ICIC Express Letters
Part B: Applications
Volume **13**, Number **4**, April **2022**

 ISSN 2185-2766
pp. **389–396**

COVID-19 CLUSTERING BY PROVINCE: A CASE STUDY OF COVID-19 CASES IN INDONESIA

Ferry Vincenttius Ferdinand[1,*], Johan Sebastian[1], Christopher Nata[2]
Friska Natalia[3] and Stevanus Adiwena[1]

[1]Department of Mathematics
[2]Department of Industrial Engineering
Pelita Harapan University
Jl. M. H. Thamrin Boulevard 1100 Lippo Village, Tangerang, Banten 15811, Indonesia
johansedbert@gmail.com; { Christopher.nata; stevanus.adiwena }@uph.edu
*Corresponding author: ferry.vincenttius@uph.edu

[3]Department of Information Systems
Universitas Multimedia Nusantara
Scientia Boulevard, Gading Serpong, Tangerang, Banten 15811, Indonesia
Friska.natalia@umn.ac.id

Received August 2021; accepted October 2021

Abstract. *The 2019 novel coronavirus disease (COVID-19) pandemic in Indonesia has caused issues in many sectors such as health, economy, and education. Several actions had been taken by the government to prevent and forestall the spread of the coronavirus infection. However, right now there are still many new cases emerging especially in cities with dense population. In the meantime, actions taken from the government are based on the classification of the severity of new cases; there are red zone, yellow zone and green zone. Therefore, mapping cities into zone is critical because it concerns the right decision to be implemented. This paper aimed to cluster the severity of each province in Indonesia based on the number of cases, recovered, and casualties using 3 clustering methods namely K-Means, K-Medoids, and Gaussian mixture model. The result shows that the most optimal clustering method is the Gaussian mixture model, while the least optimal method for clustering is the K-Means. Furthermore, it is also discovered that the cluster always changes overtime, and the cluster can shift depending on the corresponding parameter.*

Keywords: Indonesia, COVID-19, Clustering analysis, K-Means, K-Medoids, Gaussian mixture model

1. **Introduction.** The 2019 novel coronavirus disease (COVID-19) pandemic has caused various issues in different fields. These fields include health, economy, education, and others, particularly in Indonesia. The government has carried out numerous endeavors to prevent and forestall the spread of the coronavirus infection which is the mastermind behind the spread of COVID-19. However, this effort has not been effective since there are yet numerous new cases emerging, especially in huge metropolitan areas such as Jakarta, Surabaya, Manado, and Bali.

The cities with a high increase in cases are grouped into regions which are then classified into the red zone. The higher increase in COVID-19 infection cases, the higher the number of red zones. On the other hand, other areas including East Nusa Tenggara, Papua, and some parts of Sulawesi have a low rate of COVID-19 infection. These areas are included in the green zone. In addition, the *Gugus Tugas Percepatan Penanganan COVID-19* (Task Force for the Acceleration of Handling COVID-19) also classified areas with COVID-19 infection into the yellow zone which indicates areas with some local virus transmission,

DOI: 10.24507/icicelb.13.04.389

and the orange zone which indicates areas close to the red zone or with small clusters of spread. Furthermore, specific protocols are implemented to reduce the spread of viruses according to the zones. This is done to ensure that the measures taken are appropriate and effective for each specific zone. Consequently, mapping cities into zones becomes an important issue in the decision to implement appropriate policies. Therefore, this study aims to group regions based on the level of infection with respect to the number of positive cases. By observing the group of regions having above the average infection rates, it is hoped that appropriate and proper steps can be taken to prevent or reduce the infection rate of COVID-19.

According to Setiati and Azwar [1], Indonesia can improve government intervention by making policies towards the pandemic. Those policies include warning enforcement to reduce activities outside homes, lockdown implementation to reduce the virus spread, and health services improvement. However, lockdown and the reduction of activities outside homes can lead to economic disruption and instability. Therefore, the government must take measures to avoid such effects. Following Qodir, Effendi, Jubba, Nurmandi, and Hidayati [2], if an area is with an increasing positivity rate, it is hoped that the local government in the group with the area can be alert and take the necessary steps immediately. This can reduce the adverse economic impact and reduce the spread of COVID-19.

