Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 1, March 2025, pp. 26 – 40
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Algorithm Comparison of Hierarchical and Non-Hierarchical Clustering Method in Grouping Regional Poverty Variables

[1]Farhan Maulana, [2]Arie Wahyu Wijayanto
[1,2]Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia
Email: [1]222112043@stis.ac.id, [2]ariewahyu@stis.ac.id

Article Info

Article history:
Received Mar 31th, 2024
Revised May 19th, 2024
Accepted Sep 7th, 2024

Keyword:
Clustering Algorithms
Comparison
Data Mining
Hierarchical Clustering
Poverty Grouping

ABSTRACT

One of the objectives of the main Sustainable Development Goals (SDGs) is to end poverty in all forms. Although West Sumatera Province occupies ranking seventh lowest national in poverty, there is an increase amounting to 0.11 percent in September 2022 compared to March 2022. This shows the complexity of the poverty problem in the region. The Provincial Government needs to understand the poverty situation by grouping it based on characteristics in each region. This is a strategic step so that poverty reduction policies can be developed on target and efficiently according to the conditions of each region. This study aims to investigate Clustering methods, namely a non-hierarchical method represented by K-Means, Fuzzy C-means, and K-medoids also the hierarchical method, represented by Divisive Analysis (DIANA) and Agglomerative Nesting (AGNES) with complete linkage, average linkage, single linkage, and Ward's method, to group regencies/cities and compare the performance of the Clustering methods used, to get the best method using Davies Bouldin Index and Dunn index. The results of this research indicate that the divisive analysis method and agglomerative nesting, especially in complete linkage, single linkage, and Ward's method is the best Clustering method. This method works optimally when the number of clusters is equal to 3. It is hoped that our findings can support policies that are right on target and efficient in efforts to overcome poverty in West Sumatera.

Corresponding Author:
Arie Wahyu Wijayanto,
Departement of Statistical Computing,
Politeknik Statistika STIS,
Jakarta, Indonesia.
Email: ariewahyu@stis.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.29393

1. INTRODUCTION

One of the desired goals achieved in the Sustainable Development Goals (SDGs) is to end poverty in all forms, wherever it happens [1]. Poverty based on a draft from the Central Bureau of Statistics Indonesia (BPS), is the inability of somebody in a way economy to fulfill a need base like food and non-food needs are measured from side expenditure [2]. Poverty is still a problem that has not been completely resolved in Indonesia. Based on information from BPS, in September 2022, the percentage of poverty in Indonesia reached 9.57 percent, an increase of 0.03 percent compared to March 2022 [3]. West Sumatera Province in September 2022 reached the seventh lowest ranking nationally in terms of poverty percentage [3]. Even though West Sumatera Province ranked seventh lowest nationally in poverty percentage, there was an increase of 0.11 percent compared to March 2022 [3]. As a result, this province dropped from the sixth lowest ranking to the seventh lowest nationally. Even though it is faced with the challenge of increasing poverty, the West Sumatera Provincial Government remains committed to reducing the level of extreme poverty to reach zero percent by 2030 according to the SDGs target or even by 2024 according to the 2019-2024 RPJMN [4]. To achieve this goal, the West Sumatera Provincial Government needs to understand the

poverty situation in each region by grouping regencies/cities based on poverty characteristics. This is a strategic step so that poverty reduction policies can be developed on target and efficiently according to the conditions of each region.

Cluster analysis is an analytical method used to group observation objects into groups where the members have many similarities, but simultaneously have many differences from other groups [5]. In cluster analysis, there are two methods, namely the hierarchical method and the non-hierarchical method, also known as the partitioning method. The process of non-hierarchical clustering begins with determining the number of clusters and selecting the centroids first. [6]. Meanwhile, the Hierarchical method is carried out in stages or in a structured manner, without knowing in advance how many clusters will be formed [6]. In Hierarchical analysis, data grouping occurs by measuring the proximity distance between objects, which is then represented in the form of a dendrogram [7].

Previous research that has been carried out related to Clustering, namely Analyzing Groups Using Partitioning and Hierarchical Methods on Poverty Data in Indonesian Provinces in 2019, was carried out by Afira & Wijayanto [6]. In this research, the best method was obtained, namely the Hierarchical method with two clusters. Application of the Clustering Algorithm for Grouping Poverty Levels in Banten Province carried out by Munandar [8]. The best method in this research is K-Medoid with 3 clusters. Grouping of Regencies/Cities on the Island of Java Based on Poverty Factors Using the Average Linkage Approach Hierarchical Clustering was carried out by Wahyuni & Aryo [7]. Grouping Provinces in Indonesia Based on Poverty Levels Using Hierarchical Analysis Agglomerative Clustering was carried out by Widodo, et al. [9]. In this research, the best grouping method is the Ward method with 3 clusters. Cluster Analysis Based on Criminal Acts in Indonesia in 2019 was carried out by Simatupang & Wijayanto [10]. This study employs the K-Means and Fuzzy C-means method.

Previous studies have been carried out related to poverty. Research conducted by Dewi Puspita concluded that the Regency/ City Minimum Wage (UMK) and Life Expectancy Rate (AHH) had a negative and significant effect on poverty levels in Central Java Province and per capita expenditure had a positive and significant effect on poverty levels in Central Java Province [11]. Another study conducted by Nasution concluded that the consumer price index, life expectancy, local income, and literacy rates have a negative influence on the depth and severity of poverty in regencies/cities in Eastern Indonesia (KTI) and the Gini ratio and percentage of Poor people whose main jobs are in the informal sector have a positive influence on the depth and severity of poverty in KTI regencies/ cities [12]. Research conducted by Rabbani et al concluded that the factors influencing the percentage of extremely poor people in KTI during the 2010-2021 period were the percentage of people aged 15 years and over who worked in the agricultural sector, the open unemployment rate, the percentage of the illiterate population, the number school participation, average length of school, percentage of population who had health complaints during the last month, and percentage of households using electric lighting sources [13]. Research conducted by Ardian & Destanto concluded that the Human Development Index (HDI), which includes per capita income, Expected Years of Schooling, Average Years of Schooling, and Life Expectancy has a significant influence on poverty levels [14].

Based on the above research, the author aims to conduct a cluster analysis in regencies/cities in West Sumatera based on factors influencing poverty and compare several clustering methods to find the best one. This study introduces innovation by grouping regencies/cities in West Sumatra using various clustering methods. The methods used include partitioning methods such as K-Means, fuzzy c-means, and k-medoids, as well as hierarchical methods like Agglomerative Nested (AGNES) and Divisive Analysis (DIANA). The results of this research are expected to provide a clearer understanding of the characteristics of poverty in various regions of West Sumatra.

