ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 5, Issue 1, April 2025, pp. 88-100

Clustering of Provinces in Indonesia Based on Environmental Health Indicators Using K-Medoids

Widya Saputri Agustin [a,1], Safwah Ayu Mardiyyah [a,2], Qolbiyatus Syifa Az Zahra [a,3], Anggun Nur Anggreany [a,4], Edy Widodo [a,5,*]

[a] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Jl. Kaliurang KM 14.5, Yogyakarta 55584, Indonesia
[1] 22611001@students.uii.ac.id; [2] 22611101@students.uii.ac.id; [3] 22611053@students.uii.ac.id; [4] 22611056@students.uii.ac.id; [5] edywidodo@uii.ac.id*
* Corresponding author

ARTICLE INFO

Keywords
SDG
Environmental Health
Silhouette Method
Elbow Method
K-Medoids Clustering

ABSTRACT

According to the Ministry of Health of the Republic of Indonesia, key environmental health indicators include access to safe drinking water, adequate sanitation, and healthy living environments. As of 2023, only 10.21% of Indonesian households had access to safe sanitation, far from the government's 2045 target of 70%. Indonesia's ranking at 164th out of 180 countries in the 2022 environment performance index (EPI), with a score of 28.20 out of 100, further underscores the need for targeted interventions. This study aims to classify Indonesian provinces based on environmental health indicators, thereby supporting more effective policy prioritization. The k-medoids clustering algorithm was employed due to its robustness to outliers and flexibility in handling mixed data types, making it well-suited for this context. This study utilized data from 34 provinces in 2023, sourced from the Ministry of Health. These provinces were grouped into two clusters, with cluster 2 representing provinces with stronger environmental health performance. The clustering results were validated using the silhouette coefficient, confirming the quality of the groupings. Provinces in cluster 1 require greater policy attention to improve environmental health conditions. This study demonstrates the potential of robust medoids-based clustering for guiding targeted environmental health strategies in developing countries.

1. Introduction

Five main objectives of "Indonesia Emas 2045" vision are outlined in the Long-Term National Development Plan (*Rencana Pembangunan Jangka Panjang Nasional*, RPJPN) 2025-2045 vision, one of which is to enhance the competitiveness of human resources. Achieving this goal requires improvements in various sectors, one of which is health. According to the World Health Organization (WHO), environmental health encompasses all physical, chemical, and biological factors external to a person, as well as the factors that can influence human behavior. The condition and management of environmental health have the potential to impact overall health. The Ministry of Health of the Republic of Indonesia identifies key indicators of environmental health, which include access to safe

p-ISSN 2798-253X
e-ISSN 2798-3153

drinking water, adequate sanitation, and healthy districts or cities. Access to safe drinking water and basic services is a national priority, closely linked to other developmental issues such as health, poverty, and human development [1].

Environmental health conditions in Indonesia remain poor compared to other countries globally, including those in Southeast Asia. This is evident in the environmental performance index (EPI) 2022 report. The EPI evaluates 180 countries based on three major pillars: climate change performance, environmental health, and ecosystem vitality. According to the 2022 EPI, Indonesia ranked 164th out of 180 countries, with a score of 28.20 out of 100. In Southeast Asia, Indonesia was ranked 9th among 11 countries evaluated. In the 2020 EPI, Indonesia had a score of 37.08 and ranked 116th. This indicates a significant decline in rank from 116th to 164th [2].

Access to adequate sanitation is closely related to health and environmental conditions. A lack of proper sanitation can degrade water quality and contribute to an increased prevalence of stunting among children, with a correlation coefficient of 0.66 [3]. The correlation coefficient value indicates a fairly strong positive relationship between inadequate sanitation and the prevalence of stunting. This means that when access to sanitation deteriorates, the prevalence of stunting tends to increase. Although it does not indicate a causal relationship, this value suggests a significant association between the two. In Indonesia, challenges related to sanitation behavior remain significant. For instance, the government has set a target of 0% open defecation by households by 2045. However, as of 2023, 4.20% of households still practiced open defecation [4]. Additionally, the government aims to achieve 100% access to safe drinking water by 2045, yet as of 2023, only 91.72% of households had access to safe drinking water [1]. Similarly, access to safely managed sanitation remains low. By 2045, the government targets 70% of households to have access to safely managed sanitation. However, as of 2023, only 10.21% of households had achieved this target [5]. Furthermore, the government aims for 100% of households to reside in adequate housing by 2045. However, as of 2023, only 63.15% of households had access to adequate housing [6]. If these conditions persist, the 2045 targets may not be achieved, necessitating appropriate policies to address these issues. To this end, proper grouping or segmentation is required to determine policy priorities for each province based on its specific characteristics.

Given this gap, understanding regional differences in environmental health conditions is crucial for effective policymaking. Each province in Indonesia faces unique challenges that require targeted interventions rather than a one-size-fits-all approach. Therefore, proper clustering or segmentation is necessary to determine policy priorities for each province based on its specific characteristics. Cluster analysis is a statistical method used to identify natural groupings within a dataset, where objects within a group share greater similarity compared to objects in other groups [7]. In general, there are two common types of clustering algorithms: hierarchical and non-hierarchical methods [8]. Several studies have examined the clustering of provinces in Indonesia based on environmental health indicators, including those by various researchers [9]–[18]. These studies have extensively explored the clustering of provinces in Indonesia using environmental health indicators. However, most of them are limited to methods such as k-means, fuzzy c-means, and cluster ensemble, which have certain drawbacks. Among them, k-means is known for its efficiency in clustering large datasets and handling high-dimensional data. However, it is highly sensitive to outliers, which can significantly affect the position of the cluster center (centroid) and lead to less accurate clustering. Meanwhile, fuzzy c-means allows objects to belong to multiple clusters with varying degrees of membership. However, it remains vulnerable to outliers and is less effective in handling uneven data distributions, reducing the precision of cluster boundaries. On the other hand, cluster ensemble can enhance clustering accuracy but often presents challenges in computational efficiency.

