BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1257–1264
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1257-1264

CLUSTERIZATION OF REGION IN SOUTH SUMATRA BASED ON COVID-19 CASE DATA

Anita Saragih[1], Dian Cahyawati[2*], Ning Eliyati[3]

[1,2,3]*Department of Mathematics and Natural Science, Sriwijaya University*
Palembang-Prabumulih Street, km.32, Ogan Ilir, 30862, Indonesia

*Corresponding author's e-mail: * **dianc_mipa@unsri.ac.id***

ABSTRACT

Article History:

Received: 18th December 2022
Revised: 5th July 2023
Accepted: 15th July 2023

Keywords:

Agglomerative method;
Cluster analysis;
COVID-19;
Silhouette;
South Sumatra.

Based on COVID-19 case data as of July 2022, South Sumatra Province has the 15th highest rank out of 34 provinces in Indonesia, with confirmed cases totaling 82,407. This showed that the spread of COVID-19 in South Sumatra was still high. This study aimed to determine the cluster of regions in South Sumatra based on COVID-19 case data. Clustering regions used an agglomerative hierarchical method. The process began with standardizing the data, calculating the similarity distance between objects, determining the optimal number of clusters using the Silhouette method, and clustering analysis. This study found that the optimal number of clusters consisted of two clusters. The clustering process starts with objects 2 and objects 4 because these two objects have the closest similarity distance. In conclusion, objects with the closest similarity distance (in one cluster) have the same data movement (fluctuation).

How to cite this article:

A. Saragih, D. Cahyawati and N. Eliyati., "CLUSTERIZATION OF REGION IN SOUTH SUMATRA BASED ON COVID-19 CASE DATA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1257-1264, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Coronavirus is a virus that directly attacks parts of the respiratory system; more precisely, this disease is known as COVID-19 [1]. The most common symptoms for someone affected by this virus are fever, cough, prolonged fatigue, and temporary loss of the ability to recognize tastes and smells. Apart from that, there are also abnormal symptoms for someone affected by COVID-19, namely sore throat, headache, aches and pains in the joints, diarrhea, experiencing skin rashes, and eye irritation. For someone exposed to serious symptoms of COVID-19, it is indicated by difficulty breathing because this virus attacks parts of the respiratory system, causing speaking difficulty and chest pain [2], [3], [4], [5].

South Sumatra Province is ranked 15th out of 34 provinces in Indonesia, with 82,407 confirmed cases, 13 additional cases, and 237 active cases. The population in South Sumatra Province in 2021, based on BPS.go.id data, is 8,550,849 people, with a classification of the total population as female 4,190,780 people and population as male sex 4,360,069 people. The population data for South Sumatra has increased by 0.98% compared to the population census data for 2020 [6]. This shows that the spread of COVID-19 in South Sumatra is still high. The distribution of positive cases of COVID-19 for each region in South Sumatra showed that Palembang, as the center of the city, has the highest positive cases, followed by Prabumulih and Lubuk Linggau [7]. It showed that the city regions have higher positive cases than other regions. The region that has the lowest positive case was Ogan Komering Ilir (OKI).

The difference in positive case data distribution in each region was interesting to study in clustering regions with similar characteristics. That was important to make it easier for local governments to describe the data of COVID-19 in the regions to provide solutions in reducing the spread of COVID-19 in South Sumatra. The clustering of regions in South Sumatra into several clusters based on the similarities of characteristics used cluster analysis [8],[9], dan [10]. The result of clusterization was expected to be an alternative solution for the Government in developing strategies and efforts to reduce the spread of COVID-19.

The problems in this study were limited by the case of suspect-probable, suspected-confirmed, close-suspicious contact, close-confirmed-contact, case symptomatic, asymptomatic, confirmed-recovery, and confirmed-death cases data on the status of COVID-19. The agglomerative hierarchical clustering method consists of single, complete, and average linkage. A study of the methods used is described in the second part about agglomerative hierarchical clustering. Furthermore, the third section discusses district/city clustering in South Sumatra based on COVID-19 case data using hierarchical analysis. Finally, the conclusion is given in the fourth section [11].