2. MATERIAL AND METHOD

2.1 Study Area and Data Sources

This research focuses on all regencies/cities in West Sumatera Province in 2022. The data used in this research is secondary data originating from the Central Statistics Agency (BPS) and all data types are numerical. The selection of variables is based on previous research on poverty which has been described in the background. Research variables used in the study This can seen in the variable list table study.

Table 1. List of Variables Research

Variable	Description
X1	Percentage Poor Population
X2	Expenditure Real Per capita
X3	Life expectancy
X4	Average Years of Schooling
X5	Long School Expectations
X6	Open Unemployment Rate

2.2 Preprocessing

Before carrying out data analysis, a preprocessing stage was carried out on the data used. First, check for missing values in the data. The second stage is checking data types. The final preprocessing stage is data standardization. Data standardization is the transformation of data where the data units are different [16]. The goal is to standardize the data so that the mean is zero and the standard deviation is equal to one [17]. The following is the formula for data standardization:

$$z = \frac{x_i - \bar{x}}{s} \quad (1)$$

Description: z is Standardization value for the i^{th} object in the j^{th} variable; $\bar{x}$ is Average value for the ith object; s is Standard deviation of the ith object; i is Object index (1,2, ..., N); and j is Object index (1,2, ..., p).

2.3 Assumption of sample adequacy

After carrying out the preprocessing stage, the data used can be analyzed. Next, assumptions are testing on the data. At this stage, data adequacy assumptions and non-multicollinearity assumptions were tested. Sample adequacy means that the sample used can describe the population. The test used to assume data adequacy is the Kaiser Meyer Olkin (KMO) Test. Data can be said to be sufficient if the KMO value is above 0.5 [18]. The following is the formula for calculating the KMO value:

$$KMO = \frac{\sum_u^n \sum_{j \neq u}^n r_{uj}^2}{\sum_u^n \sum_{j \neq u}^n r_{uj}^2 + \sum_u^n \sum_{j \neq u}^n a_{uj}^2} \quad (2)$$

Description: r_{uj} is Simple correlation coefficient of the u^{th} variable and the j-th variable; and a_{uj} is Partial correlation coefficient of the u^{th} variable and the j^{th} variable.

2.4 Multicollierity assumption

Multicollinearity is when independent variables are strongly correlated with each other. One way to determine the presence of multicollinearity is by examining the Variance Inflation Factor (VIF) values. If the VIF value is more than 10, it can be said that multicollinearity has occurred. The following is the formula for calculating the VIF value:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3)$$

Description: R_j^2 is Coefficient of determination of the j^{th} variable.

2.5 Determining the number of clusters

After testing the adequacy of data and checking for multicollinearity assumptions, the next step involves determining the optimal number of clusters (k) using the Silhouette Coefficient method. The Silhouette Coefficient method combines cohesion and separation approaches to assess the quality of objects within a cluster [19]. The optimal number of clusters is indicated by the high value of the Silhouette coefficient [20]. The following are the stages for calculating the Silhouette coefficient:

1. Calculating the average distance of data

$$a(i) = \frac{1}{|A|-1} \sum_{m \in A, m \neq 1} d(i, m) \quad (4)$$

Description: $a(i)$ is The difference in the mean value of object i with all other objects in group A; $d(i, m)$ is The distance between the i^{th} data object and m; and A is Clusters.

2. Calculate the average distance of data i from all data in other clusters.

$$d(i, C) = \frac{1}{|C|} \sum_{m \in C} d(i, m) \quad (5)$$

Description: $d(i, m)$ is The difference in the average distance value of the ith object to all other objects in C; and C is Other clusters besides cluster A

3. Choose the smallest value of d(i,C)

$$b(i) = \min_{C \neq A} d(i, C) \quad (6)$$

The cluster B, which achieves the minimum distance (i.e., d(i,B) = b(i)), is referred to as the neighbor of object i. This represents the second-best cluster for object i.

4. Calculate the Silhouette coefficient value

$$s(i) = \frac{b(i)-a(i)}{\max a(i),b(i)} \quad (7)$$

2.6 Cluster analysis

Cluster analysis is a way to identify patterns of similarity in data by grouping similar data into groups. There are two main methods in cluster analysis, namely partitioning and Hierarchy. The Hierarchical method organizes data objects in stages, starting from a single cluster to covering all individuals or vice versa. Meanwhile, the partitioning method divides data objects into several clusters without a Hierarchical structure, either predetermined or previously estimated [21].

2.7 Agglomerative Nesting (AGNES)

Agglomerative methods can be classified into two, namely graph methods and geometric methods. Grouping methods such as complete linkage, single linkage, and average linkage fall into the category of graph methods, while Ward's method is included in the category of geometric methods [22].

The steps of the agglomerative method are as follows [22]:

1. Start by forming N groups, each group consisting of one entity, and use a symmetric distance matrix **D** = { d_{ik}}.
2. Discover the distance matrix for pairs of groups that are nearest to each other. Determine the distance between the "most similar" groups U and V as d_{UV}.
3. Recombine clusters U and V, assigning the label (UV) to the merged group. Adjust the entries in the distance matrix by (a) eliminating rows and columns associated with clusters U and V, and (b) incorporating rows and columns indicating the distances between the (UV) cluster and the remaining clusters.

Repeat Steps 2 and 3 N-1 times. Once the algorithm is complete, all objects will be in one group. Note the identity of the group being merged and the degree of distance or similarity at which the merger occurred.

1. Single linkage: The distance between two groups is determined by measuring the shortest distance between an object in one group and an object in the other group [5]. In this method, in step number 3 of the agglomerative method, the distance between this cluster and another cluster W is calculated by:

$$d_{(UV)W} = \min \{d_{UW}, d_{VW}\} \quad (8)$$

Where, the distances d_{UW} and d_{VW} represent the shortest distances between group U and W, and between group V and W.

2. Complete linkage: The measurement of the range separating two clusters relies on the maximum distance between an item within one cluster and an item within another cluster [5]. Within the agglomerative method's third step, the distance from the cluster to the other W clusters is computed as follows:

$$d_{(UV)W} = \max \{d_{UW}, d_{VW}\} \quad (9)$$

Where, the distances d_{UW} and d_{VW} represent the farthest distances between group U and W, and between group V and W.

3. Average linkage: The range between two sets is determined by averaging the gaps between items in one set and items in the other set [5]. In this process, during step 3 of the agglomerative approach, the distance from the cluster to the other W clusters is computed as:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \quad (10)$$

Where d_{ik} represents the distance between an item i belonging to group (UV) and an item k belonging to group W, and N_{UV} and N_W are the respective quantities of item in cluster (UV) and W.

