

Article Identity

Volume 1 Issue 1 Pages 1–9

October 2026

E-ISSN xxxx-xxxx

Article History

Received 20 July 2026

Revised 24 August 2026

Accepted 08 September 2026

Published 14 September 2026

Keywords

Agglomerative Ward linkage;

K-means; K-medoids; welfare

Copyright © 2026 Author(s). Replace this sentence with the official JSDA copyright and open-access license statement.

Editorial office: Journal of Statistics and Data Analytics PT. Gammabeta Edu and Tech.

To Cite this Article: Dewi Indra Setiawan, Muhammad Yahya Matdoan. Cluster Analysis of Districts and Cities in North Sumatra Province Based on Socio-Economic Indicators. Journal of Statistics and Data Analytics. 2026;1(1):1–9.

RESEARCH ARTICLE

Cluster Analysis of Districts and Cities in North Sumatra Province Based on Socio-Economic Indicators

Dewi Indra Setiawan^{1*}, Muhammad Yahya Matdoan²

¹Department of Data Science, Faculty of Science, Institut Teknologi Sumatera
Jl. Terusan Ryacudu, Way Hui, Kec. Jati Agung, Kabupaten Lampung Selatan, Lampung.

²Department of Mathematics, Faculty of Science and Technology, Universitas Pattimura
Jl. Ir. M. Putuhena, Kampus Poka, Rumah Tiga, Kec. Teluk Ambon, Kota Ambon, Maluku

Corresponding author: *dewi.setiawan@sd.itera.ac.id

Abstract. *The success of development program implementation related to community welfare in North Sumatra Province can be evaluated through several indicators, including consumption patterns, housing and environment, population, education, nutrition, health, expenditure, and employment. This study aims to group districts and cities in North Sumatra Province based on similar or homogeneous characteristics of these welfare indicators. To achieve this, both hierarchical and non-hierarchical clustering methods were applied, including k-means and k-medoids (using PAM and CLARA algorithms). Among these methods, the hierarchical clustering using Ward's linkage was found to produce the best clustering results, forming three distinct clusters. The first cluster consists of 4 districts/cities, the second cluster includes 23 districts/cities, and the third cluster comprises 6 districts/cities. These clustering results can assist local governments in North Sumatra in formulating more targeted and effective welfare policies, based on the dominant indicators present in each cluster.*

1. INTRODUCTION

Regional development refers to the utilization of available resources to achieve tangible improvements in community welfare, including income levels, employment opportunities, business opportunities, access to policy-making, competitiveness, and the improvement of the Human Development Index (HDI). Compared with other provinces across Sumatra Island, the number of poor people in North Sumatra Province is higher than in other regions. Meanwhile, the lowest number of poor people is found in the Bangka Belitung Islands Province. However, when viewed in terms of the poverty rate, Aceh Province has the highest percentage of poor people, followed by Bengkulu Province and South Sumatra Province.

This issue indicates that development policies in North Sumatra Province may not yet be properly targeted. Therefore, it is necessary to identify the characteristics of regional development based on the level of community welfare in each area in order to formulate appropriate policies and strategies to accelerate the development process. Measuring welfare requires several detailed indicators from various aspects. Public welfare indicators reflect the quality of human resources in a country, which include indicators related to consumption patterns, housing and environment, population, education, nutrition, health, expenditure, and employment [1].

To support the successful implementation of development programs, particularly those related to community welfare in North Sumatra Province, it is necessary to conduct research that groups districts and cities with similar or homogeneous characteristics in terms of welfare indicators. Such grouping can be carried out using clustering methods. Clustering methods are generally divided into two categories: hierarchical clustering and partitioning (also known as non-hierarchical clustering). Hierarchical clustering is used when the desired number of clusters is not yet known and includes methods such as complete linkage, single linkage, average linkage, centroid linkage, and Ward's linkage. Meanwhile, non-hierarchical clustering is applied when the desired number of clusters has been determined, and includes methods such as K-means and K-medoids.

Several previous studies have been conducted to determine the most appropriate clustering methods. Soft clustering and hard clustering methods are applied to group welfare indicators on Java Island [2]. The K-means algorithm was used to cluster public welfare indicators in North Sumatra Province [3]. Furthermore, research conducted by Izzaty, HG, and Devianto grouped districts and cities in West Sumatra Province based on community welfare indicators using the Silhouette coefficient for cluster validity assessment [4].

This study aims to classify districts and cities in North Sumatra Province based on the characteristics of public welfare indicators. Prior to clustering, the variables are reduced in dimension using Principal Component Analysis (PCA) to improve and accelerate the performance of clustering methods. Through this grouping, it is expected that the government can implement appropriate policies to achieve a more equitable distribution of welfare. In addition, this study also aims to compare several clustering methods that can be used to classify regions based on welfare levels in order to determine the most appropriate clustering method for grouping districts and cities in North Sumatra Province.

