International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Comparison Study of Distance Measures in Nonlinear Panel Data Clustering with K-Means Method

Muayyad[a], Indahwati[b*], Kusman Sadik[c]

[a,b,c]*IPB University, Jl. Raya Dramaga, Babakan, Dramaga District, Bogor City, West Java, Indonesia.*

[a]*Email: muayyad_muayyad@apps.ipb.ac.id,* [b]*Email: indahwati@apps.ipb.ac.id,* [c]*Email: kusmans@apps.ipb.ac.id*

Abstract

Cluster analysis is used to group objects based on the similarity of characteristics between objects. Cluster analysis is usually applied to cross-sectional data, but in this study, cluster analysis was applied to nonlinear panel data using the k-means method. The selection of the right distance measure affects the optimization of clustering. The data used in this study are simulation data and real data, so the first stage of the research was carried out by simulating data to obtain the best distance measure on nonlinear panel data. The distance measure used is Euclidean, Manhattan, Maximum, Fréchet, and Dynamic Time Warping (DTW). Based on the evaluation results of all simulation data scenarios, it can be concluded that if the data objects do not overlap and have a long time span, it is better to use the maximum distance measure. If the data objects overlap and have a short time span, then we recommend using the DTW or Fréchet distance measure. Furthermore, the implementation is carried out on Indonesian Coronavirus Disease (COVID-19) data with the aim of grouping Provinces based on the number of active positive cases. The results show that the number of clusters is optimal when three clusters are formed, with the value of the Calinski Harabatz (CH) criteria of 143,459. Cluster A consists of 30 provinces, Cluster B consists of three provinces, while Cluster C consists of one province, DKI Jakarta Province.

Keywords: Coronavirus Disease; Calinski-Harabatz; Dynamic Time Warping; Frechet; Maximum; K-means.

---

* Corresponding author.

1. Introduction

Cluster analysis is one of the techniques in the multivariate statistical analysis used to group objects based on the similarity of characteristics between objects. Objects in one group have similar characteristics compared to objects in other groups [1]. Cluster analysis is divided into two types, hierarchical cluster analysis methods, and non-hierarchical cluster analysis methods. The hierarchical cluster analysis method is usually used on data with few objects, while the non-hierarchical cluster analysis method is usually used on data with more objects of observation [2]. This study applies the non-hierarchical cluster analysis method, which is most often used in cluster analysis using the K-Means method.

In cluster analysis, one thing that needs to be considered is how to determine the best distance measure to form an optimum cluster. The distance measure in cluster analysis consists of the Euclidean, Manhattan, Maximum, Minkowski, Fréchet, Mahalanobis, Dynamic Time Warping (DTW), and other distance measures. The determination of the distance measure should adjust to the pattern or nature of the data used. Therefore, applying the appropriate distance measure according to the nature of the data used will form the optimum cluster. The nature of the data used in cluster analysis greatly influences the clustering characteristics. The nature of the data that is often encountered in the analysis is a cross-section and time series. Suppose the data used contains elements of the object of observation and elements of time. In that case, the nature of the data is called longitudinal data. Longitudinal data properties that have the same time between variables are called panel data. In general, cluster analysis is usually applied to cross-section data. In this study, the cluster analysis method was applied to panel data.

Research conducted by Reference [3] provides an implementation of k-means designed to work specifically on paths (KML) or shared paths (KML3D). The study used "Minkowski Distance" to calculate the distance used in cluster analysis on longitudinal data. Then Reference [4] used the Fréchet and Dynamic Time Warping (DTW) distance measures for longitudinal clustering data (time series) based on the shape of the curve. In subsequent research, Reference [5] conducted clustering panel data using the k-means method based on the values of the GDP growth variables and population growth of 29 countries in Europe for the period 1990 to 2017. Then research conducted by Reference [6] examines several distance measures for panel data in non-hierarchical clustering using the k-means method and the Manhattan distance measure used in clustering the human development index (HDI) data for the 2010 to 2019 time with a linear pattern or curve model.

Some of the studies mentioned applied different distance measurements when swarming. However, some of these studies have not discussed the selection of the appropriate measure of similarity or distance measure if the curve model is not linear or the panel data model is nonlinear. Nonlinear panel data is the nature of the data on the curve, which in the graph has a slope value or data pattern that is not linear based on time. Therefore, this study examines the application of the similarity measure or the best distance measure in the analysis of clustering nonlinear panel data. The application of cluster analysis in this study was used to group provinces in Indonesia based on Coronavirus Disease (COVID-19) data. The observation in this study was from March 2020 to December 2020.