To address the issue of outliers, one of the most robust nonhierarchical clustering methods is k-medoids [19]. This algorithm offers a more resilient approach by selecting actual data points as cluster centers (medoids) rather than centroids, which are easily influenced by extreme values. As a result, k-medoids is more resistant to outliers and produces more stable and representative clustering, especially in datasets with many extreme values [14]. Therefore, this study aimed to implement k-medoids clustering to group provinces in Indonesia based on environmental health indicators in 2023.

2. Method

2.1. Data

The data used in this study consisted of environmental health indicator data from 34 provinces in Indonesia in 2023. The data was obtained from the *Profil Kesehatan Indonesia 2023* published by the Ministry of Health of the Republic of Indonesia. The dataset can be accessed through the website kemkes.go.id/id/profil-kesehatan-indonesia-2023 [20]. The data included 34 provinces and 8 numerical variables. A list of the variables is presented in Table 1.

Table 1. List of Variable

Variable	Indicators
X1	Percentage of villages/subdistricts implementing community-based total sanitation by province in 2023
X2	Percentage of healthy districts/cities by province in 2023
X3	Percentage of households with access to proper drinking water by province in 2023
X4	Percentage of households with access to proper sanitation by province in 2023
X5	Percentage of public places and facilities undergoing supervision according to standards by province in 2023
X6	Percentage of food management places meeting standards by province in 2023
X7	Percentage of households living in decent housing by province in 2023
X8	Percentage of villages/subdistricts with stop open defecation behavior by province in 2023

2.2. Method of Analysis

The data analysis in this study followed several key stages to ensure accurate and meaningful results. It began with collecting secondary data from the 2023 Indonesian health profile and conducting exploratory data analysis to summarize the dataset, calculate descriptive statistics, and check for any missing values. Next, data validation was performed to identify outliers. If no outliers were found, the k-means clustering method was applied. However, if outliers were present, the k-medoids method was used to ensure reliable clustering. The analysis also included checking for multicollinearity using the variance inflation factor (VIF). If multicollinearity was detected, corrective measures were taken before proceeding.

The data was then standardized to ensure all variables had the same scale, which is essential for clustering. Clustering was conducted using the k-medoids method. The optimal number of clusters was determined using the Silhouette method and each cluster was profiled to gain insights into its characteristics. Finally, the results were interpreted, and conclusions were drawn to highlight the patterns and implications of the analysis.

2.3. No Multicollinearity Assumption

Multicollinearity refers to the presence of a strong or exact linear relationship between two or more independent variables in a regression model. It indicates that some predictors can be expressed as a linear combination of others, which can distort the estimation of coefficients. One common method to detect multicollinearity is by calculating the VIF. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value near 1 suggests no multicollinearity, while a value greater than 10 indicates a potential multicollinearity problem that warrants further investigation [21].

$$VIF_j = \frac{1}{1-R_j^2}\ ; j = 1, 2, ..., k \tag{1}$$

where k is the number of independent variables and R_j^2 is the coefficient of determination between the jth independent variable and the other independent variables.

2.4. Data Standardization

Features with large scales or high variability can significantly influence clustering results. Data standardization is an essential preprocessing step to align or control variability within the dataset. Z-score is a form of standardization used to transform a normal variable into a standard score.

Standardized variables will have a mean of 0 and a variance of 1 [22]. The z-score standardization formula is defined as:

$$z = \frac{x-\mu}{\sigma} \quad (2)$$

where z represents the standardized value, x is the original value of the data point, μ is the mean of the data for the given variable, and σ is the standard deviation.

2.5. Silhouette Method

Cluster validation indexes are often used to help determine the optimal number of k clusters during clustering. One popular index is the average silhouette width (ASW), a simple and intuitive tool for measuring cluster quality without relying on statistical model assumptions [23]. The silhouette width represents the difference between the proximity of elements within the same cluster and their distance from other clusters. The silhouette width $s(i)$ for an entity $i \in I$ is defined as:

$$s(i) \frac{b(i)-a(i)}{\max\left(a(i), b(i)\right)} \quad (2)$$

where $a(i)$ is the average distance of i to all other points in the same cluster and $b(i)$ is the smallest average distance of i to points in any other cluster [24].

2.6. Elbow Method

The elbow method is a technique used to determine the optimal number of clusters by analyzing the percentage comparison of results, where the clusters form an "elbow" at a certain point [25]. If the value of the first cluster compared to the second cluster creates an angle in the graph or shows the largest decrease, then that cluster value is considered the optimal choice [26]. Clusters in the data are identified using the within sum of squares (WSS) by minimizing the distance between points within a cluster. WSS represents the sum of squared distances of all points within a cluster [27]. The formula for calculating WSS is as follows:

$$WSS = \sum_{l=1}^{k} \sum_{i \in S_k} \sum_{j=1}^{p} \left(X_{lj} - \bar{X}_{kj}\right)^2 \quad (4)$$

where S_k represents the observations in cluster k, $\bar{X}_{kj}$ denotes the mean of object j from the cluster center for cluster k, and p is the total number of data dimensions [28].

2.7. Clustering

Cluster analysis is a multivariate technique with the primary goal of grouping objects based on the similarity of their characteristics. The characteristics of objects within a cluster have a high degree of similarity, while the characteristics between objects in one cluster and those in another cluster have a low degree of similarity. In other words, the variability within a cluster is minimal, while the variability between clusters is maximal. The difference between one cluster and another is measured using a distance system [15]. The distance measure used in this study is the Euclidean distance. The Euclidean distance is formulated as in (5).

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - y_{jk}\right)^2} \quad (5)$$

where d_{ij} is the distance between object i and object j, x_{ik} is the value of object i on variable k, x_{jk} is the value of object j on variable k, and p is the number of observed variables.

2.8. Nonhierarchical Clustering

Nonhierarchical cluster analysis is a clustering method that determines the number of clusters manually. The nonhierarchical cluster analysis technique is designed to group items, not variables, into a set of K clusters. The number of clusters, K, is predetermined to initiate the clustering procedure [29].

