Jurnal Matematika UNAND
Vol. **12** No. **3** Hal. 203 – 212
ISSN : 2303–291X
e-ISSN : 2721–9410

AN ANALYSIS OF CLUSTER TIMES SERIES FOR THE NUMBER OF COVID-19 CASES IN WEST JAVA

NURFITRI IMRO'AH[a], NUR'AINUL MIFTAHUL HUDA[b*]

[a] *Statistics Department, Universitas Tanjungpura, Pontianak, Indonesia,*
[b] *Mathematics Department, Universitas Tanjungpura, Pontianak, Indonesia.*
email : nurfitriimroah@math.untan.ac.id, nur'ainul@fmipa.untan.ac.id

Accepted October 22, 2022 Revised March 7, 2023 Published October 21, 2023

Abstract. The government may be able to develop more effective strategies for dealing with COVID-19 cases if it groups districts and cities according to the features of the number of Covid-19 cases being reported in each district or city. The data can be more easily summarized with the help of cluster analysis, which organizes items into groups according to the degree of similarity between members. Since it is possible to group more than one period together, the generation of clusters based on time series is a more efficient method than clusters that are created for each individual unit. Using a time series cluster hierarchical technique that has complete linkage, the purpose of this study is to categorize the number of instances of Covid-19 that have been found in West Java by district or city. The data that was used comes from monthly reports of Covid-19 instances compiled by West Java districts from 2020 to 2022. The Autocorrelation Function (ACF) distance cluster was utilized in this investigation to determine how closely cluster members are related to one another. According to the findings, there could be as many as seven separate clusters, each including a unique assortment of districts and cities. Cluster 3, which is comprised of three different cities and regencies, including Bandung City, West Bandung Regency, and Sumedang Regency, has an average number of cases that is 66, making it the cluster with the highest number of cases overall. A value of 0.2787590 is obtained for the silhouette coefficient as a result of the established grouping. This value suggests that the structure of the newly created cluster is quite fragile.

Keywords: ACF distance cluster, hierarchy, Silhouette coefficient

1. Introduction

Cluster analysis is a method used to group objects into relatively uniform groups [1]. The purpose of this analysis is not to associate or distinguish one object from another but to identify groups of objects with certain similarities and characteristics that distinguish them from other groups. Objects in the same group are relatively more homogeneous than objects in different groups. Objects in one group are very similar, but objects in another are very different [2]. Cluster analysis is generally used for five data types: interval, binary, nominal, ordinal, and ratio scale variables

*Corresponding Author

[3]. Nevertheless, cluster analysis can also be used to group time series data, called cluster time series analysis [4].

Cluster analysis applied to time series data has a different grouping method than cross-sectional data. It is because time series data is a series of observations that occur sequentially at regular intervals [5]. In the process of grouping time series data, many methods have been developed, including the use of a distance similarity measure that depends on the characteristics of the time series data. Time series data sets can be grouped based on the factors of each data using time series clustering, namely by grouping objects based on time series patterns. Cluster time series is helpful for grouping objects that have similar data patterns [6]. Measurement of similarity in time series clusters can use the Autocorrelation Function distance cluster. In this study, time series cluster analysis was used to categorize the number of Covid-19 cases in West Java.

West Java is the province with the second most daily Covid-19 cases after DKI Jakarta. Covid-19 cases in West Java Province have increased again. Based on data from the West Java, active cases of Corona in West Java currently reach 6,645 cases. The data on the number of COVID-19 cases was updated on Saturday (16/7/2022) at 17.00 WIB. Within a day, there were 559 additional Corona cases in West Java. The ten areas with the highest daily spread of COVID-19 cases in West Java are Depok City with 208 cases, Bandung City with 133 cases, and Bekasi City with 97 cases. Then Bogor Regency with 34 cases and Bekasi Regency with 27 cases. Furthermore, Bogor City with 24 cases, West Bandung Regency with 17 cases, and Cimahi City with 17 cases. Then Karawang Regency 7 cases and Sukabumi Regency 7 cases [7].