2. RESEARCH METHODS

The research used quantitative methods, namely the concept and theory of hierarchical cluster analysis with single linkages, complete linkages, and average linkages methods in statistics. The COVID-19 cases data were secondary data obtained from the official website [12]. The periods of the data were from January to July 2022.

The region in South Sumatra is the object of this research, which is described in Table 1:

Table 1. Object of Research

No	Region	Code
1	Kota Palembang	PLG
2	Kabupaten Ogan Komering Ilir	OKI
3	Kabupaten Lahat	LHT
4	Kabupaten Ogan Komering Ulu	OKU
5	Kabupaten Musi Banyuasin	MUBA
6	Kabupaten Muara Enim	ENIM
7	Kabupaten Musi Rawas	MURA
8	Kabupaten Banyuasin	BNS
9	Kabupaten Ogan Ilir	OI
10	Kota Prabumulih	PRB

No	Region	Code
11	Kota Pagar Alam	PGA
12	Kota Lubuk Linggau	LLG
13	Kabupaten Ogan Komering Ulu Timur	OKUT
14	Kabupaten Ogan Komering Ulu Selatan	OKUS
15	Kabupaten Empat Lawang	EMPLW
16	Kabupaten Penukal Abab Lematang Ilir	PALI
17	Kabupaten Musi Rawas Utara	MURATA

The steps of the agglomerative hierarchical cluster analysis were as follows:

1. Data Standardization

Data standardization was done before data analysis to avoid having different units and high variable differences [13]. The Z score equation is used to calculate standardized data as **Equation (1)**:

$$\boldsymbol{Z} = \frac{(\boldsymbol{x} - \mu)}{\boldsymbol{\sigma}} \tag{1}$$

where Z is the $\boldsymbol{z_{score}}$ value; x is the observed value; μ is the average; and σ is the standard deviation of the data.

2. Similarity Measure

The measurement of similarity in this study used the Euclidean distance [14]. Euclidean distance is the distance measured geometrically between two data objects. The **Equation (2)** used in measuring the Euclidean distance is:

$$\boldsymbol{d_{Euclidean}} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{a_i} - \boldsymbol{b_i})^2} \tag{2}$$

where $\boldsymbol{d_{Euclidean}}$ as Euclidean distance; $\boldsymbol{a_i}$ as a data *a* to-*i*; $\boldsymbol{b_i}$ As a data *b* to-*i*.

3. Cluster Analysis

The clustering process used the Agglomerative Hierarchical Clustering method, which consists of several functions, namely [15], [16]:

a. Single Linkages

This method uses the smallest distance in clustering each object that has similarities. For example, objects A and B correspond to each other to become clusters (AB). The next step is followed by calculating the minimum distance between cluster (AB) and cluster C, written as **Equation (3)**:

$$\boldsymbol{d_{(AB)C} = \min(d_{AC}, d_{AB})} \tag{3}$$

b. Complete Linkages

This method uses the greatest distance in clustering each object that has similarities. For example, objects A and B correspond to each other to become clusters (AB). The next step is followed by calculating the maximum distance between cluster (AB) and cluster C, written as **Equation (4)**:

$$\boldsymbol{d_{(AB)C} = \max(d_{AC}, d_{AB})} \tag{4}$$

c. Average Linkages

This method uses the average distance in clustering each object that has similarities. For example, objects A and B correspond to each other to become clusters (AB). The next step is followed by calculating the average distance between cluster (AB) and cluster C, written as **Equation (5)**:

