JOSAR, Vol. 11 No. 2 September 2026, pp. 334-346
ISSN: 2502-8521 (Print) / 2503-1155 (Online)
DOI: https://doi.org/10.35457/9wjmrq51

Cluster Analysis of Provinces Based on the Prevalence of Undernourishment Using the K-Means Algorithm

Lana Astria[1]
Santi Meliani[2]
Ade Pratama[3]
Abdul Rohman[4]

[1,2,3,4]Department of Informatics Engineering, Faculty of Computer Science and Education, Universitas Ngudi Waluyo, Semarang, Indonesia

Article Info

Article history:

Received July 24 2026
Revised August 29, 2026
Accepted September 10, 2026

Keywords:

Cluster
Data Mining
Food Security
K-Means
Prevalence of Undernourishment

ABSTRACT

Food security is a fundamental prerequisite for human development, and Indonesia still faces wide disparities in the Prevalence of Undernourishment (PoU) across provinces. This study aims to group 38 Indonesian provinces based on their average PoU for the 2018-2024 period using the K-Means Clustering algorithm to support evidence-based food security policy prioritization. Secondary PoU data compiled from official statistics underwent data cleaning (correction of naming and regional code inconsistencies), z-score standardization, and optimal cluster determination through a combination of the Elbow Method, Silhouette Score, and Davies-Bouldin Index. The analysis identified three optimal clusters (Silhouette Score = 0.593): a low-risk cluster (26 provinces, mean PoU 7.85%), a moderate-risk cluster (6 provinces, mean PoU 17.16%), and a high-risk cluster (6 provinces, mean PoU 32.13%) dominated by provinces in Papua and Maluku. This study contributes a methodological framework for handling administrative data inconsistencies arising from regional redistricting, as well as a nutrition-intervention priority map that can serve as a reference for policymakers in achieving SDG 2 (Zero Hunger).

This is an open access article under the CC BY-SA license.

Corresponding Author:
Lana Astria
Department of Informatics Engineering, Faculty of Computer Science and Education
Universitas Ngudi Waluyo
Jalan Diponegoro No. 186, Gedanganak, Ungaran Timur, Semarang, Indonesia
Email: lanasatria5@gmail.com

1. INTRODUCTION

Food security-related clustering studies have been extensively conducted in Indonesia using various socio-economic indicators. Several studies have applied K-Means clustering to classify provincial poverty levels, revealing distinct regional disparities. Bahauddin et al. (2021) identified three provincial poverty clusters, with Maluku, West Papua, and Papua consistently belonging to the highest-risk group. Similarly, Pratiwi et al. (2024) compared K-Means and K-Medoids using district- and city-level poverty data and demonstrated that both methods provide useful classifications,

although the analysis was limited to district-level observations. Rahman et al. (2021) further employed multidimensional clustering to identify poverty patterns among the bottom 40 percent of the population, emphasizing multidimensional poverty indicators rather than official food security measures. These studies demonstrate that clustering techniques are effective tools for identifying regional disparities and supporting evidence-based policy formulation.

Beyond poverty indicators, clustering techniques have also been applied to other dimensions of regional welfare. Suraya and Wijayanto (2022) compared Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means for provincial stunting management and concluded that comparative evaluation of clustering algorithms improves classification quality. Marcelina et al. (2023) utilized K-Means to classify provincial happiness indices and determined that two clusters provided the optimal partition based on the Silhouette Score. In disaster-related studies, Ashari et al. (2023) compared K-Means, Mean Shift, and DBSCAN for flood-prone areas in Jakarta and reported that Mean Shift achieved the highest clustering performance (Silhouette Score = 0.566), although K-Means also produced competitive results (Silhouette Score = 0.466). Collectively, these studies confirm that K-Means is a robust clustering algorithm that performs well across various regional development applications.

At the international level, clustering and machine learning approaches have also been employed to assess food security and Sustainable Development Goal (SDG) performance. Castelli et al. (2024) integrated machine learning and clustering techniques to evaluate the alignment among SDG 2 (Zero Hunger), SDG 12, and SDG 13 indicators in Italy, demonstrating the usefulness of clustering for sustainability assessment. Other international studies have similarly utilized clustering methods to examine SDG achievement, food waste reduction (Ahmed et al., 2020), and spatial patterns of food insecurity in Africa (Ayalew et al., 2024), highlighting the increasing importance of unsupervised learning techniques for supporting development policy and regional planning.

