STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. xxx–xxx, https://doi.org/10.59139/stattrans-2025-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Clustering based on poverty indicator data using K-Means cluster with Density-Based Spatial Clustering of Application with Noise

Sapriadi Rasyid[1], **Siswanto Siswanto**[2], **Sitti Sahriman**[3]

Abstract

The Indonesian government has implemented poverty alleviation programs, including assistance programs for the poor. Despite these efforts, the number of impoverished individuals in South Sulawesi continues to rise. To address this issue, a statistical method is necessary to cluster the poor based on error indicators for each region, serving as a reference for providing assistance. The appropriate statistical method is cluster analysis by minimizing object differences within one cluster and maximizing object differences between clusters. This study employs two methods, namely K-Means and Density-Based Spatial Clustering of Application with Noise (DBSCAN), to compare their effectiveness based on the Silhouette Coefficient. The data used for the analysis included eight poverty indicators for the South Sulawesi province in 2022. The K-Means method yielded two optimal clusters, with cluster 1 comprised of 23 regencies and cities, and cluster 2 only of Makassar City. The results of further analysis on cluster 1 consisted of eight new clusters and produced a Silhouette Coefficient of 0.507. In contrast, the DBSCAN method yielded one cluster, that encompassed 23 regencies and cities, with Makassar City identified as noise. The results of the further analysis on the clusters consisted of one cluster with three noises and produced a Silhouette Coefficient of 0.318. The study concludes that K-Means provides a higher Silhouette Coefficient and a more accurate representation of poverty clusters in South Sulawesi, which renders it a more effective tool for targeted poverty alleviation efforts.

Key words: Cluster, DBSCAN, poverty, K-Means, Silhouette Coefficient.

1. Introduction

Cluster analysis is a statistical method that can be used to cluster several objects into a class. Cluster analysis aims to maximize the similarity among objects within

[1] Department of Statistics, Hasanuddin University, Indonesia. E-mail: sapriadirasyid@gmail.com. ORCID: https://orcid.org/0000-0006-2972-7125.

[2] Corresponding author. Department of Statistics, Hasanuddin University, Indonesia. E-mail: siswanto@unhas.ac.id. ORCID: https://orcid.org/0000-0003-1934-5343.

[3] Department of Statistics, Hasanuddin University, Indonesia. E-mail: sittisahriman@unhas.ac.id. ORCID: https://orcid.org/0000-0002-9614-7132.

a cluster while minimizing the similarity between objects in different clusters (Pramana et al., 2018). However, cluster analysis assumes no multicollinearity (Hair et al., 2010). To address multicollinearity, principal component analysis (PCA) is necessary to reduce data dimensions into mutually independent principal components (Jhonson & Wichern, 2018). Thus, a combination of PCA is required for more optimal clustering results (Granato et al., 2018). Cluster analysis consists of various algorithms, two of which are K-Means and density-based spatial clustering of application with noise (DBSCAN).

K-Means is a method which needs an appropriate number of clusters denoted as k. K-Means is susceptible to noise (Huang et al., 2023). In contrast, DBSCAN is a method which clusters data based on distance density, thus identifying noise (Jing et al., 2010). Density, in this context, refers to the quantity of points found within a designated radius (Pu et al., 2021). But, DBSCAN cannot determine the number of clusters. By using the same data, the procedure and results of K-Means with DBSCAN will be different. According to Dewi & Pramita (2019), the quality of cluster analysis can be measured using the Silhouette Coefficient. Therefore, the Silhouette Coefficient test can be a reference for comparing of K-Means and DBSCAN result. This method can help the government in identifying poverty indicators in each region. This is important because the number of impoverished individuals in South Sulawesi increased by 16.86 thousand people in September 2022 compared to the previous year (BPS, 2023). By using the appropriate clustering method, the government can provide more effective and targeted assistance based on relevant poverty indicators.

Research on K-Means with the DBSCAN method has been carried out by many researchers, such as research conducted by Rais et al. (2021), which optimized K-Means clustering with PCA, resulting in two principal components and two clusters with a small Davies-Bouldin Index indicating good clustering results. Meanwhile, research on DBSCAN was performed using K-Nearest Neighbor to determine the epsilon parameter. This study used four variables and the result obtained 5 noises and produced one cluster consisting of 19 object (Akbar et al., 2021). What distinguishes this research from previous studies is that this research combines PCA to obtain comparison results of K-Means with DBSCAN using the Silhouette Coefficient on poverty indicator data of South Sulawesi Province. Besides that, determining epsilon in DBSCAN is done based on hierarchy principles. Based on this case, this study is focused on obtaining clustering results based on poverty indicators data for each region as a reference providing assistance. Results of this

research can, as hoped, assist the government in addressing poverty cases in each region in South Sulawesi.

