Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26449
Page: 651-661

Regional Clustering of CO_2 Emissions in Indonesia for Emission Policy Targeting

[1] **Sayyidah Ummi Habibah**
Mathematics Department, Universitas Negeri Surabaya, Surabaya, 60231, Indonesia

[2] **A'yunin Sofro**
Actuarial Science Department, Universitas Negeri Surabaya, Surabaya, 60231, Indonesia

Article Info

Article history:

Accepted, 15 November 2025

Keywords:

Clustering;
CO_2 Emissions;
Fuzzy K-Medoids;
K-Medoids.

ABSTRACT

Regional disparities in Indonesia's CO_2 emissions highlight the need for emissions policies tailored to regional conditions rather than uniform national policies. This study addresses this issue by applying clustering analysis to identify emission patterns across five sectors: Energy, IPPU, Agriculture, Forestry, and Waste. K-Medoids and Fuzzy K-Medoids were selected for their robustness to outliers and their ability to capture complex, cross-sectoral emission characteristics more effectively than conventional methods. The results show that the K-Medoids method produced the most reliable clustering, with a Silhouette Coefficient of 0.5981 and a Dunn Index of 0.0310, indicating a moderate cluster structure. Two clusters were identified: provinces with low emissions dominated by the forestry sector, and provinces with high emissions driven by non-forestry activities. These cluster-based patterns provide a practical basis for directing emission policy interventions according to regional characteristics.

Corresponding Author:

A'yunin Sofro,
Actuarial Science Department
Universitas Negeri Surabaya, Surabaya, Indonesia
Email: ayuninsofro@unesa.ac.id

1. INTRODUCTION

Carbon dioxide (CO_2) is a key greenhouse gas that traps solar heat in the atmosphere, helping regulate Earth's temperature. In the absence of these gases, the planet would be too cold to sustain life. However, as their concentration in the atmosphere increases, more heat becomes trapped, causing the planet's temperature to rise [1]. In recent years, the concentration of greenhouse gases has continued to escalate, reaching record levels in 2023 and committing the Earth to extended periods of warming. The World Meteorological Organization (WMO) reports that carbon dioxide levels in the atmosphere have risen by more than 10% over the past twenty years, marking the fastest increase ever recorded [2]. The negative consequences of this phenomenon are increasingly evident. Climate change has caused diverse and widespread impacts on human societies. It has also increased the occurrence and severity of natural disasters—including floods, droughts, and storms—resulting in severe damage to infrastructure, ecosystems, and human lives [3]. Therefore, efforts to mitigate global warming require not only reducing CO_2 emissions but also removing carbon dioxide from the atmosphere to achieve zero or even negative emissions. Such reductions cannot rely on a single solution; rather, they must involve multiple, synergistic strategies that integrate social, economic, environmental, and technological dimensions [4].

According to data from the Global Carbon Budget presented by Our World in Data, Indonesia recorded the highest CO_2 emissions in Southeast Asia in 2023, reaching approximately 733.22 million tons [5].

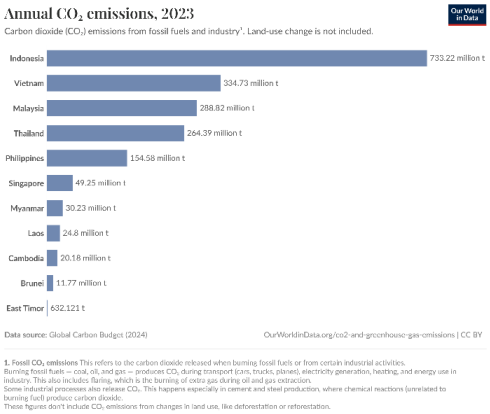

Figure 1. CO_2 emissions in Southeast Asia in 2023

It should be noted that Figure 1 represents only emissions from fossil fuels and industry, excluding those resulting from land-use change. In the broader context of national greenhouse gas accounting, Indonesia's total CO_2 emissions stem from five major sectors are energy, Industrial Processes and Product Use (IPPU), agriculture, forestry, and waste management [6]. Significant disparities exist among Indonesian provinces in terms of the sources of their carbon emissions. The composition and magnitude of consumption-based CO_2 emissions vary considerably across provinces, reflecting differences in industrial activities, energy use, and trade structures. Although the country's total carbon emissions are relatively comparable between production and consumption, their distribution by emission source or destination shows notable regional variation, indicating distinct economic and environmental characteristics across provinces. This spatial heterogeneity underscores the potential to develop mitigation strategies tailored to each region's unique conditions, rather than adopting a uniform national approach [7]. To address this complexity, clustering analysis has been applied to Indonesia's emission dataset. Clustering enables grouping regions based on similarities in emission characteristics, providing a more comprehensive understanding of regional emission patterns [8].

Several previous studies have applied clustering techniques to analyze regional emissions; most remain limited in terms of both sector coverage and methodological approach. For example, research by Zhang and Yang only focused on emissions from agricultural activities [9]. In addition, Siahaan's research classified provinces in Indonesia solely on per capita energy consumption, thereby failing to reflect the complexity of emissions from various sectors [10]. Other studies have also applied clustering techniques, but only rely on the K-Means algorithm [11]. Consequently, existing research has not been able to capture the complex and cross-sectoral nature of regional emissions. This study addresses this gap by developing a more comprehensive clustering analysis that integrates emissions from various sectors using a flexible and robust methodological approach.