Despite these advances, several important research gaps remain. First, previous studies in Indonesia have predominantly relied on proxy indicators of welfare and food security, such as poverty rates (Mayasari and Nugraha, 2023), child stunting (Suraya and Wijayanto, 2022), and happiness indices (Ayalew et al., 2024), rather than the official Prevalence of Undernourishment (PoU) indicator adopted by the Food and Agriculture Organization (FAO) for monitoring SDG Target 2.1 (FAO, 2024). Second, most existing studies have overlooked administrative data quality issues arising from regional expansion, particularly the establishment of five new provinces in Papua since 2022, which may introduce inconsistencies in provincial nomenclature and coding. Third, although K-Means has consistently demonstrated strong performance because of its conceptual simplicity, computational efficiency, and ease of interpretation (Oti et al., 2021; Chong, 2021), its application to official provincial PoU data covering all Indonesian provinces remains limited.

Therefore, this study addresses these research gaps by applying K-Means clustering to official provincial PoU data for the period 2018-2024, implementing a systematic data auditing and cleaning protocol to accommodate administrative boundary changes, determining the optimal number of clusters using multiple evaluation methods, including the Elbow Method, Silhouette Score, and Davies Bouldin Index (Shi et al., 2021; Ritonga

and Hasibuan, 2025), and providing policy-relevant classifications to support evidence-based food security interventions.

2. LITERATURE REVIEW

2.1. Prevalence of Undernourishment and SDG 2

The Prevalence of Undernourishment (PoU) is the official indicator for SDG Target 2.1.1, published annually by the FAO based on Food Balance Sheet data and estimated probability distributions of dietary energy consumption within the population (FAO, 2024). The principal advantage of PoU lies in its methodological consistency, which facilitates standardized comparisons across regions and over time. However, the indicator has limitations, as it measures only dietary energy adequacy and does not account for the quality of micronutrient intake (FAO, 2024). Studies on food security indicators suggest that data-driven and machine learning approaches are increasingly employed to assess progress toward SDG 2 (Zero Hunger) at various territorial scales, both in developed and developing countries (Castelli et al., 2024).

At the subnational level in Indonesia, disparities in PoU across provinces are closely associated with geographical conditions, connectivity, and regional economic structures. Archipelagic regions with limited food distribution infrastructure, such as Maluku and Papua, have consistently exhibited higher levels of food insecurity than regions with better accessibility, such as Java and Bali (Bahauddin et al., 2021). This pattern is also consistent with clustering results based on poverty indicators, which classify Maluku, West Papua, and Papua into the highest-risk clusters (Matdoan et al., 2024). Furthermore, it aligns with the application of K-Means Clustering at the regency and municipal levels, where the method has effectively distinguished groups of impoverished populations (Febriansyah and Muntari, 2023).

2.2. K-Means Clustering Algorithm

K-Means is a partition-based clustering algorithm that organizes data into k clusters by minimizing the Euclidean distance between data points and their nearest cluster centroids (Oti et al., 2021; Ahmed et al., 2020). The algorithm operates iteratively through two main stages: (1) assigning each data point to the cluster with the nearest centroid, and (2) updating the centroid positions based on the mean values of their cluster members until convergence is achieved (Chong, 2021; Ashabi et al., 2020). The main advantages of K-Means include its conceptual simplicity, relatively low computational complexity, and ease of interpreting clustering results (Oti et al., 2021). Nevertheless, K-Means has several limitations, including sensitivity to the selection of initial centroids, the requirement to specify the value of k in advance, and susceptibility to outliers and differences in variable scales (Ahmed et al., 2020).

K-Means has also been extensively compared with categorical clustering methods such as K-Modes and ROCK in studies involving poverty indicators (Sofyan et al., 2021), and has been integrated with decision tree approaches to identify significant factors contributing to social vulnerabilities, such as homelessness (Clark et al., 2021). Compared with other clustering techniques, including hierarchical clustering, DBSCAN, and fuzzy c-means, K-Means was selected as the primary method in this study for several academic

reasons. First, for low-dimensional data (i.e., univariate PoU data), K-Means produces consistent and easily interpretable results without requiring the additional complexity associated with density-based methods such as DBSCAN, which, although effective in detecting irregularly shaped clusters, is less suitable for data with relatively continuous distributions such as PoU (Sari and Ediwijoyo, 2023; Bajal et al., 2022). Second, compared with hierarchical clustering, which is deterministic but computationally more expensive for large datasets and less flexible when the number of observations changes, K-Means is more appropriate for provincial panel data that may experience the addition of administrative units (Suraya and Wijayanto, 2022). Third, comparative studies have demonstrated that K-Means consistently exhibits strong performance, achieving competitive Silhouette Scores and Davies Bouldin Index values relative to other algorithms across various application domains, including regional classification based on socio-economic indicators (Ashari et al., 2023; Bahauddin et al., 2021).

