ISSN: 2442-6571

Clustering stationary and non-stationary time series based on autocorrelation distance of hierarchical and k-means algorithms

Mohammad Alfan Alfian Riyadi [a,1,*], Dian Sukma Pratiwi [b,c,2], Aldho Riski Irawan [a,3], Kartika Fithriasari [a,4]

[a] *Departement of Statistics Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[b] *Departement of Actuarial Science, Bandung, Indonesia*
[1] *m.alfan.alfian.riyadi@gmail.com *;* [2] *diansukmapratiwi@gmail.com;* [3] *ir.aldhoriski@gmail.com;*
[4] *kartika_f@statistics.its.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received August 7, 2017
Revised November 18, 2017
Accepted November 21, 2017

Keywords:
Autocorrelation Distance
Hierarchical Algorithm
K-Means Algorithm
Non-Stationary Time Series
Stationary Time Series

ABSTRACT

Observing large dimension time series could be time-consuming. One identification and classification approach is a time series clustering. This study aimed to compare the accuracy of two algorithms, hierarchical cluster and K-Means cluster, using ACF's distance for clustering stationary and non-stationary time series data. This research uses both simulation and real datasets. The simulation generates 7 stationary data models and another 7 of non-stationary data models. On the other hands, the real dataset is the daily temperature data in 34 cities in Indonesia. As a result, K-Means algorithm has the highest accuracy for both data models.

I. Introduction

Time series model has many methods. The most important process was at the first step when develop the time series model. The process is the parameter identification by using autocorrelation function (ACF) plot of either stationer or non-stationer time series data. When the trend of time series and ACF plots are slightly decreased, the time series data was not stationer [1], [2]. It steps easy and fast for the small data dimension. However, it may take longer time for gigantic size of data. So, the research problem is how to find a suitable method for identification large-scale time series data.

Identification and classification method for a large scale time series data has been done by clustering the time series data. This type of clustering is different with the clustering process for cross-section data, especially in deciding the distance technique for each cluster. Manso [3] created an R package for time series clustering with stationer data, used the package for time series analysis combined with clustering and proposed a time series clustering method. D'Urso and Maharaj [4] proposed a fuzzy clustering approach based on autocorrelation function, which applied on data which have a strong autocorrelation. In other words, the process of calculating distance value becomes complex and problematic due to availability in the enormous dataset.

Therefore, the main contribution of this research is finding the most accurate based on ACF's distance for stationary and non-stationary time series data by comparing the hierarchical clustering and K-Means algorithm. The hierarchical clustering technique makes the data into groups over Dendrogram, which is containing cluster tree in different scale [5]. As in differently, k-means clustering algorithm develops the groups based on cluster center which each cluster has member depend on closest fitness value, and the cluster centers will be updated until there is no change in any of the clusters centroids [6]. The hierarchical clustering and k-means clustering is widely used because its efficiency, scalability, and simplicity. The experiment was conducted on simulated data and real data sample to see the accuracy of both methods.

II. Method

A. Stationarity Model for Time Series

The process $\{Y_i\}$ fulfilled the stationary assumption if the joint distribution of $\{Y_{i_1}\},\{Y_{i_2}\},\ldots,\{Y_{i_n}\}$ has same characteristics with the joint distribution of $\{Y_{i_1-k}\},\{Y_{i_2-k}\},\ldots,\{Y_{i_n-k}\}$ for every time points $t_1, t_2, \ldots, t_n$, and all items of lag k. Thus, when the data are univariate so Y_t will be same as that of $Y_t - k$ for all t and k. It makes $E(Y_t) = E(Y_{t-k})$ for all t and k, based on that so the mean function is constant for all periods. Then, $Var(Y_t) = Var(Y_{t-k})$ for t and k, it means that the variance is also constant over the time. In stationary definition, the Y_t and Y_s have same value as that of Y_{t-k} and Y_{s-k} from which it follows that $Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k})$ for all t, s, and k [7], [8].

