Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Cluster Modeling with K-Means on Provincial Data in Indonesia Based on Environmental Indicators

Diana Novitasari[1], Amellia Harmaimun Hidayah[2*], Rosyidatul Kamila[3], Trimono[4], Muhammad Nasrudin[5]

[1,2,3,4,5]*Veteran National Development University of East Java*
ameliahidayah02@gmail.com[2*]

Abstract

Population growth and economic activity in Indonesia significantly affect the quality of the environment. The government uses the Environmental Quality Index as a comprehensive measurement tool, which considers aspects of water, soil, and air pollution, as well as demographic variables such as population size and land area. This study aims to identify groups of 33 provinces in Indonesia based on pollution and demographic characteristics by applying the K-Means algorithm. The data, sourced from the Central Statistics Agency (BPS), underwent a series of stages: pre-processing, standardization, and evaluation using the Elbow, Silhouette Score, and Dunn Index methods. The clustering results identified two main groups. The first cluster consists of three provinces on the island of Java, which exhibit high population density and pollution levels. Meanwhile, the second cluster includes the remaining 30 provinces with more diverse characteristics. These findings are expected to support the formulation of more specific and evidence-based environmental policies.

Keywords: *Environmental Quality, K-Means, Clustering, Pollution, Population Density*

1. Introduction

Environmental quality is a fundamental element in assessing sustainable development in an area. As a comprehensive measurement tool, the government relies on the Environmental Quality Index (EQI). This index consists of several key components—such as air quality, water quality, and land cover which collectively show how environmental pressures interact with conservation efforts [1]

The increase in population and rapid economic activity in Indonesia over the past few decades has had serious consequences for environmental carrying capacity [2]. Population density often leads to higher energy consumption, intensive land use, and increased waste, all of which exacerbate pollution in the air, water, and soil. Thus, environmental pollution is the primary cause of declining quality of life. Previous studies indicate a strong negative correlation between air and water pollution and the environmental quality index, particularly in areas dominated by intensive industrial and transportation activities [2]. The size and geographical conditions of a province also contribute to the extent to which environmental impacts can be mitigated or managed effectively [3].
Therefore, the application of sustainable development principles is vital for preserving environmental quality. In this context, it is important to identify the correlation between factors.

2. Research Metode

2.1. Research Approach

This study adopts an exploratory quantitative methodology by applying a non-hierarchical cluster analysis method, namely the K-Means algorithm, to group 34 provinces in Indonesia based on pollution and environmental characteristics. Rather than testing hypotheses, this approach focuses on exploring hidden patterns or structures in data related to environmental conditions in each province [4].
The K-Means method was chosen for its advantages in computational efficiency and ease of interpretation, especially when dealing with large-scale and multivariate data [5]. The clustering process was carried out based on the similarity of values between provinces for six variables covering aspects of environmental pollution and demographic characteristics.

2.2. Data Sources and Types

The data as the basis for this study is secondary data sourced from the official portal of the Central Statistics Agency (BPS). The dataset covers 33 provinces in Indonesia with the following attributes:

Table 1: Dataset Attributes

Number	Attributes
1	Water Pollution
2	Soil Pollution
3	Air Pollution
4	Population
5	Area of Each Province
6	Environmental Quality Index

2.3. Pre-Processing

Data pre-processing is carried out to ensure that the data used is ready for accurate and efficient analysis.

2.3.1. Handling Missing Values

Each variable was examined to detect missing values. If empty values were found, imputation was performed using the mean imputation so as not to interfere with the clustering process.

2.3.2. Number Format Transformation

Some variables, such as Population and IKLH, still use decimal format with commas (","). These values are converted to numeric format (float) by replacing commas with periods (".") so that they can be read by the Python system.

2.3.3. Data Standardization

Standardization is necessary because the scales between variables differ. For example, the population size is measured in thousands, while air pollution is measured in hundreds. Therefore, Z-score standardization is performed using the following formula:

$$z = \frac{\chi_{i-}\mu}{\sigma} \quad (1)$$

Description:
z : *Z-Score* value
χ : observation value
μ : mean
σ : standard deviation
This standardization aims to ensure that each variable has an equal contribution in the cluster formation process.

