Implementation of the trimmed *k*-means clustering method in mapping the distribution of Covid-19 in Indonesia

Cite as: AIP Conference Proceedings **2563**, 050012 (2022); https://doi.org/10.1063/5.0103175
Published Online: 31 October 2022

Netti Herawati, Khoirin Nisa and Subian Saidi

AIP Conference Proceedings **2563**, 050012 (2022); https://doi.org/10.1063/5.0103175

Implementation of the Trimmed *k*-Means Clustering Method in Mapping the Distribution of Covid-19 in Indonesia

Netti Herawati[a], Khoirin Nisa[b] and Subian Saidi[c]

Department of Mathematics, University of Lampung, Bandar Lampung,
Jl. Sumantri Brojonegoro No.1, Rajabasa, Bandar Lampung, Lampung 35141, Indonesia

[a] Corresponding author: netti.herawati@fmipa.unila.ac.id
[b] Subian.saidi@fmipa.unila.ac.id
[c] khoirin.nisa@fmipa.unila.ac.id

Abstract. The Coronavirus that appeared in (COVID-19), caused by SARSCoV-2, started at Wuhan in the Hubei province of China and has spread with great speed around the world; it has caused a severe health crisis all around the world, including Indonesia. This study aims to use a clustering technique to assess the risk of the COVID-19 pandemic in Indonesia, based on data obtained between March 2020 and July 2021 in that country (http://www.covid19.go.id). Provinces in Indonesia were grouped based on COVID-19 infection rates and mortality data. Since the data con-tained some outliers, i.e. provinces with a very high number of cases, we used a robust clustering method; this method is sensitive to outliers. The analysis was performed using the Trimmed k-means clustering method. Based on the results of this study, with four provinces detected as outliers in the data, there were three optimal clusters with the maximum separation index. Cluster 1 consisted of 14 provinces, and clusters 2 and 3 consisted of 10 and 6 provinces, respectively. The four outliers, i.e. Jakarta, West Java, Central Java and East Java, formed a separate cluster.

INTRODUCTION

Indonesia's worldwide Covid 19 is due to intensive acute respiratory syndrome coronavirus 2 (SARSCov2). As of July 23, 2021, Indonesia reported 3,082,410 cases off the coast of the Philippines, the highest in Southeast Asia. Indonesia has killed 80,598 people and is ranked 0.33 in Asia and 15th in the world.

The corona virus pandemic has had a huge impact on Indonesia. Not only has it impacted the health sector, but it also hit other sectors such as the economy, industry, education, social, and tourism [1–3]. There is not a single province in Indonesia that was not affected by the coronavirus on a small or large scale. Various policies have been carried out by the government to anticipate and minimize the im-pact of the pandemic. These include setting up a task force team to accelerate the handling of COVID-19, enforcement a Large-Scale Social Restrictions (LSSR) in various regions, purchasing medical equipment, upgrading a number of referral hospitals for Covid-19 patients, distributing social assistance in the form of basic necessities and cash for people in need, exempting from electricity costs for certain groups, and providing tax incentives for the industrial sector.

Indonesia is a country with a very large population. In fact, Indonesia ranks fourth in the world as the country with the largest population after China, India and the United States. As an archipelagic country, Indonesia's population is spread across various provinces in Indonesia. The population in each province is different and continues to change. Large population growth and uneven distribution are a source of problems in Indonesia. This problem also applies when it comes to handling Covid-19 [4]. The spread of Covid-19 in Indonesia was rapid and varied from region to region. Due to the large area and large population in Indonesia, it is necessary to anticipate the transmission of the Covid-19 virus quickly, one of the efforts undertaken was to study the data of the Covid-19 cases in 34 provinces in Indonesia separately.

The 2nd Universitas Lampung International Conference on Science, Technology, and Environment (ULICoSTE) 2021
AIP Conf. Proc. 2563, 050012-1–050012-12; https://doi.org/10.1063/5.0103175
Published by AIP Publishing. 978-0-7354-4237-5/$30.00

Various studies on Covid-19 cases in Indonesia have been carried out by numerous authors to pro-vide good statistical analysis for the government as a reference in policy making. Some of those studies were modelling the data using a generalized linear model approach [5], forecasting using nonparamet-ric analysis [6], and classification using cluster analysis [7–9]. Cluster analysis seems to be the most popular technique for analysing the Covid-19 data. The technique is also predominately used for Covid-19 data analysis in worldwide countries; one can see [10–14].

Cluster analysis can be divided into two types: hierarchical clustering and non-hierarchical clustering. The difference between the two types lies in determining the number of clusters that will be generated. Hierarchical clustering is used when the number of clusters required is unknown, and non-hierarchical clustering is used when the number of clusters required is predetermined. If your data contains outliers, you need a robust way to avoid biased analysis results due to the effects of the outliers. See [15] for more information. One of the robust methods of cluster analysis is the trimmed *k*-means method. The advantage of the trimmed *k*-means method is its resistance to outliers [16,17]. In addition, cluster analysis is slightly different from other multivariate methods. This method is usually required by other multivariate methods and does not start with a particular data distribution. Cluster analysis is based on the distance matrix, which is a measure of similarity, and the most commonly used measure is the Euclidean distance. However, there is one assumption that must be made when using Euclidean distance for cluster analysis. This means that all variables are uncorrelated and this assumption is often ignored, leading to suboptimal clustering results. Principal component analysis (PCA) can be performed if the variables in the data are correlated. A cluster analysis is then performed based on the principal component evaluation of the observed values. Readers are encouraged to investigate the details of clustering analysis by PCA [18–20].

In this study, we conducted a cluster analysis of the Covid-19 data based on the number of con-firmed cases, death cases, recovered cases and mortality to incidence ratio (MIR) from 34 provinces in Indonesia using the robust trimmed k-means method and principal component scores to obtain a cluster map that was more optimal and efficient than the k-means clustering. The results of the analysis could be used as an overview of the severity and level of risk of the spread of Covid-19 in the provinces in Indone-sia using the clusters obtained.