2. METHOD

2.1. Research Method

Principal Component Analysis (PCA) is a statistical technique used to transform a large set of original variables — which are correlated with one another — into a smaller set of new, independent (uncorrelated) variables. Thus, PCA is useful for data reduction, making it easier to interpret the data. Component formation is based on two approaches: the correlation matrix and the covariance matrix [5]. The process of determining principal components using the correlation matrix is as follows:

1. Construct matrix $\mathbf{Z}$, which is the result of standardizing variable $\mathbf{X}$.
2. Construct the correlation matrix from $\mathbf{Z}$ ($\mathbf{Z}'\mathbf{Z}$). The first step in the component reduction process is to find the eigenvalues λ from the equation:

$$\left| \mathbf{Z}'_{(p \times n)} \mathbf{Z}_{(n \times p)} - \lambda_{(p \times p)} \mathbf{I}_{(p \times p)} \right| = 0 \quad (1)$$

where:

$\mathbf{Z}$: matrix resulting from the standardization of variable $\mathbf{X}$

Λ : eigenvalue matrix, which is a diagonal matrix

$\mathbf{I}$: identity matrix of size $n \times p$

The number of principal components is selected based on the eigenvalue (λ), where if $\lambda > 1$, the component is selected as a principal component [6]. Once the assumptions are met, cluster analysis is then performed. Cluster analysis is an analytical method used to group objects based on similarity of characteristics, or relative homogeneity among those objects [5]. In general, there are two methods in cluster analysis: hierarchical clustering and non-hierarchical clustering.

a. Hierarchical Clustering Method

This method begins grouping with two or more objects that have the closest similarity, then continues to the next closest objects, and so on, until the clusters form a tree-like structure with a clear hierarchy (levels) among objects, ranging from most similar to least similar [5]. Hierarchical clustering methods are differentiated based on the concept of distance between clusters. Several hierarchical techniques include Single Linkage, Complete Linkage, Average Linkage, and Ward Linkage.

Table 1. Hierarchical Clustering Distance Calculation Methods

No.	Method	Distance Calculation
1	Single Linkage	$d_{(u,v)w} = \min\{d_{uw}, d_{vw}\}$
2	Complete Linkage	$d_{(u,v)w} = \max\{d_{uw}, d_{vw}\}$
3	Average Linkage	$d_{(u,v)w} = \text{average}\{d_{uw}, d_{vw}\}$
4	Ward Linkage	Based on the sum of squares error formula

In the Ward Linkage method, the clusters formed have the smallest possible internal variance, because the distance between two clusters is based on the sum of squares of the two clusters formed. This method is also effective and is therefore frequently used compared to other hierarchical methods [7]. The clustering process is based on minimum variance within clusters [5]. This method is commonly used for objects with a small number of clusters. The process in Ward Linkage clustering is as follows:

1. Calculate the sum of squares error between two clusters using the formula:

$$SSE_{ij} = \frac{1}{2} \sum_{k=1}^p (x_{ik} - x_{jk})^2 \quad (2)$$

where:

SSE_{ij} : Sum of Squares Error between object pair i and j

x_{ik} : value of object i on variable k

x_{jk} : value of object j on variable k

2. Find the smallest SSE value between two clusters, then merge them into one cluster. Thus, from n clusters, the number systematically decreases to $n - 1$, and so on.
3. Repeat step (2) until the minimum number of clusters is obtained.

Selecting the Best Hierarchical Method

The selection of the hierarchical method in this study uses the agglomerative coefficient as a measure of the structure of the hierarchical cluster. An agglomerative coefficient value close to 1 indicates a more balanced and strong structure [8]. A value close to 0 indicates a poorer cluster structure. However, the agglomerative coefficient tends to increase as the sample size grows, so it cannot be used to compare datasets of vastly different sizes.

b. Non-Hierarchical Clustering Method

Non-hierarchical cluster analysis is a method of grouping objects into k groups, where k is determined before the analysis is conducted. This study uses K-Means Clustering and K-Medoids Clustering to assess the suitability of each method with the data used.

K-Means Clustering

K-Means clustering is the simplest non-hierarchical clustering method, which divides data into several groups where data with similar characteristics are placed in the same group and data with more different characteristics are grouped into another cluster [5]. The process in K-Means clustering is as follows:

1. Determine the number of clusters (k).
2. Randomly initialize centroid values (cluster centers) for k clusters.

3. Calculate the distance of each data point to each cluster centroid using Euclidean distance. The Euclidean distance between objects i and j is formulated as:

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (3)$$

where:

x_{ik} : value of object i on variable k

x_{jk} : value of object j on variable k

p : number of observed variables

4. Group data based on the shortest distance between each data point and the centroid.
5. Determine new centroids by calculating the mean of all data points in the same cluster center.

$$C_{kj} = \frac{x_{1j} + x_{2j} + \dots + x_{aj}}{a}, \quad j = 1, 2, \dots, p \quad (4)$$

where:

C_{kj} : center of cluster k on variable j

n_k : number of data points in cluster k

6. Return to step 3 and continue iterating until the cluster centroids are stable and no data points move between clusters (convergence).

