BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1605–1616
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1605-1616

PERFORMANCE COMPARISON OF K-MEDOIDS AND DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE USING SILHOUETTE COEFFICIENT TEST

Taufiq Akbar[1], Georgina Maria Tinungki*[2], Siswanto[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University
Perintis Kemerdekaan Street KM.10, Makassar, 90245, Indonesia.

*Corresponding author's e-mail: *georgina@unhas.ac.id*

ABSTRACT

Article History:

Received: 13th April 2023
Revised: 9th August 2023
Accepted: 17th August 2023

Keywords:

Cluster analysis:
DBSCAN;
K-Medoids;
Silhouette coefficient.

Cluster analysis is a technique for grouping objects in a database based on their similar characteristics. The grouping results are said to be good if each cluster is homogeneous, and can be validated using the silhouette coefficient test. However, the presence of outliers in the data can affect the grouping results, so methods that are robust to outliers are used, such as K-Medoids and Density-Based Spatial Clustering of Applications with Noise. The purpose of this study is to compare the results and performance of the two methods using the silhouette coefficient test on data on human development indicators in South Sulawesi Province in 2021. The results of the analysis show that K-Medoids produced 2 groups, namely the districts/cities group which has indicators of human development that consist of 21 districts/cities, and the high group, which consists of 3 districts/cities, while Density-Based Spatial Clustering of Application with Noise produces 1 group that has the same characteristics, which consists of 19 districts/cities, and the remaining 5 districts/cities are identified as noise. Based on the silhouette coefficient test, K-Medoids have a greater value than Density-Based Spatial Clustering of Application with Noise, namely 0,635 and 0,544, respectively, so that K-Medoids have better performance.

How to cite this article:

T. Akbar, G. M. Tinungki and Siswanto., "PERFORMANCE COMPARISON OF K-MEDOIDS AND DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE USING SILHOUETTE COEFFICIENT TEST," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1605-1616, September, 2023.

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

The Human Development Index (HDI) is an index that explains how residents in a region can access the results of a development to obtain education, health and income. According to [1], there are 4 indicators in measuring HDI, namely Life Expectancy (LE), Average Length of Studying (ALS), Expectation Length of Studying (ELS) and Adjusted Per Capita Expenditures [1]. Because the measure of the quality of human development is statistically measured through the HDI, the increase in HDI points in an area is the most important thing for measuring the quality of life of people, including in South Sulawesi Province. For this reason, in order to assist the local government's performance in increasing HDI points in South Sulawesi Province, districts/cities are first grouped based on their indicators through cluster analysis. In cluster analysis, there are two methods, namely hierarchical and non-hierarchical methods. According to [2], the hierarchical method is a method that groups objects that have proximity, then proceeds to the proximity of other objects to form a level cluster (hierarchy) between objects. Whereas in the non-hierarchical method, the number of clusters that will be formed is determined before the cluster process.

Density Based Spatial Clustering of Application with Noise (DBSCAN) and K-Medoids are robust cluster analysis methods for outliers in data. DBSCAN uses a hierarchical approach in determining k clusters by using the density of the data as a basis for grouping objects in the data. DBSCAN has two parameters, namely epsilon and minPoints, which are references for determining objects that are included in density or noise [3]. K-Medoids is also a robust method against outliers, but does not follow a hierarchical process in determining k clusters. K-Medoids uses representative objects or medoids as the center of the clusters that are formed [4]. Measurement of grouping results can be done using the silhouette coefficient test. Silhouette coefficient values that are closer to 1 indicate better grouping results, and values that are less than or equal to 0.25 are considered bad grouping results [5]. Therefore, the silhouette coefficient test can be a reference for comparing the performances of DBSCAN and K-Medoids.

This study aims to compare results and performance using the silhouette coefficient test of the two methods, namely DBSCAN and K-Medoids. There have been many studies on the K-Medoids and DBSCAN methods, as was done by [6] which compared the DBSCAN and K-Means methods using the Silhouette Coefficient value. The results showed that DBSCAN was better at classifying village status in Central Java in 2020. Research related to K-Medoids have also been studied by [7] in formulating the problem in his research, namely comparing the Centroid Linkage and K-Medoids methods in grouping districts/cities in South Sulawesi based on educational indicators and this comparison uses the standard deviation value which has the smallest ratio. The results show that there is no difference between Centroid Linkage and K-Medoids because they produce the same standard deviation ratio values. What distinguishes this research from previous studies is that this research compares two robust methods for grouping objects in the data and uses transformation using the Principal Component Analysis (PCA) method to overcome multicollinearity in the data. This research is also expected to be able to provide benefits, namely providing insight into robust method solutions for outliers in cluster analysis.