2.9. K-Medoids

The k-medoids method uses representative objects called medoids as the center or centroid of clusters. This method partitions data by minimizing the total dissimilarity between each object i and its closest representative object. Remaining objects are grouped with the most similar representative object and distances are calculated based on the distance between each data point [30]. The k-medoids algorithm uses actual points within a cluster to represent the center, called the medoid. A medoid is a point in the center of a cluster that has the smallest total distance to all the other points in that cluster. Because medoids are more resistant to outliers and noise, they can represent the cluster center more accurately [31]. The foundation of the k-medoids algorithm is to identify *k* clusters within *n* objects by first randomly selecting initial objects (medoids) as representatives for each cluster, then grouping the remaining objects with the most similar medoid. Throughout the clustering process, the k-medoids method consistently uses medoids as reference points [32]. The steps of the k-medoids algorithm are as follows [33].

a. Initialize the cluster centers equal to the number of clusters (k).
b. Allocate each object to the nearest cluster using the Euclidean distance, calculated using (5).
c. Randomly select an object from each cluster as a candidate for the new medoid.
d. Calculate the distance of each object in each cluster to the new medoid candidates.
e. The total deviation (*S*) is obtained by calculating the difference between the new total distance and the old total distance. If the value of *S* is less than 0, an exchange of objects with cluster data is performed to form a new set of cluster objects as the medoids.
f. Repeat steps c through e if there are still changes in the medoids. If no changes occur, the clusters and their respective members are finalized.

3. Results and Discussion

3.1. Descriptive Data Analysis

This study utilized environmental health indicator data from 34 provinces in Indonesia for the year 2023. The data was obtained from the Ministry of Health of the Republic of Indonesia. A summary of the data for each variable is presented in Table 2.

Table 2. Summary of the Variables

Variable	Min	Max	Mean	Median
X1	54.40	83.95	89.86	93.15
X2	0.00	100.00	77.39	100.00
X3	66.49	99.42	88.19	89.97
X4	43.00	96.42	82.57	83.38
X5	49.70	98.20	77.81	76.85
X6	27.70	81.50	59.93	58.50
X7	29.01	85.79	63.00	64.14
X8	30.00	100.00	67.88	68.00

The dataset revealed a few notable patterns. For example, variable X2 exhibited an unusual concentration of values at the extreme, with a median of 100.00 but a mean significantly lower at 77.39, suggesting a potential skew towards lower values despite the high median. Similarly, X6 had relatively low values, as the median and mean were both closer to the lower range, indicating a more concentrated distribution at the bottom end. On the other hand, X3 and X4 exhibited strong concentrations in the upper range, with their medians approaching 90. These patterns indicate that while some variables are relatively evenly distributed, others exhibit distinct skews or concentrations at specific points, offering insights into the data’s structure.

3.2. Outlier Detection

The outlier detection in this study was conducted using robust squared Mahalanobis distance, a method that measures the distance of each data point from the center of the distribution while considering data variability. This method identifies outliers by comparing the distance of a data point to the overall distribution; the greater the distance, the higher the likelihood that the point is considered an outlier. The outlier detection results are presented in Fig. 1.

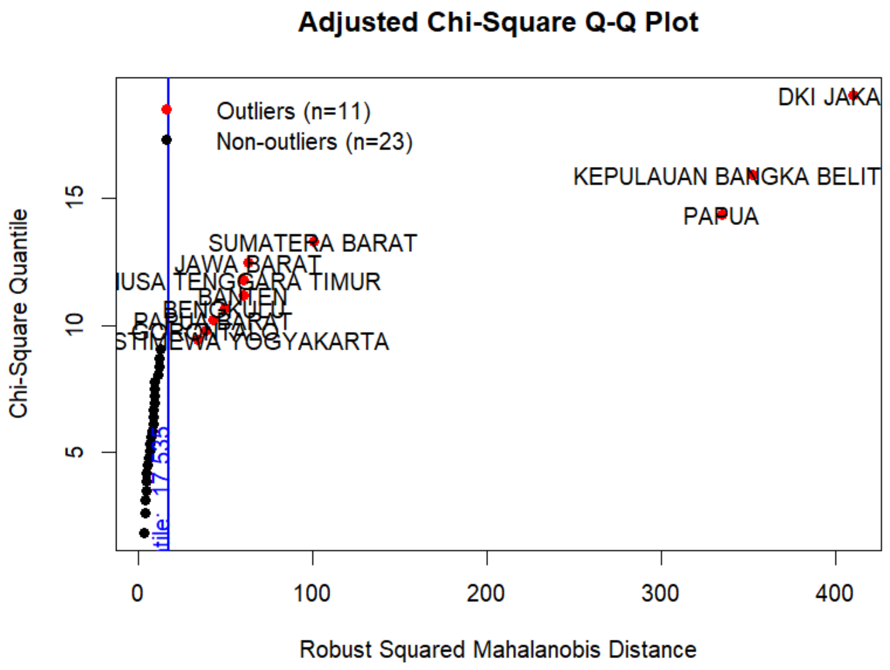


Fig. 1 Chi square Q-Q plot.

The plot in Fig. 1 shows that the red straight line represents the quantile value of 17.535. The black points on the plot indicate nonoutlier data, while the red points represent outlier data [34]. As indicated in the top-left corner, this dataset contained 23 nonoutlier data points and 11 outlier data points. Therefore, the k-medoids method was used in this study, as it can effectively handle these outliers.

3.3. No Multicollinearity Assumption

The next step involved conducting a multicollinearity test within the data. Multicollinearity assessment was performed using the VIF values. The VIF values for each variable are presented in Table 3 to Table 10.