MATERIALS AND METHODS

The data used in this study was collected from March 20, 2020 to July 23, 2021 by the Covid19 Accelerated Task Force (Indonesian SATGAS) (https://covid19.go.id /). Published Indonesian Covid19 data. Three variables were retrieved from the internet. NS. Number of confirmed Covid 19 recovery or deaths. Covid19's MIR was recorded to indicate the severity of the disease. MIR as a substitute for survival was calculated by dividing the number of deaths in each state by the cases identified in the same state and multiplying by 100. The range of MIR ranges from 0 to 100%, with 100% indicating the worst case in which all confirmed cases have died [10].

We present the summary of Covid-19 data from 34 provinces of Indonesia in Table 1. There are three variables in the data that have significantly different mean and median values, i.e. confirmed, death and recovered cases; this shows that the three variables have an asymmetric or skewed distribution. The asymmetry of the data can be caused by the natural structure of the variable or it can also be caused by the existence of outliers. The MIR of the 34 provinces in in Indonesia on 23 July 2021 was 2.531%, this value decreased compared to the previous year but it is known to be the highest among ASEAN countries.

TABLE 1. Summary of COVID-19 cases of 34 Province in Indonesia

Descriptive Statistics	Confirmed Cases	Death Cases	Recovered Cases	MIR
Minimum	7103	149	5930	0.897
1st Quartile	17790	314	13393	1.710
Median	31069	775	22580	2.250
Mean	90655	2363	71084	2.531
3rd Quartile	73331	1706	57341	2.766
Maximum	778521	17512	678992	6.568
Standard deviation	164523.3	4292.167	136554.595	1.180

Clustering is a variety of techniques for finding a subset of observations in a dataset. When grouping observations, the observations in the same group should be similar and the observations in different groups should be dissimilar. Clustering allows you to identify which observations are similar and, in some cases, classify them within. The kmeans algorithm is the simplest and most widely used clustering technique for dividing a dataset into sets of k groups, including clustering analysis of Covid19 data [7,11,21–23]. This algorithm aims to minimize the collective square

Euclidean distance between the observations and the centroid of the cluster to which they belong, but the kmeans algorithm does not give the best results. Sensitive to outliers (that is, points that are different from other data points). The outliers in the data will cause the results of cluster analysis to be inefficient. In this cases, a more robust method is needed and is described below.

Trimmed *k*-means was introduced by [16]. Using this method, one is allowed to remove a certain proportion of the possible outliers when grouping the results [24,25]; the proportion to be removed is the magnitude of the data rate to be trimmed in this method. The trimmed *k*-means method is used to overcome the outliers contained in a cluster of data to be grouped. The main concept of the trimmed kmeans method is to create a new k-cluster by removing or trimming outliers in the data. This method belongs to the classic hierarchical clustering technique, a class of methods based on "fair trimming" (determined by the data), which aims to make kmeans more robust. In addition, the generalized k-mean method uses a penalty function Φ to minimize the discrepancy between random variables (or samples of those random variables) and the set of measured k points. It is based on [16].

Let α be contained in an open unit interval, i.e. $\alpha \in (0,1)$, k a natural number, and Φ a penalty function. For every set *A* such that $P(A) \geq 1\text{-}\alpha$ and every k-set $M = \{ m_1, m_2, \ldots, m_k \}$ in R^p, let us consider the variation about *M* given *A:*

$$V_{\Phi}^{A}(M) = \frac{1}{P(A)} \int_{A}^{\Phi} \Phi\left(\inf_{i=1,\ldots,k} \|X - m_i\| \right) dP \tag{1}$$

The variation $V_{\Phi}^{A}(M)$ measures how well the set M represents the probability mass of P defined on *A,* and our job is to choose a set of given probability mass such that the variation is minimized. This is done by minimizing $V_{\Phi}^{A}(M)$ on *A* and *M* in the following way:

1. obtain the *k*-variation given A, $V_{\mathrm{k},\Phi}^{A}(M)$, by minimizing over *M*:

$$V_{\mathrm{k},\Phi}^{A}(M) = \inf_{M \subset R^p, \# M \neq k} \ V_{\Phi}^{A}(M) \tag{2}$$

2. obtain the trimmed *k*-variation, $V_{k,\Phi,\alpha}$, by minimizing over *A*:

$$V_{k,\Phi,\alpha} := V_{k,\Phi,\alpha}(X) := V_{k,\Phi,\alpha}(P_X) := \inf_{A \in \beta^p, P(A) \geq 1-\alpha} \ V_{\mathrm{k},\Phi}^{A} \tag{3}$$

We wish to obtain a trimmed set A_0 if it exists, and a k-set $M_0 = \{m_1^0, m_2^0, \ldots m_k^0\}$, if it exists, using the condition $V_{\Phi}^{A_0}(M_0) = V_{k,\Phi,\alpha}$ [16].