2. RESEARCH METHODS

2.1 Data Sources

The data used in this study is data on human development indicators for South Sulawesi Province in 2021, namely secondary data taken from the Central Statistics Agency website, www.bps.go.id. Data on human development indicators consist of 24 districts/cities and four variables, namely Life Expectancy (LE), Average Length of Studying (ALS), Expectation Length of Studying (ELS) and Adjusted Per Capita Expenditures.

2.2 Cluster Analysis

Cluster analysis is an analytical technique for determining a group of several individuals or objects based on the similarity between one object and another. In general, there are at least 3 stages in cluster analysis, namely calculating how close objects are to each other, then proceeding to the grouping process stage, and finally describing each group formed [8].

There are two ways to measure how close or far an object is from another object. The first way is to use a measure of association, which involves calculating a correlation coefficient. A higher positive correlation coefficient indicates greater similarity between the objects. The second way is to use proximity or closeness between each pair of objects to assess similarity. When measures of distance or difference are used, a smaller distance or difference represents greater similarity between the objects [8]. One of the distance measures used to see the proximity between objects is the Euclidean distance, which is calculated through **Equation (1)** [2] as follows:

$$d(i,j) = \sqrt{\sum_{p=1}^{n} \left(x_{ip} - x_{jp}\right)^2} \qquad (1)$$

where n is the number of variables, $d(i,j)$ is the distance between the-*i* object and the-*j* object, x_{ip} is the data from data the-*i* subject and the-*p* variable, x_{jp} is the data from the-*j* subject dan the-*p* variable.

According to [8], multicollinearity or correlation between variables is a violation of assumptions in cluster analysis. Multicollinearity can be interpreted as a relationship or correlation between variables. One solution that can be applied if a violation of the multicollinearity assumption occurs is to eliminate the correlation between variables by applying the Principal Component Analysis (PCA) method [8].

Principal Components Analysis (PCA) can be used to reduce the data set to a smaller dimension while providing that the least amount of information is lost and to provide a better centroid point for clustering [9]. Basically, the PCA method reduces the number of variables by transforming these old variables by eliminating the correlation between variables so as to produce new variables that are not correlated with each other, which are referred to as Principal Components (PC). The dimensionality of the original high-dimensional data is reduced through linear mapping [10]. The linear combination of KU that is formed against the original variables can be written as **Equation (2)** as follows [11]:

$$\mathrm{PC}_j = e_{1j}Z_1 + e_{2j}Z_2 + \cdots + e_{pj}Z_p;\ j = 1,2,\ldots,p \qquad (2)$$

where PC_j is the-*j* Principal Component, Z_p is the standardized *p*-variable, e_{pj} is eigenvector of the *p*-variable normalized the-*j* Principal Component.

2.3 K-Medoids

K-Medoids or commonly called the Partitioning Around Method (PAM) is a cluster analysis method that is included in a non-hierarchical approach that has been determined previously for as many as k clusters and then continues to determine representative objects (medoids) in each k. The principle is to minimize the number of dissimilarities between each object and a representative object [4]. The K-Medoids algorithm is designed to identify the medoids in a cluster, which is the center point of the cluster. Compared to K-Means, K-Medoids is more robust because it selects k representative objects that minimizes the sum of dissimilarities between data objects, while K-Means uses the sum of squared Euclidean distances between data objects [12].

According to [13], the steps in the analysis of the K-Medoids method are as follows:

1. Determine the number of *k* clusters and determine randomly representative objects (medoids) as many as *k*.
2. Calculating the Euclidean distance between each object and representative objects (medoids) using the **Equation (3)**.

$$d\left(x_{ip}, o_{mp}\right) = \sqrt{\sum_{p=1}^{n} \left(x_{ip} - o_{mp}\right)^2} \qquad (3)$$

where:

$d\left(x_{ip}, o_{mp}\right)$: the distance between the *i*-th object *x* in the *p*-variable and the *m*-th object medoid of the *p*-variable.

x_{ip} : the *i-th* object *x* the *p*-variable.

o_{mp} : the *m*-th object medoid the *p*-variable.

3. Identify each object into an appropriate cluster based on its shortest distance from the medoids.

4. Calculating the objective function, namely the sum of the shortest distances from the medoids for each object.
5. Choose k random objects that are not representative (non-medoids).
6. Calculating the Euclidean distance between each object and objects that are not representative (non-medoids) through the **Equation (4)**.

$$d(x_{ip}, o_{hp}) = \sqrt{\sum_{p=1}^{n} (x_{ip} - o_{hp})^2} \qquad (4)$$

where:

$d(x_{ip}, o_{hp})$: the distance between the *i*-th object x in the *p*-variable and the *h*-th object non-medoid of the *p*-variable.

o_{hp} : the *h*-th non-medoid object the *p*-variable.