2. Methodology

2.1. Data

There are two data used in this research, simulation data and real data. This study uses real data obtained from the web http://covid.go.id with data used COVID-19 data for each province in Indonesia with an observation time from March 2020 to December 2020 with the object of observation being each province in Indonesia with various components that will be used weekly positive cases.

2.2. Simulation Data

This study uses the R programming language to perform simulations on the data. The criteria for the simulation data in this study pay attention to the following procedures:

1. Determine the number (g) of clusters formed on nonlinear panel data with g=3 clusters.
2. Generating nonlinear panel data with (n) observation objects, (t) time period, and $f_g(t)$ is a nonlinear function.

a. Determine the number of objects of observation (n), with n = 10 objects of observation,
b. Determine the length of the time period with t = 1, 2,…,10 , t = 1,2,…,20 and t = 1,2,…,40.
c. Generating nonlinear panel data according to the number of clusters with a nonlinear function $f_g(t)$. $f_g(t)$ is a nonlinear function defined by
d. $f_a(t)$ as a function of the first cluster, $f_b(t)$ is a function of the second cluster, $f_c(t)$ is a function of the third cluster.

Table 1: Simulation Model Functions.

Model	Time	Function
Sinusoids	10	$f_a(t) = 10 \sin t;$ $f_b(t) = 7 \sin t;$ $f_c(t) = 4 \sin t$
	20	
	40	
Normal	10	$f_a(t) = N(t, t/2, 2.5) \times 60;$ $f_b(t) = N(t, t/2, 2.5) \times 40;$ $f_c(t) = N(t, t/2, 2.5) \times 20$
	20	
	40	
Exponent 1	10	$f_a(t) = \exp(0.09t);$ $f_b(t) = \exp(0.07t);$ $f_c(t) = \exp(0.05t)$
	20	
	40	
Exponent 2	10	$f_a(t) = 8 + \exp(0.1t);$ $f_b(t) = 5 + \exp(0.1t);$ $f_c(t) = 2 + \exp(0.1t)$
	20	
	40	

3. Generating residual models with Personal Variation (PV) and Residual Variation (RV). Personal Variation (PV) is the error between the observed objects in the cluster. Residual Variation (RV) is the error on each path

in the cluster formed or the error between the observed objects with time. First, generate PV and RV with a normal distribution. Second, the middle value used is 0 (zero). Thirdly, there are three variants used, namely ($\sigma^2 = 0.5$), ($\sigma^2 = 1$), and ($\sigma^2 = 2$); and the value of variance PV > RV. So, we get PV and RV: (PV=N(0,1); RV=N(0,0.5)), (PV=N(0,2); RV=N(0,0.5)) and (PV =N(0,2); RV=N(0,1))

4. Performs 1000 repetitions on the generated data.

2.3. Simulation Data Analysis Procedures

This procedure aims to obtain the best distance measure applied in clustering nonlinear panel data with K-Means. The steps of analysis on the simulation data are as follows:

1. Generating simulation data of nonlinear panel data models.
2. Perform cluster analysis using the K-Means method, with the following procedure:

a. Calculate the centroid (mean) of each cluster;
b. Calculate the distance of the object of observation to the centroid (mean) of the cluster by using the Minkowski, Fréchet, and Dynamic Time Warping (DTW) distance measure.

- Minkowski distance measure

The formula for the Minkowski distance measure is written as follows:

$$d(y_{1..}, y_{2..}) = \left[\sum_{j=1}^{T}\sum_{k=1}^{M}\left|(y_{1jx} - y_{2jx})\right|^{p}\right]^{1/p}$$

y_{ijk} is a trajectory, where i is the index of the subject $1 \cdots n$, j is the time index $1 \cdots t$, and k is the index of the trajectory variable $1 \cdots m$. When $p = 1$ then the Minkowski Distance will be equal to Manhattan Distance, $p = 2$ then the result will be that the Minkowski Distance is equal to the Euclidean Distance and the Maximum Distance is obtained when the power is on $p = +\infty$ [3].