K-Medoids Clustering

K-Medoids uses a medoid as the cluster center — that is, it uses one of the existing data points, randomly initialized, to represent the cluster [9]. The K-Medoids algorithm consists of the following 6 steps:

1. Determine the number of clusters (k), adjusted in this study based on the results of the elbow method.
2. Select arbitrary objects in each cluster to serve as medoids.
3. Allocate all objects to the cluster with the nearest medoid using Euclidean distance theory, with the formula:

$$D(a, b) = \sqrt{\sum_{k=1}^n (a_k - b_k)^2} \quad (5)$$

where:

n : number of attributes

x_k : data with attribute k

4. Calculate the Total Cost (TC). Total Cost is obtained by summing the minimum distances selected across all clusters, denoted as d_s . In general, TC can be calculated using:

$$TC = \sum_{i=1}^N d_s \quad (6)$$

5. Randomly select new medoid candidates as in step (2), then repeat steps 3 and 4. The new total cost is denoted TC'.
6. Determine the total deviation (S) by calculating the new total distance minus the previous total distance: $S = TC' - TC$. If $S > 0$, the iteration stops and the cluster members used are those from when TC was calculated. If $S < 0$, the objects must be swapped with the n cluster data to form a new set of k objects, and steps (3) through (5) are repeated.

The algorithms used in K-Medoids clustering are PAM (Partitioning Around Medoids) and CLARA (Clustering Large Application). Both algorithms share similar principles. CLARA is an extension of PAM that relies on a sampling process — it takes samples from the dataset, applies PAM to the sample, and finds the medoid from the sample. For a better approximation, CLARA takes multiple samples and returns the best output [10].

Evaluation of the Best Clustering Method

Cluster evaluation consists of external and internal evaluation. External evaluation can be performed when the clusters have labels (supervised). This study uses only internal evaluation because the data used is unlabeled. Internal evaluation uses the internal information within the data to assess clustering results. It reflects the compactness, connectivity, and separation of cluster partitions. The internal evaluations used are the Silhouette Coefficient, Dunn Index, and Connectivity Coefficient.

Internal: Silhouette Coefficient

The Silhouette Coefficient is a measure of the degree of confidence in the assignment of an observation to a cluster. A cluster is considered good if the coefficient is close to 1, and poor if it is close to -1. The Silhouette Coefficient is calculated as follows:

- Calculate the average distance between an object (i) and every other object in the same cluster (cohesion):

$$a(i) = \frac{1}{|A| - 1} \sum_{\substack{j \in A \\ j \neq i}} d(i, j) \quad (7)$$

Where j is an object other than i in the same cluster A , and $|A|$ is the number of members in cluster A .

- Calculate the average distance between object i and every object in each other cluster, then take the smallest distance (nearest neighboring cluster):

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \quad (8)$$

Where $d(i, C)$ is the average distance between object i and every object in another cluster C , where $A \neq C$. Then calculate the separation, which is the distance to the nearest neighboring cluster:

$$b(i) = \min_{C \neq A} d(i, C) \quad (9)$$

The Silhouette Coefficient is then formulated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (10)$$

Internal: Dunn Index

The Dunn Index is a measure for validating clustering results, obtained by measuring the distance between two clusters and the diameter of a cluster. It is the ratio of the largest inter-cluster distance to the smallest intra-cluster distance. The higher the Dunn Index value, the better the cluster structure [11]. The Dunn Index is formulated as:

$$DI = \min_{i=1, \dots, k} \left\{ \min_{j=i+1, \dots, k} \left(\frac{\text{disc}(c_i, c_j)}{\max_{m=1, \dots, k} \{\text{diam}(c_m)\}} \right) \right\} \quad (11)$$

where:

DI : Dunn Index

$d(i, j)$: distance between cluster i and cluster j

$\text{diam}(i)$: diameter of cluster i

Internal: Connectivity Coefficient

Compactness relates to evaluating the homogeneity of a cluster, commonly assessed using intra-cluster variance. Connectivity indicates the position of observed data within a cluster relative to its nearest neighbors. This compactness is measured using the connectivity coefficient. The connectivity value ranges from zero to ∞ . The lower the connectivity coefficient, the better the cluster formed [11]. The Connectivity Coefficient is formulated as:

$$CC = \sum_{i=1}^N \sum_{j=1}^L x_{i, nni(j)} \quad (12)$$

where:

$nni(j)$: nearest neighbor j of object i

$x_{i, nni(j)}$: approaches 0 if objects i and j are in the same cluster, and 1 otherwise

L : parameter for the number of neighbors

2.2. Data

The data used in this study is secondary data obtained from the Central Bureau of Statistics (Badan Pusat Statistik/BPS). This study uses welfare indicators sourced from the website of BPS North Sumatra Province for the year 2024. Based on the reference literature, ten variables (Table 2) were selected that influence the level of community welfare in a region.