Principal Component Analysis (PCA) is a statistical multivariate technique that aims to reduce the dimensions of the data to obtain new uncorrelated variables and retain most of the information contained in the original variables. The resulting variable is a linear combination of the original variables and is called the principal component (PC). The sum of the squares of the coefficients in the linear combination is equal to unity, and the PCs are orthogonal.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ be a random vector of a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)$ and covariance matrix $\boldsymbol{\Sigma}$ and k independent eigenvectors, $\boldsymbol{a}_k$, $k = 1,2, \ldots, k$. The principal component then can be written as follows:

$$Y = A(X - \mu) \tag{4}$$

where **A** is a *p* by *p* coefficient matrix that carries the *p*-element variable $\boldsymbol{X}$ into the *n*-element variable y. The column vectors in **A** are eigenvectors $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \ldots \mid \boldsymbol{a}_k]$. The i-th principal component can be written as:

$$y_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ki}X_k = \boldsymbol{a}_i^T \boldsymbol{X} \tag{5}$$

The covariance matrix is very sensitive to the presence of outliers, which causes PCA problems **if** the data contains outliers. To overcome this, robust principal component analysis is required. That is, we need to replace the classical sample covariance matrix S with a robust estimator. Using robust covariance matrix estimation to calculate principal components is equivalent to performing a robust PCA [26]. One of the robust estimators **of** the covariance matrix is the minimum covariance determinant [26]. One of the robust estimators **of** the covariance matrix is the minimum covariance determinant (MCD). The MCD estimator is a pair ($\overline{\boldsymbol{X}}_{MCD}, \boldsymbol{S}_{MCD}$), where $\overline{\boldsymbol{X}}_{MCD}$ is the mean vector and SMCD is the covariance matrix that minimizes the determinant of the sample covariance matrix S in the subsample containing exactly *h* members of *n* observations [27,28].

It is often necessary to validate the number of clusters obtained from the partitioning by using the cluster validity index. The validity index is a method for evaluating the results of the clustering algorithm in order to get the best number of clusters [29]. The validity index is calculated based on the following two criteria:

1. Compactness is the level of similarity of objects in the same cluster, and
2. Separation is the level of difference between objects in different clusters.

The silhouette analysis measures how well an observation is clustered, and it estimates the average distance between clusters. It contains the average value of each point in the data set, more specifically, the calculation of the value of each point is the difference between the values of separation and compactness, divided by the maximum between the two. The best number of clusters is indicated by the silhouette value which is as close to 1 as possible. Suppose there are N points in a data set. There exist clusters, p and q, whereby x_i is a point in cluster p and y_j is a point in cluster q, such that $a_{p,i}$ is the average distance of point x_i to each point in cluster p, and $d_{q,i}$ is the average distance of point x_i to every point in cluster q. For each observation, i, the silhouette width s_{x_i} is calculated as follows:

1. Calculate the average dissimilarity $a_{p,i}$ between i and all other points of the cluster to which i belongs, as follows:
$$a_{p,i} = \frac{1}{n_p}\sum_{k=1}^{n_p} d(x_i, x_k) \tag{6}$$
2. For all other clusters q, to which i does not belong, calculate the average dissimilarity $d_{q,i}$ of i to all observations of q, i.e. $d_{q,i} = \frac{1}{n_q}\sum_{j=1}^{n_q} d(x_i, y_j)$. The smallest of these $d_{q,i}$ is defined as $b_{q,i} = min\, d_{q,i}\ , q = 1, \dots, k$. The value of $b_{q,i}$ can be understood as the dissimilarity between i and its "neighbor" cluster, i.e., the nearest one to which it does not belong.
3. Finally, the silhouette width of the observation i, is defined by the formula:
$$s_{x_i} = \frac{(b_{q,i} - a_{p,i})}{max\{a_{p,i} - b_{q,i}\}}, p \neq q \tag{7}$$

The silhouette width can be interpreted as follows:
s_{x_i}>0 , means that the observation is well grouped.
s_{x_i} <0, means that the observation has been placed in the wrong cluster.
s_{x_i} =0 , means that the observation is between two clusters.
The average silhouette width is then given by:
$$SIL = \frac{1}{N}\sum_{i=0}^{N} s_{x_i} \tag{8}$$

Observations with a large s_(x_i), almost 1, are well clustered. For a maximum value of the average silhouette, SIL, the optimal clustering is obtained [30].

RESULTS

We performed a robust cluster analysis using the trimmed *k*-means method to the data by firstly addressing correlated variables and outliers using robust principal component analysis. Then, we used the robust principal component score for cluster analysis using the trimmed *k*-means method.

Outliers Detection

In this study, the method used to detect outliers is the Mahalanobis squared distance method where the *i*-th observation is identified as an outlier if $d_{MD}^2(i) > \chi_{p,1-\alpha}^2$. Based on this case, there are 4 variables studied, and we used α = 5%. Therefore, we have $\chi_p^2 = 4.(1 - 0.05) = 9.488$. Outliers are defined by comparing the results of the Mahalanobis squared distance of each object with the value of $\chi_{4,0,95}^2$; as a result, 4 outlier provinces are found, namely DKI Jakarta, West Java, Central Java, and East Java. The quantile-quantile (Q-Q) plot of the chi-square of the data is presented in Fig. 1. The result of the outlier detection is also confirmed by the panel plot of each pair between variables as shown in Fig. 2.

