Clustering Provinces in Indonesia Based on Daily Covid-19 Cases

H Fransiska*
Department of Mathematics, Bengkulu University

*E-mail : hfransiska@unib.ac.id

Abstract. The latest form of coronavirus is Coronavirus Disease 2019 (Covid-19). It was discovered in Wuhan, Hubei Province, China. The Covid-19 virus is growing very rapidly. Starting from Wuhan which then spread to the surrounding provinces to all provinces in China, and even spread to other countries such as Indonesia. Until August 17, 2020, an update from the KEMENKES RI that the development of Covid-19 in 34 Indonesian Provinces was positive (141,370), recovered (94,458), and died (6,207). This means that the case fatality rate in Indonesia is 4.39%. As a disease that is currently becoming a pandemic has a major impact in all sectors, it is very important to explore data and carry out clustering that is useful for policy making for the government such as decision making on WFH, PSBB, New Normal, or lockdown. The clustering provinces based on those similarities (homogeneity) data to find out which provinces have Covid-19 cases with similar characteristics and between clusters have different characteristics. Clusterization is expected to inform areas of high, medium, and low risk and their characteristic characteristics. Based on Covid-19 data, we used the hierarchical clustering method shows there are 4 clusters: cluster 1 (Jakarta and East Java), cluster 2 (Central Java, West Java, and South Sulawesi), cluster 3 (South Kalimantan, North Sumatra, South Sumatra, Papua, Bali, and East Kalimantan), and the provinces others in cluster 4. Validation of clustering shows Dynamic Time Warping (DTW) distance for hierarchical clustering (average linkage) is a good classification with an average silhouette value of 0.70.

1. Introduction
A new category of coronavirus is Coronavirus disease: Severe Acute Respiratory Syndrome Coronavirus 2 or abbreviated SARS-CoV-2 [1]. It was discovered in Wuhan, Hubei Province, China as mysterious pneumonia in Dec-2019. So, it called Coronavirus disease 2019 or COVID-19 This spread rapidly from Wuhan to other regions and has now spread to almost all countries and territories. The World Health Organization (WHO) declared the COVID-19 outbreak a pandemic on March 11, 2020 [2]. On March 2, 2020, the first confirmed positive Covid-19 case appeared in Indonesia (DKI Jakarta Province). Similar to other countries, the case of covid-19 has spread rapidly to all provinces in Indonesia [3]. Until August 17, 2020, an update from the Indonesian health ministry stated that the development of Covid-19 in 34 Indonesian Provinces was positive: 141,370, recovered: 94,458, and died 6,207. This means that the case fatality rate (CFR) in Indonesia is 4.39%.

Pandemic Covid-19 has a major impact in all sectors so it is very important to descriptive data and clustering data which is useful for policy making for the central government such as making decisions on Work From Home (WFH), New Normal, or lockdown [4]. One of the difficulties experienced by the government in handling COVID-19 is the level of emergency and policies
implemented by local governments. The central government knew each region has different characteristics of Covid-19 cases so that knowledge is needed about the similarities of regional characteristics in handling it. Therefore, the aim of this study is to clustering data using time series data active cases of COVID-19 in Indonesia. Clustering of time series has been shown to be efficient in providing useful information and, unlike static data, a feature’s time series contain values that have changed over time [5]. So, we use this method for analyzing clustering provinces based on daily Covid-19 data.

Time series clustering has been carried. In 2018, development of time series models with cluster analysis for the broad proportion attack of main plant-disturbing organisms of food crops in Indonesia [6]. In 2019, the implementation of clustering time series for province in Indonesia based on the rice production [7] and in 2020, development of rice price modeling in western Indonesia with the time series clustering approach [8].

2. Time Series Clustering
Clustering is the unsupervised method of grouping data patterns into clusters so that cluster patterns are closely similar to each other, but somewhat different from other cluster patterns [9]. Time series clustering is often necessary for the solution of real problems originating from a different domain to provide useful knowledge [10]. Clustering is appropriate in the absence of labeled data, irrespective of whether the data are nominal, ordinal, interval, ratio, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the data types listed above.