4. Ward's method: In this method, considerations in combining pairs of groups are those that produce the smallest increase in ESS [5]. In this method, third step of the agglomerative method, the distance between the cluster and other W clusters is calculated by:

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x})'(x_j - \bar{x}) \quad (11)$$

Where, x_j represents the measurement taken for the j^{th} item, while $\bar{x}$ denotes the average measurement across all items.

2.8 Divisive Analysis (DIANA)

The DIANA method is the opposite of the Agglomerative algorithm Hierarchical Clustering. At first, all objects are grouped in one large group. Next, in each stage, the largest group is divided into two groups, and this process continues until finally each group contains only one object [23].

The steps of the DIANA method are as follows [23]:

1. The DIANA approach follows a top-down pattern, assuming one initial group has level L(0) = n and sequence number m =0.
2. Look for the two least similar groups within the current group, denoted as (r) and (s), where the distance between (r) and (s), d[(r), (s)], is the minimum among all pairwise distances within the group.
3. The level of ordering increases as m increases by 1. The group is divided into groups (r) and (s) to form the next group with a new level of grouping: $L(m1) = d[(r)]$ and $L(m2) = d[(s)]$.
4. The distance matrix (D) gets refreshed with the incorporation of rows and columns representing groups (r) and (s). The similarity measure between the newly formed cluster, denoted as (r, s), and the previous cluster (k) is determined as follows:

$$D[(k),(r,s)] = \min d[(d),(r)], d[(k),(s)] \quad (12)$$

If all objects have become different clusters , the process is terminated; if not, go back to step 2.

2.9 K-Means

K-Means Clustering is a method of grouping that uses the initial values of centroid points to form clusters. This initial centroid value influences the next centroid value and the determination of the next cluster. If the previous cluster pattern is the same as the next cluster pattern, then the calculation is stopped. [24].

The steps of the K-Means method are as follows [5]:

1. Find out how many clusters (k) should be utilized.
2. Allocate data into clusters randomly
3. Find the center of each cluster using the equation provided by the data.

$$V_{kj} = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{N} \quad (13)$$

4. Calculate the distance from each object to each cluster center by computing the Euclidean distance between them.

$$d(X_i, X_g) = \sqrt{\sum_{j=1}^{p} (X_{ij} + X_{gj})^2} \quad (14)$$

5. Calculating the objective function

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} d(x_i, V_{kj})^2 \quad (15)$$

6. Assign each data point to the closest cluster center as formulated.

$$a_{ij} = \begin{cases} 1, s = \min\{d(x_i, V_{kj})\} \\ 0, \text{lainnya} \end{cases} \quad (16)$$

Repeat step 3 to step 6 until no moving objects or changes in objective function are found

2.10 Fuzzy C-Means

The Fuzzy C-Means method was introduced by Jim Bezdek in 1981. It is a technique for data grouping where each data point's membership level into a group is determined on a scale from 0 to 1, using Euclidean Distance. [25]. The advantage of Fuzzy C-Means lies in the accuracy of group center placement when compared to other methods [26].

The procedure for applying the Fuzzy c-means technique is outlined below. [29]:

1. Identify the number of clusters to use
2. Determining the rank/ weighting of the exponent (w)
3. Determining the maximum iteration (MaxIter)
4. Determine the stopping criteria (threshold) or ε = very small positive value
5. Create a U matrix as a starting point for determining the level of membership in a group (cluster), which functions as an initial partition matrix.

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \dots & \mu_{1i}(x_i) \\ \mu_{21}(x_1) & \mu_{21}(x_2) & \dots & \mu_{2i}(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{k1}(x_1) & \mu_{k2}(x_2) & \cdots & \mu_{ki}(x_j) \end{bmatrix} \quad (17)$$

6. Determine the center point V for every cluster.

$$V_{kj} = \frac{\sum_{k=1}^{I_k} (U_{ki})^w x_{ij}}{\sum_{k=1}^{I_k} (U_{ki})^w} \quad (18)$$

Description: V_{kj} is Centroid center point (average) cluster k^{th} j^{th} variable; U_{ki} is Degree of cluster membership k^{th} i^{th} object; x_{ij} is The value of the ith object in the cluster is for the j^{th} variable; I_k is The quantity of items belonging to the k^{th} cluster; k is Cluster index; j is Variable index; i is Object index; and w is Exponent weighting.

7. Correct the membership degree of each data in each cluster

$$U_{ki} = \left[\sum_{j=1}^{i} \left(\frac{D_{ki}}{D_{ji}} \right)^{\frac{2}{w-1}} \right]^{-1} \quad (19)$$

with:

$$D_{ki} = \sqrt{\left(\sum_{k=1}^{j} (x_{ij} - V_{kj})^2 \right)} \quad (20)$$

Description: D_{ki} is Euclidian distance clusters k^{th} object i; D_{ji} is Euclidian distance of the j^{th} variable to the ith object; j is Variable index; k is Cluster index; i is Object index; x_{ij} is The value of the i^{th} object that exists in the cluster is for the j^{th} variable; and V_{kj} is Center value (centroid/average) of the cluster k^{th} j^{th} variable

8. Determine the conditions for ending the iteration, which includes alterations in the partition matrix from the current iteration to the preceding one.

$$\Delta = |U^l - U^{l-1}| \quad (21)$$

Description: l is t^{th} iteration and U is Degree of membership

9. If $\Delta < \varepsilon$then the iteration is stopped. However, if not then repeat steps 6 to step 8

2.11 K-Medoids

The K-medoids or Partitioning Around Medoids (PAM) algorithm resembles K-Means, but it employs objects as representatives (medoids) for cluster centers, unlike K-Means which utilizes average values. K-Medoids offer an advantage in overcoming the sensitivity to noise and outliers, which is a weakness of K-Means. The clustering process of K-medoids is independent of the dataset's order, making it more stable and efficient, especially for small datasets. [28].

The process of the K-medoids technique unfolds in the following manner. [30]:

1. Begin by setting up the initial positions for k cluster centers (the number of clusters).
2. Allocate every data point or item to the closest cluster utilizing the Euclidean distance computation formula.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \tag{22}$$

3. Choose one object randomly from each cluster to be considered as a potential new medoid.
4. Find the distance from every object in every group to the new potential medoid.
5. Recompute the total deviation (S) by finding the difference between the current total distance and the previous total distance. If the outcome is negative, replace the objects with cluster data to establish a fresh group consisting of k objects as medoids.
6. Repeat steps third to fifth till the medoid is unchanged, thereby forming a cluster and enabling the identification of cluster members.