$$\boldsymbol{d_{(AB)C} = \frac{(d_{AC}, d_{AB})}{n_{ab}\, n_{ac}}} \tag{5}$$

4. Similarity Validity

Validity aims to review the accuracy and quality of the results from the clustering process that has been carried out. Validity is measured using the Silhouette coefficient by entering objects into $\boldsymbol{k}$ clusters, namely $\boldsymbol{k} \leq \boldsymbol{n}$, with $\boldsymbol{k} = \mathbf{1}$ and $\boldsymbol{k} = \boldsymbol{n}$ excluded so that it can be written as **Equation (6)**:

$$\boldsymbol{s(i)} = \frac{\boldsymbol{b(i)} - \boldsymbol{a(i)}}{\mathbf{max}\big(\boldsymbol{a(i)}, \boldsymbol{b(i)}\big)} \tag{6}$$

Where $\boldsymbol{a(i)}$ is the average dissimilarity of each i-th object to all other objects in cluster A; $\boldsymbol{b(i)}$ is the average dissimilarity of each i-th object to all objects in cluster B, if cluster A is considered non-existent; $\boldsymbol{s(i)}$ is the Silhouette coefficient with a value range of $-\mathbf{1} \leq \boldsymbol{s(i)} \leq \mathbf{1}$ for each object.

According to Kaufman and Rausseeuw (1990), the interpretation of the Silhouette coefficients can be described as **Table 2**:

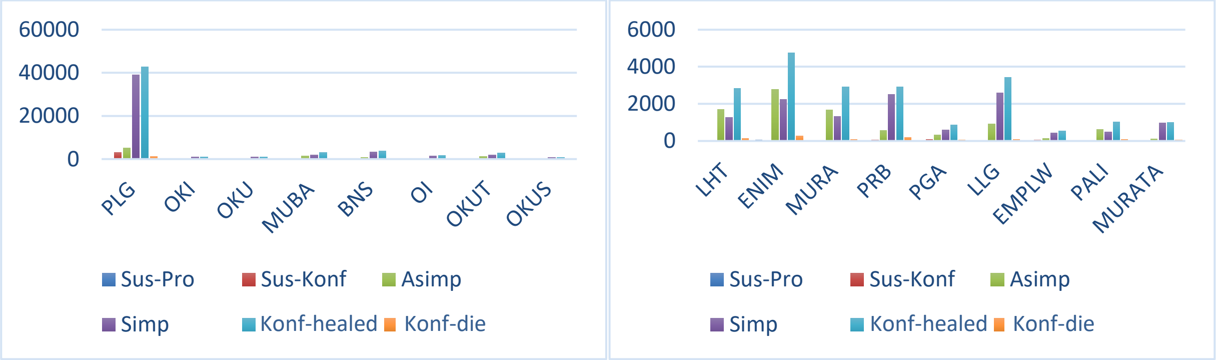
Table 2. Silhouette Coefficient

No.	Silhouette Coefficient	Classification Cluster
1.	0.71-1.00	Strong
2.	0.51-0.70	Good
3.	0.26-0.50	Weak
4.	0.00-0.25	Bad

Data Source: [17]

3. RESULTS AND DISCUSSION

The regions in South Sumatra were divided into two regions, namely the western and southern regions, in order to make it easier to display the fluctuations data. The active status of COVID-19 in South Sumatra for each object can be described as follows.



(a) (b)

Figure 2. Graph of COVID-19 Status for the Southern Region (a) and the Western Region (b) as of 31 July 2022

Figure 2 shows that the symptomatic and recovery confirmation COVID-19 case data had a significantly different number from other cases. This condition could affect the clustering process, so it was necessary to standardize the data using **Equation (1)**.

The measurement of the similarity distance used **Equation (2)**. This similarities distance process used the R application version 4.2.1. The result is displayed in **Table 2**.