2. Methodology

2.1. Data and Research Variables

This research is quantitative in nature and covers 24 regencies and cities in the South Sulawesi Province, using secondary data sourced from the Central Statistics Agency (Badan Pusat Statistik) of South Sulawesi in 2022, which is available on the official website *sulsel.bps.go.id*. The research variables consist of eight poverty indicators following Rais et al. (2021), including the human development index (X_1), population size (X_2), labor force (X_3), percentage of poor population (X_4), labor force participation (X_5), unemployment rate (X_6), population density (X_7), and per capita expenditure (X_8). These variables were selected because they are commonly used in poverty research and are considered to be significant factors in determining the socio-economic conditions of a region.

2.2. Cluster Analysis

Cluster analysis is an effort to identify groups of similarity of data objects within one cluster while minimizing similarity to other clusters. In general, clustering algorithms can be divided into different categories, such as partitional, hierarchical, and density-based (Cardeiro de Amorim & Makarenkov, 2023). Furthermore, cluster analysis methods require a measure of dissimilarity or distance. Typically, the distance often used is the Euclidean distance as defined in Equation (1) below.

$$\mathbf{d_{p,q}} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{m}} \left(\mathbf{x_{ip}} - \mathbf{x_{iq}}\right)^2} \tag{1}$$

with: $d_{p,q}$ is the distance between object p and q, x_{ip} is the i variable of object p, x_{iq} is the i variable of q , and m is the number of variables.

2.3. Assumption of Multicollinearity

Cluster analysis generally assumes that the data to be analyzed do not exhibit strong correlations between two or more variables with other variables. This is because strong correlations can lead to multicollinearity (Hair et al., 2010). In regression analysis, the Variance Inflation Factor (VIF) is used to identify the presence of multicollinearity among independent variables. Generally, multicollinearity is considered significant when VIF > 10, indicating a strong linear correlation between variables in the model

(Salmerón et al., 2020). According to Hair et al., (2010), one way to identify multicollinearity is by calculating the Variance Inflation Factor (VIF) value, which is formulated based on Equation (2) as follows:

$$\mathbf{VIF_i} = \frac{\mathbf{1}}{\mathbf{Tolerance}} = \frac{\mathbf{1}}{\mathbf{1 - R_i^2}} \tag{2}$$

with: R_i^2 is coefficient of determination from regressing variable i with others.

2.4. Principal Component Analysis

Principal Component Analysis (PCA) is an effective statistical technique for addressing multicollinearity by reducing data dimensions and the elimination of highly correlated variable, because PCA is useful for extracting data to be new variables (Kurita, 2019). These variables are called Principal Components (PC) (Festa et al., 2023). PCA can be employed for data preprocessing before applying complex statistical methods (Kherif & Latypova, 2019). In PCA, PC are new variables obtained through a linear combination of the original variables. Before transforming the data into *Principal Components*, standardization is necessary to ensure that all variables have a uniform scale. Suppose there is a set of original variables denoted as $X = (X_1, X_2, \dots, X_m)$, the standardization process is then applied to transform these variables into standardized variables $Z = (Z_1, Z_2, \dots, Z_m)$. This standardization aims to eliminate differences in scale among variables, ensuring that the analysis remains accurate and is not affected by variations in measurement units. Meanwhile, the values of Z are obtained according to Equation (3) below (González et al., 2022).

$$\mathbf{Z_i} = \frac{\mathbf{X_i - \mu_i}}{\sqrt{\boldsymbol{\sigma_i^2}}} \tag{3}$$

with: X_i is the value of variable i on the object, μ_i is the mean value of variable i, σ_i^2 is the variance of variable i, and Z_i is the standardized value of i (Bari & Kindzierski, 2018). Therefore, Equation (4) can be obtained as follows:

$$\mathbf{PC_f} = \mathbf{b_{11}Z_1} + \cdots + \mathbf{b_{mg}Z_m} \tag{4}$$

with: PC_f is the f^{th} principal component, b_{mg} is the eigenvalue of m on the g^{th} principal component. Meanwhile, the eigenvalue can be obtained using the following Equation (5):

$$\mathbf{|A - \lambda I| = 0} \tag{5}$$

Equation (5) yields characteristic roots λ_l such that $\lambda_1 > \lambda_2 > \cdots > \lambda_v$, so each characteristic λ_l depends on the value of $\mathbf{b}$ (Astutik et al., 2018). To determine the

principal components, one can follow the criteria by weighting the cumulative proportions as described in Equation (6) below (Abdi & Williams, 2010):

$$\frac{\sum_{l=1}^{u} \lambda_l}{\sum_{l=1}^{v} \lambda_l} > 0.80 \; ; u \leq v \tag{6}$$

with: $\sum_{l=1}^{u} \lambda_l$ is total variance of the first u principal component and $\sum_{l=1}^{v} \lambda_l$ is total variance.