Table 3. VIF for X1

Dependent Variable	X2	X3	X4	X5	X6	X7	X8
X1	2.15	2.49	2.43	1.50	1.19	1.80	1.90

Table 4. VIF for X2

Dependent Variable	X1	X3	X4	X5	X6	X7	X8
X2	2.28	2.55	2.32	1.50	1.42	1.86	1.82

Table 5. VIF for X3

Dependent Variable	X1	X2	X4	X5	X6	X7	X8
X3	3.84	3.71	1.98	1.23	1.44	1.50	1.89

Table 6. VIF for X4

Dependent Variable	X1	X2	X3	X5	X6	X7	X8
X4	4.41	3.98	2.34	1.38	1.72	1.88	1.95

Table 7. VIF for X5

Dependent Variable	X1	X2	X3	X4	X6	X7	X8
X5	4.41	4.17	2.35	2.24	1.58	1.85	1.95

Table 8. VIF for X6

Dependent Variable	X1	X2	X3	X4	X5	X7	X8
X6	3.01	3.40	2.38	2.40	1.36	1.86	1.95

Table 9. VIF for X7

Dependent Variable	X1	X2	X3	X4	X5	X6	X8
X7	4.05	3.95	2.19	2.33	1.41	1.65	1.91

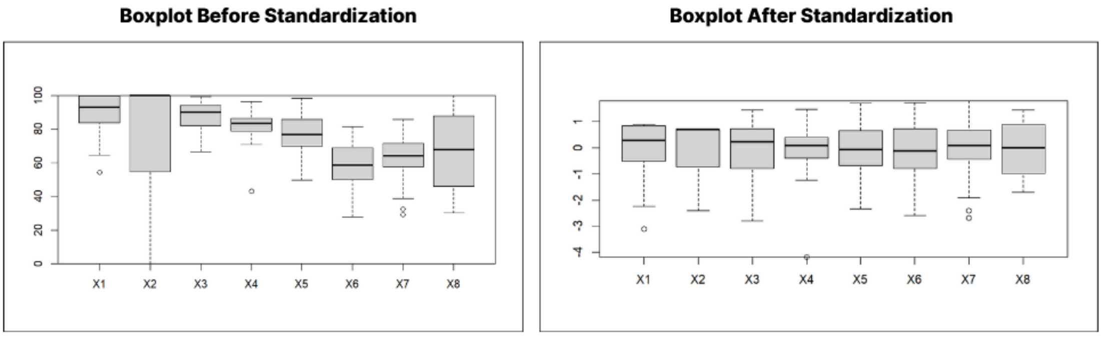
Table 10. VIF for X8

Dependent Variable	X1	X2	X3	X4	X5	X6	X7
X8	4.27	3.85	2.75	2.41	1.49	1.73	1.90

Based on Table 3 to Table 10, the VIF values for all variables are less than 10 (VIF < 10). Since all VIF values are below 10, it can be concluded that there is no multicollinearity among the environmental health indicators, and therefore, the assumption of no multicollinearity is satisfied.

3.4. Standardization

Since the variables used had different data units, the subsequent step was standardization. Standardization is necessary to ensure that variables with differing data units have a uniform scale. This is important to prevent variables with larger value ranges from dominating the analysis or model results. For instance, if one variable has values in the thousands while another has values in decimals, a machine learning model is likely to assign greater weight to the variable with a larger scale, even if it is not inherently more important. Through standardization, all variables will have a mean of zero and a standard deviation of one, ensuring that each variable contributes equally to the analysis. The boxplot visualization before and after standardization can be seen in Fig. 2.


Boxplot Before Standardization

Boxplot After Standardization

Fig. 2 Boxplot before and after standardization.

Based on the boxplot visualization before and after standardization in Fig. 2, it can be observed that the distribution of the standardized variables has become more uniform with a more consistent scale. Before standardization, some variables had much larger value ranges than others, which could affect the k-medoids algorithm in cluster formation. Variables with larger scales could dominate the distance calculations between data points, making the chosen medoid more influenced by those variables. After standardization, the data showed a more even spread, with all variables were on the same scale. This allows the k-medoids algorithm to treat all variables equally in determining the medoid and forming clusters.

3.5. Euclidean Distance

After the data was standardized, the next step was to calculate the distance between observations using the Euclidean method. Euclidean distance measures the proximity between two points in a multivariate space by calculating the square root of the sum of the squared differences between variable values. This process results in a distance matrix that indicates how far each observation is from one another based on the standardized features. The distance between observations is crucial in k-medoids, which form clusters based on the proximity between data points. In k-medoids, this distance is used to select the medoid (the central point of each cluster) and allocate observations to the cluster closest to the medoid. Therefore, calculating the distance between observations is a critical step in ensuring that the clustering process is carried out accurately and by the data structure. The result of the Euclidean distance calculation can be seen in Fig. 3.

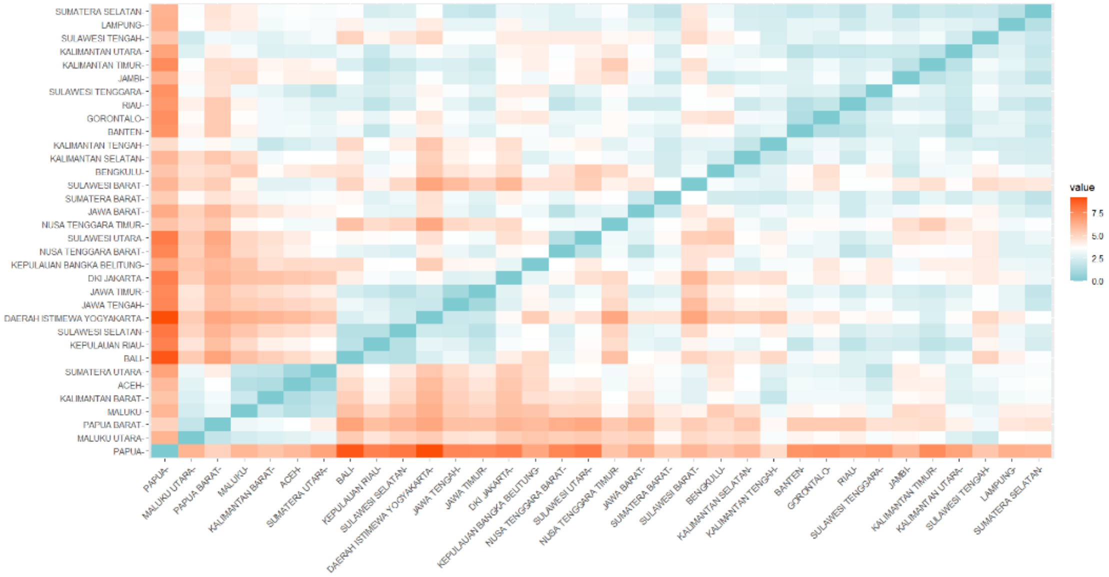


Fig. 3 Heatmap of distances between provinces.