Table 2. Distance Matrix of COVID-19 Case Data

Object	1	2	3	…	15	16	17
1	0.000						
2	11.604	0.000					
3	10.478	1.881	0.000				
4	11.509	0.122	1.778	…			
5	10.771	1.406	0.794	…			
⋮	⋮	⋮	⋮	…			

Object	1	2	3	…	15	16	17
14	11.557	0.292	1.958	…			
15	11.466	0.656	1.750	…	0.000		
16	11.145	0.930	1.222	…	0.552	0.000	
17	11.495	0.422	1.803	…	0.253	0.670	0.000

Table 2 shows the distance among objects. For example, the distance between Object 1 and Object 2 was 11.604. The distance between Object 1 and Object 3 was 10.478, and so on.

Furthermore, the clustering process using the Agglomerative Hierarchical Clustering method consists of three methods, namely single linkages, complete linkages, and average linkage. The single linkages method uses the smallest distance to cluster each object with similarities calculated using **Equation (3)**. The clustering process starts from the closest distance, so in the COVID-19 case data, the clustering process starts from Object-2 and Object-4, which had a distance of 0.122. The clustering process can be described in the form of a dendrogram. The dendrogram resulting from the clustering process on COVID-19 case data using the single linkages method is shown in **Figure 3**.

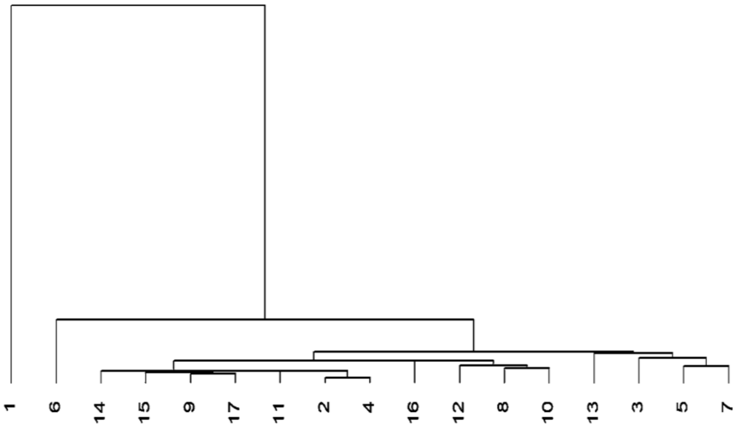


Figure 3. Dendrogram of COVID-19 cases by single linkage method

Figure 3 shows that Object 1 (PLG) was in the first cluster, and the other Objects were in the second cluster. Objects 2 (OKI) and 4 (OKU) were very close regions. The two of it have a similar number of case data. This similarity can be displayed in**Figure 4**.

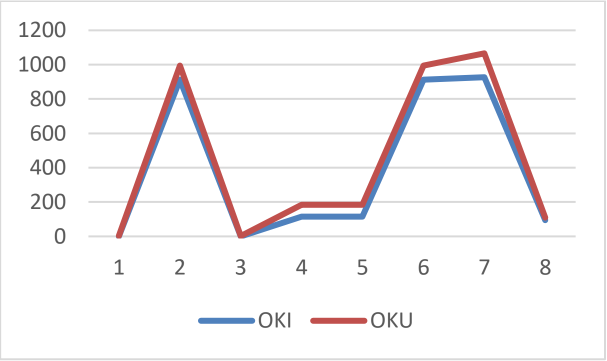


Figure 4. Comparison case data between Regency of Ogan Komering Ulu and Regency of Ogan Komering Ilir

The complete linkages method uses the greatest distance to classify each object with something in common. Using **Equation (4)**, each object results in a dendrogram in **Figure 5**.

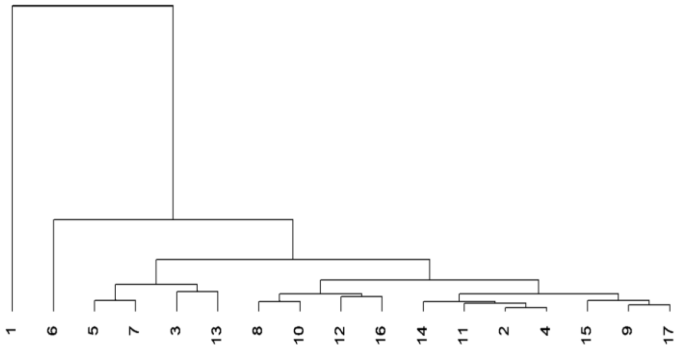


Figure 5. Dendrogram of COVID-19 cases (Complete Linkages)

Based on the dendrogram in Figure 5, it is known that object 15 (EMPTLW), object 9 (OI), and object 17 (MURATA) are in one cluster, as shown in the graph in Figure 6.