7. Determine whether each object belongs in an appropriate cluster based on its shortest distance from the non-medoids and calculate the objective function for the non-medoids.
8. Changing medoids to non-medoids is representative if the value of the medoids objective function > non-medoids objective function, whereas if the value of the medoids objective function < non-medoids objective function, then what is changed is non-medoids.
9. Perform steps 5-8 until the medoids remain unchanged.

2.4 Density Based Spatial Clustering of Application with Noise

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm for clustering based on density that generates clusters with arbitrary shapes. Density, in this context, refers to the quantity of points found within a designated radius **[14]**. DBSCAN is one of the cluster analysis methods that builds areas based on density connected. The DBSCAN model applies a simple minimum density level estimate, which is based on a minPoints threshold for the number of neighbors within the radius epsilon (ϵ) (with an arbitrary distance measure) **[15]**. Epsilon (ϵ) in DBSCAN represents the value of the threshold between neighbors which is considered to be density, and minPoints is the number of objects contained within the epsilon(ϵ) radius. In the concept of density, there are three kinds of states for each object: the object that is the core, the object that is the border, and the object that is noise **[16]**.

According to **[17]**, there are several common terms in DBSCAN, which are as follows:

a. $N_\epsilon(p^*)$: for object $p^* \in X$, the neighborhood ϵ is defined as $N_\epsilon(p^*)$ $\{x \in X | d(x, p^*) < \epsilon\}$, where $\epsilon \in \mathbb{R}^+$, X is the dataset and d is the distance function.
b. Core point : an object $p^* \in X$ is a core point if $|N_\epsilon(p^*)| \geq minPoints$ with $minPoints \in \mathbb{Z}^+$.
c. Directly reachable density : an object $p^* \in N_\epsilon$ is the directly attainable density of $x \in X$ if it satisfies $p^* \in N_\epsilon(x)$, where x is the core point object.
d. Density reachable : an object $y \in X$ is the attainable density of $x \in X$ if there is a chain of objects $p_1^*, p_2^*, \ldots, p_n^*$ with $y = p_1^*$ and $x = p^*_n$ and each $p_n^* \neq y$ is directly reachable density of p_{n-1}^*.
e. Density connected : an object $p^* \in X$ is the density connected to $q^* \in X$ if there is an object $x \in X$ so p^*, q^* is the reachable density of x.
f. Cluster : C is a group of X if $C \subset X$ and for each $p^*, q^* \in C, p^*$ and $q^*$are density connected.
g. Border point : object p^* is the border point, if $p^* \in C$ and p^* are not core points
h. Noise point : p^* is an outlier object if $p^* \notin C$.

To find the optimal epsilon value, it can be determined with the K-Nearest Neighbor (KNN) algorithm by plotting the KNN distance between each object from the nearest neighbor in the order and looking at the knee of the curve. The purpose of determining the angle of the KNN curve is so that objects located in a cluster have a small KNN distance, while objects considered noise have a large KNN distance [16]. The KNN algorithm is a classification algorithm based on the closest distance from one object to another. According to [18], the steps of KNN in determining the optimal epsilon parameters are as follows:

1. Determine the number of k nearest neighbors.
2. Calculate the distance between objects using the Equation (1).
3. Calculate the distance to the k-nearest neighbors.
4. Sort from smallest to largest, taking the calculated distance for each object.
5. Create a k-dist curve based on the comparison between the ordered distances and the sorted objects.
6. Observing the point of occurrence of an elbow or a critical distance change from the k-dist curve and taking that distance as the epsilon parameter.

Meanwhile, in determining the minPoints parameter, according to [16], at least it is the number of variables from a dataset plus one. According to [19], the DBSCAN method algorithm is as follows:

1. Determine epsilon parameter dan minPoints.
2. Randomly selects p^* object.
3. Calculating the distance between objects p^* with other x objects with the Equation (5).

$$d(x_{ij}, p_j^*) = \sqrt{\sum_{j=1}^{n} (x_{ij} - p_j^*)^2} \qquad (5)$$

where:

$d(x_{ij}, p_j^*)$: the distance between the i-th object in the j-variable and the object p^* in the j-variable.

x_{ij} : the i-th object in the j-variable.

p_j^* : object p^* in the j-th variable.

