BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 0923–0936
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp0923-0936

GROUPING PROVINCES IN INDONESIA BASED ON THE NUMBER OF VILLAGES AFFECTED BY ENVIRONMENTAL POLLUTION WITH K-MEDOIDS, FUZZY C-MEANS, AND DBSCAN

Idrus Syahzaqi [1*], Magdalena Effendi[2], Hasri Rahmawati[3], Heri Kuswanto[4], Sediono[5]

[1,5]*Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga*
Kampus C Universitas Airlangga, Surabaya, 60115, Indonesia

[2,3,4]*Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember*
Jln Raya ITS, Surabaya, 60111, Indonesia

*Corresponding author's e-mail: * **idrus.syahzaqi@gmail.com***

Article History:

Received: 23rd June 2023
Revised: 9th January 2023
Accepted: 13th March 2024
Published: 1st June 2024

Keywords:

Enviromental Pollution;
K-Medoids;
Fuzzy C-Means;
DBSCAN

ABSTRACT

Pollution can cause the environment to not function properly and ultimately harm humans and other living things. Environmental pollution is a problem that needs to be resolved because it involves the safety, health, and survival of living things. Air pollution in Pekanbaru due to a long dry season has resulted in forest fires. Then, 70% of drinking water is contaminated by fecal waste. In addition, the contamination of the land by the Chevron company resulted in residents suing the company. Until now, there has been no research that has carried out a comparison between methods for grouping villages affected by environmental pollution at the provincial level in Indonesia, so it is important to select the best method for carrying out the grouping. The limitations of this research are the use of three methods for clustering: K-Medoids, Fuzzy C-Means, and DBSCAN. The results showed that Fuzzy C-Means with five clusters have an optimal value compared to DBSCAN with an ICD rate value of 0,351. This method can be used by the government to improve the quality of villages that are clean from pollution in Indonesia, monitoring and evaluation based on the clusters formed.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

I. Syahzaqi, M. Effendi, H. Rahmawati, H. Kuswanto and Sediono., "GROUPING PROVINCES IN INDONESIA BASED ON THE NUMBER OF VILLAGES AFFECTED BY ENVIRONMENTAL POLLUTION WITH K-MEDOIDS, FUZZY C-MEANS, AND DBSCAN," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 0923-0936, June, 2024.

Copyright © 2024 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Environmental pollution is the changes that occur in the environment due to human activities or natural processes so that the quality of the environment decreases, resulting in the environment not functioning **[1].** This problem needs to be resolved because it concerns the safety, health, and survival of living things. There are three types of environmental pollution, namely air, water, and soil pollution.

Air pollution is felt to be increasing every day. Polluted air can damage the environment and potentially disrupt the respiratory health of the surrounding community. The causes of this pollution are industrial exhaust emissions and motor vehicle activities. In addition, air pollution can also be caused by nature, such as a long dry season, which results in forest fires so that the smoke pollutes the air. This case occurred in Sumatra, especially Pekanbaru City **[2]**. Apart from air, water is also an important component used by humans to carry out daily activities such as bathing, drinking, washing, cooking, and other activities. Of course, the water used must be clean in terms of quality and quantity. However, this is not supported by the water conditions in the field. The results of a study from **[3]** found that almost 70% of the 20,000 drinking water sources tested in Indonesia were contaminated with fecal waste and contributed to the spread of disease.

In addition, soil pollution has an impact on the survival of living things. Cases of soil contamination can be caused by household waste or industrial activity waste. In Indonesia, the Chevron company, which is a multinational company engaged in the oil and gas sector, has been sued by the local community because the B3 waste it produces has polluted the soil **[4]**.

Based on the background above, this research has urgency in the form of an effort to carry out sustainable development in meeting the SDGs; it is necessary to map the number of villages affected by environmental pollution at the provincial level in Indonesia so that it does not become more widespread. This role can be performed by statisticians in analyzing data related to the occurrence of environmental pollution. Data on cases of Environmental Pollution can be used for the grouping process according to the information held from accurate data so that the number of villages according to the type of environmental pollution can be identified. Several clustering methods can be applied to get grouping results from provinces that experience the most environmental pollution by type.

Cluster analysis aims to group objects based on the characteristics of the objects so that the characteristics of each group can be identified **[5]**. In this study, the Fuzzy C-Means (FCM) and Density-based Spatial Clustering of Applications with Noise (DBSCAN) methods were used to group data on many villages affected by environmental pollution at the provincial level in Indonesia. K-Medoids and DBSCAN are used because they are robust to the existence of outlier data so that all provinces can be grouped, whereas FCM is not the case. It is interesting to compare the methods. The selection of many clusters uses the Pseudo F method, while the selection of the best cluster method is based on the minimum ICD Rate value. The results of this study can be used as a reference in terms of prioritizing provinces with many villages having the most environmental pollution so that they can improve the quality of the environment and reduce the impact of environmental pollution.