2.3. Determining the Optimal Number of Clusters

The determination of the most appropriate number of clusters (k) is a critical step in K-Means Clustering, as the algorithm cannot automatically determine the optimal value of k (Oti et al., 2021; Ashabi et al., 2020). The Elbow Method evaluates the Within-Cluster Sum of Squares (WCSS) for different values of k and identifies the "elbow point," where adding additional clusters no longer results in a significant reduction in WCSS (Shi et al., 2021; Maori and Evanita, 2023). The Silhouette Score measures how well a data point is matched to its own cluster compared to other clusters, with values ranging from -1 to 1, where higher values indicate better cluster structure quality (Ashari et al., 2023); Ritonga and Hasibuan, 2025). The Davies Bouldin Index (DBI) assesses the ratio between intra-cluster dispersion and inter-cluster distance, where lower values indicate better clustering quality (Ashari et al., 2023). The combined use of these three evaluation methods is recommended to improve the validity of determining the optimal k, as each method exhibits different sensitivity to data characteristics (Ashari et al., 2023; Ashabi et al., 2020).

3. METHODS

This study applies a quantitative approach using the K-Means Clustering method as the primary analytical technique. The research stages are structured based on the Knowledge Discovery in Databases (KDD) framework, which includes data collection, data cleaning and preprocessing, data transformation, application of the K-Means algorithm, cluster quality evaluation, and result interpretation (Han et al., 2012).

3.1. Data Sources and Variables

The data used in this study are secondary panel data on the Prevalence of Undernourishment (PoU) covering 34-38 provinces in Indonesia over the period 2018-2024, obtained from official statistical sources. The dataset consists of 242 province-year observations with the following variables: Year, Province Code, Province Name, PoU (percentage), Total Population, and Number of Undernourished Population. The main variable used for clustering is PoU, in accordance with the study's objective and focus.

The Province Code variable is not used as an identifier due to its inconsistency over time (see Subsection 3.2), while the Undernourished Population variable is only used for data auditing purposes and not as a clustering feature.

3.2. Data Audit and Cleaning

A comprehensive data audit was conducted prior to the main analysis stage, revealing several important findings. First, no missing values or duplicate entries were identified based on the combination of Year and Province, indicating that the dataset is structurally complete. Second, inconsistencies were found in the Province Code variable: during 2018-2021 and 2023-2024, codes 51-55 were inconsistently used for Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, and South Kalimantan, whereas in 2022, the official codes 61-65 were used exclusively for Kalimantan provinces. Based on this finding, the Province Code variable was excluded as an identifier, and all aggregation processes used normalized province names as the primary key.

Third, inconsistencies in province naming were identified, where "Jambi" (2018-2022) was recorded as "Kota Jambi" in 2023-2024. Since no administrative entity equivalent to a province-level "Kota Jambi" exists, the naming was standardized to "Jambi" to maintain consistency in the analytical unit. Fourth, a systematic scaling error was detected in the Undernourished Population variable for the year 2024, where all 38 observations were exactly 100 times larger than the expected product of PoU and total population (with a consistent ratio of 100.00 across all observations), while the 2018-2023 data remained consistent. Since this variable is not used as a primary clustering feature, this issue does not affect the clustering results; however, it is documented to ensure methodological transparency.

Fifth, the number of provinces changed from 34 (2018-2023) to 38 (2024) due to the administrative expansion of Papua Province into five new provinces (Papua, Central Papua, South Papua, Highland Papua, and Southwest Papua), which became fully operational in 2023-2024. Since the newly established provinces have only one year of observation (2024), they are retained in the analysis as single-observation units, as their PoU values remain substantively informative for policy mapping purposes.

3.3. Unit of Analysis and Aggregation

This study focuses on provincial clustering rather than province-year clustering. Therefore, the unit of analysis is the average PoU value (2018-2024) for each province. As a result, the dataset consists of 38 observations, each representing one province. This approach was chosen for three reasons: (1) to avoid clustering provinces based solely on annual fluctuations that may not be substantively meaningful; (2) to capture medium-term structural characteristics of each province rather than short-term variations; and (3) to address changes in the number of provinces due to regional expansion.

