Time series clustering of Malaysia Air quality time series data

Mohd Aftar Abu Bakar [a,1], Fatin Nur Afiqah Suris [a,2], Noratiqah Mohd Ariff [b,3,*], Kamarulzaman Ibrahim [a,4,*], Tan Zhen Jie [a,5,*]

[a] Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[b] Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[1] tqah@ukm.edu.my
* Corresponding Author

Received 30 Janury 2022; accepted 10 March 2022; published 10 May 2022

ABSTRACT

Air quality is often associated with the area location and activities where air quality in cities is usually more polluted than in rural areas. This study aims to study the pattern of time series data from air quality stations by performing cluster analysis of air quality station based on the particulate matter 10 micrometres or less in diameter (PM10) and particulate matter 2.5 micrometres or less in diameter (PM2.5) time series data. The clusters obtained from the cluster analysis were compared with the station area category and station location. This study which uses air quality data obtained from the Department of Environment, Malaysia from 5 July 2017 until 30 June 2019, shows five types of air quality patterns in Malaysia. The results also show that none of the clusters is dominated by any station's category. Therefore, it is less appropriate to relate the air quality patterns and the station area category. However, the results show that air quality patterns were related to the station's location, where nearby stations have similar air quality patterns.

KEYWORDS
Air quality
Cluster analysis
K-means
Time series clustering.

1. Introduction

A study about air quality is critical since surrounding air quality can have short-term and long-term effects on human health and the environment. Air quality is referred to the condition of air, and it is often related to air pollution associated with the presence of toxic gas or particles in the air. In Malaysia, air quality measurement is represented by air quality index (AQI) or air pollution index (API). In general, the air quality index depends on the level of five types of pollutants in the air which are sulphur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), PM_{10} and $PM_{2.5}$. $PM_{2.5}$ refers to the atmospheric particulate matter with a diameter of smaller than 2.5 micrometres, which is about 3% of the diameter of human hair. In comparison, PM_{10} is for the particles with a diameter of smaller than 10 micrometres, also called fine particles. Typically, air quality-related analysis is vital to identify the pattern of air quality across the country and the prediction of air quality.

The clustering technique is one of the unsupervised learning techniques for clustering corresponding data into their homogenous groups. Meanwhile, time series clustering can be defined as an approach used to cluster a particular type of data that depend on time, known as time-series data. Time-series data is dynamic since the values change over time. Generally, time series clustering is implemented to uncover the pattern in the particular time series data. It has been employed in various domain, for example, [1] applied the time series clustering technique to analyse mobile games players' behaviour data. [2] implemented k-means clustering and artificial neural network modelling to do time series analysis of PM_{10} and $PM_{2.5}$, at New Zealand's coastal sites. Moreover, [3] have used storm indices, the storm intensity and storm concentration index, to cluster the storm events.

10.26555/jifo.v16i2.a25421

This study used the air quality time series data across 64 stations in Malaysia consists of 17,424 hourly observations for five types of variables which are PM$_{10}$, PM$_{2.5}$, ambient temperature (T), wind speed (WS) and humidity (H). The time series clustering has been conducted using the k-means clustering technique, where Euclidean distance is used as the similarity measure.

In the rest of this paper, the air quality data used in this study is described in next section, followed by the time series clustering method, results and discussion and the conclusion in the last section.

2. Method

2.1. Air Quality Data

Malaysia is one of the countries in Southeast Asia. Consisting of thirteen states and three federal territories, Malaysia is considered as developing country with various active industrial sectors. In this study, time series data from 64 air quality monitoring stations in Malaysia located in various industrial, city, rural and other locations were obtained from the Department of Environment Malaysia. Hourly time series data from 5 July 2017 until 30 June 2019 were used in this study. Five variables used in this study are PM$_{10}$ and PM$_{2.5}$, which play important roles in calculating API and three other meteorological parameters: ambient temperature, humidity, and wind speed. Given that there are some missing values, the imputation technique that uses linear function has been applied.

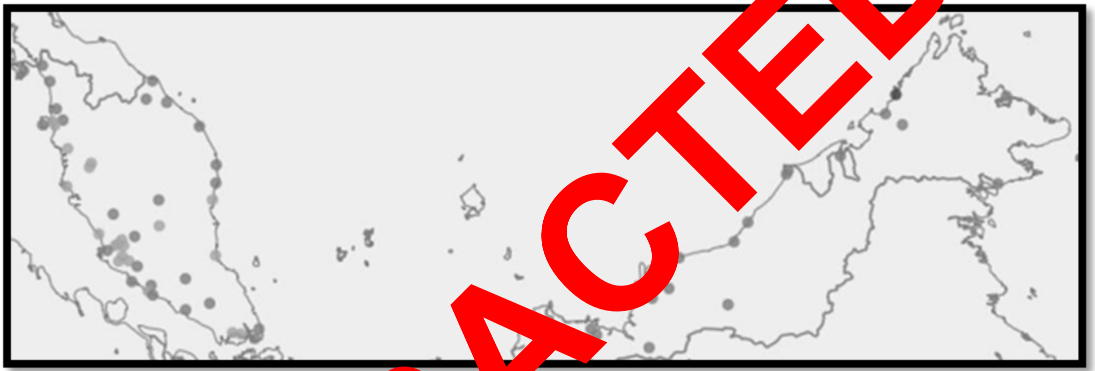

Fig 1. Location of air quality observation stations in Malaysia.