2. RESEARCH METHODS

This section contains explanations about the variables and data sources, literature review about methods, and research stages.

2.1 Variables and Data Sources

The data is accessed from Harvard Dataverse, namely Village Potential Statistics (PODES) in 2021, published by Badan Pusat Statistik **[6]**. The data is on the number of villages that are affected by the type of environmental pollution. The research unit is a province in Indonesia. The following research variables are presented in **Table 1**.

Table 1. Variable Explanation

Variable	Description	Scale	Unit
X_1	The Number of Villages Affected by Water Pollution	Ratio	Villages
X_2	The Number of Villages Affected by Soil Pollution	Ratio	Villages
X_3	The Number of Villages Affected by Air Pollution	Ratio	Villages
X_4	The Number of Villages Not Affected by Pollution	Ratio	Villages

2.2 Literature Review about Methods

Clustering methods are multivariate statistical methods that aim to group objects based on similarities and characteristics possessed by these objects. In a cluster, objects will be grouped so that each object has the most common characteristics with other objects that are in the same group.

There are two types of clustering methods: hierarchical clustering and hierarchical clustering [7]. Where is a hierarchical method with various single linkage, complete linkage, and minimum-variance approaches? The non-hierarchical method, also known as the partition method, has also been developed with various algorithmic approaches. Visible difference It is clear at the beginning of the procedure that the method is hierarchical, grouping an observation gradually. Meanwhile, the non-hierarchical method partitions the sample space.

1. Descriptive Statistics

Descriptive statistics or deductive statistics are part of statistics that studies data collected and presented so that other people can easily understand them [8]. Descriptive statistics only relate to things outlining or giving information about data, conditions, or phenomena. In other words, descriptive statistics work describes a situation, symptom, or problem. Descriptive statistics in this study are used to present the characteristics of provinces in Indonesia based on the number of villages affected by environmental pollution; in this research, descriptive statistics are presented with bar charts, boxplots, and minimum and maximum values.

2. K-Medoids

The PAM (Partitioning Around Medoid) is the algorithm represented by the cluster of the medoid. The algorithm is more popularly known as the K-Medoids. The k-Medoids algorithm is better to use than the K-Means. In K-Medoids, we can search k as an object representation to minimize the number of object inequalities in data, while in K-Means, we use sums of Euclidean distance in data objects [9]. When compared with the K-Means algorithm, the difference lies in the central point cluster. K-Means uses the average as the cluster center, whereas K-Medoids uses an object as the cluster center, and the object must represent each of the clusters.

3. Fuzzy C-Means (FCM)

The Fuzzy C-Means method is a grouping method that develops by involving fuzzy properties in its membership [10]. The FCM method reallocates data into each group by utilizing fuzzy theory [11]. The FCM uses a variable membership function u_{ik}, which refers to how most likely a data can be a member of a group. FCM introduces a variable w which is the weighting exponent of the membership function.

This variable can change the magnitude of the influence of the membership function in the clustering process using the FCM method, w has a value area greater than 1 ($w > 1$). Until now, there is no clear provision of how large the value of w is optimal in carrying out the process of optimizing a clustering problem, the value of w only affects the number of iterations in obtaining the objective function the greater the w, the faster the matrix converges.

4. DBSCAN

DBSCAN is a method for grouping data, where the algorithm is based on the density of the data. Clustering [12]. The advantages of this method is that it can detect outliers/noise, there is no need to get prefix input in the form of the number of clusters (k) such as K-Means or K-Medoids and can recognize irregular cluster shapes [13]. There are two parameters in DBSCAN, namely the minimum amount of data (minPts) within the eps radius $(\boldsymbol{\varepsilon})$. The DBSCAN algorithm generates three kinds of states for each data, namely core,

border, and noise. Core data is data where the amount of data within the eps radius is more than minPts, noise data is data where the amount of data within the eps radius is less than minPts, and boundary data is data where the amount of data within the eps radius is less than minPts but makes the neighboring data core data [14]. The DBSCAN grouping process is to calculate the distance from the center point $(\boldsymbol{p})$ to other points using the Euclidean distance [15].

