BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 0837–0848
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp0837-0848

ANALYSIS OF RAINFALL IN INDONESIA USING A TIME SERIES-BASED CLUSTERING APPROACH

A'yunin Sofro [1*], Rosalina Agista Riani[2], Khusnia Nurul Khikmah[3], Riska Wahyu Romadhonia[4], Danang Ariyanto[5]

[1,2,3,4,5]*Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*

Surabaya, Jawa Timur, 60231, Indonesia

Corresponding author's e-mail: ***ayuninsofro@unesa.ac.id**

ABSTRACT

Article History:

Received: 7th November 2023
Revised: 5th January 2024
Accepted: 6th March 2024
Published: 1st June 2024

Keywords:

Cluster;
Dynamic Time Warping;
Hierarchical;
Non-hierarchical;
Rainfall.

Indonesia has a tropical climate and has two seasons: dry and rainy. Prolonged drought can cause drought disasters, and rain can cause floods and landslides. According to information from the Meteorology, Climatology, and Geophysics Agency (BMKG), natural disasters such as floods and landslides due to heavy rains have been a severe problem in Indonesia for the past five years. Different regional characteristics can affect the intensity of rain that falls in every province in Indonesia. It can be grouped to determine which provinces have similar characteristics to natural disasters due to rainfall. Later, it can provide information to the government and the public so that they are more aware of natural disasters. So, it is necessary to research and classify provinces in Indonesia for rainfall with cluster analysis. The data used is secondary rainfall data taken from the official BMKG website. Cluster analysis of rainfall in 34 provinces in Indonesia used hierarchical and non-hierarchical methods in this study. The approach that is used in this research limits our clustering of the data. Further research with a machine learning approach is recommended. For the clustering method, the agglomerative hierarchical method includes single, average, and complete linkage. The non-hierarchical method includes k-medoids and fuzzy C-means. The cluster analysis results show that the dynamic time warping (DTW) distance measurement method with the average linkage method has the most optimal cluster results with a silhouette coefficient value of 0.813.

How to cite this article:

A. Sofro, R. A. Riani, K. N. Khikmah, R. W. Romadhonia and D. Ariyanto., "TITLE OF AR ANALYSIS OF RAINFALL IN INDONESIA USING A TIME SERIES-BASED CLUSTERING APPROACH TICLE," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 0837-0848, June, 2024.

Research Article · Open Access

1. INTRODUCTION

Indonesia is a country traversed by the equator. This causes Indonesia to have a tropical climate, which causes Indonesia to receive much sunlight because the sun is above it all year round. In addition, tropical climate areas are also influenced by atmospheric characteristics, where solar heating is extreme, so a lot of upward air movement occurs due to heating (convection). The ocean area is larger than the land area, so it has the potential to form types of clouds that can produce rain [1]. According to information from the Meteorology, Climatology, and Geophysics Agency, for the past five years, natural disasters such as floods and landslides due to heavy rains are still a severe problem in Indonesia. Different regional characteristics can affect the intensity of rain that falls in every province in Indonesia. These characteristics can be influenced by several important factors that influence the rainfall process in Indonesia, such as latitude, wind patterns, distribution of land and water areas, and mountains and high mountains [2]. These different conditions for rainfall in each region in Indonesia have consequences for various other fields, such as agriculture and economics, which have been investigated by previous research in [3]. This problem also poses challenges in assessing rainfall in Indonesia. This problem also provides its challenges in handling it. Therefore, efforts to mitigate rainfall in each region need to be made. This effort can be done by grouping regions with the same characteristics. The grouping is expected to identify the provinces in Indonesia with the highest and lowest rainfall levels to improve appropriate mitigation efforts.

Cluster analysis is a grouping of objects with similar characteristics without eliminating the natural structure of the object so that the resulting groups have meaning, such as patterns or classifications [4]. Cluster analysis aims to group data objectively into homogeneous groups where the similarities of objects within groups are minimized, and dissimilarities between groups of objects are maximized [5]. The application of cluster analysis is growing with the use of time series data (time series). Time series data is data obtained by observing sequences taken sequentially in time, which has a correlation structure between the values in each time series data [6]. Time series cluster analysis groups object based on their time series patterns. However, the choice of distance and the clustering method used must consider the structure of time series data, which is very dynamic. Calculating the distance between time series objects is one of the cornerstones of the time series clustering algorithm [7]. Currently, there have been many developments related to distance measurements as used in this study, namely Dynamic Time Warping (DTW) distance, Short Time Series (STS) distance [8], correlation-based distances [9], and Autocorrelation Function (ACF) distance [10].