Modelling stationary time series are Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA). For the general MA(q) process,

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \tag{1}$$

Moving average arises when Y_t is obtained by applying the weight $1, -\theta_1, -\theta_2, \ldots, -\theta_q$ to the variable $e_1, e_{t-1}, e_{t-2}, \ldots, e_{t-q}$ and distributing the weights and trying them to $e_{t+1}, e_t, e_{t-1}, \ldots, e_{t-q+1}$ to obtain Y_{t+1} and so on. Autoregressive process are when Y_t has linear combination with Y_{t-k}. Specifically, a path-order AR Process $\{Y_t\}$ satisfies the equation.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t \tag{2}$$

The current values of the Y_t is a linear combination of the past values of itself plus a random variables mentioned as e, as the variable which represent the other factors not explained by the model.

Assume that AR and MA, obtain a quite general time series model.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \tag{3}$$

$\{Y_t\}$ is a mixed ARMA process of orders p and q.

B. Non-Stationary Model of Mean

Time series model called non-stationary of mean if $E(Y_t) \neq E(Y_{t-k})$, and non-stationary of variance if $Var(Y_t) \neq Var(Y_{t-k})$. This paper focused on non-stationary model of mean. One of non-stationary model, in this case, is Autoregressive Integrated Moving Average (ARIMA) [9], [10]. To get a stationary model from non-stationary model, we should be differencing method. The model of ARIMA is,

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B) a_t \tag{4}$$

AR operator is $\phi_p(B) = (1 - \phi_1 B - \ldots - \phi_1 B^p)$ when the model in stationer condition and the invertible MA operator $\phi_p(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$ share no factors and d is differencing order.

C. Autocorrelation

Autocorrelation in time series means correlation between past and future value. For a stationary process $\{Z_t\}$, we have the mean $E(Z_t) = \mu$ and variance $Var(Z_t) = E(Z_t - \mu)^2 = \sigma^2$. The correlation between Z_t and Z_{t+k} as

$$\rho_k = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{Var(Z_t)}\sqrt{Var(Z_{t+k})}} = \frac{\gamma k}{\gamma 0} \tag{5}$$

Here, $Var(Z_t) = Var(Z_{t+k}) = \gamma_0$. As function of k, ρ_k is a value of correlation represent influence between the time lag (ACF) in time series analysis because they represent the correlation Z_t and Z_{t+k} from the same process, separated only by k time lags [9]–[11].

D. Cluster Time Series

Clustering is an unsupervised learning task aimed to partition a set of unlabeled data objects into homogeneous groups or clusters. Partition is performed in such a way that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criterion [12]. For time series modeling, the type of possibly used cluster is autocorrelation based distance. Let $\rho_{X_T} = (\rho_{1X_T}, \rho_{2X_T} \dots, \rho_{LX_T})'$ and $\rho_{Y_T} = (\rho_{1Y_T}, \rho_{2Y_T} \dots, \rho_{LY_T})'$ be estimated autocorrelation vector of X_T and Y_T respectively, for some L such that $\rho_{iX_T} \approx 0$ and $\rho_{iY_T} \approx 0$ for $i > L$ define a distance between X_T and Y_T as follows

$$d_{ACF}(X_T, Y_T) = \sqrt{(\hat{\rho}_{X_T} - \hat{\rho}_{Y_T})' \Omega (\hat{\rho}_{X_T} - \hat{\rho}_{Y_T})} \tag{6}$$

Where, $d_{ACF}(X_T, Y_T)$ is autocorrelation distance between X_T and Y_T, $\hat{\rho}_{X_T}$ is estimation of autocorrelation vector of X_T, $\hat{\rho}_{Y_T}$ is estimation of autocorrelation vector of Y_T, and Ω is weight matrices. While ACF distance without weight so that weighted matrices be identity matrices. If weight matrices using identity matrices, so the autocorrelation distance become

$$d_{ACFU}(X_T, Y_T) = \sqrt{(\hat{\rho}_{X_T} - \hat{\rho}_{Y_T})' (\hat{\rho}_{X_T} - \hat{\rho}_{Y_T})} \tag{7}$$

E. Cluster Algorithm

Cluster analysis is a type of data mining analysis. One of the function is reducing a cases number by grouping them into homogeneous clusters, and also can be used to recognize groups without no prior information about the number of possible groups and their membership[13]. Hierarchical cluster analysis can be divided into two types, they are agglomerative and divisive. Agglomerative hierarchical clustering separates data into its individual cluster. The first step so that the initial number of clusters equals the total number of cases [14], [15]. The present paper focused on type of hierarchical agglomerative cluster such as average linkage, complete linkage, and ward linkage.

