ComTech: Computer, Mathematics and Engineering Applications, 13(2), December 2022, 63−73 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v13i2.7413 E-ISSN: 2476-907X

Dynamic Time Warping Techniques for Time Series Clustering of Covid-19 Cases in DKI Jakarta

Meicheil Yohansa[1], Khairil Anwar Notodiputro[2]*, and Erfiani[3]

[1-3]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Jln. Raya Dramaga, Kampus IPB Dramaga Bogor, Jawa Barat 16680, Indonesia
[1]meicheilyohansa@apps.ipb.ac.id; [2]khairil@apps.ipb.ac.id; [3]erfiani@apps.ipb.ac.id

Received: 8th June 2021/ **Revised:** 27th September 2021/ **Accepted:** 28th August 2021

How to Cite: Yohansa, M., Notodiputro, K. A., & Erfiani. (2022). Dynamic Time Warping Techniques for Time Series Clustering of Covid-19 Cases in DKI Jakarta. *ComTech: Computer, Mathematics and Engineering Applications, 13*(2), 63−73. https://doi.org/10.21512/comtech.v13i2.7413

Abstract **-** The number of positive cases of Covid-19 in DKI Jakarta has contributed to the national issues, reaching 25% of the total cases in Indonesia. The research examined and modeled the distribution pattern of Covid-19 positive cases in DKI Jakarta based on 44 districts spreading over six administrative areas. The data were regarding positive Covid-19 cases in DKI Jakarta for the past year, from April 2020 to April 2021. The research related to the pattern of positive Covid-19 distribution in 44 districts was carried out by time series clustering through Dynamic Time Warping (DTW) distances and agglomerative hierarchical methods. Then, the effectiveness of the clustering process is evaluated by comparing the predicted value of Covid-19 cases between clustering and non-clustering forecast results at the city level for the next 14 days through the Autoregressive Integrated Moving Average (ARIMA) model. The results group 44 districts into 6 optimal clusters based on the pattern of positive cases of Covid-19 in each district. The highest distribution rate is in cluster A, and the lowest is in cluster F. Geographical characteristics are also indicated by clusters A, B, E, and F. Then, the results show that the Mean Average Percentage Error (MAPE) value of the clustering model ranges from 16% to 20%. The difference between MAPE values to the non-clustering model implies that the forecasting accuracy is not far apart, which is in the round of 5%−6%.

Keywords: dynamic time warping, time series, Covid-19

I. INTRODUCTION

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), known as the Covid-19 virus, has become a global pandemic and has had a significant negative impact since it was first identified in December 2019 in Wuhan, China. According to the data compiled by the Johns Hopkins Coronavirus Research Center (Johns Hopkins University, 2022), the total worldwide cases in April 2021 reached 140 million, with 3 million recorded as deaths cases. Indonesia is one of the countries that has not been spared from the Covid-19 outbreak. The massive spread of this virus has made Indonesia as one of the 20 countries with the highest positive cases. The data from John Hopkins Coronavirus Research Center have shown that Indonesia becomes the country with the highest number of active cases in the Asian region due to imbalance between new and cured cases.

The high number of Covid-19 cases in Indonesia is absolutely influenced by cases in each province. According to the data from Indonesia's Covid-19 handling task force, DKI Jakarta is the province that contributes the most to Covid-19 daily cases, which is around 25% of national cases (Pangaribuan & Munandar, 2021). Figure 1 shows the comparison of Covid-19 cases between DKI Jakarta and the whole Indonesia. This situation is inseparable from the high community mobilization in DKI Jakarta and the various momentums that cause crowds such that the spread of the Covid-19 virus increased. Various government policies have not reduced the daily increase of Covid-19 cases since the trend of positive cases in DKI Jakarta were keep increasing, which reached the highest number of 4.000 cases on February 7, 2021.

Various studies on Covid-19 cases in DKI Jakarta have been previously carried out by individual or government institutions. The majority of the studies have focused on infographics, epidemiology, and qualitative social studies related to the impact of Covid-19 on various aspects. One of the inference studies that utilize Covid-19 data in DKI Jakarta is conducted by Wiguna et al. (2020). They used the

*Corresponding Author

Covid-19 data of DKI Jakarta in the period from March 2020 to July 2020 and produced a model accuracy through the Mean Average Percentage Error (MAPE) value of 20,97%. Then, Solichin and Khairunnisa (2020) utilized Covid-19 data in DKI Jakarta at the district level. They classified the districts in DKI Jakarta into nine different clusters as measured by the sum of squares of the errors. However, the use of data that had yet to be maximized was one of the shortcomings of the previous study. The data used were only cumulative data for one day (May 13, 2020).

According to the background of the problems, the research aims to examine and forecast daily cases of Covid-19 in DKI Jakarta by including the time dimension through the use of daily data at the district level. The study of Covid-19 daily case distribution is carried out using a time series clustering technique through Dynamic Time Warping (DTW) distances with an agglomerative hierarchical method for 44 districts in Jakarta. Meanwhile, cluster-based forecasting is applied using the results of time series clustering by creating models for each cluster with Autoregressive Integrated Moving Average (ARIMA) model. A similar study is previously conducted by Fransiska (2021), revealing that the hierarchical clustering with the average linkage method produces the highest cophenetic value with the optimal number of clusters (4 clusters).

Previous studies related to clustering time series data only focus on the distance or clustering method. Then, the effectiveness of clustering is only carried out at the cluster level, without comparing the predicted results if no clustering is carried out. Similarly, evaluation is generally done through MAPE values with actual data at the cluster level. This procedure does not measure whether the forecasting results at the cluster level can represent the actual data conditions or not. This issue is very important, especially in the context of involving public policy, as done in the research.