2.12 Determining the Best Clustering Algorithms

This research involves the use of evaluation methods such as the Davies Bouldin Index and Dunn Index to measure the quality of Clustering results. Davies Bouldin Index is an internal evaluation tool that measures the quality of grouping based on cohesion and separation [30,34]. The following is Davies' calculation formula Bouldin Index:

$$DB = \frac{1}{k}\sum_{k=1}^{k} \max_{h\neq z}\{\frac{d(X_h) + d(X_z)}{d(k_h,k_z)}\} \tag{23}$$

Description : k is Number of clusters; $d(X_h)$: Distance between cluster objects h to the cluster center; $d(X_z)$ is Distance between cluster objects h to the cluster center; and $d(k_h, k_z)$ is Distance between cluster centers h and z

Dunn index is a method that also uses cohesion and separation to calculate cluster validity. A higher index value indicates a superior selection of clusters [31,35]. The following is the Dunn index calculation formula [32,36]:

$$Dunn = \min_{1\leq h\leq k}\{\min\left\{\frac{d(k_h,k_z)}{\max_{1\leq c\leq k}(d(X_c))}\right\}\} \tag{24}$$

Description: k is Number of clusters; $d(k_h, k_z)$ is Distance between cluster objects hth and cluster objects z^{th} ; $d(X_c)$ is Distance between objects in the cluster c^{th}

3. RESULTS AND ANALYSIS

3.1 Descriptive Analysis

The following is a description of the poverty variables in West Sumatera Province based on the size of concentration and the size of the distribution.

Table 2. Description Statistics

Variable	Average	Minimum	Maximum
X1	5.91	2.28	13.97
X2	11011	6567	14889
X3	70.65	64.93	74.82
X4	9.39	7.48	11.92
X5	13.85	12.51	16.54
X6	5.28	1.39	11.69

Table 1 shows a statistical description of the six research variables. It can be seen that there are differences in the units of each variable, so it is necessary to standardize all variables. Based on the calculations that have been carried out, a KMO value of 0.72 is obtained. The value obtained is greater than 0.5. Hence, it can be concluded that the assumption of sample adequacy is met.

Table 3. Variance Inflation Factor Value

Variable	Variance Inflation Factor (VIF)
X1	2.696795
X2	4.824332
X3	3.597331
X4	6.845837
X5	5.210948
X6	2.952652

According to the information provided in Table 3, there are no VIF values exceeding 10 for any of the variables, indicating that the assumption of non-multicollinearity is satisfied.

3.2 Determining the number of clusters

According to Figures 1 and 2, employing the silhouette coefficient method reveals that the ideal number of clusters for both the non-Hierarchical and Hierarchical methods in this research is 3.

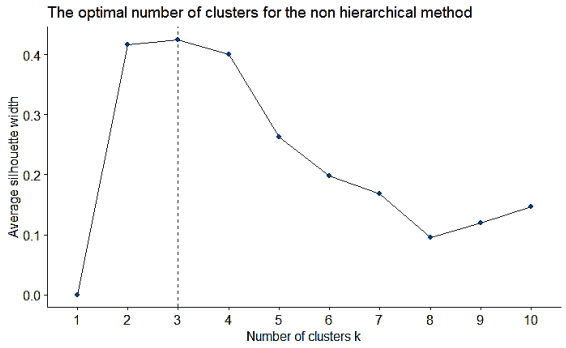

Figure 1. Determining the Number of Clusters for the Non- Hierarchical Method Using the Silhouette Coefficient Method

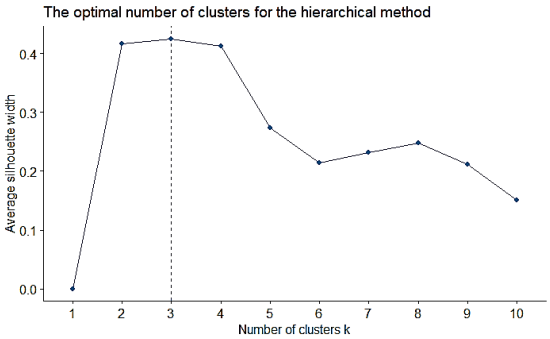

Figure 2 . Determining the Number of Clusters for the Hierarchy Method Using the Silhouette Coefficient Method

3.3 K-Means

The outcomes of the clustering conducted via the K-Means technique with three clusters can be observed in Figure 3.

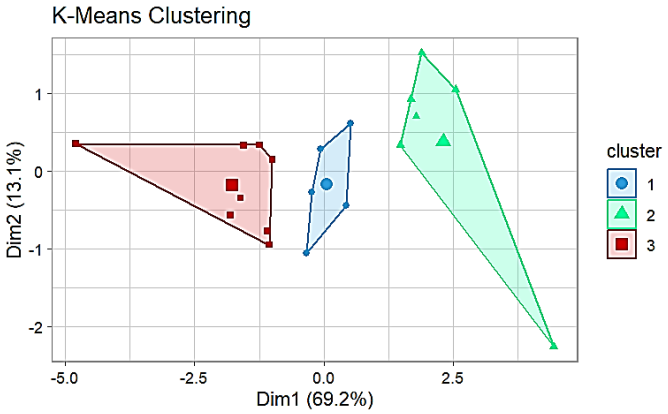

Figure 3. K- Means Cluster Plots

By using the K-Means method with 3 clusters, the clustering results were obtained as in Figure 3. Figure 3 shows that cluster 1 has 5 regencies/cities, cluster 2 has 6 regencies/cities, and cluster 3 has 8 regencies/cities. More complete information regarding cluster members can be seen in Table 4.

Table 4. Results of Grouping Regencies/ Cities Using the K-Means Method

Clusters	Num. of Member	Regency /City
1	5	Agam Regency, Sawahlunto City, Padang Pariaman Regency, Tanah Datar Regency, Dharmasraya Regency
2	6	Payakumbuh City, Solok City, Pariaman City, Bukittinggi City, Padang City, Padang Panjang City
3	8	Mentawai Island Regency, Solok Regency, Pesisir Selatan Regency, Sijunjung Regency, Pasaman Barat Regency, Solok Selatan Regency, Lima Puluh Kota Regency, Pasaman Regency

3.4 Fuzzy C-Means

The outcomes of clustering performed utilizing the fuzzy c-means technique with three clusters can be observed in Figure 4.

