Jurnal TIERS_Artikel Yeni Rahkmawati.docx

by ragam statistika

Submission date: 04-May-2024 11:39AM (UTC+0530)
Submission ID: 2370560738
File name: Jurnal_TIERS_Artikel_Yeni_Rahkmawati.docx (293K)
Word count: 2766
Character count: 14915

TIERS Information Technology Journal
Vol.4, No.2, December 2023, pp. 00~00
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: https://doi.org/10.38043/tiers.v4i1.xxxx

Clustering Time Series Using Dynamic Time Warping Distance in Provinces in Indonesia Based on Rice Prices

Yeni Rahkmawati[1], Selvi Annisa[2]
yeni.rahkmawati@ulm.ac.id[1], selvi.annisa@ulm.ac.id[2]
[1,2]Statistics Study Program, Lambung Mangkurat University, South Kalimantan, Indonesia

ABSTRACT

Rice is a food commodity that is a basic need for Indonesian people. Since the end of 2022, the price of rice has continued to increase until it broke the highest price record from February to March 2024. The price of rice in each province in Indonesia is different. This can happen because rice center provinces will distribute their rice production to other regions to meet rice needs. The grouping of provinces in Indonesia based on rice prices over time is an interesting thing to research. The analysis method used to group similar objects into groups for time series data is called clustering time series. The distance that can be used to measure the closeness of two-time series is the Dynamic Time Warping (DTW) distance. The clustering analysis used is the single, complete, average, Ward, and median linkage method. The results of the analysis show that time series clustering in provinces in Indonesia based on rice prices is best using average linkage hierarchical clustering. The average linkage method has a cophenetic correlation coefficient value of 0.9692, meaning that clustering using the DTW distance with the average difference is very good. The resulting clusters contained 5 clusters which had different characteristics between the clusters.

Keywords:Clustering time series; Dynamic time warping; Hierarchical clustering; Rice Prices, Indonesia

Article Info

Accepted : xx-xx-20xx
Revised : xx-xx-20xx
Published Online : xx-xx-20xx

Correspondence Author:

Yeni Rahkmawati
Statistics Study Program,
Lambung Mangkurat University,
A.Yani St., Km. 36, Banjarbaru, 70714
Email: yeni.rahkmawati@ulm.ac.id

1. INTRODUCTION

Rice is a food commodity that is a basic need for the people of Indonesia. A survey by the Central Statistics Agency shows that around 98.3% of Indonesians consume rice. Rice is an inelastic commodity, meaning that price changes do not cause changes in consumer demand. If availability decreases, prices will soar, which can be unaffordable for consumers [1]. Since the end of 2022, rice prices have been increasing, breaking the record for the highest price from February to March 2024. Medium rice costs Rp 14,000 per kg, and premium rice is Rp 18,000 per kg. In addition, the price of rice in each province in Indonesia is different. This can be because the province, which is the center of rice, will distribute its rice production to other regions to meet rice needs. Provinces that are rice-producing centers such as East Java, Central Java, West Java, South Sumatra, etc. So, the price of rice in the rice center province will affect the price of rice in

the area that is the destination of distribution because there are additional transportation and labor costs to distribute it [2].

Clustering of provinces in Indonesia based on rice prices over time is an interesting thing to research. Several studies on provincial clustering in Indonesia based on rice prices have been carried out, such as research [2] by those who carried out the development of rice price modeling in the western part of Indonesia with a time series clustering approach using DTW distance. In addition, [3] it also conducts provincial clustering based on rice prices using correlation distance. A method of data analysis used to group similar objects together in groups is called clustering. If the data used is time series data, then the clustering time series can be used.