Clustering time series analysis is used by grouping objects based on their time series patterns [5]. If all their function values do not change over time, or change negligibly, data is considered static. One of clustering methods developed for handling various static data is hierarchical methods. A technique of hierarchical clustering operates by grouping data items into a cluster tree. Time series clustering, like static data clustering, involves a clustering algorithm or process to shape clusters given a collection of unlabeled data objects, but the method of distance selection and clustering is in accordance with the time series data structure which is very dynamic in nature [5]. Some algorithms or procedures for general purpose clustering that were used in clustering studies of the time series.

2.1. Distance Measurement
To find the similarity or dissimilarity between two time series, the distance measure was used. There are two distance measurements available: the Euclidean distance metric and the distance measurement with Dynamic Time Warping (DTW) [11]. We used DTW because DTW is a design measure of similarity. DTW breaks the one-to-one alignment constraint, and also accepts time series that are not equal in length. The DTW distance is the minimum distance where the cumulative distance of each element in the matrix is the minimum of the three surrounding neighbors [8]. The goal of the DTW distance is to find a mapping \( r \) between the series in order to minimize a particular distance measure between the coupled observations \((X_{ai}, Y_{bi})\). Let \( M \) be the set of all decision sequences of \( m \) pairs that preserve the order of observations in the form of

\[
r = ( (X_{a_1}, Y_{b_1}), \ldots, (X_{a_m}, Y_{b_m}) )
\]

with \( a_i, b_j \in \{1, \ldots, T\} \) such that \( a_1 = b_1 = 1 \), \( a_m = b_m = T \) and \( a_{i+1} = a_i \) or \( a_i + 1 \) and or \( b_{i+1} = b_i \) or \( b_i + 1 \), for \( i \in \{1, \ldots, m - 1\} \). The definition of the distance from DTW is given by [12]

\[
D_{DTW}(X_T, Y_T) = \min_{r \in M} \left\{ \sum_{i=1,2,\ldots,m} |X_{ai} - Y_{bi}| \right\}
\]

2.2. Clustering tools: A Hierarchical clustering
Hierarchical clustering approaches arrange knowledge on the basis of acceptable proximity measures in the hierarchical structure. That is, similarity indices and dissimilarity measurements for distance
measurements [13]. There are two types of clustering: agglomerative and divisive approaches. The method of agglomerative hierarchical clustering is more common than the method of division [5]. With 1-clusters, agglomerative clustering begins. In these strategies, each of which requires exactly one data point, a sequence of merge operations are followed next, eventually forcing all objects into the same class. [13]. The following procedure can summarize the general agglomerative clustering [14]:

1. Begin with the singleton clusters of I.
2. Look for the minimum distance in the distance matrix and update the cluster.
3. Update the distance matrix by calculating the distances between Step two and the other clusters of the clusters;
4. Repeat measures 2 and 3 until there is just one cluster.

Several strategies for describing distance functions: Single linkage (SL), complete linkage (CL), and average linkage (AL). In calculating their inter-cluster distance, these methods utilize all points of a pair of clusters, and they are also called graph methods [13].

2.3. Evaluation of Technical Dissimilarity Measures
One of the clustering technique dissimilarity measures is the cophenetic correlation coefficient [15]. The coefficient of cophenetic correlation is the coefficient of correlation between the original dissimilarity matrix elements (Euclidean distance) and the elements generated by the dendrogram (Cophenetic matrix based on distance measures and connectedness methods used). The cophenetic correlation coefficient formula is:

$$r_{coph} = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{Cik} - \bar{d}_C)}{\sqrt{\sum_{i<k}(d_{ik} - \bar{d})^2} \sum_{i<k}(d_{Cik} - \bar{d}_C)^2}$$  (3)

Where:
- $r_{coph}$ = Cophenetic correlation coefficient
- $d_{ik}$ = Euclidean distance of $i$-th and $k$-th objects
- $\bar{d}$ = Average $d_{ik}$
- $d_{Cik}$ = Cophenetic distance of $i$-th and $k$-th object
- $\bar{d}_C$ = Average $d_C$.

The DTW distance measure with the linkage method that produces the largest cophenetic value is the best solution for the hierarchical clustering method [7].

