International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 4, No. 2, September 2025, pp. 121∼134
ISSN: 2828-5611
DOI: doi.org/10.30812/ijecsa.v4i2.5363

Clustering Regency in Kalimantan Island Based on People's Welfare Indicators Using Ward's Algorithm with Principal Component Analysis Optimization

Eva Lestari Ningsih, Siti Mahmuda, Memi Nor Hayati
Universitas Mulawarman, Samarinda, Indonesia

Article Info

Article history:

Received July 18, 2025
Revised July 28, 2025
Accepted August 2, 2025

Keywords:

People's Welfare Indicators
Principal Component Analysis
Silhouette Coefficient
Ward's Algorithm

ABSTRACT

Cluster analysis is used to group objects based on similar characteristics, so that objects in one cluster are more homogeneous than objects in other clusters. One method that is widely used in hierarchical clustering is Ward's algorithm. This method works by minimizing the sum of squared distances between objects in one cluster (within-cluster variance) to produce optimal clustering. However, one important assumption in using this method is that there is no high correlation between variables, or in other words, the data must be free from multicollinearity. Multicollinearity can distort distance calculation, resulting in less accurate clustering results. To overcome this problem, a Principal Component Analysis (PCA) approach is used to reduce the dimension and eliminate the correlation between variables by forming several mutually independent principal components. This research aims to cluster 56 districts/cities in Kalimantan Island based on 19 indicators of people's welfare in 2023, using Ward's algorithm optimized through PCA. Validation of clustering results is done using the Silhouette Coefficient value to assess the quality of clustering. This research method is a combination of Principal Component Analysis (PCA) and hierarchical clustering using Ward's algorithm. PCA was applied to reduce 19 welfare-related indicators into four principal components that retained most of the essential information in the dataset. The clustering process based on these components resulted in two optimal clusters, as determined by a Silhouette Coefficient value of 0.651, which indicates a moderately strong cluster structure. The results of this research indicate that the first cluster comprises 47 districts/cities characterized by relatively low welfare levels. In comparison, the second cluster comprises 9 districts/cities with comparatively higher welfare conditions. These findings imply the existence of considerable disparities in welfare among regions on Kalimantan Island. The results can be used as a reference for policymakers in formulating more targeted and equitable development strategies.

Copyright ©2025 The Authors.
This is an open access article under the CC BY-SA license.

Corresponding Author:

Siti Mahmuda,
Universitas Mulawarman, Samarinda, Indonesia,
Email: sitimahmuda@fmipa.unmul.ac.id

How to Cite: E. L. Ningsih, S. Mahmuda, and M. N. Hayati, "Clustering Regency in Kalimantan Island Based on People's Welfare Indicators Using Ward's Algorithm with Principal Component Analysis Optimization", *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 4, no. 1, pp. 121–134, Sep. 2025. doi: 10.30812/ijecsa.v4i2.5363.

1. INTRODUCTION

Cluster analysis is a technique that combines objects based on similar characteristics. Objects in one group have a high level of similarity (homogeneity) while objects in other groups have many differences (heterogeneity). The main objective of cluster analysis is to group objects into several groups that have significant differences, so that each group consists of objects that have relatively similar characteristics. Cluster analysis has two approaches: non-hierarchical and hierarchical [1]. The hierarchical method starts by grouping two or more objects with the closest similarity, then moves on to other objects with the second closest similarity, and so on, until the cluster forms a hierarchical cluster. On the other hand, non-hierarchical methods start by finding the desired number of clusters and then perform the clustering process without following a hierarchical process [2]. The hierarchical method also has the advantage of making the clustering formation visually visible, so that it is easy to understand and can help in choosing the optimal cluster. Several algorithms can be used in hierarchical methods, including single linkage, complete linkage, average linkage, centroid linkage, and ward [3].

Ward's algorithm is a popular cluster analysis technique that focuses on minimizing within-group variance by minimizing the sum of squares of differences between data within each cluster [4]. When compared to other hierarchical approaches, such as single linkage, complete linkage, and average linkage, Ward has the advantage of forming homogeneous groups by minimizing variation within clusters. When compared to other hierarchical methods such as single linkage, complete linkage, and average linkage, Ward's method produces clusters that are more compact and clearly separated. This advantage is supported by internal validation results using the Silhouette Coefficient and the Davies–Bouldin Index, which show that objects within each cluster have a high degree of similarity and good separation between clusters. Moreover, stability validation metrics, including the Average Proportion of Non-overlap (APN) and the Average Distance Between Means (ADM), also demonstrate that the clustering results remain stable even under minor data perturbations [5].

There is an assumption of cluster analysis that needs to be considered, namely, the presence or absence of a strong correlation relationship between research variables. This can cause the analysis results to be inaccurate and difficult to understand properly if there is a strong correlation between the independent variables. If this happens, one way to overcome this correlation is to use Principal Component Analysis (PCA) [6]. PCA is a statistical method for reducing the dimensionality of data by transforming correlated variables into several independent principal components without losing important information. PCA is useful when data has many correlated variables and the calculations are based on the eigenvalues and eigenvectors of the covariance or correlation matrix [7]. After the clustering results are obtained, the next step is to perform validation to assess the quality of the clusters formed. One commonly used cluster validation method is the Silhouette Coefficient (SC). SC measures the quality of clustering by combining the concepts of cohesion, namely the closeness between objects in one cluster, and separation, namely the separation between clusters. This method is also used to assess the extent to which each object actually fits into the cluster it belongs to [8].