Indonesia's emissions profile shows significant regional disparities due to diverse economic activities and different land-use patterns. Consequently, formulating region-specific and data-driven policies has become essential to achieving the country's emission reduction targets. This study aims to answer how the clustering of CO_2 emissions in various regions of Indonesia can provide insights for developing targeted and effective emission reduction strategies. To achieve this objective and ensure methodological rigor, this study provides a comprehensive and multivariate approach to analyzing regional emissions by integrating five key sectors—Energy, IPPU, Agriculture, Forestry, and Waste—into a unified cluster analysis framework. Methodologically, this study applies methods that are more robust and flexible than K-Means, namely K-Medoids and Fuzzy K-Medoids. These methods offer robustness to outliers and flexibility for overlapping data, allowing more reliable representation of regional emission patterns [12]. Once regional emission patterns are identified, the clustering results group provinces with similar characteristics, providing a practical basis for more targeted emissions policy interventions.

2. RESEARCH METHOD

2.1 CO_2 Emissions

Carbon dioxide emissions are the release of CO_2 gas into the atmosphere, mainly caused by human activities such as the combustion of carbon-based materials, including biomass. These emissions play a major role in contributing to global climate warming. In Indonesia, CO_2 emission sources are classified into five main sectors, namely [13]:

Energy

The energy sector covers emissions generated from the combustion of fossil fuels in power generation, transportation, and other energy uses.

Industrial Processes and Product Use (IPPU)

The industrial sector covers emissions from industrial processes and product use that do not involve the combustion of fuels.

Agriculture

The agriculture sector covers emissions from agricultural activities such as land management, enteric fermentation, and fertilizer use.

Forestry

The forestry covers emissions and carbon sequestration resulting from changes in forest cover, deforestation, and land management.

Waste

The waste sector covers emissions generated from solid and liquid waste management processes, both domestic and industrial.

2.2 Data Preprocessing

The first step is data preprocessing, which included data cleaning and standardization. Data cleaning was performed to address missing values in order to maintain the quality of the dataset. Next, data standardization is performed, which is the process of changing the values in the dataset so that they have a uniform scale and format. The purpose of standardization is to prevent variables with different value ranges from dominating the analysis, especially in methods that are sensitive to scale differences. The technique applied is Min-Max Scaling, which converts the original vale into range between 0 and 1. The Min-Max Scaling formula is written as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

X shows the initial data value, X_{min} is the smallest value in the column, while X_{max} is the largest value in the same column [14].

2.3 Elbow Method

The Elbow Method is the oldest method for identifying the ideal number of clusters (k) in a dataset. The basic concept involves setting k = 2 as the initial candidate for the ideal number of clusters, then gradually increasing k by 1 until reaching a predetermined maximum value for the potential optimal estimates. For each k, the total within-cluster sum of squared distances (WSS) is calculated, which measures how closely data points in a cluster are grouped around the cluster center. The optimal number of clusters is typically chosen as the k value where the decrease in WSS begins to slow down, forming an "elbow" in the WSS plot [15].

Mathematically, let WSS_j represent the total squared distances within the j-th cluster, and let n be the total number of clusters. Then, the overall WSS for n clusters is defined as:

$$WSS(n) = \sum_{j=1}^{n} (WSS_j) \tag{2}$$

This formula sums the WSS of all clusters, allowing the selection of the k value that best balances cluster compactness and simplicity [16]

2.4 K-Medoids

The K-Medoids method is also called the Partitioning Around Medoids (PAM) algorithm, works by selecting one object from each cluster as a medoid. This medoid is the point with the smallest total distance to all other points within the same cluster, thus representing that cluster best [17]. Compared to K-Means, K-Medoids is more resistant to noise and outliers because it uses actual data point as cluster centers [18]. In addition, the use of actual data objects as medoids contributes to more stable clustering performance, particularly when the dataset contains non-normal distributions or features with varying scales. Although this robustness offers clearer interpretability—especially in applications such as document clustering—the algorithm is generally more computationally demanding due to the pairwise distance calculations required for medoid selection [19].

To determine the optimal medoid, it is necessary to measure the distance between objects in the cluster. The distance between objects i and j is calculated using a measure of dissimilarity. Although various dissimilarity measures can be used, Euclidean distance is often chosen because of its simplicity, intuitive interpretation, and

ability to be applied to various types of data, especially numerical data and continuous variables. Therefore, Euclidean distance is very popular in distance-based clustering methods such as K-Means, hierarchical clustering, and other similar algorithms [20]. The Euclidean distance is formulated as follows [21]:

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})^2} \tag{3}$$

In Equation (3), d_{ij} denotes the measure of separation between objects i and j, x_{ia} represents the attribute value of object i on variable a, x_{ja} indicates the attribute value of object j on variable a, and p signifies the total count of variables under observation.

The K-Medoids algorithm operates according to the following procedure [22]:

Input:

Desired count of clusters (k)

A dataset consisting of n objects

Output: k clusters, where the total dissimilarity between each object and its nearest medoid is minimized.

