Universitas Ahmad Dahlan

Bulletin of Applied Mathematics and Mathematics Education
Volume 5, Number 2, October 2025 pp. 97-102
https://dx.doi.org/10.12928/bamme.v5i2.14502
e-ISSN 2776-1029, p-ISSN 2776-1002

Shopping pattern segmentation: HAC versus K-Means performance analysis

Nur Arina Hidayati*, Uswatun Khasanah

Universitas Ahmad Dahlan, Jl. Ahmad Yani, Tamanan, Banguntapan, Bantul, DIY 55191 Indonesia

*Corresponding e-mail: nur.hidayati@pmat.uad.ac.id

ARTICLE INFO

Article History
Received 20 June 2025
Revised 7 October 2025
Accepted 8 October 2025

Keywords
Consumer analytics
Clustering technique
HAC
K-Means

How to cite this article:
Hidayati, N. A., & Khasanah, U. (2025). Shopping pattern segmentation: HAC versus K-Means performance analysis. *Bulletin of Applied Mathematics and Mathematics Education*, *5*(2), 97-102.

ABSTRACT

Despite widespread use in consumer analytics, clustering techniques remain underutilized for analyzing household basic food commodity consumption patterns, particularly for developing localized retail strategies and targeted food security policies in resource-constrained contexts. This study addresses this practical gap by systematically comparing Hierarchical Agglomerative Clustering (HAC) and K-Means performance on essential consumption patterns across seven commodities: bread, vegetables, fruit, meat, poultry, milk, and wine. Using dual validation metrics, Silhouette Coefficient and Davies-Bouldin Index, we evaluate clustering effectiveness specifically for small-scale household datasets typical of regional food policy environments. HAC demonstrated superior cluster stability (Silhouette score = 0.2936, DBI = 0.8977) compared to K-Means (0.2912, 0.9871), enabling identification of three actionable consumption segments, namely budget-conscious households with economical protein consumption, high spender households with premium patterns across categories, and balanced/selective households preferring bread and wine. These empirically-derived segments provide implementable frameworks for food subsidy targeting, inventory optimization in local retail contexts, and nutrition intervention program design. The findings demonstrate that methodologically rigorous clustering analysis yields policy-relevant household segmentation even with constrained data, offering practical guidance for evidence-based food security interventions where basic commodity consumption directly informs resource allocation decisions.

Introduction

Understanding household consumption patterns has become increasingly important in supporting effective food policies and targeted retail marketing strategies. Analysis of family shopping patterns can identify consumption characteristics that reflect lifestyle, preferences, and differences in spending priorities. Unlike commercial retail analytics that primarily serves profit optimization, household-level consumption segmentation directly informs decisions on food subsidy targeting, nutrition intervention programs, and emergency food distribution systems (FAO, 2006; Smith & Subandoro, 2007; Hidayati, 2013). In resource-constrained contexts, particularly in developing regions, policymakers require cost-effective analytical methods that can reliably segment

households based on essential food commodity consumption without demanding extensive computational infrastructure or complex multi-dimensional data (Hadley & Crooks, 2012). Clustering-based approaches are commonly used to group households into segments with similar characteristics without requiring initial class labels. Common methods include Hierarchical Agglomerative Clustering (HAC) and K-Means, which are widely applied for consumer segmentation (Maulana et al., 2021; Safitri et al., 2025). While comparative studies demonstrate that K-Means excels with large-scale numerical datasets due to computational efficiency (Likas et al., 2003) and HAC performs better on smaller datasets (Kaushik & Mathur, 2014).

Previous research has examined the characteristics of family shopping patterns and demonstrated that cluster analysis can reveal significant consumption preference variations across households (Hidayati, 2013). Recent studies have also utilized K-Means methods for customer segmentation across various sectors, including e-commerce and retail, showing its effectiveness in finding consumer segments with different behaviors (Apriyanto & Sitio, 2025; Fatrilia et al., 2023; Jihan et al., 2025; Rahma et al., 2025; Safitri et al., 2025; Siagian et al., 2025). RFM (Recency, Frequency, Monetary) model-based approaches combined with K-Means have proven to provide relevant strategic insights in developing data-driven m arketing strategies (Jihan et al., 2025; Rahma et al., 2025). On the other hand, research using hierarchical approaches, particularly HAC, has been used in customer grouping based on demographic attributes and shopping behavior with promising results (Maulana et al., 2021).

Several studies have specifically compared the performance of K-Means and Hierarchical Clustering, showing that K-Means is superior in handling large and numerical datasets, while Hierarchical Clustering is more suitable for small datasets and can produce better cluster quality under certain conditions (Kaushik & Mathur, 2014).