Cluster analysis is divided into two methods, namely, the hierarchical method and the non-hierarchical method. Cluster analysis using the hierarchical approach is divided into two, namely the agglomerative method and the divisive method. Hierarchical cluster analysis with agglomerative method consists of single linkage, average linkage, complete linkage, and ward method. The cluster analysis method uses a non-hierarchical method, namely K-Means, K-Medoids, and Fuzzy C-means [11]. Relevant previous studies, on average, only used one of the hierarchical methods with the most frequently used distance measures. According to previous research, the hierarchical method is selected in this study, as this approach is robust to outliers [12]. Another previous research also states that non-hierarchical is the best clustering approach because it is simple and computationally faster [13]. Combining cluster analysis with the hierarchical method with renewable distance measures can be performed to get better results so that it can be seen which hierarchy method and distance measure are the best for applying cluster analysis in this study.

This study uses cluster analysis to determine the grouping of provinces in Indonesia regarding rainfall from January 2018 to December 2022. The grouping of provinces in Indonesia needs to be carried out as information and prevention material for the community. The aim is to increase public awareness of natural disasters. Rainfall data is time series data, so cluster analysis uses distance measures. The distance measures used are Dynamic Time Warping (DTW) distance, Short Time Series (STS) distance, and Autocorrelation Function (ACF) distance. For the cluster analysis method used, the agglomerative hierarchical method includes single linkage, average linkage, and complete linkage. The non-hierarchical approach includes K-Medoids and Fuzzy C-means. From the cluster analysis using four distance measures, the optimal number of clusters will be determined using the elbow method, followed by four agglomerative hierarchical methods. The results of each method and the distance measures will be validated with silhouette coefficients to determine the most optimal cluster results. Provincial groups in Indonesia will later know the most optimal results based on the level of rainfall intensity. So that later, it can be an estimate to find out which areas are vulnerable to natural disasters based on rainfall intensity.

2. RESEARCH METHODS

2.1 Rainfall

Rainfall is the height of rainwater that collects in a flat place and does not evaporate, seep, or flow. The unit of rainfall used in Indonesia is the millimeter (mm). The data is taken from the official website of the Meteorology, Climatology, and Geophysics Agency (BMKG). The data was used for 34 provinces from 34 BMKG observation stations. One millimeter of rainfall means that in one square meter on a flat place, one millimeter of water is accommodated, or one liter of water is accommodated [14]**.** There are four categories of rainfall divided by BMKG, namely [15]:

1. 0 – 100mm/month, low category
2. 100 – 300mm/month, medium category
3. 300 – 500mm/month, high category
4. >500mm/month, very high category

Topographical factors and regional weather systems are essential in the amount and spatial rainfall pattern in an area [16]. In addition, the characteristics of different regions can also affect the intensity of rain that falls in every province in Indonesia. The characteristics of these different areas can be identified by applying a cluster analysis.

2.2 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) distance is an algorithm used to find the distance between two time series with the same or different lengths. Dynamic Time Warping (DTW) distance uses a dynamic programming technique to find all possible paths and choose the one that results in the minimum distance between two-time series using a distance matrix, where each element in the matrix is the minimum cumulative distance of its three neighbors.

Suppose there are two time series $X = x_1, x_2, \dots, x_i, \dots, x_n$ with length n and $Y = y_1, y_2, \dots, y_j, \dots, y_m$ with length m. First, create a matrix D of size $n \times m$ where for each element $n \times m$ in the matrix D is the difference between x_i and y_j, which is written in **Equation (1)** as follows:

$$d_{i,j} = |x_i - y_j| \tag{1}$$

Where $i = 1,2, \dots, n$ and $j = 1,2, \dots, m$. After getting the cumulative distance in the form of d_{ij}, then add the minimum value of the three elements adjacent to the element (i,j), namely $\{d_{(i-1)(j-1)}, d_{(i-1)j}, d_{i(j-1)}\}$, where $0 < i \leq n$ and $0 < j \leq m$ so that matrix E is formed. It can be defined in **Equation (2)** elements (i,j) in matrix E as follows:

$$E_{i,j} = d_{ij} + \min \{d_{(i-1)(j-1)}, d_{(i-1)j}, d_{i(j-1)} \tag{2}$$

After obtaining matrix E, the next step is to determine the distance between the two-time series X and Y with **Equation (3)** as follows [17]:

$$d_{DTW}(X,Y) = \min_{\forall w \in p} \left\{ \sum_{i,j=1}^{k} E_{i,j} \right\} \tag{3}$$

Where, p is a set of all possible warping paths, $E_{i,j}$ is elements (i,j) in matrix E, and k is the length of the warping path.

2.3 Short Time Series (STS) Distance

Short Time Series Distance (STS) distance was introduced by Möller-Levet, Klawonn, Cho, and Wolkenhauer [18], which measures the similarity of DNA microarray time series data. Microarray is a pattern obtained by analyzing the function and expression of many genes simultaneously and in one experiment. Moller's goal was to determine distances capable of capturing differences in form, determined by relative expressive changes and corresponding temporal information. Suppose there are two-time series data $X = \{x_0, x_1, \dots, x_{N-1}\}$ and $Y = \{y_0, y_1, \dots, y_{N-1}\}$ the STS distance is defined in **Equation (4)** as follows:

$$d_{STS}(X,Y) = \sqrt{\sum_{k=0}^{N-1} \left(\frac{y_{k+1}-y_k}{t_{k+1}-t_k} - \frac{x_{k+1}-x_k}{t_{k+1}-t_k} \right)} \tag{4}$$

Where t_k is the time of each point in data X and Y.