2.5. K-Means

K-Means is a type of the non-hierarchical cluster analysis techniques as it is susceptible to the selection of the initial clustering center (Liu et al., 2023). In simple terms, the K-Means algorithm can be performed as follows (Huang et al., 2023):

1. Determine *k* as the number of clusters to be formed.
2. Randomly initialize *k* parameters, which are the initial centroids of the clusters.
3. Calculate the distance of each data point to each selected centroid using Equation (1). Each data point chooses the nearest centroid.
4. Calculate the mean value of the data points that chose the same centroid. This value becomes the new centroid.
5. Repeat step 3 and 4 if the position of the new centroid and the old centroid is not the same. However, if the positions of the new and old centroids are the same, the clustering process is considered complete.

2.6. Density-Based Spatial Clustering of Application with Noise

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is often considered the most popular density-based clustering algorithm (Chowdhury et al., 2023). DBSCAN is a clustering method based on the concept of density, which can be determined by a single density condition. The following are terms in DBSCAN: Epsilon is the DBSCAN radius used to determine density connectivity. As core points, the number of points within their neighborhood must be greater than or equal to the minimum points (Jing et al., 2019). Point p is density-reachable from a point q if there is a chain of points $p_1, p_2, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{1+1} is direct density-reachable from p_1 (Zhang et al., 2022).

In simple terms, the DBSCAN algorithm requires two parameters: epsilon and minimum points (Starczewski & Cader, 2019). When determining the minimum points, according to Hahsler et al. (2019), it is typically at least the number of variables in the analyzed dataset plus one. This approach aims to adjust the parameter according to the complexity of the data dimensions to ensure that each formed cluster has a sufficiently significant density. However, DBSCAN has evolved by creating

a representation without requiring an epsilon radius by incorporating hierarchical clustering processes within the density concept (Stewart & Al-Khassaweneh, 2022). The DBSCAN algorithm is as follows (Nurhaliza & Mustakim, 2021):

1. Initialize the minimum points parameter.
2. Choose a random starting point, p.
3. Calculate all point distances using Equation (1) for density reachability with respect to p. If a point satisfies the core point condition, the number of points within its neighborhood is equal to or greater than the minimum points parameter, it forms a cluster. If p is a border point and no points are density-reachable from p, move to the next point.
4. Repeat 2-4 step for each observed point until all objects are identified.

2.7. Silhouette Coefficient

The *Silhouette Coefficient* (SC) is a widely used metric for evaluating clustering performance by assessing the compactness and separation of clusters. It measures how well each data point fits within its assigned cluster compared to other clusters (Řezanková, 2019). Moreover, SC is used to evaluate the clustering results as per the following Equation (7):

$$\mathbf{SC} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{p=1}}^{\mathbf{n}} \frac{\mathbf{b(p)} - \mathbf{a(p)}}{\mathbf{max(a(p); b(p))}} \tag{7}$$

The *Silhouette Coefficient* (SC) is a widely used metric for evaluating clustering performance by assessing the compactness and separation of clusters. It measures how well each data point fits within its assigned cluster compared to other clusters

with: $a(p)$ is the average distance of object p's characteristics to all objects within the same cluster, and $b(p)$ is the average distance of object p's characteristics to all objects within a different cluster (Batool & Hening, 2021).

SC has an interval range of $-1 \leq SC \leq 1$, and the evaluation criteria for the SC method can be shown in Table 1 as follows (Batool & Hening, 2021):

Table 1: Evaluation Criteria for SC Method

Interval SC Score	Interpretation	Interval SC Score	Interpretation
$0.70 < SC \leq 1.00$	Strong Structure	$0.25 < SC \leq 0.50$	Weak Structure
$0.50 < SC \leq 0.70$	Medium Structure	$SC \leq 0.25$	No Structure

Source: Batool & Hening, 2021.

