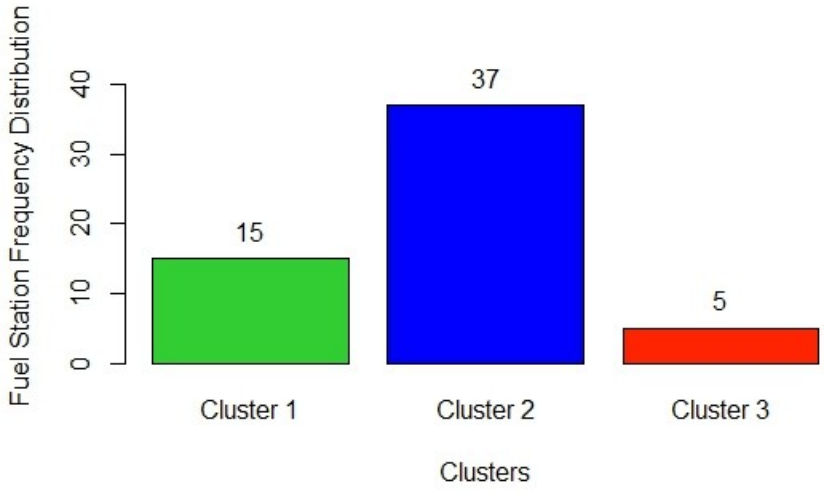


Source: from the authors (2025).

The scree plot reveals a noticeable decline deceleration, indicating that the optimal number of clusters according to the Elbow method is 3 (three).

Using the K-means method, we can observe the distribution of fuel stations across different clusters. This segmentation is visualized in Figure 2.

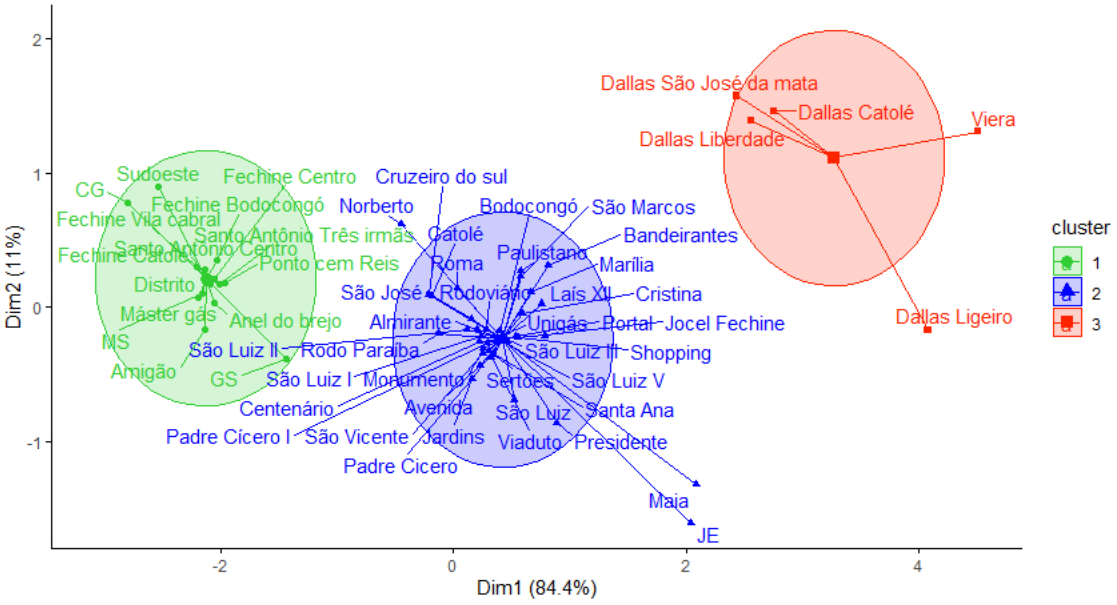
Figure 2: Distribution of fuel stations across identified clusters



Source: from the authors (2025).

Figure 3 displays the geographic distribution of fuel stations, organized according to clusters formed by the K-means method. Each color represents a distinct group.