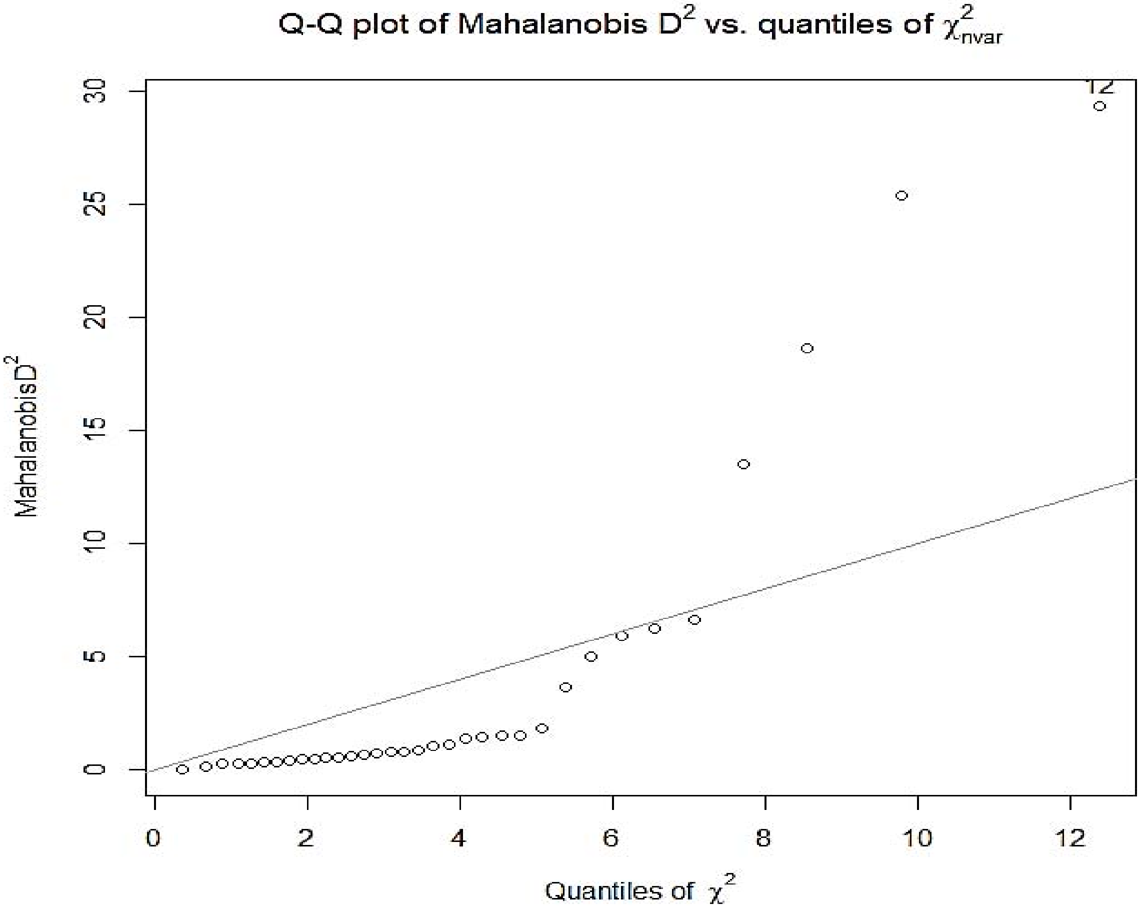

FIGURE 1. Q-Q plot of squared Mahalanobis distance vs. chi-squared quantiles

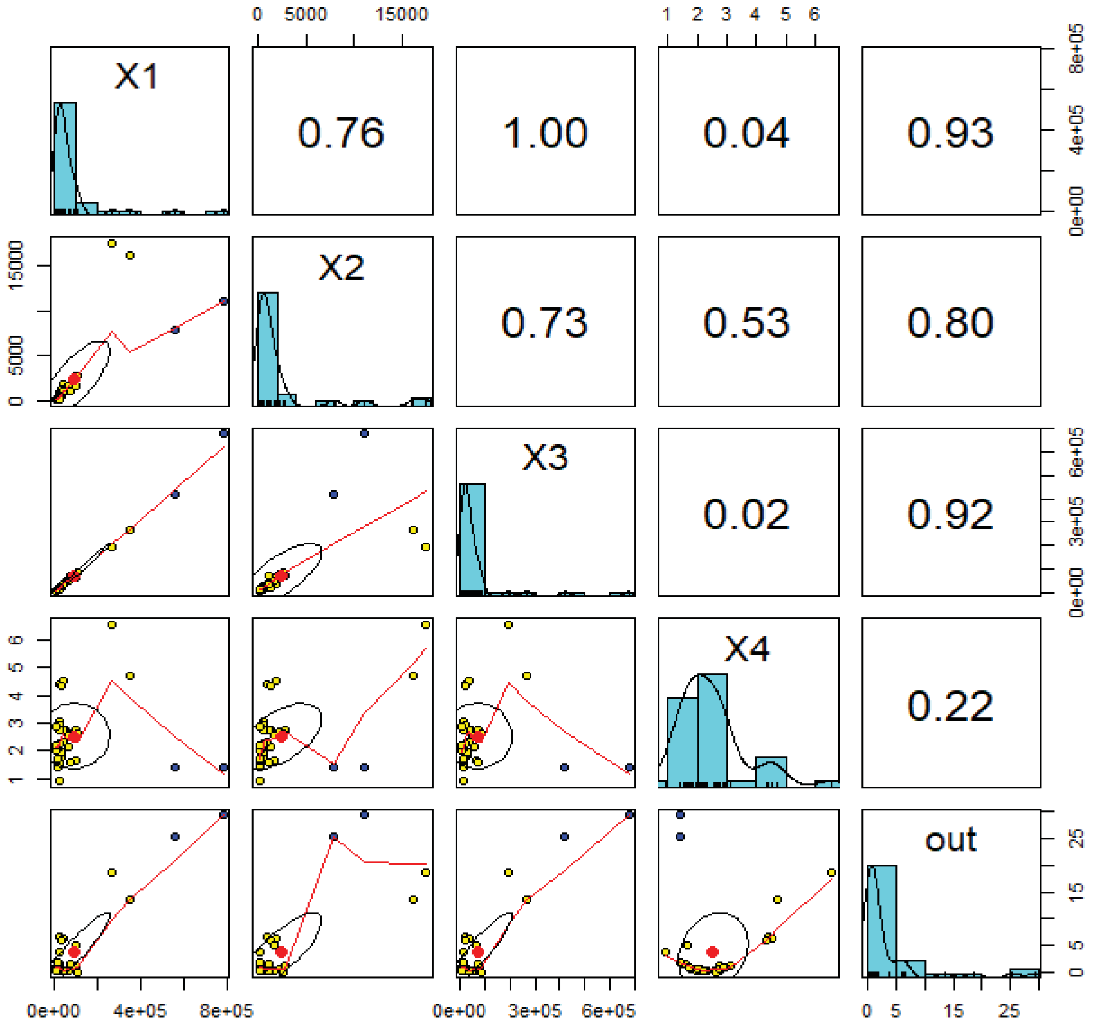

FIGURE 2. Pair panels plot of Covid-19 Data in Indonesia

The four provinces identified as outliers with their attribute presented in Table 2. The data in the ta-ble show that DKI Jakarta had the highest confirmed cases, while East Java had the lowest confirmed cases. However, the MIR of East Java was the highest among the four provinces.

