TIERS Information Technology Journal
Vol.4, No.2, December 2023, pp. 115~121
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i2.5081

Clustering Time Series Using Dynamic Time Warping Distance in Provinces in Indonesia Based on Rice Prices

Yeni Rahkmawati[1], Selvi Annisa[2]
yeni.rahkmawati@ulm.ac.id [1], selvi.annisa@ulm.ac.id [2]
[1,2]Statistics Study Program, Universitas Lambung Mangkurat, South Kalimantan, Indonesia

ABSTRACT

Rice is a food commodity that is a basic need for Indonesian people. Since the end of 2022, average rice prices in Indonesia have been increasing, breaking the record for the highest price from August to October 2023. The price of rice in each province in Indonesia is different. This can happen because rice center provinces will distribute their rice production to other regions to meet rice needs. The grouping of provinces in Indonesia based on rice prices over time is an interesting thing to research. The analysis method used to group similar objects into groups for time series data is called clustering time series. The distance that can be used to measure the closeness of two-time series is the Dynamic Time Warping (DTW) distance. The clustering analysis used is the single, complete, average, Ward, and median linkage method. The results of the analysis show that time series clustering in provinces in Indonesia based on rice prices is best using median linkage hierarchical clustering. The median linkage method has a cophenetic correlation coefficient value of 0.899064, meaning that clustering using the DTW distance with the median difference is very good. The resulting clusters contained 3 clusters which had different characteristics between the clusters. There are 2 clusters that can be of concern to the government, because there are clusters that have rice prices that have always been high in the last period and there are provincial clusters that have rice prices that are very diverse or can be said to be unstable.

Keywords:Clustering Time Series; Dynamic Time Warping; Hierarchical Clustering; Rice Prices, Indonesia

Article Info

Accepted : 08-12-2023
Revised : 12-09-2023
Published Online : 25-12-2023

Correspondence Author:

Yeni Rahkmawati
Statistics Study Program,
Universitas Lambung Mangkurat,
A.Yani St., Km. 36, Banjarbaru, 70714
Email: yeni.rahkmawati@ulm.ac.id

1. INTRODUCTION

Rice is a food commodity that is a basic need for the people of Indonesia. A survey by the Central Statistics Agency shows that around 98.3% of Indonesians consume rice. Rice is an inelastic commodity, meaning that price changes do not cause changes in consumer demand [1]. If availability decreases, prices will soar, which can be unaffordable for consumers [2]. Since the end of 2022, average rice prices in Indonesia have been increasing, breaking the record for the highest price from August to October 2023, premium rice price reaches Rp 20,000 per kg. The trend in premium rice prices in Indonesia from January 2023 from October 2023 can be seen in the time series plot in Figure.

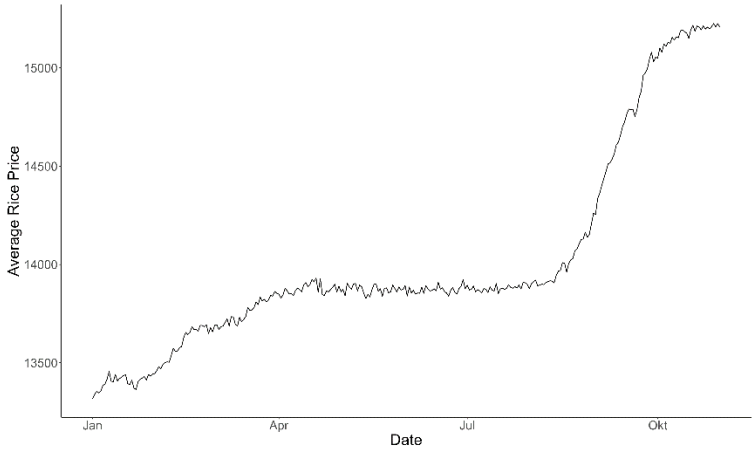

Figure 1. Average Rice Price in Indonesia from January to October, 2023

High price of rice has such a negative impact on the majority of the Indonesian population, with as much as 80% of the people feeling the effects. It's essential to find ways to address this issue and ensure that everyone has access to affordable food [3]. In addition, the price of rice in each province in Indonesia is different. This can be because the province, which is the center of rice, will distribute its rice production to other regions to meet rice needs. Provinces that are rice-producing centers such as East Java, Central Java, West Java, South Sumatra, etc. So, the price of rice in the rice center province will affect the price of rice in the area that is the destination of distribution because there are additional transportation and labor costs to distribute it [4].