Table 2. Research Variables

No.	Variable	Data Type
1	Life Expectancy	Quantitative
2	Expected Years of Schooling	Quantitative
3	Mean Years of Schooling	Quantitative
4	Adjusted Per Capita Expenditure	Quantitative
5	Economic Growth	Quantitative
6	Total Labor Force	Quantitative
7	Open Unemployment Rate	Quantitative
8	Number of People in Poverty	Quantitative
9	Population Density	Quantitative
10	Home Ownership	Quantitative

2.3. Analysis Method

The analysis stages are as follows:

1. Collecting welfare indicator data for the year 2024 from 33 regencies/cities in North Sumatra Province.
2. Data normalization.
3. Performing Principal Component Analysis.
4. Determining the optimal number of principal components for clustering.
5. Cluster analysis using the hierarchical method:
 - Determining the initial number of clusters using the elbow and silhouette methods.
 - Performing hierarchical cluster analysis (Single Linkage, Average Linkage, Complete Linkage, and Ward Linkage).
 - Evaluating the best hierarchical clustering method using the agglomerative coefficient, cophenetic correlation, and dendrogram criteria.

- Grouping regencies/cities based on the best hierarchical clustering method.
- Cluster analysis using the K-Means method: Determining the initial number of clusters using the elbow and silhouette methods.
 - Performing K-Means cluster analysis according to the optimal number of clusters.
 - Grouping regencies/cities based on the K-Means clustering method.
 - Calculating the degree of variation within each cluster and how well the clustering results are in terms of compactness.
 - Cluster analysis using the K-Medoids method: Determining the initial number of clusters using the elbow and silhouette methods.
 - Performing K-Medoids cluster analysis using the PAM and CLARA algorithms.
 - Grouping regencies/cities based on the K-Medoids PAM and CLARA clustering methods.
 - Comparing the K-Medoids PAM and CLARA clustering methods for the optimal number of clusters.
 - Comparing the best clustering method between the Hierarchical and Non-Hierarchical methods (K-Means and K-Medoids).

3. RESULTS AND DISCUSSION

3.1. Dimension Reduction with PCA

PCA was applied to reduce the dimensionality of the data by forming several components that are linear combinations of the variables used in the analysis. Only components with an eigenvalue greater than 1 were retained for further analysis. In this study, three components were formed with eigenvalues greater than 1 (Table 3). The cumulative variance of the three components is 72.65 percent, meaning that the three principal components are able to explain approximately 73 percent of the total variation in the original data. The analysis then proceeds with clustering modeling using various methods, with the principal component scores as the input variables.

Table 3. PCA Component Result

No.	Component	Eigenvalue	Cumulative Variance
1	Component 1	4.11390	41.14%
2	Component 2	1.76925	58.83%
3	Component 3	1.38174	72.65%

3.2. Cluster Analysis Using Hierarchical Method

The first step is to determine the optimal number of clusters (k) to be used for clustering the data with the agglomerative hierarchical method. This study uses the Elbow Method and Silhouette Method to determine the optimal value of k, resulting in the plots presented in Figure 1. Based on the Elbow Method plot, it can be seen that the largest gradient break occurs when the number of clusters is three. However, the Silhouette Method indicates that the optimal number of clusters is two. This study proceeds with three clusters, as two clusters is considered too few. A comparison of the best hierarchical method among Single Linkage, Complete Linkage, Average Linkage, and Ward Linkage was then conducted using the Agglomerative Coefficient and Connectivity Coefficient as evaluation criteria.

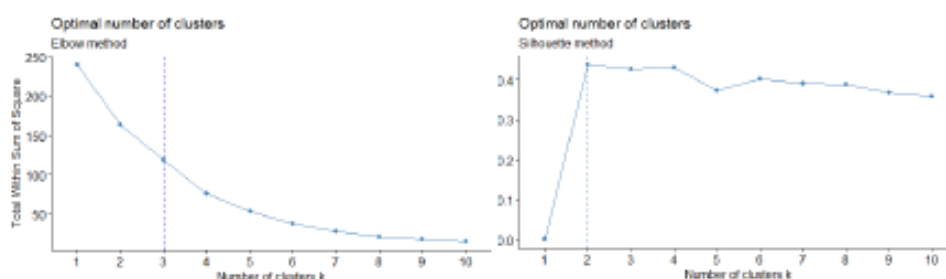


Figure 1. Elbow Method and Silhouette Method for Agglomerative Hierarchical Clustering

Table 4 presents the evaluation of hierarchical clustering methods using the Agglomerative Coefficient (AC) and Connectivity Coefficient (CC) criteria. Based on the Agglomerative Coefficient criterion, the best hierarchical method is Complete Linkage, followed by Ward Linkage. Meanwhile, based on the Connectivity Coefficient criterion, the best hierarchical methods are Average Linkage and Single Linkage.

Table 4. Comparison of Hierarchical Methods Based on AC and CC Criteria

Hierarchical Method	Agglomerative Coefficient	Connectivity Coefficient
Single Linkage	0.8255569	0.882321
Complete Linkage	0.9031195	0.807929
Average Linkage	0.8646245	0.904301
Ward Method	0.9029421	0.720407

The next step in determining the best hierarchical method is by using a dendrogram. Figure 2 shows the dendrogram results for the 3 best methods based on the Agglomerative Coefficient and Connectivity Coefficient criteria, namely the dendrograms for the Complete Linkage, Ward Linkage, and Average Linkage methods. As can be seen from the dendrogram structure presented in Figure 2, the Complete Linkage and Average Linkage methods are less stable compared to the Ward Linkage method. The Ward Linkage method is considered more stable and representative in forming clusters. To further clarify the best method, a visualization of the resulting clusters was performed.