3. Result and Discussion

3.1. Multicollinearity Test

Cluster analysis must satisfy the assumption of non-multicollinearity to ensure that the weights of each variable are balanced. Therefore, VIF calculations are performed based on Equation (2), which can be presented in Table 2 as follows:

Table 2: VIF Score Each Variable

Variable	VIF	Description	Variable	VIF	Description
X_1	4.152	Not Significant	X_5	3.031	Not Significant
X_2	128.993	Significant	X_6	34.406	Significant
X_3	130.826	Significant	X_7	19.184	Significant
X_4	2.202	Not Significant	X_8	4.097	Not Significant

Source: data processed.

Based on Table 2 show the VIF values indicate that for variables $X_2 = 128.993 > 10$, $X_3 = 130.826 > 10$, $X_6 = 34.406 > 10$, and $X_7 = 19.184 > 10$, and the general criterion that multicollinearity is considered significant when the VIF > 10. Therefore, it is concluded that there is multicollinearity in the data. The presence of multicollinearity leads to imbalanced weights in the analysis results, making the presented information inaccurate, such as in the calculation of distances between objects. The solution to this problem is to use PCA.

3.2. Principal Component Analysis

The data on poverty indicators in South Sulawesi Province have different units of measurement. Therefore, to determine the Principal Components (PC), a correlation matrix is used. Additionally, differences in units of measurement can result in inconsistent cluster analysis results. The solution to this problem is to transform the data into the same units of measurement using the PCA according to Equation (3). The number of PC formed is based on the cumulative diversity proportion of the PC variables, which should be at least around 80%. The calculation of the cumulative diversity proportion is calculated according to Equation (6), and the results of the cumulative diversity proportion calculation are shown in Table 3 as follows:

Table 3: Cumulative Diversity Proportion Each PC

Principal Component	λ	Cumulative Diversity Proportion	Principal Component	λ	Cumulative Diversity Proportion
PC_1	4.778	59.723%	PC_5	0.191	97.906%
PC_2	**1.686**	**80.797%**	PC_6	0.144	99.710%
PC_3	0.767	90.383%	PC_7	0.019	99.952%
PC_4	0.411	95.519%	PC_8	0.004	100.00%

Source: data processed.

The cumulative diversity proportion in Table 3 show value greater than 80% indicates that these two principal components capture most of the information contained in the original variables, allowing the analysis to proceed with just these two components without significant loss of information. Hence, it can be concluded that PC_1 and PC_2 meet the criteria for forming two principal components that explain a total variance of 80.797% in the original variables. The principal components are as follows according to Equation (4):

$$\mathbf{PC_1 = 0.349Z_1 - 0.374Z_2 - 0.368Z_3 - 0.255Z_4 - 0.203Z_5 + 0.431Z_6 + 0.427Z_7 + 0.359Z_8} \quad (8)$$

$$\mathbf{PC_2 = 0.284Z_1 - 0.377Z_2 - 0.407Z_3 - 0.461Z_4 - 0.502Z_5 + 0.207Z_6 + 0.110Z_7 - 0.301Z_8} \quad (9)$$

Equation (8) is formed from an eigenvalue of 4.778, while Equation (9) is formed from an eigenvalue of 1.686. So, Equations (8) and (9) become the new variables that will be further analyzed using both K-Means cluster and DBSCAN. The results of this research can provide information that K-Means clustering and DBSCAN can be combined with PCA to create new mutually independent variables called PC.

3.3. K-Means

The initial step in K-Means is to determine the optimal k value or the number of clusters, as shown in Table 4 below:

Table 4: SC Score Each k

k	SC Score	k	SC Score	k	SC Score	k	SC Score
1	0.000	4	0.400	7	0.279	10	0.459
2	0.772	5	0.382	8	0.411	11	0.431
3	0.436	6	0.426	9	0.483	12	0.313

Source: data processed.

Based on Table 4, the best SC is obtained when $k = 2$. The next step is to determine the initial centroids randomly. Given that the selection of centroids is random, this study adopts an approach by selecting the minimum and maximum values of each Principal Component (PC). Specifically, the first centroid is determined based on the minimum value of each PC, while the second centroid is determined based on the maximum value of each PC, as shown in Table 5 below:

Table 5: Initial Centroid

k	o_1	o_2
1	-1.958	-2.486
2	9.439	1.990

Source: data processed.

The next step is to calculate the distance between objects based on Equation (1) for each data point based on the selected initial centroids. Each data point selects the nearest centroid. After that, the new centroids are determined by calculating the average of the data points that selected the same centroid, as shown in Table 6 below:

Table 6: Centroid of the second iteration

k	o_1	o_2
1	-0.410	-0.060
2	9.439	1.371

Source: data processed.