Clustering of provinces in Indonesia is an interesting thing to research. Several studies on clustering in provinces in Indonesia, such as: Ahmar, et.al [5] Cluster Province in Indonesia using K-Means. Apart from that, research on the clustering of provinces in Indonesia based on rice prices has also been carried out, such as research by S. U. Wijaya and Ngatini [4] those who carried out the development of rice price modeling in the western part of Indonesia with a time series clustering approach using Dynamic Time Wrapping (DTW) distance. In addition, Ulinnuha, et.al [6] it also conducts provincial clustering based on rice prices using correlation distance. A method of data analysis to identify groups of objects, or clusters, that are more similar to each other than to other clusters is clustering [7]. If the data used is time series data, then the clustering time series can be used. Clustering of time series data is the unsupervised classification of a set of unlabeled time series into groups or clusters, where all the sequences are grouped in the same cluster. The clusters should be coherent or homogeneous [8].

A time series data clustering is a cluster that pays attention to the dynamic nature of time series data. The use of distance in clustering time series data is divided into three categories: raw data, feature data, and model parameters. Distance Raw Data is the distance obtained based on the original data. Distance Featured represents the distance on the representation of the characteristics of the data [9]. Distance-based example Featured The time series is Auto-Correlation Function (ACF) distance used by Sakoe and Chiba [10]. The model parameter distance is the distance of the coefficient of the time series model. The Euclidean distance is the most common choice. This measure only applies to small-scale and equal-length time series, which limits the scope of its application. Furthermore, time shifts are inevitable in time series data, which is an intractable issue for Euclidean distance. Thus, it can be seen that Euclidean distance is not the optimal choice for clustering time series. Therefore, in this work, we select Dynamic Time Warping (DTW) distance as the similarity measure. In time series analysis, DTW is the most well-known algorithm, and it is used exclusively for measuring similarity between two temporal sequences that may vary in speed. Taking into account time shifts, this algorithm calculates an optimal match between two-time series and thus can compute the similarity more accurately [11].

Dynamic Time Wrapping (DTW) is one method for calculating the distance between two-time series data. Dynamic Time Wrapping (DTW) is the calculated distance of the optimal warping path between two-time series [12]. DTW distances are more realistically used in measuring the similarity of a pattern than using only linear measurement algorithms such as Euclidean, Manhattan, and other measurement algorithms [13]. One clustering analysis that can be used is hierarchical clustering analysis. This method is used to group observations in a structured manner based on their similar nature, and the number of desired clusters is not yet known. There are two methods of hierarchical clustering: aglomerative and divisive. The hierarchical method of merging is obtained by combining observations or groups gradually so that, in the end, only one group is obtained. On the contrary, the method of separation in the hierarchical method begins by forming one large group consisting of all observations. The large group is then separated into smaller groups until one group

.

only has one observation. Cluster objects in a hierarchical algorithm using the linkage method (Linkage). Some of the linkage methods used are single, complete, average, Ward, and median linkage methods [14]. So, based on the description above, this study will group provinces in Indonesia based on rice prices using DTW distance and several hierarchical clustering. This research also compares several clustering methods, with the measure of goodness used being the cophenetic correlation coefficient.

2. RESEARCH METHOD

The data in this study used secondary data obtained from the National Food Agency (website: https://badanpangan.go.id/). The variable used is the price of premium rice measured in 34 provinces in Indonesia. The price of premium rice is the price prevailing at retail traders. The data is daily data from January 1, 2023 to October 31, 2023.