Based on the clustering results, it can be seen more clearly that the Agglomerative Ward Linkage method groups regions based on welfare levels better than the Complete Linkage method, as the three clusters are more distinctly visualized with clearer boundaries (Figure 3).

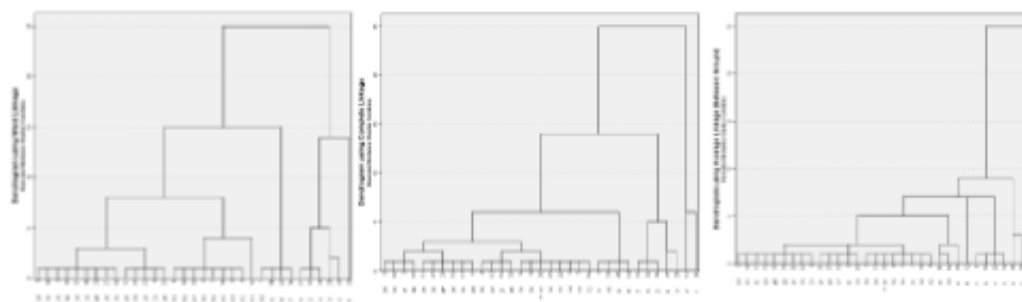


Figure 2. Dendrograms for Ward Linkage (a), Complete Linkage (b), and Average Linkage (c)

Therefore, it can be concluded that the best Agglomerative hierarchical method for grouping welfare levels based on regencies/cities in North Sumatra Province is the Agglomerative Ward Linkage method.

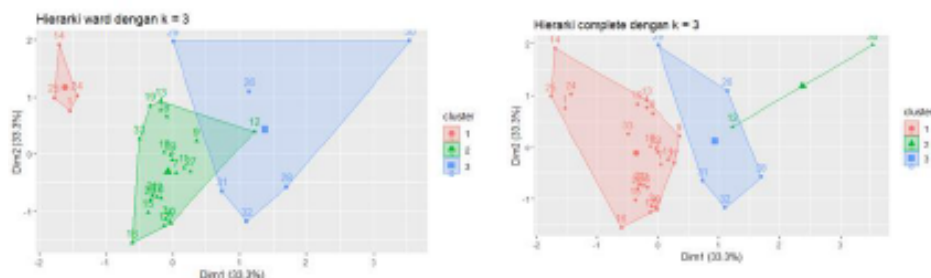


Figure 3. Clustering Using Agglomerative Ward Linkage and Complete Linkage with $k = 3$

3.3. Cluster Analysis Using the K-Means Method

In cluster analysis using the K-Means method, the first step is to determine the appropriate value of k for the clustering process. The Elbow Method and Silhouette Method were used to determine the optimal value of k . Based on the Elbow Method plot (Figure 4), it can be seen that the largest gradient break occurs when the number of clusters is four. This is consistent with the Silhouette Method, which also indicates that the optimal number of clusters is four.

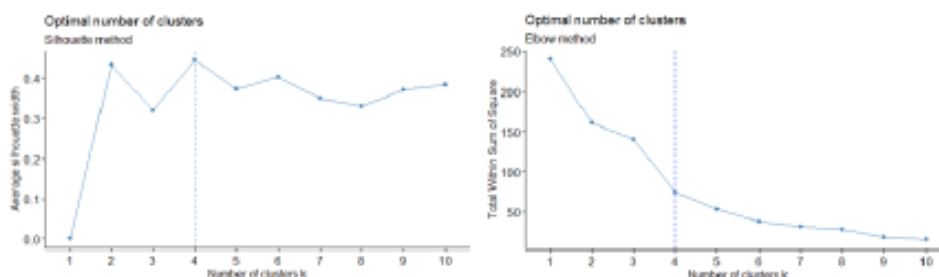


Figure 4. Elbow Method and Silhouette Method for K-Means Clustering

The clustering results using the K-Means method with 4 clusters are presented in Figure 5. Based on the clustering results, it can be seen that the first cluster consists of 5 regions, the second cluster consists of 22 regions, the third cluster consists of 4 regions, and the fourth cluster consists of 2 regions. Analysis of the clustering results shows the Within-Cluster Sum of Squares (WCSS) values for each cluster as follows: 21.61 (Cluster 1), 37.03 (Cluster 2), 2.22 (Cluster 3), and 13.63 (Cluster 4).