Based on Fig. 3, some provinces with a considerable distance are represented by dark orange, while provinces with closer proximity are shown in blue. The white color indicates a medium level of distance. For example, East Java and South Sumatra provinces are colored blue, indicating that these two provinces are geographically close to each other. On the other hand, Bali and Papua provinces are represented in dark orange, signifying a considerable distance between them.

3.6. Number of Optimal Clusters

After calculating the Euclidean distance, the subsequent step was to determine the optimal number of clusters using the silhouette method and the elbow method. The visualization of the optimal number of clusters can be seen in Fig. 4 and Fig. 5.

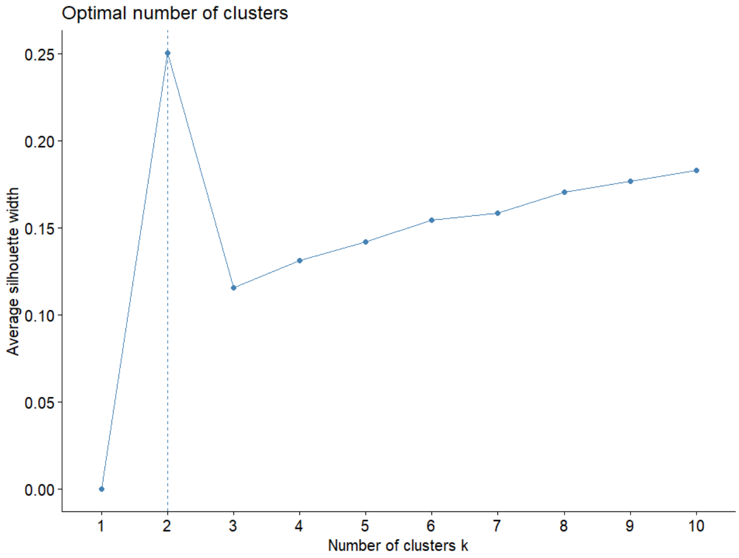


Fig. 4 Optimization number of clusters using silhouette method.

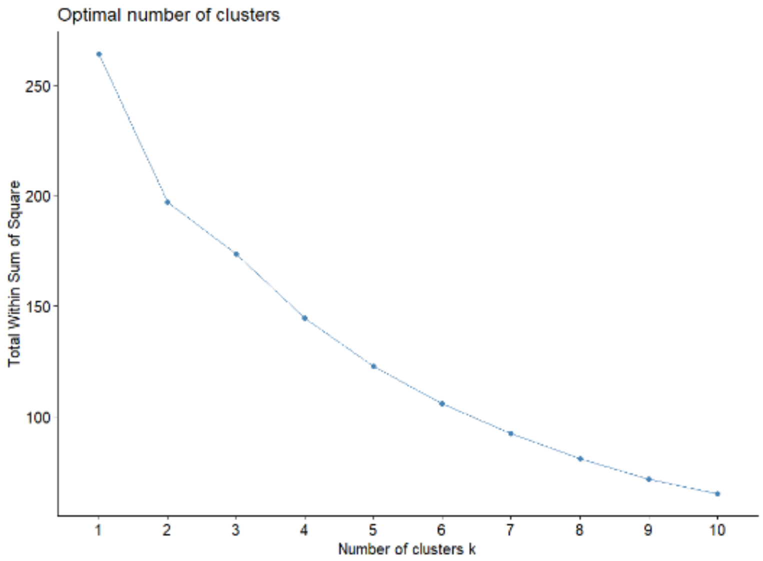


Fig. 5 Optimization number of clusters using the elbow method.

Fig. 4 shows the optimal cluster results based on the silhouette method. The recommended optimal number of clusters, or *k*, is when *k* = 2. This is indicated by the vertical line on the x-axis at *k* = 2. The optimization using the elbow method in Fig. 5 shows a noticeable bend at k = 2, indicating that the optimal *k* value based on the WSS calculation is *k* = 2. The calculation results from the silhouette method and the elbow method both indicate an optimal number of clusters as *k* = 2. Therefore, in this study, the provinces in Indonesia were divided into two clusters based on environmental health indicators for the year 2023.

3.2. K-Medoids Clustering Analysis

Based on the results of the k-medoids clustering, the researcher identified two clusters. Cluster 1 consisted of 10 provinces, while cluster 2 consisted of 24 provinces. The results of the cluster analysis can be seen in Fig. 6 and Table 11.

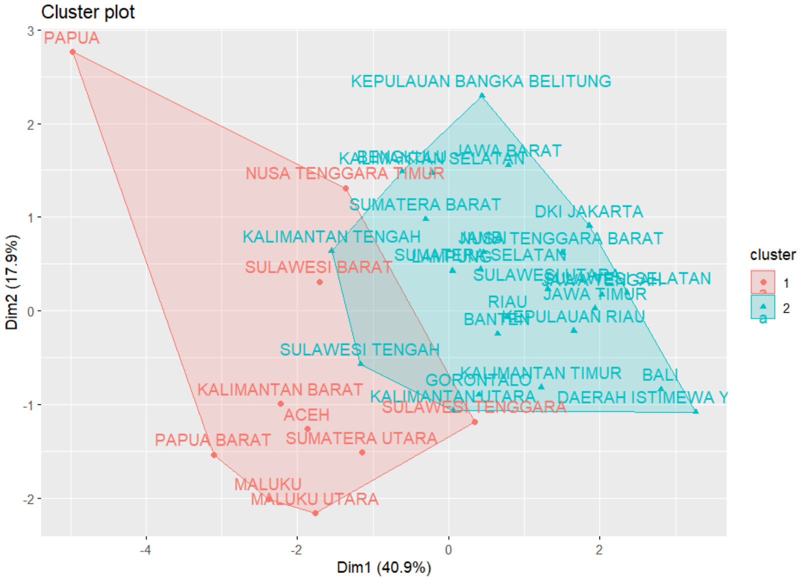


Fig. 6 Plot of the k-medoids clustering results.

In Fig. 6 above, the plot of the k-medoids clustering results displays two different colors representing each cluster. Members of cluster 1 are shown in red, while members of cluster 2 are shown in blue. The members and the number of members in each cluster can be seen in Table 11.

