SAINSMAT: Journal of Applied Sciences, Mathematics, and Its Education *ISSN: 2776-3641 (online)*
Vol. 12 No. 1 (2023)
https://doi.org/10.35877/sainsmat1775

Application of K-Medoids Algorithm in Provincial Grouping in Indonesia Based On Case of Environmental Pollution

Muh. Hizbul Zainul Muttaqim, Ruliana, & Zulkifli Rais*

Department of Statistics, Universitas Negeri Makassar, Makassar, 90223, Indonesia

Abstract

Cluster analysis is a method for grouping objects that have the same characteristics. One of the methods in cluster analysis used to group data is the K-Medoids method. In this study the K-Medoids method was applied to classify provinces in Indonesia based on environmental pollution. The variables used are: the number of sub-districts/villages that experience water pollution from factory waste, the number of sub-districts/villages that experience water pollution from household waste, the number of sub-districts/villages that experience soil pollution from factory waste, the number of sub-districts/villages that experience soil pollution from household waste, the number of sub-districts/villages that experience air pollution from factory waste and the number of sub-districts/villages that experience air pollution from household waste. Based on the Davies Bouldin Index, the 2 best clusters were obtained where the first cluster consisted of 31 provinces which had low environmental pollution and the second cluster consisted of 3 provinces which had high environmental pollution.

Keywords: Cluster analysis, K-medoids, environmental pollution

Received: 16 April 2022 Revised: 21 August 2023 Accepted: 13 January 2023

1. Introduction

Data mining is the process of finding correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques (Larose, 2014). Data mining can be divided into 6 groups, namely description, estimation, prediction, classification, clustering, and association (Kusrini & Luthfi, 2009). Clustering analysis is an analysis used to separate data into several groups according to their respective characteristics (Eko, 2012). In general, the clustering method is divided into two, namely Hierarchical Clustering and Non-Hierarchical Clustering.

The K-Medoids method is a method that aims to reduce the sensitivity of the resulting partitions with respect to the extreme values contained in the dataset, the use of medoids is not based on the observed mean belonging to each cluster (Supriyadi et al., 2021). The K-Medoids method is similar to the K-Means method, and K-Means is very sensitive to outliers so that it can be overcome by K-Medoids (Safitri et al., 2021). Following are the steps of the K-Medoids algorithm (Sindi et al., 2020) namely: determine the cluster center as many as k then randomly select objects in each cluster as new medoid candidates after that calculate the distance of each object in each cluster with the new medoid candidate using the Euclidean distance measure then mark the closest distance of the object to the medoid and calculate the total then do the medoid iteration until there is no movement of objects between clusters. Based on the previous explanation, the researcher wants to apply the K-Medoids method to cases of environmental pollution in Indonesia.

* Corresponding author.
E-mail address: zulkifli.rais89@unm.ac.id

One of the factors causing environmental pollution in Indonesia is the destruction of nature by residents. World Population Day which falls on July 11, and with a population that continues to increase every year. Environmental destruction is carried out due to lack of attention to ecosystems, which is not uncommon for us to see, this is because the higher the population level causes environmental pollution, both water, soil and air pollution. Environmental pollution is the entry or inclusion of living things, energy substances, and/or other components into the environment or changes to the environmental order by human activities or by natural processes so that the quality of the environment drops to a certain level which causes the environment to become less or unable to function anymore (Sipayung et al., 2020). Classification of environmental pollution in Indonesia can be done to find out the results of grouping provinces that experience the highest levels of environmental pollution by type, so the K-Medoids method can be used.

2. Literature Review

2.1. Data Mining

Data Mining is the process of finding useful information automatically in large data stores. Data mining techniques are used to explore large databases to find new and useful patterns that may not be known by others. Data mining is the process of finding interesting patterns and knowledge from large amounts of data (Han dkk., 2012). In general, data mining methods are divided into two, namely Supervised Learning and Unsupervised Learning.

2.2. Clustering

Clustering is part of data mining science that is unsupervised. Clustering is the process of dividing data into groups or clusters based on their level of similarity. Clustering is work that separates data or vectors into a number of groups or clusters according to their respective characteristics (Larose, 2014).