This paper begins by discussing the COVID-19 pandemic situation in Indonesia and clustering done by previous researchers. It then describes the 3 clustering methods used in this study which include K-Means, K-Medoids, and GMM. The study further explores and compares the clusters resulted by each method to assess the respective method's performance.

2. **Problem Statement and Preliminaries.** September 15, 2020, marks the 225,000th COVID-19 positive case with 161,000 patients recovered, 8,965 deaths, and 3,507 daily new cases. It should also be noted that the number of daily new cases in Indonesia is much higher compared to other Southeast Asian countries, such as Malaysia and Singapore with 23 and 34 daily new cases consecutively [3,4]. In a simulation conducted by Aldila et al. [5], it was indicated that the implementation of social distancing, in the long run, could significantly reduce the COVID-19 infection rate in Jakarta, Indonesia. However, not doing so may potentially increase the number of positive cases. In China, Cai et al. [6] made a categorization of COVID-19 cases based on a shopping center in Wenzhou. In contrast to the previous findings, Cai et al. concluded that even without close contact for a long period of time, virus transmission intensity in the area remains high. It implies that the coronavirus spreads through indirect transmission.

There are a lot of clustering researches such as Friska et al. comparing a few methods in Indonesia [7]. Specifically, in COVID-19 cases, there are a lot of researches using K-Means and K-Medoids. In 2020, Virgantari implemented K-Means clustering using provincial daily COVID-19 cases [8]. In April 2021, Utomo analyzed provincial cases on January 2021 in Indonesia and implemented K-Means and K-Medoids algorithms. The two clustering methods were then compared using the Davies-Boulden index which resulted in K-Means method being a better algorithm in clustering the spread of COVID-19 in Indonesia [9]. K-Means is also used by Mahmudan for analyzing COVID-19 in smaller scope: district and/or city in Central Java [10] and Abdulla et al. for analyzing provincial results [11]. While an unpopular method such as the Gaussian mixture model could be applied to analyzing COVID-19 numbers in Indonesia, there are yet recognized results on the implementation of the model.

Therefore, clustering methods applied in this study will include K-Means clustering, K-Medoids, and Gaussian Mixture Model (GMM). However, since the data are collected daily, they are classified as time-series data. In clustering time-series data, the number

of features is equal to the number of the data itself. Thus, the process of clustering will be relatively costly and time-consuming. Other than that, the Euclidean method that is generally used in determining the distance does not account the delay that might occur in time-series data. Therefore, before using the clustering methods, the data will be transformed using Fast Fourier Transform (FFT) to obtain and reduce the frequency dimension [12]. The data consist of n days of observation. Then, each data will be transformed into complex number which contains real and imaginary part which will produce new vector with $2n$ dimension. After being transformed, the dimension will be reduced using Nyquist-Shannon sampling theorem [13].

K-Means clustering will be applied to making each data sample fall into K-subgroups without overlapping. The method uses an iterative algorithm to partition the data. The data are then divided into different subgroups (clusters) based on the Euclidean distance of the centroid and data point. If the distance between two data points to a centroid is relatively small, then the data points can be considered similar. Thus, a point Xi can be assigned into a cluster Ci [14]. In effect, each data in one group has similarities with the others and maximizes differences between groups, according to Bradley et al., 1998 in Sonbhadra et al. [15]. In addition, the K-Means method has the advantage of efficiency in grouping a large amount of data [16]. Despite the practical use of K-Means, it is still sensitive to outliers. Supposing it is the case that some data have an extreme value, it may lead to distortion in the data distribution. K-Means uses the mean value of the distribution as the centroid, while K-Medoids uses a medoid, which is the most centrally located object in the cluster. The main difference between K-Means and K-Medoids is the reference point [14]. K-Means uses the mean value as the centroid, while K-Medoids uses representative object defined as the medoid as the reference point. Furthermore, the Gaussian mixture model is the sum of several weighted Gaussian distributions, a basic probability distribution [17]. Before the implementation of clustering, cluster analysis must be conducted by the researcher beforehand. Cluster analysis is a technique used to group objects based on their characteristics and determine the number of clusters. There are several methods to determine the number of clusters. In this study, the K-Means method will use the Calinski-Harabasz (CH) index as it is the best method [18]. The K-Medoids method will implement the gap-statistic method. On the other hand, for the GMM method, the BIC criterion will be used. The evaluation of the clustering methods can be done by using the within-cluster variance and between-cluster variance [19]. Smaller within-cluster variance and greater between-cluster variance indicate the better method.