2.4 Autocorrelation Function (ACF) Distance

Galeano and Pella **[19]** investigated the relationship of time series data using the Autocorrelation Function (ACF) distance. Suppose there are two time series data, X and Y. Where $\hat{\rho}_X = \left(\hat{\rho}_{1,X}, \hat{\rho}_{2,X} \dots, \hat{\rho}_{k,X}\right)^t$ and $\hat{\rho}_Y = \left(\hat{\rho}_{1,Y}, \hat{\rho}_{2,Y}, \dots, \hat{\rho}_{k,Y}\right)^t$. is the autocorrelation vector representation of the estimation results from lag 1 to lag k where $\hat{\rho}_{i,X} \approx 0$ and $\hat{\rho}_{i,Y} \approx 0$ for $i > k$. The Autocorrelation Function (ACF) is a function used to explain the correlation between X_t and $X_{(t+k)}$ of the same process and only separated by the k-th time lag. Suppose there is time series data $X = X_1, X_2, \dots, X_n$, then ACF can be written in **Equation (5)** as follows:

$$\rho_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{(t+k)} - \bar{X})}{\sum_{i=1}^{n}(X_t - \bar{X})^2} \tag{5}$$

Once the autocorrelation vector is known in **Equation (5)**, the distance between the two-time series can be written in **Equation (6)** as follows:

$$d_{ACF}(X,Y) = \sqrt{\sum_{i=1}^{k}(\hat{\rho}_X - \hat{\rho}_Y)^2} \tag{6}$$

Where, $d_{ACF}(X,Y)$ is a set of all possible warping paths and $\bar{X}$ is the average of the time series.

2.5 Hierarchical Method

The hierarchical method is divided into two methods, namely the agglomerative and divisive methods. Grouping with the agglomerative method starts with separate objects. Then, clusters are formed by grouping objects into increasingly large groups. This process is continued until all objects are members of a single cluster. Whereas the divisive method works the other way around, grouping starts with all objects grouped into a single group. Then, the groups are separated until each object is in a separate group [**20]**.

The agglomerative method consists of single linkage, average linkage, and complete linkage. The single linkage method uses the minimum distance rule between clusters. Determining the distance between clusters using the single linkage method can be done by looking at the distance between the two existing clusters and then choosing the closest distance. Then, look for the closest distance, and the next object is combined into the group. And so on until all the objects are in one large group. If two objects are U and V to be grouped, then a cluster (UV) is obtained with the distance between the two objects denoted d_{UV}. To find the distance between the cluster (UV) and cluster W or other clusters, the Equation used to determine the distance between the two is written in **Equation (7)** as follows:

$$d_{(UV)W} = \min\,(d_{UW}, d_{VW}) \tag{7}$$

Where d_{UW} and d_{VW} values are the minimum distances between clusters U and W and clusters V and W.

The average linkage method is a hierarchical cluster method that groups based on the average between objects. This method is considered more stable and less biased because it uses an average. Calculation of the distance between groups is written in **Equation (8)** as follows:

$$d_{\,(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{n_{(UV)} n_W} \tag{8}$$

Where $n_{(UV)}$ is the number of cluster members (UV), and n_W is the number of cluster members W.

The complete linkage method uses the maximum distance rules between clusters. Grouping with the complete linkage method begins with determining the object with the closest distance. The next step is to combine these objects by looking at the farthest or maximum distance. So that it can be written in **Equation (9)** as follows:

$$d_{(UV)W} = \max\left(d_{UW}, d_{VW}\right) \tag{9}$$

d_{UW} is the farthest distance from clusters U and W, while d_{VW} is the farthest from clusters V and W.

2.6 Non-Hierarchical Method

Non-hierarchical methods are used for object grouping, where the number of clusters to be formed can be determined beforehand. The methods included in the non-hierarchical method are the k-medoids and fuzzy C-means methods. The k-medoids method uses k as the number of initial cluster centers randomly generated at the beginning of the clustering process. Any object closer to the cluster center will be grouped and form a new cluster. The algorithm then randomly determines a new cluster center from each previously formed cluster and recalculates the distance between the object and the resulting new cluster center. The distance between objects i and j is calculated using the dissimilarity measurement function, one of which is the Euclidean Distance Function, as shown in the following form **[21]**:

$$d_{ij} = \sqrt{\sum_{a=1}^{p}(x_{ia} - x_{ja})^2} \tag{10}$$

Where, d_{ij} is a distance between objects i and j. x_{ia} is the value of object i when variable to a. x_{ja} is the value of object j when variable to a and p is total observed variables. The K Medoids algorithm is as follows.