Cluster analysis can be used in a variety of fields, one of which is to categorize regions according to the level of people's welfare. People's welfare is one of the important measures in assessing the quality of life and the level of development of a region. Welfare can be defined as the fulfillment of the community's basic needs, which are reflected through various indicators, such as employment, poverty, education, health, population, income, consumption patterns, and housing and environmental conditions [9]. According to [10], The level of welfare is a key indicator of successful development because it includes aspects of security, prosperity, and quality of life. In Indonesia, improving welfare is the main objective of economic development as stated in the Preamble of the fourth paragraph of the 1945 Constitution. One of the areas of concern is the island of Kalimantan. According to the Central Bureau of Statistics (2019), Kalimantan has shown significant progress in recent years, with the population expected to reach 20 million by 2025. However, this growth poses new challenges, such as unequal population distribution, which could widen the income gap between regions. Therefore, understanding the specific characteristics and problems in each province and formulating appropriate development strategies are important steps in realizing equitable welfare for its people [11].

Several studies related to cluster analysis have been conducted previously. The study by [12] classified sub-districts in West Sulawesi based on education indicators. Before clustering, the researchers examined the assumption of no strong correlation between variables. Due to multicollinearity in three variable pairs, Principal Component Analysis (PCA) was applied before performing Ward's method, which resulted in three clusters. However, some gaps have not been resolved by previous research, namely the absence of cluster validation methods such as the Silhouette Coefficient, and the limited application of clustering to a broader and more diverse set of welfare indicators. The study by [13] grouped Indonesian provinces based on the economic impact of the Covid-19 pandemic using various hierarchical methods and distance measures. Ward's method with Euclidean distance produced the best result, forming six clusters with a Silhouette value of 0.48. Nonetheless, this study focused only on macroeconomic impacts and did not incorporate dimensionality reduction techniques like PCA or explore regional welfare disparities at the district/city level. The difference between this research and the previous one is the integration of PCA for handling multicollinearity, the use of the Silhouette Coefficient for cluster validation, and the focus on multidimensional welfare indicators at the district/city level in Kalimantan.

This research aims to determine the number of principal components (PCs) from the results of variable reduction using Principal Component Analysis (PCA), obtain the optimal number of clusters (K) using the PCA-based Ward algorithm validated by the Silhouette Coefficient, and generate district/city groupings in Kalimantan based on indicators of people's welfare. The contribution of this research to the development of science lies in its methodological integration for handling high-dimensional socioeconomic data in clustering analysis. Practically, this study contributes by providing data-driven insights into the welfare levels of districts/cities in Kalimantan, which the government can use as a basis for formulating more targeted and effective development policies.

2. RESEARCH METHOD

This research is a non-experimental study that uses secondary data in the form of people's welfare indicators in 2023 obtained from the official website of the Central Statistics Agency (BPS) at https://www.bps.go.id. The data used covers the five provinces on the island of Kalimantan, namely West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, and North Kalimantan. The indicators used represent various dimensions of welfare, such as poverty, education, health, employment, and economic conditions. This data is the basis for conducting a cluster analysis to group kabupaten/kota in Kalimantan based on similar welfare characteristics. The variables used in this study are indicators of people's welfare, as shown in Table 1. Meanwhile, the research flow is shown in Figure 1.

Table 1. Research Variables

Notation	Variables	Notation	Variables
X_1	Open Unemployment Rate	X_{11}	Percentage of Households with Access to Adequate Sanitation
X_2	Labor Force Participation Rate	X_{12}	Health Complaints
X_3	Total Labor Force	X_{13}	Population Density
X_4	Poverty Line	X_{14}	Gross Regional Domestic Product per capita at constant prices
X_5	Percentage of Poor Population	X_{15}	Per capita expenditure
X_6	Poverty Depth Index	X_{16}	Average Expenditure per Capita on Food per Month
X_7	Poverty Severity Index	X_{17}	Tenure Status of Owned Residential Building
X_8	Average Years of Schooling	X_{18}	Number of Crimes
X_9	Expected Years of Schooling	X_{19}	Human Development Index
X_{10}	Percentage of Households with Access to Adequate Drinking Water		

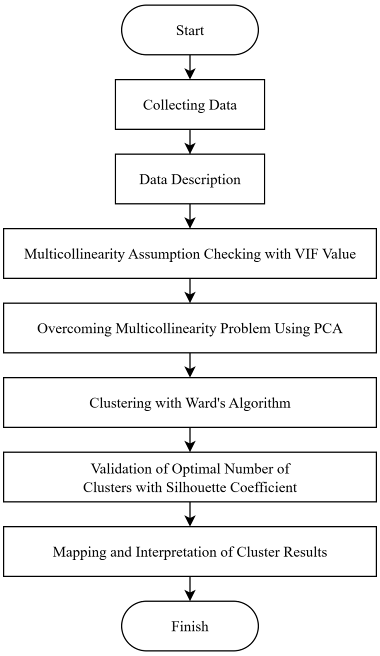
Start
Collecting Data
Data Description
Multicollinearity Assumption Checking with VIF Value
Overcoming Multicollinearity Problem Using PCA
Clustering with Ward's Algorithm
Validation of Optimal Number of Clusters with Silhouette Coefficient
Mapping and Interpretation of Cluster Results
Finish

Figure 1. Research Flowchart

The analysis technique in this study was to use total sampling, where all members of the population were used as samples, and the following data analysis methods were used:

1. Conduct descriptive statistical analysis.
2. Checking the assumption of correlation between research variables using the Variance Inflation Factor (VIF) value using Equation (1). If there is a high correlation, PCA is performed. But if the assumptions are met, cluster analysis is done directly.

$$VIF_l = \frac{1}{1-R_l^2} \tag{1}$$

where,

$$R_l^2 = 1 - \frac{\sum_{i=1}^{n}\left(x_{il}-\hat{x}_{il}\right)^2}{\sum_{i=1}^{n}\left(x_{il}-\bar{x}_{il}\right)^2}\ ; i = 1,\ 2,\ldots,n \text{ and } l = 1,2,\ \ldots,\ p \tag{2}$$

In the formula, x_{il} indicates the actual observation of the i^{th} object on the l^{th} variable, $\hat{x}_{il}$ is the predicted value, and $\bar{x}_{il}$ is the average value of the observations for the respective variable.