5. Pseudo F Statistics

Pseudo F criteria can be chosen to select the optimum number of clusters. The highest Pseudo F value indicates that the number of groups used for grouping data has been optimal [16]. The following equation of Pseudo F value are presented:

$$\boldsymbol{Pseudo\ F} = \frac{\left(\frac{\boldsymbol{R^2}}{\boldsymbol{c-1}}\right)}{\left(\frac{\boldsymbol{1-R^2}}{\boldsymbol{n-c}}\right)} \tag{1}$$

$$\boldsymbol{R^2} = \frac{(\boldsymbol{SST-SSW})}{\boldsymbol{SST}} \tag{2}$$

where:

SST	:	the total sum of the squares of the distances sample to the overall mean
SSW	:	the total sum of the squares of the distances sample to the group mean
n	:	the number of samples
c	:	the number of groups

6. ICD Rate

The ICD (Internal Cluster Dispersion) rate criteria can be used to determine the best cluster method. The smaller the ICD rate value indicates the result the better the grouping. The ICD rate value is the level of dispersion within the cluster. The following equation of ICD rate are presented below [17].

$$\boldsymbol{Pseudo-F} = \frac{\left(\boldsymbol{R^2}/\boldsymbol{c-1}\right)}{\left(\boldsymbol{1-R^2}/\boldsymbol{n-c}\right)} \tag{3}$$

2.3 Research Stages

The methods used in this research are K-Medoids, FCM, and DBSCAN. This method is used to group data on many villages affected by environmental pollution at the provincial level in Indonesia. The following research stages are presented below.

1. Describe the characteristics of the data.
2. Perform multivariate outlier assumption tests with 50% bivariate normal contours.
3. Testing KMO and Bartlett tests.
4. If there is a correlation between variables, carry out the reduction variable using Principal Component Analysis (PCA).
5. Grouping provinces using the K-Medoids, FCM, and DBSCAN methods. As for determining the optimum number of clusters in each method with Pseudo F statistics values.
6. Comparing the results of clustering K-Medoids, FCM, and DBSCAN with ICD rate values.
7. Perform multivariate normal assumption tests and tests homogeneity. Determine the difference in characteristics on each cluster using nonparametric one-way MANOVA and one-way ANOVA.

3. RESULTS AND DISCUSSION

3.1 Characteristic of Provinces in Indonesia Based on Types of Enviromental Pollution

Descriptive statistics aim to show an overview of the problem. Descriptions related to the number of villages affected by environmental pollution in Indonesia can be seen from the size of the concentration and distribution of data. The measure of centering uses the average, minimum, and maximum values. Meanwhile, the measure of the spread uses the standard deviation value.

Table 2. Data Characteristics

Variable	Mean	Standard Deviation	Minimum	Maximum
X_1	314.2	339.3	16	1310
X_2	44.1	51.4	4	224
X_3	166	199.9	15	781
X_4	2058	1906	173	6932

From **Table 2**, the average village affected by water pollution is 314.2 villages whereas Central Java Province is the province with the most villages affected by water pollution (1310 villages), while Kep. Riau is a province with the fewest villages affected by water pollution (16 villages).

For soil pollution, it is known that the average village affected by soil pollution is 44,1 villages whereas Central Java Province is the province with the most villages affected by soil pollution (224 villages), while Kep. Riau is a province with the fewest villages affected by soil pollution (4 villages).

For air pollution, it is known that the average village affected by air pollution is 166 villages. Central Java Province is the province with the most villages affected by air pollution (781 villages). In comparison, Bali is a province with the fewest villages affected by air pollution (15 villages).

From **Table 2**, the average village not affected by pollution is 2058 villages where East Java Province is the province with the most villages not affected by pollution (6932 villages), while DKI Jakarta is a province with the fewest villages not affected by pollution (173 villages). From a very large standard deviation value, it shows that the data variation is getting wider. This means that villages affected by water pollution, soil pollution, and air pollution and not affected by pollution in every province of Indonesia are very diverse.

3.2 Checking Multivariate Outlier and Multivariate Normal Assumption

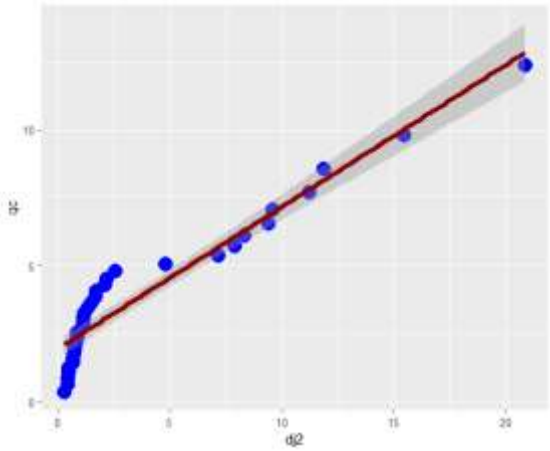

Figure 1. Multivariate Outlier Scatterplot

From **Figure 1**, the data distribution doesn't follow the red line, and it's known that there are six outlier points (Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua). In this study, abnormalities due to outliers in the data were not resolved because grouping would be carried out using the K-Medoids, FCM, and DBSCAN methods which are robust to outliers.