1. Determine the number of clusters are generated.
2. Determines the medoid by as much as k.
3. Calculates the distance between each item and the medoid. The Euclidean distance may be computed using **Equation (10)**.
4. Choose a randomly selected medoid candidate from each cluster.
5. Calculates the distance between each object and the new medoid in each cluster.
6. The total deviation (S) is obtained by subtracting the current total distance from the previous total distance. If $S > 0$, the clustering procedure is complete, and cluster members are acquired from each medoid. If $S < 0$, create a group of k new objects as medoid by updating the object with cluster data.
7. Return to Stage 3 until the medoid shows no change.

Fuzzy C-means (FCM) is a data clustering technique in which the existence of each data point in a cluster is determined by its degree of membership. Following is the FCM clustering algorithm **[22]**:

1. The input data to be clustered X is a matrix of size $n \times p$ (n is the number of data samples and p is attribute of each data). X_{kj} is k-th sample data $(k = 1,2,\ldots,n)$ and j-th attributes $(j = 1,2,3,.,m)$.
2. Determine:
   a. c: number of clusters
   b. m: weighting rank
   c. Max_{Iter}: maximum iteration
   d. ξ: the slightest error expected
   e. $P_0 = 0$: initial Objective Function
   f. $t = 1$: initial iteration
3. Generate random numbers $(\mu_{ik} = 1,2,\ldots,c; k = 1,2,\ldots,n)$ as elements of the initial partition matrix U.

$$U_0 = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \ldots & \mu_{1c}(x_c) \\ \vdots & \vdots & & \vdots \\ \mu_{n1}(x_1) & \mu_{n2}(x_2) & \ldots & \mu_{nc}(x_c) \end{bmatrix} \tag{11}$$

The partition matrix in fuzzy clustering must meet the following conditions:

$$\mu_{ik} = [0,1]; (1 \leq i \leq c; 1 \leq k \leq n) \tag{12}$$

$$\sum_{i=1}^{n} \mu_{ik} = 1; 1 \leq i \leq c \tag{13}$$

$$0 < \sum_{i=1}^{c} \mu_{ik} < c; 1 \leq k \leq n \tag{14}$$

Count the sum of each column (attribute):

$$Q_j = \sum_{i=1}^{c} (\mu_{ik}) \tag{15}$$

with $j = 1,2,3, \dots, m$. Then count:

$$\mu_{ik} = \frac{\mu_{ik}}{Q_j} \tag{16}$$

4. Calculate the k-th cluster center V_{ij}, where $i = 1,2,3, \dots, c$ and $j = 1,2,3, \dots, m$.

$$V_{ij} = \frac{\sum_{k=1}^{n} ((\mu_{ik})^m * X_{kj})}{\sum_{k=1}^{n} (\mu_{ik})^m} \tag{17}$$

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & \ddots & \vdots \\ v_{c1} & \cdots & v_{cm} \end{bmatrix} \tag{18}$$

5. Calculate the objective function at the t-th iteration, P_t, using the following Equation:

$$P_t = \sum_{k=1}^{n} \sum_{i=1}^{c} \left(\left[\sum_{j=1}^{m} (X_{kj} - V_{ij})^2 \right] (\mu_{ik})^m \right) \tag{19}$$

6. Calculate changes in the partition matrix:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{p} (X_{kj} - V_{ij})^2 \right]^{\frac{-1}{p-1}}}{\sum_{i=1}^{c} \left[\sum_{j=1}^{p} (X_{kj} - V_{ij})^2 \right]^{\frac{-1}{p-1}}} \tag{20}$$

7. Check stop condition:
   a. If $(|Pt - Pt - 1| < \xi)$ or ($t <$ maximum iteration) then stop;
   b. If not, then $t = t + 1$. Repeat step 4.

2.7 Elbow Method

The elbow method is used to determine how many clusters to choose by looking at the percentage of the results of the comparison between the number of clusters that will form an elbow at a point [23]. The elbow method works by selecting a cluster value and then adding the cluster value to be used as a data model in determining the best cluster, and the resulting calculation presentation becomes a comparison between the numbers of added clusters. The results of the different percentages of each cluster value can be shown using graphics as a source of information. If the value of the first cluster with the value of the second cluster gives a corner in the graph or the value has the most significant decrease, then the value of the cluster is the best [24]. To get a comparison to calculate the SSE (Sum Square Error) of each cluster value. The greater the number of K clusters, the smaller the SSE value. The algorithm for obtaining SSE values is like Equation (21) below [25]:

$$SSE = \sum_{K=1}^{K} \sum_{X_i} \| x_i - c_k \|^2 \tag{21}$$

Where, K is the number of c clusters, x_i is object data distance to i, and c_k is the center of the i-th cluster.