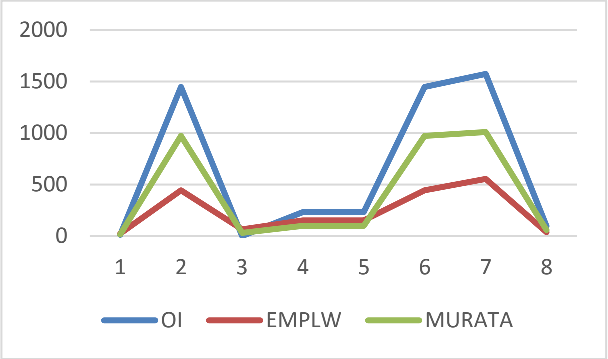


Figure 6. Comparison between Regency of Empat Lawang, Regency of Ogan Ilir, and Regency of North Musi Rawas

The average linkages method uses the average distance to classify each object that has similarities using Equation (5) The dendrogram resulting from the clustering process on the COVID-19 case data using the average linkages method is shown in Figure 7.

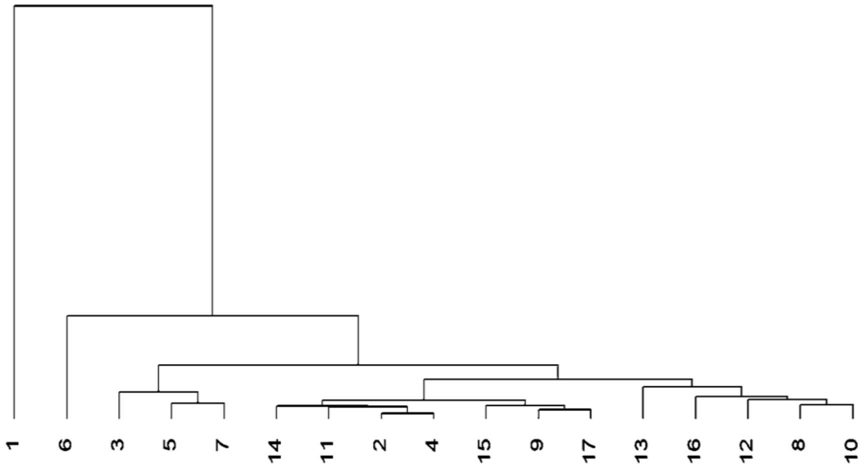


Figure 7. Dendrogram of COVID-19 cases (Average Linkages)

Based on the dendrogram in Figure 7, it is known that object 13 (OKUT), object 16 (PALI), object 12 (LLG), object 8 (BNS), and object 10 (PRABU) are in one cluster, meaning that these objects also have a great resemblance shown by the graph in Figure 8.

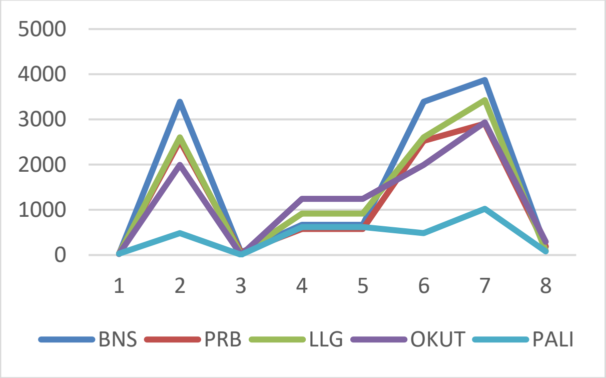


Figure 8. Comparison between Regency of East Ogan Komering Ulu, Regency of Penukal Abab Lematang Ilir, Regency of Lubuk Linggau , Regency of Banyuasin, and Regency of Prabumulih

Next is to determine the optimal number of clusters that can be formed. The optimal number of clusters determined using the R 4.2.1 application with the Silhouette method is obtained, as shown in Figure 9.