TABLE 2. List of outliers in the Covid-19 Data in Indonesia

Province	Confirmed Cases	Death Cases	Recovered Cases	MIR
DKI Jakarta	778521	11131	678992	1.430
West Java	556181	7917	421977	1.423
Central Java	343210	16195	267511	4.719
East Java	266638	17512	194233	6.568

The correlation coefficients between variables in the data are shown in Table 3. It can be seen that there were correlations between variables, with the maximum value is the correlation between the con-firmed cases and the recovered cases, i.e. 0.997. Some other correlation coefficients were moderate, with the values between 0.53 and 0.76.

TABLE 3. Correlation matrix between variables

	Confirmed Cases	Death Cases	Recovered Cases	MIR
Confirmed Cases	1.000	0.758	0.997	0.036
Death Cases	0.758	1.000	0.732	0.533
Recovered Cases	0.997	0.732	1.000	0.015
MIR	0.015	0.533	0.015	1.000

Robust Cluster Analysis using Robust Principal Scores

A robust PCA using MCD estimator was performed to obtain new uncorrelated variables from the data containing outliers. The coefficients of the linear combinations of each PC are shows in Table 4.

TABLE 4. Robust Principal Component Coefficients

Variables	PC1	PC2	PC3	PC4
Confirmed Cases	-0.561	-0.253	-0.099	0.781
Death Casess	-0.571	0.153	-0.671	-0.446
Recovered Cases	-0.558	-0.257	0.681	-0.398
MIR	-0.215	0.920	0.274	0.179

The trimmed k-means clustering was done using all principal components to retain all the information in the data. We conducted α-trimmed k-means clustering with α=0.1 and the number of clusters k = 2,3,4,5,6,7, and 8. Using α=0.1 means that 10% of the trimmed data are observations that were not part of the cluster that was formed. The optimal clustering was then selected based on the validation indices resulting from each of the k clusters. The comparison of the average silhouette indexes of all k can be seen through their plots in Fig. 3.

In Fig. 3, it can be seen that the average silhouette index reached the maximum value at k=3 with the value equals 0.54; this means that the optimal cluster number for the data was k=3.

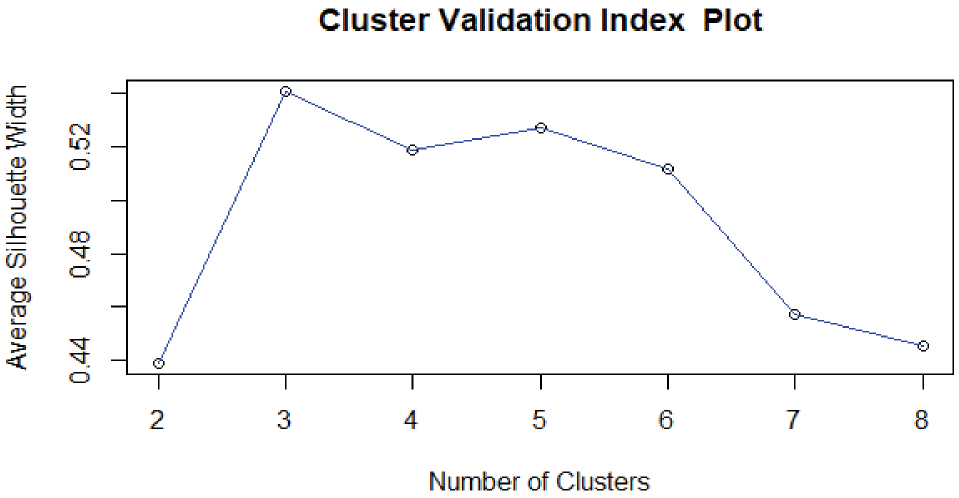

FIGURE 3. Plot of average silhouette index for k=2, …, 8

The three clusters resulting from the trimmed k-means clustering in Fig. 4 are Cluster 1, Cluster 2 and Cluster 3 (the other cluster in the figure, i.e. Cluster 0, is containing the four provinces detected as outliers; they will be analyzed further). Since the silhouette width of the observations in each cluster are positif ($s_{x_i} > 0$), then all observations in the three clusters are well grouped. Based on the silhouette index, the optimal number of clusters for the α-trimmed clustering (i.e. the data excluding 4 outliers) is k=3. Nevertheless, this result is in accordance with the separation index as shown in Figure 4. The result of the trimmed k-means clustering is shown in Fig. 5.