A time series data cluster is a cluster that pays attention to the dynamic nature of time series data. The use of distance in clustering time series data is divided into three categories: raw data, feature data, and model parameters. Distance Raw Data is the distance obtained based on the original data. Distance Featured represents the distance on the representation of the characteristics of the data [4]. Distance-based sample Featured The time series is distance ACF used by [5]. The model parameter distance is the distance of the coefficient of the time series model. One distance that can be used to measure the proximity of two-time series is the Dynamic Time Warping (DTW) distance. Dynamic Time Wrapping (DTW) is one method for calculating the distance between two-time series data. Dynamic Time Wrapping (DTW) is the calculated distance of the optimal warping path between two-time series. DTW distances are more realistically used in measuring the similarity of a pattern than using only linear measurement algorithms such as Euclidean, Manhattan, and other measurement algorithms [6]. One clustering analysis that can be used is hierarchical clustering analysis. This method is used to group observations in a structured manner based on their similar nature, and the number of desired clusters is not yet known. There are two methods of hierarchical clustering: aglomerative and Divisive. The hierarchical method of merging is obtained by combining observations or groups gradually so that, in the end, only one group is obtained. On the contrary, the method of separation in the hierarchical method begins by forming one large group consisting of all observations. The large group is then separated into smaller groups until one group only has one observation. Cluster objects in a hierarchical algorithm using the linkage method (Linkage). Some of the linkage methods used are single, complete, average, Ward, and median linkage methods [7]. So, based on the description above, this study will group provinces in Indonesia based on rice prices using DTW distance and hierarchical clustering.

2. RESEARCH METHOD

2.1 Data

The data in this study used secondary data obtained from the National Food Agency (website: https://badanpangan.go.id/). The variable used is the price of premium rice measured in 38 provinces in Indonesia. The data is daily data from January 18, 2024 to April 22, 2024.

2.2 Method

The procedures used in this study are as follows:

1. Data Exploration
   Data exploration is carried out on premium rice price data with a line box chart to see the distribution of data and data diversity.
2. Calculates Dynamic Time Warping (DTW) distance on premium rice price data
   Dynamic Time Wrapping (DTW) is a crucial method in our study. It calculates the distance between two-time series data, providing a measure of dissimilarity that is not dependent on a specific model approach. This dynamic distance is determined by comparing two-time series data and attempting to find the optimal compressible curve between them, a technique known as time series data clustering. [8]

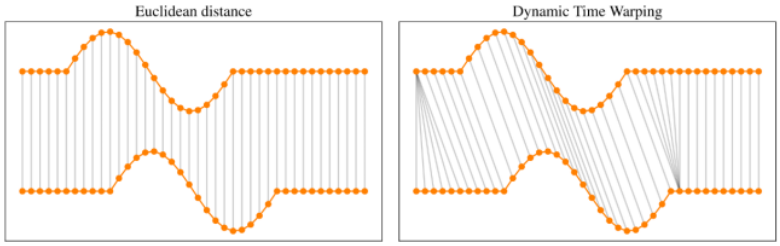

Figure 1. Illustration between Euclidean Distance and DTW Distance

DTW distance can be calculated using the formula below:

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

With $S = s_1, s_2, \dots, s_n$ dan $T = t_1, t_2, \dots, t_m$,, is a time series contained in a matrix of size n x m. $W = w_1, w_2, \dots, w_k$ is the possibility of an arch path that maps or realigns the members of S and T so that the distance between them is minimum. The distance δ can be $\delta(i,j) = |s_i - t_j|$ and w_k refers to a point $(i,j)_k$ on the path of the k-th arch.

3. Perform hierarchical hordes using the Single, Complete, Average, Ward and Median linkage.
4. Calculates the cophenetic correlation coefficient.
   Cophenetic correlation coefficient is the correlation coefficient between the original element of the inequality matrix (Euclidean distance matrix) and the element generated by the dendrogram (Cophenetic matrix based on distance measures and the connectedness method used). The formula for calculating the Cophenetic correlation coefficient is as follows: [9]

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

   With r_{coph} cophenetic correlation coefficient; d_{ik} i-th and k-th Euclidean distances; $\bar{d}$: average distance d_{ik}; d_{Cik}: i-th and k-th cophenetic distances; $\bar{d}_c$: average distance d_{Cik}. The value of the cophenetic correlation coefficient ranges between -1 and 1. The closer to the value of 1, the better the resulting cluster.
5. Select the hierarchical clustering method using the highest Cophenetic correlation value.
6. Make a dendogram for the best hierarchical cluster method.
7. Calculate the optimum number of cluster using the silhouette coefficient.
   Silhouette coefficient is a comparison between the size of the proximity of objects in one cluster and the size of the proximity between the clusters formed. The formula in the measurement of cluster accuracy, namely:

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

   where a(i) is the average distance between object i and all objects in the same group (Intracluster), while b(i) is the average distance between object i and all objects in the nearest cluster (nearest cluster). Coefficient silhouette has a range of values for each object. Subjective interpretation of coefficient quantities $-1 \le s(i) \le 1$ silhouette as in Table 1 [10].