2.8 Silhouette Coefficient

The silhouette coefficient measures accuracy in grouping a time series and is commonly used to determine grouping quality. The silhouette coefficient calculation step starts with finding $\boldsymbol{a_i^j}$, namely the average distance of the i-th data with all data in the same cluster. It is assumed that the i-th data is in cluster A. Kaufman and Rousseeuw [26] write $\boldsymbol{a_i^j}$ in Equation (22) as follows:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq 1}} d(x_i^j, x_r^j) \tag{22}$$

Where, j is cluster, i is index data $(i = 1,2, \dots, m_j)$, $\boldsymbol{a_i^j}$ **is** the average distance of the i-th data with all data in the same cluster, M_j is the amount of data in the j-th cluster, and $\boldsymbol{d(x_i^j, x_r^j)}$ is the distance between the i-th data and the j-th data in one cluster j.

Next, calculate the value of $\boldsymbol{b_i^j}$, the minimum value of the i-th data mean with all data in different clusters. Then, $\boldsymbol{b_i^j}$ is written in Equation (23) as follows:

$$b_i^j = \min_{\substack{n=1,\dots,k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq 1}}^{m_n} d(x_i^j, x_r^n) \right\} \tag{23}$$

Where, j is cluster, i is index data $(\boldsymbol{i = 1, 2, \dots, m_j})$, $\boldsymbol{b_i^j}$ **is** the average distance of the i-th data with all data in a different cluster, M_n is the amount of data in the n-th cluster, and $\boldsymbol{d(x_i^j, x_r^n)}$ is the distance between the i-th data and the j-th data in one cluster n. After $\boldsymbol{a_i^j}$ and $\boldsymbol{b_i^j}$ are known, the next step is to calculate $\boldsymbol{SI_i^j}$, which is written in Equation (24) as follows:

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \tag{24}$$

Where, $\boldsymbol{SI_i^j}$ **is** the i-th silhouette index data in one cluster, $\boldsymbol{a_i^j}$ is the average distance of the i-th data with all data in the same cluster, and $\boldsymbol{b_i^j}$ **is** the average distance of the i-th data with all data in a different cluster. The silhouette coefficient value of each object in a cluster is a measure that shows how closely the data are grouped in one cluster. The value of $\boldsymbol{SI_i^j}$ is in the range:

$$-1 < SI_i^j \leq 1 \tag{25}$$

The value of $\boldsymbol{SI_i^j}$ is close to -1, indicating that the distance between objects in $\boldsymbol{a_i^j}$ is much greater than $\boldsymbol{b_i^j}$, so it is said that there was a misclassification or doubt in the grouping that was done. Meanwhile, if the value of $\boldsymbol{SI_i^j}$ is close to 1, the distance between objects in $\boldsymbol{a_i^j}$ is much smaller than $\boldsymbol{b_i^j}$, so the grouping is said to be done well. Then, use Equation (26) for the calculation to get $\boldsymbol{SI_j}$ as follows:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \tag{26}$$

Where, $\boldsymbol{SI_j}$ is the average Silhouette Index cluster j, $\boldsymbol{SI_i^j}$ is the i-th silhouette index data in one cluster, M_j is the amount of data in the j-th cluster, and i is index data $(\boldsymbol{i = 1, 2, \dots, m_j})$. The calculation formula for obtaining the global SI value is written in Equation (27) below:

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j \tag{27}$$

Where, SI is the average Silhouette Index of the dataset, $\boldsymbol{SI_j}$ **is** the average Silhouette Index cluster j, and k is number of clusters. The final step is to determine the silhouette coefficient (SC), obtained by finding the maximum value of the Global Silhouette Index from the number of clusters 2 to the number of $q - 1$ clusters. SC is written in Equation (28) as follows:

$$SC = {}^{max}_{\ \ k} SI(k) \tag{28}$$

Where, SC is silhouette coefficient, SI is the average Silhouette Index of the dataset, and k is the k-th cluster $(\boldsymbol{k = 2, 3, \dots, q - 1})$ where q is the number of clusters. The silhouette coefficient criteria set by Kaufman and Rousseeuw [26] are presented in Table 1 as follows:

Table 1. Silhouette Coefficient Criteria

Silhouette Coefficient Value	Cluster Criteria
0.71 – 1.00	Strong
0.51 – 0.70	Good
0.26 – 0.50	Weak
0.00 – 0.25	Bad

So, we give simple data for the illustration:

Table 2. Data of Illustration

Data	X_1	X_2	Cluster
A	1	5	1
B	5	8	1
C	3	4	2
D	1	2	2

- The average distance (cohesion) between A and B $= \sqrt{((1-5)^2+(5-8)^2)} = \sqrt{41}$.
- The average distance between data point A and all data points in the other cluster:
  Distance between A and C $=\sqrt{((1-3)^2+(5-4)^2)} = \sqrt{5}$.
  Distance between A and D $=\sqrt{((1-1)^2+(5-2)^2)} = \sqrt{9}$.
  Minimum average distance $=\frac{(\sqrt{5}+\sqrt{9})}{2} \approx 2.24$.
- Silhouette coefficient or $SI = \frac{(2.24-\sqrt{41})}{\max(\sqrt{41},2.24)} = -0.32$.

There are illustrations of calculating silhouettes manually.