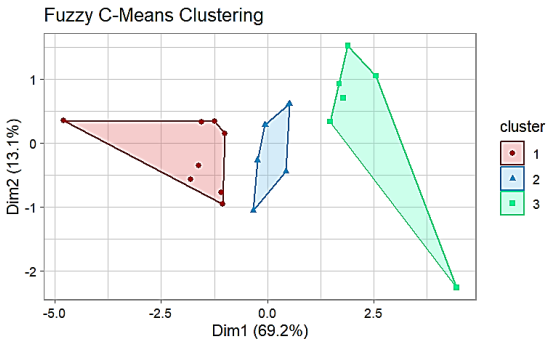

Figure 4. Fuzzy C-Means Cluster Plots

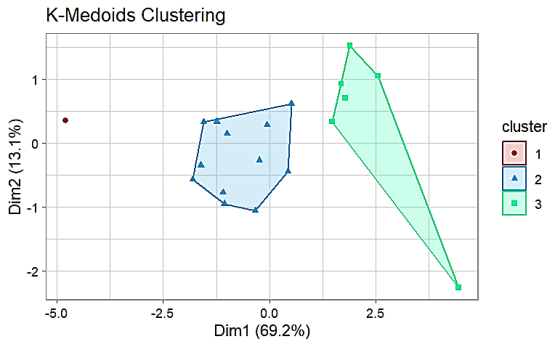

Figure 5 . K-Medoids Cluster Plots

Based on Figure 4, it is revealed that cluster 1 has 8 regencies/cities, cluster 2 has 5 regencies/cities, and cluster 3 has 6 regencies/cities. More detailed information regarding members of each cluster can be seen in Table 5.

Table 5. Results of Grouping Regencies/ Cities Using the Fuzzy C-Means Method

Clusters	Num. of Member	Regency/ City
1	8	Mentawai Island Regency, Solok Regency, Pesisir Selatan Regency, Sijunjung Regency, Pasaman Barat Regency, Solok Selatan Regency, Lima Puluh Kota Regency, Pasaman Regency
2	5	Agam Regency, Sawahlunto City, Padang Pariaman Regency, Tanah Datar Regency, Dharmasraya Regency
3	6	Payakumbuh City, Solok City, Pariaman City, Bukittinggi City, Padang City, Padang Panjang City

3.5 K-Medoids

The results of Clustering which have been carried out using the K-Medoids method with the number of clusters equal to 3, can be seen in Figure 5. Based on Figure 5, it can be seen that cluster 1 has one regency/City, cluster 2 has twelve regencies/cities, and cluster 3 has six regencies/cities. More specific details regarding the members of each cluster can be seen in Table 6.

Table 6. Results of Grouping Regencies/ Cities Using the K-Medoids Method

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	12	Padang Pariaman Regency, Pasaman Barat Regency, Solok Regency, Solok Selatan Regency, Sawahlunto City, Agam Regency, Pasaman Regency, Dharmasraya Regency, Lima Puluh Kota Regency, Tanah Datar Regency, Sijunjung Regency, Pesisir Selatan Regency
3	6	Payakumbuh City, Padang Panjang City, Solok City, Padang City, Bukittinggi City, Pariaman City

3.6 Agglomerative Nesting

1. Single linkage: The results Agglomerative nesting clustering that has been carried out using the single linkage method can be seen in Figure 6.

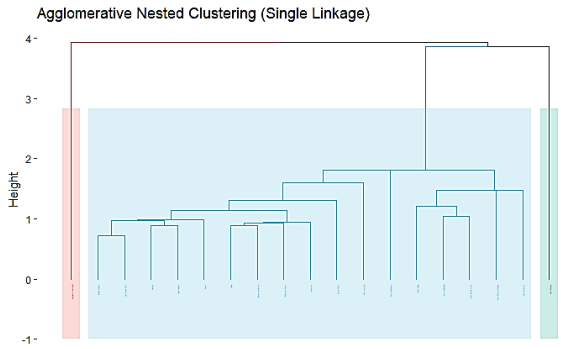

Figure 6. Dendrogram Agglomerative Nesting Clustering (Single linkage)

Based on Figure 6, information is obtained that cluster 1 consists of 1 regency/city, cluster 2 consists of 17 regencies/cities, and cluster 3 consists of 1 regency/City. Information regarding cluster members is in Table 7.

Table 7. Results Agglomerative Nesting With the Single Linkage Method for Grouping Regencies/ Cities

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	17	Padang Pariaman Regency, Pasaman Regency, Solok City, Payakumbuh City, Pariaman City, Sawahlunto City, Solok Regency, Tanah Datar Regency, Agam Regency, Pasaman Barat Regency, Dharmasraya Regency, Padang Panjang City, Solok Selatan Regency, Bukittinggi City, Sijunjung Regency, Pesisir Selatan Regency, Lima Puluh Kota Regency
3	1	Padang City

2. Complete linkage: The results Agglomerative Nesting clustering that has been carried out using the complete linkage method can be seen in Figure 7.

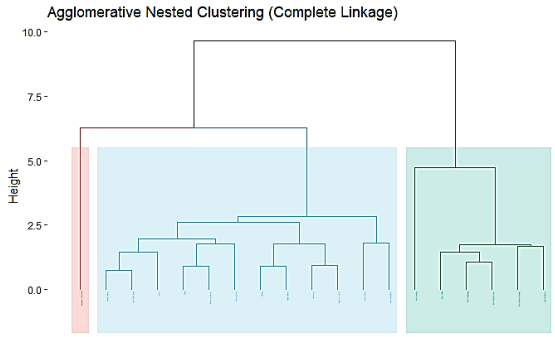

Figure 7. Dendrogram Agglomerative Nesting Clustering (Complete linkage)

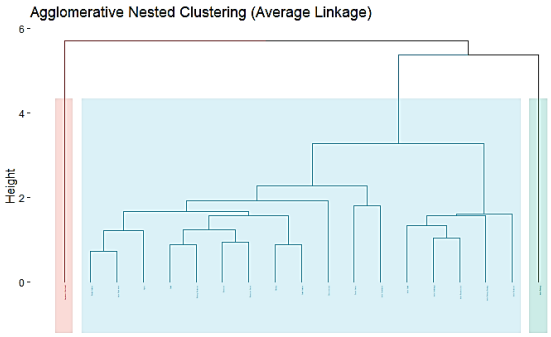

Figure 8. Dendrogram Agglomerative Nesting Clustering (Average linkage)

Based on Figure 7, information is obtained that cluster 1 has 1 regency/City, cluster 2 has 12 regencies/cities, and cluster 3 has 6 regencies/cities. Information regarding cluster members is in Table 8.