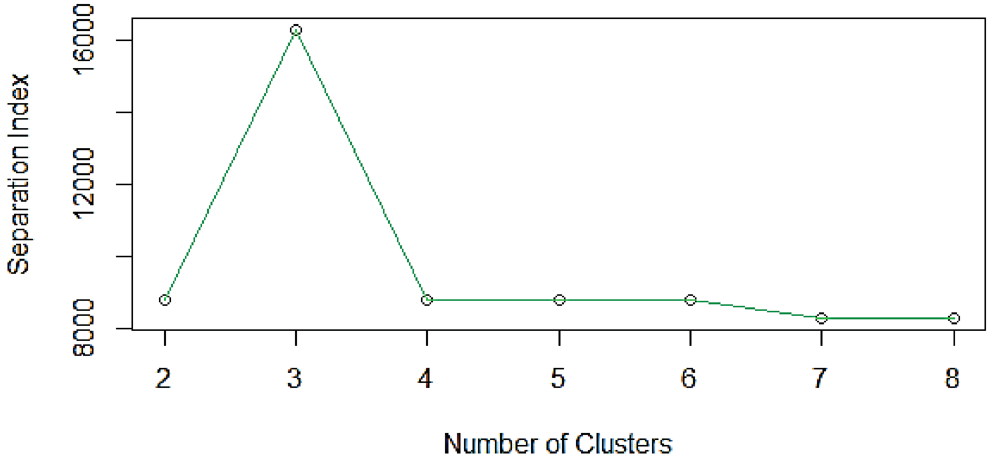

FIGURE 4. Plot of separation index for k=2, …, 8

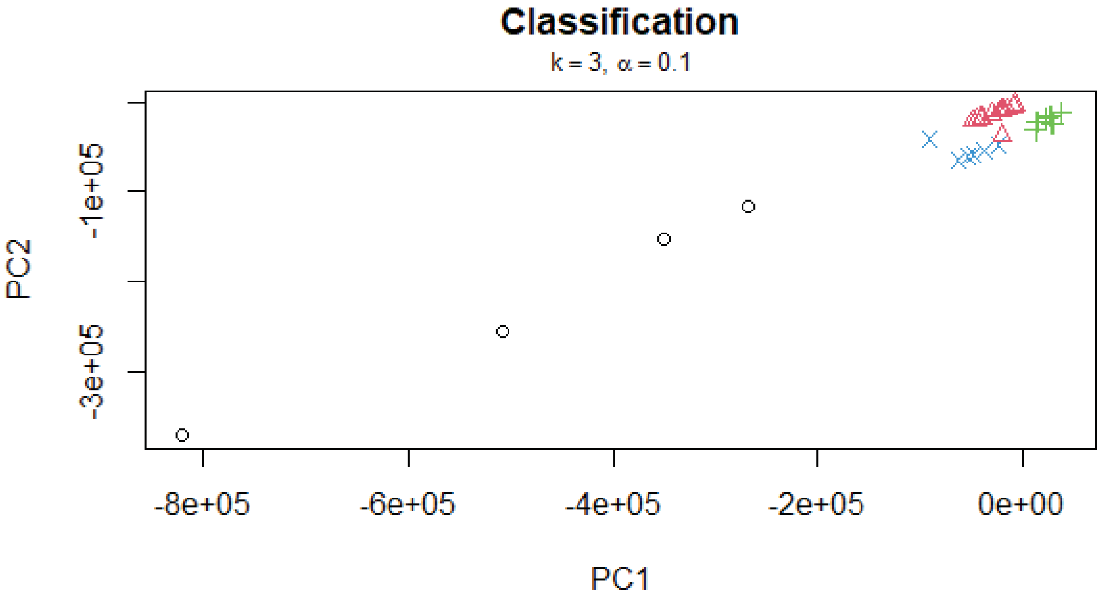

FIGURE 5. Trimmed *k*-means classification plot

Combining the outlying clusters with the three clusters of trimmed *k*-means clustering, we have 4 clusters, the member of each clusters are presented in Table 5 and grapically in Fig. 6.

TABLE 5. Members of Cluster

Cluster	Cluster size	Province
1	14	North Sumatera, South Sumatera, Lampung, Riau Island, Bali, West Nusa Tenggara, West Kalimantan, South Kalimantan, North Kalimantan, Central Sulawesi, South East Sulawesi, West Sulawesi, Maluku Utara, Papua
2	10	Aceh, Jambi, Bengkulu, Bangka Belitung Island, East Nusa Tenggara, Central Kalimantan, North Sulawesi, Gorontalo, Maluku, West Papua
3	6	West Sumatera, Riau, Banten, DI Yogyakarta, East Kalimantan, South Sulawesi
outliers	4	DKI Jakarta, West Java, Central Java, and East Java