Based on Table 6, it is found that the closest distance values have not changed, indicating that the cluster iteration process is stopped, meaning that the clustering results have been obtained. The result is that cluster 1 consists only of Makassar City, while the other 23 cities and regencies are in cluster 2. This indicates that Makassar City has data characteristics in the poverty indicators that are significantly different from the others. Therefore, analysis continues with clustering without including Makassar City. The analysis is carried out with the same procedure, resulting in an optimal number of clusters of eight. The results of K-Means cluster can be presented in Figure 1 below:

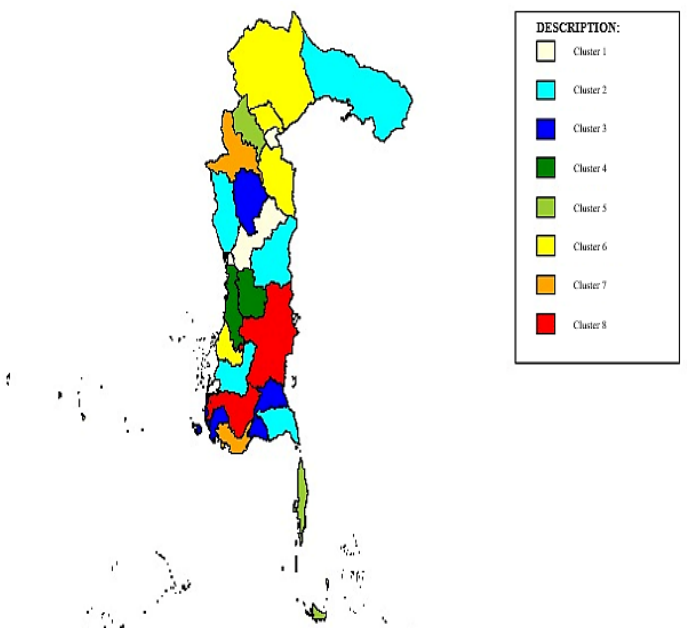

Figure 1: Further clustering results using K-Means

Source: data processed.

Based on Figure 1, it can be concluded that further clustering using K-Means cluster produce eight clusters, with cluster 1 consisting of Sidrap, Palopo city, and Pare-Pare city. Cluster 2 consists of Bulukumba, Maros, Wajo, Pinrang, and Luwu Timur. Cluster 3 consists of Takalar, Sinjai, Bantaeng, and Enrekang. Cluster 4 consists of Barru and Soppeng. Cluster 5 consists of Kepulauan Selayar and Toraja Utara. Cluster 6 consists of Luwu, Luwu Utara, and Pangkep. Cluster 7 consists of Tana Toraja and Jeneponto. Cluster 8 consists of Gowa and Bone.

3.4. Density-Based Spatial Clustering of Application with Noise

The initial step in DBSCAN is to determine the minimum points based on the number of analyzed variables, which in this case is two PC plus one, resulting in a minimum of three points. The determination of the hierarchy can be done by gradually selecting object p^*. The initial object p^* is the one with the smallest epsilon when the minimum points are three. Based on the data, the initial p^* object is Luwu Timur. Then, the next p^* has the second smallest epsilon, which is 0.437, in the case of Wajo and Maros. This hierarchy results in Wajo as the first new core point, having border points: Maros, Luwu Timur, and Pinrang. Meanwhile, Maros, as the second new core point, has border points: Wajo, Luwu Timur, and Bulukumba. The next step is to determine the clusters formed based on the principles of density achieved (reachable) and connected density (connectivity). Therefore, the direct arrived density for each core point in the second hierarchy is:

$p_{20}^* = \{p_8, p_{13}, p_{15}\}$;
$p_{13}^* = \{p_8, p_{20}, p_{15}\}$;
$p_8^* = \{p_{20}, p_{13}, p_2\}$.

Object $p_{20}^*, p_{13}^*, p_8^*$ are core points, and if $\exists\ p_{20}^*$ is not a member of p_8^* and vice versa $\in p_{20}^* \cup p_8^*$, then p_{20}^* and p_8^* are considered as density reached. Based on the connectivity principle, if p_{20}^* and p_8^* are considered density reached, then all members of p_{20}^* and p_8^* are considered as connected density. The same process in determining clusters is repeated until all objects have been determined. Thus, based on the concept of the second hierarchy, it still forms one cluster with Luwu Timur Regency, Maros Regency, and Wajo Regency as core points. Meanwhile, the border points consist of Bulukumba and Pinrang. Additionally, 19 other regencies and cities are considered as noise.