2.4. Cluster validity criteria: Silhouette criterion
The cluster validity criterion was introduced by Rousseeuw to evaluate the number of clusters using the silhouette criterion (1987) [13]. Consider a unit $i \in (1, \ldots, I)$ that belongs to cluster $p \in (1, \ldots, C)$. Let the average or squared euclidean distance of $i$-th unit to all the other units belonging to cluster $p$ be denoted by $a_{ip}$. Also, Let this unit's average distance to all the units belonging to some other $q$ ($q \neq p$) cluster be labelled $d_{iq}$. At last, let $b_{ip}$ be a minimum of $d_{iq}$ measured over $q = 1, \ldots, c$, $q \neq p$, reflecting the $i$-th unit's dissimilarity to its nearest neighbouring cluster. Then, the $i$-th object silhouette is described as follows:

$$S_i = \frac{b_{ip} - a_{ip}}{\max[a_{ip}, b_{ip}]}$$  (4)

Where the equation is a concept of normalization. Obviously, the greater the $S_i$ value, the better the classification to the $c$-th cluster of the $i$-th unit. The silhouette identified as the $S_i$ mean over $i=1, \ldots, I$ is [13]:
\[ SIL = \frac{1}{n} \sum_{i=1}^{n} S_i \]  

(5)

The criteria for the accuracy and quality of the clustering result based on the silhouette coefficient value are as follows: 0.71 – 1.00 (Strong), 0.51 – 0.70 (Good), 0.26 – 0.50 (Weak), and 0 – 0.25 (Bad) Classification [16]

3. Methods
The data used in this study is the total of daily positive cases of Covid-19 in Indonesia. This data from kawal information about COVID-19 (https://kawalcovid19.id/). The total of daily positive cases of Covid-19 data is time-series data. This study used data from March 18, 2020 to September 17, 2020.

The initial step in this study is to describe daily positive cases of Covid-19 in 34 provinces. The next step is cluster analysis for time series data which there are some stages: Distance measure used Dynamic Time Warping (DTW), Clustering used Hierarchical methods (Single linkage (SL), complete linkage (CL), and average linkage (AL)), Evaluation of Dissimilarity Measures used the cophenetic correlation coefficient, and then cluster validity criteria used Silhouette criterion. The results obtained at this stage is the clustering of 34 Provinces in Indonesia based on the total of daily positive cases of Covid-19.

4. Results and Discussion
4.1. Descriptive Statistic
The number of daily Covid-19 cases in Indonesia (March 18, 2020-September 17, 2020) reached 232,628. These cases spread in 34 Provinces. Plot for number of daily Covid-19 can be seen in Figure 1:

![Figure 1](Image)

Figure 1. Ts plot of Covid-19 cases in 34 provinces Indonesia.

This plot shows of daily Covid-19 in Indonesia always increase every day. As the capital city of Indonesia, the number of covid-19 case in Jakarta is very different and highest from the others.

The descriptive statistics of covid-19 cases in Indonesia:
Table 1. Descriptive statistics of daily Covid-19 in Indonesia

| Province  | Mean   | Q1    | Median  | Q3    |
|-----------|--------|-------|---------|-------|
| Aceh      | 407.00 | 11.30 | 37.50   | 432.50|
| Bali      | 1954.00| 243.00| 862.00  | 3519.00|
| Banten    | 1334.00| 428.30| 1318.00 | 1885.80|
| Babel     | 116.60 | 19.25 | 147.00  | 195.25 |
| Bengkulu  | 139.99 | 12.00 | 105.00  | 235.00 |
| DIY       | 488.70 | 114.30| 276.00  | 769.00 |
| Jakarta   | 15214.00| 4414.00| 9433.00 | 22498.00|
| Jambi     | 122.66 | 33.50 | 109.00  | 169.00 |
| Jabar     | 4352.00| 1046.00| 2731.00 | 6679.00|
| Jateng    | 5178.00| 769.00 | 2369.00 | 9803.00|
| Jatim     | 12855.00| 1057.00| 8725.00 | 22863.00|
| Kalbar    | 277.90 | 68.50 | 282.00  | 387.00 |
| Kaltim    | 1186.00| 156.00| 401.00  | 1533.00|
| Kalteng   | 1007.40| 157.00| 720.00  | 1787.50|
| Kalsel    | 3241.00| 183.00| 2267.00 | 6227.00|
| Kaltara   | 194.96 | 122.00| 171.00  | 284.00 |
| Kep Riau  | 357.50 | 89.80 | 263.00  | 493.00 |
| NTB       | 1186.20| 254.80| 993.00  | 2141.30|
| Sumsel    | 1919.00| 163.00| 1569.00 | 3452.00|
| Sumbar    | 835.30 | 185.30| 693.50  | 969.80 |
| Sulut     | 1363.00| 45.00 | 749.00  | 2727.00|
| Sumut     | 2322.00| 119.00| 982.00  | 4179.00|
| Sultra    | 541.30 | 64.00 | 309.00  | 788.50 |
| Sulsel    | 5012.00| 583.00| 3283.00 | 9720.00|
| Sulteng   | 142.30 | 59.00 | 172.00  | 213.00 |
| Lampung   | 186.30 | 50.00 | 170.00  | 290.50 |
| Riau      | 548.20 | 47.00 | 131.00  | 493.50 |
| Malut     | 746.20 | 43.30 | 332.50  | 1560.80|
| Maluku    | 706.80 | 23.00 | 528.50  | 1135.00|
| Papbar    | 301.30 | 43.00 | 216.50  | 460.00 |
| Papua     | 1674.00| 240.00| 1331.00 | 3114.00|
| Sulbar    | 149.90 | 44.00 | 101.50  | 238.80 |
| NTT       | 94.86  | 4.75  | 108.00  | 151.00 |
| Gorontalo | 619.10 | 15.00 | 214.00  | 1312.50|