In general, the methods in cluster analysis are divided into two, namely Hierarchical Clustering and Non-Hierarchical Clustering. The hierarchical method for the number of groups to be formed has not yet been determined. The non-hierarchical method is a grouping method where the clusters to be formed are determined first, so that objects will be grouped into k predetermined groups. (Sindi dkk., 2020).

2.3. K-Medoids Algorithm

K-Medoids algorithm or another term Partitioning Around Medoids (PAM). Calculations on K-Medoids do not refer to the average (mean) value of all data in each cluster, the goal is to reduce the outliers or sensitivity of the resulting clusters in the dataset (Supriyadi dkk., 2021).

The steps of the K-Medoids algorithm are as follows (Sindi dkk., 2020):

1) Initialize k cluster centers (number of clusters).
2) Allocate each data (object) to the nearest cluster using the euclidean distance measure equation with the equation.

$$euclidean(a,b) = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2} \quad (2.1)$$

3) Randomly select objects in each cluster as new medoid candidates.
4) Calculate the distance between each object in each cluster and the new medoid candidate.
5) Calculate the total deviation (S) by calculating the value of the new total distance – the old total distance. if S < 0, then swap objects with data clusters to form a new set of k objects as medoids.
6) Repeat steps 3 to 5 until there is no medoid change, so that a cluster and each cluster member is obtained.

2.4. Davies Bouldin Index

The Davies Bouldin Index (DBI) was first introduced by scientists named David L. Davies and Donald W. Bouldin in 1979. This index is used to measure the validity of clusters in the clustering method. DBI maximizes the distance between clusters and minimizes the distance between the data and the cluster center point. The smaller the DBI value indicates that the better the cluster is. The formula for calculating the Davies Bouldin Index is as follows (Azrahwati dkk., 2022):

$$DBI = \frac{1}{k}.\sum_{i=1}^{k} max_{i \neq j}\left(R_{i,j}\right) \tag{2.2}$$

Where

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{2.3}$$

$$ssw_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j \,.\, c_i) \tag{2.4}$$

and

$$SSB_{i,j} = d(c_i, c_j) \tag{2.5}$$

Where SSW (*Sum of Square Within-cluster*) is the sum of the proximity of the data to the cluster center point, SSB (*Sum of Square Between-cluster*) is the distance between cluster center points, $R_{i,j}$ is to find out the comparison value between clusters to-i and cluster to-j, k is the number of data in the cluster, c_i is the cluster centroid to-i, m_i is the number of data in the to-i, $d(x_j, c_i)$ is the euclidean distance of each data to the centroid, $d(c_i, c_j)$ is the distance between centroids.

2.5. Environmental Pollution

The definition of the living environment is confirmed based on Law No. 32 of 2009, the environment is a spatial unit of all objects, forces, conditions, living things, including humans and other living things. Based on the above understanding, it can be simply concluded that the environment is a unit that includes all aspects of life. According to BPS, environmental pollution is something that causes changes to the environment, either directly or indirectly, which can endanger the health, safety and welfare of humans, which usually occurs for a long time. This disturbance can occur by itself (natural process) or caused by human activity. The types of environmental pollution are water pollution, soil pollution and air pollution (Badan Pusat Statistik, 2021).

3. Methods

3.1. Data Source

The data used in this study is secondary data sourced from the publication of the Central Bureau of Statistics. The Environmental Pollution Data used is 2021 data.

3.2. Variable Operational Defenition

1) Many Villages/ward have experienced water pollution from factory waste (X_1)

   Many villages/ward have experienced water pollution from factory waste such as liquid or solid chemicals, fuel residue, oil and oil spills.

2) Many villages/ward have experienced water pollution from household waste (X_2)

Many villages/ward experience water pollution from household waste such as organic waste (food scraps), inorganic waste (plastic, glass, cans) and chemicals (detergent, batteries).

3) Many villages/ward have experienced soil contamination from factory waste (X_3)

   Many villages/ward have experienced soil contamination from factory waste such as chemicals, both in liquid and solid forms. The majority of this solid waste is generated from factory waste, for example in a pulp factory.

4) Many villages/ward have experienced soil contamination from household waste (X_4)

   The large number of villages/ward that experience soil contamination from household waste is generally more caused by domestic waste, for example originating from household activities, such as washing and disposing of garbage.