4. Define p^* objects including core points with the following conditions:

$$|N_\epsilon(p^*)| \geq minPoints$$

5. Repeat steps 2-4 until all objects have been processed.
6. Forming cluster C with 2 density conditions, namely reachable density and connected density.

2.5 Silhouette Coefficient

Measuring how good or bad the results of the clusters that are formed are evaluated. The evaluation was carried out using the silhouette coefficient method, with the silhouette coefficient calculation steps as follows [20]:

1. Calculate the average distance from an object, for example, the i-th object with all other objects that are in one cluster with the Equation (6).

$$a(i) = \frac{1}{n_A - 1} \sum_{j \in A, j \neq i}^{n_A} d(i,j) \qquad (6)$$

where n_A the number of objects in one cluster A.

2. Calculate the average distance from the i-th object to all objects in other clusters, for example C are all clusters formed other than A, then can be calculated through Equation (7) as follows:

$$b(i) = \min_{C \neq A} \left(\frac{1}{n_C} \sum_{j \in C}^{n_C} d(i,j) \right) \qquad (7)$$

where n_C the number of objects in one cluster C.

3. Calculate the silhouette value with the following Equation (8).

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i); b(i)\big)} \qquad (8)$$

4. Calculating the Silhouette Coefficient (SC) value with the Equation (9).

$$SC = \frac{1}{n} \sum_{i=1}^{n} s(i) \qquad (9)$$

with SC values in the interval $-1 \leq SC \leq 1$.

The SC value can describe how closely the objects are similar in the cluster. If the value of SC is close to 1, the clustering results will be better. Otherwise, if the SC value is close to -1, then the clustering results will be worse [5].

Table 1. Silhouette Coefficient Criteria

SC	Criteria
$\mathbf{0.7 < SC \leq 1}$	Strong Structure
$\mathbf{0.5 < SC \leq 0.7}$	Good Structure
$\mathbf{0.25 < SC \leq 0.5}$	Weak Structure
$\mathbf{0.25 \leq SC}$	Unstructured

Table 1 shows that the SC values can be interpreted as the results of the clusters formed into 4 structural categories, namely categories with strong structures in the interval 0,71-1,00, good or reasonable structures in the interval 0,51-0,70, weak structures in the interval 0,26-0,50, and unstructured is in the interval ≤ 0,25.

3. RESULTS AND DISCUSSION

3.1 Data Preparation

The human development indicator data for South Sulawesi Province in 2021 is indicator data that measures the level of welfare of human development in districts/cities through 4 dimensions of measurement: life expectancy (X_1), expectation length of studying (X_2), average length of studying (X_3), and adjusted per capita spending (X_4). Descriptively, the data on human development indicators in South Sulawesi in 2021 can be presented in Table 2.

Table 2. Descriptive Statistics on Human Development Indicator Data in South Sulawesi Province in 2021

Statistics	Variable			
	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$
Minimum	66.49	12.05	6.60	7434
Quartile 1	67.94	12.82	7.81	9505
Median	69.79	13.15	7.97	10995
Mean	69.58	13.30	8.27	11016
Quartile 3	70.70	13.61	8.39	12017
Maximum	73.41	15.58	11.43	17097

The first step in data preparation is to detect any outliers. The presence of outliers in the cluster analysis can lead to inaccurate clustering results. To find out whether there are outliers in the data, identification of outliers is carried out through descriptive and hypothesis testing. To see more clearly whether the data contains outliers or not, a box plot is presented visually, as shown in Figure 1.

