Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 1 2023, 39 - 45
DOI: *https://doi.org/10.22487/27765660.2023.v3.i1.16408*
eISSN: 2776-5660

APPLICATION OF TIME SERIES CLUSTER ANALYSIS IN CLUSTERING THE CENTRAL JAVA PROVINCE BASED ON THE POVERTY DEPTH INDEX

Zulfanita Dien Rizqiana[1*]
[1]UIN Raden Mas Said Surakarta

e-mail: *zulfanita.dr@staff.uinsaid.ac.id*

ABSTRACT

Poverty is a problem that continues to be faced, especially in developing countries such as Indonesia. Poverty is included in one of the Sustainable Development Goals (SDGs) programs, which is related to hunger and health. The time series data can be clustered based on the characteristics of the time series data and adjusted to the time series pattern. The choice of distance and method used must be adjusted to the dynamic structure of time series data. The purpose of this research is to cluster districts/cities in Central Java Province based on the poverty depth index value from 2017 to 2022. The variable that used in this research is the Poverty Depth Index of 35 districts in Central Java Province from 2017 to 2022. This research used cluster time series with DTW similarity measurment. Based on theDTW and cophenetic coefficient correlation value using three linkage methods, the average linkage method has the highest cophenetic coefficient correlation value of 0.8017988. Testing the quality of clusters using the silhouette coefficient using DTW distance and average linkage method and 2 clusters are included in the good cluster category with a silhouette coefficient value of 0.60. The resulting clusters using the DTW distance and average linkage method are cluster 1 consisting of 25 districts / cities and cluster 2 consisting of 10 districts.

Keywords: *Poverty, DTW, Average linkage, Cluster Time Series*

Cite: Rizqiana, Z. D., (2023). *Application of Time Series Cluster Analysis in Clustering the Central Java Province Based on the Poverty Depth Index*. *Parameter: Journal of Statistics*, *3*(1), 39-45, https://doi.org/10.22487/27765660.2023.v3.i1.16408.

[*]Received: Jun 2023; Reviewed: Jun 2023; Published: Jun 2023

INTRODUCTION

Poverty is an issue that continues to be faced, especially by developing countries like Indonesia. Poverty is a complex problem with multiple underlying factors. It refers to the condition in which an individual is unable to meet their basic needs (Nafi'ah, 2021). Poverty is included as one of the Sustainable Development Goals (SDGs), specifically related to hunger and health. This indicates that poverty impacts every aspect of life and needs to be addressed seriously. Some government programs that have been implemented to reduce poverty include the Family Hope Program (PKH), the Smart Indonesia Card (KIP) provided for school-age children, and the Healthy Indonesia Card (KIS) (Lestari dan Busnetty, 2022).

Badan Pusat Statistik (2023) recorded that there is at least a 14.38% poverty rate in rural areas and an 11.98% poverty rate in urban areas. Meanwhile, in Java Island, Central Java Province ranks second with the highest percentage of poverty at 10.98% (Badan Perencanaan Pembangunan Nasional, 2023). When viewed based on another poverty indicator, namely the poverty depth index, Central Java Province also ranks second after Yogyakarta Special Region (DIY), with a rate of 1.77% in the first semester of 2022 and 1.75% in the second semester of 2022 (Badan Pusat Statistik, 2022). Different demographic characteristics will greatly influence the poverty rate in each region. These distinct demographic characteristics can be identified using a statistical analysis tool called cluster analysis. This analysis aims to assist government policies or intervention programs to be more focused and targeted.

Cluster analysis is a multivariate analysis tool used to group n objects into k clusters ($k \leq n$) based on their characteristics. Clustering is performed based on the similarity or dissimilarity between objects. Objects within the same cluster are more similar to each other compared to objects in different clusters. (Suhaeni *et al*., 2018). Cluster analysis can be applied to time series data. The time series data can be clustered based on the characteristics of the time series and adjusted to the temporal patterns. The choice of distance metric and clustering method should be tailored to the dynamic structure of the time series data. (Munthe, 2019). Several previous studies have utilized time series cluster analysis, such as the research conducted by Buaton *et al.* (2019), Dani *et al.* (2020), dan Soleha *et al.* (2022). The previous research did not discussed about clustering of Districtin Central Java based on the depth of proverty. Research by (Soleha et al., 2022) discuss about clustering Province in Indonesia based on non oil and gas export valeu, research by (Buaton et al., 2019) discussed about clutering time series with manhattan distance. Therefore, the objective of this research is to cluster the districts and cities in Central Java Province based on the n values of the poverty depth index from 2017 to 2022.