Figure 3: Geographic distribution of fuel stations by cluster

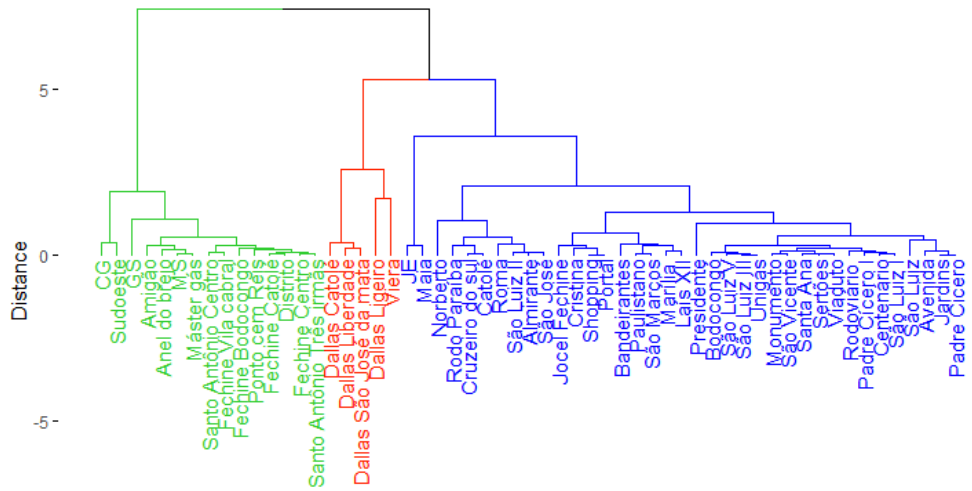


Source: from the authors (2025).

Hierarchical method

Figure 4 displays the dendrogram showing clusters formed by the complete linkage method, following the same criterion used to determine the cluster number in the non-hierarchical approach (i.e., 3 clusters)

Figure 4: Dendrogram



Source: from the authors (2025).

Descriptive statistics by cluster

Table 1 presents descriptive statistics of regular gasoline prices across the identified clusters, highlighting key group characteristics.

Table 1: Statistics for regular gasoline by cluster

Statistics	cluster 1	cluster 2	cluster 3
Minimum	4,257	4,294	4,318
Median	4,283	4,327	4,326
Mean	4,282	4,328	4,343
SD	0,012	0,017	0.030
CV	0,003	0,004	0.007
Maximum	4,308	4,388	4,386

Source: from the authors (2025).

The statistical analysis reveals that cluster 3, comprising only five fuel stations, exhibited the highest mean price for regular gasoline in 2019, suggesting these stations maintained consistently elevated pricing. The mean values indicate that cluster 1 offered the most competitive pricing for regular gasoline, with the lowest average price. cluster 2 showed intermediate pricing, while cluster 3 maintained the highest prices.

Table 2 presents descriptive statistics of ethanol prices across clusters, highlighting key group characteristics.

Table 2: Statistics for ethanol by cluster

Statistics	cluster 1	cluster 2	cluster 3
Minimum	3,209	3,268	3,478
Median	3,235	3,299	3,499
Mean	3,232	3,312	3,523
SD	0,013	0,031	0,068
CV	0,004	0,009	0,019
Maximum	3,257	3,391	3,641

Source: from the authors (2025).

The observed pattern for regular gasoline happens for ethanol: Cluster 3 shows the highest mean ethanol price and maximum values across all reported statistics, while Cluster 1 maintains the lowest mean price - mirroring the gasoline pricing structure.

Conclusion

The results identified three distinct clusters grouping stations with similar pricing patterns. The hierarchical method employed Euclidean distance and complete linkage, while the non-hierarchical approach utilized k-means clustering. This consistency across methodologies - with no station reassignments between clusters - confirms the robustness of the cluster analysis in this study.

As highlighted by Silva (2021), Campina Grande's PROCON could implement cluster analysis in their monthly fuel price surveys. This approach would help consumers identify stations with the most affordable prices, potentially reducing their gasoline and ethanol expenditures.

References

ABUSHILAH, S. F.; ABBAS, R. H. Performance evaluation of some clustering algorithms under different validity indices. **Mathematical Modelling of Engineering Problems**, v. 10, n. 4, p. 1271–1280, 2023.

ALBUQUERQUE, M. A. de; NASCIMENTO, E. R. do; BARROS, K. N. N. de O.; BARROS, P. S. N. Comparison between similarity coefficients with application in forest sciences. **Research, Society and Development**, v. 11, n. 2, p. e48511226046–e48511226046, 2022.

ALVES, K. A.; ALBUQUERQUE, O. de S.; SOUSA, A. L. de; JÚNIOR, G. de M. Análise de dados dos planos de desenvolvimento institucional do instituto federal do pará (2009-2023) utilizando o algoritmo de aprendizado de máquina k-means. **Cuadernos de Educación y Desarrollo**, v. 16, n. 11, p. e6259–e6259, 2024.

BUCCIANTI, A.; GOZZI, C. Cluster analysis and classification. In: **Encyclopedia of Mathematical Geosciences**. [S.l.]: Springer, 2023. p. 127–133.

CABEZAS, L. M.; IZBICKI, R.; STERN, R. B. Hierarchical clustering: Visualization, feature importance and model selection. **Applied Soft Computing**, Elsevier, v. 141, p. 110303, 2023.