2.4. Data Analysis Methods

2.4.1. K-Means Algorithm

As a non-hierarchical clustering method, the K-Means algorithm is applied to group data into k clusters based on the proximity of the distance between data. The main objective of this algorithm is to minimize variation within each cluster (intra-cluster) and maximize differences between clusters (inter-cluster). The clustering process in this study was carried out through the following series of stages :

1) Determining the Number of Clusters (k)
   In the initial stage, it is important to determine the number of clusters to be formed. In this study, the value of k was determined empirically using two auxiliary methods, namely the Elbow method and the Silhouette Coefficient, to obtain the optimal number of clusters.
2) Initializing the Initial Centroid
   After the value of k is determined, the system will randomly select k initial points that will serve as the temporary centers of each cluster (initial centroids).
3) Calculating the Distance to the Centroid
   Each object in the dataset will be calculated for its distance to all centroids using the Euclidean metric, which is formulated as follows:

$$dEuclidean = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (2)$$

   Description:
   a_i : i-th value of the first object
   b_i : i-th value of the second object
   n : the number of data variables being compared
4) Grouping Objects Based on Nearest Centroid
   Each data point is then allocated to the cluster that has the closest distance to the centroid based on the previous calculation results.
5) Centroid Update

After all data has been clustered, the centroids are updated by calculating the average of all data points in each cluster. The centroid update formula is expressed as follows:

$$C_k = \frac{1}{n_k}\sum d_i \quad (3)$$

Description:
C_k : new centroid position for the k-th cluster
n_k : number of data or members in the kt-h cluster
d_i : i-th data value belonging to the kt-h cluster

6) Iteration until Convergent
Steps 3 to 5 are performed iteratively until the centroid position does not change significantly or has reached a predetermined maximum number of iterations.

2.4.2. Determining the Number of Clusters (k)

Determining the number of clusters (k) is an important step in the clustering process using the K-Means algorithm. The value of k chosen will greatly affect the grouping results, as it is directly related to the representation of the data patterns formed. Therefore, the selection of the number of clusters must be done carefully, taking into account the quality of the resulting clusters. In this study, the optimal number of clusters was determined empirically using two common approaches: the Elbow Method and the Silhouette Coefficient. These methods help evaluate cluster structure based on data distribution and proximity, thereby providing a strong foundation for determining the number of clusters that best align with the data characteristics.

a) Elbow Method
The Elbow Method works by calculating the Within-Cluster Sum of Square (WCSS) value for various *k* values, then visualizing it in graph form. WCSS measures the total squared distance between each data point and its cluster centroid. The lower the WCSS value, the better the clustering. However, continuously increasing the number of clusters will result in a negligible decrease in WCSS after a certain point. The inflection point on the WCSS graph is called the "elbow" and is considered the optimal number of clusters [6]

b) Silhouette Coefficient
The Silhouette Coefficient measures how close a data point is to its cluster compared to other nearby clusters. This coefficient ranges from -1 to 1, where higher values indicate better clustering [7]. Silhouette Coefficient Formula:

$$\mathrm{s(i)} = \frac{b(i)-a(i)}{\max\,(a(i),b(i))} \quad (4)$$

Description:
$a(i)$: average distance between i-th data and all other data in the same cluster
$b(i)$: average distance between i-th data and all other data in the next closest cluster
The k value that produces the highest average Silhouette Coefficient is considered the optimal number of clusters.

2.4.3. Evaluating Clustering Results

Evaluation of clustering results is done to assess the extent to which the method used is able to form good data groups. In general, the quality of clustering is considered good if each cluster has homogeneous members (internally compact) and between clusters are significantly different from each other (externally separated). One of the internal evaluation methods used in this study is the Dunn Index. This metric was introduced by J.C. Dunn in 1974 as a measuring tool to determine the quality of clusterization without the need to know the class label on the data. The main principle of the Dunn Index is to compare the minimum distance between clusters with the maximum diameter of any cluster [8]. The Dunn Index value is calculated by the following formula:

$$DI = \frac{\min d(_{c_p,c_q})}{\max diam(c_r)} \quad (5)$$

Description:
$d(_{c_p,c_q})$: the minimum distance between clusters c_p and c_q, usually calculated using Euclidean distance.
$diam(c_r)$: the maximum diameter of cluster c_r, i.e. the maximum distance between two points in the cluster.
A high Dunn Index value indicates that the clusters are far apart and the members within each cluster are quite compact. Conversely, a low value may indicate overlap between clusters or too wide a spread of data within a cluster. In this study, the Dunn Index is used to compare clustering results from various methods, as well as a consideration in determining the best number of clusters based on the highest index value.