Complete linkage is one of clustering methods which use the maximum distance between the data. This measure is similar to the single linkage measure, the difference is single linkage using the minimum distance [16], [17]. The formula of complete linkage cluster is,

$$d_{(IJ)K} = \max\{d_{IJ}, d_{JK}\} \tag{8}$$

Where d_{IJ} and d_{JK} are farthest distance between cluster I-J and J-K [13].

Average linkage have the rules using Unweighted Pair Group Method using Arithmetic Average (UPGMA). To overcome the limitations of single and complete linkage [18] proposes measure the average between the data. This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters.

$$d_{(IJ)K} = \frac{\Sigma a . \Sigma b . d_{ab}}{N_{IJ} N_K} \tag{9}$$

Where d_{ab} is distance of object from cluster (IJ) to object b of cluster K, N_{IJ} is count of (IJ) cluster's item, N_K is count of (IJ) and K cluster's item.

Ward's method also called the incremental sum of squares method, uses the within cluster (square) distances and the between-cluster (squared) distance. Formulas for Ward's distance is [19],

$$d_{(I,J)} = \sum_{i \in I \cup J} \left\| \vec{x}_i - \vec{m}_{I \cup J} \right\|^2 - \sum_{i \in I} \left\| \vec{x}_i - \vec{m}_I \right\|^2 - \sum_{i \in J} \left\| \vec{x}_i - \vec{m}_J \right\|^2 \quad (10)$$

$$d_{(I,J)} = \frac{n_I n_J}{n_I + n_J} \left\| \vec{m}_I - \vec{m}_J \right\|^2 \quad (11)$$

Non-hierarchical clustering techniques is one of method in clustering analysis which required to design number of group items before doing the clustering process [15]. On the other hand, K-means required to set the number of K before running the process. Afterwards, the algorithm allowed the objects to be clustered based on the nearest centroid. The centroid was calculated using the mean formula between the objects in each cluster. The procedure of k-means can be defined as:

1. Set the number of groups with k groups.
2. Process every object to choose one the closest distance to the centroid.
3. Use ACF distance to recalculate the mean of each cluster to be set as the new centroid.
4. Repeat Step 2 and 3 until no more reassignments for each objects.

F. Datasets

1) Simulated Data

The simulation study is conducted by generating 7 data models stationary and 7 data models non-stationary. The generated stationary and non-stationary models are presented on Table 1.

Table 1. Simulation models

Stationary	Non Stationary
AR(1)	ARIMA(1,1,0)
AR(2)	ARIMA(2,1,0)
MA(1)	ARIMA(0,1,1)
MA(2)	ARIMA(0,1,2)
ARMA(1,1)	ARIMA(1,1,1)
ARMA(2,1)	ARIMA(2,1,1)
ARMA(2,2)	ARIMA(2,1,2)

Each time series model is generating by 10 different parameters with length of the data (t) is 150. At first, there will be 140 models dataset generation time series with each length (t) 150. Then the model that has been determined, repeated 10 times.

2) Real Dataset

Real data used in this research is the temperature data (C^o) daily in 34 cities in Indonesia. The period is from January 1st until June 30th, 2016. The dataset is obtained from the website of Indonesian Agency for Meteorological Climatological and Geophysics (BMKG).

III. Results and Discussion

A. Simulated Data Process

Each raised-simulated-data will be calculated based on the accuracy of cluster predetermined algorithm. This research simulate are four algorithms: Complete Linkage, Average Linkage, Ward

Linkage, and K-Means. Weight matrices in this research is matrices identity, which is not a weight for each autocorrelation.