In this table, the descriptive of covid-19 case from March 18, 2020, until September 17, 2020 in Aceh Province has an mean value 407 cases which has the Q1 11 cases, median 37 cases and Q3 432 cases. As well as 33 provinces other which can be seen in the table.
4.2. Clustering

The initial step in the time series clustering is to calculate the distance measurement. This distance is used DTW distance. The results of the calculation of the DTW distance of the covid-19 cases in 34 Provinces are Aceh-Bali (89.575), Banten-Aceh (8295) and etc. The minimum DTW distance is Sulteng-Babel (555) and the maximum DTW distance Sulteng-Jakarta (2758122). So the first steps clustering is make Sulteng and Babel as one cluster.

The next step is clustering data used Hierarchical (single linkage (SL), complete linkage (CL) and average linkage (AL)). Based on calculations using R, we has the dendogram:

![Dendogram of hierarchical clustering](image)

**Figure 2.** Dendogram of hierarchical clustering. (a) single linkage (SL), (b) complete linkage (CL) and (c) average linkage (AL)
From Figure 2, it can be seen that the cluster of Covid-19 cases is 4 cluster. The next step is calculate cophenetic value for evaluation of technical dissimilarity measures. The results of the calculation of the cophenetic value of the three linkage methods are as follows:

| No | Linkage Methods | Cophenetic Value |
|----|-----------------|------------------|
| 1  | SL              | 0.69             |
| 2  | CL              | 0.56             |
| 3  | AL              | 0.62             |

The results of the cophenetic value of three linkage processes are shown in Table 2. These results indicate that average linkage is the best methods because cophenetic value is largest (0.69). Then we use alternative cluster with no weight for the dynamic behavior (raw DTW) as comparing. We has the dendogram are follows:

![Cluster Dendrogram](image)

**Figure 3.** Dendogram of hierarchical clustering with no weight for the dynamic behavior (a) single linkage (SL), (b) complete linkage (CL) and (c) average linkage (AL)
From Figure 3, it can be seen that the cluster of daily Covid-19 cases is 4 cluster. The cophenetic value for evaluation of technical dissimilarity measure are as follows:

**Table 3.** Cophenetic value of three linkage methods for cluster with no weight for the dynamic behavior

| No | Linkage Methods | Cophenetic Value |
|----|-----------------|------------------|
| 1  | SL              | 0.94             |
| 2  | CL              | 0.94             |
| 3  | AL              | 0.95             |

Table 3 shows the results of the cophenetic value which indicate that average linkage is the best methods because cophenetic value is largest (0.95).

One of the methods to estimate the cluster quality is by looking at the silhouette value.

Based on Figure 4, the number of clusters is 4 cluster. Therefore we used 4 clusters with the average linkage method for final clustering number of daily covid-19 cases in Indonesia. The distribution of the clustering and silhouette values can be seen in Figure 5:

**Figure 4.** Optimal number of cluster used silhouette value

**Figure 5.** Silhouette plot for daily Covid-19 cases In Indonesia

Based on the criteria for the accuracy and quality of the clustering, on figure 5 we have cluster 1 (0.56) and cluster 4 (0.66) as good classification, cluster 2 (0.91) and cluster 3 (0.83) as strong classification. Average of silhouette value for 4 cluster is 0.70 this meaning this clustering is a good classification.
Based on the number of positive covid-19 provinces in Indonesia, the grouping of provinces indicates that high mobility cities such as Jakarta and East Java are in one cluster. This cluster can also be said to be the cluster with the highest covid-19 risk, and the fourth cluster is provinces where Covid-19 risk is quite low because these provinces in this group are low mobility dan small provinces.