2.5. Research Flowchart

The research flowchart illustrates the systematic flow carried out in the process of clustering provinces in Indonesia based on pollution indicators and environmental characteristics. The preparation of the flowchart also helps clarify the structure of the research work and

makes it easier for readers to understand the entire process taken. A detailed flowchart of the steps of this research can be seen in Figure 1, which visualizes the process from data collection to evaluation and visualization of clustering results.

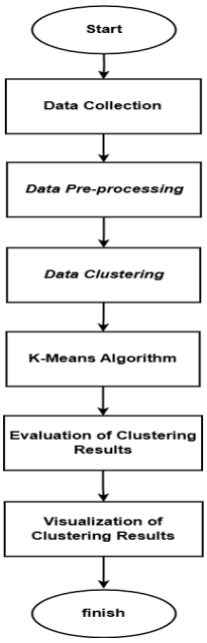

Fig. 1: Flowchart of Provincial Clustering Process Based on Pollution and Environmental Indicators

2.6. Visualizing Clustering

Visualization of clustering results aims to facilitate understanding of data grouping patterns in two-dimensional space. Given that the research data consists of six variables, dimensional reduction is carried out by utilizing the Principal Component Analysis (PCA) method so that the data can be visualized more intuitively. The PCA method summarizes information from the original variables into two principal components that retain most of the data variance. The results are then visualized in the form of a scatter plot, where each dot represents one province, and different colors indicate the cluster to which the province belongs. This visualization facilitates the identification of separation between clusters, density within each group, as well as relative relationships between provinces based on similarities in pollution and environmental characteristics.

3. Results and Discussion

3.1. Data Description

The data used in this study was obtained from environmental management documentation from 33 provinces in Indonesia. The secondary data consists of numerical variables, including Water Pollution, Soil Pollution, Air Pollution, Population Size, Area Size, and Environmental Quality Index (EQI). Each variable represents the environmental and demographic characteristics of each province, which will be used in the cluster analysis process.

3.2. Data Pre-processing and Descriptive Statistics

a. Handling Missing Values
   All numeric variables were checked for missing values, and the results showed that the entire table contained no missing values, so no imputation was necessary. This indicates that the data is ready for further analysis.
b. Descriptive Statistics

Table 2: Descriptive Statistics

Variable	Mean	Standard Deviation	Minimal	Maximal
Water Pollution	324.909091	358.549627	21.000000	1366.000000
Soil Pollution	25.363636	29.900688	-26.000000	122.000000
Air Pollution	140.363636	143.235780	4.000000	583.000000
Population	8395.797879	11865.581186	739.800000	50345.200000
Area of Each Province	46680.933333	38729.865870	660.980000	153443.910000
Environmental Quality Index	72.820606	5.693205	54.650000	84.220000

Based on descriptive statistics, all variables show considerable variation between provinces. Water, soil, and air pollution have moderate average values but with large standard deviations, indicating significant differences in pollution levels. The population and area of provinces also vary greatly, from small to very large. The Environmental Quality Index (EQI) is relatively stable across all provinces, with an average of around 72.8 and low dispersion, indicating fairly uniform environmental conditions nationwide.

c. Data Transformation and Standardization

Numerical data was selected for analysis. Transformation was necessary to standardize the scale of variables, especially when there were differences in units between variables such as population size and environmental quality index. Standardization was performed using the Standard Scaler method, which involves transforming the data to have a mean of zero and a standard deviation of one so that all variables contribute equally in the clustering process.