3. Perform PCA analysis with the following steps:

   a. Standardize the Zscore observation data using Equation (3).

$$Z_{il} = \frac{x_{il}-\bar{x}_{il}}{s_l} \tag{3}$$

with the average for each variable using Equation (4).

$$\bar{x}_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}\ ; i = 1,\ 2,\ \ldots,n \text{ and } l = 1,2,\ \ldots,\ p \tag{4}$$

and standard deviation using Equation (5).

$$s_l = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{il}-\bar{x}_l\right)^2} \tag{5}$$

where Z_{il} represents the standardized value of the ith observation on the lth variable, x_{il} is the original (unstandardized) value of the i$^{\text{th}}$ observation on the lll$^{\text{th}}$ variable, $\bar{x}_l$ is the mean of the l^{th} variable, and s_l is the standard deviation of the lth variable.

   b. Calculate the correlation coefficient between variables using Equation (6).

$$r_{lm} = \frac{n\left(\sum_{i=1}^{n} Z_{il}Z_{im}\right) - \left(\sum_{i=1}^{n} Z_{il}\right)\left(\sum_{i=1}^{n} Z_{im}\right)}{\sqrt{n\left(\sum_{i=1}^{n} Z_{il}^2\right) - \left(\sum_{i=1}^{n} Z_{il}\right)^2}.\sqrt{n\left(\sum_{i=1}^{n} Z_{im}^2\right) - \left(\sum_{i=1}^{n} Z_{im}\right)^2}}\ ; l,\ m = 1,\ 2,\ \ldots,p \tag{6}$$

where r_{lm} is the correlation coefficient between the standardized data of the lth and mth variables, Z_{il} represents the standardized value of the ith observation on the l^{th} variable, Z_{im} is the standardized value of the ith observation on the mth variable, and n is the total number of data observations used.

   c. Create a correlation matrix based on the correlation coefficient using Equation (7).

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p1} & \cdots & r_{pp} \end{bmatrix} \tag{7}$$

Based on [14], to provide an assessment of whether or not the relationship between a variable is strong as shown in Table 2.

Table 2. Research Variables

Interval Coefficient	Criteria
0.000-0.199	Very weak (negligible)
0.200-0.399	Weak
0.400-0.599	Medium
0.600-0.799	Strong
0.800-1.000	Very Strong

d. Calculate eigenvalues based on Equation (8) and eigenvectors according to Equation (9).

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0 \tag{8}$$

and eigenvectors as per Equation (9).

$$(\mathbf{R} - \lambda\mathbf{I})\,\vec{v} = 0 \tag{9}$$

with,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{10}$$

where λ represents the eigenvalue, $\vec{v}$ denotes the eigenvector, and $\mathbf{I}$ is the identity matrix.

e. Determine the number of principal components formed by looking at the eigenvalue ≥ 1.

f. Form a correlation matrix component that shows the magnitude of the variable correlation to the component score formed using Equation (11).

$$r_{X_l, PC_t} = \vec{v}_{it}\sqrt{\lambda_t} \tag{11}$$

where, PC_t refers to the t^{th} principal component, X_l represents the l^{th} original variable, $\vec{v}_{it}$ denotes the ith eigenvector of the tth principal component, and λ_t is the eigenvalue associated with the tth principal component.

g. Calculate the principal component (PC) score using Equation (12).

$$PC_{it} = \vec{v}_{i1}Z_{i1} + \vec{v}_{i2}Z_{i2} + \ldots + \vec{v}_{ip}Z_{ip} \tag{12}$$

4. Perform clustering on observation objects using the ward algorithm with the following steps:

a. Calculate the distance matrix between objects using the squared Euclidean distance according to Equation (13).

$$\begin{aligned} d^2(x_{il}, x_{jl}) &= \textstyle\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \;\; ; l = 1,\ 2,\ \ldots, p \\ &= (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2 \end{aligned} \tag{13}$$

where, $d^2(x_{il}, x_{jl})$ represents the squared Euclidean distance between the ith and jth observation data, p is the number of variables, x_{il} refers to the ith observation on the l^{th} variable, and x_{jl} denotes the j^{th} observation on the l^{th} variable.

b. Calculate the sum of Square Error (SSE) value for the combination of two pairs of clusters using Equation (14) and then select the smallest value to combine.

$$\begin{aligned} I_{ij} = SSE_{ij} \quad &= \tfrac{n_i \times n_j}{n_i + n_j} \textstyle\sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2 \\ &= \tfrac{n_i \times n_j}{n_i + n_j} d^2\left(x_{il}, x_{jl}\right) \end{aligned} \tag{14}$$

where, I_{ij} represents the distance between the i^{th} and j^{th} observation data, n_i is the number of members in the i^{th} cluster, and n_j is the number of members in the j^{th} cluster.

c. Continue until n clusters are formed.

5. Calculate the silhouette coefficient value to see the optimal number of clusters with the following steps:

a. Calculate the average distance of the ith object to all objects in the same cluster using Equation (15).

$$a_i = \frac{1}{n_k - 1} \sum_{j=1}^{n_k - 1} d(i, j) \; ; k = 1,\ 2, \ldots, K \tag{15}$$

b. Calculate the average distance of the i^{th} object with each different cluster object according to Equation (16) and then select the smallest value with Equation (17).

$$d_i\left(k\right) = \frac{1}{n_k} \sum_{j=1}^{n_k} d(i, j) \tag{16}$$

$$b_i = \min d_i\left(k\right) \tag{17}$$

c. Calculate the silhouette value for each i^{th} object denoted by $SC_1(i)$ using Equation (18).