Table 1. Subjective Interpretation of Coefficients Silhouette

Silhouette coefficient	Interpretation
0.71 – 1.00	There is a strong cluster structure
0.51 – 0.70	Reasonable cluster structure
0.26 – 0.50	Weak cluster structure, very likely pseudo
0.00 – 0.25	There is no significant cluster structure

8. Identify the cluster membership.

3. RESULTS AND DISCUSSION

The data exploration aims to determine the distribution of premium rice price data for each province in Indonesia, which is presented in the line box chart in Figure 2. The distribution of rice price data can be seen from the location of the line box chart. Most provinces in Indonesia have rice prices between Rp 14,000 - Rp 20,000, but provinces on Papua and the Maluku Islands have higher premium rice prices than provinces

ISSN: 2723-4533 / E-ISSN: 2723-4541

on other islands. In these provinces, the price of premium rice ranges from Rp. 16,000 to Rp. 36,000. In addition to the data distribution, the diversity in the data can be seen from the width of the box in each province. The width of the boxes between quartiles in most provinces is almost the same, indicating that the diversity of premium rice price data in most provinces is homogeneous. However, several provinces have a larger box width, such as Gorontalo, Central Sulawesi, Southeast Sulawesi, Central Papua, and South Papua. These provinces tend to have a higher diversity of premium rice prices than others.

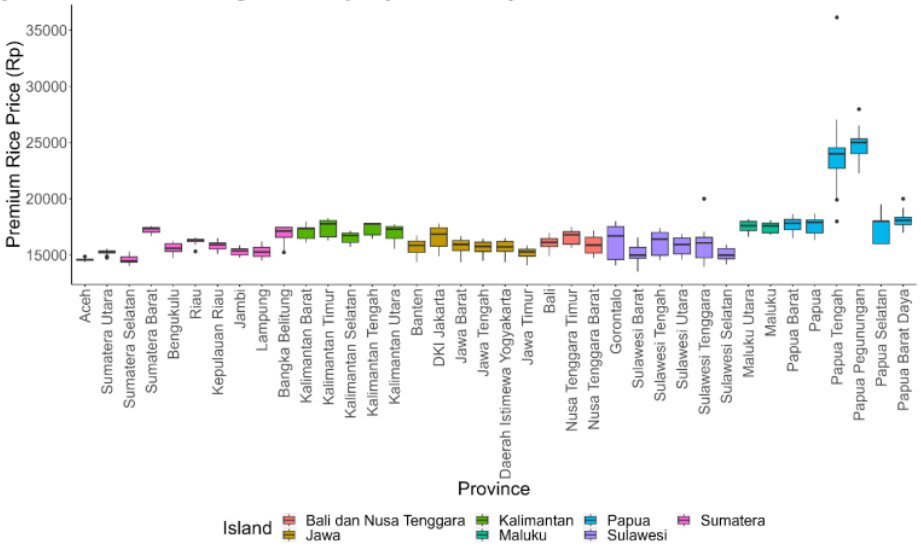

Figure 2. Boxplot of Premium Rice Prices in 38 Provinces in Indonesia

After exploring the data, then a hierarchical cluster was carried out. The hierarchical cluster in this study began by calculating the distance between each province and other provinces using the DTW dissimilarity measure. In simple terms, distance measurement can be started by measuring the distance between Nanggroe Aceh Darussalam and North Sumatra, Nanggroe Aceh Darussalam and West Sumatra, measuring the distance between Nanggroe Aceh Darussalam and Southwest Papua. The results of DTW distance measurement are presented in the table below.