3. **Main Results and Analysis.** The main result for this paper is the cluster of provinces. After transforming data using Fast Fourier Transform, we examine the data using 3 different methods for 8 different time periods, which include March 2020, March-April 2020, March-May 2020, March-June 2020, March-July 2020, March-August 2020, March-September 2020, March-October 2020. For each period, we compare each method using 3 parameters: cases, recoveries, and casualties. Figures 1-6 show clusters that are obtained using the combination of the 3 parameters. Furthermore, Table 1 and Table 2 are the examples of the data collected for each period for each method. Figure 1 represents clusters made by K-Means method for the period March 2020 that results in six clusters. The result of K-Medoids method is the same as K-Means method as shown in Figure 3. Although Figure 5 also results in six clusters, the clusters made by K-Medoids method and GMM methods are different. For March-October 2020 period, K-Means and K-Medoids produced the same cluster, while GMM method produced a different result. For further analysis of each figure, Table 1 and Table 2 show the cluster's mean which is used to infer the corresponding cluster. As an example, for March 2020 period, it is shown from Table 1 that there are 6 clusters made from K-Means, K-Medoids, and GMM. As an example,

for March period, Cluster 1 has the highest mean of cases, recoveries, and casualties. This cluster consists of only 1 province, which is Jakarta, the capital of Indonesia where the first case was found. On the other hand, the least mean of cases and casualties are from Cluster 6 but the recoveries are varying between Cluster 3, Cluster 5, and Cluster 6. However, Cluster 3 of GMM method and Cluster 5 of K-Means and K-Medoids methods consist of the same province which is South Sulawesi. Based on Figure 1, Figure 3, and Figure 5, it could be analyzed that those clusters consist of provinces with fewer health facilities than the others and therefore the recoveries number is less. While for March-October 2020 period, Table 2 represents Figure 2, Figure 4, and Figure 6. Based on K-Medoids and K-Means methods, only two clusters are made. The first cluster consists of Jakarta and East Java, which includes Surabaya, the second biggest city in Indonesia. The second cluster consists of the rest of the other provinces. On the other hand, for the GMM method, five clusters are made as shown in Figure 6. For all methods, the first cluster which consists of Jakarta, the capital of Indonesia, still has the highest mean for all cases, casualties, and recoveries. Further comparison of all 3 clustering methods for each period, the ratios of within-between cluster variance are used and calculated which are shown in Table 3. The lower value of this ratio indicates the better method.

From Table 3, if we look at cases parameter, we get the average $\frac{Wvar}{Bvar}$ ratios for K-Means, K-Medoids, and GMM methods are 0.074748, 0.078509, and 0.057935 respectively.