Table 8. Results Agglomerative Nesting With the Complete Linkage Method for Regency /City Grouping

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	12	Solok Selatan Regency, Lima Puluh Kota Regency, Padang Pariaman Regency, Solok Regency, Agam Regency, Pasaman Regency, Pasaman Barat Regency, Sawahlunto City, Pesisir Selatan Regency, Tanah Datar Regency, Sijunjung Regency, Dharmasraya Regency
3	6	Padang Panjang City, Pariaman City, Padang City, Bukittinggi City, Payakumbuh City, Solok City

3. Average linkage: The results Agglomerative nesting clustering that has been carried out using the average linkage method can be seen in Figure 8.

Based on Figure 8, information is obtained that cluster 1 has 1 regency/City, cluster 2 has 17 regencies/cities, and cluster 3 has 1 regency/City. Information regarding cluster members is in Table 9.

Table 9. Results Agglomerative Nesting With Average Linkage Method for Grouping Regencies/ Cities

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	17	Padang Pariaman Regency, Pasaman Regency, Solok City, Payakumbuh City, Pariaman City, Sawahlunto City, Solok Regency, Tanah Datar Regency, Agam Regency, Pasaman Barat Regency, Dharmasraya Regency, Padang Panjang City, Solok Selatan Regency, Bukittinggi City, Sijunjung Regency, Pesisir Selatan Regency, Lima Puluh Kota Regency
3	1	Padang City

4. Ward's Method: The results Agglomerative nesting clustering that has been done using ward's method can be seen in Figure 9.

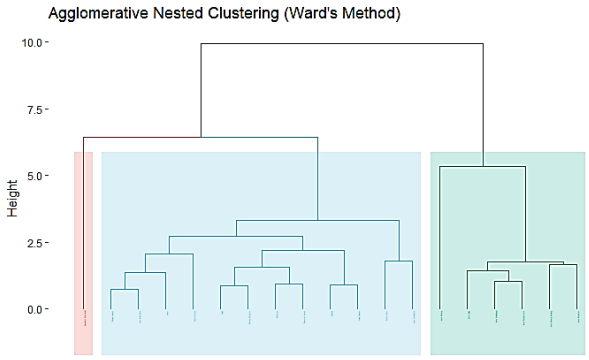

Figure 9. Dendrogram Agglomerative Nesting Clustering (Ward's Method)

Based on Figure 9, information is obtained that cluster 1 has 1 regency/City, cluster 2 has 12 regencies/cities, and cluster 3 has 6 regencies/cities. Information regarding cluster members is in Table 10.

Table 10. Results Agglomerative Nesting With Ward's Method For Grouping Regencies/ Cities

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	12	Solok Selatan Regency, Lima Puluh Kota Regency, Padang Pariaman Regency, Solok Regency, Agam Regency, Pasaman Regency, Pasaman Barat Regency, Sawahlunto City, Pesisir Selatan Regency, Tanah Datar Regency, Sijunjung Regency, Dharmasraya Regency
3	6	Padang Panjang City, Pariaman City, Padang City, Bukittinggi City, Payakumbuh City, Solok City

3.7 Divisive Analysis

The result divisive analysis clustering that has been carried out can be seen in Figure 10.

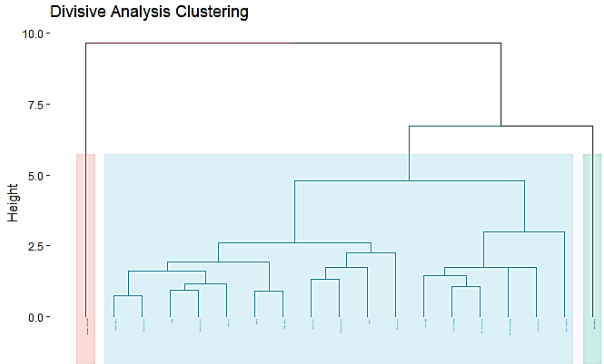

Figure 10. Dendrogram Divisive Analysis Clustering

Based on Figure 10, information is obtained that cluster 1 consists of 1 regency/City, cluster 2 consists of 17 regencies/cities, and cluster 3 has 1 regency/City. Information regarding cluster members is in Table 11.

Table 11. Results of Grouping Regencies/ Cities Using The Divisive Analysis Method

Clusters	Num. of Member	Regency /City
1	1	Mentawai Island Regency
2	17	Padang Pariaman Regency, Pasaman Regency, Solok City, Payakumbuh City, Pariaman City, Sawahlunto City, Solok Regency, Tanah Datar Regency, Agam Regency, Pasaman Barat Regency, Dharmasraya Regency, Padang Panjang City, Solok Selatan Regency, Bukittinggi City, Sijunjung Regency, Pesisir Selatan Regency, Lima Puluh Kota Regency
3	1	Padang City

3.8 Determine the best method

The Clustering method for grouping regencies/cities in this study was obtained by evaluating the cluster results that were obtained. There are 2 methods of evaluating Clustering results used, namely Davies Bouldin Index and Dunn index. Based on the calculations that have been carried out, the results obtained are as in Table 12 and Table 13.

Table 12. Evaluation of The Results of the Regency/ City Groupings in West Sumatera Province Part 1

Method	Davies Bouldin Index	Dunn Index
K-Means	1,380	0.187
Fuzzy C-Means	1,380	0.187
K-Medoids	0.714	0.383
Divisive Analysis	0.344	0.806

Table 13. Evaluation of The Results of Regency/ City Groupings in West Sumatera Province Part 2

Agglomerative Nesting Method	Davies Bouldin Index	Dunn Index
Single linkage	0.344	0.806
Complete linkage	0.714	0.383
Average linkage	0.344	0.806
Ward's Method	0.714	0.383

Based on Tables 12 and 13, it can be seen that the Clustering results have a Davies value The lowest Bouldin Index and the highest Dunn index are the divisive analysis method and Agglomerative Nesting for the single linkage, complete linkage, and Ward's method. So it can be concluded that the best Clustering method in this research is the divisive analysis method and Agglomerative Nesting for the single linkage, complete linkage, and Ward's method.

3.9 Discussion

3.9.1. West Sumatera Regency/ City Grouping Results

The thematic map based on the best Clustering method can be seen in Figure 11.