Table 11. Cluster Members

Cluster	Number of Provinces	Provinces
1	10	Aceh, West Kalimantan, Maluku, North Maluku, East Nusa Tenggara, Papua, West Papua, Wst Sulawesi, Southeast Sulawesi, North Sumatra
2	24	Bali, Banten, Bengkulu, D.I. Yogyakarta, DKI Jakarta, Gorontalo, Jambi, West Java, Central Java, East Java, South Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Bangka Belitung Islands, Riau Islands, Lampung, West Nusa Tenggara, Riau, South Sulawesi, Central Sulawesi, North Sulawesi, West Sumatra, South Sumatra

The map visualization for each cluster is shown in Fig. 7, which was generated using the QGIS software. The red color represents provinces included in cluster 1, while the blue color represents provinces included in cluster 2.

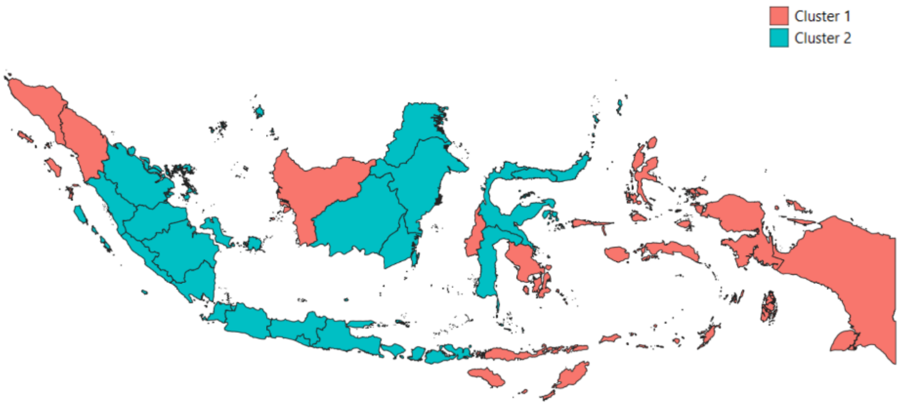


Fig. 7 Visualization map for each cluster.

3.7. Cluster Profiling and Interpretation

After obtaining the cluster members, the next step was to calculate the average for each cluster, the results of which can be seen in Table 12. These average calculations were used for data profiling, aiming to examine the characteristics of each cluster.

Table 12. Cluster Profiling

Cluster	X1	X2	X3	X4	X5	X6	X7	X8
1	77.74	32.54	85.70	76.64	71.79	55.21	59.47	49.63
2	94.91	96.07	89.22	85.04	80.33	61.89	64.46	75.48

In Table 12, the green color represents the group with high environmental health indicator characteristics, while the yellow color indicates the group with low environmental health indicators. Cluster 1, which has lower average values on almost all environmental health indicators, reflects a group with environmental conditions that are less supportive of health. In contrast, cluster 2 shows better environmental health indicators with higher values on all measured variables. This reflects that cluster 2 is the group with higher environmental health.

Based on Fig. 7, it can be observed that the provinces in each cluster exhibit significant geographical differences between cluster 1 and cluster 2. Cluster 1 consists of the provinces of Aceh, West Kalimantan, Maluku, North Maluku, East Nusa Tenggara, Papua, West Papua, West Sulawesi, Southeast Sulawesi, and North Sumatra. Geographically, Cluster 1 is predominantly located in the eastern regions of Indonesia, including Maluku, East Nusa Tenggara, Papua, and Sulawesi, as well as several provinces in western Indonesia, such as Aceh and North Sumatra. These areas generally feature diverse geographical characteristics, ranging from mountainous regions and archipelagos to remote inland areas. Such geographical conditions often pose challenges to infrastructure development, accessibility, and public services, including environmental health. For instance, the presence of numerous islands and remote areas can hinder the implementation of sanitation programs, water management, and public facility supervision.

Conversely, cluster 2 encompasses the provinces of Bali, Banten, Bengkulu, D.I. Yogyakarta, DKI Jakarta, Gorontalo, Jambi, West Java, Central Java, East Java, South Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Bangka Belitung Islands, Riau Islands, Lampung, West Nusa Tenggara, Riau Islands, South Sulawesi, Central Sulawesi, North Sulawesi, West Sumatra, and South Sumatra. Cluster 2 is more dominant in the western and central regions of Indonesia. Provinces in cluster 2 generally have better infrastructure access, including transportation networks, water distribution, and sanitation services. The geographical conditions of these regions, such as lowland areas and easier access to economic and administrative centers, support the more optimal implementation of environmental health programs.

The profiling results indicate significant differences between cluster 1 and cluster 2 based on environmental health indicators. Cluster 1 represents areas with environmental conditions that are less supportive of health. This is evident from the lower implementation of the community-based total sanitation program (X1) and the limited number of villages/subdistricts practicing stop open defecation behavior (X8). Additionally, the number of healthy districts/cities (X2) and community access to proper drinking water facilities (X3) and proper sanitation (X4) are also lower compared to cluster 2. Supervision of public facilities (X5) and food management places (X6) is also inadequate, further compounded by substandard housing conditions (X7).

In contrast, cluster 2 represents areas with better environmental health conditions. These areas have successfully implemented the community-based total sanitation program more broadly (X1) and have a higher number of villages/subdistricts practicing stop open defecation behavior (X8). The number of healthy districts/cities (X2) in cluster 2 is also more prominent, along with better access to proper drinking water facilities (X3) and proper sanitation (X4). Moreover, supervision of public facilities (X5) and food management places (X6) meets higher standards, and housing conditions are more adequate (X7).