Then, a multivariate normal assumption test was performed. The results showed that the proportion of the squared value of the distance was smaller than the chi-squared value of 70,588%. With this proportion value, it can be concluded that the data does not meet the multivariate normal distribution assumption because the proportion of squared distance values that are in the normal 50% bivariate contour is not in the 45% -55% interval.

3.3 Variabel Reduction with PCA

For further analysis, it is necessary to test data adequacy and variable dependencies.

Table 3. Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.688
Bartlett's Test of Sphericity	*Approx. Chi-Square*	148.522
	Df	6
	Sig.	0.000

Data adequacy can be achieved with the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO). From the test results in **Table 3**, it's known that the KMO value is 0,688. The KMO value obtained is greater than 0,5, which means the data is sufficient and worthy of further analysis carry on.

Variable dependency testing is carried out to find out whether the dependent variable (correlated) is carried out using Bartlett's Sphericity test. From **Table 3**, variables used in the study are correlated (dependent) because they have a p-value smaller than α (0.05), so that can proceed to principal component analysis for reduction into several factors. The following is an equation based on the 2 main components that have been formed.

$PC1 = 0.951\ X_1 + 0.915\ X_2 + 0.951\ X_3 + 0.890\ X_4$

$PC2 = -0.242\ X_1 - 0.346\ X_2 + 0.187\ X_3 + 0.414\ X_4$

3.4 Using K-Medoids to Group Provinces in Indonesia

Before grouping the K-Medoids method, the Pseudo F statistic value is calculated to decide the optimal number of clusters to be selected in classifying the number of villages affected by pollution according to province in Indonesia. Pseudo F statistics calculations are carried out for clusters totaling 2 to 5. The Pseudo F statistics values obtained are shown in **Table 4**.

Table 4. Pseudo F Statistics Value of K-Medoids Method

Number of Cluster	Pseudo F-Statistic
2	23.889
3	24.592
4	**34.474**
5	32.894

From **Table 4**, the Pseudo F statistics value is used to determine the optimal number of clusters. The optimum cluster is selected from the higher the Pseudo F statistics value. The highest Pseudo F statistics value is achieved by using 4 clusters, and then the K-Medoids method will use 4 clusters. Then **Table 5** below shows the details of the provinces in each cluster.

Table 5. Grouping of K-Medoids

Cluster	Member of Cluster
1	Aceh, North Sumatra, Riau, South Sumatra, Lampung, East Nusa Tenggara, South Sulawesi, Southeast Sulawesi, West Papua, and Papua.
2	West Sumatra, Jambi, Bengkulu, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, DI Yogyakarta, Banten, Bali, West Nusa Tenggara, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Gorontalo, West Sulawesi, Maluku and North Maluku.
3	West Java, Central Java, and East Java.
4	Central Kalimantan and West Kalimantan

After obtaining the results of grouping provinces in Indonesia based on the number of villages affected by pollution using the K-Medoids method, the next step is to determine the characteristics of each cluster formed, which will be shown in **Table 6** below.

Table 6. Characteristic of Clusters with K-Medoids

Cluster	Water Pollution	Soil Pollution	Air Pollution	Not Affected by Pollution
1	673	72	339	5227
2	163	8	71	1307
3	308	39	229	2580
4	99	31	47	374

From **Table 6**, cluster 1 is a province with the highest number of villages affected by pollution, namely water pollution, soil pollution, air pollution, and no pollution. Cluster 3 members are provinces with a number of villages affected by water pollution, soil pollution, air pollution, and no high pollution. Cluster 2 members are provinces with a relatively high number of villages affected by water pollution, air pollution, and no pollution, while the number of villages affected by soil pollution is low. Cluster 4 members are provinces with a low number of villages affected by pollution, namely water pollution, air pollution, and no pollution, while the number of villages affected by soil pollution is quite high.

3.5 Using FCM to Group Provinces in Indonesia

To perform grouping with the FCM Clustering method, then first, the Pseudo F statistic value is calculated to decide the optimal number of clusters to be selected in classifying provinces in Indonesia based on the number of villages affected by pollution. Pseudo F statistics calculations are performed for the clusters totaling 2 to 5. The Pseudo F statistics values obtained are shown in **Table 7**.