Algorithm steps:

1. Initialization: Randomly choose k data point from the data as the first medoid.
2. Assignment: For this step, assign each remaining data points to the closest medoid using the chosen distance metric.
3. Updating:
   1) Randomly pick an object that isn't a medoid.
   2) Swap the selected object with the current medoid.
   3) Calculate the total cost (difference) of this new configuration.
   4) Select the set of medoids that gives the lowest total cost for the next iteration.
4. Termination: End the process when the stopping criteria are satisfied (until the medoid remains unchanged); if not, go back to Step 2 and repeat the steps.

2.5 Fuzzy K-Medoids

Fuzzy K-Medoids is a clustering method that groups data into clusters based on a distance criterion, computed from the cluster centers derived from the data values. The primary difference between the Fuzzy K-Medoids and FCM (Fuzzy C-Means) algorithms lies in how the cluster centers are determined. In the FCM approach, the central point of the cluster may be positioned anywhere within the domain of discussion (U), whereas in the Fuzzy K-Medoids method, it corresponds to an actual data point, the medoid [23]. A medoid refers to an object that represents the center point of a cluster [24].

The Fuzzy K-Medoids algorithm assesses the distance criterion by computing the dataset's cluster centers. In this approach, the updated membership matrix M_u is initially generated using the FCM procedure to determine the medoid. Subsequently, within each cluster, the data point corresponding to the highest membership value is selected as the medoid. The objective function of the Fuzzy K-Medoids method is expressed in (4) [25].

$$P_t = \sum_{i=1}^{m} \sum_{k=1}^{n} (d^2(c_k, x_i)(\delta_{ik})^r) \tag{4}$$

Where P_t denotes the objective function at the t-th iteration, δ_{ik} represents the membership degree in the matrix M_u, r is the fuzziness parameter ($r \geq 2$), and $d^2(c_k, x_i)$ indicates the distance between the i-th data point and the k-th cluster center.

The preliminary membership degree δ_{ik} within the M_u matrix is computed using Equations (5) and (6) within the Fuzzy K-Medoids framework for determining cluster centers.

$$M_U = [\delta_{ik}]_{m \times n}, \sum_{k=1}^{c} \delta_{ik} = 1, 1 \leq i \leq n \tag{5}$$

$$\delta_{ik} = [0,1], i = 1,2,\ldots,m;\ k = 1,2,3,4,\ldots,c \tag{6}$$

During each iteration, the membership matrix δ_{ik} is recalculated and updated according to Equation (7).

$$\delta_{ik} = \frac{[d^2(c_k, x_i)]^{\frac{-1}{r-1}}}{\sum_{j=1}^{c} [d^2(c_k, x_i)]^{\frac{-1}{r-1}}} \tag{7}$$

Once the M_u membership matrix is obtained, the cluster center is determined using Equation (8).

$$C_k = \frac{\sum_{i=1}^{m}(\delta_{ik})^r y_i}{\sum_{i=1}^{m}(\delta_{ik})^r} \tag{8}$$

In addition, Fuzzy K-Medoids offers greater accuracy in handling overlapping data points. Unlike Fuzzy K-Means, which defines cluster prototypes as synthetic objects derived from weighted-average calculations, Fuzzy K-Medoids selects a set of objects that actually exist in the data as cluster prototypes (medoids). Moreover, this method also introduces an extra cluster, referred to as a noise cluster, to accommodate outlier objects with high membership degrees that fall outside the k main clusters [12].

2.6 Silhouette Coefficient

The silhouette coefficient is an internal validity metric used to evaluate the quality of clustering. This metric considers both the intra-cluster distance (compactness) and the inter-cluster distance (seprataion) [26]. The calculation begins by determining the average distance between a data point x_i^j and all other points within the same cluster j. This quantity is defined as [21]:

$$a_i^j = \frac{1}{m_j - 1}\sum_{\substack{r=1 \\ r\neq i}}^{m_j} d(x_i^j, x_r^j) \tag{9}$$

Where, j denotes the cluster index, i is the data index $(i = 1,2,\ldots,m_j)$ indicates the data point within cluster j, a_i^j denotes the mean distance between the i-th data point and all other points within the same cluster, m_j is the total number of points in the cluster j, and $d(x_i^j, x_r^j)$ is the distance between the i-th and the r-th data points in the same cluster.

Next, the inter-cluster distance is computed by identifying the lowest mean distance between x_i^j and all points belonging to any other cluster $n \neq j$. This written as:

$$b_i^j = \min_{\substack{n=1,\ldots,k \\ n\neq j}} \left(\frac{1}{m_n}\sum_{r=1}^{m_n} d(x_i^j, x_r^n) \right) \tag{10}$$

Where, k denotes the total number of clusters, m_n is the number of data points in cluster n, and $d(x_i^j, x_r^n)$ is the distance between the data point x_i^j in cluster j and the point x_r^n in cluster n.