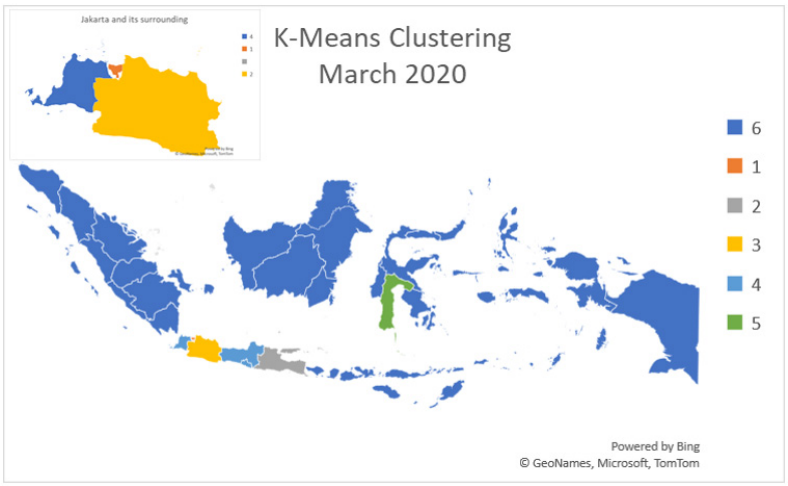

FIGURE 1. (color online) Clustering by K-Means method for March 2020 period

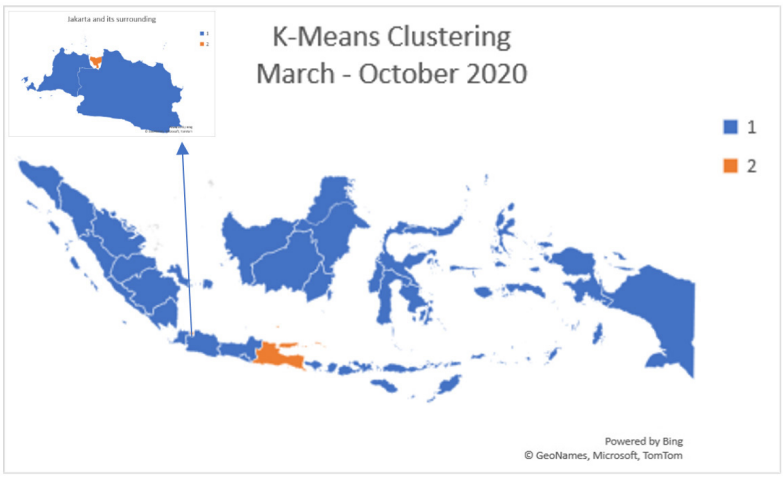

FIGURE 2. (color online) Clustering by K-Means method for March-October 2020 period