Table 7. Pseudo F Statistics of FCM

Number of Cluster	Pseudo F-statistic
2	17.85985
3	28.63626
4	39.05637
5	**39.52189**

The Pseudo F statistics value obtained as shown in Table 4. The optimum cluster is selected from the higher the Pseudo F statistics value. The higher the Pseudo F statistics value, the more optimal it will be clusters. By using 5 clusters, the highest Pseudo F statistics value is obtained so that the FCM method will use 5 clusters. **Table 8** shows details of provinces in each cluster.

Table 8. Grouping with FCM

Cluster	Member of Cluster
1	West Sumatra, Kep. Bangka Belitung, Jambi, Kep. Riau, DI Yogyakarta, Banten, DKI Jakarta, Bali, West Nusa Tenggara, South Kalimantan, East Kalimantan, North Kalimantan, Gorontalo, West Sulawesi, and North Maluku.
2	Riau, South Sumatra, Bengkulu, Lampung, East Nusa Tenggara, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Maluku and West Papua.
3	West Kalimantan and Central Kalimantan.
4	Aceh, North Sumatra, and Papua.
5	East Java, Central Java, and West Java.

After obtaining the results of grouping provinces in Indonesia based on the number of villages affected by pollution using the FCM clustering method, the next step is to determine the characteristics of each cluster formed.

Table 9. Characteristic of Clusters with FCM

Cluster	Water Pollution	Soil Pollution	Air Pollution	Not Affected by Pollution
1	153.1	18.4	62.7	739.3
2	187.9	25.5	130.4	2012.7
3	662.5	123	123	1148
4	438.3	63.3	303.3	5409.3
5	1226.3	169	704.7	6070.7

Based on **Table 9**, it's known that cluster 1 members are provinces with the lowest numbers of villages affected by pollution and not affected by pollution. Members from Cluster 2 are provinces with a low number of villages affected by water, soil, and air pollution, while the number of villages not affected by pollution is moderate. Cluster 3 members are provinces with a high number of villages affected by water and soil pollution. The number of villages affected by air pollution is moderate, while those not affected by air pollution are low. Cluster 4 members are provinces with a moderate number of villages affected by water and soil pollution. The number of villages affected by air pollution and that are not affected by pollution is high. Cluster 5 members are provinces with very high numbers of villages affected by pollution and not affected by pollution.

3.6 Using DBSCAN to Group Provinces in Indonesia

The clustering uses the DBSCAN method, before the grouping is done using the DBSCAN method, the Pseudo F-statistic value is calculated to decide the optimal eps and minPts values to be selected in grouping the number of villages affected by pollution according to provinces in Indonesia. Pseudo F statistics calculations were carried out based on eps values from 0,5 to 0,7, and 2 minPts were selected. The Pseudo F statistics values obtained are shown in **Table 10**.

Table 10. Pseudo F statistics of DBSCAN

Eps	MinPts	Pseudo F-statistic
0,5	**2**	**29.605**
0,6	2	15.072
0,7	2	27.236

The Pseudo F statistics values are obtained as in **Table 10**. The optimum cluster is selected from the higher the Pseudo F statistics value. The higher the Pseudo F statistics value, the more optimal the cluster will be. The highest Pseudo value is found in the eps value of 0.5 and 2 minPts, which then results in 2 clusters. With these results, the DBSCAN method will use 2 clusters.

Table 11. Grouping with DBSCAN

Cluster	Member of Cluster
1	West Sumatra, Riau, Jambi, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, DI Yogyakarta, Banten, Bali, West Nusa Tenggara, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua.
2	West Kalimantan, and Central Kalimantan.
Outlier	South Sumatra, Aceh, North Sumatra, East Java, Central Java, West Java, and East Nusa Tenggara.

After obtaining the results of grouping provinces in Indonesia based on the number of villages affected by pollution using the DBSCAN method, the next step is to determine the characteristics of each cluster formed.

Table 12. Characteristic of Clusters with DBSCAN

Cluster	Water Pollution	Soil Pollution	Air Pollution	Not Affected by Pollution
1	160.2	18.7	81.1	1139.4
2	662.5	123	123	1148

Based on **Table 12**, it is known that the members of cluster 1 are provinces with a lower number of villages affected by pollution and not affected by pollution than cluster 2. Then there are 7 provinces that are included in the outliers.

3.7 Comparison Analysis of Province Grouping Results in Indonesia

The results of each of the best cluster methods were then compared to choose the optimal cluster method. The best cluster method is selected based on the minimum ICD Rate value. The following **Table 13** shows a comparison between FCM and DBSCAN.