Based on the general evaluation methods described, the research novelty lies in how to evaluate the effectiveness of the clustering process. Evaluation of clustering effectiveness is handled by modeling the clustering results and comparing forecasting accuracy with non-clustering models at the city level. Its effectiveness is measured by the difference between the two models' prediction errors compared to the actual data. If the prediction based on the clustering procedure gives results that are not much different from the non-cluster model, the clustering process can be said to be quite effective.

II. METHODS

The data are time-series for Covid-19 cases in DKI Jakarta at the district level in the period of April 2020 to April 2021. Daily positive cases are chosen because it represents the addition of Covid-19 cases without being affected by recovered or death cases. The data are obtained through the official website of DKI Jakarta's Covid-19 handling task force. Table 1 describes the structure of the data used.

Figure 1 Time Series Plot for Covid-19 Daily Case in Indonesia and DKI Jakarta

Table 1 The Example of Data Structure

Date	District			
	Cakung	**Ciracas**	...	**Tebet**
01/04/2020	0	1	...	0
02/04/2020	0	1	...	3
⋮	⋮	⋮	...	
04/04/2021	6	9	...	8

The importance of clustering in the research is to find groups of districts with a similar pattern of distribution of Covid-19 cases so that the mapping of policy implementation can be more accurate based on the results of this clustering. The clustering technique used is average linkage agglomerative hierarchical clustering. Then, a clustering process that starts with one object in a set is combined with other objects with similar characteristics (Sammour, Othman, Rus, & Mohamed, 2019). This process continues until the optimal conditions are met and produces one cluster as the endpoint. An overview of the agglomerative hierarchical method is shown in Figure 2.

The distance method used to measure the similarity between time series is the DTW method, which calculates the shortest distance. It includes all trajectories in the matrix of the two series compared (Dong & Liu, 2018). DTW is generally used in speech recognition. Its algorithm compares two data series and performs calculations to find the optimum path between two data series through the data alignment process (Wang, Lyu, Shi, & Liang, 2018). In other words, the DTW distance is the minimum distance between two series by considering the possibility of a point shift (Han et al., 2020). As the advantages of DTW, this distance method is used because it can accommodate the condition of Covid-19 data in each sub-district that is identified for the first time at different times. In addition, the DTW distance also works for non-linear series, such as Covid-19 data. The following equation defines the DTW distance (Wan, Chen, & Shi, 2017). It shows S,T ($S = s_1, s_2, \ldots s_i, \ldots, s_n$ and $T = t_1, t_2, \ldots t_j, \ldots, t_m$) as some time series, $W(w_1, w_2, \ldots w_k)$ as possible curvature path that remaps the members of S and T so that the distance between them is minimum, $\delta_{(i,j)}$ as $|s_i - t_j| + \min\{\delta_{(i-1,j-1)}, \delta_{(i,j-1)}, \delta_{(i-1,j)}\}$, and w_k as points $(i,j)_k$ on the $k-th$ curve path.

$$d_{DTW}(S,T) = min_W = \left[\sum_{k=1}^{p} \delta(w_k)\right] \quad (1)$$

Based on the DTW algorithm described, DTW is one of the non-linear methods used to calculate the distance between two-time series sequences. Those may have different lengths (Puspita & Zulkarnain, 2020). As an illustration, Figure 3 shows the comparison between alignment with Euclidian distance (Figure 3a) and DTW distance (Figure 3b).

Alignment with Euclidian distance has a one-to-one correspondence principle. The *i*-th observation in the first series will be compared with the *i*-th observation in the other series. On the other hand, the DTW method allows for alignment with many-to-one or one-to-many principles so that the *i*-th observation in one series can be compared with a different order in another series.

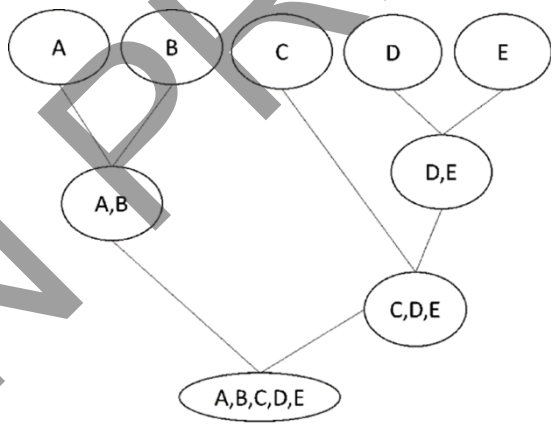
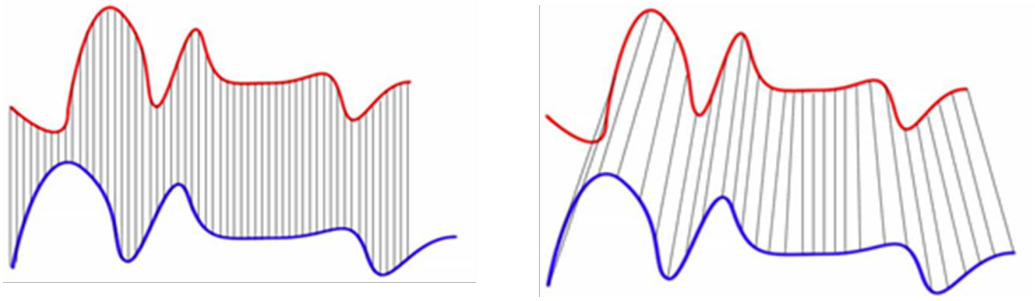


Figure 2 Agglomerative Hierarchical Clustering



(a) (b)

Figure 3 Illustration Comparison of Euclidian Distance and Dynamic Time Warping (DTW)

Then, the dissimilarity measurement becomes an essential issue in clustering because it will determine the optimization of the cluster formed. The cophenetic correlation coefficient can be used to evaluate the goodness of the dissimilarity measurement method in the clustering process of time series data. The cophenetic correlation coefficient formula can be written as follows (Novidianto & Dani, 2020).

$$c = \frac{\sum_{i<j}(d(i,j)-\bar{d})(v(i,j)-\bar{v})}{\sqrt{[\sum_{i<j}(d(i,j)-\bar{d})^2][\sum_{i<j}(v(i,j)-\bar{v})^2]}} \quad (2)$$

The $d(i, j)$ is the dissimilarity between the i-th and i-th observations. Then, d is the average dissimilarity. Meanwhile, $v(i, j)$ is the distance between the i-th dendrogram point, and the j-th is the dendrogram point. The v is the average distance on the dendrogram. If the value of the cophenetic correlation coefficient gets closer to one, the distance method is suitable.