The government's efforts to prevent the transmission of Covid-19 and break the chain of transmission are by self-isolate for those who are exposed to the virus or who travel outside the city or abroad island, stay home, and always wash hands and use a mask when outside the home. However, despite the enactment of recommendations or Social Restrictions Large-Scale (PSBB), Covid-19 cases continue to grow daily. Every day the spread of Covid-19 cases in West Java continues to increase. The government needs to establish more effective policies to handle Covid-19 cases. So a solution that is required to find out this problem is to create a system that can provide information about grouping data for Covid-19 sufferers in clusters. Grouping areas or districts/cities that have a level of vulnerabilities the higher spread of Covid-19 is one of the ways to limit the activity of the spread of Covid-19 increasingly widespread to reduce the increase in coronavirus cases.

This study aims to classify the number of Covid-19 cases by district in West Java using a time series cluster hierarchical method with complete linkage. The data used is monthly data from Covid-19 cases by the district in West Java from 2020 to 2022. The data is analyzed using a time series cluster hierarchical method with complete linkage based on the ACF distance cluster distance.

2. Cluster Time Series

Cluster analysis is a method of multivariate analysis to classify objects of observation into several groups. In these so obtained groups, objects in one group have many similarities. Meanwhile, with other group members, there are many differences [8]. The measure of similarity is the most crucial thing in cluster analysis. For the case of clusters in time series data, one measure of similarity that can be used is the ACF distance cluster [9]. Given time series $\{x_t, t = 1, 2, 3, \cdots, T\}$, let $\hat{\rho}_{x_r} = (\hat{\rho}_1, \hat{\rho}_2, \cdots, \hat{\rho}_R)$ is the autocorrelation function estimated from time series x_t from lag 1 to lag R such that $\hat{\rho}_i \cong 0$ for $i > R$. The distance between the two time series x_t and y_t can be determined as [10]:

$$d_{ACF}(x_t, y_t) = \sqrt{(\hat{\rho}_{x_r} - \hat{\rho}_{y_r})'\Omega(\hat{\rho}_{x_r} - \hat{\rho}_{y_r})}, \tag{2.1}$$

where $d_{ACF}(x_t, y_t)$ is distance of autocorrelation vektor x_t and y_t; $\hat{\rho}_{x_r}$ is autocorrelation estimation function x_r; $\hat{\rho}_{y_r}$ is autocorrelation estimation function y_r; Ω is weight matrix; and r is time lag. If the ACF distance does not use weights , then the weight matrix is in the form of an identity matrix. So the distance equation ACF is [11]:

$$d_{ACF}(x_t, y_t) = \sqrt{(\hat{\rho}_{x_r} - \hat{\rho}_{y_r})'(\hat{\rho}_{x_r} - \hat{\rho}_{y_r})}. \tag{2.2}$$

An important connection exists between estimating the auto-correlation vector x_r and the stochastic process that underlies its calculation. A stochastic process is defined as a collection of random variables arranged in time and is defined as a set of points that may be discrete or continuous. Random variable at time t with $X(t)$ if time is continuous or with X_t if time is discrete. The continuous stochastic process is described as $\{X(t), -\infty < t < \infty\}$ while the discrete stochastic process is described as $\{X_t, t = \cdots, -2, -1, 0, 1, 2, \cdots\}$. The estimation of the auto-correlation vector x_r is explained in

$$\hat{\rho}_{x_r} = \frac{c_s}{c_0} = \frac{\frac{1}{T}\sum_{s-1}^{T-s}(x_t - \bar{x})(x_{t+s} - \bar{x})}{\frac{1}{T}\sum_{t-1}^{T-1}(x_t - \bar{x}^2)} = \frac{\sum_{s-1}^{T-s}(x_t - \bar{x})(x_{t+s} - \bar{x})}{\sum_{t-1}^{T-1}(x_t - \bar{x}^2)}.$$