5) Many villages/ward have experience air pollution from factory waste (X_5)

   Many villages/ward experience air pollution from factory waste, such as the air in factory industrial areas which has a lot of smoke from factory chimneys. This smoke contains harmful gases such as carbon dioxide, carbon monoxide and methane.

6) Many villages/ward have experience air pollution from household waste (X_6)

   Many villages/ward experience air pollution from household waste, for example in household activities such as cooking using firewood, burning garbage, using air conditioners and refrigerators which emit exhaust substances in the form of freon gas (CFC).

3.3. Data Analysis Techniques

The data analysis techniques in this study are:

1) Conduct descriptive analysis of the variables used.
2) Doing clustering with the K-Medoids algorithm. The steps are:
   a. Determines the number of clusters.
   b. Randomly select objects in each cluster as new centroid candidates.
   c. Calculate the distance of each object in each cluster to the new centroid candidate. The distance measure used is the Euclidean distance.
   d. Mark the object's closest distance to the centroid and calculate the total.
   e. Determines cluster members against temporary centroids.
   f. Perform centroid iteration. The iteration is carried out until there is no movement of objects between clusters.
3) Determine the best cluster using the Davies Bouldin Index.
4) Perform interpretation of the results of each cluster formed.

4. Results and Discussion

4.1. Descriptive Analysis

Table 1. Statistics of Environmental Pollution

Variable	Minimum	Maximum	Average	Median	Standar Deviation
X_1	6	778	181	97	197,15
X_2	5	532	132	71	153,69
X_3	0	112	20	10	27,21
X_4	1	86	18	9	22,72
X_5	1	85	19	11	21,59
X_6	2	548	99	36	132,68

4.2. The K-Medoids Method Cluster

The K-Medoids cluster method can be used to classify provinces in Indonesia based on environmental pollution. The stages of grouping are as follows:

a. Number of clusters

The grouping of provinces based on environmental pollution will be made into 2 to 5 clusters, this is done so that later the best number of clusters will be determined in grouping provinces based on environmental pollution.

b. Determines the initial cluster center or initial centroid

For example, 2 groups will be created, then 2 initial cluster centers or centroids will be determined. It is shown in Table 2 that the provinces of Central Java and Central Sulawesi are the initial cluster centers or initial centroids.

Table 2. Early Centroids

Province	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**
Aceh	241	109	6	3	26	115
Sumatera Utara	460	212	40	18	38	182
Jawa Tengah	**778**	**532**	**112**	**86**	**83**	**529**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sulawesi Tengah	**75**	**51**	**10**	**12**	**8**	**36**
Papua Barat	17	22	10	3	1	8
Papua	267	25	88	4	60	17

c. Allocate data to the nearest cluster starting center

The process of grouping data into a cluster can use a distance measure. The minimum distance between data and a certain centroid will determine the data group. The distance measure used is the Euclidean distance. The formulas and calculations to find the shortest distance are as follows:

$$d(a,b) = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2}$$

For example, an example is used to calculate the Euclidean distance in Data 1 and this step is repeated for 34 provinces.

$$\mathrm{C}_1 = \sqrt{(241-75)^2 + (109-51)^2 + (6-10)^2 + (3-12)^2 + (26-8)^2 + (115-36)^2}$$

$$= \sqrt{(27556) + (3364) + (16) + (81) + (324) + (6241)} = 193{,}861$$

$$\mathrm{C}_2 = \sqrt{(241-778)^2 + (109-532)^2 + (6-112)^2 + (3-86)^2 + (26-83)^2 + (115-529)^2}$$

$$= \sqrt{(288369) + (178929) + (11236) + (6889) + (3249) + (171396)} = 812{,}446$$

Based on the results of calculating the Euclidean distance for 34 provinces, this is summarized in Table 3. The province that has the smallest Euclidean distance is included in the first or second cluster.

Table 3. Initial Centroid Calculation Results

Province	C1	C2	Explanation
Aceh	193,8608	812,4457	1
Sumatera Utara	444,1824	579,4532	1
Jawa Tengah	995,0497	0	2
⋮	⋮	⋮	⋮
Sulawesi Tengah	0	995,0497	1
Papua Barat	71,54719	1065,213	1
Papua	216,2244	887,7742	1

d. Recalculating clusters with new cluster centers or new centroids in order to obtain convergent groups which are convergent, i.e. no group members change.