MATERIALS AND METHODS

The data used in this research is secondary data from the official website of the Badan Pusat Statistika of Central Java Province. The variable that used in this research is the Poverty Depth Index of 35 districts in Central Java Province from 2017 to 2022. This research used cluster time series method for answering research aims. Cluster time series analysis is used to cluster objects with dynamic data.

Cluster Analysis

Cluster analysis is an analytical tool used to group objects based on their similarities. Objects within a cluster exhibit a high level of similarity, while the similarity between clusters is low. (Yusfar *et al*., 2020). The advantages of cluster analysis are that it can handle a large and relatively large amount of data, and it can be applied to data with minimal ordinal measurement scale. However, the disadvantages of cluster analysis are its subjective nature as researchers base the analysis results on dendrograms, the difficulty in determining the number of clusters formed for heterogeneous data, significant variations between different methods used, and an increasing error rate with a larger number of observations. (Anggraini and Arum, 2021)

Cluster Time Series Analysis

Cluster time series analysis is used to cluster objects with dynamic data. In this time series cluster analysis, the algorithms used for static data are modified to be applicable to time series data. According to Yanti dan Rahardiantoro (2018) There are three categories of time series cluster analysis: (1) raw data-based approach, which involves calculating distances between clusters, (2) approach that eliminates outlier data and reduces data dimensions, followed by distance calculation and clustering, and (3) approach that utilizes pre-formed models for clustering. This research utilizes the first approach.

Similarity Measurement

The similarity measurement utilizes Dynamic Time Warping (DTW) distance. DTW distance is an algorithm used to compare two data series and calculate the optimal alignment between them. DTW distance is a generalization of classic algorithms that compare continuous value sequences with discrete value sequences (Munthe, 2019). This research used three similarity measurement method which are *average linkage, complete linkage, dan centroid linkage.* The DTW formula followed by.

$$d_{DTW} = min\frac{\sum_{k=1}^{K} w_k}{K}$$

with d_{DTW} is DTW distance, w_k is warping path k.

Cluster Validity

The accuracy and quality of the formed clusters are determined using the cophenetic correlation coefficient and the Silhouette coefficient. The cophenetic correlation coefficient measures the correlation between the Euclidean distance matrix and the cophenetic matrix (based on the distance metric and linkage method used). The value of the cophenetic correlation coefficient ranges from -1 to 1. The limitations in determining the formed clusters are that k ≤ n, with k = 1 and k = n being excluded. The criteria for the adequacy and quality of clustering based on the Silhouette Coefficient can be expressed as follows (Kaufman dan Rousseeuw, 1990).

Table 1. Clustering Category Based on Silhoutte Coefficient

Silhoutte Coefficient	Clustering Category
0.71 – 1.00	Strong Cluster
0.51 – 0.70	Good Cluster
0.26 – 0.50	Weak Cluster
0.00 – 0.25	Bad Cluster

RESULTS AND DISCUSSION

Descriptive Analysis

Before conducting time series cluster analysis, the initial step in data analysis is to describe the poverty depth index in 35 regencies/cities in Central Java from 2017 to 2022. The following is a descriptive analysis of the poverty depth index in Central Java Province from 2017 to 2022.