$$SC_1\left(i\right) = \frac{b_i - a_i}{\max a_i, b_i} \; ; i = 1,\ 2, \ldots, n \tag{18}$$

d. Calculate the average value $SC_1(i)$ of all objects belonging to the cluster into $SC_2(k)$ with Equation (19).

$$SC_2\left(k\right) = \frac{1}{n_k} \sum_{j=1}^{n_k} SC_1(i) \tag{19}$$

e. Calculate the SC_{global} value according to the formula in Equation (20).

$$SC_{global} = \frac{\sum_{k=1}^{K} \left(n_k \times SC_2(k)\right)}{\sum_{k=1}^{K} n_k} \tag{20}$$

Where, a_i is the average distance of the ith observation data to all other data within the same cluster, while $d_i(k)$ denotes the average distance of the i^{th} observation to all data in other clusters. The minimum of these distances is represented by b_i. The silhouette coefficient for each ith observation is denoted by SC_i, while $SC_2(k)$ represents the silhouette coefficient for each kth cluster. The global silhouette coefficient is denoted as SC_{global}. Additionally, n_k is the number of data points in the kth cluster, and K indicates the total number of clusters.

To interpret the results of cluster evaluation with the silhouette coefficient method, the categories shown in Table 3 are used as follows [15].

Table 3. Categories Silhouette Coefficient

Silhouette Coefficient	Interpretation
0.71 – 1.00	Strong cluster
0.51 – 0.70	Good or suitable cluster
0.26 – 0.50	Weak cluster
$<$ 0.25	Cannot be called a cluster

6. Mapping and interpretation of cluster results.

3. RESULT AND ANALYSIS

3.1. Descriptive Statistical Analysis

Descriptive statistics are used to describe and present data in a concise and systematic form, making it easier to understand the information contained therein. This includes measures such as minimum, maximum, mean and standard deviation values for each welfare indicator. These summary statistics provide an initial picture of the distribution and diversity of the data, which is important as a first step before further analysis is undertaken. The results obtained are shown in Table 4.

Table 4. Descriptive Statistics

Variable	Minimum	Maximum	Mean	Standard Deviation
X_1	2.070	8.920	4.339	1.722
X_2	62.950	75.880	69.438	3.772
X_3	15,131.000	654,362.000	176,190.589	130,620.661
X_4	365,262.000	854,967.000	586,288.018	110,300.595
X_5	2.310	11.380	5.889	2.142
X_6	0.140	2.470	0.745	0.437
X_7	0.010	0.830	0.169	0.141
X_8	6.350	11.650	8.576	1.184
X_9	11.340	15.390	12.897	0.848
X_{10}	48.980	99.910	79.222	13.831
X_{11}	56.660	98.060	81.696	10.609
X_{12}	7.020	39.420	25.549	7.455
X_{13}	2.000	6,775.000	374.964	1,187.802
X_{14}	10,437.000	229,770.000	54,660.036	49,010.091
X_{15}	7,787.000	17,659.000	11,567.982	2,237.866
X_{16}	539,231.290	1,084,778.000	779,403.037	110,652.659
X_{17}	65.360	96.890	84.585	8.184
X_{18}	25.000	1,805.000	382.893	365.576
X_{19}	66.060	82.320	72.241	4.136

3.2. Checking the Assumption of Non-Multicollinearity

Multicollinearity checking is done using the Variance Inflation Factor (VIF) value to determine whether there is a high linear relationship between independent variables, which can interfere with the further analysis process, especially in cluster analysis. The calculation results are shown in Table 5.

Table 5. Descriptive Statistics

$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	$\mathbf{X_7}$	$\mathbf{X_8}$	$\mathbf{X_9}$	$\mathbf{X_{10}}$
2.953	3.655	6.223	3.277	**11.364***	**52.083***	**29.155***	**13.947***	6.812	2.256
$\mathbf{X_{11}}$	$\mathbf{X_{12}}$	$\mathbf{X_{13}}$	$\mathbf{X_{14}}$	$\mathbf{X_{15}}$	$\mathbf{X_{16}}$	$\mathbf{X_{17}}$	$\mathbf{X_{18}}$	$\mathbf{X_{19}}$	
3.288	1.602	3.075	2.961	8.532	2.488	4.148	**12.987***	**31.746***	

Notes: (*) VIF value $\geq$ 10

Based on Table 5, six variables $X_5, X_6, X_7, X_8, X_{18}$, and X_{19} have Variance Inflation Factor (VIF) values greater than 10 indicating the presence of multicollinearity. Multicollinearity can lead to distortion in cluster formation. Therefore, Principal Component Analysis (PCA) was applied to reduce the number of variables and eliminate correlations among them before performing cluster analysis.

3.3. Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction method that forms new uncorrelated variables to overcome multicollinearity. The application steps are as follows:

1. Conduct data standardization
Data standardization is necessary to equalize the scale of variables, prevent the dominance of large-scale variables, and ensure more accurate analysis results. This study uses Z_{score} standardization in Equation (3), which changes the data so that it has a mean of 0 and a standard deviation of 1, so that the variables are in the same range (see in Table 6).

Table 6. Data Standardization Results

Districts/Cities	Z_1	Z_2	$\cdots$	Z_{19}
Sambas	0.407	0.870	$\cdots$	-0.397
Bengkayang	-0.824	0.926	$\cdots$	-0.655
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Kota Tarakan	0.529	-0.874	$\cdots$	1.279

2. Calculating the correlation coefficient
To determine how strong the relationship between two variables is, a correlation coefficient is calculated. Each variable is paired with another variable, resulting in correlation values for each pair. These values are then organized into a matrix called a correlation matrix, which provides an overall picture of the linear relationship between variables in the dataset. The correlation value is obtained using Equation (7).