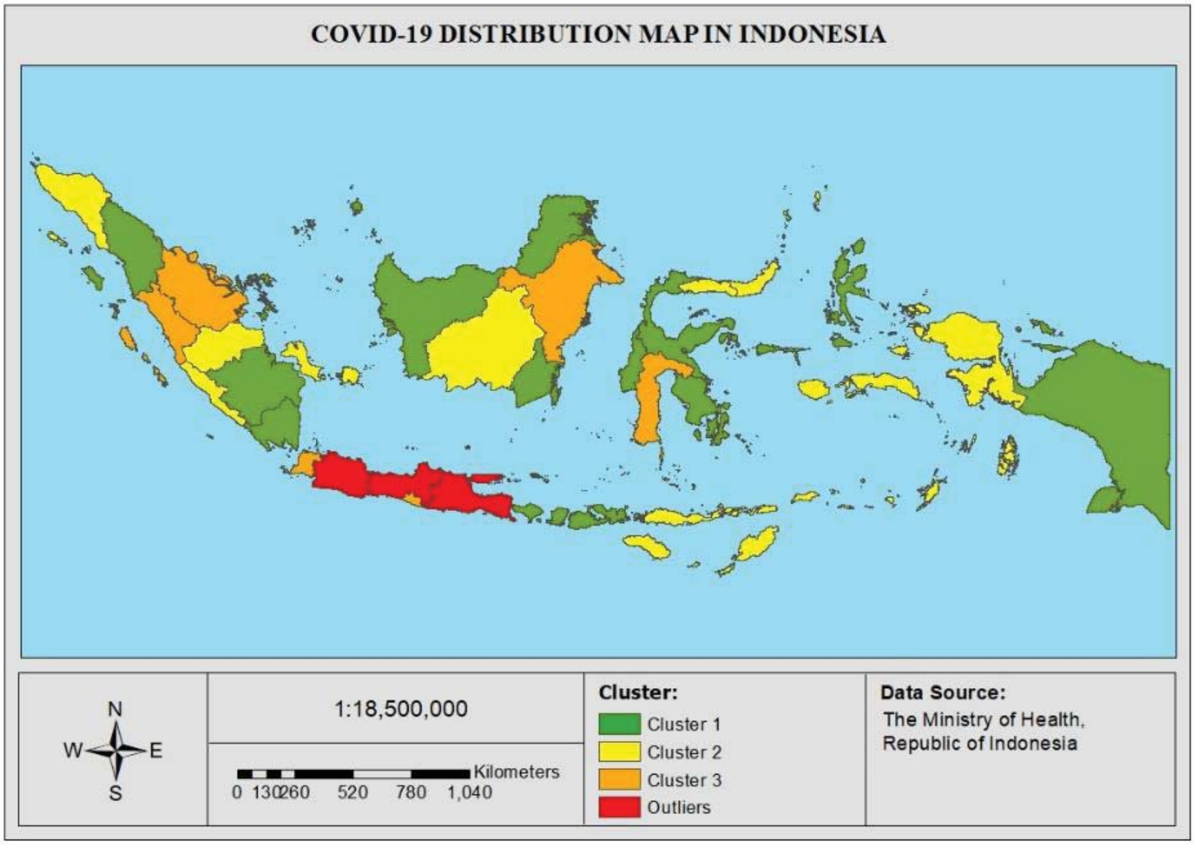

FIGURE 6. Covid-19 cluster result map in Indonesia between March 2020 and July 2021

To find out the characteristics of each cluster, it was necessary to do further analysis on each cluster by calculating the cluster centroid i.e. the mean value of each variable of the observations in the cluster $(\overline{\boldsymbol{X}}_{(x)k})$. Furthermore, these values are compared to the robust centroid of the entire data $(\boldsymbol{\mu}_{(x)})$ as shown in Table 6. If $\overline{\boldsymbol{X}}_{(x)k} \leq \boldsymbol{\mu}_{(x)}$, where *x* is the clustering variables, then it could be interpreted that the mean value of the variables in the cluster was "low", on the one hand. On the other hand, if $\overline{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$, then the mean value of the variables in the cluster could be interpreted as "high" and "very high" if $\overline{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$ of the 3 clusters. The result is shown in Table 7.

TABLE 6. Cluster Centroid

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	28708.8	787.9	21238.2	2.58
2	20519.3	453.7	15341.4	2.24
3	88434.7	2004.8	67230.7	2.24
Outliers	371262.6	14269.5	341566.9	3.54
Data ($\boldsymbol{\mu}_{(x)}$)	26391.3	648.9	20053.9	2.40

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	High	High	High	High
2	Low	Low	Low	Low
3	High	High	High	Low
Outliers	Very High	Very High	Very High	Very High

DISCUSSION

The results of data clustering on Covid19 in Indonesia from March 2020 to July 2021 show that the pandemic spread caused by coronavirus 2 (SARSCov2) varies from state to state. Of the 34 provinces in Indonesia, 4 provinces, DKI Jakarta, West Java, Central Java and East Java, have been found to have much higher deaths, recovery cases and MIRs than other provinces. Four states based on Mahalanobis distance squares are recognized as outliers. A trimmed kmeans clustering technique was used to solve this problem. This method can isolate outliers in the data and provide the best cluster. In the Indonesian Covid 19 pandemic data, the trimmed kmeans clustering technique formed three clusters based on the maximum isolation index. Cluster 1 consisted of 14 states, and clusters 2 and 3 consisted of 10 and 6, respectively. On the one hand, the outlier states DKI Jakarta, West Java, Central Java and East Java formed separate clusters outside the above three clusters. These results are consistent with preceded studies that have reported the trimmed *k*-means clustering method was robust to outliers [16,17,24,25].Similar results were obtained by [31] who studied a trimmed clustering Based l_1-Principal Component Analysis Model for Image Classification and Clustering Problems with Outliers. They showed that the trimming clustering method worked well in dealing with outliers.

For future work, we shall study whether the trimmed *k*-means method can overcome outliers in high dimensional data.

CONCLUSION

The three optimal clusters of the Covid 19 pandemic in Indonesia were discovered between March 2020 and July 2021 based on the maximum isolation index using the trimmed *k*-means clustering method. Cluster 1 consisted of 14 states, and clusters 2 and 3 consisted of 10 and 6, respectively. On the other hand, the four states of DKI Jakarta, West Java, Central Java and East Java were outliers and formed separate clusters outside the above three clusters. This indicates that the four states had more confirmed cases, recovered mortality, and higher mortality (MIR) than the other states in Indonesia. Researchers studying cluster analysis with outlier data are advised to use the kmeans trim clustering method instead of the *k*-means method.

REFERENCES

1. Livana, Ph.; Suwoso, R.H.; Febrianto, T.; Kushindarto, D.; Aziz, F. Dampak Pandemi Covid-19 Bagi Perekonomian Masyarakat Desa. Indones. J. Nurs. Health Sci. 2020, 1, 37–48.
2. Fahrika, A.I.; Roy, J. Dampak pandemi covid 19 terhadap perkembangan makro ekonomi di indonesia dan respon kebijakan yang ditempuh. INOVASI 2020, 16, 206–213, doi:10.29264/jinv.v16i2.8255.
3. Rohmah, S.N. Adakah Peluang Bisnis Di Tengah Kelesuan Perekonomian Akibat Pandemi Corona? ADALAH 2020, 4, 63–74.
4. Muhyiddin, M. Covid-19, New Normal, Dan Perencanaan Pembangunan Di Indonesia. J. Perenc. Pembang. Indones. J. Dev. Plan. 2020, 4, 240–252, doi:10.36574/jpp.v4i2.118.
5. Saidi, S.; Herawati, N.; Nisa, K. Modeling with Generalized Linear Model on Covid-19: Cases in Indonesia. Int. J. Electron. Commun. Syst. 2021, 1, 25–32.