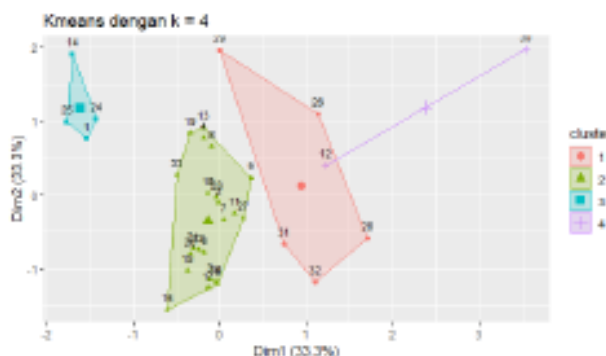


Figure 5. Regional Clustering Results Using K-Means with $k = 4$

The Within-Cluster Sum of Squares (WCSS) describes the degree of internal variation within each cluster; the smaller the value, the more homogeneous the members within that cluster.

Table 5. WCSS Values and Between-Cluster Proportion

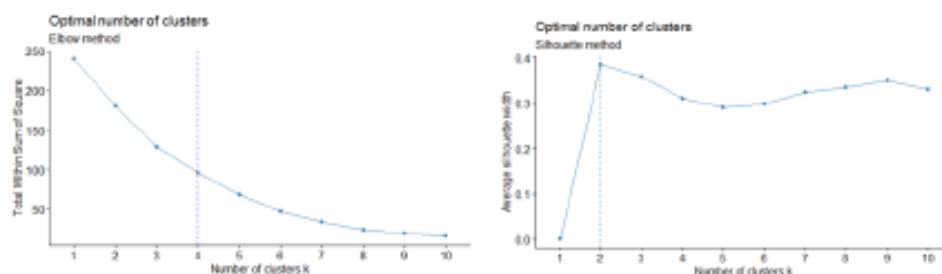
Cluster	WCSS
1	21.61
2	37.03
3	2.22
4	13.63
Total	74.49
Between-Cluster Proportion	68.9%

Note: $\text{Between_SS}/\text{Total_SS}$.

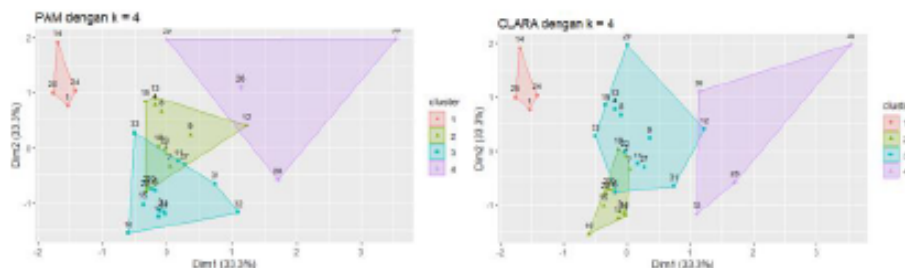
In addition, the ratio of Between-Cluster Sum of Squares (BCSS) to Total Sum of Squares (TSS) was recorded at 68.9%, indicating that approximately 68.9% of the total variation in the data can be explained by the separation between clusters (Table 5). This value indicates that the clustering results are of reasonably good quality, with a relatively clear cluster structure.

3.4. Cluster Analysis Using the K-Medoids Method

The K-Medoids algorithms used are PAM and CLARA. As with the previous methods, the first step is to determine the optimal number of clusters using the Elbow and Silhouette methods for both the PAM and CLARA algorithms. Based on the analysis using the Elbow and Silhouette methods, the optimal number of clusters for the PAM and CLARA algorithms are 4 clusters and 2 clusters, respectively. However, this study proceeds with 4 clusters, as 2 clusters is considered too few.

**Figure 6.** Comparison of Elbow and Silhouette Methods for PAM and CLARA Algorithms of K-Medoids

The clustering results of regions in North Sumatra using the PAM and CLARA algorithms are presented in Figure 7. Based on the clustering results, the first cluster consists of 4 regions, the second cluster consists of 12 regions, the third cluster consists of 13 regions, and the fourth cluster consists of 4 regions.

**Figure 7.** Clustering Results of Regions Using PAM and CLARA Algorithms of K-Medoids

The evaluation between PAM and CLARA was conducted using the between-cluster proportion criterion ($\text{Between_SS}/\text{Total_SS}$). The evaluation results are presented in Table 6. Based on Table 6, the ratio of Between-Cluster Sum of Squares (BCSS) to Total Sum of Squares (TSS) for PAM ($k = 4$) is 65.56%, indicating that approximately 65.56% of the total variation in the data can be explained by the separation between clusters. Meanwhile, for CLARA ($k=4$), the between-cluster proportion shows a value of 63.88%, meaning that 63.88% of the total variation in the data can be explained by the separation between clusters. These values indicate that the clustering results of PAM ($k=4$) are better than those of CLARA ($k = 4$).

Table 6. Between-Cluster Proportion of PAM and CLARA Algorithms

Clustering Method	Between-Cluster Proportion
PAM ($k = 4$)	65.56%
CLARA ($k = 4$)	63.88%

3.5. Evaluation of the Best Clustering Method

The evaluation of the best clustering method was conducted to compare all clustering methods for grouping regencies/cities in North Sumatra based on socioeconomic indicators in 2024. In this study, the best clustering method was evaluated using the Silhouette Coefficient, Dunn Index, and Connectivity Coefficient. The evaluation results are presented in Table 7.