3.4. Data Standardization

Since K-Means relies on Euclidean distance for clustering, data standardization is generally not required when only a single variable is used, as there is no scale discrepancy

among variables. However, in this study, z-score normalization was still applied as a standard procedure in K-Means analysis. This ensures reproducibility of the results and allows future extensions of the model with additional variables that may have different scales, such as population size.

3.5. Determination of Optimal Number of Clusters

The optimal number of clusters was determined using three evaluation methods simultaneously: the Elbow Method (evaluating k = 1 to 10 based on WCSS), Silhouette Score (evaluating k = 2 to 7), and the Davies–Bouldin Index (evaluating k = 2 to 7). The K-Means algorithm was executed using k-means++ initialization with 20 random initializations (n_init = 20) for each value of k to minimize the risk of convergence to local optima.

3.6. K-Means Algorithm and Cluster Evaluation

K-Means Clustering was selected for several reasons. First, for low-dimensional (univariate) data, K-Means produces stable and easily interpretable clusters without requiring additional density parameters as in DBSCAN (Sari and Ediwijoyo, 2023). Second, K-Means is more adaptable to changes in the number of observational units compared to hierarchical clustering (Suraya and Wijayanto, 2022). Third, K-Means has demonstrated competitive performance in socio-economic regional clustering studies in Indonesia (Ashari et al., 2023; Bahauddin et al., 2021). After determining the final model, cluster quality was evaluated using the Silhouette Score and the Davies Bouldin Index, and each cluster was interpreted based on centroid values and member characteristics.

3.7. Tools and Analytical Environment

The analysis was conducted using the Python programming language, with the panda's library for data manipulation, scikit-learn for implementing the K-Means algorithm as well as computing the Silhouette Score and Davies Bouldin Index, and matplotlib for data visualization.

4. RESULTS AND DISCUSSION

4.1. Descriptive Statistics

After aggregating the average PoU values from 2018 to 2024 for each province, 38 observations were obtained (one observation per province), as shown in table 1. The descriptive statistics indicate an average PoU of 13.15%, with a standard deviation of 9.52%, a minimum value of 2.37% (DKI Jakarta), and a maximum value of 37.69% (Central Papua). The relatively large standard deviation, which is approximately 72% of the mean, indicates substantial variability in undernourishment across Indonesian provinces. This wide dispersion suggests that food security conditions are highly heterogeneous, reflecting differences in socio-economic development, geographic accessibility, agricultural productivity, infrastructure, and food distribution systems among regions. Furthermore, the interquartile range (Q3-Q1 = 6.38 percentage points)

ISSN: 2502-8521 (Print) / 2503-1155 (Online)
DOI: https://doi.org/10.35457/9wjmrq51

indicates that the middle 50% of provinces still exhibit considerable variation in PoU levels, suggesting that disparities are not limited to a few extreme provinces but are also present among provinces with moderate food security conditions. A skewness value of 1.29 indicates that the distribution is right-skewed, where most provinces have low to moderate PoU values, while only a small number of provinces exhibit very high PoU levels, forming a long right tail. This pattern is consistent with previous studies on regional inequality in food security using poverty indicators (Bahauddin et al., 2021; Matdoan et al., 2024).

Table 1. Descriptive statistics

Statistic	Value (%)
Number of Provinces (n)	38
Mean	13.15
Standard Deviation	9.52
First Quartile (Q1)	7.73
Median (Q2)	9.97
Third Quartile (Q3)	14.11
Minimum	2.37 (DKI Jakarta)
Maximum	37.69 (Central Papua)
Skewness	1.29

4.2. Data Preprocessing and Standardization

The data preprocessing stage produced a clean dataset consisting of 38 provincial observations with no missing values or duplicates, and with standardized province names (including correction from “Kota Jambi” to “Jambi”). The PoU variable was then standardized using z-score normalization to obtain a mean of 0 and a standard deviation of 1, following standard procedures for Euclidean distance-based algorithms such as K-Means.

4.3. Determination of Optimal Number of Clusters

Figure 1 presents the Elbow Method results for k = 1 to 10. The graph shows a sharp decrease in WCSS from k = 1 (38.00) to k = 2 (7.01) and k = 3 (4.14), followed by a gradual flattening after k = 3, indicating that the elbow point is located at k = 3.