- Fréchet distance measure

The formula for the Fréchet distance measure is written as follows [7]

$$d_{Frechet}(y_1, y_2) = \min_{m \in M}\left(\max_{j = 1,2,\cdots t}\left|y_{1j\cdot} - y_{2j\cdot}\right|\right)$$

where, $m = (y_{11\cdot}, y_{21\cdot}), (y_{12\cdot}, y_{22\cdot}) \cdots (y_{1t\cdot}, y_{2t\cdot})$

- Dynamic Time Warping (DTW) distance measure

The formula for the Dynamic Time Warping (DTW) distance measure is written as follows [8]:

$$d_{DTW}(y_1, y_2) = \sum_{i=1}^{N}\sum_{j=1}^{T}\left|y_{1i} - y_{2j}\right| + min\begin{cases} dDTW(i-1,j) \\ dDTW(1,j-1) \\ dDTW(i-1,j-1) \end{cases}$$

where,

$$d_{DTW}(i-1,j) = d(y_{1(i-1),}y_{2j}) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left|y_{1(i-1)} - y_{2j}\right|$$

c. Clustering objects that have the closest distance to the cluster mean.
d. Repeat steps a to d so that no object of observation moves clusters.

3. Calculating the accuracy value of the Minkowski distance measure, Fréchet, and Dynamic Time Warping distance measure with the Confusion matrix to estimate objects included in the true and false groups [9].

$$Accuracy = \frac{TP}{number\ of\ data}$$

4. Determine the best distance measure between the Minkowski distance measure, Fréchet distance, and the DTW distance measure based on the greatest accuracy value with the evaluation criteria at point 3.

2.4. Real Data Procedure

This stage is the application of the clustering method with the best distance measure. The steps of analysis on real data are as follows:

1. Data exploration
2. Perform cluster analysis using the K-Means clustering method based on the best distance measure from the simulation results with the optimal number of clusters.
3. Evaluating the K-Means clustering method using the Calinski-Harabasz (CH) method. The CH validity index is formulated as follows [10]

$$CH = \frac{trace\ (SSB)}{trace\ (SSW)} \times \frac{N-k}{k-1}$$

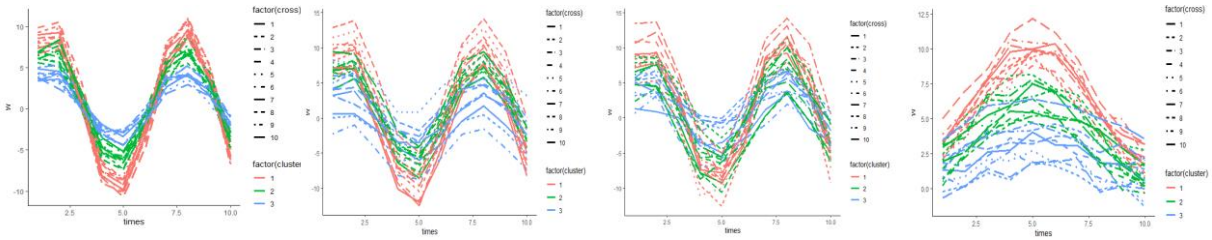
N = Number of objects, k = Number of k-clusters, Trace (SSB) = Trace value of the variance matrix between clusters, and Trace (SSW) = Trace value of the variance matrix within clusters.

4. Conclude the results of the evaluation of the best clustering method.

3. Result and Discussion

3.1. Exploration of Generated Data

First, an exploration of the generated data is carried out to ensure the generation data is in accordance with the scenario in the Stages of generating simulation data subsection. There are four general models of functions used in the simulation, namely the sinusoidal model, normal distribution, exponent 1, and exponent 2.

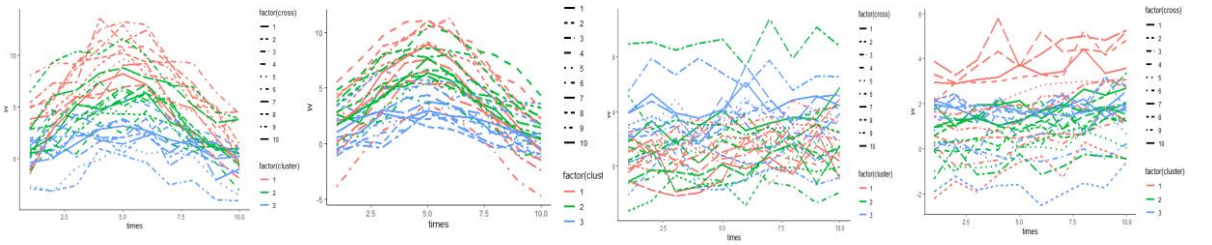


(A1). Sinusoidal Model PV=N(0,1);RV= N(0,0.5)