After obtaining a_i^j and b_i^j, the Silhouette Index (SI) for each data point is computed as:

$$SI_i^j = \frac{b_i^j - a_i^j}{max\{a_i^j, b_i^j\}} \tag{11}$$

A higher value of SI_i^j approaching 1 indicates that a data point fits well within its assigned cluster and has minimal similarity to points in other clusters. In contrast, values close to -1 reflect poor clustering performance, which may imply overlapping clusters or data points placed in the wrong cluster. The overall Silhouette score for cluster j is calculated by taking the average of all Silhouette values for the data points within that cluster. The interpretation of these Silhouette scores follows the classification presented in Table 1 [27].

Table 1. Silhouette Coefficient

Silhouette Coefficient	Category
0.71 – 1.00	Strong
0.51 – 0.70	Moderate
0.26 – 0.50	Weak
0.00 – 0.25	Bad

2.7 Dunn Index

The Dunn Index is a metric used to assess the quality of clustering results by maximizing the distance between clusters (intercluster) and minimizing the distance within clusters (intracluster). This index combines the aspects of cohesion and separation in cluster evaluation. Cohesion describes how tightly objects are grouped within a cluster, which is measured by the cluster diameter, while separation indicates how far apart two different

clusters are. To calculate cohesion, the distance between pairs of objects within a cluster is measured, and the maximum distance value is taken as the cluster diameter. The formula used in the Dunn Index calculation can be expressed as follows:

The cluster diameter is calculated using the following formula:

$$diam(C_k) = \max_{x,y \in C_k} \{d_{xy}\} \tag{12}$$

This formula shows that the diameter of a cluster C_k is determined by the maximum distance between two objects x and y within that cluster.

Meanwhile, to calculate the distance between two clusters, the following equation is used:

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d_{xy}\} \tag{13}$$

This means that the distance between two clusters c_i and c_j is determined by the smallest distance between pairs of objects from both clusters.

From the two formulas, the following Dunn Index (DI) equation is obtained [28]:

$$DI = \min_{i=1,\dots,N} \left\{ \min_{j=i+1,\dots,N} \left(\frac{d(c_i, c_j)}{\max_{k=1,\dots,N} \{diam(C_k)\}} \right) \right\} \tag{14}$$

Where DI is the Dunn Index value used to assess the quality of data clustering results. The value $d(c_i, c_j)$ indicates the distance between clusters i and j, while N is the total number of clusters. The symbols x and y represent objects in clusters c_i and c_j, respectively. Meanwhile, $diam(C_k)$ indicates the diameter of cluster k, which is a measure that reflects how compact a cluster is. Overall, these components are used to calculate the Dunn Index as a measure of cluster separation and compactness.

2.8 Mann-Whitney Test

The Mann-Whitney test is a nonparametric method used to test differences between two independent groups with ordinal, interval, or ratio scale data, especially when the data does not meet the assumption of normal distribution. Although it is generally considered to test the difference in medians between groups, some experts argue that this test can also be used to compare means. This test is also often referred to as the Wilcoxon Rank Sum Test [29].

The steps in performing the Mann-Whitney test are as follows [30]:

1) State the research hypothesis and determine the significance level (α).
2) Rank the data as a whole without regard to sample groups or categories.
3) Sum the ranks for each group and calculate the U statistics using the formula [31]:
4)

$$U_1 = n_{X_1} n_{X_2} + \frac{n_{X_2}(n_{X_2} + 1)}{2} - R_{min} \tag{15}$$

$$U_2 = n_{X_1} n_{X_2} - U_1 \tag{16}$$

Where n_{X_1} denotes the number of subjects in variable X_1 while n_{X_2} is the number of subjects in variable X_2. Furthermore, R_{min} refers to the smallest rank given to each group. These values are used to calculate the U statistic, which then becomes the basis for drawing conclusions from the Mann-Whitney test.

5) Draw statistical conclusions about the null hypothesis based on the obtained U value.

2.9 Data Source

This study uses a quantitative research approach and secondary data obtained from the National Greenhouse Gas Inventory System (SIGN-SMART KLHK), managed by the Ministry of Environment and Forestry [32]. The data covers 2019-2023 and include records from 38 provinces in Indonesia. The emission data represent total annual greenhouse gas emissions, expressed in tons of carbon dioxide equivalent (CO_2-eq) for each province. Emissions are estimated through an inventory-based approach using regional activity data and national emission factors, following the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. The analyzed variables are Energy, Industrial Processes and Product Use (IPPU), Agriculture, Forestry, and Waste.

The entire data processing and analysis were conducted using RStudio, a statistical computing and visualization software environment. To begin the research process, data preprocessing was conducted, which included data cleaning and data standardization. After preprocessing, the ideal number of clusters was identified

using the Elbow method. Subsequently, cluster analysis was performed using the K-Medoids and Fuzzy K-Medoids methods. Finally, the quality of the resulting clusters was validated through the Silhouette Coefficient to assess their accuracy and determine the most appropriate clustering method. The following section explains the results and analysis in detail.

3. RESULT AND ANALYSIS

The preliminary phase of this study involves data collection and preprocessing. The preprocessing stage includes data cleaning to address missing values and data standardization to ensure uniform scales across variables. Next, use the Elbow Method criterion to identify the most suitable number of clusters. Figure 2 below presents the results of the cluster number identification.