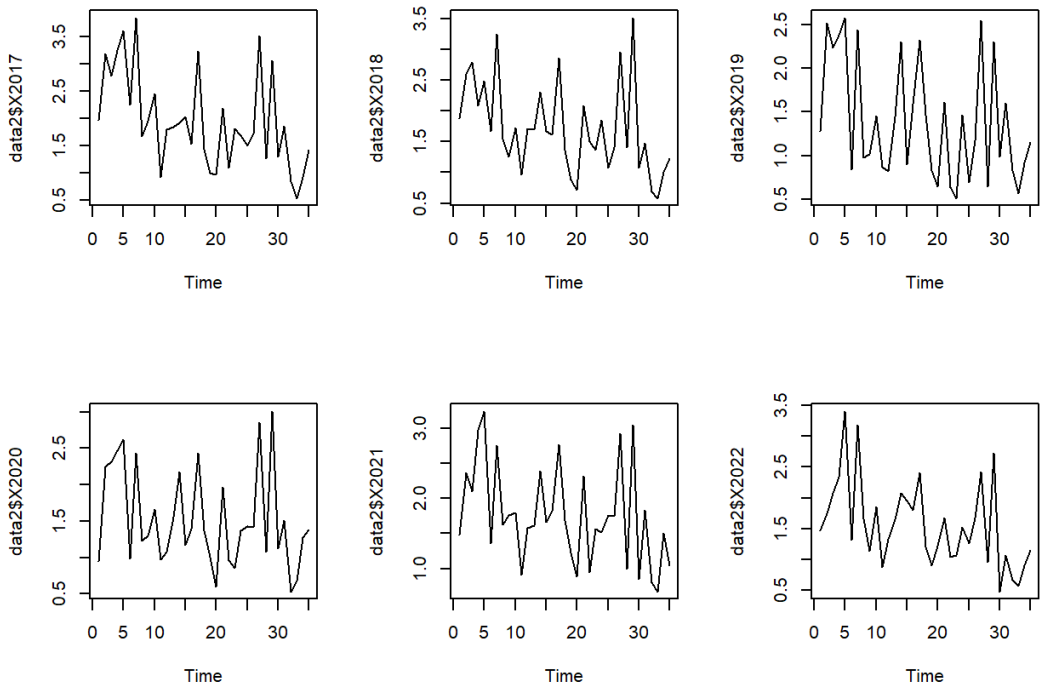

Figure 1. Index Depth of Poverty Plot Cental Java Province 2017 – 2022

Based on Figure 1 above, we can observe the distribution of the trend of the poverty depth index from 2017 to 2018. The poverty depth index is one of the indicators used to assess the level of poverty in a region. The poverty depth index measures the average expenditure inequality of each individual relative to the poverty line. (BPS, 2023). The poverty depth index in Central Java Province in Figure 1 exhibits a similar trend over time. However, there is a slight difference in the trend in 2022 compared to previous years.

Clustering With DTW Distance

The next step is to create a dendrogram for time series cluster analysis using the DTW distance. The creation of this dendrogram involves the application of various linkage methods, such as average linkage, complete linkage, and centroid linkage. The following are the resulting dendrograms and their respective cutoff points.