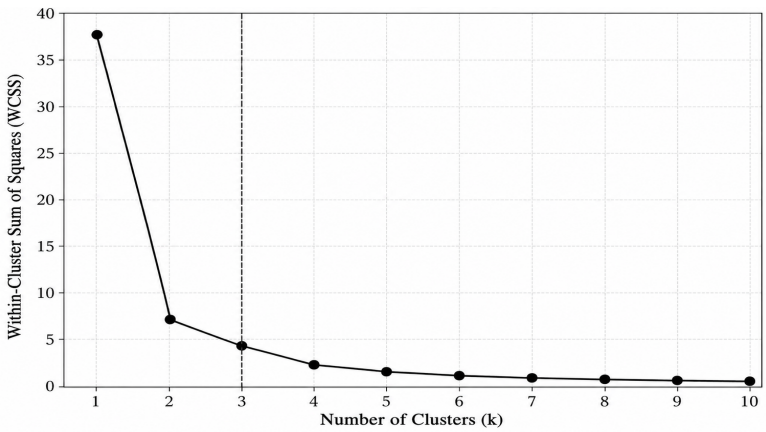

Figure 1. Elbow method results for cluster determination

ISSN: 2502-8521 (Print) / 2503-1155 (Online)
DOI: https://doi.org/10.35457/9wjmrq51

Figure 2 presents the Silhouette Score values for k = 2 to 7. The highest score is obtained at k = 2 (0.755), while k = 3 (0.593) and k = 4 (0.597) remain relatively comparable and still indicate a strong clustering structure (Silhouette Score > 0.5 indicates reasonable to strong cluster structure) (Ritonga and Hasibuan, 2025). The Davies Bouldin Index (DBI) for k = 3 is 0.558, indicating good clustering quality (lower values indicate better clustering).

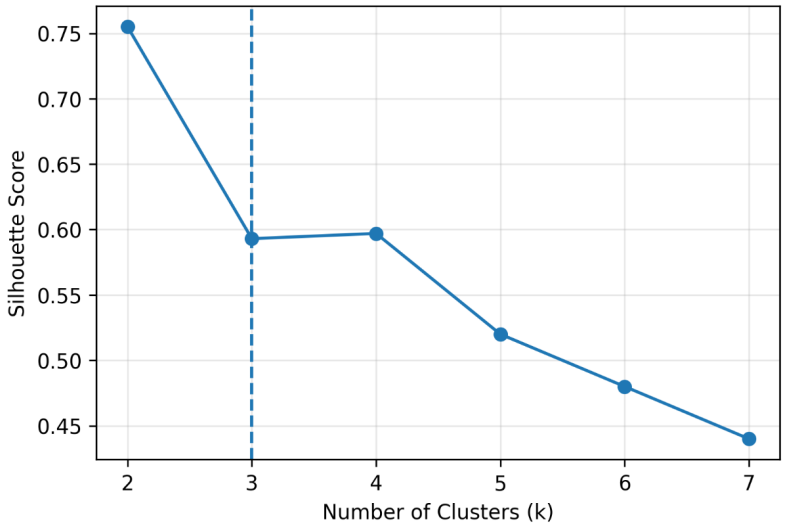

Figure 2. Silhouette score results for determining the number of clusters source

Although k = 2 provides the highest statistical performance, it only produces a binary classification ("high" and "low"), which is not sufficiently informative for policy differentiation. Based on the trade-off between statistical strength and interpretability commonly applied in clustering studies for public policy (Suraya and Wijayanto, 2022), this study selects k = 3 as the final solution. This decision is supported by: (1) a clear elbow at k = 3; (2) a still-strong Silhouette Score (0.593); and (3) meaningful policy-relevant interpretation resulting in three risk groups (low, medium, high) aligned with food security policy needs.

4.4. K-Means Results and Centroid Table

The implementation of K-Means with k = 3 produced three clusters, with centroid characteristics shown in Table 2.

Table 2. Cluster centroids and membership size

Cluster	PoU Centroid (%)	Number of Provinces	Risk Category
0	7.85	26	Low
1	17.16	6	Medium
2	32.13	6	High

Table 3 summarizes the clustering results for all 38 Indonesian provinces based on their average Prevalence of Undernourishment (PoU) values using the K-Means algorithm (k = 3). The analysis classified six provinces into the high cluster (PoU: 27.26-

ISSN: 2502-8521 (Print) / 2503-1155 (Online)
DOI: https://doi.org/10.35457/9wjmrq51

37.69%), six provinces into the medium cluster (PoU: 13.40-23.35%), and twenty-six provinces into the low cluster (PoU: 2.37-11.32%). Provinces in the high cluster are predominantly located in eastern Indonesia, indicating relatively severe food insecurity conditions, whereas most provinces belong to the low cluster, reflecting comparatively better food security status. These findings reveal substantial regional disparities in undernourishment and provide a basis for prioritizing food security policies according to provincial characteristics rather than implementing a uniform national approach.