Table 4. New Centroids

Province	X_1	X_2	X_3	X_4	X_5	X_6
Nusa Tenggara Timur	76	174	2	4	4	114
Jawa Tengah	778	532	112	86	83	529

e. After obtaining the convergent results as shown in Table 5. It can be seen that the Euclidean distance calculation results have not changed, meaning that the calculation process has converged.

Table 5. New Centroid Calculation Results

Province	C1	C2	Explanation
Aceh	220,2158	812,4457	1
Sumatera Utara	471,2186	579,4532	1
Jawa Tengah	1023,087	0	2
⋮	⋮	⋮	⋮
Sulawesi Tengah	47,98958	995,0497	1
Papua Barat	61,10646	1065,213	1
Papua	224,7977	887,7742	1

f. Calculating the total deviation (S)

Based on the grouping results for the initial centroid and the new centroid, the next step is to calculate the total deviation (S) by calculating the total distance from the new centroid minus the total distance from the old centroid.

$$Total\ Deviation\ (S) = New\ Total\ Distance - Long\ Total\ Distance$$
$$Total\ Deviation\ (S) = 4877{,}985 - 4402{,}124$$
$$Total\ Deviation\ (S) = 475{,}8609$$

The centroid calculation process will continue to repeat itself until the data grouping does not change if s < 0. Based on the results above, the total deviation (S) is not less than 0, so the centroid does not change.

The following shows the plot of grouping results starting from groups of 2 to groups of 5.

Based on Figure 1, it can be seen that the first cluster consists of 31 provinces, namely (Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung, Riau Islands, DKI Jakarta, DI Yogyakarta, Banten, Bali, Nusa Tenggara West, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Papua). Meanwhile, the second cluster consists of 3 provinces (West Java, Central Java and East Java).

Based on Figure 2, it can be seen that the first cluster consists of 11 provinces, namely (Aceh, North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, South Sulawesi, Papua).

The second cluster consists of 20 provinces (West Sumatra, Riau, Bengkulu, Bangka Belitung, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Sulawesi Southeast, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua). And for the third cluster consists of 3 provinces (West Java, Central Java and East Java).