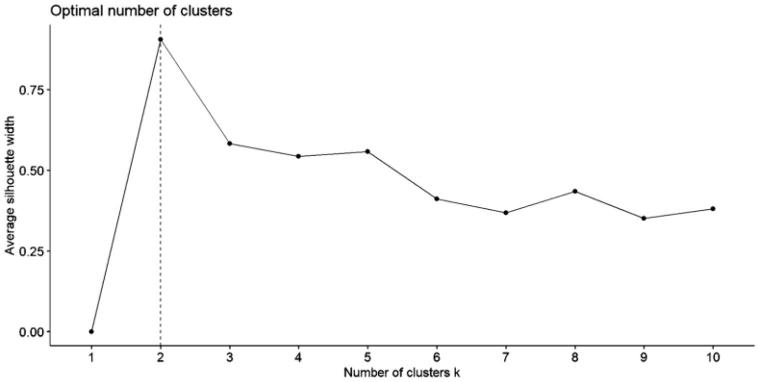


Figure 9. Optimal Number of Clusters

In Figure 9, it is known that the optimal number of clusters for the Regency/City clustering process in South Sumatra based on the COVID-19 case is two clusters with members in cluster one consisting only of Palembang City and cluster two consisting of Pagar Alam City, Lubuk Linggau City, Prabumulih City, Ogan Ilir Regency, Ogan Komeril Ilir Regency, East Ogan Komering Ulu, South Ogan Komering Ulu, Ogan Komering Ulu, Lahat Regency, Musi Banyuasin Regency, North Musi Rawas Regency, Musi Rawas Regency, Pali Regency, Empat Lawang Regency, Regency Banyuasin, and Muara Enim Regency.

4. CONCLUSIONS

Based on the processing of COVID-19 case data using cluster analysis in the R 4.2.1 application, it is known that the optimal number of clusters that can be formed in the clustering process of COVID-19 case data consists of two clusters. This means there are two groups of regencies/cities in South Sumatra based on the COVID-19 case data. The first group only consists of Palembang City. In contrast, in the second cluster, there are Pagar Alam City, Lubuk Linggau City, Prabumulih City, Ogan Ilir Regency, Ogan Komeril Ilir Regency, East Ogan Komering Ulu, South Ogan Komering Ulu, Ogan Komering Ulu, Lahat Regency, Musi Banyuasin Regency, North Musi Rawas Regency, Musi Rawas Regency, Pali Regency, Empat Lawang Regency, Banyuasin Regency and Muara Enim Regency. The results of the similarity measurement show that Object 2 and 4 have the closest distance, namely 0.122, so the clustering process in the COVID-19 case data starts from Object 2 and 4.

REFERENCES

[1] J. B. Soriano, S. Murthy, J. C. Marshall, P. Relan, and J. V. Diaz, “A clinical case definition of post-COVID-19 condition by a Delphi consensus,” *Lancet Infect. Dis.*, vol. 22, no. 4, pp. e102–e107, 2022, doi: 10.1016/S1473-3099(21)00703-9.

[2] D. Handayani, F. Isbaniah, E. Burhan, and H. Agustin, “Penyakit Virus Corona 2019,” *REPIROLOGI Indones.*, vol. 40, no. 2, 2020.

[3] M. I. Rizki, T. A. Taqqiyuddin, and J. J. Cerelia, “K-Medoids Clustering dengan Jarak Dynamic Time Warping dalam Mengelompokkan Provinsi di Indonesia Berdasarkan Kasus Aktif COVID-19,” *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 685–

692, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/.