5. Conclusion
A time-series clustering approach for clustering provinces in Indonesia based on daily covid-19 cases result that there were 4 clusters: cluster 1 (Jakarta and East Java), cluster 2 (Central Java, West Java, and South Sulawesi), cluster 3 (South Kalimantan, North Sumatra, South Sumatra, Papua, Bali, and East Kalimantan), and the provinces others in cluster 4. Validation of clustering shows Dynamic Time Warping (DTW) distance for hierarchical clustering (average linkage) is a good classification with an average silhouette value of 0.70.

References
[1] Bikdeli B, Madhavan M V, Gupta A, Jimenez D, Burton J R, Der Nigoghossian C, Chuich T, Nouri S N, Dreyfus I, Driggin E, and others 2020 Pharmacological agents targeting thromboinflammation in COVID-19: review and implications for future research Thromb. Haemost. 120 1004
[2] Cdc Covid-19 Response Team 2020 “Severe Outcomes Among Patients with Coronavirus Disease 2019 (Covid-19) — United States, February 12–March 16, 2020,” MMWR. Morb. Mortal. Wkly. Rep., vol. 69, no. 12, pp. 343–346
[3] Kementerian Kesehatan RI 2020 “Pedoman Pencegahan dan Pengendalian Coronavirus Disease (Covid-19),” pp. 1–214
[4] Zahrotunnimah Z 2020 “Langkah Taktis Pemerintah Daerah Dalam Pencegahan Penyebaran Virus Corona Covid-19 di Indonesia” SALAM J. Sos. dan Budaya Syar-i, vol. 7, no. 3, Mar
[5] Liao T W 2005 “Clustering of time series data - A survey,” Pattern Recognit., vol. 38, no. 11, pp. 1857–1874
[6] Setiawan I 2018 “Pengembangan Model Deret Waktu Dengan Analisis Gerombol ( Studi Kasus : Organisme Pengganggu Tumbuhan ( Opt ) Utama Tanaman Pangan Di Indonesia ),”
[7] Munthe A D 2019 “Penerapan Clustering Time Series Untuk Menggerombolkan Provinsi Di Indonesia Berdasarkan Nilai Produksi Padi,” J. Litbang Sukowati L, vol. 2, no. 1, pp. 1–11
[8] Wijaya S U and Ngatini N N 2020 “Pengembangan Pemodelan Harga Beras di Wilayah Indonesia Bagian Barat dengan Pendekatan Clustering Time Series,” Limits J. Math. Its Appl., vol. 17, no. 1, p. 51
[9] Xiong Y and Yeung D Y 2004 “Time series clustering with ARMA mixtures,” Pattern Recognit., vol. 37, no. 8, pp. 1675–1689
[10] Corduas M and Piccolo D 2008 “Time series clustering and classification by the autoregressive metric,” Comput. Stat. Data Anal., vol. 52, no. 4, pp. 1860–1872
[11] Niennattrakul V and Ratanamahatana C A 2007 “On clustering multimedia time series data using k-means and dynamic time warping,” Proc. - 2007 Int. Conf. Multimed. Ubiquitous Eng. MUE 2007, no. January 2007, pp. 733–738
[12] Montero P and Vilar J A 2014 “TSclust: An R package for time series clustering,” J. Stat. Softw.
[13] Maharaj E A, D’Urso P, and Caiado J 2019 Time Series Clustering and Classification
[14] Xu R, and Wunsch, D C 2010 Clustering algorithms in biomedical research: a review IEEE reviews in biomedical engineering, 3, 120-154.
[15] Adinugroho W 2018, “Pengembangan Model Peramalan Harga Minyak Goreng Dengan Pendekatan Clustering Time Series Widayayanto Adinugroho,”
[16] Kaufman L and Rousseuw P J 1990 Finding Groups In Data : An Introduction to Cluster Analysis