Table 3. Complete list of 38 provinces with average PoU values and cluster assignments

No	Province	Average PoU (%)	Cluster Category
1	Central Papua	37.69	High
2	Papua	34.77	High
3	Maluku	32.44	High
4	North Maluku	31.38	High
5	South Papua	29.26	High
6	Highland Papua	27.26	High
7	West Papua	23.35	Medium
8	Southwest Papua	20.53	Medium
9	West Kalimantan	17.89	Medium
10	North Kalimantan	14.32	Medium
11	Gorontalo	13.46	Medium
12	East Nusa Tenggara	13.40	Medium
13	Lampung	11.32	Low
14	Central Java	11.20	Low
15	Bangka Belitung Island	11.08	Low
16	Riau	10.88	Low
17	Southeast Sulawesi	10.72	Low
18	Jambi	10.42	Low
19	East Kalimantan	10.11	Low
20	DI Yogyakarta	9.82	Low
21	Central Kalimantan	9.68	Low
22	Central Sulawesi	9.63	Low
23	Bengkulu	9.08	Low
24	Aceh	9.02	Low
25	West Sulawesi	8.80	Low
26	East Java	8.48	Low
27	South Sumatera	8.47	Low
28	South Sulawesi	8.23	Low
29	Riau Island	7.57	Low
30	North Sumatera	6.93	Low
31	West Sumatera	6.58	Low
32	North Sulawesi	5.31	Low
33	West Java	4.87	Low
34	Bali	4.54	Low
35	South Kalimantan	3.36	Low
36	West Nusa Tenggara	2.95	Low
37	Banten	2.55	Low
38	DKI Jakarta	2.37	Low

4.5. Cluster Characteristics and Interpretation

Low Cluster (26 provinces, mean PoU 7.85%). This cluster includes provinces in Java, Bali, parts of Sumatra, and parts of Sulawesi. DKI Jakarta (2.37%), Banten (2.55%), and West Nusa Tenggara (2.95%) represent the lowest PoU values. This cluster is characterized by relatively good infrastructure connectivity, diversified economic structures, and more balanced food distribution access. It indicates provinces that are relatively close to achieving SDG 2 targets, although continued efforts are still required.

Medium Cluster (6 provinces, mean PoU 17.16%). This cluster consists of West Papua, Southwest Papua, West Kalimantan, North Kalimantan, Gorontalo, and East Nusa Tenggara. These provinces generally exhibit challenging geographical conditions (border regions, islands, or mountainous areas) with lower connectivity compared to the Low Cluster but not yet reaching severe food insecurity levels. This cluster represents a vulnerable group requiring preventive policy interventions.

High Cluster (6 provinces, mean PoU 32.13%). This cluster includes Central Papua (37.69%), Papua (34.77%), Maluku (32.44%), North Maluku (31.38%), South Papua (29.26%), and Highland Papua (27.26%). The dominance of Papua and Maluku regions in this cluster is consistent with previous studies using poverty indicators (Bahauddin et al., 2021; Matdoan et al., 2024). Structural characteristics include island and mountainous geography, severe infrastructure constraints in food logistics, high poverty rates, and limited access to basic services. This cluster represents the main priority for national food security interventions.

The dominance of geographically remote and topographically challenging regions in the High Cluster indicates that the primary drivers of high PoU are structural rather than temporary, requiring long-term interventions focused on infrastructure development and food distribution systems rather than short-term food assistance alone.

4.6. Comparison with Previous Studies

This study shows several similarities and differences compared to prior research. First, compared to Bahauddin et al. (2021), who identified Maluku, West Papua, and Papua as the highest poverty cluster, this study finds a very similar clustering pattern (Papua regions and Maluku provinces in the High Cluster), reinforcing the external validity of the results despite using different indicators (PoU vs poverty rate). This supports the structural correlation between poverty and food insecurity.

Second, unlike Suraya and Wijayanto (2022), which compared multiple clustering methods for stunting indicators, this study uses a single method (K-Means) but strengthens validation through a combination of three evaluation metrics (Elbow, Silhouette, and DBI), whereas many previous studies rely on a single metric (Ayalew et al., 2024; Pratiwi et al., 2024).