Furthermore, objects that are clustered will be grouped into k clusters. The limit for the number of clusters that can be formed in the research is $k \leq (n - 1)$, where n is the number of districts, and $k = 1$ is excluded. So, the limitation for the number of clusters formed can be written as $2 \leq k \leq (n - 1)$. The formula used to measure cluster accuracy can be written as follows (Řezanková, 2018).

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (3)$$

It shows $s(i)$ as silhouette coefficient, $a(i)$ as the average dissimilarity of each i-th object to all other objects in group A, and $b(i)$ as the average dissimilarity of each i-th object to all other objects in cluster B, if cluster A is considered non-existent.

This silhouette coefficient $s(i)$ will be the indicator to determine how many clusters are the most optimal for clustering objects. This silhouette coefficient value ranges from -1 to 1. The silhouette coefficient ranges from -1 to 1. If the value is close to 1, it means that objects are well clustered. Then, if the value is close to 0, the objects are possible to be assigned to another cluster. Meanwhile, if the silhouette coefficient value is close to -1, the objects are misclassified. Every number of clusters, which are $2 \leq k \leq (n - 1)$, will have its silhouette coefficient. The optimal number of clusters will be selected based on the best silhouette coefficient. Based on Kaufman and Rousseeuw (1990), the silhouette coefficient can be interpreted subjectively related to the strength of the cluster structure formed. These interpretations are presented in Table 2.

The results of time series clustering of k clusters will be modeled to predict positive cases of Covid-19 at the cluster level. The model used is the ARIMA model. ARIMA model is a generalized model of the Autoregressive Moving Average (ARMA), which combines the Autoregressive (AR) process and Moving Average (MA) processes and builds a composite model of the time series (Siami-Namini, Tavakoli, & Siami Namin, 2018). The ARIMA model is generally written in the following equation (Atique et al., 2019). It describes Yit as a daily number of covid-19 in cluster i, B as the backshift operator, ϕ_p as an autoregressive parameter, Θ_q as the moving average parameter, d as differencing level, and e_{it} as model error for i-th cluster.

$$\phi_p(B)(1-B)^d Y_{it} = \theta_q(B)e_{it} \quad (4)$$

The identification of the ARIMA model can be made through the identification of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) graphs (Benvenuto, Giovanetti, Vassallo, Angeletti, & Ciccozzi, 2020). The ACF plot is used to determine the moving average order. Meanwhile, the PACF plot determines the order of autoregression.

Generally, the analysis procedures are divided into two main stages: clustering and forecasting. The following procedure describes the steps in each stage of the analysis. The first stage is the clustering process. The clustering process begins by calculating the dissimilarity measurement for every two of the 44 districts in Jakarta based on daily Covid-19 cases using the DTW distance. The calculation of the distance between two districts produces a 44 × 44 distance matrix which will be used to perform hierarchical agglomerative clustering. Based on this matrix, two districts with the minimum distance will be grouped into one temporary cluster. Then, the distance matrix is evaluated using the average linkage method. This procedure is continued until 44 districts are in one cluster, and a dendrogram is formed as hierarchical clustering.

Table 2 Silhouette Coefficient Interpretation

Silhouette Coefficient	Interpretation
0,71−1,00	Strong cluster structure
0,51−0,70	Reasonable cluster structure
0,26−0,50	Weak cluster structure
0−0,25	No meaningful cluster structure

Furthermore, the DTW measurements will be evaluated through the cophenetic correlation coefficient. The closer it is to one, the better the distance measurement method used will be. The following procedure is to determine the optimum number of clusters based on Kaufman's interpretation of the silhouette coefficients and calculate the 95% confidence interval through the bootstrapping procedure. Bootstrap is a method that counts on different approaches to estimate sampling distribution without knowing the underlying observations. The bootstrap method replaces the random samples from the original data, and the interested parameter estimation will be computed (Phuenaree & Sanorsap, 2017). The 44 districts will be grouped into *k* optimum clusters, and each cluster will be interpreted.

The second stage is forecasting. Forecasting is based on the clustering results in the previous stage, so the model used in this forecasting process will be referred to as a cluster-based model. The data used for forecasting is cluster representative data calculated using the prototype, which is the average of the members of each cluster. As an illustration, if cluster A has five members, the representative data for cluster A is calculated based on the daily average of the five members of cluster A. The modeling stages in the research follow the ARIMA modeling stage. It begins with model identification to obtain several tentative models. The model identification stage includes checking the data stationarity in the variance and mean. If the data are not stationary in the variance, they need to be transformed. Meanwhile, the differencing is done if the data are not stationary in the mean.

Based on several tentative models, the best model is selected based on the complexity and the smallest Akaike's Information Criterion (AIC) value. AIC is a mathematical method for evaluating how well a model fits the generated data. If a model with a low AIC has a high complexity, the best model chosen has a lower complexity even though the AIC value is higher than others. Furthermore, the best-selected model will be tested for feasibility by testing the model's residual, namely autocorrelation and normality of the residual. The statistics used to evaluate these two things are L-Jung Box and Jarque Bera. Each of them is a statistic for autocorrelation and the normality test of the model residual. After confirming that the best model is a feasible model, the next step is to make predictions at the cluster level. The goodness of the model in forecasting is measured through the value of MAPE (Montgomery, Jennings, & Kulahci, 2016).