Table 2. DTW Distance Matrix

Provinces	Aceh	Sumatera Utara	⋯	Papua Selatan	Papua Barat Daya
NAD	0	6710.551	⋯	30262.02	34014.5
Sumatera Utara	6710.551	0	⋯	23986.95	27531.34
⋮	⋮	⋮	⋱	⋮	⋮
Papua Selatan	30262.02	23986.95	⋯	0	10817.55
Papua Barat Daya	34014.5	27531.34	⋯	10817.55	0

Then, DTW distances are used to group provinces using the single, complete, average, ward, and median linkage methods. Comparison of the five linking methods using the measure of goodness cophenetic correlation coefficient. The cophenetic correlation coefficient measures the usefulness of using a distance or dissimilarity in the time series data cluster. This measure is obtained from the correlation between the cophenetic distance from the tree diagram and the distance of the original object used to create the tree diagram. The value of the cophenetic correlation has a range, the value of $-1<r<1$the cophenetic correlation close to 1 means that the resulting cluster is perfect. A comparison of several links presented in Table 3 is obtained.

Table 3. Cophenetic Correlation Coefficient in the Hierarchical Cluster Method

Method	Cophenetic Correlation Coefficient
Single	0.9489
Complete	0.9477
Average	0.9692*

Ward	0.6763
Median	0.9617

Based on Table 3, the highest cophenetic correlation coefficient is obtained from the cluster method using an average linkage of 0.9692, meaning that referring to Table 1, the presence of a strong cluster structure or cluster using the distance of DTW with the average linkage is outstanding. The clustering results can be illustrated by the tree diagram (dendogram) in Figure 3.

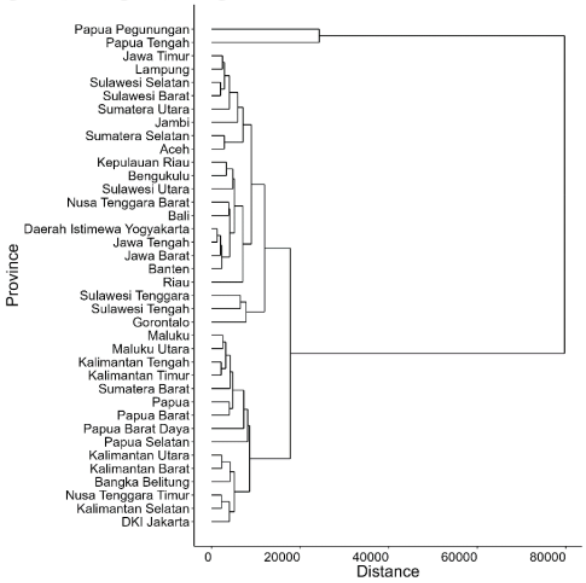

Figure 3. Hierarchical Cluster Analysis Dendogram Using DTW Distance

The cluster results obtained in Figure 3 show that the number of clusters that can be formed is 2 to 38. So, the next step in obtaining the best cluster requires determining the optimal number of clusters. The optimal cluster lot can be determined based on the maximum value of the silhouette coefficient on the number of clusters 2 to 10 presented in the plot in Figure 4.

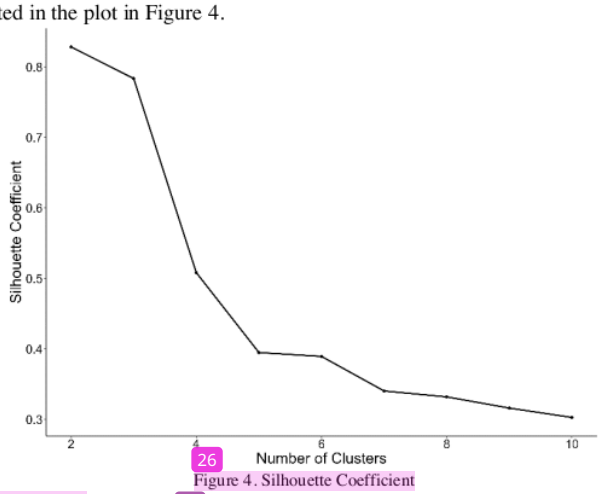

Figure 4. Silhouette Coefficient

The optimal number of clusters obtained can be determined through the subjective interpretation of the silhouette coefficient for each possible number of clusters. Based on the silhouette coefficient in Figure 4, it shows its maximum value is in many clusters k = 2. However, if only two clusters are formed, the spread of objects in the cluster has not been seen because it collects in a particular group. As for the subjective criterion, a cluster can be formed in 5 clusters. The results of clustering with the number of clusters k=5 are presented in Table 4.