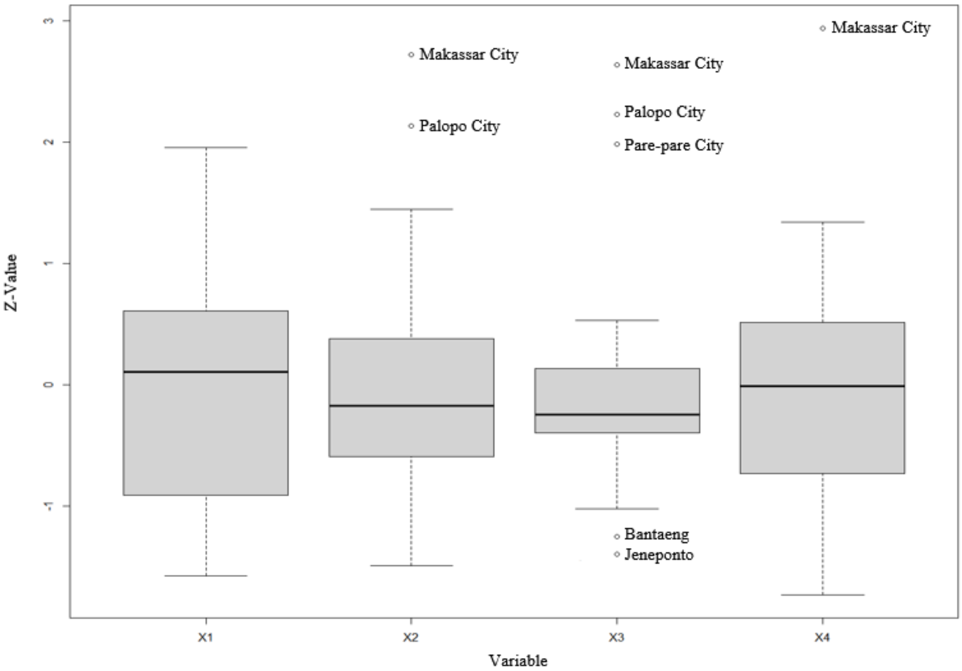


Figure 1. Box plot of human development indicator data in South Sulawesi in 2021.

The Regencies/cities included in the outlier in **Figure 1**, are objects contained in the X_2, X_3, and X_4 variables. Each regency/city that is included as an outlier, namely Makassar City and Palopo City on the X_2 variable; Jeneponto, Bantaeng, Pare-Pare City, Palopo City, and Makassar City on the X_3 variable and Makassar City on the X_4 variable. If there are outliers in the data, if further analyzed using a cluster analysis approach, it is better to use methods that are robust against outliers, such as K-Medoids and DBSCAN. Next is to transform the data using the Principal Component Analysis (PCA) method, which is used to overcome multicollinearity by forming uncorrelated Principal Components (PC). The number of KU formed is determined based on the cumulative proportion of the cumulative diversity of the PC variables, which ranges from 80%-90% with as many as 2 PC formed.

3.2 Cluster Process

3.2.1 K-Medoids

The k value, or optimum number of clusters, can be determined through the silhouette coefficient test by taking the Silhouette Coefficient (SC) value, which is the value obtained from the largest average silhouette index of the number of clusters that may be formed. The number of k formed is taken with the largest SC average value based on **Figure 2**, which is 2 with a SC value of 0,635.

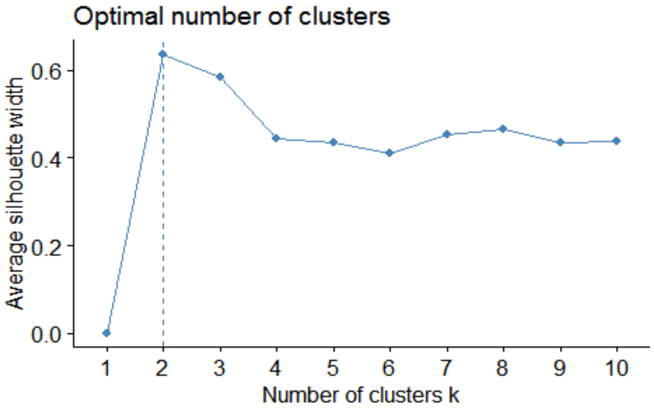


Figure 2. Selection of the optimum number of clusters in the K-Medoids method is based on the silhouette coefficient value.

Based on the selected representative objects, the results of clusters from the K-Medoids method with k = 2 can be presented in **Figure 3**.