Third, compared to Ashari et al. (2023), where K-Means achieved a Silhouette Score of 0.466 in a multidimensional flood-risk context, this study achieves a higher score (0.593 at $k = 3$). This difference can be explained by the univariate nature of PoU data, which produces clearer cluster separations compared to multidimensional and overlapping feature spaces.

Fourth, compared to Castelli et al. (2024), which examined SDG clustering in a high-income country context (Italy) with relatively low and homogeneous PoU values,

this study reveals much greater inequality in Indonesia (range: 2.37%-37.69%), reflecting more complex food security challenges in a developing archipelagic country.

Fifth, although this study does not use spatial methods like those applied in African food insecurity hotspot analysis (Ayalew et al., 2024), it still reveals implicit geographical patterns, as the High Cluster is dominated by Papua and Maluku regions. This suggests that geographical determinants can still be captured indirectly through attribute-based clustering approaches.

The main contribution of this study lies in combining rigorous data auditing procedures with policy-relevant interpretation, an approach that is still rarely explicitly integrated in Indonesian government panel data clustering studies.

4.7. Policy Implications for Food Security

The clustering results have tiered policy implications. For High Cluster provinces, interventions should be structural and long-term, including transportation infrastructure development, food logistics system improvement, locally adapted food production systems, and comprehensive nutrition intervention programs. For Medium Cluster provinces, policies should focus on prevention through early warning systems and infrastructure strengthening. For Low Cluster provinces, policies should emphasize maintaining existing achievements and ensuring equitable food access for vulnerable groups that may not be visible in aggregated provincial data.

5. CONCLUSION

This study successfully clustered 38 provinces in Indonesia based on average Prevalence of Undernourishment (PoU) from 2018 to 2024 using the K-Means Clustering method. The analysis involved several stages, including data auditing, cleaning, standardization, and optimal cluster determination. The results identified three clusters: Low (26 provinces, mean PoU 7.85%), Medium (6 provinces, mean PoU 17.16%), and High (6 provinces, mean PoU 32.13%), with High Cluster provinces predominantly located in Papua and Maluku regions.

Low Cluster provinces generally have strong connectivity and diversified economies, while Medium Cluster provinces face geographical constraints but not severe food insecurity. High Cluster provinces exhibit structural limitations in food distribution infrastructure and therefore require long-term policy interventions. The results can assist policymakers in prioritizing food security and nutrition programs in alignment with SDG 2 (Zero Hunger).

This study contributes in three aspects: (1) methodological contribution through a structured data audit protocol addressing administrative inconsistencies; (2) substantive contribution by using official PoU indicators rarely applied in clustering studies; and (3) practical contribution by producing a policy-oriented prioritization map.

Limitations of this study include the use of a single feature (PoU), which can be expanded with additional contextual variables. Future research may apply spatial or spatio-temporal clustering methods to capture temporal dynamics more effectively and compare K-Means with alternative clustering algorithms to enhance result robustness. Continuous data validation is also recommended to ensure data accuracy and reliability.

ACKNOWLEDGMENTS

The author would like to thank the official statistical data provider institutions for making the data used in this study available.

REFERENCES

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020). The Systematic Review of K-Means Clustering Algorithm. *Proceedings 9th International Conference on Networks, Communication and Computing (ICNCC 2020)*, Tokyo, Japan. https://doi.org/10.1145/3447654.3447657

Ashari, F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing*, *7*(1), 95-103. https://doi.org/10.30871/jaic.v7i1.4947

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1-12. https://doi.org/10.3390/electronics9081295

Ayalew, M. M., Dessie, Z. G., Zewotir, T., & Mitiku, A. A. (2024). Exploring the spatial and spatiotemporal patterns of severe food insecurity across Africa (2015-2021). *Scientific Reports*, *14*(29846), 1-14. https://doi.org/10.1038/s41598-024-78616-8

Bahauddin, A., Fatmawati, A., & Sari, F. P. (2021). Analisis Clustering Provinsi Di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Algoritma K-Means. *Jurnal Manajemen Informatika dan Sistem Informasi*, *4*(1), 1-8. https://doi.org/10.36595/misi.v4i1.216

Castelli, T., Mocenni, C., & Dimitri, G. M. (2024). A machine learning approach to assess Sustainable Development Goals food performances: The Italian case. *PLOS One*, *19*(1), 1-20. https://doi.org/10.1371/journal.pone.0296465