The estimation of the autocorrelation vector y_r is

$$\hat{\rho}_{y_r} = \frac{c_s}{c_0} = \frac{\frac{1}{T}\sum_{s-1}^{T-s}(y_t - \bar{y})(y_{t+s} - \bar{y})}{\frac{1}{T}\sum_{t-1}^{T-1}(y_t - \bar{y}^2)} = \frac{\sum_{s-1}^{T-s}(y_t - \bar{y})(y_{t+s} - \bar{y})}{\sum_{t-1}^{T-1}(y_t - \bar{y}^2)}.$$

2.1. *Complete Linkage Method*

The complete linkage method in determining the distance between clusters is done by looking at the distance between two clusters with the closest maximum distance combined. This process is repeated until only one cluster remains. The complete linkage algorithm is a hierarchical algorithm to form clusters based on the furthest distance between objects [12]. The complete linkage algorithm begins by selecting the most significant distance in the matrix $\mathbf{D} = \{d_{ij}\}$, then combining the corresponding objects such as U and V to obtain clusters (UV). The next step is to find the distance between (UV) and other clusters, for example, W. The equation used

to determine the distance between the clusters (U, V) and W is in the following equation [13]:

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}, \tag{2.3}$$

where d_{UW} is distance between U and W cluster; d_{VW} is distance between V and W cluster.

2.2. *Cluster Validation*

One of the problems in cluster analysis is determining the optimum number of clusters. Therefore, in conducting cluster analysis, performing a cluster validation test is necessary to resolve the optimum number of clusters. Cluster validation tests were also carried out to determine the results of the cluster groups formed that could explain and represent the population in general [14]. The validity of this study uses a silhouette index. The silhouette coefficient is used to see the quality and strength of the cluster and how well or poorly an object is placed in a cluster [15]. This method is a combination of the way of separation and cohesion. It is necessary to calculate the silhouette coefficient value from the i-th data to calculate the silhouette index value. The silhouette coefficient value is obtained by finding the maximum value of the global silhouette index value from the number of clusters 2 to the number of clusters $n - 1$. The formula for finding the Silhouette Coefficient value is [16]:

$$SC = \max_{k} SI(k), \tag{2.4}$$

where SC is Silhouette Coefficient; SI is average Silhouette Index of dataset; and k is number of cluster. While the calculation of the value of the Silhouette Index is notated [17]:

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j, \tag{2.5}$$

where SI is average Silhouette Index of dataset; SI_j is average Silhouette Index of cluster j; and k is number of cluster.

The silhouette index is calculated as the degree of confidence in the clustering process on an observation. The cluster formed is said to be good if the index value is close to 1. Meanwhile, if the index value is relative to -1, then the cluster formed is said to be not good. Based on this, the greater the silhouette index value, the better or optimum cluster formed. Table 1 presents the subjective criteria for grouping measurements based on the Silhouette Coefficient (SC) [18].

3. Case Study

This study uses data on Covid-19 cases in 27 districts/cities in West Java Province. Data were collected monthly from April 2020 to June 2022 from the National Disaster Management Agency (BNPB). An overview of the data used is presented in Figure 1.

Table 1: Silhoutte Coefficient Measurement Criteria

Value of SC	Criteria
$0,71 - 1,00$	Strong structure
$O,51 - 0,70$	Good structure
$0,26 - 0,50$	Weak structure
$\leq 0,25$	Bad structure

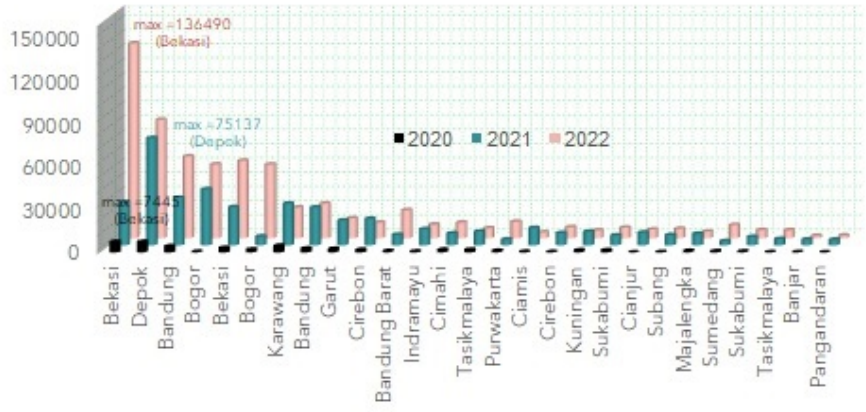

Figure. 1: The number of Covid-19 cases in West Java by district/city.