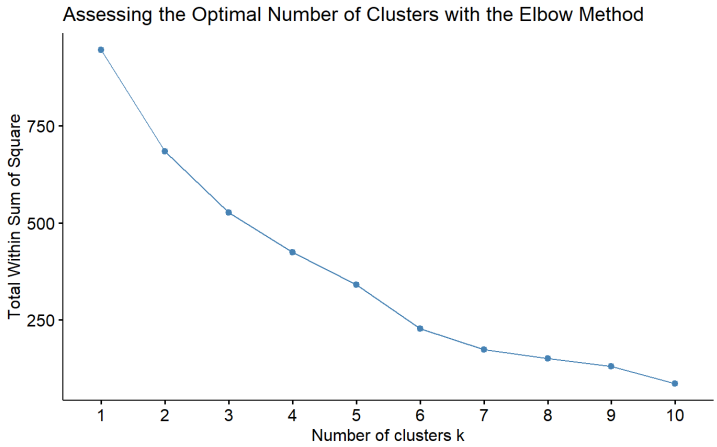

Figure 2. Elbow Method Plot

According to the Elbow method, this study's data is best grouped into two clusters. Thus, the clustering results obtained with the K-Medoids algorithm using the ideal number of clusters are shown in Table 2 below.

Table 2. Result K-Medoids Clustering

Cluster	Provinces	Number of members
Cluster 1	Aceh, Sumatera Utara, Sumatera Barat, Riau, Kepulauan Riau, Jambi, Sumatera Selatan, Bengkulu, Kepulauan Bangka Belitung, Banten, DKI Jakarta, DI Yogyakarta, Bali, Nusa Tenggara Barat, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Gorontalo, Sulawesi Tengah, Sulawesi Barat, Sulawesi Tenggara, Maluku Utara, Maluku, Papua Barat Daya, Papua Barat, Papua Tengah, Papua Pegunungan, Papua Selatan, Papua	31
Cluster 2	Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, Kalimantan Selatan, Nusa Tenggara Timur, Sulawesi Selatan	7

To understand the characteristics and differences among clusters, the average values of variables for each cluster are calculated as follows.

Table 3. Average Value of the Variables using K-Medoids

	Cluster 1	Cluster 2
Energy	-0.191	1.018
IPPU	-0.229	1.219
Agriculture	-0.328	1.749
Forestry	0.074	-0.392
Waste	-0.236	1.261

The table shows that Cluster 1 consists of provinces with lower emissions in Energy, IPPU, Agriculture, and Waste, and slightly higher values in Forestry, indicating that overall emissions are low and dominated by the Forestry sector. In contrast, Cluster 2 comprises provinces with significantly higher emissions in Energy, IPPU, Agriculture, and Waste, but lower Forestry values, highlighting areas with high non-forestry emissions. Furthermore, the Mann-Whitney test results indicate that the differences between clusters are statistically significant across all sectors ($p < 0.05$), confirming that the clustering effectively distinguishes provinces based on their emission characteristic.

The second method used is Fuzzy K-Medoids clustering. This method represents a fuzzy clustering approach that distinguishes itself from conventional clustering methods. The key distinction of Fuzzy K-Medoids is that each data object is assigned a membership degree before the clustering process. The outcomes of Fuzzy K-Medoids clustering are presented in the table below.

Table 4. Result Fuzzy K-Medoids Clustering

Cluster	Provinces	Number of members
Cluster 1	Aceh, Sumatera Barat, Jambi, Bengkulu, Kepulauan Riau, Kepulauan Bangka Belitung, Banten, DKI Jakarta, DI Yogyakarta, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Gorontalo, Sulawesi Tengah, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tenggara, Maluku Utara, Maluku, Papua Barat Daya, Papua Barat, Papua Tengah, Papua Pegunungan, Papua Selatan, Papua	29
Cluster 2	Sumatera Utara, Riau, Lampung, Sumatera Selatan, Jawa Barat, Jawa Tengah, Jawa Timur, Kalimantan Tengah, Kalimantan Selatan	9

Then, the average results for the variables between clusters are as follows.

Table 5. Average Value of the Variables using Fuzzy K-Medoids

	Cluster 1	Cluster 2
Energy	-0.237	0.834
IPPU	-0.182	0.640
Agriculture	-0.129	0.456
Forestry	-0.271	0.955
Waste	-0.357	1.257

The table above shows the standard average values for variables in each cluster. Cluster 1 shows negative values for all variables, indicating provinces with low levels of activity or emissions in all sectors. Conversely, Cluster 2 shows positive values for all variables, indicating provinces with high levels of activity or emissions in all sectors. Furthermore, the Mann-Whitney test results indicate that the differences between clusters are statistically significant across all sectors ($p < 0.05$), confirming that the clustering effectively distinguishes provinces based on their emission characteristic.

After the clustering procedure with K-Medoids and Fuzzy K-Medoids is complete, the clustering results need to be evaluated to assess their quality and accuracy. In this study, the evaluation is conducted using the Silhouette Coefficient and the Dunn Index. Table 6 below shows the evaluation results for both methods using the Silhouette Coefficient and the Dunn Index.