(A2). Sinusoidal Model PV=N(0,2); RV= N(0,0.5)

(A3). Sinusoidal Model PV=N(0,2); RV= N(0,1)

(B1). Normal distribution model PV=N(0,1);RV= N(0,0.5)

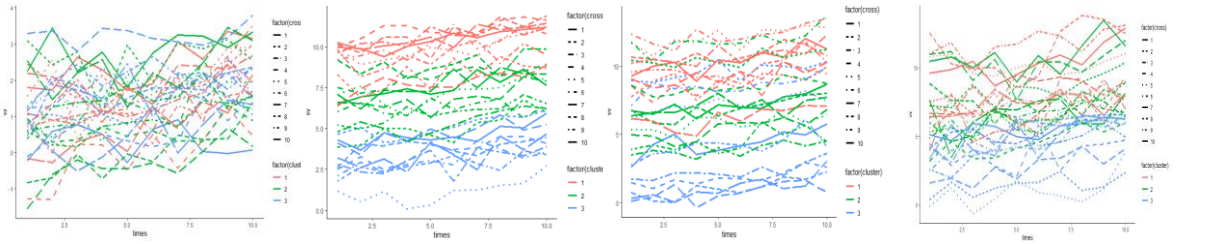


(B2). Normal distribution model PV=N(0,2);RV= N(0,0.5)

(B3). Normal distribution model PV=N(0,2);RV= N(0,1)

(C1). Exsponent 1 model PV=N(0,1);RV= N(0,0.5)

(C2). Exponent 1 model PV=N(0,2);RV= N(0,0.5)



(C3). Exponent 1 model PV=N(0,2);RV= N(0,1)

(D1). Exponent 2 model PV=N(0,1);RV= N(0,0.5)

(D2). Exponent 2 model PV=N(0,2);RV= N(0,0.5)

(D3). Exponent 2 model PV=N(0,2);RV= N(0,1)

Figure 1: Plot Simulation of Generation Data Scenario.

Figure 1 illustrates some of the scenarios generated by the sinusoidal model, normal distribution, exponent 1, and exponent 2. The data in each model is generated with 3 clusters. The data generated in nonlinear panel data is based on time, and there are ten objects of observation in the cluster. In Figures A1, B1, C1, and D1 generated by PV=N(0,1) and RV=N(0,0.5) for time 10, it is clearly seen between the clusters of images A2, B2, C2, and D2. In Figures 1b, 2b, 3b, and 4b generated with PV=N(0,2) and RV=N(0,1) for time 10, the generation data tend to overlap due to the increase in the value of the variance in PV. In Figures A3, B3, C3, and D3 generated

with PV=N(0,2) and RV=N(0,1) for time 10, the generation data tends to overlap due to the increase in the value of the variance in PV and object fluctuations due to increased variance in RV. In the sinusoidal model generation data, the longer the time used, the more waves will be formed. At times 20 and 40 formed three waves and six waves, respectively. The shorter the time used in the normal distribution model generation data, the wider the normal curve formed. In the generation data of the exponential model 1, when used at a time period of 10, it appears that the objects overlap each other; if a time period of 20 is used, the exponential pattern will be clearly visible, in the period of 40 the exponential pattern is clearer and between clusters formed. In the generation data of the exponential model 2, when used with a time period of 10, it appears that several objects overlap each other, if time periods of 20 and 40 are used, it is clear that the exponential pattern and between the clusters that coalesce at the end of the trajectory is clearly visible. Based on the generation data scenario, the complexity or overlap between clusters is strongly influenced by the PV variance value, and fluctuations in an object are influenced by the RV variance value. The greater the PV value used, the more complicated or overlapping between objects. As with PV, the greater the RV value used, the more fluctuations in the object or the higher the resulting wave.