Based on Figure 1 as a whole, it can be seen that there was an increase in the number of Covid-19 cases from April 2020 to June 2022. Most cases during the research period were in Bekasi, namely 136,490 cases that occurred in 2022. Cases of Covid-19 new variant of Omicron increased in Bekasi City. In this regard, the Bekasi City Government issued a Circular regarding the Implementation (PPKM) of Level 3 Corona Virus Disease 2019 Restrictions on Community Activities from February 8 to 14, 2022. The circular regulates the tightening of community activities. The Bekasi City Government continues to increase vaccination with priority for the elderly. In addition, the strengthening of 'Testing, Tracing and Treatment' in implementing PPKM is also carried out. The Bekasi City Government noted that the highest daily positive cases of Covid-19 occurred on February 12, 2022, with the daily confirmed number of 3,359 cases. However, for four consecutive days, daily positive cases of Covid-19 in Bekasi City experienced a downward trend. On February 15, 2022, there were 3,175 cases. Then, on February 16, 2022, there were 3,100 cases; on February 17, 2022, there were 3,088 cases, and on February 20, 2022, the trend of daily positive cases of Covid-19 was decreasing, namely 2,527 cases.

The similarity measure used for time series clusters is the ACF distance cluster

calculated based on Equation 2.1. The data with the smallest ACF distance cluster value is the distance between Bekasi and Purwakarta, with a value of 0.073819. It means that Bekasi and Purwakarta have the highest similarity measure. Meanwhile, the most significant ACF distance cluster value is the distance between Bogor and Bandung, with a value of 0.721069. It means that Bogor and Bandung have the lowest similarity measure.

The silhouette index is used to measure how close the similarity of objects in a cluster. In other words, the silhouette index also shows how precisely the objects have been grouped so that the larger the silhouette index value, the more similar the objects in a cluster are. The silhouette index value ranges from -1 to 1. If it is closer to 1, then the number of clusters is the most optimal. Table 2 presents in detail the global silhouette index values obtained.

Table 2: The Global Silhouette Index Values

The number of clusters	The average silhouette index score
2	0.2445460
3	0.2012234
4	0.2328841
5	0.2603082
6	0.2707557
7	**0.2757890**
8	0.2538347
9	0.2542306
10	0.2399109

Based on Table 2, the optimum silhouette index is found in the 7 clusters formed. It is because the highest silhouette index value is 0.2757890. Therefore, it is determined that the hierarchical cluster formed is cluster 7. It can also be seen in the silhouette score plot in Figure 2 which shows that $k = 7$ is the most optimal.

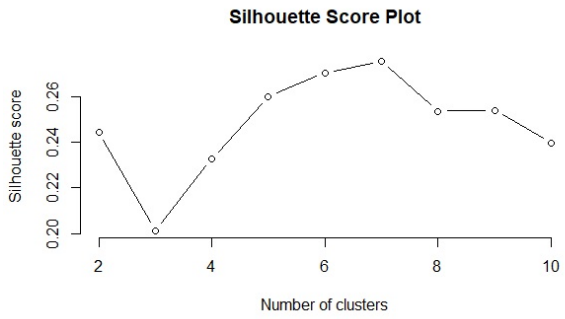

Figure. 2: Silhouette Score Plot.

The silhouette coefficient is the maximum value of the global silhouette index value from the number of clusters 2 to the number of clusters 10. Based on the silhouette index value, the silhouette coefficient value is 0.2757890. It means that the 7 clusters formed have a weak structure.

The complete linkage method is a method that groups the two clusters that have the furthest distance first. The cluster formed in the complete linkage method using the ACF distance cluster can be seen from the results of the dendrogram and dendrogram cutting in Figure 3.