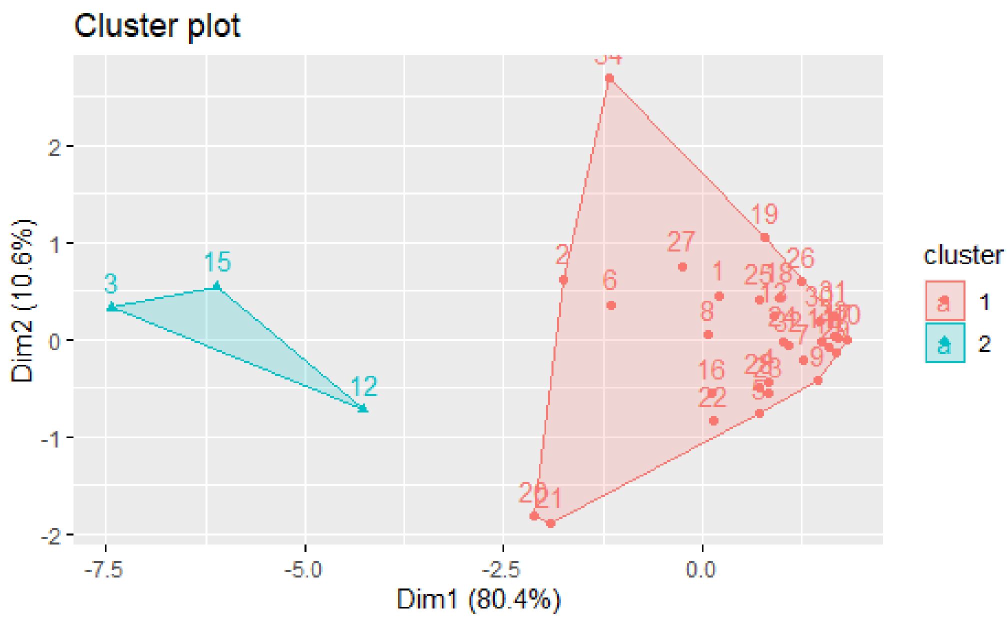

Figure 1. Plot K = 2

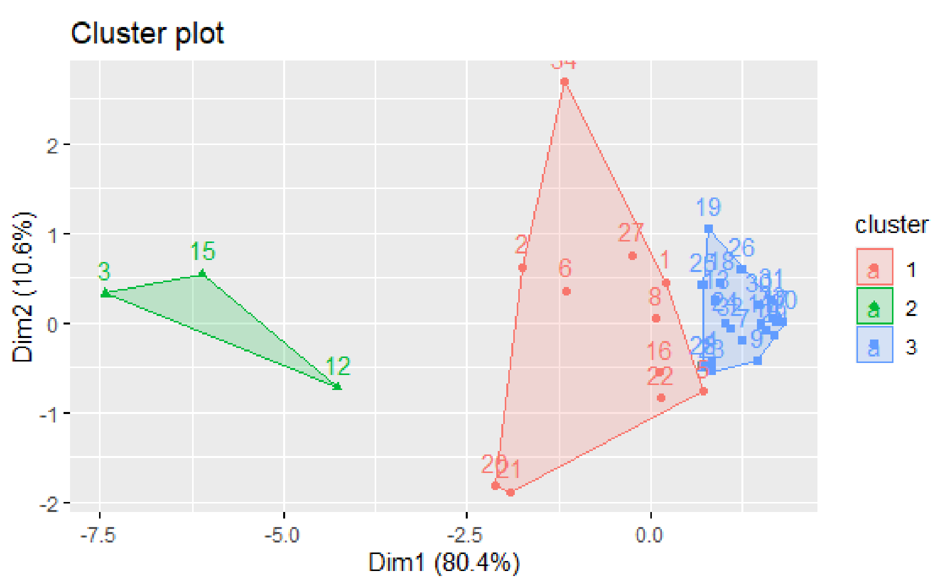

Figure 2. Plot K = 3

Based on Figure 3, it can be seen that the first cluster consists of 9 provinces, namely (Aceh, North Sumatra, Jambi, South Sumatra, Lampung, Banten, South Kalimantan, South Sulawesi, Papua). The second cluster consists of 20 provinces (West Sumatra, Riau, Bengkulu, Bangka Belitung, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West

Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Sulawesi Southeast, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua). The third cluster consists of 3 provinces (West Java, Central Java and East Java). And for the fourth cluster consists of 2 provinces (West Kalimantan and Central Kalimantan).