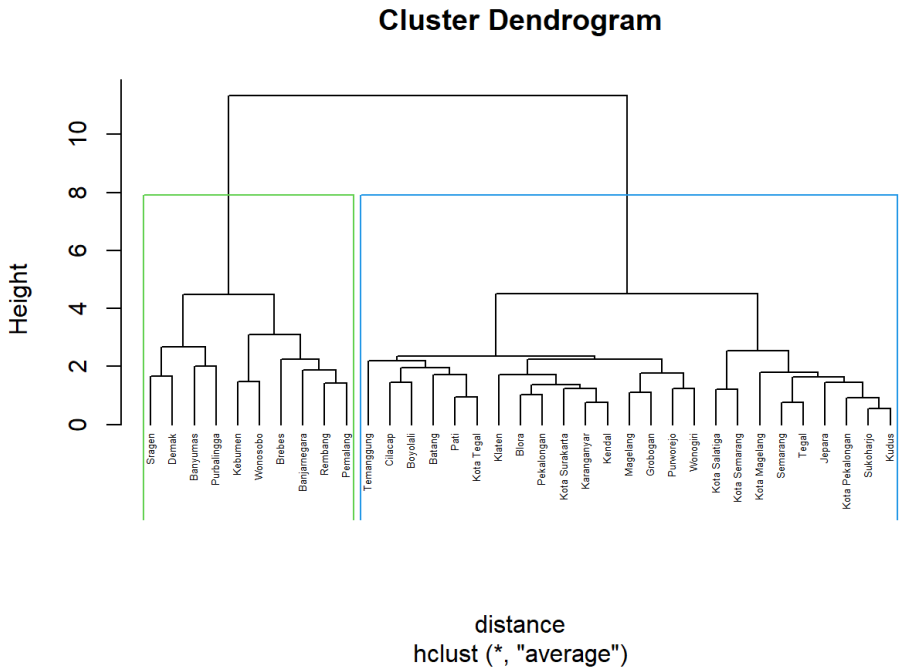

Figure 2. Dendogram Clustering of *Average linkage*

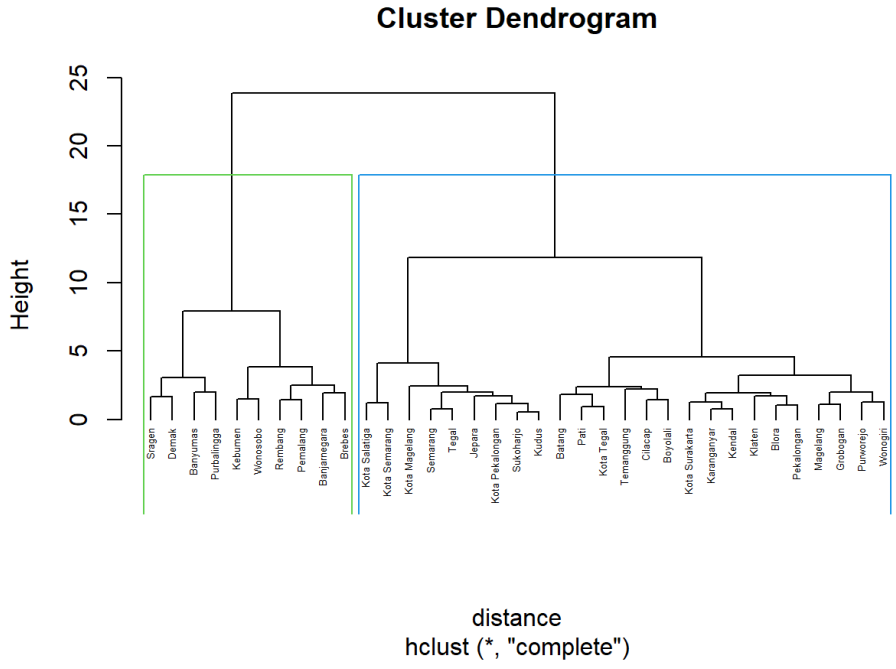

Figure 2. Dendogram Clustering of *Complete linkage*

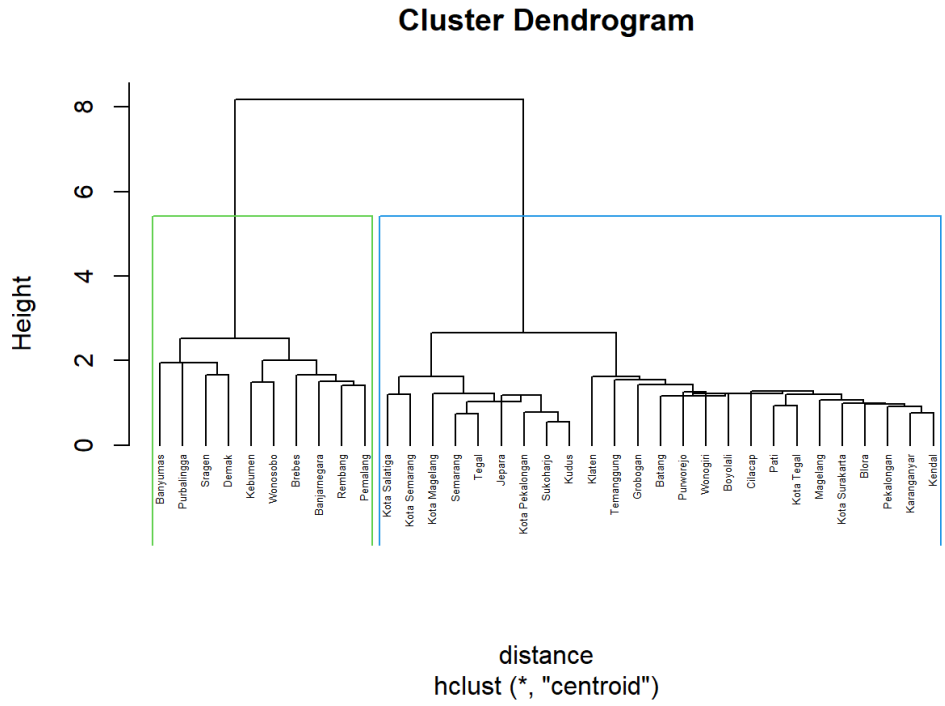

Figure 3. Dendogram Clustering of *Centroid linkage*

Based on Figure 2, using the average linkage method, two clusters are formed. The first cluster consists of 25 Districts/Cities, including Temanggung, Cilacap, Boyolali, Batang, Pati, Tegal City, Klaten, Blora, Pekalongan, Surakarta City, Karanganyar, Kendal, Magelang, Grobogan, Purworejo, Wonogiri, Salatiga City, Semarang City, Magelang City, Semarang, Tegal, Jepara, Pekalongan City, Sukoharjo, and Kudus. The second cluster consists of 10 Districts/Cities, including Sragen, Demak, Banyumas, Purbalingga, Kebumen, Wonosobo, Brebes, Banjarnegara, Rembang, and Pemalang.