3.3. Determining the Number of Clusters (k)

a. Elbow Method

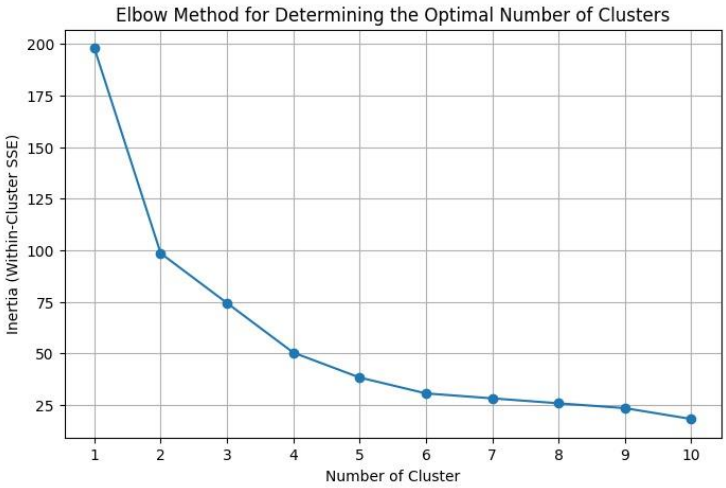

Fig. 2: Elbow Method

Based on the Elbow graph, there is a sharp decline from k=1 to k=2, and it begins to level off after k=4. This indicates that the "elbow" point is around k = 4, because after that the decline in inertia is insignificant. Therefore, we can conclude that the optimal number of clusters is k = 4.

b. Silhoutte Score

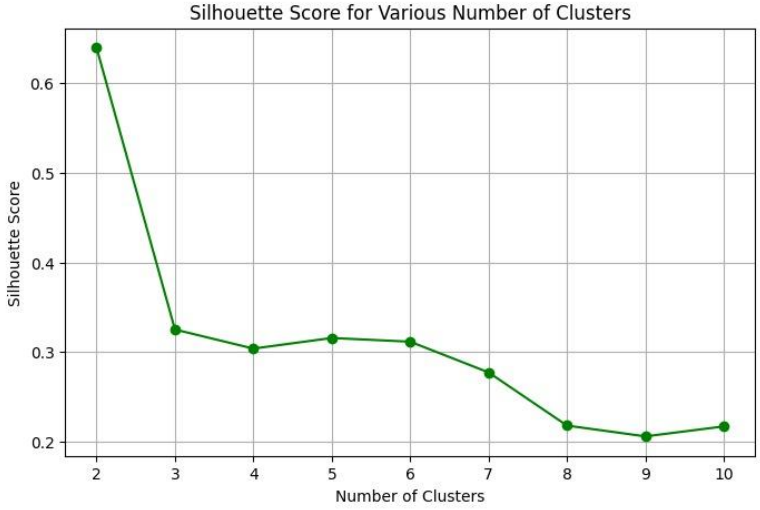

Fig. 3: Silhouette Score for Determining Various Cluster Numbers

Figure 3 clearly illustrates that optimal clustering quality is achieved when the number of clusters (k) is 2, as indicated by the highest Silhouette Score value at that point. This indicates that the data is most effectively divided and the clusters formed are most clearly separated when there are only two groups. The consistent decrease in the Silhouette Score after k=2 shows that increasing the number of clusters actually reduces internal cohesion and/or external separation between clusters, thereby lowering the overall quality of the clustering results.

Table 3: Silhouette Score Results

Cluster	2	3	4	5	6	7	8	9	10
Silhouette Score	0.64	0.33	0.30	0.32	0.31	0.28	0.22	0.21	0.22

The highest Silhouette Score was obtained when the number of clusters was k=2 with a score of 0.6395, indicating the best cluster separation quality compared to other numbers of clusters. After k=2, the score tended to decrease significantly, indicating that adding more clusters actually reduced the cohesion and separation between clusters. Although the use of two clusters may be considered sufficiently simple to describe the complexity of data related to pollution and environmental quality, the evaluation results show that the cluster structure at k=2 remains capable of effectively capturing the main differences between provinces. Therefore, considering cluster quality and interpretive simplicity, the optimal number of clusters selected in this study is two clusters.

3.4. K-Means Clustering Results

Table 4: K-Means Clustering Results

Cluster	Provinces in Indonesia		
1	Aceh	DKI Jakarta	South Kalimantan
	North Sumatera	DI Yogyakarta	North Sulawesi
	West Sumatera	Banten	Central Sulawesi
	Riau	Bali	South Sulawesi
	Jambi	West Nusa Tenggara	Southeast Sulawesi
	South Sumatera	East Nusa Tenggara	Gorontalo
	Bengkulu	West Kalimantan	West Sulawesi
	Lampung	Central Kalimantan	Maluku
	Bangka Belitung Islands	East Kalimantan	North Maluku
	Riau Islands	North Kalimantan	West Papua
0	West Java	Central Java	East Java

The clustering process using the K-Means algorithm and two clusters resulted in two groups of provinces in Indonesia based on the indicators analyzed. The first cluster (Cluster 0) consists of three provinces, namely West Java, Central Java, and East Java. The second cluster (Cluster 1) includes the remaining 30 provinces, including Jakarta Special Capital Region, Yogyakarta Special Region, Banten, and provinces in Sumatra, Kalimantan, Sulawesi, Bali, Nusa Tenggara, Maluku, and West Papua. This clustering indicates that provinces on the island of Java are grouped into a separate cluster, while the remaining provinces are distributed across the second cluster.