3. Form a correlation matrix
After the correlation coefficient between variables is calculated, the correlation matrix is formed based on Equation (7) as follows:

$$\mathbf{R} = \begin{bmatrix} 1 & -0.616 & 0.358 & \cdots & 0.579 \\ -0.616 & 1 & -0.460 & \cdots & -0.565 \\ 0.358 & -0.460 & 1 & \cdots & 0.463 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.579 & -0.565 & 0.463 & \cdots & 1 \end{bmatrix}_{19\times 19}$$

Referring to Table 2, the correlation between the open unemployment rate (X_1) and the labor force participation rate (X_2) is -0.616. This value indicates a strong negative relationship between the two variables. This means that when the labor force participation rate increases, the open unemployment rate tends to decrease and vice versa.

4. Determine eigenvalues and eigenvectors
The eigenvalue is calculated based on Equation (8) and the eigenvector from Equation (9). The results of the eigenvalue (λ) calculation are shown below.

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \vdots \\ \lambda_{19} \end{pmatrix} = \begin{pmatrix} 8.048 \\ 2.481 \\ 1.940 \\ 1.292 \\ 0.941 \\ \vdots \\ 0.011 \end{pmatrix}_{19\times 1}$$

Based on the eigenvalue, nineteen eigenvalues were obtained. The formation of principal component (PC) is selected based on the criteria $\lambda \geq 1$. So from the nineteen eigenvalues obtained, only four eigenvalues meet the criteria, namely $\lambda_1 = 8.048$, $\lambda_2 = 2.481$, $\lambda_3 = 1.940$, and $\lambda_4 = 1.292$. Furthermore, the eigenvectors are obtained as follows:

$$\bar{v}_1 = \begin{pmatrix} 0.232 \\ -0.218 \\ 0.177 \\ \vdots \\ 0.366 \end{pmatrix}_{19\times 1}, \bar{v}_2 = \begin{pmatrix} -0.087 \\ 0.137 \\ -0.258 \\ \vdots \\ -0.051 \end{pmatrix}_{19\times 1}, \bar{v}_3 = \begin{pmatrix} -0.199 \\ 0.167 \\ -0.344 \\ \vdots \\ 0.036 \end{pmatrix}_{19\times 1}, \bar{v}_4 = \begin{pmatrix} -0.159 \\ 0.441 \\ 0.121 \\ \vdots \\ 0.039 \end{pmatrix}_{19\times 1}$$

5. Calculating the components of the correlation matrix
The correlation component describes how strong the relationship is between a variable and the resulting principal component score. This component is the result of projecting the original data into a new space formed by eigenvectors. Each object (district/city) has a score on each component that can be used as a concise representation of the original data. The correlation matrix components are obtained based on Equation (11).

6. Form the principal component equation
The next step is to form the equation of each principal component (PC). Based on the eigenvector results, Equation (12) can be written as follows:

$$\begin{aligned} PC_{i,1} &= 0.232Z_{i,1} - 0.218Z_{i,2} + \ldots + 0.335Z_{i,19} \\ PC_{i,2} &= -0.087Z_{i,1} + 0.137Z_{i,2} + \ldots - 0.051Z_{i,19} \\ PC_{i,3} &= -0.199Z_{i,1} + 0.167Z_{i,2} + \ldots + 0.035Z_{i,19} \\ PC_{i,4} &= -0.159Z_{i,1} + 0.441Z_{i,2} + \ldots + 0.039Z_{i,19} \end{aligned}$$

7. Calculating the principal component score
Based on the PC equation that has been formed, the results of the PC score calculation are shown in Table 7.

Table 7. Reduction Results Using PCA

Districts/Cities	PC_1	PC_2	PC_3	PC_4
Sambas	-1.517	-0.269	-2.232	2.332
Bengkayang	-2.632	0.324	-0.944	1.062
⋮	⋮	⋮	⋮	⋮
Kota Tarakan	3.572	-1.027	1.018	-0.099

3.4. Clustering with Ward's Algorithm

Clustering is performed using a new dataset of PCA-reduced results. The stages in clustering using the ward algorithm are as follows:

1. Calculating the squared Euclidean distance matrix
The distance between objects in the observation can be calculated using the squared Euclidean distance, which is by summing the squares of the differences of each variable value between two objects. This distance is used as the basis in determining the similarity between objects. The results of the calculation of the Euclidean square distance matrix using Equation (13) are shown in Table 8.

Table 8. Squared Euclidean Distance Matrix

Districts/Cities	Sambas	Bengkayang	...	Kota Tarakan
Sambas	0	4.866	⋯	42.940
Bengkayang	4.866	0	⋯	45.509
⋮	⋮	⋮	⋱	⋮

Districts/Cities	Sambas	Bengkayang	. . .	Kota Tarakan
Kota Tarakan	42.940	45.509	· · ·	0

2. Calculating the SSE value for the combination of two pairs of clusters
   After obtaining the distance matrix between objects, the next step is to calculate the Sum of Squared Errors (SSE) value for each possible merger of two clusters. SSE is used to measure the total variation in the cluster, where a smaller value indicates a higher level of closeness or similarity of characteristics between objects. The calculation of the SSE value based on Equation 14 revealed that the merger with the smallest SSE value first occurred between Hulu Sungai Selatan Regency and Hulu Sungai Tengah Regency, which amounted to 0.099. In iteration 2, the cluster had the smallest SSE value when merged with Hulu Sungai Utara District, which amounted to 0.170. This low SSE value indicates that the three districts have high similarity and form a relatively homogeneous cluster. This merging pattern can also be seen visually in Figure 2, the dendrogram of the clustering results using the ward algorithm, where the three districts are joined in the same branch with a relatively short separation distance.