It can be observed that cluster 2 represents regions with better sanitation and environmental management compared to cluster 1. Indicators such as the implementation of community-based total sanitation (X1), the number of healthy districts/cities (X2), public facility supervision (X5), and the behavior of stopping open defecation (X8) are the primary distinguishing factors. Among these, the most significant differentiating indicator is the implementation of community-based total sanitation (X1), where cluster 2 has implemented this program more extensively compared to cluster 1, which remains very limited. Furthermore, the number of healthy districts/cities (X2) is significantly higher in cluster 2, reflecting stronger efforts to create a health-supportive environment. Public facility supervision (X5) in cluster 2 is conducted more meticulously and adheres to higher standards, whereas in cluster 1, it remains inadequate. Another striking variable is the behavior of stopping open defecation (X8), where cluster 2 has more villages/sub-districts adopting this practice, reflecting better sanitation behavior and awareness. The regions within cluster 1 require enhanced attention through well-targeted policies and strategic interventions to improve access to sanitation services, environmental management, and the quality of public facilities. These efforts are essential to bridging the gap and aligning the conditions in cluster 1 with the advancements achieved by Cluster 2.

4. Conclusion

The implementation of k-medoids cluster analysis to group the provinces in Indonesia was conducted based on eight environmental health indicator variables. According to the silhouette method, the optimal number of clusters was determined to be 2. The significant differences between these two clusters indicate an imbalance in the distribution of environmental health across the regions analyzed.

In cluster 1, the low indicator values may signal areas with environmental issues that require more attention. Meanwhile, in cluster 2, the high indicator values reflect areas with healthier environments, which could serve as models or benchmarks for other regions. The government needs to identify and provide additional attention to the areas classified in cluster 1, which have lower environmental health indicators. For the provinces in cluster 1, various intervention programs can be implemented, such as improving access to clean water, promoting awareness about waste disposal and open defecation, as well as health and hygiene education campaigns. These efforts aim to reduce the regional disparities and achieve a more equitable distribution of environmental health across all provinces in Indonesia.

References

[1] Kementerian Kesehatan Republik Indonesia, "Profil Kesehatan Indonesia Tahun 2019." 2020. [Online]. Available: https://kemkes.go.id/id/profil-kesehatan-indonesia-2019

[2] N. Siddiqui, "India in the Environmental Performance Index," *Econ. Polit. Wkly.*, vol. 57, no. 25, Jun. 2022.

[3] F.E. Linder, "National Health Survey," Science, vol. 127, no. 3309, pp. 1275–1279, May 1958, doi: 10.1126/science.127.3309.1275.

[4] Badan Pusat Statistik, "Persentase Rumah Tangga yang Masih Mempraktikkan Buang Air Besar Sembarangan (BABS) di Tempat Terbuka menurut Provinsi dan Klasifikasi Desa (Persen), 2023-2024," 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MjE3NiMy/persentase-rumah-tangga-yang-masih-mempraktikkan-buang-air-besar-sembarangan-babs-di-tempat-terbuka-menurut-provinsi-dan-tipe-daerah.html

[5] Badan Pusat Statistik, "Persentase Rumah Tangga Menggunakan Layanan Sanitasi yang Dikelola Secara Aman Menurut Provinsi dan Tipe Daerah (Persen), 2023-2024," 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MjE3OSMy/persentase-rumah-tangga-menggunakan-layanan-sanitasi-yang-dikelola-secara-aman-menurut-provinsi-dan-tipe-daerah--persen-.html

[6] Badan Pusat Statistik, "Persentase Rumah Tangga yang Memiliki Akses Terhadap Hunian yang Layak Menurut Klasifikasi Desa (Persen), 2021-2023," 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MTI0MiMy/persentase-rumah-tangga-yang-memiliki-akses-terhadap-hunian-yang-layak-menurut-klasifikasidesa.html

[7] U. Rahardja, Q. Aini, and M. Iqbal, "Analisis cluster dalam pengelompokan provinsi di indonesia berdasarkan variabel penyakit menular menggunakan metode complete linkage, average linkage dan ward," *InfoTekJar, J. Nas. Inform. Teknol. Jar.*, vol. 5, no. 1, pp. 40–43, Mar. 2020, doi: 10.30743/infotekjar.v5i1.2464

[8] H. Hikmah, F. Fardinah, L. Qadrini, and E. Tande, "Analisis klaster pengelompokan kecamatan di sulawesi barat berdasarkan indikator pendidikan," Saintifik, vol. 8, no. 2, pp. 188–196, 2022, doi: 10.31605/saintifik.v8i2.383.

[9] Mukidin, "Clustering tingkat kesehatan lingkungan berdasarkan data penyehatan lingkungan pemukiman menggunakan metode fuzzy c – means (studi kasus: dinas kesehatan kab. cirebon)," *J. Ilm. Indonesia*, vol. 4, no. 2, pp. 22–31, 2019, doi: 10.36418/syntax-literate.v4i2.551

[10] A.P. Irfan, S. Aminah, S. Cokrowibowo, and N. Zulkarnaim, "Clustering wilayah berdasarkan data kesehatan lingkungan menggunakan fuzzy c-means," *J. Comput. Inf. Syst.*, vol. 1, no. 2, pp. 12–22, Apr. 2020, doi: 10.31605/jcis.v1i2.609.

[11] N.S. Belinda, I.R. HG, and H. Yozza, "Penerapan analisis cluster ensemble dengan metode rock untuk mengelompokkan provinsi di Indonesia berdasarkan indikator kesejahteraan rakyat," *J. Mat. UNAND*, vol. 8, no. 2, pp. 108–119, Aug. 2019, doi: 10.25077/jmu.8.2.108-119.2019.

[12] M.R. Ikhsanudin and A.W. Wijayanto, "Perbandingan pengelompokkan provinsi di Indonesia menurut kualitas lingkungan hidup menggunakan metode hierarki dan partisi," *J. Sist. Teknol. Inf.*, vol. 12, no. 1, pp. 155–163, 2024, doi: 10.26418/justin.v12i1.71495.

[13] T.R. Mayasari, "Pengelompokkan provinsi berdasarkan variabel kesehatan lingkungan dan pengaruhnya terhadap kemiskinan di Indonesia tahun 2018*," J. Siger Mat.*, vol. 1, no. 1, pp. 24–30, Mar. 2020, doi: 10.23960/jsm.v1i1.2471.

[14] A. Fadlurohman and I.M. Nur, "Pengelompokan provinsi di Indonesia berdasarkan indikator perumahan dan kesehatan lingkungan menggunakan metode k-medoids," in *Pros. Semin. Nas. UNIMUS,* 2023, pp. 1168–1180.