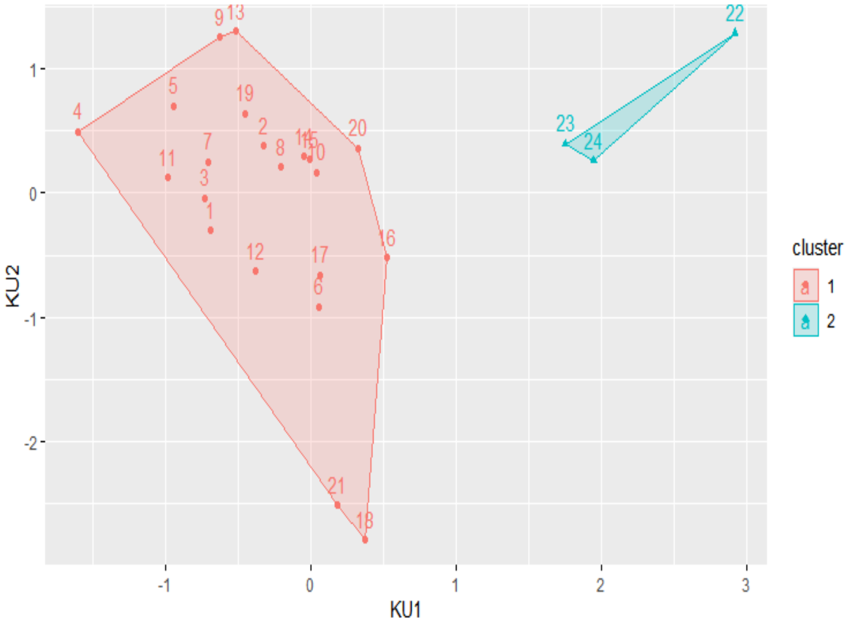


Figure 3. **Cluster Results on the K-Medoids Method.**

Figure 3. shown that the number of clusters formed by the K-Medoids method is as many as two, with the number of members in cluster 1 being 21 districts/cities and the number of members in cluster 2 being 3 districts/cities. The grouping of this method can be roughly seen, namely groups of districts/cities that are urban areas (cluster 2) and those that are not urban areas (cluster 1). The two clusters formed, namely cluster 1 and cluster 2, are presented descriptively in **Table 3** and **Table 4** respectively to see the characteristics of each cluster.

Table 3. **Descriptive Characteristics of Cluster 1 with the K-Medoids Method**

Statistics	**Variable**			
	X_1	X_2	X_3	X_4
Minimum	66.49	12.05	6.60	7434
Quartile 1	67.48	12.77	7.78	9504
Median	69.07	13.06	7.92	10632
Mean	69.31	13.05	7.88	10495
Quartile 3	70.45	13.39	8.24	11736
Maximum	73.41	13.86	8.91	12886

Table 4. **Descriptive Characteristics of Cluster 2 with the K-Medoids Method**

Statistics	**Variable**			
	X_1	X_2	X_3	X_4
Minimum	70.92	14.51	10.65	13117
Quartile 1	71.11	14.80	10.79	13452
Median	71.31	15.09	10.94	13786
Mean	71.45	15.06	11.01	14667
Quartile 3	71.72	15.34	11.19	15442
Maximum	72.13	15.58	11.43	17097

Each variable from the characteristics of cluster 2 in **Table 4** has an average value higher than the average value of the characteristics of cluster 1 in **Table 3**, so that in this case, cluster 2 can be interpreted as a district/city that has high human development indicators, while cluster 1 can be interpreted as districts/cities that have moderate human development indicators. The grouping results of this method can also be seen roughly, namely the group of districts/cities that belong to urban areas and those that are not urban areas, which means that the quality of life of people in urban areas is better than people who live in non-urban areas.

3.2.2 Density Based Spatial Clustering of Application with Noise

The first step of the DBSCAN method is to determine the epsilon and minPoints values. The optimum epsilon parameter value can be determined through the distance of the nearest neighbor or K-Nearest Neighbor (KNN) through the visualization of the KNN plot, which is made based on the closest k distance of each sorted object. If the 5 nearest neighbors are selected, then based on the calculation results of the distance to the 5 nearest neighbors, it can be presented in **Figure 4.**

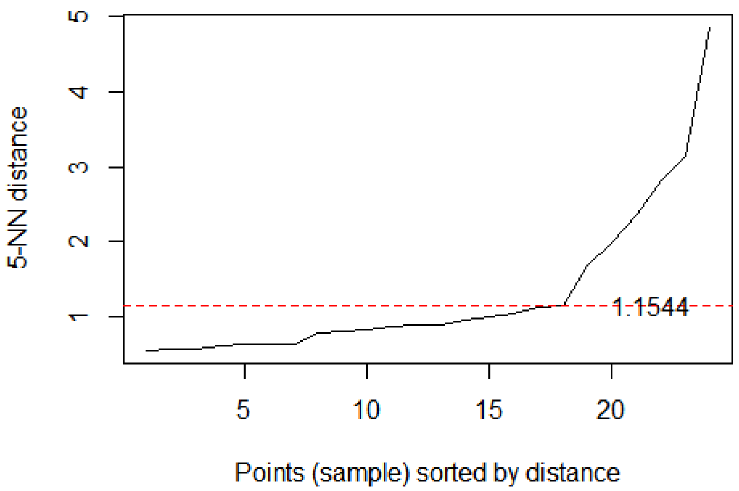


Figure 4. KNN plots of ordered distances.

Figure 4, the KNN plot shows no significant change in distance after the distance is 1,1544, so that distance is used as the value of the epsilon parameter. Meanwhile, the minPoints parameter is set at 3, because the number of variables involved in the analysis is 2 plus 1. After that, the cluster process is carried out, with the cluster results shown in **Figure 5**.