The procedures used in this study are as follows:

1. Data Exploration
   Data exploration is carried out on premium rice price data with a line box chart to see the distribution of data and data diversity.
2. Calculates Dynamic Time Warping (DTW) distance on premium rice price data.
   Dynamic Time Wrapping (DTW) is a crucial method in our study. It calculates the distance between two-time series data, providing a measure of dissimilarity that is not dependent on a specific model approach. DTW distance is a suitable measure to evaluate the similarities/dissimilarities of time series concerning their shape information [15]. This dynamic distance is determined by comparing two-time series data and attempting to find the optimal compressible curve between them, a technique known as time series data clustering [16].

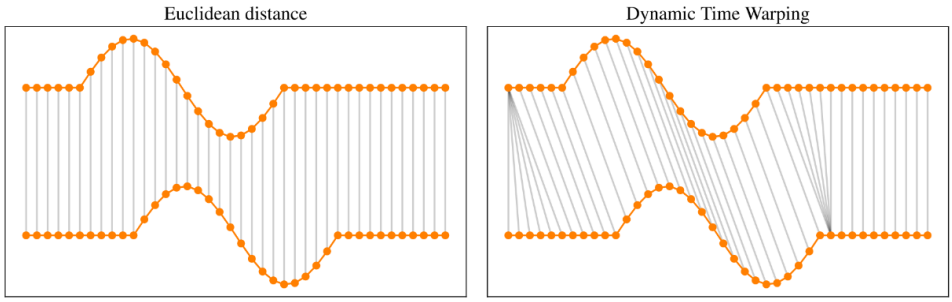

Figure 2. Illustration between Euclidean Distance and DTW Distance

DTW distance can be calculated using the formula below:

$$\mathrm{DTW}(S,T) = \min_W \left[\sum_{k=1}^{p} \delta(w_k) \right] \quad (1)$$

With $S = s_1, s_2, \ldots, s_n$ dan $T = t_1, t_2, \ldots, t_m$,, is a time series contained in a matrix of size n x m. $W = w_1, w_2, \ldots, w_k$ is the possibility of an arch path that maps or realigns the members of S and T so that the distance between them is minimum. The distance δ can be $\delta(i,j) = |s_i - t_j|$ and w_k refers to a point $(i,j)_k$ on the path of the k-th arch.

3. Perform hierarchical clustering using the Single, Complete, Average, Ward, and Median linkage.

4. Calculates the cophenetic correlation coefficient.
   Cophenetic correlation coefficient is the correlation coefficient between the original element of the inequality matrix (Euclidean distance matrix) and the element generated by the dendrogram (Cophenetic matrix based on distance measures and the connectedness method used) [17]. The formula for calculating the Cophenetic correlation coefficient is as follows:

$$r_{coph} = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{Cik} - \bar{d}_c)}{\sqrt{\left(\sum_{i<k}(d_{ik} - \bar{d})^2\right)\left(\sum_{i<k}(d_{Cik} - \bar{d}_c)^2\right)}} \quad (2)$$

With r_{coph}: cophenetic correlation coefficient; d_{ik} i-th and k-th Euclidean distances; $\bar{d}$: average distance d_{ik}; d_{Cik}: i-th and k-th cophenetic distances; $\bar{d}_c$: average distance d_{Cik}. The value of the cophenetic correlation coefficient ranges between -1 and 1. The closer to the value of 1, the better the resulting cluster.

5. Select the hierarchical clustering method using the highest cophenetic correlation coefficient [18].
6. Make a dendogram for the best hierarchical cluster method.
   Dendrogram clustering is also known as hierarchical clustering algorithms. This clustering falls into two categories: top-down or bottom-up. The bottom-up algorithm categorizes each data or object as a single cluster and then merges with its pair until all clusters are merged and become a single cluster. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the objects, the leaves being the clusters with only one sample. Another category uses a top-down algorithm known as divisive [19].
7. Calculate the optimum number of cluster using the silhouette coefficient.
   Silhouette coefficient is a comparison between the size of the proximity of objects in one cluster and the size of the proximity between the clusters formed. The formula in the measurement of cluster accuracy, namely:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

   where a(i) is the average distance between object i and all objects in the same group (Intracluster), while b(i) is the average distance between object i and all objects in the nearest cluster (nearest cluster) [20]. Coefficient silhouette has a range of values for each object. Subjective interpretation of coefficient quantities $-1 \leq s(i) \leq 1$ silhouette as in Table 1 [21].