Table 6. Result Silhouette Coefficient

Method	Silhouette Coefficient	Dunn Index	Category
K-Medoids	0.5981	0.03104339	Moderate
Fuzzy K-Medoids	0.5217	0.007688153	Moderate

From Table 6, it can be seen that clustering with the K-Medoids method yields a higher Silhouette Coefficient and Dunn Index than clustering with the Fuzzy K-Medoids method. Therefore, it can be said that clustering results using the K-Medoids method are more accurate than those using Fuzzy K-Medoids. Accordingly, it can be inferred that the K-Medoids clustering method achieves the highest accuracy in grouping CO_2 emission data in Indonesia from 2019 to 2023. Furthermore, a cross-year validation was conducted to assess the robustness of the clustering results, where the model trained on data from 2019-2022 was tested using data from 2023. The validation produced a consistency value of 0.9736842 (97%), indicating that most provinces remained in the same cluster across years. In total, 37 of 38 provinces maintained the same cluster assignment, with only Kalimantan Selatan showing shift from Cluster 1 to Cluster 2, likely due to an increase in its non-forestry emission components in 2023. This high level of consistency demonstrates that the K-Medoids method generates stable and generalizable clustering patterns over time.

The results of clustering using the K-Medoids method show that Cluster 1 represents regions with low emission levels and a dominance of the forestry sector. This pattern is supported by the characteristics of several provinces included in this cluster, such as Central Kalimantan and Papua, both of which have extensive forest cover and a significant role for the forestry sector in carbon sequestration. For example, Central Kalimantan has 12.27 million hectares of forest cover, or about 80.10% of its total area, indicating great potential for reducing

emissions through carbon sequestration [33]. Meanwhile, Papua has primary forest cover reaching 61.85% of its total area, making it one of the regions with the most extensive and sustainable tropical forest ecosystems in Indonesia [34]. This condition reinforces the interpretation that the provinces in Cluster 1 function as low-emission regions that ecologically serve as carbon sinks, with a strong dominance of the forestry sector.

On the contrary, Cluster 2 describes areas with high emission levels, one of which is characterized by intensive economic and industrial activity in various sectors. This pattern is in line with the characteristics of provinces such as East Java, Central Java, and West Java, which are recorded as having the largest number of manufacturing industries in Indonesia. Based on data from the Central Statistics Agency in 2023 [35], the number of medium and large manufacturing companies reached 977,471 in East Java, followed by 862,926 in Central Java, and 641,639 in West Java. The high concentration of industrial activity in these three provinces contributes significantly to increased emissions, particularly from the energy and industrial process sector (IPPU), thus supporting the characteristics of Cluster 2 as a region with high emissions in all sectors. The following is a map of the K-Medoids cluster analysis results.

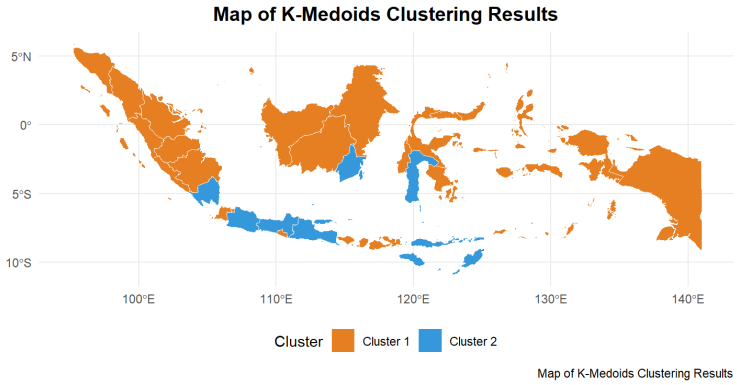

Figure 3. Map of K-Medoids Clustering Results

The map above shows the results of K-Medoids clustering generated in RStudio, with Cluster 1 marked in orange and Cluster 2 in blue. Cluster 1, which include provinces such as Central Kalimantan and Papua, generally shows lower average emissions. On the other hand, Cluster 2, which consists mainly of provinces and regions in Java such as East Java, Central Java, and West Java, shows higher emissions. This visualization highlights regional variations in CO_2 emission patterns across Indonesia.

Based on these findings, policy recommendations can be tailored to the characteristics of each cluster. Cluster 1 regions can prioritize forest protection and enhancement of carbon sequestration, reflecting their emission profiles that are strongly influenced by land-use and forestry dynamics. In contrast, Cluster 2 regions require mitigation strategies focused on non-forestry sectors, including energy efficiency improvements, cleaner industrial operations, and a gradual transition toward renewable energy. Aligning policies with these cluster-specific characteristics enables more precise targeting of emission reduction efforts across provinces.

4. CONCLUSION

According to the result of CO_2 emission cluster analysis in Indonesia during the 2019–2023 period, the K-Medoids method produced the most ideal grouping, with a Silhouette Coefficient of 0.5981 and a Dunn Index of 0.03104339, which is considered moderate structure. The clustering results show two distinct regional characteristics. Cluster 1 consists of provinces with relatively low emission levels, dominated by the forestry sector. In contrast, Cluster 2 includes provinces with high emission levels, influenced by non-forestry sectors such as energy, industry, and waste. These results are expected to serve as a basis for the government in establishing more targeted emission control policies tailored to each region's characteristics.