[4] A. Abunayan, B. Aljadaan, M. Almudayfir, S. Alshareef, and A. alamer, "The effect of COVID-19 on orthopedic elective/emergency procedures in a tertiary hospital Riyadh Saudi Arabia. A cross-sectional study," *Ann. Med. Surg.*, vol. 81, no. July, p. 104331, 2022, doi: 10.1016/j.amsu.2022.104331.

[5] P. Flores-Pérez, N. Gerig, M. I. Cabrera-López, J. L. de Unzueta-Roch, T. del Rosal, and C. Calvo, "Acute bronchiolitis during the COVID-19 pandemic," *Enfermedades Infecc. y Microbiol. Clin. (English ed.)*, vol. 40, no. 10, pp. 572–575, 2022, doi: 10.1016/j.eimce.2021.06.005.

[6] BPS Sumatra-Selatan, *Provinsi Sumatra Selatan Dalam Angka 2022*. 2022.

[7] Corona.sumselprov.go.id, "Sumatra Selatan Tanggap COVID-19."http://corona.sumselprov.go.id/index.php?module= home&id=1 (accessed Aug. 30, 2022).

[8] D. N. P. Sari and Y. Sukestiyarno, "Analisis Cluster dengan Metode K-Means pada Persebaran Kasus COVID-19 Berdasarkan Provinsi di Indonesia," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, pp. 602–610, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/.

[9] A. Sucipto, "Klasterisasi Calon Mahasiswa Baru Menggunakan Algoritma K-Means," *J. Sci. Tech*, vol. 5, no. 2, pp. 50–56, 2019.

[10] N. Ulinnuh and R. Veriani, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage , Average Linkage dan Ward," *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 5, no. 1, pp. 101–108, 2020.

[11] H. Fransiska, "Clustering Provinces in Indonesia Based on Daily COVID-19 Cases Clustering Provinces in Indonesia Based on Daily COVID-19 Cases," *J. Phys. Conf. Ser.*, no. March, 2021, doi: 10.1088/1742-6596/1863/1/012015.

[12] dinkes.sumselprov.go.id, "13 Terkonfirmasi Covid19 Sumsel 28/9/2022 di Sumatra Selatan." https://dinkes.sumselprov.go.id/2022/09/kata-kemkes-ada-enam-aktivitas-fisik-yang-pas-untuk-dilakukan-13-terkonfirmasi-covid19-sumsel-28-9-2022/ (accessed Sep. 30, 2022).

[13] Basri and Syarli, "Ahp-Standar Score: Pendekatan Baru Dalam Sistem Pemeringkatan," *J. Keteknikan dan Sains – LPPM UNHAS*, vol. 1, no. 1, pp. 1–6, 2018.

[14] A. Aditya *et al.*, "Perbandingan pengukuran jarak Euclidean dan Gower pada klaster k-medoids," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, pp. 1–7, 2021, doi: 10.14710/jtsiskom.2021.13747.

[15] A. Inguanzo *et al.*, "Hierarchical cluster analysis of multimodal imaging data identifies brain atrophy and cognitive patterns in Parkinson's disease," *Park. Relat. Disord.*, vol. 82, pp. 16–23, 2021, doi: 10.1016/j.parkreldis.2020.11.010.

[16] A. Caggiano, F. Napolitano, and R. Teti, "Hierarchical cluster analysis for pattern recognition of process conditions in die sinking EDM process monitoring," *Procedia CIRP*, vol. 99, pp. 514–519, 2021, doi: 10.1016/j.procir.2021.03.071.

[17] L. P. W. Adnyani and P. R. Sihombing, "Analisis Cluster Time Series Dalam Pengelompokan Provinsi Di Indonesia Berdasarkan Nilai PDRB," *Lppm Bina Bangsa*, vol. 1, no. 1, pp. 47–54, 2021, doi: 10.46306/bay.v1i1.5.