Table 1. Subjective Interpretation of Coefficients Silhouette

Silhouette coefficient	Interpretation
0.71 – 1.00	There is a strong cluster structure
0.51 – 0.70	Reasonable cluster structure
0.26 – 0.50	Weak cluster structure, very likely pseudo
0.00 – 0.25	There is no significant cluster structure

8. Identify the cluster membership.

3. RESULTS AND DISCUSSION

The data exploration aims to determine the distribution of premium rice price data for each province in Indonesia, which is presented in the line box chart in Figure 3. The distribution of rice price data can be seen from the location of the line box chart. Provinces in Indonesia have rice prices between Rp 11,210 - Rp 20,000 from 1 january 2023 to 31 october 2023. Some provinces have boxes located higher than others, which shows that the premium rice price range in those provinces is higher than that of others. Provinces that have a high rice price range, for example, include provinces on the islands of Kalimantan, Maluku, and Papua. Apart from that, most provinces have outlier data seen from the plot of dots outside the boxplot, meaning that the price of premium rice from 1 January 2023 to 31 October 2023 in that province has relatively high fluctuating prices. In addition to the data distribution, the diversity in the data can be seen from the width of the box in each province. The width of the boxes between quartiles in most provinces is almost the same, indicating that the diversity of premium rice price data in most provinces is homogeneous. However, province has a larger box width, such as Papua. This province tends to have a higher diversity of premium rice prices than others.

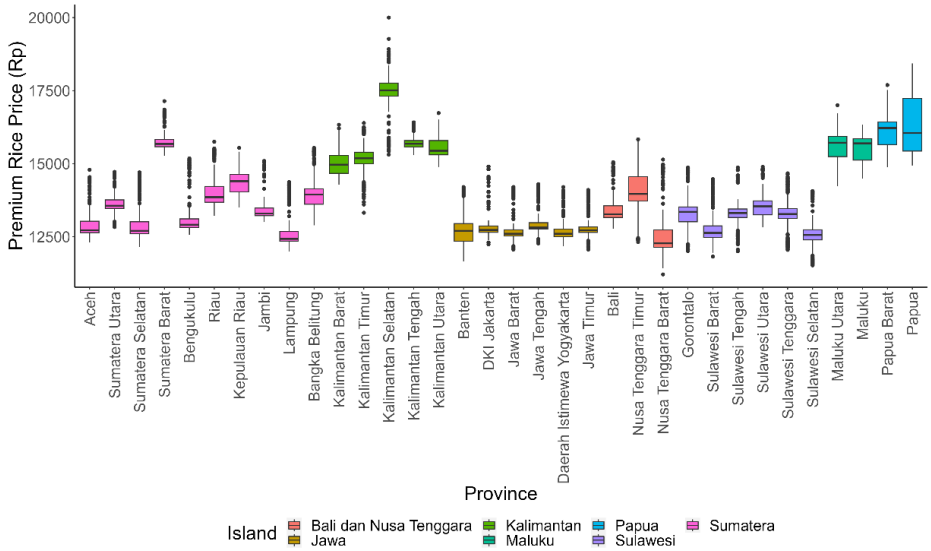

Figure 3. Boxplot of Premium Rice Prices in 34 Provinces in Indonesia

.

After exploring the data, then a hierarchical cluster was carried out. The hierarchical cluster in this study began by calculating the distance between each province and other provinces using the DTW dissimilarity measure. In simple terms, distance measurement can be started by measuring the distance between Nanggroe Aceh Darussalam and Utara Sumatra, Nanggroe Aceh Darussalam and Sumatra Barat, measuring the distance between Nanggroe Aceh Darussalam and Papua. The results of DTW distance measurement are presented in the Table below.