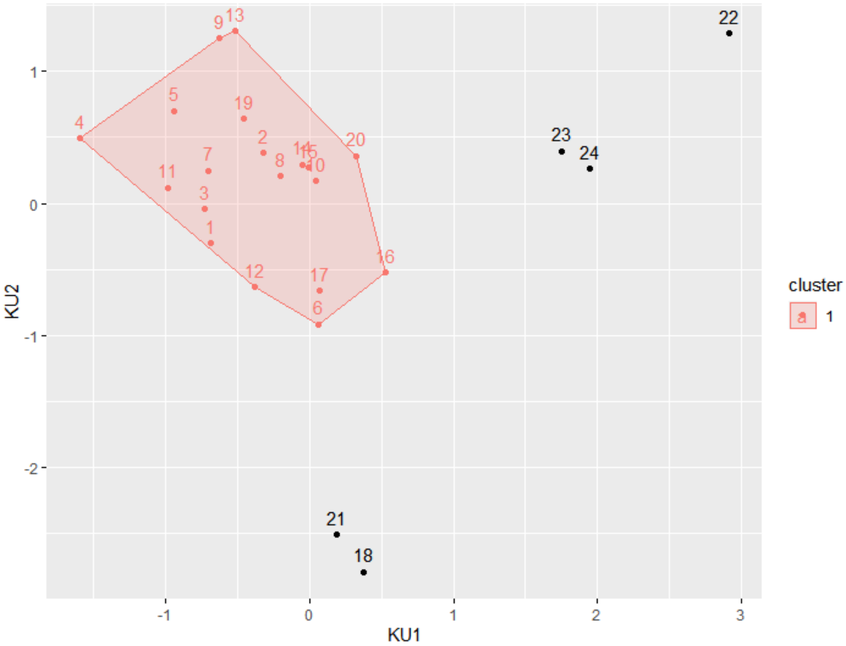
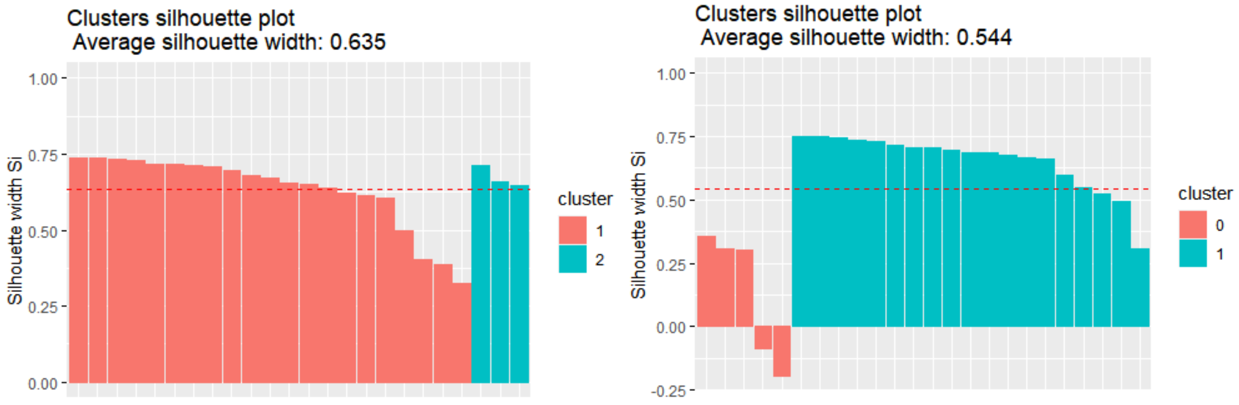


Figure 5. Cluster results on the DBSCAN method.

Figure 5 shown that the number of clusters formed by the DBSCAN method is one, with the number of members in cluster 1 being 19 districts/cities that have the same characteristics, while the remaining 5 districts/cities are identified as noise or districts/cities that do not belong to any cluster in the data.

3.3 Cluster Performance Evaluation

After the clustering process for each method is carried out, the results of each cluster method need to be evaluated using the Silhouette Coefficient (SC) test. The goal is to find out how good the structure of a cluster that has been built is. To see a comparison of the performance of the K-Medoids and DBSCAN methods, a silhouette coefficient test was carried out for each method, whose calculation results can be seen in Figure 6.



(a) (b)

Figure 6. Silhouette coefficient value on the method cluster results, (a) K-Medoids, (b) DBSCAN.

Based on Table 1, the Silhouette Coefficient Criteria, it can be said that the cluster structure produced by the K-Medoids and DBSCAN methods meets the criteria for a good structure because the SC values are at intervals of 0,5 and 0,7, but K-Medoids has a silhouette coefficient value that is larger than DBSCAN, so in this case the K-Medoids method has better performance than DBSCAN.