The same procedure is repeated to determine core points by choosing the smallest third object and so on, carried out step by step based on the reachable density and

connectivity principles until all objects have been declared as core points. The cluster extraction process is further illustrated in Figure 2 below:

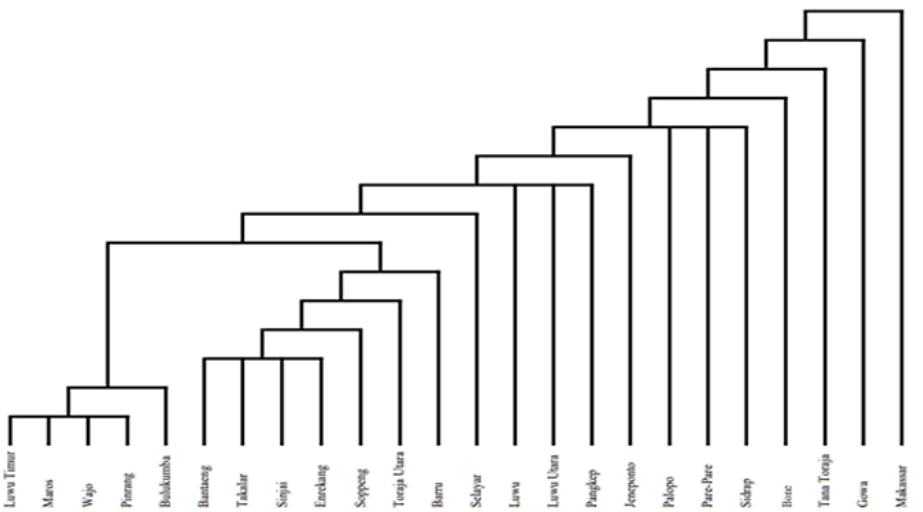

Figure 2: Extraction of DBSCAN Based on Hierarchy

Source: data processed.

Based on Figure 2, the SC values for each hierarchy level (denoted as h) can be shown in Table 7 below:

Table 7: SC Score Each Hierarchy Level

h	SC Score	h	SC Score	h	SC Score	h	SC Score
1	0.000	4	0.521	7	0.310	10	0.027
2	0.772	5	0.472	8	0.202	11	0.009
3	0.627	6	0.341	9	0.179	12	-0.015

Source: data processed.

Based on Table 7, the best hierarchy level is level 2, which results in one noise point, namely Makassar city, while the other 23 districts and cities form a single cluster. This indicates that Kota Makassar has different poverty indicator characteristics compared to the others. Therefore, the analysis continues with clustering without including Kota Makassar. The analysis is performed using the same procedure, the best hierarchy at the fourth level with one cluster and three noise points consisting of Bone, Gowa, and Tana Toraja. So, the results of this research can provide information that determining

epsilon in DBSCAN can be represented through the hierarchical principle, whereas the DBSCAN results in this research can be presented in Figure 3 below:

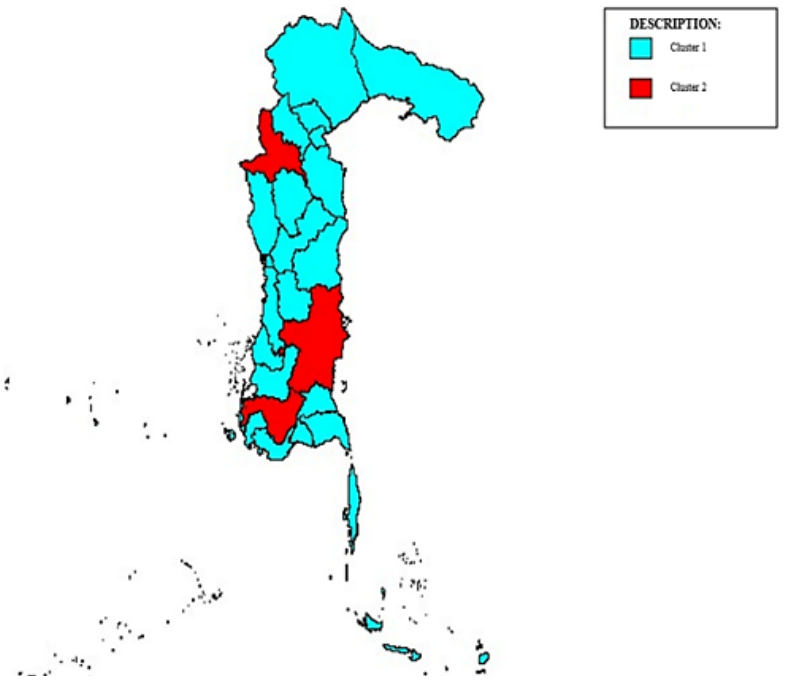

Figure 3. Further clustering results using DBSCAN

Source: data processed.