2.9 Data Source

The data used in this study is secondary rainfall data (mm) for all provinces in Indonesia from January 2018 to December 2022. The data was taken from the official website of the Meteorology, Climatology, and Geophysics Agency. The observational data to be used is data from 34 provinces from 34 observation stations. The initial stage of this research is to collect rainfall data for 34 provinces for five years, starting from 2018 to 2022. The next stage is to calculate distances using four distance measurement methods, namely dynamic time warping (DTW) distance, short time series (STS) distance, and autocorrelation function (ACF) distance. Then, determine the best number of clusters for each cluster analysis method using the elbow method. The next step is to carry out cluster analysis using an agglomerative hierarchical method consisting of single linkage, average linkage, and complete linkage. Afterwards, proceed with cluster analysis with non-hierarchical methods, namely k-medoids and fuzzy C-means. After obtaining the results of cluster analysis, cluster validation was carried out to find out the distance measurement method and the best cluster analysis method with the most optimal cluster results. Next will be the interpretation of the most optimal cluster results based on rainfall in Indonesia.

3. RESULTS AND DISCUSSION

The first stage of this research is collecting data. Then, calculations were carried out using the distance measurement method. Each distance measurement method measures the respective distance between two-time series, which means that each Province will calculate its distance from one other Province from the first Province to the 34th Province. After calculating the distance, then is done is to determine the number of clusters with the elbow method. The results of determining the best number of clusters are presented in Figure 1 below.

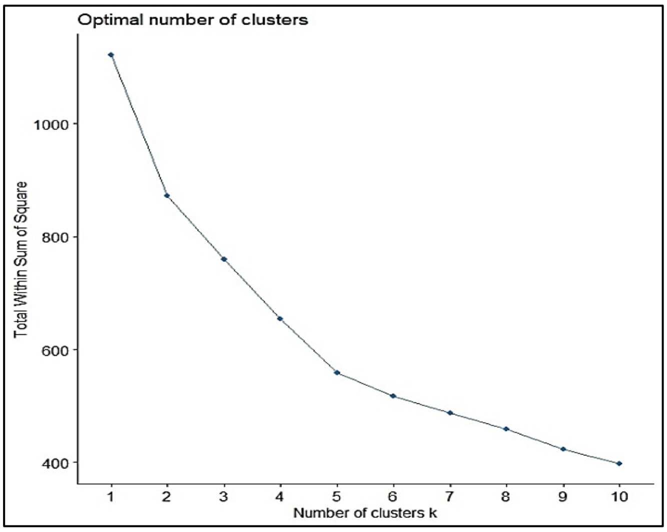

Figure 1. Result of The Elbow Method

The results of the elbow method show that the best grouping in this study is grouped into five clusters using hierarchical methods and non-hierarchical methods. After measuring the distance and knowing the number of clusters formed, the next step is conducting cluster analysis using hierarchical and non-hierarchical methods. The first are three distance measurement methods and three agglomerative hierarchical cluster analysis methods, namely single linkage, average linkage, and complete linkage. Then for non-hierarchical methods, namely k-medoids and fuzzy C-means. Later, each of these results will look for the optimal cluster results by comparing the silhouette coefficient values. Comparison of silhouette coefficient values for dynamic time warping (DTW) distance, short time series (STS) distance, and autocorrelation function (ACF) distance with each agglomerative hierarchical cluster analysis method is presented in Table 3.

Table 3. Comparison of Silhouette Coefficients using Hierarchical Method

Distance Measure Method	Agglomerative Hierarchical Method	Silhouette Coefficient
Dynamic Time Warping (DTW) distance	Single Linkage	0.788
	Complete Linkage	0.690
	Average Linkage	0.813
Autocorrelation Function (ACF) distance	Single Linkage	0.726
	Complete Linkage	0.723
	Average Linkage	0.748
Short Time Series (STS) distance	Single Linkage	0.770
	Complete Linkage	0.707
	Average Linkage	0.794

Table 3 shows that the most optimal cluster analysis results are the dynamic time warping (DTW) distance with the average linkage method because it has the best silhouette coefficient value of 0.813, which is included in the strong category. The distance autocorrelation function (ACF) method with the average linkage method has the best silhouette coefficient value of 0.748, which is included in the strong category. Furthermore, the short time series (STS) distance method with the average linkage method has the best silhouette coefficient value, 0.794, which is included in the strong category.

The results of the comparison of silhouette coefficient values in the non-hierarchical method are presented in Table 4.

Table 4. Comparison of Silhouette Coefficients using the Non-Hierarchical Method

Non-Hierarchical Method	Silhouette Coefficient
K-Medoids	0.132
Fuzzy C-means	0.120

Table 4 shows that the most optimal cluster results are in the non-hierarchical method, namely the K-Medoids, because it has a silhouette coefficient value of 0.132, which is in the wrong category. After knowing the most optimal cluster results from each distance measurement method using the hierarchical method and the most optimal cluster results using the non-hierarchical method, validate the most optimal cluster results from all cluster analysis methods using silhouette coefficients. A comparison of silhouette coefficients is presented in Table 5.