Table 2. DTW Distance Matrix

Provinces	Aceh	Sumatera Utara	⋯	Papua Barat	Papua
NAD	0	12558.27	⋯	54953.57	57804.12
Sumatera Utara	12558.27	0	⋯	43233.49	46785.62
⋮	⋮	⋮	⋱	⋮	⋮
Papua Barat	54953.57	43233.49	⋯	0	12202.93
Papua	57804.12	46785.62	⋯	12202.93	0

Then, DTW distances are used to group provinces using the single, complete, average, Ward, and median linkage methods. Comparison of the five linking methods using the measure of goodness cophenetic correlation coefficient. The cophenetic correlation coefficient measures the usefulness of using a distance or dissimilarity in the time series data cluster. This measure is obtained from the correlation between the cophenetic distance from the tree diagram and the distance of the original object used to create the tree diagram. The value of the cophenetic correlation has a range, the value of $-1 < r < 1$ the cophenetic correlation close to 1 means that the resulting cluster is perfect. A comparison of several links presented in Table 3 is obtained.

Table 3. Cophenetic Correlation Coefficient in the Hierarchical Cluster Method

Method	Cophenetic Correlation Coefficient
Single	0.846707
Complete	0.859601
Average	0.86429
Ward	0.835614
Median	0.899064*

Based on Table 3, the highest cophenetic correlation coefficient is obtained from the cluster method using median linkage of 0.899064, meaning that referring to Table 1, the presence of a strong cluster structure or cluster using the distance of DTW with the median linkage is outstanding. The clustering results can be illustrated by the tree diagram (dendogram) in Figure 4.

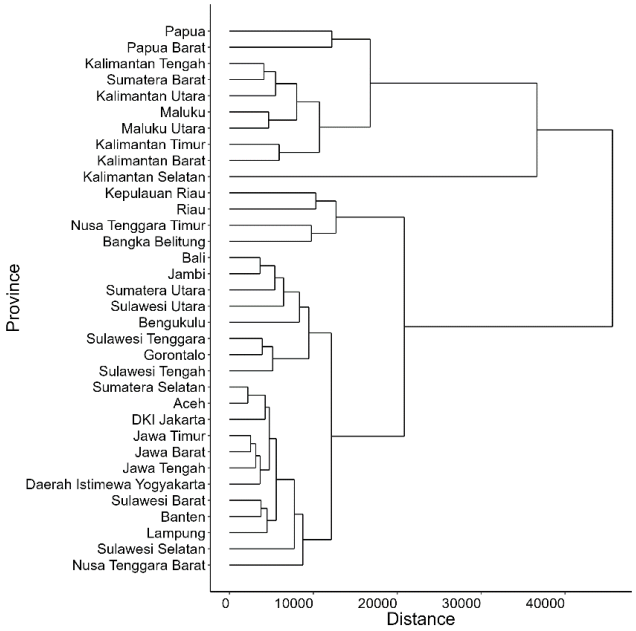

Figure 4. Hierarchical Cluster Analysis Dendogram Using DTW Distance

The cluster results obtained in Figure 4 show that the number of clusters that can be formed is 2 to 34. So, the next step in obtaining the best cluster requires determining the optimal number of clusters. The optimal cluster lot can be determined based on the maximum value of the silhouette coefficient on the number of clusters 2 to 10 presented in the plot in Figure 5.