[15] W. Wijayanti, I.R. HG, and F. Yanuar, “Penggunaan metode fuzzy c-means untuk pengelompokan provinsi di Indonesia berdasarkan indikator kesehatan lingkungan,” *J. Mat. UNAND*, vol. 10, no. 1, pp. 129–136, Jan. 2021, doi: 10.25077/jmu.10.1.129-136.2021.

[16] M.S. Kudadiri, P. Silvianti, and F.M. Afendi, “Pengelompokan provinsi berdasarkan capaian indikator kesehatan lingkungan di Indonesia tahun 2020,” *Xplore, J. Stat.*, vol. 11, no. 3, pp. 191–202, Sep. 2022, doi: 10.29244/xplore.v11i3.879.

[17] F.S. Pratiwi, S. Sudarno, and A. Rusgiyono, “Penerapan response based unit segmentation in partial least square (REBUS-PLS) untuk analisis dan pengelompokan wilayah (studi kasus: kesehatan lingkungan perumahan di Provinsi Jawa Tengah),” *J. Gaussian*, vol. 9, no. 3, pp. 364–375, Aug. 2020, doi: 10.14710/j.gauss.v9i3.28927.

[18] S.P. Dewi and M. Bin Othman, “Implementation of cluster k-means for the East Java environmental health areas grouping in 2017,” *J. Biometrika Kependud.*, vol. 9, no. 1, pp. 1–9, Jun. 2020, doi: 10.20473/jbk.v9i1.2020.1-9.

[19] R.E. Sihombing, D. Rachmatin, and J.A. Dahlan, “Program aplikasi bahasa R untuk pengelompokan objek menggunakan metode k-medoids clustering,” *J. EurekaMatika*, vol. 7, no. 1, pp. 58–79, May 2019, doi: 10.17509/jem.v7i1.17888.

[20] Kementerian Kesehatan Republik Indonesia, “Profil Kesehatan Indonesia 2023,” 2024. [Online]. Available: https://kemkes.go.id/id/profil-kesehatan-indonesia-2023

[21] N.S. Ibrahim, “Analisis diskriminan linear robust dengan penduga minimum covariance determinant (Studi kasus: Indeks kerentanan pangan menurut kabupaten/kota di Indonesia tahun 2023),” *Emerg. Stat. Data Sci. J.*, vol. 2, no. 2, pp. 264–279, 2024, doi: 10.20885/esds.vol2.iss.2.art20.

[22] I. Bin Mohamad and D. Usman, “Standardization and its effects on k-means clustering algorithm,” *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, 2013, doi: 10.19026/rjaset.6.3638.

[23] F. Batool and C. Hennig, “Clustering with the average silhouette width,” *Comput. Stat. Data Anal.*, vol. 158, Jun. 2021, doi: 10.1016/j.csda.2021.107190.

[24] P.A. Rizaldi, M. Hakimah, and T. Indriyani, “Penentuan jurusan siswa SMA menggunakan metode k-means ++,” in *Semin. Nas. Sains Teknol. Terap*, 2022, pp. 1–7.

[25] A.T. Rahman, W. Wiranto, and R. Anggrainingsih, “Coal trade data clustering using k-means (case study PT. Global Bangkit Utama),” *ITSMART, J. Teknol. Inf.*, vol. 6, no. 1, pp. 24–31, Jun. 2017, doi: 10.20961/itsmart.v6i1.11296.

[26] E. Muningsih and S. Kiswati, “Sistem Aplikasi berbasis optimasi metode elbow untuk penentuan clustering pelanggan,” *Joutica*, vol. 3, no. 1, pp. 117–124, Apr. 2018, doi: 10.30736/jti.v3i1.196.

[27] B.S.A. Arif, A. Rusgiyono, and A. Hoyyi, “Pengelompokan provinsi-provinsi di Indonesia menggunakan metode ward (Studi kasus: Produksi tanaman pangan di Indonesia tahun 2018),” *J. Gaussian*, vol. 9, no. 1, pp. 112–121, Feb. 2020, doi: 10.14710/j.gauss.v9i1.27528.

[28] T.M. Kodinariya and P.R. Makwana, “Review on determining number of cluster in k-means clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 2321–7782, Nov. 2013.

[29] M. Afriana, S. Nugroho, and F. Pachri, “Penentuan awal keanggotaan analisis klaster non hirarki (K-means),” Undergraduate thesis, Math. Study Progr., Universitas Bengkulu, Bengkulu, Indonesia, 2023.

[30] A. Rofifah, R. Goenjantoro, and D. Yuniarti, “Perbandingan pengelompokan k-means dan k-medoids pada data potensi kebakaran hutan/lahan berdasarkan persebaran titik panas (Studi kasus: Data Titik panas di Indonesia pada 28 April 2018),” *J. Eksponensial*, vol. 10, no. 2, pp. 143–152, Nov. 2019.

[31] N. Sureja, B. Chawda, and A. Vasant, “An improved k-medoids clustering approach based on the crow search algorithm,” *J. Comput. Math. Data Sci.*, vol. 3, Jul. 2021, 2022, doi: 10.1016/j.jcmds.2022.100034.

[32] [S. Defiyanti, M. Jajuli, and N. Rohmawati, “Optimalisasi K-medoid dalam pengklasteran mahasiswa pelamar beasiswa dengan cubic clustering criterion,” *J. Nas. Teknol. Sist. Inf.*, vol. 3, no. 1, pp. 211–218, Apr. 2017, doi: 10.25077/teknosi.v3i1.2017.211-218.

[33] C. Astria, A.P. Windarto, and D. Hartama, “Penerapan k-medoid pada rumah tangga yang memiliki sumber penerangan listrik PLN berdasarkan provinsi,” KOMIK (*Konferensi Nas. Teknol. Inf. Komput.*), vol. 3, no. 1, pp. 604–609, Oct. 2019, doi: 10.30865/komik.v3i1.1667.

[34] M. Muhajir, *Modul Praktikum Statistika Multivariat Terapan*. Yogyakarta, Indonesia: Universitas Islam Indonesia, 2021.