Table 5. Comparison of Silhouette Coefficients on Hierarchical and Non-Hierarchical Method

Cluster Analysis Method	Method	Silhouette Coefficient
Hierarchical Method	Dynamic Time Warping (DTW) distance with the average linkage method	0.813
	Autocorrelation Function (ACF) distance with the average linkage method	0.748
	Short Time Series (STS) distance method with the average linkage method	0.794
Non-Hierarchical Method	K-Medoids	0.132

The results show the comparison of the silhouette coefficient values in Table 5, it is known that the most optimal cluster results are the dynamic time warping (DTW) distance measurement method with the average linkage method with a silhouette coefficient value of 0.813 which is included in the strong category. The results of cluster analysis of 34 provinces in Indonesia on rainfall using four distance measurement methods and four cluster analysis methods obtained 5 clusters with the most optimal cluster results using the dynamic time warping (DTW) distance measurement method with the average linkage method. The map of cluster analysis results is presented in Figure 2.

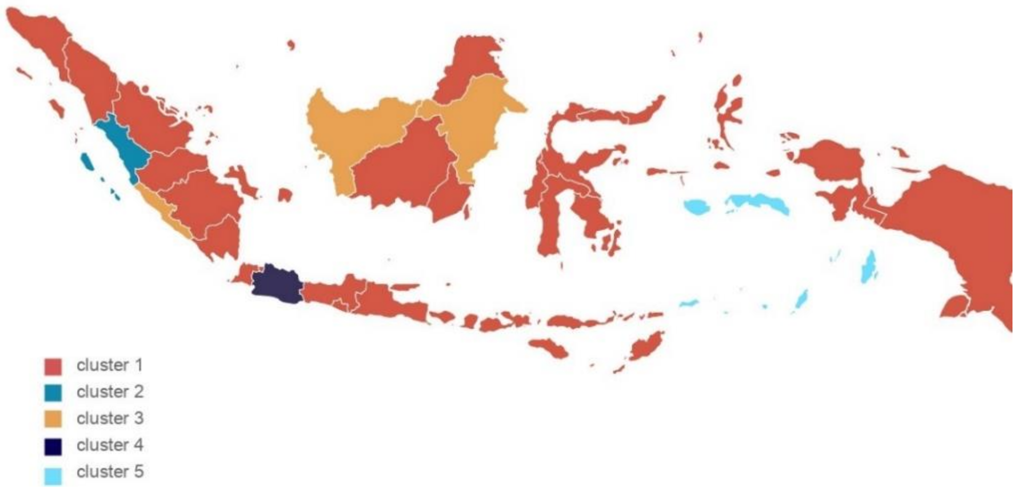

Figure 2. Map of Cluster Analysis Results

Based on Figure 2, the results of cluster 1, marked in red, consist of the provinces of Aceh, North Sumatra, Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, Central Kalimantan, South Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, North Maluku, West Papua, and Papua. Cluster 2 is marked in dark blue, namely the Province of West Sumatra. Cluster 3, marked in orange, consists of the provinces of Bengkulu, West Kalimantan, and East Kalimantan. Cluster 4 is marked in purple, namely, the Province of West Java, and Cluster 5 is marked in light blue, the Province of Maluku.

4. CONCLUSIONS

Based on the results of rainfall cluster analysis research in Indonesia from 2018 to 2022, the grouping of each distance measurement method and hierarchical and non-hierarchical cluster analysis methods was formed into 5 clusters, with the category for cluster 1 being provinces that are very vulnerable to natural disasters due to rainfall. Cluster 2 is a province that has a reasonably vulnerable condition, Cluster 3 is a province that has a relatively vulnerable condition, Cluster 4 is a province that has a reasonably safe condition, and Cluster 5 is a province that has a safe condition. This study's findings can assist Indonesian parties prevent and mitigate monsoon catastrophes. Future research can improve the clustering process by including machine learning and additional criteria to achieve more accurate findings. Based on the results of cluster analysis using hierarchical methods and non-hierarchical methods, the most optimal cluster results were obtained using the dynamic time warping (DTW) distance measurement method with the average linkage method with a silhouette coefficient value of 0.813, which is included in the strong category.

REFERENCES

[1] B. Suhardi, A. Adiputra, and R. Avrian, “Kajian Dampak Cuaca Ekstrem Saat Siklon Tropis Cempaka dan Dahlia di Wilayah Jawa Barat,” *Jurnal Geografi, Edukasi dan Lingkungan (JGEL)*, vol. 4, no. 2, pp. 61–67, 2020.

[2] Tukidi, “Karakter Curah Hujan Di Indonesia,” *Jurnal Geografi*, vol. 7, no. 2, pp. 136–145, 2010.

[3] M. Kotz, A. Levermann, and L. Wenz, “The effect of rainfall changes on economic production,” *Nature*, vol. 601, no. 7892, pp. 223–227, 2022.