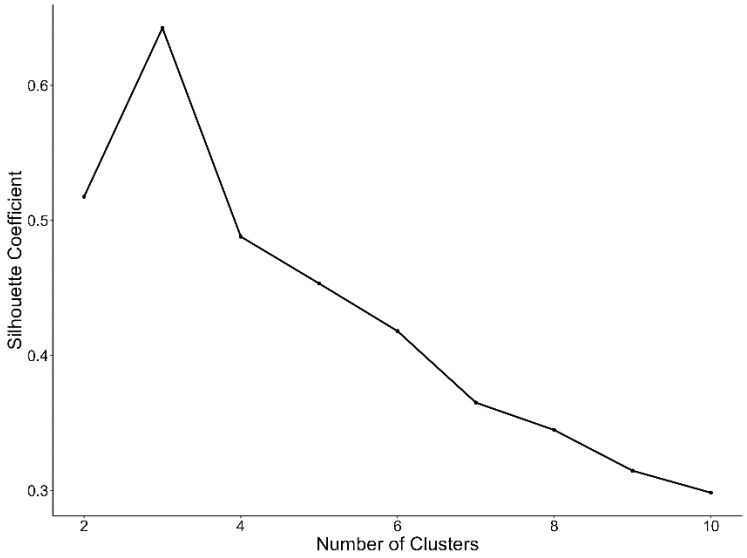

Figure 5. Silhouette Coefficient

The optimal number of clusters obtained can be determined through the subjective interpretation of the silhouette coefficient for each possible number of clusters. Based on the silhouette coefficient in Figure 5, it shows its maximum value is in many clusters k = 3. So, the clusters that can be formed are 3 clusters. The results of clustering with the number of clusters k=3 are presented in Table 4.

Table 4. Cluster of Provinces in Indonesia based on Premium Rice Prices

Clusters	Number of Clusters	Member of Clusters
Cluster 1	24	Aceh, Sumatera Utara, Sumatera Selatan, Bengukulu, Riau, Kepulauan Riau, Jambi, Lampung, Bangka Belitung, Banten, DKI Jakarta, Jawa Barat, Jawa Tengah, Daerah Istimewa Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Timur, Nusa Tenggara Barat, Gorontalo, Sulawesi Barat, Sulawesi Tengah, Sulawesi Utara, Sulawesi Tenggara, Sulawesi Selatan
Cluster 2	9	Sumatera Barat, Kalimantan Barat, Kalimantan Timur, Kalimantan Tengah, Kalimantan Utara, Maluku Utara, Maluku, Papua Barat, Papua
Cluster 3	1	Kalimantan Selatan

Table 4 presents the results of hordes divided into 3 clusters. Cluster 1 have 3 provinces which have almost the same characteristics; namely, the price of rice in these 24 provinces tends to be stable and not too high compared to other clusters. Cluster 2 consists of 9 provinces which tend to have quite high rice prices compared to cluster 1. Furthermore, cluster 3, there is only one province, namely Kalimantan Selatan. This province have a high distribution of rice prices compared to other provinces and high variability. So, from the hordes above, it can be seen that provincial groups must be the government's attention so that rice prices can be controlled and evenly distributed in each province.

4. CONCLUSION

Time series clustering in Indonesian provinces based on rice prices is best used by median linkage hierarchical clustering. The average linkage method has a cophenetic correlation coefficient of 0.899064 meaning clustering using the DTW distance with the median linkage is very good. The resulting cluster has 3 groups with different characteristics. There are clusters that have the characteristics of provinces with high rice prices, and there are also clusters of provinces with unstable prices which can be a concern for the government in making policies, especially in controlling rice prices.

.

REFERENCES

[1] M. Asaad, "Economic Policies on RiceCommodity and Welfare," *Economic Journal of Emerging Markets,* vol. 2, no. 1, pp. 13-29, 2010.

[2] E. Siswanto, B. M. Sinaga and Harianto, "Dampak Kebijakan Perberasan pada Pasar Beras dan Kesejahteraan," *Jurnal Ilmu Pertanian Indonesia (JIPI),* vol. 23, no. 2, pp. 93-100, 2018.

[3] N. McCulloch, "RICE PRICES AND POVERTY IN INDONESIA," *Bulletin of Indonesian Economic Studies,* vol. 44, no. 1, pp. 45-64, 2008.