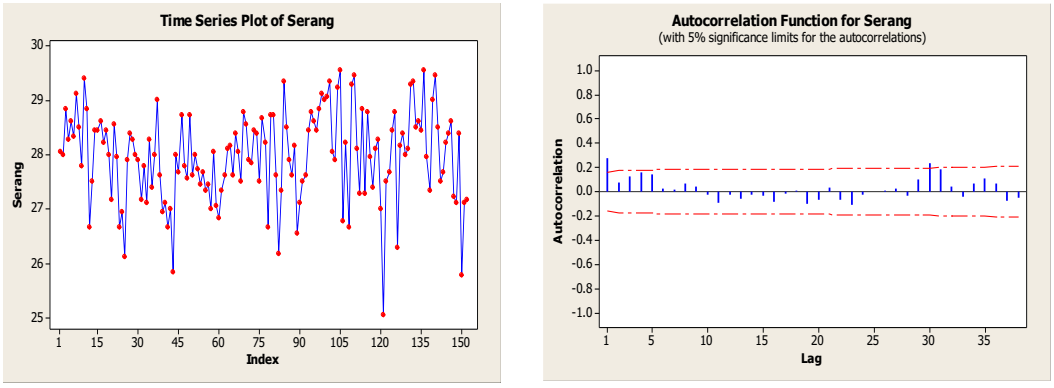
Table 2 shows the accuracy of each algorithms where the K-Means algorithm has higher accuracy than a hierarchical algorithm which is equal to 84.13286%. Therefore, K-Means algorithm is better to classify the stationary and non-stationary data than the other algorithms.

Table 2. Result Accuracy Simulation Study Overall (%)

Simulations	**Algorithm**			
	Complete	***Average***	***Ward***	***K-Mean***
1	80.71429	81.42857	83.57143	82.85714
2	70.71429	79.28571	78.57143	82.85714
3	73.57143	73.57143	73.57143	83.57143
4	81.42857	82.14286	81.42857	85
5	79.28571	80	82.14286	82.14286
6	82.85714	84.28571	84.28571	82.85714
7	87.85714	85	85	86.42857
8	85.71429	85.71429	84.28571	85
9	80	81.42857	80	84.28571
10	79.28571	86.42857	85.71429	86.42857
Average	80.14286	81.92857	81.85714	84.14286

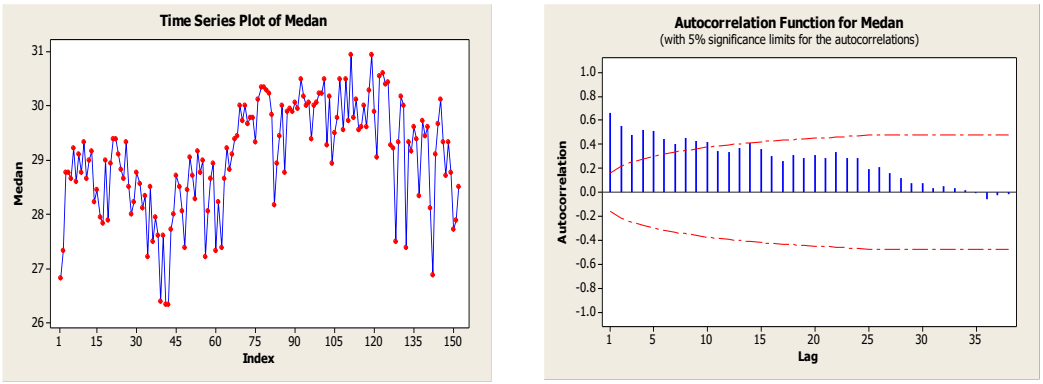
B. Real Dataset Process

In real dataset, to identify data characteristics was done by analyzing time series plot and autocorrelation function plots. Fig. 1 and 2 show a time series plot and ACF's plot of stationary and non-stationary data of some cities in Indonesia. Fig. 1(a) and 1(b) present the temperature in Serang city, one of the observed examples city. It has stationary trend because the time series plot does not indicate graphic trend and the plot autocorrelation is not decrease gradually. Fig. 2(a) and 2(b) show the temperature in another city, Medan. It is recognized as s non-stationary trend because it has deterministic trend which indicated by constant trend and the plot autocorrelation has a slow steady change.