6. Setiawan, S.S., Netti Herawati, Khoirin Nisa, Eri Nonparametric Modeling Using Kernel Method for the Estimation of the Covid-19 Data in Indonesia During 2020. Int. J. Math. Trends Technol. IJMTT.
7. Abdullah, D.; Susilo, S.; Ahmar, A.S.; Rusli, R.; Hidayat, R. The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic Based on COVID-19 Data. Qual. Quant. 2021, 1–9, doi:10.1007/s11135-021-01176-w.
8. Indraputra, R.A.; Fitriana, R. K-Means Clustering Data COVID-19. J. Tek. Ind. 2020, 10, 275–282, doi:10.25105/jti.v10i3.8428.
9. Virgantari, F.; Faridhan, Y.E. K-Means Clustering of COVID-19 Cases in Indonesia's Provinces. 2020, 7.
10. Vahabi, N.; Salehi, M.; Duarte, J.D.; Mollalo, A.; Michailidis, G. County-Level Longitudinal Clustering of COVID-19 Mortality to Incidence Ratio in the United States. Sci. Rep. 2021, 11, 3088, doi:10.1038/s41598-021-82384-0.
11. Rojas, F.; Valenzuela, O.; Rojas, I. Estimation of COVID-19 Dynamics in the Different States of the United States Using Time-Series Clustering. medRxiv 2020, 2020.06.29.20142364, doi:10.1101/2020.06.29.20142364.
12. Maugeri, A.; Barchitta, M.; Basile, G.; Agodi, A. Applying a Hierarchical Clustering on Principal Components Approach to Identify Different Patterns of the SARS-CoV-2 Epidemic across Italian Regions. Sci. Rep. 2021, 11, 7082, doi:10.1038/s41598-021-86703-3.
13. Kumar, S. Use of Cluster Analysis to Monitor Novel Coronavirus-19 Infections in Maharashtra, India. Indian J. Med. Sci. 2020, 72, 44–48, doi:10.25259/IJMS_68_2020.
14. Choi, Y.-J.; Park, M.-J.; Park, S.J.; Hong, D.; Lee, S.; Lee, K.-S.; Moon, S.; Cho, J.; Jang, Y.; Lee, D.; et al. Types of COVID-19 Clusters and Their Relationship with Social Distancing in the Seoul Metropolitan Area, South Korea. Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis. 2021, 106, 363–369, doi:10.1016/j.ijid.2021.02.058.
15. Gallegos, M.T.; Ritter, G. A Robust Method for Cluster Analysis. Ann. Stat. 2005, 33, 347–380, doi:10.1214/009053604000000940.
16. Cuesta-Albertos, J.A.; Gordaliza, A.; Matran, C. Trimmed K-Means: An Attempt to Robustify Quantizers. Ann. Stat. 1997, 25, 553–576.
17. Garcia-Escudero, L.A.; Gordaliza, A. Robustness Properties of k Means and Trimmed k Means. J. Am. Stat. Assoc. 1999, 94, 956–969, doi:10.2307/2670010.
18. Larasati, S.D.A.; Nisa, K.; Herawati, N. Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators. J. Phys. Conf. Ser. 2021, 1751, 012021, doi:10.1088/1742-6596/1751/1/012021.
19. Meng, S.; Fu, Y.; Liu, T.; Li, Y. Principal Component Analysis for Clustering Temporomandibular Joint Data. In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID); December 2015; Vol. 1, pp. 422–425.
20. Rahman, A.S.; Rahman, A. Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. Water 2020, 12, 781, doi:10.3390/w12030781.
21. Untoro, M.C.; Anggraini, L.; Andini, M.; Retnosari, H.; Nasrulloh, M.A. Penerapan metode k-means clustering data COVID-19 di Provinsi Jakarta. Teknol. J. Ilm. Sist. Inf. 2021, 11, 59–68, doi:10.26594/teknologi.v11i2.2323.
22. Utomo, W. The Comparison of K-Means and k-Medoids Algorithms for Clustering the Spread of the Covid-19 Outbreak in Indonesia. Ilk. J. Ilm. 2021, 13, 31–35, doi:10.33096/ilkom.v13i1.763.31-35.
23. Hutagalung, J.; Ginantra, N.L.W.S.R.; Bhawika, G.W.; Parwita, W.G.S.; Wanto, A.; Panjaitan, P.D. COVID-19 Cases and Deaths in Southeast Asia Clustering Using K-Means Algorithm. J. Phys. Conf. Ser. 2021, 1783, 012027, doi:10.1088/1742-6596/1783/1/012027.
24. García-Escudero, L.; Gordaliza, A.; Matrán, C.; Mayo, A. A General Trimming Approach to Robust Cluster Analysis. Ann. Stat. 2008, 36, 1324–1345, doi:10.1214/07-AOS515.
25. García-Escudero, L.; Gordaliza, A.; Matrán, C.; Mayo, A. A Review of Robust Clustering Methods. Adv. Data Anal. Classif. 2010, 4, 89–109, doi:10.1007/s11634-010-0064-5.
26. Nisa, K.; Herawati, N.; Setiawan, E.; Nusyirwan Robust Principal Component Analysis Using Minimum Covariance Determinant Estimator.; November 30 2006.
27. Rousseeuw, P.J.; Driessen, K.V. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 1999, 41, 212–223, doi:10.1080/00401706.1999.10485670.
28. Hubert, M.; Debruyne, M.; Rousseeuw, P.J. Minimum Covariance Determinant and Extensions. WIREs Comput. Stat. 2018, 10, e1421, doi:10.1002/wics.1421.

29. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J. Comput. Appl. Math. 1987, 20, 53–65, doi:10.1016/0377-0427(87)90125-7.
30. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of Internal Clustering Validation Measures. 2010 IEEE Int. Conf. Data Min. 2010, doi:10.1109/ICDM.2010.35.
31. Lam, B.S.Y.; Choy, S.K. A Trimmed Clustering-Based L1-Principal Component Analysis Model for Image Classification and Clustering Problems with Outliers. Appl. Sci. 2019, 9, 1562, doi:10.3390/app9081562.