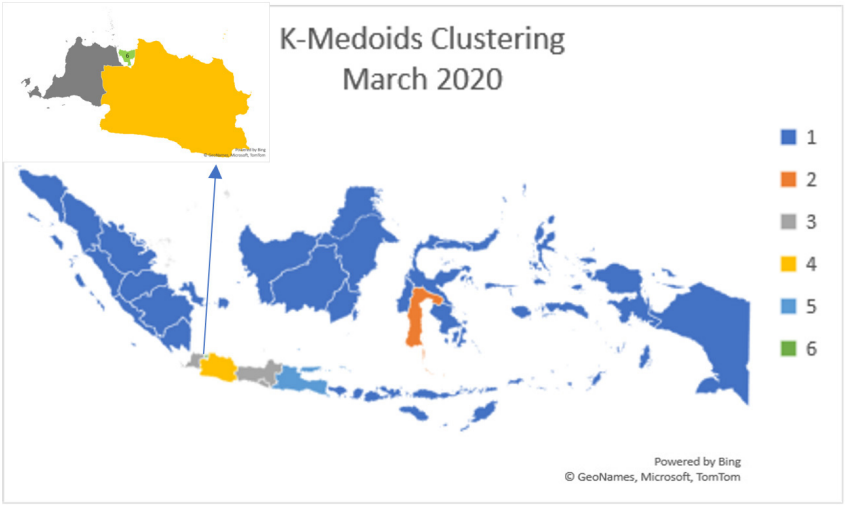

FIGURE 3. (color online) Clustering by K-Medoids method for March 2020 period

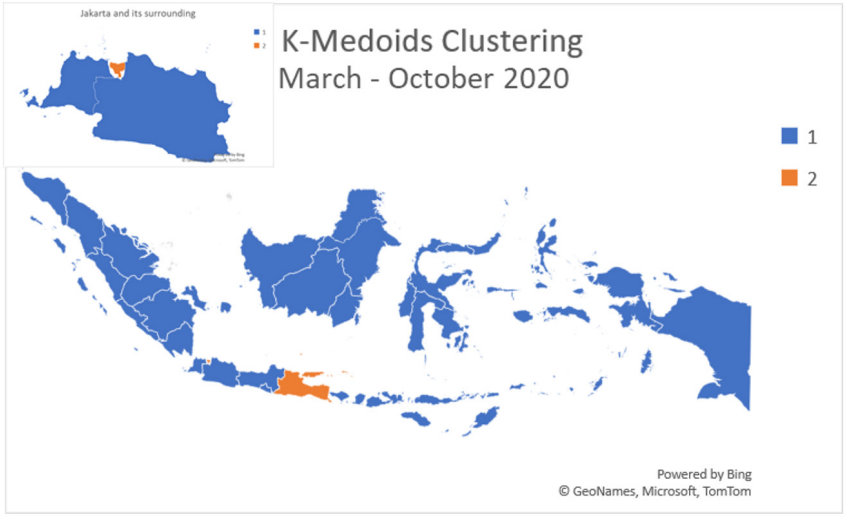

FIGURE 4. (color online) Clustering by K-Medoids method for March-October 2020 period

FIGURE 5. (color online) Clustering by Gaussian mixture modelling method for March 2020 period

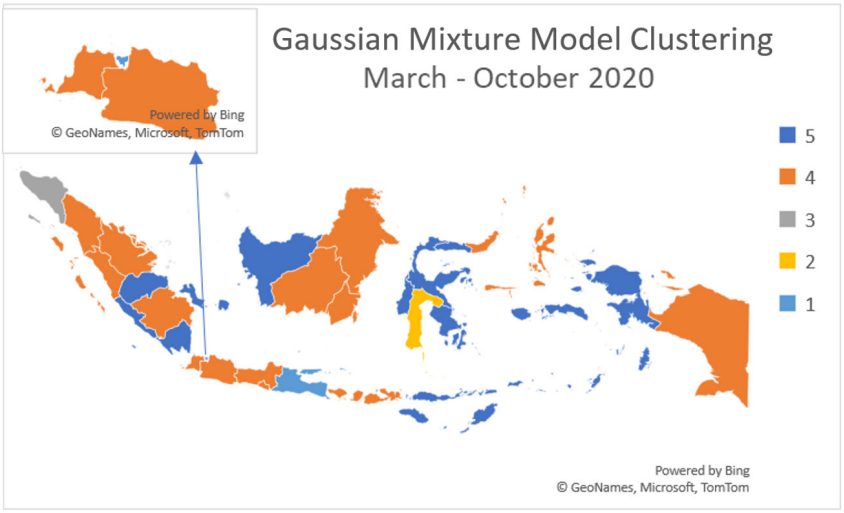

FIGURE 6. (color online) Clustering by Gaussian mixture modelling method for March-October 2020 period

TABLE 1. Cluster recapitulation period March 2020

March										
		Cases per Day			Recoveries Recovered per Day			Casualties per Day		
		K-Means	K-Medoids	GMM (Gaussian Mixture Model)	K-Means	K-Medoids	GMM (Gaussian Mixture Model)	K-Means	K-Medoids	GMM (Gaussian Mixture Model)
Mean	Cluster 1	47.64	47.64	47.64	3.09	3.09	3.09	5.91	5.91	5.91
	Cluster 2	7.09	7.09	7.09	1.45	1.45	1.45	0.64	0.64	0.64
	Cluster 3	13.55	13.55	4.55	0.91	0.91	0.00	1.27	1.27	0.18
	Cluster 4	5.94	5.94	8.18	0.06	0.06	0.09	0.27	0.27	0.27
	Cluster 5	4.55	4.55	3.26	0.00	0.00	0.16	0.18	0.18	0.30
	Cluster 6	0.41	0.41	0.38	0.01	0.01	0.00	0.03	0.03	0.01
Within Variance/ Between Variance		0.004621	0.004621	0.012840	0.000441	0.000411	0.010700	0.000331	0.000331	0.005160