Table 13. Grouping with DBSCAN

Methods	ICD Rate
K-Medoids (4 Clusters)	0.386
FCM (5 Clusters)	**0.351**
DBSCAN (2 Clusters)	0.520

Based on **Table 13**, the FCM method is better than the DBSCAN. This was seen from the smaller ICD rate value, which means that FCM is very suitable for use in grouping provinces in Indonesia based on the number of villages affected by environmental pollution. The following grouping provinces in Indonesia based on the number of villages affected by pollution using FCM can be seen in **Figure 2**

Figure 2. Map of Province in Indonesia using FCM

To improve the quality of villages that are clean from pollution in Indonesia, monitoring and evaluation can be carried out by the Government in each province based on the clusters formed. Cluster 1 is a province with very low numbers of villages affected by pollution. Cluster 2 is provinces with a low number of villages affected by water, soil, and air pollution, while the number of villages not affected by pollution is moderate.

In provinces in cluster 3, there needs to be an emphasis on reducing water and soil pollution. The Government needs to urge people to buy biodegradable products, reduce the amount of plastic, and reduce the use of pesticides and fertilizers in agricultural activities. In addition, the Government also needs to implement strict regulations regarding proper waste management, urge the public to use environmentally friendly materials, not throw garbage in rivers or other water sources, and routinely carry out efforts to clean water sources and plant trees. Action in each area. Available land.

For provinces in cluster 4, there needs to be an emphasis on reducing air pollution. The government needs to implement a vehicle emission test policy to measure the level of combustion efficiency in engines. The government can also improve mass transportation services to be environmentally friendly and provide pathway facilities for pedestrians. For industry, the government can oversee compliance with emission quality standards, report emissions continuously or continuously, and be integrated with the reporting system at the Ministry of Environment and Forestry. In addition, the government must continue to build and develop city parks, urban forests, and botanical gardens as the lungs of the city.

Meanwhile, provinces in cluster 5 can emphasize reducing water, soil, and air pollution. The government in this province can take the same action as the provinces in clusters 3 and 4. For further research, it is recommended not to use the random matrix as the initial membership degree matrix in the FCM method.

3.8 Determination of Characteristic Differences using Nonparametric One-Way MANOVA and One-Way ANOVA

Before carrying out the MANOVA and ANOVA tests, a multivariate normal assumption test and the homogeneity of the covariance matrix were tested. From the result obtained in **Figure 1**, it is known that the proportion of values in the squared distance, which is smaller than the chi-squared value, is 0,7058. With the value of these proportions, it can be concluded that the multivariate normal distribution assumption is not met in these data, so the test will be carried out using nonparametric one-way MANOVA.

The result of the next test is to look at the value of Box's M. From the Box's M test, the p-value is 0.000, and the F-value is 8.658. The results show that the p-value < α (5%) rejects H_0, which means the variance covariance matrix is non-homogeneous. The nonparametric one-way MANOVA test yields a W^2 of 43.6758 which indicates that the W^2 value is greater than the chi-square value (26.2962). This means that there are differences in the characteristics of each cluster that is formed. Then, proceed with nonparametric one-way ANOVA analysis, which will also display a boxplot. Boxplots are one way in descriptive statistics to describe the distribution of data graphically from numerical data. Here are boxplots of each variable, along with a review of the nonparametric one-way ANOVA that has been performed.

1. Water Pollution

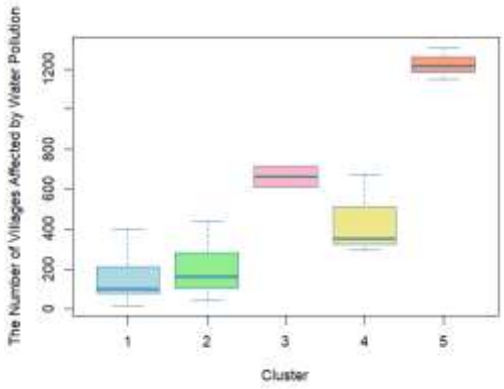

Figure 3. Boxplot Based on the Number of Villages Affected by Water Pollution

In **Figure 3**, clusters 1, 2, 3, 4, and 5 have different medians and ranges of data distribution. In terms of the number of villages affected by water pollution, it is known that cluster 5 is the dominant cluster, which is higher than the other clusters, while cluster 1 is the cluster which tends to be lower compared to other clusters.