Based on Figure 3, using the complete linkage method, two clusters are formed. The first cluster consists of 25 Districts/Cities, including Salatiga City, Semarang City, Magelang City, Semarang, Tegal, Jepara, Pekalongan City, Sukoharjo, Kudus, Batang, Pati, Tegal, Temanggung, Cilacap, Boyolali, Surakarta City, Karanganyar, Kendal, Klaten, Blora, Pekalongan, Magelang, Grobogan, Purworejo, and Wonogiri. The second cluster consists of 10 Districts/Cities, including Sragen, Demak, Banyumas, Purbalingga, Kebumen, Wonosobo, Rembang, Pemalang, Banjarnegara, and Brebes.

Based on Figure 4, using the centroid linkage method, two clusters are formed. The first cluster consists of 25 Districts/Cities, including Salatiga City, Semarang City, Magelang City, Semarang, Tegal, Jepara, Pekalongan City, Sukoharjo, Kudus, Klaten, Temanggung, Grobogan, Batang, Purworejo, Boyolali, Cilacap, Pati, Tegal City, Magelang, Surakarta City, Blora, Pekalongan, and Karanganyar, Kendal. The second cluster consists of 10 Districts/Cities, including Banyumas, Purbalingga, Sragen, Demak, Kebumen, Wonosobo, Brebes, Banjarnegara, Rembang, and Pemalang.

Clustering Validity

After conducting time series cluster analysis based on DTW and creating a dendrogram, the next step is to calculate the clustering validity to determine the most optimal similarity measurement based on *Cophenetic* Correlation.

Table 2. Coefficient Value of *Cophenetic* Correlation

Distance	Similarity Measurement		
	Average	Complete	Centroid
DTW	0.8017988	0.7917858	0.8003558

Based on Table 2, it can be observed that each linkage method has different *Cophenetic* Correlation coefficient values with DTW distance. The cophenetic correlation coefficient ranges from -1 to 1, where a higher value indicates a better measure of similarity in the clustering process. Based on the cophenetic

correlation coefficient values of the three linkage methods mentioned above, it can be concluded that average linkage has the highest coefficient value, which is 0.8017988.

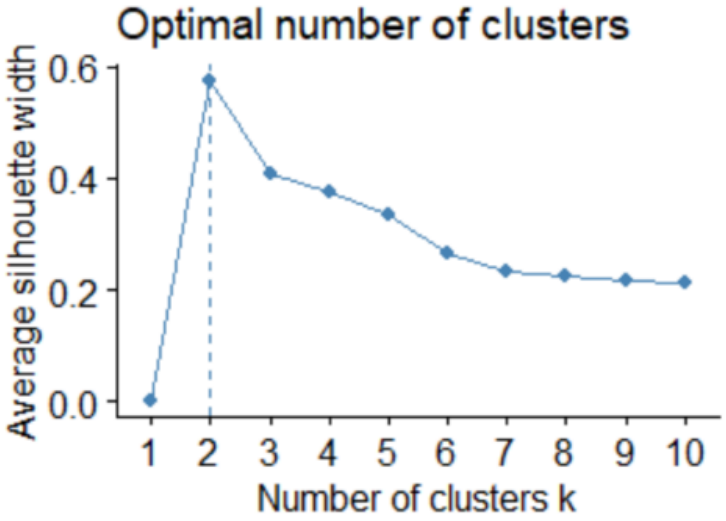

Figure 4. Number Cluster Optimum Plot with *Silhoutte* Coefficient

The next step is to test the quality of the clustering in the study using the silhouette coefficient. This quality testing is conducted to determine whether the number of clusters, which is 2 in this case, is representative or not. Based on the silhouette coefficient values in Figure 5 above, it can be observed that the 2-cluster solution has an average silhouette coefficient of 0.6. The clusters formed in each dendrogram with 3 similarity measurements form the same clusters, namely the cluster with high poverty depth index and the cluster with low poverty index. The cluster with high poverty depth index had 10 District/Cities and the cluster with low poverty index had 25 District/Cities. This average silhouette coefficient value indicates that the clustering using average linkage with DTW distance measurement falls into the "good cluster" category.