Future work could link cluster types to policy effectiveness or disaster vulnerability indices to provide deeper insights into the practical implications of emission patterns. Moreover, future research could be expanded by conducting multi-year trend analyses or integrating regional socioeconomic indicators to better understand the underlying factors that drive emission disparities across provinces.

5. REFERENCES

[1] "Carbon Dioxide - Earth Indicator," National Aeronautics and Space Administration (NASA). Accessed:

Oct. 21, 2025. [Online]. Available: https://science.nasa.gov/earth/explore/earth-indicators/carbon-dioxide/
[2] "Greenhouse gas concentrations surge again to new record in 2023," World Meteorogical Organization. Accessed: Oct. 21, 2025. [Online]. Available: https://wmo.int/news/media-centre/greenhouse-gas-concentrations-surge-again-new-record-2023
[3] L. J. R. Nunes, "The Rising Threat of Atmospheric CO2: A Review on the Causes, Impacts, and Mitigation Strategies," *Environ. - MDPI*, vol. 10, no. 4, 2023, doi: 10.3390/environments10040066.
[4] N. V. Lobus, M. A. Knyazeva, A. F. Popova, and M. S. Kulikovskiy, "Carbon Footprint Reduction and Climate Change Mitigation :," *J. Carbon Res.*, vol. 9, no. 4, p. 120, 2023, doi: 10.3390/c9040120.
[5] H. Ritchie and M. Roser, "CO_2 emissions," Our World in Data. Accessed: Oct. 21, 2025. [Online]. Available: https://ourworldindata.org/co2-emissions
[6] "Third Biennial Update Report Under the United Nations Framework Convention on Climate Change," *Minist. Environ. For.*, 2021.
[7] I. A. Rum, A. Tukker, R. Hoekstra, A. de Koning, and A. A. Yusuf, "Exploring carbon footprints and carbon intensities of Indonesian provinces in a domestic and global context," *Front. Environ. Sci.*, vol. 12, no. October, pp. 1–15, 2024, doi: 10.3389/fenvs.2024.1325089.
[8] T. R. Noviandy *et al.*, "Environmental and Economic Clustering of Indonesian Provinces: Insights from K-Means Analysis," *Leuser J. Environ. Stud.*, vol. 2, no. 1, pp. 41–51, 2024, doi: 10.60084/ljes.v2i1.181.
[9] Z. Zaijie and Y. Weifang, "Agricultural Carbon Emissions in Hubei Province and County-level Carbon Emission Research," *J. Yunnan Agric. Univ. Sci.*, vol. 17, no. 2, pp. 134–140, 2023, doi: 10.12371/j.ynau(s).202209104.
[10] A. T. A. A. Siahaan, "Clustering Indonesian Provinces Based on Per Capita Energy Consumption Using the K-Means Algorithm," *IJICS (International J. Informatics Comput. Sci.*, vol. 9, no. 1, pp. 1–6, 2025, doi: 10.30865/ijics.v9i1.8875.
[11] William, L. Bayuaji, N. J. Perdana, and T. Handhayani, "Mapping Indonesia's Regions Based on Carbon Emissions Using the K-Means Algorithm," *ICoCSETI 2025 - Int. Conf. Comput. Sci. Eng. Technol. Innov. Proceeding*, pp. 200–205, 2025, doi: 10.1109/ICoCSETI63724.2025.11020099.
[12] M. M. Madbouly, S. M. Darwish, N. A. Bagi, and M. A. Osman, "Clustering Big Data Based on Distributed Fuzzy K-Medoids: An Application to Geospatial Informatics," *IEEE Access*, vol. 10, pp. 20926–20936, 2022, doi: 10.1109/ACCESS.2022.3149548.
[13] "Laporan Inventarisasi Gas Rumah Kaca (GRK) dan Monitoring, Pelaporan, Verifikasi (MPV) Tahun 2024," *Kementeri. Lingkung. Hidup Dan Kehutan.*, vol. 10, 2024.
[14] A. L. Firmansyah, B. I. N, and Z. Arif, "Optimasi K-Means Clustering Pada Data Harga Mangga Menggunakan Particle Swarm Optimization," *J. Teknol. Sist. Inf.*, vol. 6, no. September, pp. 245–259, 2025, doi: 10.35957/jtsi.v6i2.13158.
[15] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.
[16] A. Jimoh Jacob, J. Daniel, and C. Samuel Aneke, "Determination Of Optimal Number Of Clusters Using Gap Statistics And Elbow Methods," *Int. Multiling. J. Sci. Technol.*, vol. 9, no. 3, pp. 7361–7366, 2024, [Online]. Available: www.imjst.or
[17] H. Řezanková, "Different approaches to the silhouette coefficient calculation in cluster evaluation," *21st Int. Sci. Conf. AMSE*, no. September, pp. 1–10, 2018.
[18] Annisa Nadaa Shabrina, M. Afdal, and Siti Monalisa, "Comparison Of K-Means, K-Medoids, and Fuzzy C-Means Algorithms for Clustering Drug User's Addiction Levels," *J. Sist. Cerdas*, vol. 6, no. 2, pp. 113–122, 2023, doi: 10.37396/jsc.v6i2.313.
[19] A. P. Putra, J. Tshivana, and E. Rilvani, "PERBANDINGAN TEORITIS DAN EKSPERIMEN ALGORITMA K-MEANS DAN K-MEDOIDS DALAM KLASTERISASI DATA," *Kohesi J. Multidisiplin Saintek*, vol. 10, no. 2, pp. 1–24, 2025.
[20] Z. Shapcott, "An Investigation into Distance Measures in Cluster Analysis," no. April, pp. 1–38, 2024.
[21] A. Sofro, R. A. Riani, K. N. Khikmah, R. W. Romadhonia, and D. Ariyanto, "Analysis of Rainfall in Indonesia Using a Time Series-Based Clustering Approach," *Barekeng*, vol. 18, no. 2, pp. 837–848, 2024, doi: 10.30598/barekengvol18iss2pp0837-0848.
[22] 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, no. April, p. 100034, 2022, doi: 10.1016/j.jcmds.2022.100034.
[23] D. A. Dewi, S. Surono, R. Thinakaran, and A. Nurraihan, "Hybrid Fuzzy K-Medoids and Cat and Mouse-Based Optimizer for Markov Weighted Fuzzy Time Series," *Symmetry (Basel).*, vol. 15, no. 8, 2023, doi: 10.3390/sym15081477.
[24] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "METODE k-MEDOIDS CLUSTERING DENGAN VALIDASI SILHOUETTE INDEX DAN C-INDEX," *J. GAUSSIAN*, vol. 8, pp. 161–170, 2019.
[25] H. K. Sivaraman and R. Leburu, "Energy-efficient clustering and routing using fuzzy k-medoids and