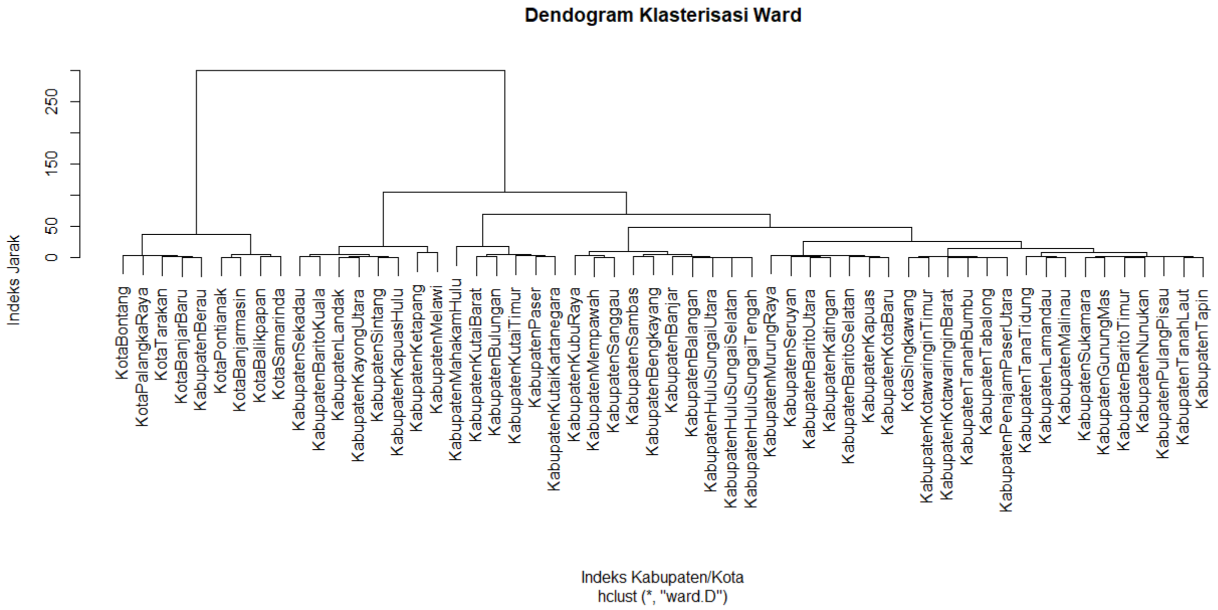

Figure 2. Clustering Dendrogram Using Ward's Algorithm

3.5. Validation of the Optimal Number of Clusters Using the Silhouette Coefficient

Cluster validation aims to assess the quality of the clustering that has been done and determine the optimal number of clusters. One commonly used internal validation method is the Silhouette Coefficient (SC), which measures how similar an object is to its own cluster compared to other clusters. Given an example of calculation for the number of two clusters with the following steps:

1. Calculate the average distance of the i^{th} object to all objects in the same cluster. An example of calculating the average distance from the first data, namely Sambas Regency, based on Equation 15, is as follows.

$$\begin{aligned} a_{Sambas} &= \tfrac{1}{46}\left(4.866 + 10.099 + \ldots + 14.183\right) \\ &= 20.034 \end{aligned}$$

2. Calculate the average distance from Sambas Regency to all objects in other clusters based on Equation 16 as follows.

$$\begin{aligned} d_{Sambas}\left(1\right) &= \tfrac{1}{9}\left(75.692 + 50.655 + \ldots + 49.940\right) \\ &= 61.285 \end{aligned}$$

Because in this calculation, there are only two clusters, the value of $d_i\,(k) = b_i$.

$$b_{Sambas} = 61.285$$

3. Calculate the SC value by using the a_{sambas} and b_{sambas} values previously obtained in accordance with Equation (18).

$$\begin{aligned} SC_1(Sambas) &= \frac{b_{Sambas} - a_{Sambas}}{\max a_{Sambas}, b_{Sambas}} \\ &= \frac{61.285 - 20.034}{\max 20.034\ ,\ 61.285} \\ &= 0.673 \end{aligned}$$

The calculation of SC values for other districts/cities was done using the same formula as for the Sambas District. The results of the SC value for each district/city are presented in Table 9.

Table 9. Local SC Value of Each District/City

Districts/Cities	**Nilai SC**
Sambas	0.673
Bengkayang	0.806
Landak	0.782
⋮	⋮
Kota Tarakan	0.598

Next, calculate the Local SC average of all objects in the cluster using Equation (19).

$$SC_2\,(1) = \frac{1}{47}(0.673 + 0.806 + \ldots + 0.802) = 0.642$$

$$SC_2\,(2) = \frac{1}{9}(0.766 + 0.738 + \ldots + 0.598) = 0.699$$

4. Calculating the overall cluster average using Equation (20).

$$\begin{aligned} SC_{global} &= \frac{\sum_{k=1}^{2}(n_k \times SC_2(k))}{\sum_{k=1}^{2} n_k} \\ &= \frac{(n_1 \times SC_2(1)) + (n_2 \times SC_2(2))}{n_1 + n_2} \\ &= \frac{(47 \times 0.642) + (9 \times 0.699)}{47 + 9} \\ &= \frac{36.462}{56} \\ &= 0.651 \end{aligned}$$

Based on the calculation results, the global Silhouette Coefficient value for the number of two clusters is 0.651. This value indicates that the clustering results have good quality, where each object tends to be closer to its own cluster members and there is a clear separation between clusters. Similar calculations were also carried out for the number of clusters of three, four, and five. The results of the Silhouette Coefficient value for each number of clusters are presented in Table 10.

Table 10. Comparison of Validation Results Based on SC Value

Number of Clusters	**SC Value**	**Interpretation**
2	0.651	Good
3	0.496	Weak
4	0.504	Weak
5	0.516	Good

Based on Table 10, the highest Silhouette Coefficient value was obtained in the formation of two clusters, which amounted to 0.651. Therefore, it can be concluded that the most optimal grouping of districts/cities in Kalimantan Island based on community welfare indicators is two clusters.