[4] S. U. Wijaya and Ngatini, "Pengembangan Pemodelan Harga Beras di Wilayah Indonesia," *Limits: Journal of Mathematics and Its Applications,* vol. 17, no. 1, pp. 51-66, 2020.

[5] A. S. Ahmar, R. Hidayat, D. Napitupulu, R. Rahim, Y. Sonatha and M. Azmi, "Using K-Means Clustering to Cluster Provinces in Indonesia," in *nd International Conference on Statistics, Mathematics, Teaching, and Research* , 2018.

[6] M. Ulinnuha, F. M. Afendi and I. M. Sumertajaya, "Study of Clustering Time Series Forecasting Model for," *Indonesian Journal of Statistics and Its Applications* , vol. 6, no. 1, pp. 50-62, 2022.

[7] M. Z. Rodriguez, C. H. Comin, D. Casanova, D. R. Amancio, M. B. Odemir, L. d. F. Costa and F. A. Rodrigues, "Clustering algorithms: A comparative," *PLoS ONE* , vol. 14, no. 1, 2019.

[8] S. Rani and G. Sikka, "Recent Techniques of Clustering of Time Series Data: A," *International Journal of Computer Applications (,* vol. 52, no. 15, 2012.

[9] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition,* vol. 38, p. 1857 – 1874, 2005.

[10] P. D'Urso and E. A. Maharaj, "Autocorrelation-based fuzzy clustering of time series," *Fuzzy Sets and Systems,* vol. 160, no. 24, pp. 3565-3589, 2009.

[11] Y. Liu, J. Chen, S. Wu, Z. Liu and H. Chao, "Incremental fuzzy C medoids clustering of," *PLoS ONE* , vol. 13, no. 5, 2018.

[12] L. Liu, W. Li and H. Jia, "Method of Time Series Similarity Measurement Based on Dynamic Time Warping," *Tech Science Press* , vol. 57, no. 1, pp. 97-106, 2018.

[13] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for," *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING,* Vols. VOL. ASSP-26, no. 1, 1978.

[14] A. A. Mattjik and I. M. Sumertajaya, Sidik Peubah Ganda dengan Menggunakan SAS, Bogor: IPB Press, 2013.

[15] H. Izakian, W. Pedrycz and I. Jamal, "Fuzzy clustering of time series data using dynamic time warping distance," *Engineering Applications of Artificial Intelligence,* vol. 39, p. 235–244, 2015.

[16] D. J. Bemdt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," *Knowledge Discovery in Databases Workshop,* pp. 359-370, 1994.

[17] P. R. Carvalho, C. S. Munita and A. L. Lapolli, "Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient," *BRAZILIAN JOURNAL OF RADIATION SCIENCES* , vol. 7, no. 2, pp. 1-14, 2019.

[18] Iis, I. Yahya, G. N. A. Wibawa, Baharuddin, Ruslan and L. Laome, "PENGGUNAAN KORELASI COPHENETIC UNTUK PEMILIHAN METODE CLUSTER BERHIERARKI PADA MENGELOMPOKKAN KABUPATEN/KOTA BERDASARKAN JENIS PENYAKIT DI PROVINSI SULAWESI TENGGARA TAHUN 2020," in *PROSIDING SEMINAR NASIONAL SAINS DAN TERAPAN* , Manado, 2010.

[19] S. Azri, U. Ujang and A. A. Rahman, "DENDROGRAM CLUSTERING FOR 3D DATA ANALYTICS IN SMART CITY," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. XLII, no. 4, pp. 247-253, 2018.

[20] H. ŘEZANKOVÁ, "DIFFERENT APPROACHES TO THE SILHOUETTE COEFFICIENT CALCULATION IN CLUSTER EVALUATION," in *DIFFERENT APPROACHES TO THE SILHOUETTE COEFFICIENT,* Kutná Hora,, 2018.

[21] Y. Rahkmawati, I. M. Sumertajaya and M. N. Aidi, "Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach," *International Journal of Scientific and Research Publications,* vol. 9, no. 9, pp. 439-443, 2019.