However, there are research gaps that need to be addressed: first, lack of studies that directly compare segmentation results between HAC and K-Means on simple yet representative household shopping data with basic necessity attributes. Second, previous research mostly focuses on large e-commerce or modern retail sector data with complex transaction attributes, while daily household consumption patterns with limited but important attributes have not been extensively explored. Third, cluster quality evaluation aspects using Silhouette Score or Davies-Bouldin Index in the context of household data are relatively rarely reported.

This study focuses on seven essential food commodities—bread, vegetables, fruit, meat, poultry, milk, and wine—deliberately selected based on the Food and Agriculture Organization's household expenditure structure framework and dietary diversity guidelines (Kennedy et al., 2011; FAO, 2018), which identifies these as representative indicators of protein, carbohydrate, and micronutrient consumption patterns. These commodities collectively account for approximately 60-70% of typical household food expenditure in middle-income contexts (Muhammad et al., 2011) and serve as proxy variables for nutritional adequacy assessment (Ruel, 2003).

Based on this background, this research aims to apply Hierarchical Agglomerative Clustering and K-Means methods to household shopping data and compare the segmentation results obtained from both methods with evaluation based on cluster validity indices. This study is expected to contribute to the literature related to consumption pattern segmentation and provide practical insights for developing data-based household needs provision strategies.

Method

Research design

This research uses a quantitative approach with unsupervised learning methods based on

0.12928/bamme.v5i2.14502

clustering. Two algorithms used are Hierarchical Agglomerative Clustering (HAC) and K-Means, which are compared for their results on household shopping data.

Data and variables

This study utilizes the French Food Data, a well-established benchmark dataset for clustering analysis (Lebart et al., 1982). The dataset consists of average food expenditures for 12 different household types in France, categorized by occupation (manual workers = MA, employees = EM, managers = CA) and number of children (2, 3, 4, or 5 children). Each household type's consumption is measured across seven essential food commodities: bread, vegetables, fruits, meat, poultry, milk, and wine. Expenditure values represent average spending in French francs per household category. Each attribute is expressed in numerical form representing the number of purchase units per period.

Data preprocessing

Preprocessing steps include data cleaning through removal of missing or extreme values, normalization using min-max scaling so that each attribute is in the range [0,1] (Han et al., 2011), and outlier detection as K-Means algorithm is sensitive to extreme values (Kaushik & Mathur, 2014).

Clustering algorithms

Hierarchical Agglomerative Clustering (HAC)

HAC starts with each object as one cluster, then the two most similar clusters are iteratively merged until the desired number of clusters is reached. Ward's method is used as linkage to minimize the sum of squared distances within clusters. Distance between objects is calculated using Euclidean formula:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (1)$$

where x_i and x_i are two objects with p attributes.

K-Means clustering

K-Means groups data into k clusters with the goal of minimizing variation within clusters. The iterative process involves determining the number of clusters k = 3, random selection of initial centroids, assignment of each object to the nearest centroid, and centroid update until convergence (Selim & Ismail, 1984).

Cluster quality evaluation

Cluster quality is evaluated using Silhouette Coefficient and Davies-Bouldin Index. The Silhouette Coefficient is calculated as:

$$s(i) = \frac{b\,(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (2)$$

where :

$a(i)$ = the average distance of object i to other objects in the same cluster

$b\,(i)$ = is the average distance of object i to objects in the nearest different cluster

The value of $s(i)$ is in the range [-1,1], with values approaching 1 indicating good clusters. Davies-Bouldin Index measures the ratio of inter-cluster distance to the size of intra-cluster spread (Sinaga & Yang, 2020).

Results and discussion

Hierarchical Agglomerative Clustering (HAC) results

The application of Hierarchical Agglomerative Clustering method with Ward linkage on household shopping pattern data resulted in three clusters determined based on dendrogram interpretation.