Next, the clustering process is evaluated by comparing the predicted values at the city level. Comparisons are made through another ARIMA model that is built without going through the clustering results. Thus, the predicted data at the cluster level will be aggregated at the city level based on the district's origin in each cluster.

III. RESULTS AND DISCUSSIONS

Cumulatively, the distribution of positive Covid-19 data in each district in the initial period of the pandemic moves in a similar range of numbers. The difference in the number of daily COVID-19 cases between districts has begun to be seen since August 2020. The massive spread of the Covid-19 virus increasingly shows differences in the daily patterns in each district. For example, Duren Sawit district has the highest cumulative number until the end of the observation period, reaching 13.500 cumulative cases. In comparison, the district with the lowest cumulative number is Kepulauan Seribu Selatan, with 240 cumulative cases.

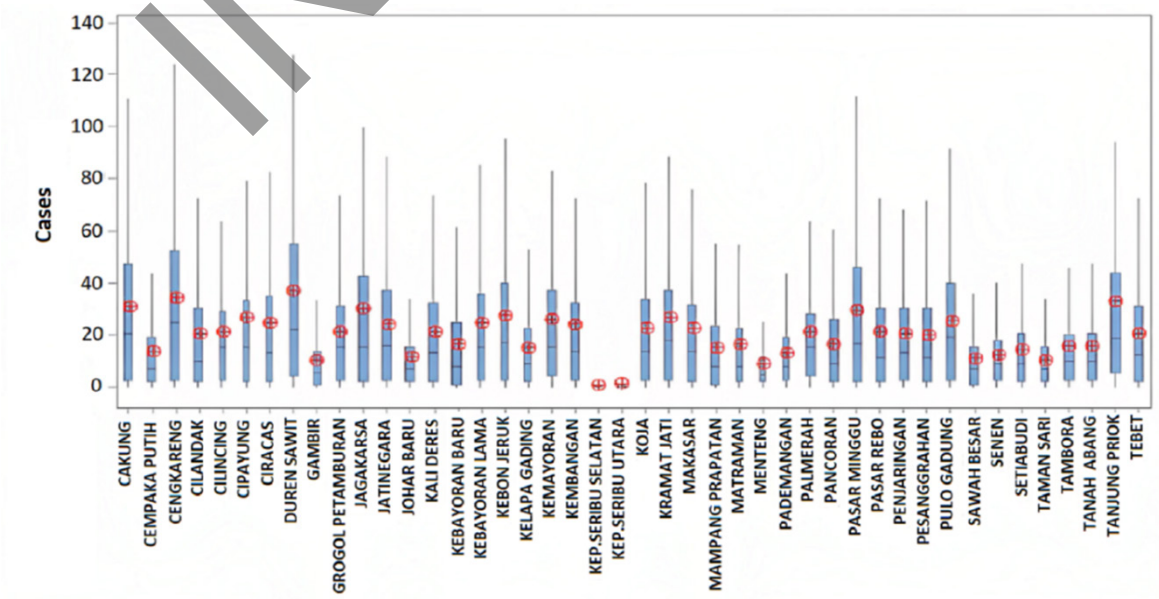


Figure 4 Box-Plot for Covid-19 Daily Cases in DKI Jakarta per District

Based on daily positive data per district, data distribution for all districts has a negative skew or skewed tendency to the left. This condition means that the frequency of daily positive numbers in each district tends to accumulate in high positive numbers. Based on Figure 4, the highest average daily positive case is in Duren Sawit district, followed by the Cengkareng district, with an average difference of about three cases per day.

Time-series clustering groups districts in DKI Jakarta based on the similarity in the distribution of Covid-19 cases in each region. Each district with the same distribution pattern will be grouped in one cluster, so the clusters will be heterogeneous (Sulastri, Usman, & Syafitri, 2021). The analysis results show that the value of the cophenetic coefficient for the DTW distance used is 0,93. This value indicates that the clustering results are pretty good (Carvalho, Munita, & Lapolli, 2019).

The optimal number of clusters in the silhouette coefficient is measured. It produces the highest silhouette coefficient value, which is 0,94, for a total of 6 clusters. Based on Kaufman's subjective interpretation of the silhouette coefficient, a silhouette coefficient close to one indicates a strong cluster structure. In addition, a bootstrapping procedure is also carried out to determine the confidence interval for the number of optimal clusters. The bootstrapping procedure is done by 25 times of resampling the Covid-19 data in different ranges of date and calculating the silhouette coefficient from each dataset, and a 5% confidence interval is obtained ($5 \leq k \leq 7$)

Based on the optimal cluster criteria and the confidence interval obtained, it determines that the number of clusters used is six, so the districts are distributed into six different clusters. Table 3 shows the members of each group formed.

Clustering time series data can be implemented based on several things, such as clustering based on certain statistical features, data-based, model-based, and others. In the research, the clustering process in 44 districts is carried out based on the similarity of the original data (data-based clustering). Identifying the 6 clusters formed can be done based on the similarity in the number of positive cases of Covid-19 in each district. Figure 5 shows the pattern of cumulative cases in each cluster.

Cluster A has the highest average daily cases by having the highest cumulative number in April 2021, reaching 13.000 cases. On the other hand, cluster F, consisting of two districts from the Kepulauan Seribu region, has the fewest daily cases, about 200 and 300 cases for Kepulauan Seribu Selatan and Kepulauan Seribu Utara, respectively. Next, cluster B has a case distribution of 6.000 cases at the end of the observation period. Meanwhile, clusters C and D have a cumulative total distribution of 8.000 and 10.000 cases. Next, cluster E has the lowest distribution rate outside Kepulauan Seribu, with around 4.000 cases.

Further identification is carried out by grouping the six clusters' results based on six regions that geographically divide the districts in DKI Jakarta. The results are shown in Figure 6. In general, all cities and regencies in DKI Jakarta are divided into four clusters, except for the Kepulauan Seribu district, which forms its cluster, namely cluster F. The result is obtained by returning each group member to their original area according to six administrative cities in DKI Jakarta. As an illustration, suppose that East Jakarta consists of 10 sub-districts, and it turns out that the ten sub-districts are spread into four different clusters.