3.2. Simulation Data Accuracy

The simulation data generated based on the scenario will be evaluated. The K-means method with the Manhattan, Euclidean, Maximum, Fréchet, and Dynamic Time Warping (DTW) distance measure will be used to cluster the data. The accuracy of these distances will be seen from the value of the clustering accuracy. The following will explain the evaluation of the accuracy of the distance measure on the simulation data for each scenario in each generation data model. In Table 2, there are nine scenarios for each generation data model based on time, PV, and RV. Table 2 shows that each scenario in the sinusoidal generation data model has the highest accuracy value in the maximum distance measure. The greatest accuracy value is when using PV=N(0,1) and RV=N(0,0.5) because the generation data is made clear between the clusters at any time (10, 20, and 40). Meanwhile, when increasing the value of the PV variance, the accuracy is close to 70% each time. The highest accuracy value is at the maximum distance in 7 scenarios in the normal generation data model. Meanwhile, in the other two scenarios, the DTW and frechet distance measure have the highest accuracy value when the time is 10 at PV=N(0,2) when the clusters overlap each other. The exponential 1 generation data model has the highest accuracy at the maximum distance in each scenario at time 40. Meanwhile, in other scenarios (times 10 and 20), the best distance measure is DTW, but the accuracy value is not much different from the maximum distance measure. In the exponential 2 generation data model, when PV and RV are increased, the best distance measure is DTW and Fréchet each time, but when PV=N(0,1) and RV=N(0,0.5), the best distance measure is the Manhattan and Euclidean distances. The Euclidean, Manhattan, and Maximum distance measure have accuracy values that are not much different/like the exponential 2 generation model. After completing all distance measure s with 36 scenarios from 4 generation data models, the maximum distance measure is the best distance measure with a total of 20 highest accuracy values compared to other distance measure. Table 2 shows that when the generation data tends to be separated between clusters, the accuracy value gets bigger in each time period (10, 20, and 40). Meanwhile, when the generation overlaps between clusters, the accuracy value decreases each time (10, 20, and 40). Based on the evaluation results of all scenarios, it can be concluded that if the data objects do not overlap and have a long period of time, it is better to use the maximum distance measure.

However, suppose the data objects overlap and have a short time. In that case, it is better to use the DTW or Fréchet distance measure.

Table 2: The results of the average accuracy of the distance measure for each scenario.

Model	Time	Variance		Accuracy				
		PV	RV	Euclidean	Manhattan	Maximum	Fréchet	DTW
Sinusoids	10	1	0.5	0.969	0.943	0.986	0.822	0.827
		2	0.5	0.679	0.665	0.699	0.654	0.654
		2	1	0.666	0.656	0.678	0.650	0.650
	20	1	0.5	0.985	0.964	0.993	0.833	0.839
		2	0.5	0.684	0.671	0.704	0.673	0.671
		2	1	0.682	0.675	0.696	0.672	0.674
	40	1	0.5	0.984	0.965	0.993	0.821	0.826
		2	0.5	0.682	0.668	0.704	0.666	0.665
		2	1	0.678	0.671	0.692	0.673	0.671
Normal	10	1	0.5	0.791	0.782	0.812	0.807	0.811
		2	0.5	0.576	0.567	0.591	0.592	0.596
		2	1	0.564	0.562	0.573	0.584	0.581
	20	1	0.5	0.728	0.709	0.770	0.706	0.710
		2	0.5	0.509	0.499	0.557	0.512	0.505
		2	1	0.517	0.512	0.540	0.525	0.525
	40	1	0.5	0.646	0.620	0.722	0.528	0.521
		2	0.5	0.463	0.453	0.550	0.439	0.438
		2	1	0.461	0.455	0.516	0.448	0.446
Exponent 1	10	1	0.5	0.431	0.430	0.433	0.437	0.440
		2	0.5	0.423	0.422	0.423	0.426	0.426
		2	1	0.424	0.425	0.425	0.427	0.429
	20	1	0.5	0.510	0.506	0.535	0.520	0.549
		2	0.5	0.444	0.443	0.456	0.448	0.449
		2	1	0.444	0.444	0.450	0.451	0.458
	40	1	0.5	0.995	0.984	1.000	0.904	0.898
		2	0.5	0.867	0.834	0.962	0.783	0.786
		2	1	0.867	0.846	0.944	0.797	0.819
Exponent 2	10	1	0.5	0.861	0.862	0.857	0.849	0.854
		2	0.5	0.659	0.660	0.659	0.660	0.663
		2	1	0.661	0.661	0.658	0.665	0.668
	20	1	0.5	0.889	0.889	0.887	0.878	0.877
		2	0.5	0.658	0.659	0.656	0.662	0.659
		2	1	0.656	0.656	0.653	0.665	0.663
	40	1	0.5	0.893	0.893	0.891	0.868	0.855
		2	0.5	0.659	0.659	0.657	0.657	0.650
		2	1	0.664	0.664	0.658	0.671	0.666