3.5. Silhouette Coefficient

The next step after the clustering process of each method is to evaluate the clustering results obtained with the Silhouette Coefficient according to Equation (7). The purpose is to assess how well the clusters that have been constructed perform. The comparison of the K-Means and DBSCAN hierarchical-based methods using the Silhouette Coefficient without Makassar City can be presented in Table 8 as follows:

Table 8. Comparison of SC Values for K-Means and DBSCAN

Method	SC Value
K-Means	0.507
DBSCAN	0.318

Source: data processed.

Based on Table 8, referring to Table 1, it can be concluded that K-Means achieves a good clustering result because it falls within the range of $0.50 < SC \leq 0.70$. On the other hand, DBSCAN obtains a weak clustering result as it falls within the range of $0.25 < SC \leq 0.50$.

3.6. Profiling Cluster

Based on the clustering results, the best method is K-Means. The profiles and explanations of the characteristics of each cluster are as follows:

a. Cluster 1 consists of Sidrap, Kota Palopo, and Kota Pare-Pare. This cluster has the highest population density. However, it also has the highest Human Development Index (IPM), per capita expenditure, and relatively low poverty indicators, making it the most prosperous cluster among the others.
b. Cluster 2 consists of Bulukumba, Maros, Wajo, Pinrang, and Luwu Timur. This cluster has a relatively high IPM and per capita expenditure, as well as relatively low poverty indicators, categorizing it as a prosperous cluster.
c. Cluster 3 consists of Takalar, Sinjai, Bantaeng, and Enrekang. This cluster has relatively low IPM and per capita expenditure. It also has relatively low poverty indicators, making it moderately prosperous.
d. Cluster 4 consists of Barru and Soppeng. This cluster has relatively low IPM and per capita expenditure compared to other clusters. It also has the lowest average percentage of poor population and relatively low poverty indicators, categorizing it as prosperous based on these indicators.
e. Cluster 5 consists of Kepulauan Selayar and Toraja Utara. Despite having the lowest population, workforce, and unemployment figures, this cluster has a relatively low IPM and per capita expenditure compared to other clusters, categorizing it as relatively poor.
f. Cluster 6 consists of Luwu, Luwu Utara, and Pangkep. This cluster has the highest percentage of poor population and the second-highest workforce participation rate, categorizing it as a poor region.
g. Cluster 7 consists of Tana Toraja and Jeneponto. This cluster has the lowest IPM and per capita expenditure and the second-highest percentage of poor population and workforce participation rate, categorizing it as poor.
h. Cluster 8 consists of Gowa and Bone. This cluster has the highest population, workforce, and unemployment figures and the second-lowest IPM, categorizing it as a poor region.

Based on these characteristics, clusters 5, 6, 7, and 8 require special attention from the government based on the analysis. Meanwhile, cluster 1 is the most prosperous cluster. The results of this research are useful for the government to identify poverty indicators for each region in South Sulawesi.

4. Conclusions

This study obtained that K-Means cluster obtained the optimal number of clusters, which is two. Cluster 1 consists only of Kota Makassar, while cluster 2 consists of 23 districts and cities. Further analysis revealed eight optimal clusters. On the other hand, the Density-Based Spatial Clustering of Application with Noise produced one large cluster with 23 districts and cities, classifying Kota Makassar as noise. Further analysis at the best hierarchy level yielded three noise points, including Gowa, Bone, and Tana Toraja, while the other districts and cities belong to a single cluster. K-Means clustering provided more effective groupings of the poverty indicator data for South Sulawesi Province in 2022 compared to Density-Based Spatial Clustering of Application with Noise. This is evident in the higher Silhouette Coefficients obtained, with values of 0.507 and 0.318, respectively. K-Means clustering achieved better groupings, while Density-Based Spatial Clustering of Application with Noise resulted in weaker clustering. So, the study concludes that the K-Means method is more effective than DBSCAN in helping the government to group the poverty characteristics of each region so that it can overcome poverty cases in South Sulawesi Province.

Acknowledgement

Thanks to the Department of Statistics, Universitas Hasanuddin, along with my colleagues and academic staff, for their invaluable support and contributions in the preparation of this research article.

References

Abdi, H., Williams, L. J., (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433–459.