The clusters that provide the most characteristics for the geographical area are clusters B, E, and F. There are five members of cluster B in the South Jakarta area. In contrast, in clusters E and F, it can be seen that all members are from Central Jakarta and Kepulauan Seribu, respectively. Other geographical characteristics can be identified in cluster A. The cluster tends to be on the outskirts of the city or district bordering areas outside DKI Jakarta.

Table 3 Cluster Members Recapitulation

Cluster	Number of Members	Members
A	6	Cakung, Cengkareng, Duren Sawit, Jagakarsa, Pasar Minggu, Tanjung Priok
B	11	Cempaka Putih, Kebayoran Baru, Mampang Prapatan, Matraman, Kelapa Gading, Pademangan, Pancoran, Setia Budi, Tambora, Tanah Abang, Tebet
C	7	Cilandak, Cilincing, Grogol Petamburan, Kali Deres, Palmerah, Penjaringan, Pesanggrahan
D	11	Cipayung, Ciracas, Jatinegara, Kebayoran Lama, Kebon Jeruk, Kembangan, Koja, Kramat Jati, Makasar, Pasar Rebo, Pulo Gadung
E	7	Gambir, Johar Baru, Kemayoran, Menteng, Sawah Besar, Senen, Taman Sari
F	2	Kepulauan Seribu Utara, Kepulauan Seribu Selatan

Moreover, modeling at the cluster level (cluster-based model) is based on the clustering results. It results in six clusters. The modeling is done by applying the ARIMA model to each cluster with time-series data for each cluster represented by the average of the members of each cluster or known as a prototype (Fontes, Santos, Embiruçu, & Aragão, 2021). Through this procedure, six ARIMA models are formed, which are used to forecast the number of Covid-19 cases in DKI Jakarta at the cluster level. The modeling stage begins by dividing the data into two parts: training data and testing data. The best period for modeling as cross-validation results is July 2020 to April 2021, with the last month of data used as testing data.