3.3. Real Data Exploration

Real data exploration is carried out at an early stage to find out the pattern of the data against time. The variable component consists of 34 objects (34 provinces) with a time period of 44 weeks from March to December 2020. The data pattern for each object of observation fluctuates with time or is like an exponential distribution pattern; for more details, see Figure 2 below.

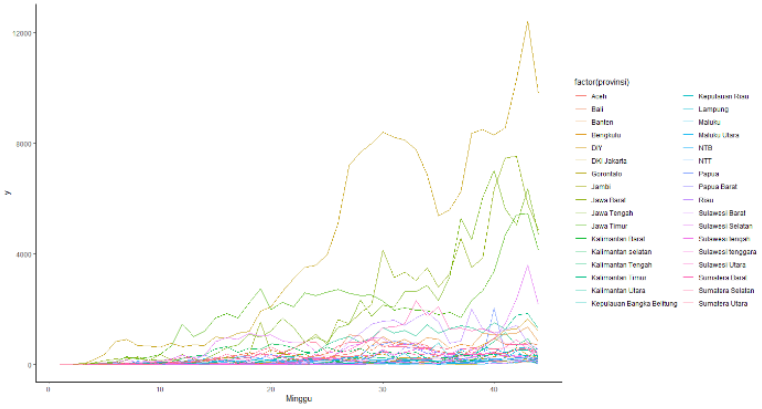


Figure 2: The data plot on the number of active cases for each province in Indonesia.

Table 3: Descriptive statistics.

Min	1st Qu.	2Median	Mean	3rd Qu.	Max
1941	3805.750	9207	21839	18134.5	183248

Table 3 above shows the lowest number of COVID-19 cases in 1941 (West Sulawesi) and the highest number of cases 183,248 (DKI Jakarta). Nationally, the mean number of COVID-19 cases in Indonesia is 21,839.

3.4. Real Data Result

Based on the simulation results, the K-Means method with the maximum distance measure is able to provide the best results in clustering nonlinear panel data if the data do not overlap each other. The clustering method is divided into two types of hierarchical and non-hierarchical clustering methods, K-Means method is a type of non-hierarchical clustering. Thus, at an early stage, it is necessary to determine the number of clusters [11]. The number of clusters tested was g = 2 to g = 10. The optimum number of clusters was obtained from the Calinski-Harabatz (CH) value.

Table 4: Calinski-Harabatz (CH) criteria values for each number of clusters.

Number of Clusters	2	**3**	4	5	6	7	8	9	10
CH	85.997	**143.549**	108.517	105.536	137.306	138.429	133.433	127.913	123.072

Table 4 shows a comparison of the values of the Calinski-Harabatz (CH) criteria for each number of clusters that have been tested. Table 4 shows that the highest CH value was obtained when three clusters were formed. Therefore, the optimum number of clusters for clustering data on the number of active COVID-19 cases in each province in Indonesia using the K-Means method and the maximum distance is when three clusters are formed.

3.5. Group Member Identification

The optimum clusters produced were three clusters based on the CH criteria value in the previous stage. Figure 3 shows the results of the clustering of provinces on the data on the number of active cases for each province in Indonesia. Provinces that are in one cluster have similar characteristics of the number of active COVID-19 cases every week, see Figure 3 below;

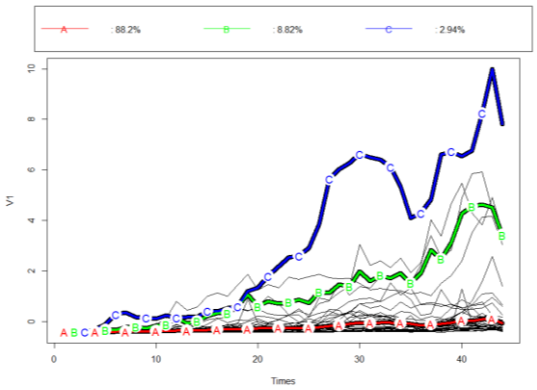


Figure 3: The Plot of clustering results for each province in Indonesia.