Chong, B. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, *4*(5), 37-40. https://doi.org/10.25236/AJCIS.2021.040506

Bajal, E., Katara, V., Bhatia, M., & Hooda, M. (2022). A Review of Clustering Algorithms: Comparison of DBSCAN and K-mean with Oversampling and t-SNE. *Recent Patents on Engineering*, *16*(2), 1-15. https://doi.org/10.2174/1872212115666210208222231

Clark, A. Y., Blumenfeld, N., Lal, E., Darbari, S., Northwood, S., & Wadpey, A. (2021). Using k-means cluster analysis and decision trees to highlight significant factors leading to homelessness. *Mathematics*, *9*(17), 2021, 1-14. https://doi.org/10.3390/math9172045

Food and Agriculture Organization of the United Nations. (2024). Indicator 2.1.1 - Prevalence of undernourishment. https://www.fao.org/sustainable-development-goals-data-portal/data/indicators/2.1.1-prevalence-of-undernourishment/en

Febriansyah, F. & Muntari, S. (2023). Penerapan Algoritma K-Means untuk Klasterisasi Penduduk Miskin pada Kota Pagar Alam. *JISKa: Jurnal Informatika Sunan Kalijaga*, *8*(1), 66-77. https://doi.org/10.14421/jiska.2023.8.1.66-77

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann.

Maori, N. A. & Evanita, E. (2023). Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, *14*(2), 277-288. https://doi.org/10.24176/simet.v14i2.9630

Marcelina, D., Kurnia, A., & Terttiaavini, T. (2023). Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 293-301. https://doi.org/10.57152/malcom.v3i2.952

Matdoan, M. Y., Igo, L., Rumeon, R., Fadhilah, R, & Laamena, N. S. (2024). Penerapan Algoritma K-Means Untuk Klusterisasi Kabupaten/Kota Berdasarkan Tingkat Kemiskinan di Kepulauan Maluku dan Papua. *Jurnal Sains Matematika dan Statistika*, *10*(1), 1-9. https://doi.org/10.24014/jsms.v10i1.21260

Mayasari, S. N. & Nugraha, J. (2023). Implementasi K-Means cluster analysis untuk mengelompokkan kabupaten/kota berdasarkan data kemiskinan di Provinsi Jawa Tengah tahun 2022. *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, *3*(2), 317-329. https://doi.org/10.24002/konstelasi.v3i2.7200

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive Review of K-Means Clustering Algorithms. *IJASRE: International Journal of Advances in Scientific Research and Engineering*, *7*(8), 64-69, 2021. https://doi.org/10.31695/IJASRE.2021.34050

Pratiwi, G. R., Wahiddin, D., Awal, E. E., & Fauzi A. (2024). Klasterisasi tingkat kemiskinan kabupaten/kota di Indonesia menggunakan algoritma K-Means dan K-Medoids. *Jurnal Algoritma*, *21*(2), 197-208.

Rahman, M. A., Sani, N. S., Hamdan, R., Othman, Z. A., & Bakar, A. A. (2021). A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. *PLOS One*, *16*(8), 1-25. https://doi.org/10.1371/journal.pone.0255312

Ritonga, P. K. & Hasibuan, M. S. (2025). Analisis Perbandingan Silhouette dengan Elbow pada Algoritma K-Means dan DBSCAN. *METIK JURNAL*, *9*(1), 64-71. https://doi.org/10.47002/metik.v9i1.1027

Sari, F. D. R. & Ediwijoyo, S. P. (2023). Clustering Analysis Using K-Medoids on Poverty Level Problems in Central Java by District/City. *KnE Social Sciences*, *2023*, 78-87. https://doi.org/10.18502/kss.v8i9.13321

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(31), 1-16. https://doi.org/10.1186/s13638-021-01910-w

Sofyan, H., Iqbal, M., Marzuki, M., & Muhammad, M. (2021). The comparison of k-modes clustering and ROCK clustering to the poverty indicator in Samadua Subdistrict, South Aceh. *IOP Conference Series: Materials Science and Engineering*, *1087*(2021), 1-8. https://doi.org/10.1088/1757-899X/1087/1/012085

Suraya, G. R. & Wijayanto, A. W. (2022). Comparison of hierarchical clustering, k-means, k-medoids, and fuzzy c-means methods in grouping provinces in Indonesia according to the special index for handling stunting. *Indonesian Journal of Statistics and Its Applications*, *6*(2), 180-201. https://doi.org/10.29244/ijsa.v6i2p180-201