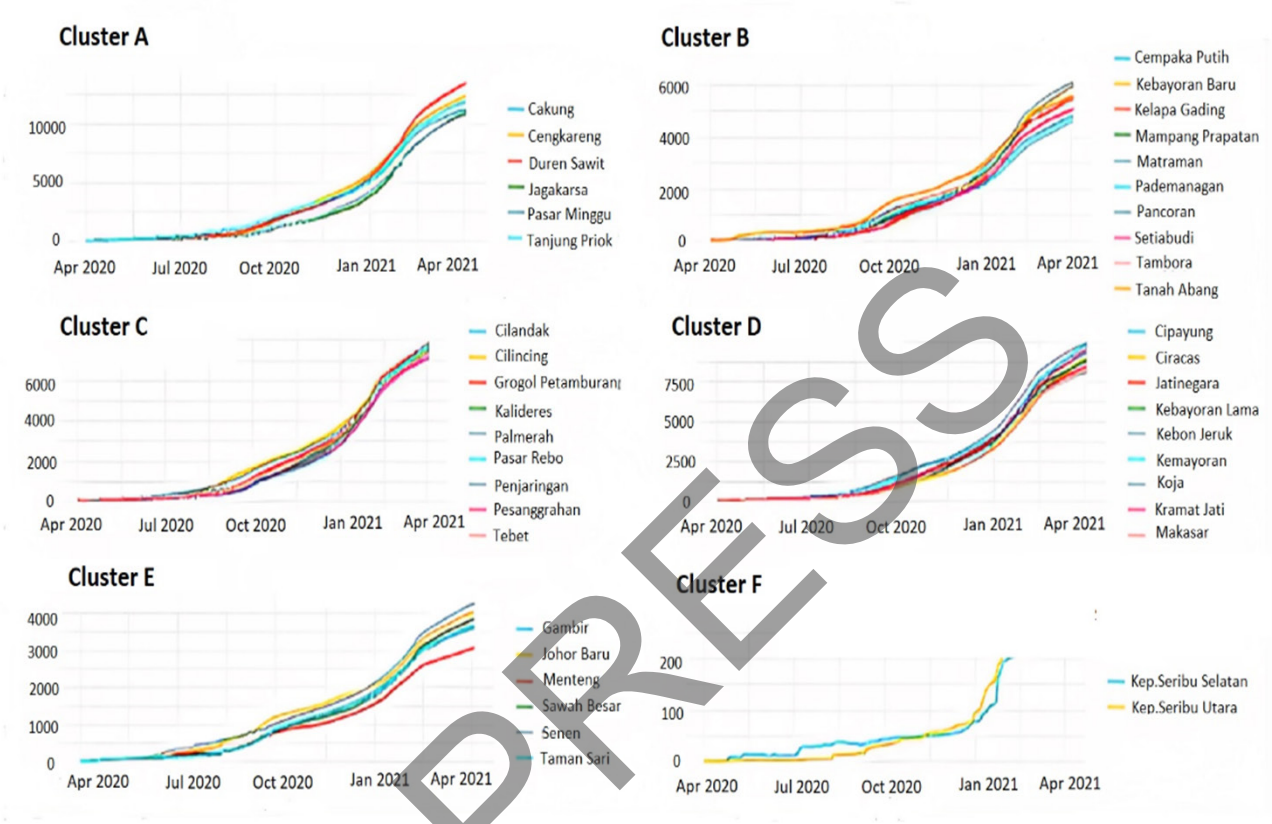


Figure 5 The Pattern of Cumulative Cases of Covid-19 Based on Cluster

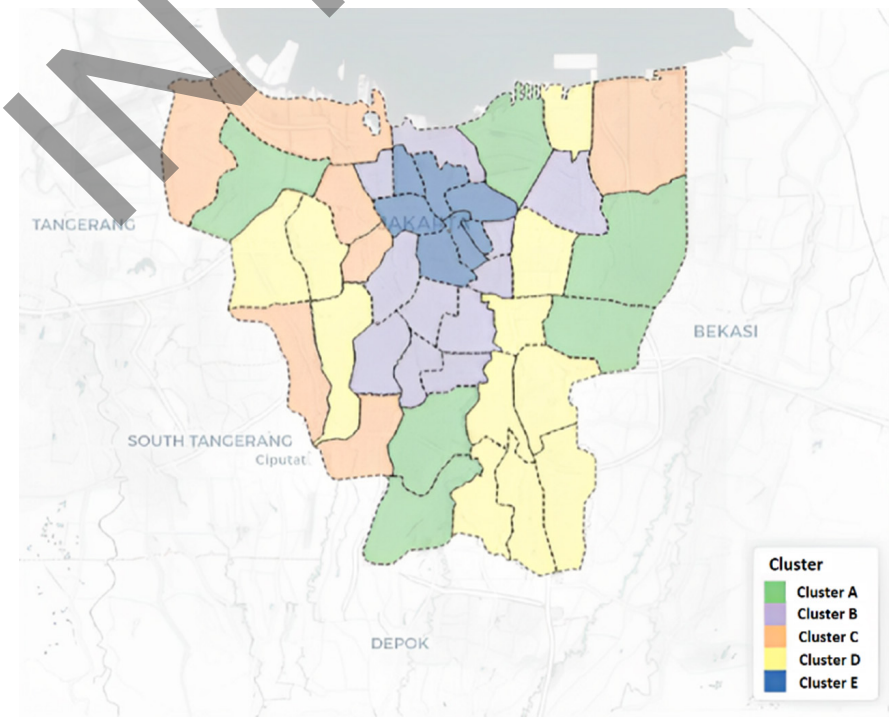


Figure 6 Distribution Mapping of Covid-19 Cases by Cluster

The highly fluctuated daily data indicates the need for handling the problem of data stationery in variance. The handling step before modeling is done by applying the Box-Cox transformation to the data based on the optimum value for each cluster (Yati, Devianto, & Asdi, 2013). The time series plot for the transformation results for each cluster is shown in Figure 7.

Furthermore, the transformed data are modeled using the ARIMA model, and the best model is obtained, as presented in Table 4. The significant values for the L-Jung Box and Jarque Bera tests in Table 4 indicate no autocorrelation in the model remainder, and the model residuals are normally distributed. The MAPE value shows the percentage of model error in predicting Covid-19 cases at the cluster level. The smaller the MAPE value is, the better the accuracy of the model is in making predictions. Meanwhile, the limit of the MAPE value for a model categorized as good in making predictions is 20% (Hämäläinen, Jauhiainen, & Kärkkäinen, 2017).

Evaluation of the performance of the ARIMA model as a result of clustering in predicting Covid-19 cases is carried out by comparing it to the ARIMA model formed without going through the clustering process at the city levels (non-cluster-based model). The predictive value of the clustering model is grouped by the city from which each district belongs to the cluster. The best models for the six cities are ARIMA (5,1,5) for East Jakarta, ARIMA (6,1,6) for West Jakarta, ARIMA (3,1,6) for Central Jakarta, ARIMA (4,1,3) for South Jakarta, ARIMA (2,1,5) for North Jakarta, and ARIMA (4,1,5) for Kepulauan Seribu. A comparison of accuracy values between the two groups of models is measured through the MAPE measure, as presented in Table 5.

Table 5 provides information that non-cluster models or models produced without going through the clustering process have better performance with lower MAPE values for all regions. This condition is reasonable, considering that the cluster model is obtained by estimating the mean values of the cluster members. The result is interesting because the MAPE difference between the two model groups is below 10%, except for the Kepulauan Seribu region. It shows that the performance of the clustering model is not much different from the model without clustering, so the clustering procedure can be said to be effective

Table 4 Summary of the Best ARIMA Models for Each Cluster

Cluster	Best Model	P-Value		MAPE
		L-Jung Box	Jarque Bera	
A	ARIMA(5,1,6)	0,24	0,100	12,64%
B	ARIMA(3,1,2)	0,15	0,100	15,60%
C	ARIMA(5,1,5)	0,15	0,100	10,48%
D	ARIMA(4,1,6)	0,10	0,100	16,19%
E	ARIMA(5,1,3)	0,29	0,100	21,92%
F	ARIMA(4,1,9)	0,07	0,175	-

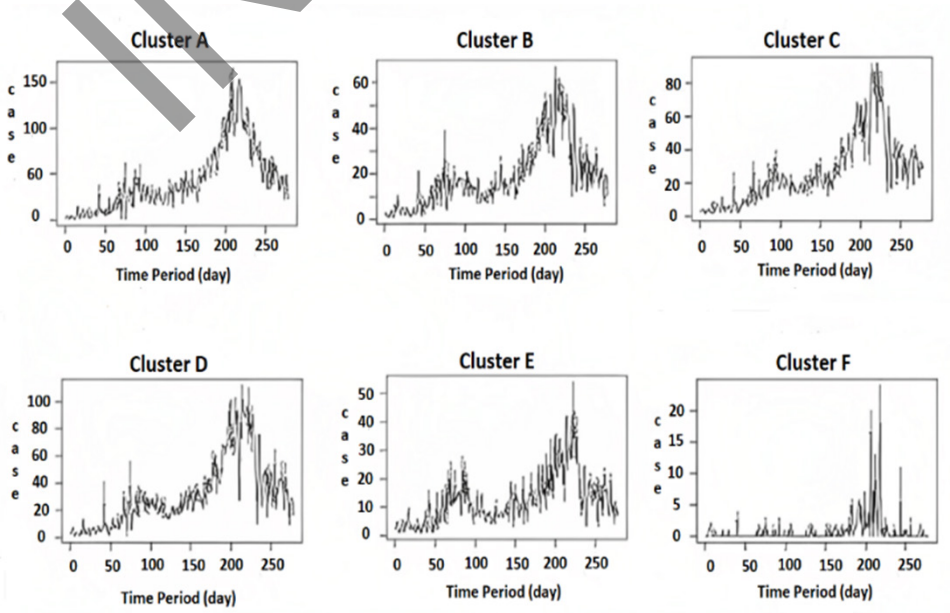


Figure 7 Time Series Plot for Transformation

with reasonable accuracy. Kepulauan Seribu region has a fairly large MAPE because the actual value in Kepulauan Seribu is very small, so an error in predicting one positive number will contribute to a huge error. The comparison of the 14-day prediction plot between the non-cluster model and the cluster model against the actual data is presented in Figure 8.