Table 7. Comparison of Clustering Methods Based on Internal Evaluation

Clustering Method	Connectivity Coefficient	Dunn Index	Silhouette Coefficient
Ward Linkage ($k = 3$)	13.8452	0.1263	0.4268
K-Means ($k = 4$)	23.8913	0.0736	0.3093
PAM ($k = 4$)	26.6690	0.0377	0.3106
CLARA ($k = 4$)	29.9647	0.1114	0.2417

Based on the evaluation results in Table 7, it can be concluded that the best method for grouping the welfare levels of the population by regency/city in North Sumatra Province is the Agglomerative Ward Linkage method, with a maximum Dunn Index of 0.1263, a minimum Connectivity Coefficient of 13.8452, and a maximum Silhouette Coefficient of 0.4268. Therefore, the subsequent interpretation will use the clustering results from the Agglomerative Ward Linkage method.

3.6. Interpretation of Clustering Results from the Best Method

The clustering results from the Agglomerative Ward Linkage method form three clusters with regencies/cities grouped as presented in Table 8. Based on Table 8, it can be concluded that Cluster 1 consists of 4 regencies/cities, where the regions in this cluster are underdeveloped or peripheral areas characterized by a high number of people in poverty, low life expectancy, low mean years of schooling, very low purchasing power, and low population density.

Table 8. Clustering Results of Districts and Cities Based on the Best Method

Cluster	Regency/City
Cluster 1	Nias, Nias Selatan, Nias Utara, Nias Barat
Cluster 2	Mandailing Natal, Tapanuli Selatan, Tapanuli Tengah, Tapanuli Utara, Toba, Labuhan Batu, Asahan, Simalungun, Dairi, Karo, Deli Serdang, Langkat, Humbang Hasundutan, Pakpak Bharat, Samosir, Serdang Bedagai, Batu Bara, Padang Lawas Utara, Padang Lawas, Labuhanbatu Selatan, Labuanbatu Utara, Tanjungbalai, Gunungsitoli
Cluster 3	Sibolga, Pematangsiantar, Tebing Tinggi, Medan, Binjai, Padangsidimpuan

Cluster 2 consists of 23 districts/cities, characterized as areas that are not underdeveloped but have not yet reached the level of major cities. The distinctive features of these areas include moderate life expectancy, expected years of schooling, and mean years of schooling, moderate purchasing power, medium population density, and a fairly varied open unemployment rate. Cluster 3 consists of 6 districts/cities and represents major urban areas with abundant facilities and high education levels, but urban problems such as urban poverty, high unemployment, and low home ownership have emerged. The characteristics of Cluster 3 include high life expectancy, high mean years of schooling, high purchasing power, very high population density, high open unemployment rate, high economic growth, but low home ownership.

Table 9. Mean Values of Each Variable by Cluster

Cluster	Life Expectancy (Years)	Expected Years of Schooling (Years)	Mean Years of Schooling (Years)	Purchasing Power (IDR)	Economic Growth
1	72.13	13.15	6.81	4,083.50	3.81
2	73.45	13.44	9.53	10,503.78	4.77
3	74.46	14.13	11.21	11,041.50	4.44
North Sumatra	73.9	13.49	9.93	11,460	5.03

The characteristics described above were concluded based on the composition of districts/cities grouped within each cluster. To verify this, a univariate analysis was conducted using mean values (Table 9). Based on Table 9, it can be seen that the average life expectancy in Clusters 1 and 2 is below the life expectancy of North Sumatra Province. The Expected Years of Schooling variable for Clusters 1 and 2 is also below the provincial average. The same applies to mean years of schooling, economic growth, open unemployment rate, and population density. Based on this analysis, it can be concluded that Cluster 3 represents urban or major city areas.

Table 9. Mean Values of Each Variable by Cluster (Continued)

Cluster	Total Labor Force (People)	Open Unemployment Rate (%)	Number of People in Poverty (People)	Population Density (People/km ²)	Home Ownership (%)
1	114,225.00	2.24	31,382.50	166.04	90.46
2	232,323.40	4.06	36,628.70	333.07	76.30
3	297,858.70	6.91	43,341.67	4,992.13	58.32
North Sumatra	8,181.16	5.60	1,228.01	215.13	74.75

4. CONCLUSION

The comparison of cluster analysis methods shows that the Ward Linkage method with $k = 3$ is the best method for grouping the socioeconomic indicators of districts/cities in North Sumatra Province. The clustering results indicate that the regions of North Sumatra can be mapped across a spectrum of socioeconomic development: from underdeveloped areas, to developing areas, to metropolitan areas. This classification can serve as a consideration in the formulation of regional development policies, such as: infrastructure and accessibility prioritization for Cluster 1, improvement of human resource quality and industrialization for Cluster 2, and urbanization management and social inequality reduction for Cluster 3. With this approach, regional policies can be more precisely targeted in accordance with the characteristics of each area.

Funding Information

The authors state no funding is involved.