CHATTAMVELLI, R. Measures of association. In: **Correlation in Engineering and the Applied Sciences: Applications in R**. [S.l.]: Springer, 2024. p. 1–54.

CHAVENT, M.; KUENTZ, V.; LIQUET, B.; SARACCO, J. Clustofvar: An r package for the clustering of variables (version 1.2). 2025. Disponível em: ⟨https://CRAN.R-project.org/package=ClustOfVar⟩.

CRISPIM, D. L.; FERNANDES, L. L.; ALBUQUERQUE, R. L. d. O. Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do marajó-pa. **Revista Principia**, v. 1, n. 46, p. 145–154, 2019.

DALLAL, G. F. C. et al. Flutuações no preço do petróleo e seus impactos em indicadores econômicos nacionais. Florianópolis, SC., 2024. Disponível em: ⟨https://repositorio.ufsc.br/handle/123456789/261480⟩.

de Vries, A.; RIPLEY, B. D. **ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'**. [S.l.], 2024. R package version 0.2.0. Disponível em: ⟨https://andrie.github.io/ggdendro/⟩.

FALQUETO, A. A.; CEZAR, L. C. SegmentaÇÃo via machine learning: Proposta de clusterizaÇÃo de consumidores do e-commerce de uma empresa multinacional do varejo esportivo. **HOLOS**, v. 4, dez. 2022. Disponível em: ⟨https://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/12032⟩.

FERREIRA, R.; PAIM, F. d. P.; RODRIGUES, V.; CASTRO, G.; RODRIGUES, U. V. G. S. Análise de cluster não supervisionado em r: agrupamento hierárquico. 2020. Disponível em: ⟨https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1126478/1/5360.pdf⟩.

GAO, C. X.; DWYER, D.; ZHU, Y.; SMITH, C. L.; DU, L.; FILIA, K. M.; BAYER, J.; MENSSINK, J. M.; WANG, T.; BERGMEIR, C. et al. An overview of clustering methods with guidelines for application in mental health research. **Psychiatry Research**, Elsevier, v. 327, p. 115265, 2023.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. [S.l.]: Bookman editora, 2009.

KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**. [S.l.], 2020. R package version 1.0.7. Disponível em: ⟨https://CRAN.R-project.org/package=factoextra⟩.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. **cluster: Cluster Analysis Basics and Extensions**. [S.l.], 2025. R package version 2.1.8.1 — For new features, see the 'NEWS' and the 'Changelog' file in the package source). Disponível em: ⟨https://CRAN.R-project.org/package=cluster⟩.

MALHOTRA, N. K. **Pesquisa de Marketing-: uma orientação aplicada**. [S.l.]: Bookman Editora, 2019.

OLIVEIRA, P. L. S. de; RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S. Identificação de pesquisas e análise de algoritmos de clusterização para a descoberta de perfis de engajamento. **Revista Brasileira de Informática na Educação**, v. 30, p. 01–19, 2022.

PAZ, H. O. d. Método de agrupamento multinível para dados mistos. Universidade Federal da Bahia, 2024. Disponível em: ⟨https://repositorio.ufba.br/handle/ri/40414⟩.

PEREIRA, L. G. Clusterização como técnica de apoio à decisão para um marketplace eletrônico logístico. 2023. Disponível em: ⟨https://repositorio.unifei.edu.br/jspui/handle/123456789/3965⟩.

PETRÓLEO, G. N. E. B. AGÊNCIA NACIONAL do. **Cartilha do posto revendedor de combustíveis**. 2017. Disponível em: ⟨https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/cartilhas-e-guias/arq/cartilhapostorevendedor6ed.pdf⟩.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: ⟨https://www.R-project.org/⟩.

RAN, X.; XI, Y.; LU, Y.; WANG, X.; LU, Z. Comprehensive survey on hierarchical clustering algorithms and the recent developments. **Artificial Intelligence Review**, Springer, v. 56, n. 8, p. 8219–8264, 2023.

SILVA, W. E. A. **Análise de cluster aplicada aos dados de preços de combustíveis na cidade de Campina Grande - PB**. 31 p. — Universidade Estadual da Paraíba, Campina Grande, 2021. Trabalho de Conclusão de Curso (Graduação em Estatística). Disponível em: ⟨http://dspace.bc.uepb.edu.br/jspui/handle/123456789/25640⟩.

WANG, F.; ZHOU, G.; XIE, J.; FU, B.; YOU, H.; CHEN, J.; SHI, X.; ZHOU, B. An automatic hierarchical clustering method for the lidar point cloud segmentation of buildings via shape classification and outliers reassignment. **Remote Sensing**, MDPI, v. 15, n. 9, p. 2432, 2023.

This page was intentionally left blank – formatting consistency.