[4] I. Ayundari and Sutikno, “Penentuan Zona Musim di Mojokerto Menurut Karakteristik Curah Hujan Dengan Metode Time Series Based Clustering,” *INFERENSI*, vol. 2, no. 2, pp. 63–70, 2019.

[5] T. W. Liao, “Clustering of time series data—a survey,” *Pattern Recognit*, vol. 38, no. 11, pp. 1857–1874, 2005.

[6] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis : Forecasting and Control Fourth Edition*. New Jersey : John Wiley and Sons Inc., 2008. doi: 10.1142/9789813148963_0009.

[7] A. Sardá-Espinosa, “Time-series clustering in R Using the dtwclust package,” *R Journal*, vol. 11, no. 1, pp. 1–45, 2019, doi: 10.32614/rj-2019-023.

[8] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan, “Adaptively constrained dynamic time warping for time series classification and clustering,” *Inf Sci (N Y)*, vol. 534, pp. 97–116, 2020.

[9] G. Marti, F. Nielsen, M. Bińkowski, and P. Donnat, “A review of two decades of correlations, hierarchies, networks and clustering in financial markets,” *Progress in information geometry: Theory and applications*, pp. 245–274, 2021.

[10] M. A. A. Riyadi, D. S. Pratiwi, A. R. Irawan, and K. Fithriasari, “Clustering stationary and non-stationary time series based on autocorrelation distance of hierarchical and k-means algorithms,” *International Journal of Advances in Intelligent Informatics*, vol. 3, no. 3, pp. 154–160, 2017.

[11] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Pearson Education, Inc., 2007.

[12] R. M. Shukla and S. Sengupta, “Scalable and robust outlier detector using hierarchical clustering and long short-term memory (lstm) neural network for the internet of things,” *Internet of Things*, vol. 9, p. 100167, 2020.

[13] S. Maulidani, N. Ihsan, and Sulistiawaty, “Analisis pola dan intensitas curah hujan berdasakan data observasi dan satelit tropical rainfall measuring missions (trmm) 3b42 v7 di makassar,” *Jurnal sains dan pendidikan fisika*, vol. 11, no. 1, pp. 98–103, 2015.

[14] Supriyati, B. Tjahjono, and S. Effendy, “ANALISIS POLA HUJAN UNTUK MITIGASI ALIRAN LAHAR HUJAN GUNUNG API SINABUNG,” *Jurnal Ilmu Tanah dan Lingkungan*, vol. 20, no. 2, pp. 95–100, 2018.

[15] Enyew and Steeneveld, “Analysing the Impact of Topography on Precipitation and Flooding on the Ethiopian Highlands,” *Journal of Geology & Geosciences*, vol. 3, no. 6, 2014, doi: 10.4172/2329-6755.1000173.

[16] V. Niennattrakul and C. A. Ratanamahatana, “On clustering multimedia time series data using k-means and dynamic time warping,” *Proceedings - 2007 International Conference on Multimedia and Ubiquitous Engineering, MUE 2007*, pp. 733–738, 2007, doi: 10.1109/MUE.2007.165.

[17] C. S. Möller-Levet, F. Klawonn, K. H. Cho, and O. Wolkenhauer, “Fuzzy clustering of short time-series and unevenly distributed sampling points,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2810, no. 0, pp. 330–340, 2003, doi: 10.1007/978-3-540-45231-7_31.

[18] P. Galeano and D. Pella, “Multivariate Analysis in Vector Time Series,” *Resenhas, the Journal of the Institute of Mathematics and Statistics of the University of Sao Paolo*, vol. 4, pp. 383–403, 2000.

[19] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis: Fifth Edition*. United Kingdom: John Wiley and Son, Ltd., 2011.

[20] Z. Mustofa and Iman Saufik Suasana, “Algoritma Clustering K-Medoids Pada E-Government Bidang Information and Communication Technology Dalam Penentuan Status Edgi,” *Jurnal Teknologi Informasi Dan Komunikasi*, vol. 9, no. 1, pp. 1–10, 2020, doi: 10.51903/jtikp.v9i1.162.

[21] W. Sanusi, A. Zaky, and B. N. Afni, “Analisis Fuzzy C-means dan Penerapannya Dalam Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Faktor-faktor Penyebab Gizi Buruk,” *Journal of Mathematics, Computations, and Statistics*, vol. 2, no. 1, p. 47, 2019, doi: 10.35580/jmathcos.v2i1.12458.

[22] P. Bholowalia and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” *Int J Comput Appl*, vol. 105, no. 9, pp. 975–8887, 2014.

[23] N. Putu, E. Merliana, and A. J. Santoso, “Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means,” pp. 978–979, 2015.

[24] Irwanto, Y. Purwananto, and R. Soelaiman, “Optimasi Kinerja Algoritma Klasterisasi K-Means,” *Jurnal Teknik ITS*, vol. 1, no. 1, pp. 197–202, 2012.

[25] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.*, vol. 47, no. 2. 1991. doi: 10.2307/2532178.