3.6. Mapping and Interpretation of Cluster Results

After the clustering process is carried out and the optimal number of clusters is determined based on the Silhouette Coefficient value, the next step is to map and interpret the cluster results. Mapping is done to visualize the distribution of areas in each cluster so that spatial patterns between areas can be observed more clearly. The map of the distribution of areas based on the clustering results is presented in Figure 3.

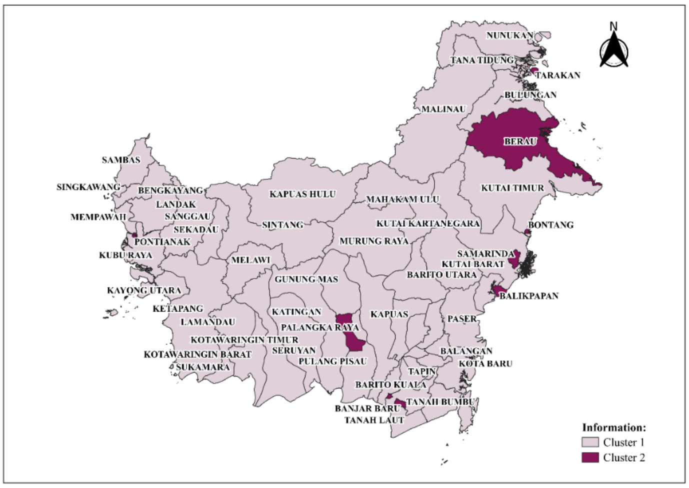

Figure 3. Distribution Map of Optimal Grouping Results

Figure 3 shows the results of grouping 56 districts/cities in Kalimantan into two clusters based on indicators of people's welfare. A total of 47 districts/cities belong to Cluster 1, while 9 districts/municipalities belong to Cluster 2. This map shows the differences in welfare characteristics between regions, where Cluster 2 is dominated by large cities on the island of Kalimantan that have more advanced socio-economic conditions. Furthermore, the characteristics of each cluster were interpreted by comparing the average value of the variables in each cluster. The calculation results are presented in Table 11.

Table 11. Cluster Results

Variable	Cluster		Variable	Cluster	
	1	2		1	2
X_1	3.984	6.190	X_{11}	79.742	91.903
X_2	70.145	65.743	X_{12}	26.061	22.873
X_3	154,870.766	287,527.444	X_{13}	59.021	2,024.889
X_4	558,968.191	728,958.222	X_{14}	48,103.806	88,898.122
X_5	6.180	4.368	X_{15}	10,905.702	15,026.556
X_6	0.797	0.474	X_{16}	763,740.210	861,197.800
X_7	0.184	0.089	X_{17}	87.134	71.270
X_8	8.175	10.670	X_{18}	271.340	965.444
X_9	12.612	14.387	X_{19}	70.730	80.133
X_{10}	76.298	94.493			

Based on Table 11, it can be seen that there are significant differences between the mean values of the variables in cluster 1 and cluster 2. Cluster 2 shows higher mean values compared to cluster 1 in most of the variables analyzed. Cluster 1 has a lower unemployment rate and a slightly higher labor force participation rate compared to cluster 2. However, cluster 1 shows a higher poverty rate with a higher percentage of poor people, poverty depth index, and poverty severity index than cluster 2, indicating greater economic inequality. Cluster 2 excels in education, with a higher average and expected years of schooling, reflecting a better quality of education. Access to basic facilities such as safe drinking water and proper sanitation is also better in cluster 2, which implies better health conditions with a lower percentage of the population experiencing health complaints. Cluster 2 also shows higher economic well-being with higher Gross Regional Domestic Product per capita, expenditure per capita, and expenditure on food. However, population density per square kilometer is higher in cluster 2, reflecting a trend towards more significant population concentration. Although crime rates are higher in cluster 2, the human development index remains higher, indicating a better quality of life. Overall, cluster 1 can be categorized as a cluster with a lower level of welfare, while cluster 2 shows characteristics of an area with a higher level of welfare. The difference in average values between the variables in these two clusters illustrates the inequality in welfare between the grouped regions.

The findings of this research are that the initial 19 variables related to people's welfare were successfully reduced into four principal components using Principal Component Analysis (PCA), which retained most of the essential information in the dataset. Based on these components, clustering using Ward's algorithm produced two optimal clusters, with a Silhouette Coefficient value of 0.651, indicating a fairly strong cluster structure. Cluster 1 consists of 47 districts/cities with relatively lower welfare levels, while Cluster 2 consists of 9 districts/cities with higher welfare levels. These differences are reflected in indicators such as poverty rate, education level, access to basic services, and economic capacity. The results of this research are in line with previous studies, such as [12], which demonstrated that applying PCA before clustering can effectively address multicollinearity and improve the separation between clusters. Furthermore, the study by [13] supports the use of Ward's method with Euclidean distance, which yielded the best Silhouette score compared to other hierarchical clustering methods. These findings strengthen the methodological foundation of this research and confirm the validity of the clustering results obtained.

4. CONCLUSION

Based on the analysis, it can be concluded that the application of Ward's algorithm optimized with Principal Component Analysis (PCA) can reduce the data into four principal components (PC) with eigenvalues ≥ 1, namely λ_1 = 8.048; λ_2 = 2.481; λ_3 = 1.940; and λ_4 = 1.292. These four principal components were used as the basis in the clustering process. The clustering results showed that the optimal number of clusters was two with a Silhouette Coefficient value of 0.651, indicating a fairly good quality of cluster separation. Cluster 1 consists of districts/cities that tend to have higher values on the variables of labor force participation rate, percentage of poor people, poverty depth index, poverty severity index, percentage of population with health complaints, and tenure status of owned residential buildings. These characteristics indicate that cluster 1 is an area group with a relatively low level of welfare. Meanwhile, cluster 2 is characterized by higher values on the variables of open unemployment rate, total labor force, poverty line, average years of schooling, expected years of schooling, access to drinking water and proper sanitation, population density, Gross Regional Domestic Product per capita at constant prices, expenditure per capita, average food expenditure per capita per month, number of crimes, and human development index. In general, these characteristics suggest that cluster 2 comprises areas with a higher level of welfare. The results of this clustering provide an overview of the welfare characteristics between regions in Kalimantan Island, so that it can be used as a reference for the government or policymakers in formulating development programs that are more targeted based on the conditions of each cluster.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to all those who have helped in the preparation and completion of this research. Thanks also go to the Faculty of Mathematics and Natural Sciences, Mulawarman University, for providing facilities and opportunities to conduct this research. Thanks to the editors and anonymous reviewers for their efforts to read carefully, provide interesting comments, and make useful suggestions, which have substantially improved the quality of this work.