Akbar, T., Tinungki, G. M. and Siswanto, (2023). Performance of K-Medoids and Density Based Spatial Clustering of Application with Noise Using Silhouette Coefficient Test. *Barekeng: J. Math. & App*, 17(3), pp. 1605–1616.

Astutik, S., Solimun and Darmanto, (2018). Analisis Multivariat: Teori dan Aplikasinya dengan SAS. UB Press.

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

Batool, F., Hennig, C., (2021). Clustering with the Average Silhouette Width. *Computational Statistics and Data Analysis*, 158(107190), pp. 1–18.

BPS, (2023, March). *Profil Kemiskinan di Sulawesi Selatan*. *https://sulsel.bps.go.id*

Chowdhury, S., Helian, N. and Cordeiro de Amorim, R., (2023). Feature weighting in DBSCAN using reverse nearest neighbours. *Pattern Recognition*, 137(109314), pp. 1–15.

Cordeiro de Amorim, R., Makarenkov, V., (2023). On k-means iterations and Gaussian clusters. *Neurocomputing*, 553(126547), pp. 1–10.

Dewi, D. A. I. C., Pramita, D. A. K., (2019). Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Matrix*, 9(3), pp. 102–109.

Festa, D., Novellino, A., Hussain, E., Bateson, L., Casagli, N., Confuorto, P., Soldato, M. D. and Raspini, F., (2023). Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering. *International Journal of Applied Earth Observation and Geoinformation*, 118, pp. 1–13

González, C. A. D, Calderón, Y. M. M, Cruz, N. A. M and Sandoval, L. E. P., (2022). Typologies of Colombian off-grid localities using PCA and clustering analysis for a better understanding of their situation to meet SDG-7. *Cleaner Energy Systems*, *3*(100023), pp. 1–16.

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L. and Maggio, R. M., (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology,* 72, pp. 83–90.

Hahsler, M., Piekenbrock, M. and Doran, D., (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software,* 91, pp. 1–30.

Hair, J. F. J. R., Black, W. C., Babin, B. J. and Anderson, R. E., (2010). Multivariate Data Analysis (7th ed.). Pearson Education Inc.

Huang, Q., Chen, S. and Li, Y., (2023). Selection of seismic noise recording by K-means. *Case Studies in Construction Materials*, 19 (e02363), pp 1–16.

Jing, W., Zhao, C. and Jiang, C., (2019). An Improvement Method of DBSCAN Algorithm on Cloud Computing. *Procedia Computer Science*, 147, pp. 596–604.

Johnson, R. A., Wichern, D. W., (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Kherif, F., Latypova, A., (2019). Principal Component Analysis. *In Machine Learning: Methods and Applications to Brain Disorders*, pp. 209–225.

Kurita, T., (2019). Principal Component Analysis (PCA). *In Computer Vision: a Reference Guide*, pp. 1–4.

Liu, G., Ji, F., Sun, W. and Sun, L., (2023). Optimization design of short-circuit test platform for the distribution network of integrated power system based on improved K-means clustering. *Energy Reports*, 9, pp. 716–726.

Nurhaliza, N., Mustakim, (2021). Pengelompokan Data Kasus Covid-19 di Dunia Menggunakan Algoritma DBSCAN. *IJIRSE*, *1*(1), 1–8.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I. and Nooraeni, R., (2018). Data Mining Dengan R (Konsep Serta Implementasi). *In Media*.

Pu, G., Wang, L., Shen, J. and Dong, F., (2021). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Sci Technol*, 26(2), pp. 146–153.

Rais, M., Goejantoro, R. and Prangga, S., (2021). Optimalisasi K-Means Cluster dengan Principal Component Analysis pada Pengelompokan Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Tingkat Pengangguran Terbuka. *Jurnal Eksponensial*, 12(2), pp. 129–135.

Řezanková, H. A. N. A., (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. *In 21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics*, pp. 1–10.

Salmerón, R., García, C. B. and García, J., (2020). Variance Inflation Factor and Its Influence on Regression Models. *Journal of Statistical Computation and Simulation, 90*(12), pp. 1–15.

Starczewski, A., Cader, A., (2019). Determining the eps parameter of the DBSCAN algorithm. *In Artificial Intelligence and Soft Computing: 18th International Conference*, pp. 420–430.

Stewart, G., Al-Khassaweneh, M., (2022). An Implementation of the HDBSCAN Clustering Algorithm. *Applied Sciences*, 12(2405), pp. 1–21.

Zhang, R., Qiu, J., Guo, M., Cui, H. and Chen, X., (2022). An Adjusting Strategy after DBSCAN. *IFAC-PapersOnLine*, *55*(3), pp. 219–222.