Author Contributions Statement

Dewi Indra Setiawan: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing-original draft, writing-review & editing, visualization, supervision, project administration. Muhammad Yahya Matdoan: Methodology, validation, formal analysis, investigation, data curation, writing-review & editing. All authors discussed the results and contributed to the final manuscript.

Conflict of Interest Statement

Authors state no conflict of interest.

Data Availability

The data that support the findings of this study are publicly available on the North Sumatra BPS Portal and can be accessed at: <https://sumut.bps.go.id/id/statistics-table?subject=519&sortBy=date%2Ctitle&sortOrder=desc%2Casc>

REFERENCES

- [1] BPS, “Welfare Indicators of Indonesia 2018,” BPS, 2018. [Online]. Available: <https://www.bps.go.id/publication/2018/11/28/f6adb407ea72d9b66776a270/indikator-kesejahteraan-rakyat-2018.html>.
- [2] N. Thamrin and A. W. Wijayanto, “Comparison of soft and hard clustering: A case study on welfare level in cities on Java Island,” *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 141–160, 2021.
- [3] Y. L. Dakhi and B. A. Ningsi, “Clustering of districts and cities in North Sumatra Province based on people’s welfare indicators using K-means algorithm,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 993–1003, 2024.
- [4] U. Izzaty, I. R. HG, and D. Devianto, “Clustering of regencies/cities in West Sumatra Province based on community welfare indicators using Silhouette Coefficient validity,” *Jurnal Matematika UNAND*, vol. 9, no. 2, pp. 192–198, 2020.
- [5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice Hall, 2002.
- [6] J. Supranto, *Multivariate Analysis: Meaning and Interpretation*, 2nd ed. Jakarta, Indonesia: Rineka Cipta, 2010.
- [7] P. Govender and V. Sivakumar, “Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019),” *Atmospheric Pollution Research*, vol. 11, no. 1, pp. 40–56, 2020.
- [8] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 2009.
- [9] A. Meiriza and E. Ali, “Comparison of K-means and K-medoids algorithms for grouping BPJS Ketenagakerjaan programs,” *The Indonesian Journal of Computer Science*, vol. 12, no. 2, 2023.
- [10] T. Gupta and S. P. Panda, “A comparison of k-means clustering algorithm and CLARA clustering algorithm on iris dataset,” *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 4766–4768, 2018.
- [11] G. Brock, V. Pihur, S. Datta, S. Datta, et al., “clValid: An R package for cluster validation,” *Journal of Statistical Software*, 2011.
- [12] N. R. Handitia and A. Y. Sofro, “Cluster analysis of sub-districts in Surabaya City based on education data for 2022–2023,” *Jurnal Gaussian*, vol. 13, no. 2, pp. 351–362, 2024.
- [13] S. Soemartini and E. Supartini, “K-means cluster analysis for grouping regencies/cities in West Java based on community indicators,” in *Proc. National Conference on Mathematics Research and Learning*, 2017, pp. 144–154.
- [14] Bappelitbang Provinsi Sumatera Utara, *Rencana Strategis Bappelitbang Provinsi Sumatera Utara Tahun 2024–2026*. Medan, Indonesia: Bappelitbang Provinsi Sumatera Utara, 2024.
- [15] Badan Perencanaan Pembangunan Penelitian dan Pengembangan Provinsi Sumatera Utara, *Rencana Kerja Pemerintah Daerah (RKPD) Provinsi Sumatera Utara Tahun 2024*. Medan, Indonesia: Bappelitbang Provinsi Sumatera Utara, 2024.
- [16] G. Gunadi, “Implementation of the K-means clustering algorithm to analyze print service sales transactions at the Print on Demand (POD) Unit of Gramedia Printing,” *STMIK Widuri*, vol. 8, no. 2, 2022.
- [17] N. W. Hidayat, U. Ghoni, and F. A. Tyas, “Perbandingan algoritma K-means dan K-medoids dalam pengelompokan kabupaten/kota berdasarkan Indeks Pembangunan Manusia di Provinsi Jawa Tengah,” in *Proc. Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media (CENTIVE)*, vol. 3, no. 1, pp. 663–672, Dec. 2023.
- [18] N. T. Luchia, H. Handayani, F. S. Hamdi, D. Erlangga, and S. F. Octavia, “Comparison of K-means and K-medoids on poor data clustering in Indonesia,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 2, pp. 35–41, Oct. 2022.
- [19] M. E. Ariyani and A. W. Wijayanto, “Perbandingan kinerja clustering non-hierarchical pada Indeks Daya Saing Daerah di Provinsi Jawa Tengah Tahun 2022,” *Jurnal Ilmu Komputer*, vol. 17, no. 2, 2024.
- [20] A. Ikhrum, S. Mutia, P. Akbar, and C. Umam, “Penerapan algoritma K-medoids untuk pengelompokan provinsi di Indonesia berdasarkan indikator kemiskinan: Akses perumahan dan kesejahteraan,” *Emerging Statistics and Data Science Journal*, vol. 3, no. 2, pp. 583–596, 2025.