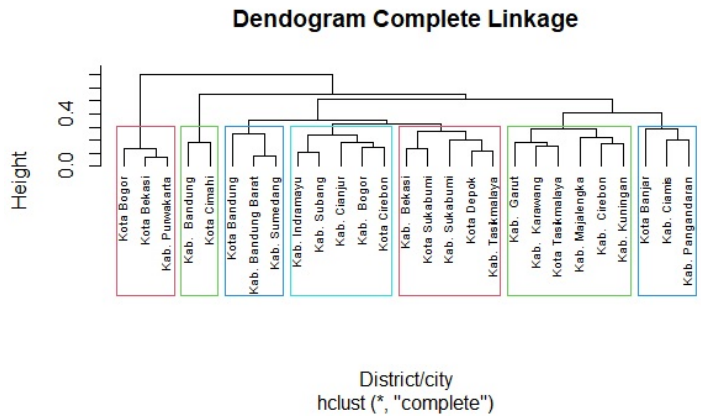

Figure. 3: Dendrogram Complete Linkage.

The results of cluster formation with the complete linkage method with each cluster consisting of members are shown in Table 3.

The average obtained from each cluster is used to see its characteristics. The following Table 4 presents the features of each cluster.

Based on Table 4, it can be seen that the average number of new Covid-19 cases in 7 clusters experienced significant differences. The highest average occurred in cluster 3, which only consisted of 3 regencies/cities, namely Bandung City, West Bandung Regency, and Sumedang Regency. Districts/cities in cluster 3 in mid-2021 had the status of standby 1. This status was determined based on the Bed Occupancy Rate (BOR) of patients, which reached 84.19 percent. This figure exceeds the WHO and national provisions of $60-70$ percent. In addition, these regencies/cities are in the red zone of the alert level. The red zone area plus high BOR can be an indicator of setting alert one because it is in an agglomeration area that influences each other. Therefore, the government enforces Work From Home (WFH) for seven days and urges no tourists to visit until the situation is under control. In addition, the government is also maximizing vaccination in the zone to pursue group immunity. This condition is based on the spike in new cases of COVID-19, which has been proven due to the long holiday of Eid al-Fitr 1442 H and the lack of discipline of the community in implementing health protocols.

Table 3: The Results of Cluster Formation

Cluster	The districts/city	The number of districts/city
1	Kota Bekasi	
	Kota Bogor	
	Kab. Purwakarta	3
2	Kota Depok	
	Kab. Bekasi	
	Kab. Sukabumi	
	Kota Sukabumi	
	Kab. Tasikmalaya	5
3	Kota Bandung	
	Kab. Bandung Barat	
	Kab. Sumedang	3
4	Kab. Bogor	
	Kab. Indramayu	
	Kota Cirebon	
	Kab. Cianjur	
	Kab. Subang	5
5	Kab. Karawang	
	Kab. Garut	
	Kab. Cirebon	
	Kota Tasikmalaya	
	Kab. Kuningan	
	Kab. Majalengka	6
6	Kab. Bandung	
	Kota Cimahi	2
7	Kab. Ciamis	
	Kota Banjar	
	Kab. Pangandaran	3

Table 4: The Characteristics of Each Cluster

Cluster	The average of each cluster
1	34.666667
2	57.400000
3	**66.000000**
4	19.000000
5	3.333333
6	42.000000
7	2.666667

4. Conclusion

The results of this study suggest that the time series cluster analysis performed using the ACF distance cluster yields seven clusters, each of which contains a unique set of members. This finding can be deduced from the discussion contained within this study. Cluster 3 is comprised of three different cities and regencies in West

Java. These cities and regencies are Bandung City, West Bandung Regency, and Sumedang Regency. The cluster's average number of cases is 66, making it the one with the greatest number of cases overall. The extended holiday for Eid al-Fitr in 1442 H and the absence of community discipline in the implementation of health regulations are the likely causes of the highest average number of cluster 3 cases. A value of 0.2787590 is obtained for the silhouette coefficient as a result of the established grouping. This score suggests that the structure of the newly created cluster is quite fragile.