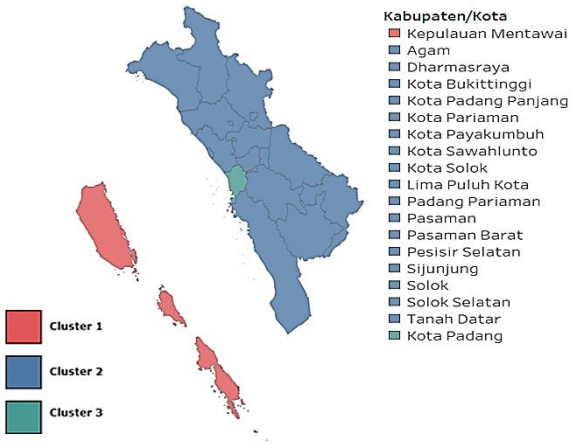

Figure 11. Thematic map resulting from grouping regencies/ cities based on the best method

Based on Figure 11, it can be seen that the high level of poverty in the Mentawai Islands Regency can be caused by its geographical location which is far from other regencies/cities in West Sumatera. This is by Tobler's first law of geography which states that the relationship between locations or phenomena on the earth's surface influences each other, especially if the locations are close together, [33]. In other words, things that are close to each other tend to be more strongly related than things that are far away. From the clusters

formed, the category of poverty level is determined based on the average of each cluster which can be seen in Table 14. In carrying out profiling, the average used is the average of the data before standardization.

Table 14. Profiling of Regency/ City Grouping Results Using the Best Method

Variable	Average		
	Cluster 1	Cluster 2	Cluster 3
Percentage Poor Population	13.97	5.53	4.26
Expenditure Real Per capita	6567	11044.24	14889
Life expectancy	64.93	70.80	73.93
Average Years of Schooling	7.48	9.37	11.60
Long School Expectations	12.89	13.75	16.54
Open Unemployment Rate	1.39	5.13	11.69

In Table 14, the orange color shows the highest average value and the blue color shows the lowest average value of each variable compared to the entire cluster. Based on Table 14, the characteristics of each cluster are obtained as follows:

1. Cluster 1 is the cluster categorized as having the highest poverty level. This cluster is a group of regencies/cities that have the highest percentage of poor people compared to other groups. Per Capita Real Expenditure, Average Years of Schooling, Life Expectancy, Expected Years of Schooling, and the Open Unemployment Rate are the lowest among other groups.
2. Cluster 2 is a cluster that is categorized as having a medium poverty level. This cluster is a group of regencies/cities that have a Percentage of Poor Population, Real Per Capita Expenditure, Average Years of Schooling, Life Expectancy, Expected Years of Schooling, and Open Unemployment Rate which are in a moderate position among other groups.
3. Cluster 3 is the cluster categorized as having the lowest poverty level. This cluster is a group of regencies/cities that have the lowest percentage of poor people compared to other groups. Real Expenditure Per Capita, Expected Years of Schooling, Life Expectancy, Average Years of Schooling, and the Open Unemployment Rate are the highest among other groups.

3.9.2. Hierarchical vs Non-Hierarchical

The hierarchical and non-hierarchical clustering methods each have their own advantages and disadvantages. Based on the research conducted, it can be seen that the hierarchical method is more optimal for relatively small data sets compared to the non-hierarchical method. However, the non-hierarchical method is very suitable for large-sized data due to its higher speed compared to the hierarchical method [37]. Nevertheless, the weakness of this method lies in the need to determine the number of clusters and centroids beforehand, as well as the clustering results that may depend on the order of data observations [37]. Therefore, if the data is small, the hierarchical method is more recommended. Conversely, if the data is large, the non-hierarchical method is more recommended. However, the best clustering method actually depends on the analysis needs and specific data characteristics.

4. CONCLUSION

Based on the silhouette coefficient method, the optimal number of clusters for the hierarchical method and non-hierarchical method in this study is 3. In this research, the best Clustering method is the divisive analysis method and agglomerative nesting, especially in complete linkage, single linkage, and Ward's method because it has a Davies value The lowest Bouldin Index, and the highest Dunn index value. The final results obtained for Cluster 1 were 1 regency/City, Cluster 2 was 17 regencies/cities, and Cluster 3 was 1 regency/City. The characteristic of Cluster 1 is a high level of poverty, the characteristic of Cluster 2 is a moderate level of poverty, and the characteristic of Cluster 3 is a low level of poverty. By understanding the characteristics of poverty in each regency/City, it is hoped that the West Sumatera provincial government can make more appropriate policies to overcome the issue of poverty. In future research, it is recommended to consider additional Clustering methods to improve the quality of analysis as well as more in-depth geographic analysis.

REFERENCES

[1] Badan Pusat Statistik. [Online]. Available: https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html
[2] United Nations. Sustainable Development Goals. [Online]. Available: https://www.un.org/sustainabledevelopment/poverty/
[3] Badan Pusat Statistik, "Berita Resmi Statistik" Berita Resmi Statistik No.07/01/Th. XXVI, Januari 2023.
[4] Portal Resmi Provinsi Sumatera Barat. (2022, 10). Tindak Lanjuti Arahan Presiden untuk nol Kemiskinan Ekstrim 2024, Gubernur Sumbar Buka Rakor Penanggulangan Kemiskinan Tahun 2022. [Online]. Available :

https://sumbarprov.go.id/home/news/22102-tindak-lanjuti-arahan-presiden-untuk-nol-kemiskinan-ekstrim-2024-gubernur-sumbar-buka-rakor-penanggulangan-kemiskinan-tahun-2022

[5] Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (6th ed.). Pearson, Prentice Hall.

[6] Afira, N., & Wijayanto, A. W. (2021). Analisis Cluster dengan Metode Partitioning dan Hierarki pada Data Informasi Kemiskinan Provinsi di Indonesia Tahun 2019. Komputika : Jurnal Sistem Komputer, 10(2), 101–109. https://doi.org/10.34010/komputika.v10i2.4317

[7] Wahyuni, S., & Jatmiko, Y. A. (2018). Pengelompokan Kabupaten/Kota Di Pulau Jawa Berdasarkan Faktor-Faktor Kemiskinan Dengan Pendekatan Average linkage Hierarchical Clustering. Jurnal Aplikasi Statistika Dan Komputasi Statistik.

[8] Munandar, Tb. A. (2022). Penerapan Algoritma Clustering Untuk Pengelompokan Tingkat Kemiskinan Provinsi Banten. JSiI (Jurnal Sistem Informasi), 9(2), 109–114. https://doi.org/10.30656/jsii.v9i2.5099

[9] Widodo, E., Ermayani, P., Laila, L. N., & Madani, A. T. (2021). Pengelompokkan Provinsi di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Analisis Hierarchical Agglomerative Clustering (Indonesian Province Grouping Based on Poverty Level Using Hierarchical Agglomerative Clustering Analysis). Prosiding Seminar Nasional Official Statistics.

[10] Simatupang, M. D., & Wijayanto, A. W. (2021). Analisis Klaster Berdasarkan Tindakan Kriminalitas Di Indonesia Tahun 2019. Jurnal Statistika Industri Dan Komputasi, 6(1), 10–19.

[11] Puspitasari, Dewi. (2018). Faktor-faktor Yang memengaruhi Kemiskinan di Jawa Tengah Tahun 2012-2016: Regresi Data Panel Spasial. [Skripsi]. Jakarta: Politeknik Statistika STIS.