4. CONCLUSIONS

Based on the characteristics of descriptive statistics, there are 2 clusters formed by the K-Medoids method, which are categorized as districts/cities with moderate and high development indicators. Of the 24 districts/cities in South Sulawesi Province, only 3 districts/cities are included in the high category, namely Makassar City, Pare-pare City, and Palopo City; the rest of the districts/cities are included in the moderate category. While the groups formed by the DBSCAN method are as many as 1 cluster and some noise. The districts/cities identified as producing noise are Tana Toraja, North Toraja, Makassar City, Pare-pare City, and Palopo City, while the rest belong to a cluster that has the same characteristics. The K-Medoids method has better performance in grouping data on human development indicators in South Sulawesi Province in 2021 compared to DBSCAN because it has a greater Silhouette Coefficient (SC) value than DBSCAN, namely 0,635 and 0,544, respectively. If the SC value of the cluster results is close to 1, the better the structure of the cluster.

REFERENCES

[1] Badan Pusat Statistik Provinsi Sulawesi Selatan, *Indeks Pembangunan Manusia Provinsi Sulawesi Selatan 2021*. Sulawesi Selatan: Badan Pusat Statistik Provinsi Sulawesi Selatan, 2021.

[2] N. Arsih, N. Hajarisman, and S. Darwis, “Metode Pengclusteran Berbasis Densitas Menggunakan Algoritma DBSCAN,” *Prosiding Statistika*, pp. 153–163, 2016.

[3] I. Daqiqil Id, “Modifikasi DBSCAN (Density-Based Spatial Clustering with Noise) Pada Objek 3 Dimensi,” *Jurnal Komputer Terapan* , vol. 3, no. 1, pp. 41–52, May 2017, [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/article/view/954

[4] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Penerbit INFORMATIKA, 2017.

[5] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, “Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering,” *Techno. Com*, vol. 20, no. 2, pp. 186–197, 2021.

[6] M. M. Putri, C. Dewi, E. Permata Siam, G. Asri Wijayanti, N. Aulia, and R. Nooraeni, “Comparison of DBSCAN and K-Means Clustering for Grouping the Village Status in Central Java 2020 Komparasi DBSCAN dan K-Means Clustering pada Pengelompokan Status Desa di Jawa Tengah Tahun 2020,” vol. 17, no. 3, pp. 394–404, 2021.

[7] N. A. Raja, “Implemetasi Algoritma Centroid Linkage dan K-Medoids dalam Mengelompokkan Kabupaten/Kota di Sulawesi Selatan berdasarkan Indikator Pendidikan,” Universitas Hasanuddin, Sulawesi Selatan, 2020.

[8] J. F. Hair Jr, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis (Seventh Edition)*. 2014.

[9] C. Zhu, C. U. Idemudia, and W. Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,” *Inform Med Unlocked*, vol. 17, 2019.

[10] M. A. Bari and W. B. Kindzierski, “Ambient volatile organic compounds (VOCs) in Calgary, Alberta: Sources and screening health risk assessment,” *Science of the Total Environment*, vol. 631, pp. 627–640, 2018.

[11] R. Remesan and J. Mathew, *Hydrological data driven modelling: A case study approach*. 2015.

[12] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," in *Physics Procedia*, 2016.

[13] Athifaturrofifah, R. Goejantoro, and D. Yuniarti, "Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Studi Kasus: Data Titik Panas di Indonesia Pada 28 April 2018)," *Jurnal EKSPONENSIAL*, vol. 10, no. 2, 2019.

[14] G. Pu, L. Wang, J. Shen, and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," *Tsinghua Sci Technol*, vol. 26, no. 2, pp. 146–153, 2021.

[15] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[16] M. Hahsler, M. Piekenbrock, and D. Doran, "Dbscan: Fast density-based clustering with R," *J Stat Softw*, vol. 91, 2019.

[17] I. Cordova and T. S. Moh, "DBSCAN on Resilient Distributed Datasets," in *Proceedings of the 2015 International Conference on High Performance Computing and Simulation, HPCS 2015*, 2015.

[18] N. Rahmah and I. S. Sitanggang, "Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra," in *IOP Conference Series: Earth and Environmental Science*, 2016.

[19] A. I. Khurun'in, "Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Tingkat Sebaran Pengangguran Menggunakan Metode Density Based Spatial Clustering Algorithm with Noise (DBSCAN)," UIN Sunan Ampel Surabaya, Surabaya, 2021.

[20] H. Řezanková, "Different approaches to the silhouette coefficient calculation in cluster evaluation," in *21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics*, 2018, pp. 1–10.