CONCLUSION

The time series clustering analysis in this study aims to cluster the Districts/Cities in Central Java Province based on the poverty depth index values from 2017 to 2022. Based on the correlation coefficient values using three linkage methods, the average linkage method has the highest cophenetic correlation coefficient of 0.8017988. The quality testing of the clusters using the silhouette coefficient with DTW distance and the average linkage method shows that the 2-cluster solution falls into the "good cluster" category with a silhouette coefficient value of 0.60. The resulting clusters using DTW distance and average linkage method consist of Cluster 1 with 25 Districts/Cities and Cluster 2 with 10 Districts/Cities.

REFERENCES

Badan Perencanaan Pembangunan Nasional. (2023). *No Title*. https://simreg.bappenas.go.id/home/pemantauan/tk

Buaton, R., Zarlis, M., Mawengkang, H., & Effendi, S. (2019). Clustering Time Series Data Mining dengan Jarak Kedekatan Manhattan City. *Prosiding Seminar Nasional Riset Information Science (SENARIS), 1*(September), 1155. https://doi.org/10.30645/senaris.v1i0.129

Dani, A. T. R., Wahyuningsih, S., & Rizki, N. A. (2020). Pengelompokkan Data Runtun Waktu menggunakan Analisis Cluster (Studi Kasus: Nilai Ekspor Komoditi Migas dan Nonmigas Provinsi Kalimantan Timur Periode Januari 2000-Desember 2016). *Jurnal EKSPONENSIAL*, *11*(1), 29–38.

Kaufman, Leonard; Rousseeuw, P. J. (1990). *Finding Groups In Data An Introduction to Cluster Analysis*. A John Wiley & Sons.

Lestari, W. I. (2022). Faktor-Faktor Yang Mempengaruhi Kemiskinan Per Provinsi Di Indonesia Dalam Perspektif Islam. *Jurnal Ilmiah Ekonomi Islam*, *8*(03), 3136–3144. https://jurnal.stie-aas.ac.id/index.php/jei/article/view/6208%0Ahttps://jurnal.stie-aas.ac.id/index.php/jei/article/download/6208/2820

Lisa Anggraini, Prizka Rismawati Arum, M. S. (2021). *Analisis Cluster Menggunakan Algoritma K-Means Pada Provinsi Sumatera Barat Berdasarkan Indeks Pembangunan Manusia Tahun 2021*. 11. https://prosiding.unimus.ac.id/index.php/semnas/article/viewFile/1214/1211

Munthe, A. D. (2019). Penerapan Clustering Time Series Untuk Menggerombolkan Provinsi Di Indonesia Berdasarkan Nilai Produksi Padi. *Jurnal Litbang Sukowati : Media Penelitian Dan Pengembangan*, *2*(2), 11. https://doi.org/10.32630/sukowati.v2i2.61

Nafi'ah, B. (2021). Analisis Faktor-Faktor Yang Dapat Mempengaruhi Pengentasan Kemiskinan Di Indonesia (2016- 2019). *Jurnal Ilmiah Ekonomi Islam*, *7*(2), 953–960. https://doi.org/10.29040/jiei.v7i2.2206

Pusat Statistik, B. (2022). *No Title*. https://www.bps.go.id/indicator/23/503/1/indeks-kedalaman-kemiskinan-p1-menurut-provinsi-dan-daerah.html

Pusat Statistik, B. (2023). *Berita Resmi Statistik*. https://www.bps.go.id/website/images/Kemiskinan-Sep-2022-ind.jpeg

Soleha, H. A., Nurmayanti, W. P., Hidayaturrohman, U., Haiban, R., Tgkh, J., Abdul, M. Z., & No, M. (2022). *Penerapan Clustering Time Series pada Pengelompokan Provinsi di Indonesia (Studi Kasus : Nilai Ekspor Non Migas di Indonesia Tahun 2016-2020) Program Studi Statistika , Universitas Hamzanwadi*. *15*(2), 286–291.

Suhaeni, C., Kurnia, A., & Ristiyanti, R. (2018). Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil). *Jurnal Media Infotama*, *14*(1). https://doi.org/10.37676/jmi.v14i1.469

Yanti, Y., & Rahardiantoro, S. (2018). *(Studi Kasus Penggerombolan Provinsi di Indonesia Berdasarkan*. 13–22.

Yusfar, A. A., Tiro, M. A., & Sudarmin, S. (2020). Analisis Cluster Ensemble dalam Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Kinerja Pembangunan Ekonomi Daerah. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *3*(1), 31. https://doi.org/10.35580/variansiunm14626