TABLE 2. Cluster recapitulation period March-October 2020

March-October 2020										
		Cases per Day			Recoveries Recovered per Day			Casualties per Day		
		K-Means	K-Medoids	GMM (Gaussian Mixture Model)	K-Means	K-Medoids	GMM (Gaussian Mixture Model)	K-Means	K-Medoids	GMM (Gaussian Mixture Model)
Mean	Cluster 1	350.41	350.41	350.41	307.59	307.59	307.59	13.33	13.33	13.33
	Cluster 2	34.66	34.66	82.07	26.98	26.98	73.10	1.08	1.08	2.08
	Cluster 3			33.06			22.59			1.19
	Cluster 4			53.93			41.65			1.80
	Cluster 5			9.37			7.22			0.19
Within Variance/ Between Variance		0.303887	0.303887	0.199507	0.284043	0.284043	0.185082	0.189698	0.189698	0.128268

Therefore, it can be concluded that the best clustering method for cases parameter is GMM. Then, if we look at recovery's parameter, we get the average $\frac{Wvar}{Bvar}$ ratios for K-Means, K-Medoids, and GMM methods are 0.082934, 0.067251, and 0.119297 respectively. Thus, the best method for clustering recoveries parameter is K-Medoids. Lastly, if we look at casualties, we got the average $\frac{Wvar}{Bvar}$ ratios for K-Means, K-Medoids, and GMM methods are 0.139709, 0.135006, and 0.103945 respectively. Therefore, the best method to cluster casualty's parameter is GMM.

Table 3. Recapitulation clustering methods for each period

Period	K-Means			K-Medoids			Gaussian Mixture Model		
	Wvar/Bvar			Wvar/Bvar			Wvar/Bvar		
	Cases	Recoveries	Casualties	Cases	Recoveries	Casualties	Cases	Recoveries	Casualties
Mar	0.004621	0.000441	0.000331	0.004621	0.000441	0.000331	0.012839	0.010698	0.005516
Mar-Apr	0.000722	0.004038	0.000701	0.008275	0.018525	0.008463	0.007843	0.009880	0.011152
Mar-May	0.006374	0.009478	0.008070	0.037234	0.015711	0.046189	0.070081	0.521280	0.030118
Mar-Jun	0.028631	0.151280	0.087928	0.020307	0.005094	0.004426	0.010348	0.091964	0.051891
Mar-Jul	0.032598	0.033228	0.298276	0.032598	0.033228	0.298276	0.021525	0.025466	0.182626
Mar-Agu	0.041053	0.031690	0.290119	0.041053	0.031690	0.290119	0.021193	0.014986	0.217981
Mar-Sep	0.180094	0.149275	0.242547	0.180094	0.149275	0.242547	0.120141	0.095021	0.204009
Mar-Oct	0.303887	0.284043	0.189698	0.303887	0.284043	0.189698	0.199507	0.185082	0.128268

4. **Conclusions.** From this research, it can be concluded that there is a possibility to use several clustering methods for COVID-19 cases in Indonesia. A data transformation is required before applying the time-series clustering procedure into the data. Based on 8 periods and 3 methods: K-Medoids, K-Means, and GMM, it is concluded that GMM is the best method. On the other hand, K-Means method is the least satisfying technique according to the $\frac{Wvar}{Bvar}$ ratio. Furthermore, as for future works, other methods of time-series clustering can be used for analysis of the development of COVID-19 cases in Indonesia and the results can be compared. Clustering based on regency and city also can be considered for future study.