Figure 3 shows the results of clustering Provinces that have similar movements or characteristics of active COVID-19 case data for 44 weeks. In Figure 3, DKI Jakarta Province is separated from other provinces. Table 5 shows the provincial members belonging to each cluster.

Table 5: Details of the Members of each Cluster.

Clusters	Number of members	Cluster members
A	30	West Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, South Sumatra, West Papua, Papua, NTT, NTB, North Maluku, Maluku, Lampung, Riau Islands, Bangka Belitung islands, North Kalimantan, Central Kalimantan, South Kalimantan, West Kalimantan, Jambi, Gorontalo, Bengkulu, Aceh, West Sumatra, Riau, East Kalimantan, Yogyakarta, South Sulawesi, North Sumatra, Banten, and Bali.
B	3	West Java, Central Java and East Java.
C	1	DKI Jakarta

The table above shows the members of each cluster. Cluster A has the most members, which is 30 members. cluster B has three members. Meanwhile, cluster C has one member of DKI Jakarta Province.

4. Conclusion

The best distance measure is obtained based on the evaluation of all scenarios. The optimum distance measure for clustering nonlinear panel data using the K-Means method recommends the maximum distance measure and DTW. Maximum distance measure when objects do not overlap and have a long time period, while the DTW distance measure overlaps and has a short time period.

Based on the simulation data, the maximum distance measure is applied to the k-means method to group provinces in Indonesia because the objects do not overlap and have a long time period of 44 weeks. The application of the k-means method using the maximum distance measure on Covid-19 data with the variable used is the number of active Covid-19 cases in each province in Indonesia, the optimum number of clusters is three clusters with a CH value of 143,549. Cluster A has 30 members, Cluster B has three members, while Cluster C has one member from DKI Jakarta province. The plot of clustering results shows that DKI Jakarta Province is separated from other provinces.

5. Suggestions

In this study, only four models of generation data were used. Therefore, the next step is to generate data with several scenarios with various functions and can add several distance measures as a comparative study of accuracy. Then, on real data using more than one variable.

References

[1] Mattjik, AA, Sumertajaya, I,M. 2011. *Sidik Peubah Ganda dengan Menggunakan SAS*. Bogor: IPB Press.

[2] Sumertajaya IM, Erfiani, Putri WDY. 2007. Analisis Gerombol Menggunakan Metode Two Step Cluster (Studi kasus: data Potensi Desa Sensus Ekonomi 2003 wilayah Jawa Barat). *Forum Statistika dan Komputasi*. 12(1): 18-23.

[3] Genolini C, Alacoque X, Sentenac M, Arnaud C. 2015. KML and KML3D: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*. 65(4): 2-10. doi: 10.18637/jss.v065.i04.

[4] Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. 2016 kmlShape: An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes. *PLoS ONE* 11(6):1-12. doi: 10.1371/journal.pone.0150738

[5] Bilgic E, Baydar V. 2018. *Panel Data Clustering with R: An Application on Macroeconomic Variables of European Countries*. 19, 258.

[6] Sugiono, Adella Sari Cahyani. 2020. *Kajian Perbandingan Beberapa Jarak untuk Data Panel dalam Penggerombolan Tak Berhiraki* [Tesis]. Bogor: Institut Pertanian Bogor.

[7] Montero P, Villar Jose A. 2014. TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software*. 65 (4): 2-18.

[8] Liu L, Li W, Jia H. 2018. Method of Time Series Similarity Measurement Based on Dynamic Time Warping. *CMC*. 57(1):97-106.

[9] Gorunescu, F. 2011. *Data Mining: Concepts, Model and Techniques*. Berlin, Jerman: Springer.

[10] Baarsch J, Celebi ME. 2012. Investigation of Internal Validity Measures for K-Means Clustering. *International Multiconference Of Engineers And Computer Scientists.*1:14-16. LA: Louisiana Board of Regents.

[11] Johnson RA, Wichern DW. 2002. *Applied Multivariate Statistical Analysis 6th Edition*. New Jersey: Prentice-Hall International.