REFERENCES

[1] T. Apriliana and E. Widodo, "Analisis Cluster Hierarki untuk Pengelompokan Provinsi di Indonesia berdasarkan Jumlah Base Transceiver Station dan Kekuatan Sinyal," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 3, no. 2, pp. 286–296, Dec. 22, 2023. DOI: 10.24002/konstelasi.v3i2.7143.

[2] D. Andiani, S. D. Rahayu, and A. Riana, "Analisis Teknik non-Hierarki untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Indikator Kesejahteraan Rakyat 2020," *JRMST: Jurnal Riset Matematika dan Sains Terapan*, vol. 2, no. 1, pp. 21–28, Aug. 25, 2022.

[3] P. Purnomo *et al.*, *Analisis Data Multivariat*. Banyumas: Omera Pustaka, 2022.

[4] R. A. Andyani *et al.*, "Aplikasi Metode Ward dengan Berbagai Pengukuran Jarak (Studi Kasus: Klasifikasi Tingkat Perekonomian di Indonesia)," *Jurnal Ilmiah Kampus Mengajar*, pp. 177–190, Oct. 28, 2024. DOI: 10.56972/jikm.v4i2.208.

[5] N. Afira and A. W. Wijayanto, "Analisis Cluster dengan Metode Partitioning dan Hierarki pada Data Informasi Kemiskinan Provinsi di Indonesia Tahun 2019," *Komputika : Jurnal Sistem Komputer*, vol. 10, no. 2, pp. 101–109, Sep. 2, 2021. DOI: 10.34010/komputika.v10i2.4317.

[6] R. Fikri *et al.*, "Pengelompokan Kabupaten/Kota di Indonesia Berdasarkan Informasi Kemiskinan Tahun 2020 Menggunakan Metode K-Means Clustering Analysis," *Seminar Nasional Teknik dan Manajemen Industri*, vol. 1, no. 1, pp. 190–199, Dec. 1, 2021. DOI: 10.28932/sentekmi2021.v1i1.76.

[7] G. Enzellina and D. Suhaedi, "Penggunaan Metode Principal Component Analysis dalam Menentukan Faktor Dominan," *Jurnal Riset Matematika*, vol. 2, no. 2, pp. 101–110, Dec. 20, 2022. DOI: 10.29313/jrm.v2i2.1192.

[8] Y. P. Anggriani, A. Arif, and F. Febriansyah, "Implementasi Algoritma K-Means Clustering dalam Menentukan Blok Tanaman Sawit Produktif pada PT Arta Prigel," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, pp. 1820–1825, Apr. 18, 2024. DOI: 10.36040/jati.v8i2.9225.

[9] N. A. Nabilah, H. Perdana, and E. Sulistianingsih, "Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Kesejahteraan Masyarakat dengan Algoritma K-Means++," *BIMASTER : Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 13, no. 3, Mar. 21, 2024. DOI: 10.26418/bbimst.v13i3.77795.

[10] N. Oktaviani, A. Fauzan, and G. Widyastuti, "Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Tingkat Kesejahteraan Masyarakat Menggunakan K-Means Cluster," *Emerging Statistics and Data Science Journal*, vol. 2, no. 2, pp. 290–301, Jun. 30, 2024. DOI: 10.20885/esds.vol2.iss.2.art22.

[11] W. S. F. Hariadi, S. Martha, and H. Perdana, "Klasterisasi Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Kesejahteraan dengan Two-Step Cluster," *BIMASTER : Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 14, no. 1, pp. 29–36, Jan. 31, 2025. DOI: 10.26418/bbimst.v14i1.91668.

[12] H. Hikmah *et al.*, "Analisis Klaster Pengelompokan Kecamatan di Sulawesi Barat Berdasarkan Indikator Pendidikan," *SAINTIFIK*, vol. 8, no. 2, pp. 188–196, Jul. 29, 2022. DOI: 10.31605/saintifik.v8i2.383.

[13] I. N. Hasanah and A. Sofro, "Analisis Cluster Berdasarkan Dampak Ekonomi di Indonesia Akibat Pandemi Covid-19," *MATHunesa: Jurnal Ilmiah Matematika*, vol. 10, no. 2, pp. 239–248, Jul. 6, 2022. DOI: 10.26740/mathunesa.v10n2.p239-248.

[14] D. Dwitasari and M. R. Yudhanegara, "Analisis Klaster untuk Hubungan antara Kemampuan Komunikasi Matematis dengan Kemampuan Pemecahan Masalah Menggunakan K-Means," *Jurnal Educatio FKIP UNMA*, vol. 10, no. 3, pp. 1025–1033, Sep. 29, 2024. DOI: 10.31949/educatio.v10i3.8234.

[15] G. F. Ibanez, G. W. Wiriasto, and R. Rosmaliati, "Kombinasi Principal Component Analysis dengan Algoritma K-Means untuk Klasterisasi Data Stunting," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 5, no. 1, pp. 131–141, Aug. 22, 2024. DOI: 10.30865/klik.v5i1.1977.