Based on Figure 8, the prediction period for Kepulauan Seribu looks different compared to other regions. Predictions for the Kepulauan Seribu region are carried out for the next 28 days and are measured every four days. This procedure is implemented as a strategy so that the MAPE value for Kepulauan Seribu can still be calculated and compared. Although the MAPE value for Kepulauan Seribu is quite large, it is very reasonable because the daily cases in Kepulauan Seribu are very low. However, the prediction performance of the two models can still be said to be good, especially the performance of the clustering results.

Table 5 Performance Comparison Between Non-Cluster-Based Models and Cluster-Based Models

Region	MAPE		Differences
	Non-Cluster	Cluster-Based	
East Jakarta	13,76%	18,25%	4,49%
West Jakarta	12,26%	17,67%	5,41%
Central Jakarta	14,28%	19,59%	5,31%
South Jakarta	10,45%	16,77%	6,32%
North Jakarta	11,14%	17,14%	6,00%
Kepulauan Seribu	24,76%	56,67%	31,89%

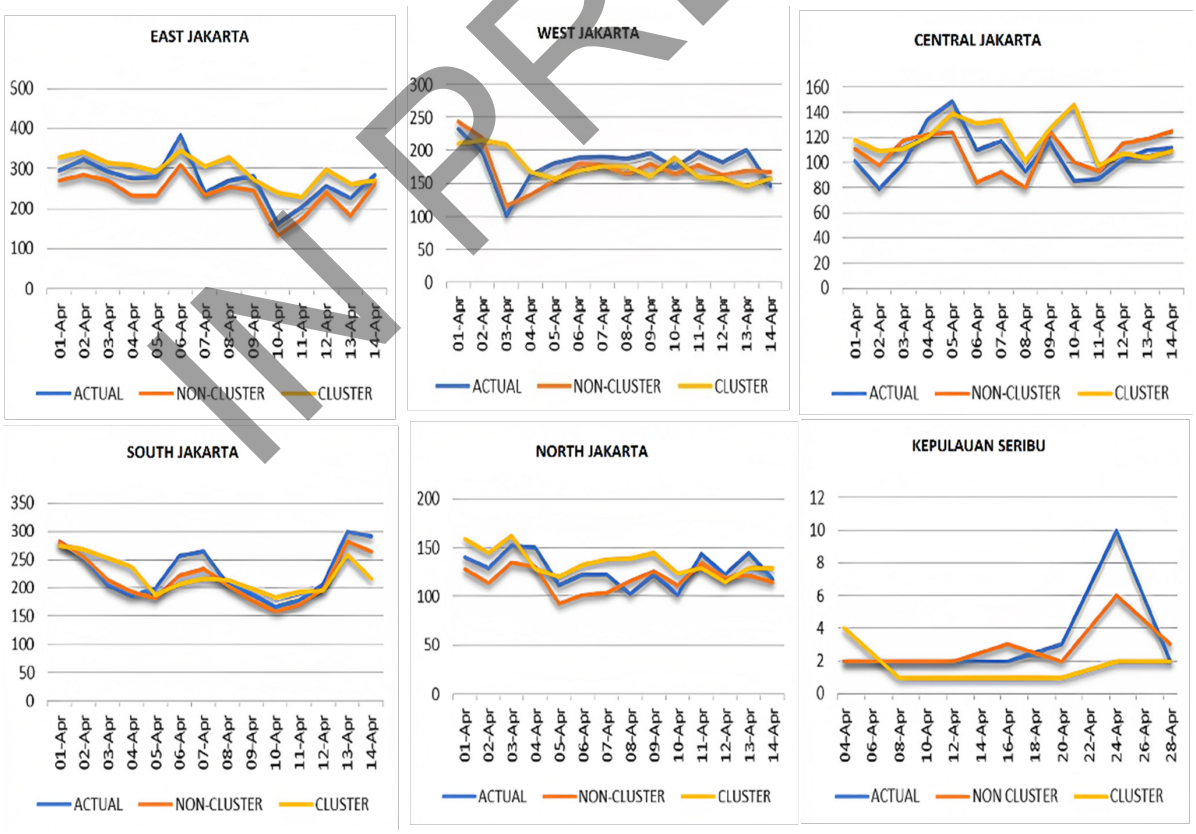


Figure 8 Time Series Plots for 14 Days Forecasting

IV. CONCLUSIONS

The distribution pattern of Covid-19 cases in DKI Jakarta is very massive, such that not a single district is free from the spread of this virus. The high and low distribution of daily Covid-19 cases in each district makes the districts in DKI Jakarta can be clustered into six optimal clusters. Clustering procedure was done analytically using the hierarchical agglomerative clustering method through DTW distance.

Some of the characteristics of each cluster have been interpreted interestingly. Cluster A records the highest distribution of Covid-19 cases. The members in this cluster tend to be on the border between DKI Jakarta and other areas. Meanwhile, the cluster with the lowest daily case distribution is cluster F, a cluster that contains only the Kepulauan Seribu region. Other geographical characteristics recorded are in clusters B and E. Cluster B members tend to come from South Jakarta, while cluster E are districts from Central Jakarta.