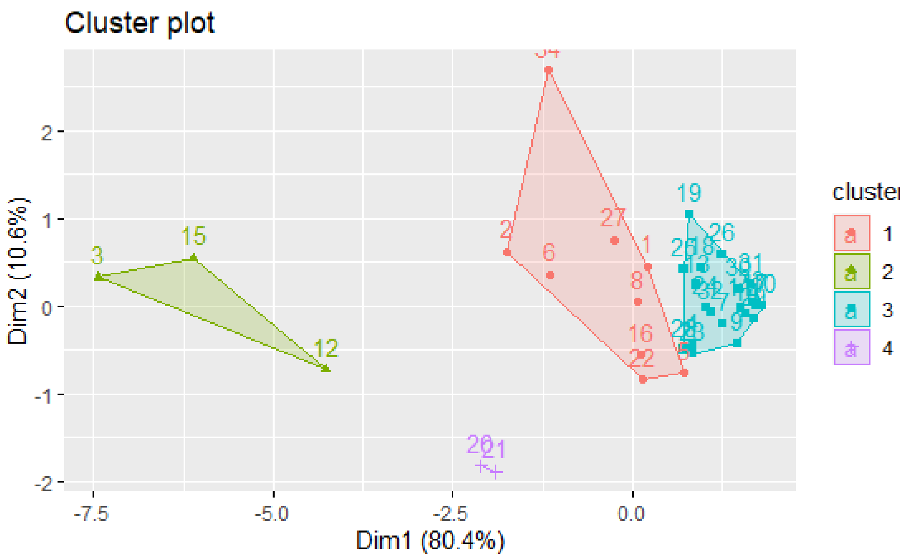

Figure 3. Plot K = 4

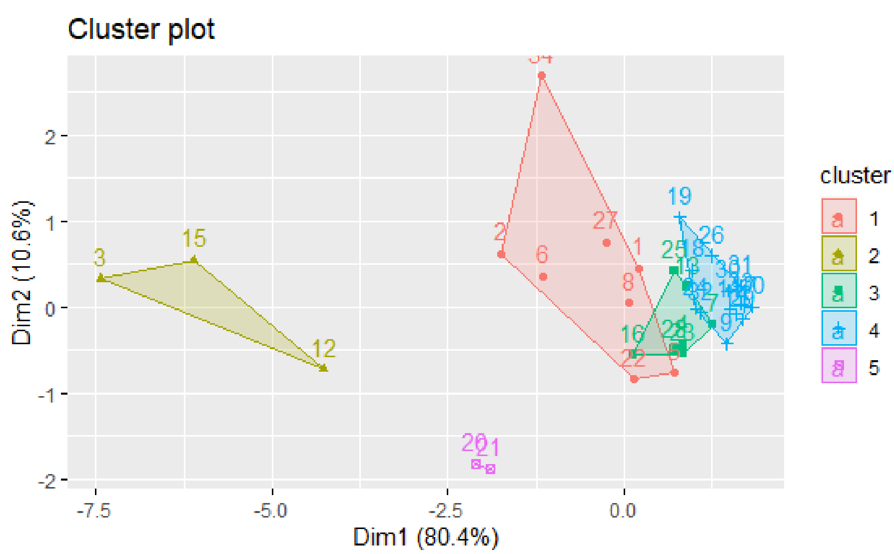

Figure 4. Plot K = 5

Based on Figure 4, it can be seen that the first cluster consists of 8 provinces namely (Aceh, North Sumatra, Jambi, South Sumatra, Lampung, South Kalimantan, South Sulawesi, Papua). The second cluster consists of 7 provinces (West Sumatra, Riau, Bengkulu, Banten, East Kalimantan, North Sulawesi, Southeast Sulawesi). The third cluster consists of 14, namely (Bangka Belitung, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, North Kalimantan, Central Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua). The fourth cluster consists of 3 provinces (West Java, Central Java and East Java). And for the fifth cluster consists of 2 provinces (West Kalimantan and Central Kalimantan).

4.3. Determination of the Best Cluster

Table 6. Results of the Davies Bouldin Index

Cluster	Davies Bouldin Index
K = 2	0,3051427
K = 3	0,6701511
K = 4	0,5882908
K = 5	0,7621123

Based on Table 6, the smallest DBI results were obtained when K = 2, which was 0,3051427, so that the best number of clusters was 2 groups. The following is shown in Figure 5 the results of the final grouping of provinces in Indonesia based on environmental pollution, namely 2 groups:

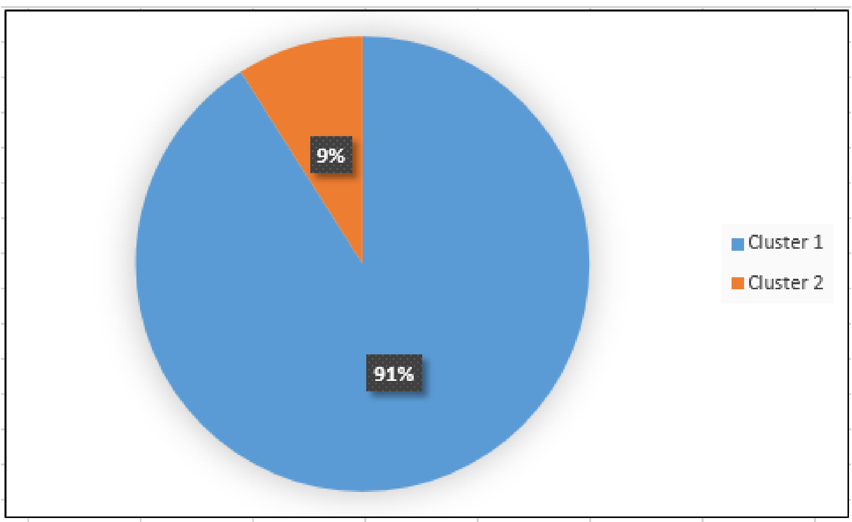

Figure 5. Number and Percentage of Clusters Formed

4.4. Interpretation of Province Situation Based on Cluster Results

By using the K-Medoids method, after the groups were formed, namely as many as 2 clusters, to see the state of the provinces based on the results of the groupings formed, then from all research objects as many as 34 provinces in Indonesia, the average of each variable of environmental pollution ($\bar{X}$) was taken and each group also takes the average value of each variable ($\bar{X}_c$).