Based on the results of the nonparametric ANOVA test, the results obtained were a p-value of 0.002 and an *H* of 16.427, so from these values, it is known that the *H* value > chi-square value (9.488). Then, reject H_0, and it can be concluded that the variable of the number of villages affected by water pollution has an influence on formation of provincial groupings in Indonesia.

2. Soil Pollution

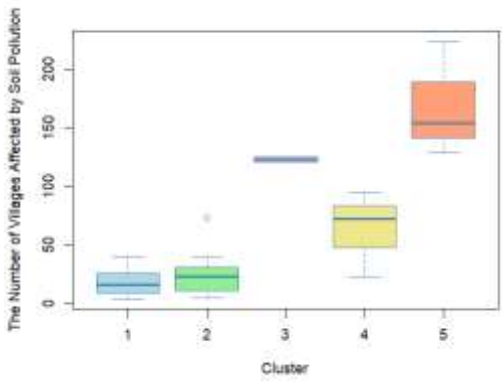

Figure 4. Boxplot Based on the Number of Villages Affected by Soil Pollution

In **Figure 4**, clusters 1, 2, 3, 4, and 5 have different medians and ranges of data distribution. In the characteristics of the number of villages affected by soil pollution, it is known that cluster 5 is the dominant cluster, which is higher than the other clusters, while cluster 1 is the cluster which tends to be low compared to other clusters. There are outlier values in cluster 2, namely in the provinces South Sumatra, which have villages exposed to far more air pollution than other provinces in their clusters.

From the results of the nonparametric ANOVA test, the results obtained were a p-value of 0.003 and an *H* of 15.769, so that from these values it is known that the *H* value > chi-square value (9.488). Then, reject H_0, and it can be concluded that the variable of the number of villages affected by soil pollution has an influence on formation of provincial groupings in Indonesia.

3. Air Pollution

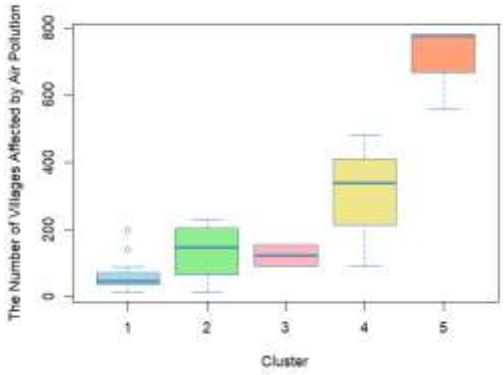

Figure 5. Boxplot Based on the Number of Villages Affected by Air Pollution

In **Figure 5**, clusters 1, 2, 3, 4, and 5 have different medians and ranges of data distribution. In the characteristics of the number of villages affected by air pollution, it is known that cluster 5 is the dominant cluster, which is higher than the other clusters, while cluster 1 is the cluster tends to be low compared to other clusters. There are outlier values in cluster 1, namely in the provinces of Banten and South Kalimantan, which have villages exposed to far more air pollution than other provinces in their clusters.

From the results of nonparametric ANOVA test, the results obtained were a p-value of 0.002 and an *H* of 16.513, so that from these values it is known that the *H* value > chi-square value (9.488). Then, reject H_0, and it can be concluded that the variable of the number of villages affected by air pollution has an influence on formation of provincial groupings in Indonesia.

4. Not Affected by Pollution

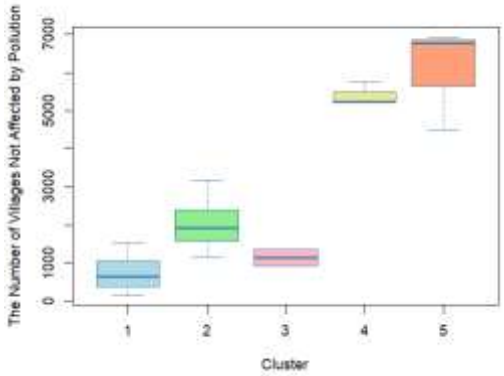

Figure 6. Boxplot Based on the Number of Villages Not Affected by Pollution

In **Figure 6**, it can be seen that clusters 1, 2, 3, 4, and 5 have different medians and ranges of data distribution. In the characteristics of the number of villages that are not affected by pollution, it is known that cluster 5 is the dominant cluster, which is higher than the other clusters, while cluster 1 is the cluster tends to be low compared to other clusters.

From the results of nonparametric ANOVA test, the results obtained were a p-value of 0.000 and an *H* of 26.469 so that from these values it is known that the *H* value > chi-square value (9.488). Then, reject H_0, and it can be concluded that the variable of the number of villages not affected by pollution has an influence on formation of provincial groupings in Indonesia.