(a) (b)

Fig. 1. Time Series plot and ACF plot for stationary data



(a) (b)

Fig. 2. Time series plot and ACF plot for non-stationary data

Based on the identification of times series data models, there were classified that 11-time series data were non-stationary and 23-time series data identified as stationery (Table 3).

Table 3. Type Time Series

Type Data	Count
Non Stationary	11
Stationary	23

Table 4 shows the accuracy results of each algorithm which is identified that K-Means algorithm has the highest accuracy for distinguishing stationary and non-stationary data with accuracy 85.29412 %. The other three hierarchical algorithms have the same accuracy, 82.35294 %.

Table 4. Result Accuracy of Real Dataset

Algorithm	Accuracy (%)
Average	82.35294
Complete	82.35294
Ward	82.35294
K-Means	85.29412

IV. Conclusion

This research focuses on data simulation and uses clustering method to generate the data. There were seven models to build data simulation for each stationary and non-stationary data. The best models was applied to be used in real case dataset and compared the result based on accuracy. Then time series model was generated by ten different parameters with 150 periods. The real data used in this research were daily temperature data in 34 cities in Indonesia**.** The experiment on simulated data and real dataset shown that the K-Means algorithm has the highest accuracy in both data models, stationary and non-stationary data, with accuracy 84.13286% in simulated data and 85.29412% real dataset. Thus, it can be concluded that K-Means is the best algorithm for classifying stationary and non-stationary time series data.

Acknowledgment

We would like to thanks to Institut Teknologi Bandung and Institut Teknologi Sepuluh Nopember for financial support.

References

[1] A.-H. Homaie-Shandizi, V. P. Nia, M. Gamache, and B. Agard, “Flight deck crew reserve: From data to forecasting,” *Eng. Appl. Artif. Intell.*, vol. 50, pp. 106–114, Apr. 2016.

[2] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting methods and applications*. John wiley & sons, 2008.

[3] P. Manso, "M., A Package for Stationary Time Series Clustering," Master thesis, Universidade da Coruna, 2013.

[4] P. D'Urso and E. A. Maharaj, "Autocorrelation-based fuzzy clustering of time series," *Fuzzy Sets Syst.*, vol. 160, no. 24, pp. 3565–3589, 2009.

[5] U. Habib, K. Hayat, and G. Zucker, "Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering," *Complex Adapt. Syst. Model.*, vol. 4, no. 1, pp. 1–20, 2016.

[6] S. G. Khawaja, M. U. Akram, S. A. Khan, and A. Ajmal, "A novel multiprocessor architecture for k-means clustering algorithm based on network-on-chip," in *Multi-Topic Conference (INMIC), 2016 19th International*, 2016, pp. 1–5.

[7] J. D. Cryer and K.-S. Chan, *Time Series Analysis with applicaitons in R*, 2nd ed. New York: Springer-Verlag New York, 2008.

[8] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering--A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, 2015.

[9] W. W. S. Wei and others, *Time series analysis: univariate and multivariate methods*. Pearson Addison Wesley, 2006.

[10] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 339–367, 2017.

[11] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, 2014.

[12] P. Montero and J. A. Vilar, "Tsclust: An r package for time series clustering," *J. Stat. Softw.*, vol. 62, no. 1, pp. 1–43, 2014.

[13] J. Gunnarsson, "Portfolio-based segmentation and consumer behavior empirical evidence and methodological issues," Stockholm School of Economics, 1999.

[14] M. J. Norusis, *PASW Statistics 18 Statistical Procedures Companion*. Prentice Hall, 2010.

[15] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 3026–3037, 2014.

[16] R. A. Johnson, D. W. Wichern, and others, *Applied multivariate statistical analysis*, vol. 4. Prentice-Hall New Jersey, 2014.

[17] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 2nd ed. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg, 2007.

[18] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*, 1st ed. USA: W. H. Freeman and Company, 1973.

[19] A. M. Paul, *The cult of personality testing: How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstand ourselves*. Simon and Schuster, 2010.