Table 7. Group Descriptive Based on Average

		$\bar{X}_c$	
Variable	$\bar{X}$	**Cluster 1**	**Cluster 2**
X_1	181,18	129,22	718
X_2	132,24	95,87	508
X_3	19,91	14,03	80,67
X_4	17,82	13,29	64,67
X_5	18,62	13,90	67,33
X_6	98,82	62,93	469,67

Provinces that are included in cluster 1 are characterized by variables that have a low average value in the number of villages/ward that experience water pollution from factory waste, the number of villages/ward that experience water pollution from household waste, the number of villages/ward that experience pollution land from factory waste, the number of villages/ward experiencing soil pollution from household waste, the number of villages/ward experiencing air pollution from factory waste and the number of villages/ward experiencing air pollution from household waste.

Provinces in cluster 2 are grouped based on variables that have a high average value for each variable. So that the provinces in this cluster are the groups with the highest environmental pollution from the other groups.

5. Conclusion

The conclusions reached from this research can be given into several important points as follows:

1) Based on the results of a descriptive analysis of cases of environmental pollution in Indonesia, it was found that Central Java Province had the highest environmental pollution in terms of the number of villages/ward that experienced water pollution from factory waste (X_1), the number of villages/ward that experienced water pollution from household waste (X_2), the umber of villages/ward experienced soil pollution from factory waste (X_3) & the number of villages/ward experienced soil pollution from household waste (X_4) and East Java Province the highest in the variable the number of villages/ward experienced air pollution from factory waste (X_5) & the number of villages/ward that experienced air pollution from household waste (X_6).
2) The best cluster results are 2 where the first cluster consists of 31 provinces and the second cluster consists of 3 provinces. Provinces in the first cluster are provinces with low environmental pollution, while in the second cluster are provinces with high environmental pollution.

References

Azrahwati, Nusrang, M., Aidid, M. K., & Rais, Z. (2022). *K-Means Cluster Analysis for Grouping Districts in South Sulawesi Province Based on Village Potential. 2*(2), 73–82.

Badan Pusat Statistik. (2021). *Badan Pusat Statistik* (hal. 335–358). https://doi.org/10.1055/s-2008-1040325

Eko, P. (2012). Data mining konsep dan aplikasi menggunakan satscan. In CV Andi Offset. https://elibrary.bsi.ac.id/readbook/200350/data-mining-konsep-dan-aplikasi-menggunakan-matlab

Han, J., Kamber, M., & Oei, J. (2012). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining.pdf. In M. K. Novita Indriyani (Ed.), CV *Andi Offset* (hal. 114). CV Andi Offset.

Larose, D. T. (2014). Discovering Knowledge in Data: an Introduction to Data Mining. In *Journal of the American Statistical Association* (Vol. 100, Nomor 472). https://doi.org/10.1198/jasa.2005.s61

Safitri, P. N., Aristawidya, R., & Faradilla, S. B. (2021). Klasterisasi Faktor-Faktor Kemiskinan Di Provinsi Jawa Barat Menggunakan K-Medoids Clustering. *Journal of Mathematics Education and Science*, 4(2), 75–80. https://doi.org/10.32665/james.v4i2.242

Sindi, S., Ningse, W. R. O., Sihombing, I. A., Zer, P. P. P. A. N. . F. iImi R. H., & Hartama, D. (2020). *Analisis algoritma k-medoids clustering dalam pengelompokan penyebaran covid-19 di indonesia*. 4(1), 166–173.

Sipayung, A. T., Saifullah, & Winanjaya, R. (2020). Penerapan Metode K-Means Dalam Mengelompokkan Banyaknya Desa/ Kelurahan Menurut Keberadaan Permukiman Di Bantaran Sungai Berdasarkan Provinsi. *Brahmana : Jurnal Penerapan Kecerdasan Buatan*, 2(1), 49–56. https://doi.org/10.30645/brahmana.v2i1.48

Supriyadi, A., Triayudi, A., & Sholihati, I. D. (2021). *Perbandingan algoritma k-means dengan k-medoids pada pengelompokan armada kendaraan truk berdasarkan produktivitas*. 06, 229–240.