3.5. Interpretation and Discussion

The clustering results produced two groups of provinces, with West Java, Central Java, and East Java grouped into a separate cluster (Cluster 0), while the other 30 provinces were grouped into Cluster 1. This composition is consistent with the descriptive statistical results, which show significant variation in the variables of population size and pollution levels. The three provinces in Cluster 0 have characteristics of high population density, massive economic activity, and relatively higher pollution levels.

Meanwhile, provinces in Cluster 1 exhibit more diverse characteristics, both in terms of population size, land area, and environmental conditions, with a tendency toward lower density and pollution levels. Although the highest Silhouette Score was obtained at k = 2, the cluster structure that formed still reflects significant differences between regions, particularly between provinces with high urbanization levels and other provinces. These findings can serve as a basis for formulating more targeted environmental policies that align with the characteristics of each region.

4. Conclusion

This study successfully grouped 33 provinces in Indonesia into two main clusters, based on environmental pollution indicators (water, soil, and air) and demographic characteristics (population and land area) using the K-Means algorithm. The clustering results show that West Java, Central Java, and East Java are grouped separately, characterized by high population density and pollution levels. Meanwhile, the remaining 30 provinces fall into the second cluster, exhibiting more diverse characteristics and generally showing lower pollution levels and population density.

Evaluations using the Elbow method, Silhouette Score, and Dunn Index confirmed that the selection of two clusters resulted in an optimal, clear, and representative grouping structure. These findings demonstrate the effectiveness of the K-Means method in identifying hidden patterns in environmental data, making it a strong foundation for the formulation of data-driven environmental management policies.

Acknowledgement

The authors would like to express their sincere gratitude to the Department of Data Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, for the academic support and guidance throughout the completion of this study. Appreciation is also extended to the Central Statistics Agency (BPS) for providing access to the environmental and demographic data that formed the basis of this research. We are especially thankful to our supervisor, colleagues, and reviewers whose insightful feedback significantly improved the quality of this paper. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] B. K. Dewi and L. Fitria, "ANALISIS INDEKS KUALITAS LINGKUNGAN HIDUP (IKLH) DI DKI JAKARTA TAHUN 2019-2021," *Ilmiah Indonesia*, vol. 33, no. 1, pp. 1–12, 2022.

[2] Z. A. Hidayati and Zakianis, "Analisis Faktor-Faktor Yang Mempengaruhi Indeks Kualitas Lingkungan Hidup (IKLH) Di Indonesia Tahun 2017-2019," *Jurnal Medika Utama*, vol. 3, no. 2, p. 2329, 2022, [Online]. Available: http://jurnalmedikahutama.com/index.php/JMH/article/view/456

[3] Y. B. Kondolele and B. Mustari, "Faktor Penentu Kualitas Lingkungan Hidup pada Pusat Populasi Indonesia Yusliaty Bubun Kondolele 1 , Bakhtiar Mustari 2 1," *Ekonomika dan Dinamika Sosial*, vol. 4, pp. 71–93, 2025.

[4] I. Ferdiansyah, B. Huda, and A. Hananto, "Analisis Clustering Menggunakan Metode K-Means Pada Kemiskinan Di Jawa Timur Tahun 2020," vol. 4, pp. 858–869, 2024.

[5] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 293–301, 2023, doi: 10.57152/malcom.v3i2.952.

[6] M. Guntara and N. Lutfi, "Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method," *JuTI "Jurnal Teknologi Informasi,"* vol. 2, no. 1, p. 43, 2023, doi: 10.26798/juti.v2i1.944.

[7] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.

[8] R. A. Sary, N. Satyahadewi, and W. Andani, "Application of K-Means++ With Dunn Index Validation of Grouping West Kalimantan Region Based on Crime Vulnerability," *Barekeng*, vol. 18, no. 4, pp. 2283–2292, 2024, doi: 10.30598/barekengvol18iss4pp2283-2292.