Modeling at the cluster level yields excellent performance with MAPE ranging from 10% to 20%. Likewise, the evaluation of clustering results is carried out through comparisons to non-cluster-based models. The modeling performance on the results of clustering aggregated to the city level has a MAPE difference of less than 10% compared to the non-cluster-based model. It means that the clustered data has a strong structure to represent individual district data. It will undoubtedly be very useful for related parties in formulating policies for handling Covid-19 in DKI Jakarta based on differences in the distribution in each district.

However, the scope of the research is limited by two things. The first one is the range of time used in the research. It is limited to April 2020−April 2021. The second is the distance method, which is limited to the DTW method only. For future research, the use of a different and longer time for the dataset can be done since the Covid-19 data is time series data that are updated daily. In addition, the distance method can be modified and compared to another method to find the characteristics and the most suitable distance method to handle similar data like the Covid-19 dataset. Future research can also be conducted in another province in Indonesia.

REFERENCES

Atique, S., Noureen, S., Roy, V., Subburaj, V., Bayne, S., & Macfie, J. (2019). Forecasting of total daily solar energy generation using ARIMA: A case study. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 114–119). IEEE. https://doi.org/10.1109/CCWC.2019.8666481

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*(April), 1–4. https://doi.org/10.1016/j.dib.2020.105340

Carvalho, P. R., Munita, C. S., & Lapolli, A. L. (2019). Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient. *Brazilian Journal of Radiation Sciences*, *7*(2A), 1–14. https://doi.org/10.15392/bjrs.v7i2a.668

Dong, G., & Liu, H. (Eds). (2018). *Feature engineering for machine learning and data analytics*. CRC Press.

Fontes, C. H., Santos, I. C., Embiruçu, M., & Aragão, P. (2021). Pattern reconciliation: A new approach involving constrained clustering of time series. *Computers & Chemical Engineering*, *145*(February), 1–23. https://doi.org/10.1016/j.compchemeng.2020.107169

Fransiska, H. (2021). Clustering provinces in Indonesia based on daily COVID-19 cases. *Journal of Physics: Conference Series*, *1863*, 1–9. https://doi.org/10.1088/1742-6596/1863/1/012015

Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms*, *10*(3), 1–14. https://doi.org/10.3390/a10030105

Han, T., Peng, Q., Zhu, Z., Shen, Y., Huang, H., & Abid, N. N. (2020). A pattern representation of stock time series based on DTW. *Physica A: Statistical Mechanics and Its Applications*, *550*(July), 1–12. https://doi.org/10.1016/j.physa.2020.124161

Johns Hopkins University. (2022). *Coronavirus resource center*. Retrieved from https://coronavirus.jhu.edu/

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, Inc.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2016). *Introduction to time series analysis and forecasting*. Wiley.

Novidianto, R., & Dani, A. T. R. (2020). Analisis klaster kasus aktif COVID-19 menurut provinsi di Indonesia berdasarkan data deret waktu. *Jurnal Aplikasi Statistika dan Komputasi Statistik*, *12*(2), 15–24. https://doi.org/10.34123/jurnalasks.v12i2.280

Pangaribuan, M. T., & Munandar, A. I. (2021). Kebijakan pemerintah DKI Jakarta menangani pandemi Covid-19. *Government: Jurnal Ilmu Pemerintahan*, *14*(1), 1–9.

Phuenaree, B., & Sanorsap, S. (2017). An interval estimation of Pearson's correlation coefficient by bootstrap methods. *Asian Journal of Applied Sciences, 05*(03), 623–627.

Puspita, P. E., & Zulkarnain. (2020). A practical evaluation of dynamic time warping in financial time series clustering. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 61–68). IEEE. https://doi.org/10.1109/ICACSIS51025.2020.9263123

Řezanková, H. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. In *21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics* (pp. 1–10).

Sammour, M., Othman, Z. A., Rus, A. M. M., & Mohamed, R. (2019). Modified dynamic time warping for hierarchical clustering. *International Journal on Advanced Science, Engineering and Information Technology*, *9*(5), 1481–1487. https://doi.org/10.18517/ijaseit.9.5.7079

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1394–1401). IEEE. https://doi.org/10.1109/ICMLA.2018.00227

Solichin, A., & Khairunnisa, K. (2020). Klasterisasi persebaran virus Corona (COVID-19) di DKI Jakarta menggunakan metode k-means. *Fountain of Informatics Journal*, *5*(2), 52–59.

Sulastri, S., Usman, L., & Syafitri, U. D. (2021). K-prototypes algorithm for clustering schools based on the student admission data in IPB University. *Indonesian Journal of Statistics and Its Applications*. *5*(2), 228–242. https://doi.org/10.29244/ijsa.v5i2p228-242

Wan, Y., Chen, X. L., & Shi, Y. (2017). Adaptive cost dynamic time warping distance in time series analysis for classification. *Journal of Computational and Applied Mathematics*, *319*(August), 514–520. https://doi.org/10.1016/j.cam.2017.01.004

Wang, W., Lyu, G., Shi, Y., & Liang, X. (2018). Time series clustering based on dynamic time warping. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 487–490). https://doi.org/10.1109/ICSESS.2018.8663857

Wiguna, H., Nugraha, Y., Rizka, F., Andika, A., Kanggrawan, J. I., & Suherman, A. L. (2020). Kebijakan berbasis data: Analisis dan prediksi penyebaran COVID-19 di Jakarta dengan metode Autoregressive Integrated Moving Average (ARIMA). *Jurnal Sistem Cerdas*, *3*(2), 74–83.

Yati, E., Devianto, D., & Asdi, Y. (2013). Transformasi Box-Cox pada analisis regresi linier sederhana. *Jurnal Matematika UNAND*, *2*(2), 115–122. https://doi.org/10.25077/jmu.2.2.115-122.2013