4. CONCLUSIONS

Based on the results of the analysis and discussion, it can be concluded as follows.

1. Optimum cluster results with Fuzzy C- Means clustering there are 5 clusters, K-Medoids with 4 clusters, and DBSCAN method clustering has 2 clusters and 7 outlier provinces.
2. The best method based on the lowest ICD rate value on the grouping of provinces in Indonesia on the number of villages affected by environmental pollution is using the FCM method.
3. This research has limitations in data, so future research can use the latest data to better describe the condition of environmental pollution in the villages of each Indonesian province. Apart from that, in future research we can use more advanced classification methods such as random forest or naive Bayes.
4. To improve the quality of villages that are clean from pollution in Indonesia, monitoring and evaluation can be carried out by the Government in each province based on the clusters formed.

ACKNOWLEDGMENT

The authors give high appreciation to Harvard Dataverse and BPS RI, which has provided data and supported this research.

REFERENCES

[1] M. S. Muslimah and S. Si, “Dampak pencemaran tanah dan langkah pencegahan,” *J. Penelit. Agrisamudra*, vol. 2, no. 1, pp. 11–20, 2017.

[2] UNICEF., “Indonesia: Nearly 70 percent of household drinking water sources contaminated by faecal waste,” 2022. https://www.unicef.org/indonesia/press-releases/indonesia-nearly-70-cent-household-drinking-water-sources-contaminated-faecal-waste (accessed Feb. 07, 2022).

[3] F. Yazid and M. Affandes, “Clustering Data Polutan Udara Kota Pekanbaru dengan Menggunakan Metode K-Means Clustering,” *J. CoreIT*, vol. 3, no. 2, p. 1, 2017.

[4] M. Syukur, “Menelisik Dugaan Pencemaran Lingkungan oleh Perusahaan Minyak di Riau,” *LIPUTAN 6*, 2021. https://www.liputan6.com/regional/read/4600332/menelisik-dugaan-pencemaran-lingkungan-oleh-perusahaan-minyak-di-riau

[5] R. Sitepu, I. Irmeilyana, and B. Gultom, “Analisis cluster terhadap tingkat pencemaran udara pada sektor industri di Sumatera Selatan,” *J. Penelit. sains*, vol. 14, no. 3, 2011.

[6] C. B. of Statistics, *Village Potential Statistics*. JAKARTA: BPS, 2021. [Online]. Available: http://www.bps.go.id/

[7] Joseph F. Hair, *Multivariate Data Analysis*, vol. 7. 2010. doi: 10.3390/polym12123016.

[8] N. Nasir and S. Sukmawati, “Analysis of Research Data Quantitative and Qualitative,” *Edumaspul J. Pendidik.*, vol. 7, no. 1, pp. 368–373, 2023.

[9] D. Marlina, N. F. Putri, A. Fernando, and A. Ramadhan, “Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak,” *J. CoreIT*, vol. 4, no. 2, p. 64, 2018.

[10] A. Aryandani, N. Solimun, A. A. Rinaldo Fernandes, and A. Efendi, “Implementation of FCM in Investor Group in the Stock Market Post-Covid-19 Pandemic,” *WSEAS Trans. Math.*, vol. 21, pp. 415–423, 2022.

[11] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[12] M. Zhang, “Use density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify galaxy cluster members,” in *IOP conference series: earth and environmental science*, 2019, vol. 252, no. 4, p. 42033.

[13] B. N. Sari and A. Primajaya, “Penerapan clustering DBSCAN untuk pertanian padi di kabupaten Karawang,” *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 28–34, 2019.

[14] B. S. Ashari, S. C. Otniel, and R. Rianto, "Perbandingan Kinerja K-Means dengan DBSCAN untuk Metode Clustering data Penjualan Online Retail," *J. Siliwangi Seri Sains dan Teknol.*, vol. 5, no. 2, pp. 64–67, 2019.

[15] N. Nurhaliza and M. Mustakim, "Pengelompokan Data Kasus Covid-19 di Dunia Menggunakan Algoritma DBSCAN: Clustering of Data Covid-19 Cases in the World Using DBSCAN Algorithms," *Indones. J. Inform. Res. Softw. Eng.*, vol. 1, no. 1, pp. 1–8, 2021.

[16] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]," *IEEE Trans. Automat. Contr.*, vol. 42, no. 10, pp. 1482–1484, 1997.

[17] S. A. Mingoti and J. O. Lima, "Comparing SOM neural network with FCM, K-means and traditional hierarchical clustering algorithms," *Eur. J. Oper. Res.*, vol. 174, no. 3, pp. 1742–1759, 2006.