[12] Nasution, Taupiq H. T. (2022). Faktor-faktor Yang Memengaruhi Kedalaman Dan Keparahan Kemiskinan Kabupaten/kota di Kti Tahun 2020. [Skripsi]. Jakarta: Politeknik Statistika STIS.

[13] Rabbani, H. M. (2022). Analisis Faktor-faktor Yang Memengaruhi Kemiskinan Ekstrem Menurut Provinsi Di Kawasan Timur Indonesia Tahun 2010-2021. [Skripsi]. Jakarta: Politeknik Statistika STIS.

[14] Ardian, D., & Rizqi Destanto, M. (2021). Pengaruh Faktor Sosial Ekonomi terhadap Kemiskinan di Provinsi Jawa Barat (The Effect of Socio-Economic Factor on Poverty Level in West Java Province). Prosiding Seminar Nasional Official Statistics, 377–384.

[15] Saputra, Danu A. (2021). Penerapan Algoritma Machine Learning untuk Pengelompokan Kesejahteraan Sosial (Studi Kasus : Kawasan Timur Indonesia Tahun 2019). [Skripsi]. Jakarta: Politeknik Statistika STIS.

[16] Kusumaningtyas, C. A. (2018). Analisis Cluster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Papua Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2017. [Skripsi]. Jakarta: Politeknik Statistika STIS.

[17] Palanu, H. (2017). Analisis Pengelompokan Karakteristik Konsumen Penggunawarnet Dengan Metode Ward's (Studi Kasus : Warnet Turbo Net). [Skripsi]. Yogyakarta: Universitas Islam Indonesia.

[18] Ningrat, D. R., Maruddani, D. A. I. dan Wuryandari, T. (2016). Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan data Obligasi Korporasi. Jurnal Gaussian, 5(4), 641-650.

[19] Eldo, H. (2020). Penentuan Cluster Terbaik K-Means Menggunakan Algoritma Silhouette. [Tesis]. Medan: Universitas Sumatera Utara.

[20] Zulfa, F. (2019). Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Pendidikan Menggunakan K-Means Dan K-Medoids. [Skripsi]. Semarang: Universitas Muhammadiyah Semarang.

[21] R. Xu and D. Wunsch, "Clustering", John Wiley & Sons, vol. 10., 2008.

[22] Gan, G., Ma, C., and Wu, J., Data Clustering: Theory, Algorithms, and Applications, ASA SIAM series on Statistics and Applied Probability, Philadelphia, 2007.

[23] Patnaik, A. K., Bhuyan, P. K., & Krishna Rao, K. V. (2016). Divisive analysis (DIANA) of Hierarchical Clustering and GPS data for level of service criteria of urban streets. Alexandria Engineering Journal, 55(1), 407–418. https://doi.org/10.1016/j.aej.2015.11.003

[24] Prasetyo, E., "Data Mining: Konsep dan Aplikasi Menggunakan Matlab", Yogyakarta: Andi Offset, 2012.

[25] T. S. Jaya, "Sistem Pemilihan Perumahan dengan Metode Kombinasi Fuzzy CMeans Clustering dan Simple Additive Weighting," J. Sist. Inf. Bisnis, vol. 1, no. December, pp. 153–158, 2018, doi: 10.21456/vol1iss3pp153-158.

[26] Yohannes, "Analisis Perbandingan Algoritma Fuzzy C-Means," in Annual research Seminar, 2016, vol. 2, no. 1, pp. 151–155.

[27] Bezdek, J. C., Enrlich, R., dan Full, W., "FCM : The Fuzzy C-Means Clustering Algortihm," Computer & Geosciences, 10(2-3), 191-203, 1984.

[28] Kaur, Noor K., Kaur, Usvir., & Singh, Dr. Dheerendra., 2014. K-Medoids Clustering Algorithm –A Review. International Journal of Computer Application and Technology (IJCAT). ISSN. 2349-1841 Vol. 1, Issue 1. April 2014.

[29] Riyanto, B. (2019). Penerapan Algoritma K-Medoids Clustering Untuk Pengelompokkan Penyebaran Diare Di Kota Medan (Studi Kasus: Kantor Dinas Kesehatan Kota Medan). KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer),3(1). https://doi.org/10.30865/komik.v3i1.1659

[30] Dewi, D. A. I. C., dan Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow Dan Silhouette Pada Algoritma Clustering K-Medoids Dalam Pengelompokan Produksi Kerajinan Bali. Matrix : Jurnal Manajemen Teknologi dan Informatika, 9(3), 102-109

[31] Sarjanako, R. J. (2016). Penerapan Fuzzy C-Means Clustering Untuk Mengoptimalkan Penentuan Media Promosi. Jurnal Ilmiah Teknologi dan Informasi (SNATi), 1-6.

[32] Satoso, B. D., Khotimah, B. K. dan Muhammad, A. (2015). Pengelompokan Tingkat Kesehatan Masyarakat Menggunakan Shelf Organizing Maps Dengan Cluster Validation Idb dan I-Dunn. Seminar Nasional Aplikasi Teknologi Informasi (SNATi), 1-6.
[33] Schabenberger O, Gotway CA. 2005. Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC.
[34] Thamrin, N,, & Wijayanto, A. W. (2021). Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island: Analisis cluster dengan menggunakan hard clustering dan soft clustering untuk pengelompokkan tingkat kesejahteraan kabupaten/kota di pulau Jawa. Indonesian Journal of Statistics and Its Applications, 5(1), 141-160.
[35] Damayanti, A. R., & Wijayanto, A. W. (2021). Comparison of Hierarchical and Non-Hierarchical Methods in Clustering Cities in Java Island using the Human Development Index Indicators year 2018. Eigen Mathematics Journal, 4(1).
[36] Suraya, G. R., & Wijayanto, A. W. (2022). Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting. Indonesian Journal of Statistics and Its Applications, 6(2).
[37] Sigit Nugroho, P. (2008). Statistika Multivariat Terapan. Bengkulu: UNIB Press.

BIBLIOGRAPHY OF AUTHORS

Farhan Maulana is a bachelor degree student at Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia. His research interest are data science and data analysis.

Arie Wahyu Wijayanto is a Lecturer at Department of Computational Statistics Politeknik Statistika STIS, Jakarta, Indonesia. He also currently serves as Head of the Center for Research and Community Service and Assistant Professor at Politeknik Statistika STIS. He completed his Doctor of Engineering in Computer Science from Tokyo University of Technology, Japan. His special interest lies in the field of data science, big data analytics, and geospatial artificial intelligence.