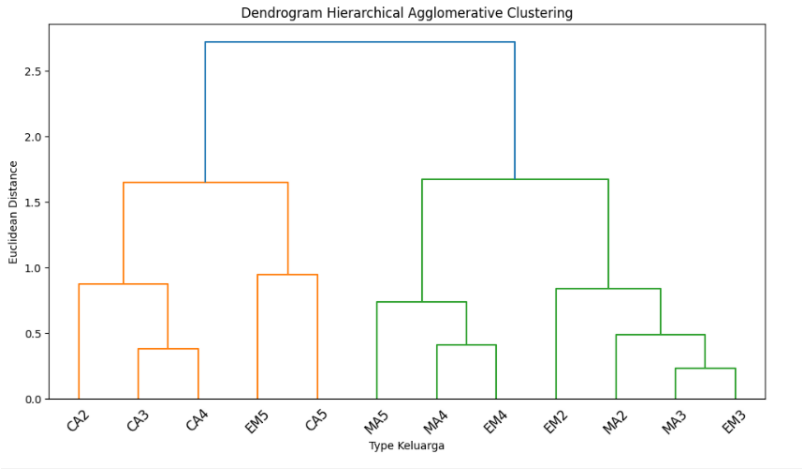


Figure 1. Dendogram of Hierarchial Agglomerative Clustering

The cluster membership consists of
Cluster 1: CA2, CA3, CA4, EM5, CA5
Cluster 2: MA4, EM4, MA5 and
Cluster 3: MA2, EM2, MA3, EM3.

The analysis of the average characteristics of each attribute in each cluster is shown in Table 1.

Table 1. Average shopping per attribute for each cluster (HAC)

Cluster	Bread	Vegetables	Fruits	Meat	Poultry	Milk	Wine
1	354.25	539.50	369.75	1,493	548.75	282.25	363.75
2	446.00	909.67	732.33	2,447	1,154.67	369.33	302.33
3	521.00	779.40	476.80	1,865	795.80	412.40	412.20

Based on Table 1, Cluster 2 has the highest consumption level in almost all categories, particularly meat and vegetables, categorized as high spender. Cluster 1 shows lower animal protein consumption with simple shopping patterns in main categories. Cluster 3 has relatively balanced consumption patterns but with higher preferences for bread and wine.

Cluster quality evaluation resulted in Silhouette Score = 0.2936 and Davies-Bouldin Index = 0.8977. The Silhouette value approaching 0.3 indicates that cluster separation is quite good, although there is still some overlap between clusters. The relatively small DBI value indicates that the formed clusters have sufficiently clear and compact separation.

K-Means results

Clustering using K-Means with the same number of clusters (k = 3) resulted in different cluster member composition from HAC. Cluster quality evaluation gave Silhouette Score = 0.2912 and DBI = 0.9871, which is slightly lower compared to HAC. This difference shows that on small and relatively homogeneous datasets, hierarchical approaches produce slightly more stable clusters compared to K-Means partition methods (Kaushik & Mathur, 2014).

10.12928/bamme.v5i2.14502

Comparison of HAC and K-Means results

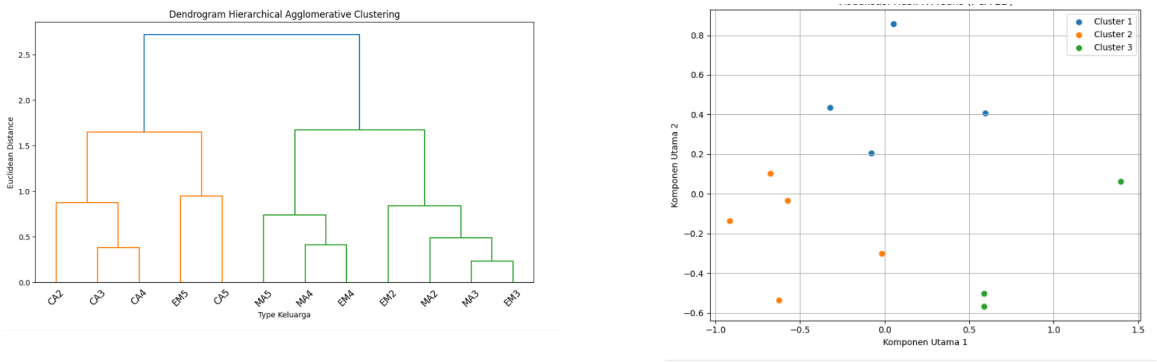
A comparison of the two methods is shown in Table 2.

Table 2. Performance comparison

Method	Silhouette Score	Davies-Bouldin Index
HAC	0.2936	0.8977
K-Means	0.2912	0.9871

The results show that HAC has slightly better performance than K-Means in separating family shopping pattern groups. This aligns with literature stating that hierarchical methods are often more stable for small datasets, while K-Means is more efficient for large datasets but sensitive to initial center point selection (Kaushik & Mathur, 2014).

Visualization of the clustering results using PCA dimension reduction is shown in Figure 2, which illustrates the distribution of data points across the two principal components. It can be seen that the results of HAC clustering provide a slightly clearer separation of groups compared to K-Means, which is consistent with the quantitative evaluation results.