Acknowledgment. This work is fully supported by Universitas Pelita Harapan. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] S. Setiati and M. K. Azwar, COVID-19 and Indonesia, *Acta Med. Indones. – Indones. J. Intern. Med.*, vol.52, no.1, pp.84-89, 2020.

[2] Z. Qodir, G. N. Effendi, H. Jubba, A. Nurmandi and M. Hidayati, COVID-19 and chaos in Indonesia social-political responsibilities, *International Research Association for Talent Development and Excellence*, vol.12, no.1, pp.4629-4642, 2020.

[3] G. Tugas, *Peta Sebaran*, https://covid19.go.id/peta-sebaran, Accessed on 15-09-2020.

[4] OCHA, *Indonesia: Coronavirus (COVID-19) Subnational Cases*, https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province, Accessed on 30 September 2020.

[5] D. Aldila, S. H. Khoshnaw, E. Safitri, Y. R. Anwar, A. R. Bakry, B. M. Samiadji, D. A. Anugerah, M. F. G. Alfarizi, I. D. Ayulani and S. N. Salim, A mathematical study on the spread of COVID-19 considering social distancing and rapid assessment: The case of Jakarta, Indonesia, *Chaos, Solitons and Fractals*, 2020.

[6] J. Cai, W. Sun, J. Huang, M. Gamber, J. Wu and G. He, Indirect virus transmission in cluster of COVID-19 cases, Wenzhou, China, *Emerging Infectious Diseases*, vol.26, no.6, pp.1343-1345, 2020.

[7] F. Natalia, R. I. Desanti and F. V. Ferdinand, Prediction and visualization of flood occurences in Tangerang using K-Medoids, DBScan, and X-Means clustering algorithms, *2019 5th International Conference on New Media Studies (CONMEDIA)*, 2019.

[8] Virgantari, Fitria and Y. E. Faridhan, K-Means clustering of COVID-19 cases in Indonesia's provinces, *ADRI International Journal of Engineering and Natural Science*, vol.5, no.2, pp.34-39, 2020.

[9] W. Utomo, The comparison of K-Means and K-Medoids algorithms for clustering the spread of the COVID-19 outbreak in Indonesia, *ILKOM Jurnal Ilmiah*, vol.13, no.1, pp.31-35, 2021.

[10] A. Mahmudan, Clustering of district or city in Central Java based COVID-19 case using K-Means clustering, *Jurnal Matematika, Statistika dan Komputasi*, vol.17, no.1, pp.1-13, 2020.

[11] D. Abdullah et al., The application of K-Means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data, *Quality & Quantity*, pp.1-9, 2021.

[12] P. Roelofsen, *Time Series Clustering*, Master Thesis, Vrije Universiteit, Amsterdam, Netherlands, 2018.

[13] T. Strohmer and J. Tanner, Implementations of Shannon's sampling theorem, a time-frequency approach, *Sampling Theory in Signal and Image Processing*, vol.4, no.1, pp.1-17, 2005.
[14] P. Arora, Deepali and S. Varshney, Analysis of K-Means and K-Medoids algorithm for big data, *Procedia Computer Science*, vol.78, pp.507-512, DOI: 10.1016/j.procs.2016.02.095, 2016.
[15] S. K. Sonbhadra, S. Agarwal and P. Nagabhushan, Target specific mining of COVID-19 scholarly articles using one-class, *Chaos, Solitons and Fractals*, 2020.
[16] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Education, Inc., 2007.
[17] C. Fang and A. L. Ralescu, Online Gaussian mixture model for concept modeling and discovery, *International Journal of Intelligent Technolgies and Applied Statistics*, vol.2, no.1, pp.59-75, 2008.
[18] W. G. Milligan and M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, vol.50, no.2, pp.159-179, 1985.
[19] M. J. Bunkers, J. R. Miller Jr. and A. T. DeGaetano, Definition of climate regions in the Northern Plains using an objective cluster modification technique, *Journal of Climate*, vol.9, no.1, pp.130-146, 1966.