Table 4. Cluster of Provinces in Indonesia based on Premium Rice Pri

Clusters	Number of Clusters	Member of Clusters
Cluster 1	18	Aceh, Sumatera Utara, Sumatera Selatan, Bengukulu, Riau, Kepulauan Riau, Jambi, Lampung, Banten, Jawa Barat, Jawa Tengah, Daerah Istimewa Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Barat, Sulawesi Barat, Sulawesi Utara, Sulawesi Selatan
Cluster 2	15	Sumatera Barat, Bangka Belitung, Kalimantan Barat, Kalimantan Timur, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Utara, DKI Jakarta, Nusa Tenggara Timur, Maluku Utara, Maluku, Papua Barat, Papua, Papua Selatan, Papua Barat Daya
Cluster 3	3	Gorontalo, Sulawesi Tengah, Sulawesi Tenggara
Cluster 4	1	Papua Tengah
Cluster 5	1	Papua Pegunungan

Table 4 presents the results of hordes divided into 5 clusters. Group 1 18 provinces have almost the same characteristics; namely, the price of rice in these 18 provinces tends to be stable and not too high compared to other gangs. Gerombol 2 consists of 15 provinces that tend to have pretty high rice prices. Group 3 is a cluster with provinces with high rice prices, meaning that rice prices tend to be less stable in these three provinces. Furthermore, in hordes 4 and 5, there is only one province each, namely Central Papua in group 4 and Mountain Papua in group 5. These two provinces have a high distribution of rice prices compared to other provinces. So, from the hordes above, it can be seen that provincial groups must be the government's attention so that rice prices can be controlled and evenly distributed in each province.

4. CONCLUSION

Time series clustering in Indonesian provinces based on rice prices is best used by average linkage hierarchical clustering. The average linkage method has a cophenetic correlation coefficient of 0.9692, meaning clustering using the DTW distance with the average linkage is very good. The resulting cluster has 5 groups with different characteristics. There are gangs with provincial characteristics and high rice prices, and there are also gangs with unstable prices, which can be a concern for the government in making policies.

REFERENCES

[1] E. Siswanto, B. M. Sinaga och Harianto, ”Dampak Kebijakan Perberasan pada Pasar Beras dan Kesejahteraan,” *Jurnal Ilmu Pertanian Indonesia (JIPI),* vol. 23, nr 2, pp. 93-100, 2018.

[2] S. U. Wijaya och Ngatini, ”Pengembangan Pemodelan Harga Beras di Wilayah Indonesia,” *Limits: Journal of Mathematics and Its Applications,* vol. 17, nr 1, pp. 51-66, 2020.

[3] M. Ulinnuha, F. M. Afendi och I. M. Sumertajaya, ”Study of Clustering Time Series Forecasting Model for,” *Indonesian Journal of Statistics and Its Applications ,* vol. 6, nr 1, pp. 50-62, 2022.

[4] T. W. Liao, ”Clustering of time series data—a survey,” *Pattern Recognition,* vol. 38, p. 1857 – 1874, 2005.

[5] P. D’Urso och E. A. Maharaj, ”Autocorrelation-based fuzzy clustering of time series,” *Fuzzy Sets and Systems,* vol. 160, nr 24, pp. 3565-3589, 2009.

[6] H. Sakoe och S. Chiba, ”Dynamic Programming Algorithm Optimization for,” *IEEE Transactions On Acoustics, Speech, And Signal Processing,* Vol. %1 av %2VOL. ASSP-26, nr 1, 1978.

[7] A. A. Mattjik och I. M. Sumertajaya, Sidik Peubah Ganda dengan Menggunakan SAS, Bogor: IPB Press, 2013.

[8] D. J. Bemdt och J. Clifford, ”Using Dynamic Time Warping to Find Patterns in Time Series,” *Knowledge Discovery in Databases Workshop,* pp. 359-370, 1994.