(a) (b)

Figure 2. Dendrogram resulting from clustering using the HAC method (a), and visualization of K-means results (b)

Further interpretation of shopping characteristics per cluster shows three distinct patterns: Budget-Conscious households with more economical consumption patterns and relatively low spending for meat, poultry, and milk; High Spender households with high consumption in almost all categories, especially meat and vegetables, showing more premium shopping patterns; and Balanced/Selective households with relatively balanced consumption patterns but with higher preferences for bread and wine, reflecting certain typical consumption styles.

Conclusion

The results show that HAC provides slightly better cluster quality than K-Means, both in terms of Silhouette Score and Davies-Bouldin Index. The identified segmentation patterns offer valuable insights for retailers and policymakers in understanding diverse household consumption behaviors. Cluster characteristic interpretation provides practical insights about household consumption pattern segmentation that can be utilized for food policy planning, marketing strategies, and product stock management. Future research could explore larger datasets and additional clustering validation metrics to further validate these findings.

References

Apriyanto, B., & Sitio, S. L. M. (2025). Penerapan K-Means dalam menganalisis pola pembelian pelanggan pada data transaksi e-commerce. *Bit-Tech*, *7*(3), 790–797. https://doi.org/10.32877/bt.v7i3.2195

Fatrilia, E. I., Safaat, I., Maharani, E., & Rihastuti, S. (2023). Segmentasi konsumen berdasarkan pola pembeli dengan sistem cluster. *Seminar Nasional AMIKOM Surakarta (SEMNASA)*.

FAO. (2006). *Food Security Policy Brief, Issue 2*. Food and Agriculture Organization.

Hadley, C., & Crooks, D. L. (2012). Coping and the biosocial consequences of food insecurity in the 21st century. *American Journal of Physical Anthropology*, *149*(S55), 72-94

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd Edition.* The Morgan Kaufmann Series in Data Management Systems.

Hidayati N. (2013). Analisis karakteristik pola belanja keluarga dengan analisis klaster. *AdMathEdu, 3*(1).

Jihan, A., Prihartono, W., & . F. (2025). Segmentasi konsumen di pasarmu.id menggunakan K-means clustering berdasarkan model RFM. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(2). https://doi.org/10.23960/jitet.v13i2.6327

Kaushik, M., & Mathur, M. B. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. *International Journal of Software & Hardware Research in Engineering, 2*(6), 93-98.

Kennedy, G., Ballard, T., & Dop, M. C. (2011). *Guidelines for measuring household and individual dietary diversity*. FAO.

Lebart, L., Morineau, A., & Fénelon, J.-P. (1982). French food data: Average expenditures on food for several different types of families in France [data set]

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global K-means clustering algorithm. *Pattern Recognition, 36*(2), 451-461.

Maulana, R., Adi Putra Pratama, D., Nugraha, N., & Rahmasari, A. (2021). Implementasi algoritma Hierarchical Clustering untuk klasterisasi data pelanggan mall. *Gunung Djati Conference Series, 3*.

Muhammad, A., Seale Jr, J. L., Meade, B., & Regmi, A. (2011). International evidence on food consumption patterns: An update using 2005 international comparison program data. *Technical Bulletin TB-1929*. U.S. Department of Agriculture.

Rahma, A. A., Faqih, A., & Rinaldi, A. R. (2025). Optimalisasi strategi pemasaran melalui segmentasi pelanggan dengan analisis RFM dan algoritma K-Means untuk bisnis ritel. *JIKO (Jurnal Informatika dan Komputer), 9*(2), 338. https://doi.org/10.26798/jiko.v9i2.1737

Ruel, M. T. (2003). Operationalizing dietary diversity: A review of measurement issues and research priorities. *The Journal of Nutrition, 133*(11), 3911S-3926S.

Safitri, H., Putri Lenggo Geni, S., Merry, F., & Wati, M. (2025). Penerapan K-Means Clustering untuk segmentasi konsumen e-commerce berdasarkan pola pembelian. *JUKI : Jurnal Komputer Dan Informatika*, *7*.

Selim, S. Z., & Ismail, M. A. (1984). K-Means-type algorithms: A generalized convergence theorem and characterization of local optimality. In *IEEE Transactions on Pattfrn Analysis and Machine Intelligence* (Issue 1).

Smith, L. C., & Subandoro, A. (2007). *Measuring food security using household expenditure surveys*. International Food Policy Research Institute.

Siagian, R., Pratama, E., Lubis, F., Priscillia, S., & Ramadhani, F. (2025). Segmentasi pelanggan dengan K-Means untuk strategi pemasaran yang efektif. *JATI Jurnal Mahasiswa Teknik Informatika*, *9*.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

10.12928/bamme.v5i2.14502