5. Acknowledgment

The authors would like to express their gratitude to everyone who contributed to the writing of this paper; however, each individual contributor cannot be named.

References

[1] Knote, R., Janson, A., Sllner, M., Leimeister, J. M., 2019, Classifying Smart Personal Assistants: An Empirical Cluster Analysis, https://doi.org/10.24251/HICSS.2019.245

[2] Ali, A., Sheng-Chang, C., 2020, Characterization of well logs using K-mean cluster analysis, *Journal of Petroleum Exploration and Production Technology*, Vol. **10**(6): 2245 – 2256

[3] Abualigah, L. M., Khader, A. T., and Hanandeh, E. S., 2018, Hybrid clustering analysis using improved krill herd algorithm, *Applied Intelligence*, Vol. **48**(11): 4047 – 4071

[4] Gao, Z., Wei, S., Wang, L., Fan, S., 2020, Exploring the Spatial-Temporal Characteristics of Traditional Public Bicycle Use in Yancheng, China: A Perspective of Time Series Cluster of Stations, *Sustainability*, Vol. **12**(16): 6370

[5] Younan, M., Houssein, E. H., Elhoseny, M., Ali, A. E. A., 2020, Improved Models for Time Series Cluster Representation Based Dynamic Time Warping, *2020 15th International Conference on Computer Engineering and Systems (ICCES)*: 1 – 6

[6] Alexander, C., Shi, L., Akhmametyeva, S., 2018. Using Quantum Mechanics to Cluster Time Series, https://doi.org/10.48550/arXiv.1805.01711

[7] National Disaster Management Agency (BNPB), 2022

[8] Shaukat, M. A., Shaukat, H. R., Qadir, Z., Munawar, H. S., Kouzani, A. Z., Mahmud, M. A. P., 2021, Cluster Analysis and Model Comparison Using Smart Meter Data, *Sensors*, Vol. **21**(9): 3157

[9] Pappu, A. R., Kar, S., Kadu, S., 2022, ACF/PACF-Based Distance Measurement Techniques for Detection of Blockages in Impulse Lines of a Pressure Measurement Circuit for Nuclear Reactors: 1017 – 1029

[10] Setiawan, I., Sumertajaya, I. M., Afendi, F. M., 2021, Predicting and forecasting of time series models using cluster analysis, *Journal of Physics: Conference Series*, Vol. **1763**(1): 012035

[11] Anastasiou, A., Hatzopoulos, P., Karagrigoriou, A., and Mavridoglou, G., 2021, Causality Distance Measures for Multivariate Time Series with Applications, *Mathematics*, Vol. **9**(21): 2708

[12] Xinyi, C., 2022, Comparison between Complete and Wards Linkage Method in

Hierarchical Clustering Analysis on Cancer Omics Dataset. *2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB)*: 73 – 77

[13] Mattiev, J., and Kavsek, B., 2021, Distance based clustering of class association rules to build a compact, accurate and descriptive classifier, *Computer Science and Information Systems*, Vol. **18**(3): 791 – 811

[14] Kumar, S. S., Ahmed, S. T., Vigneshwaran, P., Sandeep, H., Singh, H. M., 2021, RETRACTED ARTICLE: Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets, *Journal of Ambient Intelligence and Humanized Computing*, Vol. **12**(7): 7581 – 7594

[15] Dinh, D.-T., Fujinami, T., Huynh, V.-N., 2019, Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient: 1 – 17

[16] Jin-Heng, G., Jia-Xiang, L., Zhen-Chang, Z., and Han-Yu, L., 2022, CDBSCAN: Density clustering based on silhouette coefficient constraints, *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*: 600 – 605

[17] Nidheesh, N., Nazeer, K. A. A., Ameer, P. M., 2020, A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data, *Neural Computing and Applications*, Vol. **32**(15): 11459 – 11476

[18] Wang, Z., Wang, H., 2021, Global Data Distribution Weighted Synthetic Oversampling Technique for Imbalanced Learning, *IEEE Access*, Vol. **9**: 44770 – 44783