[9] Iis, I. Yahya, G. N. A. Wibawa, Baharuddin, Ruslan och L. Laome, ”Penggunaan Korelasi Cophenetic Untuk Pemilihan Metode Cluster Berhierarki pada Mengelompokkan Kabupaten/Kota Berdasarkan Jenis Penyakit di Provinsi Sulawesi Tenggara Tahun 2020,” *Prosiding Seminar Nasional Sains Dan Terapan*, Manado, 2010.

[10] Y. Rahkmawati, I. M. Sumertajaya och M. N. Aidi, ”Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach,” *International Journal of Scientific and Research Publications,* vol. 9, nr 9, pp. 439-443, 2019.

Jurnal TIERS_Artikel Yeni Rahkmawati.docx

ORIGINALITY REPORT

19%	13%	15%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	scik.org Internet Source	4%
2	journal.undiknas.ac.id Internet Source	3%
3	Rina Trisminingsih, Sarah Shanaz Shaztika. "ST-DBSCAN clustering module in SpagoBI for hotspots distribution in Indonesia", 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2016 Publication	1%
4	www.growingscience.com Internet Source	1%
5	H Fransiska. "Clustering Provinces in Indonesia Based on Daily Covid-19 Cases", Journal of Physics: Conference Series, 2021 Publication	1%
6	Weng, S.S.. "Mining time series data for segmentation by using Ant Colony	1%

	Optimization", European Journal of Operational Research, 20060916 Publication	
7	Izakian, Hesam, Witold Pedrycz, and Iqbal Jamal. "Fuzzy clustering of time series data using dynamic time warping distance", Engineering Applications of Artificial Intelligence, 2015. Publication	1 %
8	depositonce.tu-berlin.de Internet Source	1 %
9	ijece.iaescore.com Internet Source	1 %
10	Submitted to University of Leeds Student Paper	<1 %
11	5dok.net Internet Source	<1 %
12	www.hindawi.com Internet Source	<1 %
13	"Highlighting the Importance of Big Data Management and Analysis for Various Applications", Springer Science and Business Media LLC, 2018 Publication	<1 %
14	Li, Gang. "Application of Improved K-Means Clustering Algorithm in Customer	<1 %

Segmentation", Applied Mechanics and Materials, 2013.
Publication

15	www.degruyter.com Internet Source	<1%
16	commons.und.edu Internet Source	<1%
17	digitalarchive.boun.edu.tr Internet Source	<1%
18	era.library.ualberta.ca Internet Source	<1%
19	jurnal.stmikroyal.ac.id Internet Source	<1%
20	Edmond H. C. Wu, Philip L. H. Yu. "ICLUS: A robust and scalable clustering model for time series via independent component analysis", International Journal of Systems Science, 2006 Publication	<1%
21	Hasih Pratiwi, Sri S. Handajani, Irwan Susanto, Senot Sangadji, Renny Meilawati, Indah S. Khairunnisa. "Hierarchical Clustering Algorithm for Analyzing Risk of Earthquake on Sumatra Island", 2021 International Conference on Electrical, Computer,	<1%

Communications and Mechatronics Engineering (ICECCME), 2021
Publication

---

22 Rani Nooraeni, Jimmy Nickelson, Eko Rahmadian, Nugroho Puspito Yudho. "New recommendation to predict export value using big data and machine learning technique", Statistical Journal of the IAOS, 2022
Publication

<1 %

---

23 technodocbox.com
Internet Source

<1 %

---

24 Mark Last, Rafael Carel, Dotan Barak. "Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization", Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), 2007
Publication

<1 %

---

25 Rezaie, Mohsen. "Weakly Supervised Performance Evaluation of Trajectory Clustering", Ecole Polytechnique, Montreal (Canada), 2023
Publication

<1 %

---

26 Sydney A. Barnard, Soon M. Chung, Vincent A. Schmidt. "Content-based clustering and visualization of social media text messages",

<1 %

2017 International Conference on Data and Software Engineering (ICoDSE), 2017
Publication

27 Vit Niennattrakul. "Exact indexing for massive time series databases under time warping distance", Data Mining and Knowledge Discovery, 02/16/2010 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Jurnal TIERS_Artikel Yeni Rahkmawati.docx

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6