adaptive ranking-based wireless sensor network," *Int. J. Reconfigurable Embed. Syst.*, vol. 13, no. 3, pp. 774–785, 2024, doi: 10.11591/ijres.v13.i3.pp774-785.

[26] D. T. Dinh, T. Fujinami, and V. N. Huynh, "Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient," *Commun. Comput. Inf. Sci.*, vol. 1103 CCIS, pp. 1–17, 2019, doi: 10.1007/978-981-15-1209-4_1.

[27] F. P. Purba, K. Roder, E. Simamora, and H. Nasution, "Implementation of Fuzzy C-Means (FCM) and Fuzzy Possibilistic C-Means (FPCM) for Clustering District/City Based on Health Services and Infectious Diseases in North Sumatera," *ZERO J. Sains, Mat. dan Terap.*, vol. 8, no. 2, p. 47, 2025, doi: 10.30829/zero.v8i2.23480.

[28] Hidayatullah, S. Martha, and S. Aprizkiyandari, "ANALISIS K-MEANS MENGGUNAKAN METODE DUNN INDEX DALAM MENENTUKAN JUMLAH CLUSTER OPTIMAL (Studi Kasus: Indikator Pendidikan SMA di Indonesia Tahun 2022)," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 13, no. 3, pp. 303–310, 2024.

[29] D. Selphia, M. Fathurrahman, M. Susilawati, N. Pratiwi, and R. Purnami, "PENERAPAN UJI MANN-WHITNEY DALAM PERBANDINGAN PRESTASI AKADEMIK MAHASISWA STATISTIKA UNIVERSITAS," *J. Eksbar*, vol. 2, no. 1, pp. 19–28, 2024.

[30] T. Sriwidadi, "PENGGUNAAN UJI MANN-WHITNEY PADA ANALISIS PENGARUH PELATIHAN WIRANIAGA DALAM PENJUALAN PRODUK BARU," *BINUS Bus. Rev.*, vol. 2, no. 2, pp. 751–762, 2011.

[31] A. Damanhuri and A. Solikin, "IMPLEMENTASI UJI MANN-WHITNEY DALAM EVALUASI PRESTASI HASIL BELAJAR DALAM KEGIATAN PELATIHAN SAILS-UINSA DI FAKULTAS SYARIAH DAN HUKUM UINSA," *Didakt. J. Pendidik. dan Ilmu Pengetah.*, vol. 23, no. 1, pp. 40–47, 2023.

[32] "Sistem Inventarisasi Gas Rumah Kaca Nasional-Sederhana, Mudah, Akurat, Ringkas, dan Transparan," Kementerian Lingkungan Hidup Dan Kehutanan. Accessed: Oct. 11, 2025. [Online]. Available: https://signsmart.menlhk.go.id/

[33] A. K. A. Tandir, P. Hergianasari, and S. S. Hadiwijoyo, "KEMITRAAN MULTI PIHAK DALAM PELESTARIAN EKOSISTEM HUTAN DI KALIMANTAN TENGAH TAHUN 2016-2020," vol. 4, no. 4, pp. 2059–2072, 2024.

[34] Bappeda Provinsi Papua, "Analisis Kerangka Pembangunan Provinsi Papua 2021," *Pemerintah Drh. Provinsi Papua*, pp. 1–186, 2022.

[35] "Jumlah Perusahaan Industri Skala Mikro dan Kecil Menurut Provinsi (Unit), 2023," Badan Pusat Statistik. Accessed: Oct. 25, 2025. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NDQwIzI=/jumlah-perusahaan-industri-skala-mikro-dan-kecil-menurut-provinsi.html