2.2. Time Series Clustering

Clustering is an unsupervised machine learning technique in which data will be collected based on the similarity of its characteristics. In this context, every group or cluster comprises objects with similar attributes [4]. The cluster analysis has been widely utilised in air quality study. For example, [5] have used multidimensional scaling to reduce the dimensions of air quality data and applied spatial clustering to design its air quality display system.

Time series clustering is a type of analysis used to cluster time series data that is dynamic. Many areas of study have applied time series clustering to group time series with similar characteristics, for example, in biology [6], climate [7], finance [8], psychology [9].

Time series clustering can be defined as an unsupervised partitioning process of the dataset with n time-series data $X = \{x_1, x_2, \cdots, x_n\}$ into $C = \{C_1, C_2, \cdots, C_n\}$ where a homogeneous time series is included in the same group based on the measure of similarity. Meanwhile, C_i is known as a cluster where $X = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ where $i \neq j$. In general, time series clustering is branched into three different types: whole time-series clustering, subsequence time-series clustering, and time point clustering. For this study, the whole time series clustering will be used since it is one of the typical time series clustering used by many researchers and since this study involved several series of time series data. The whole time-series clustering consists of four main components, which are dimension reduction, similarity or dissimilarity measures, clustering prototypes and clustering algorithms. Thus, before performing the clustering, these components have to be decided first.

2.3 Similarity Measures

Various similarity measures can be employed in time-series clusterings, such as the Minkowski, Manhattan, Euclidean distance (ED) and Dynamic Time Warping. Let $X = \{x_{ij}: 1, \cdots, I; j = 1, \cdots, J\} =$

$\{x_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ})' : i = 1, \cdots, I\}$ as data matrix where x_{ij} represents the *j*-th variable that is observed on *i*-th object and x_i represents the *i*-th observation vector. This study applied Euclidean distance as the similarity measure in clustering with the following function:

$$d_{il} = \sqrt{\sum_{j=1}^{J} (x_{ij} - x_{lj})^2} ; i, l = 1, \cdots, I \qquad (1)$$

2.4 Cluster Prototypes

Generally, there are three time-series cluster prototypes: time-series medoid, time-series average, and local search prototype. The medoid of a time series has its constraint where the selected time series is mandatory to be in the particular data set. For the time-series average, the limitation occurs as the length of the time series must be equal, and the matrix distance must not be elastic. Meanwhile, the local search prototype refers to a combination of the medoid and the time series average where the medoid is applied first to overcome the inadequacy of starting points in the time series average. Since the air quality time series data used in this study have the same time series length and the similarity measure used is the Euclidean distance, a non-elastic matrix distance, the time series average has been chosen as the prototype of time series clustering in this study.

2.5 K-Means Clustering Algorithm

The k-means clustering is one of the common algorithms used in cluster analysis given its simplicity and ease of implementation [10]. According to Hartigan dan Wong [11], the k-means clustering algorithm divides m objects with n dimension into k groups where k ≤ n by minimising the sum of squares in each group. According to Maharaj et al. [12], the objective function of the k-means clustering algorithm is as follows:

$$Min: \sum_{c=1}^{C} \sum_{i=1}^{I} u_{ic} d_{ic}^2 \qquad (2)$$

where u_{ic} refers to the degree of membership of the *i*-th object in the *c*-th cluster; $u_{ic} = \{0, 1\}$ where $u_{ic} = 1$ indicates the *i*-th object is in the *c*-th cluster while $u_{ic} = 0$ indicates the opposite. Meanwhile, d_{ic}^2 refers to the similarity measure where in this study, Euclidean distance is used as the similarity measure. Hence, the objective function is as follows:

$$Min: \sum_{c=1}^{C} \sum_{i=1}^{I} u_{ic} \|x_i - c_i\|^2 \qquad (3)$$

This algorithm aims to minimise the distance between points with their cluster centroid. For time-series clustering, the objective function becomes:

$$Min: \sum_{c=1}^{C} \sum_{i=1}^{I} \sum_{j=1}^{J} u_{ic} \|x_{ij} - c_{ij}\|^2 \qquad (4)$$

where x_{ij} and c_{ij} respectively refer to the variables and clusters *i* with time series of length *j*.

3. Results and Discussion

Three cluster analysis have been performed in this study. The first two cluster analysis is a univariate time series clustering involving only a single time series data from each station where the variables considered are PM_{10} and $PM_{2.5}$. The last cluster analysis was a multivariate time series cluster analysis conducted on five variables from each station involving both PM_{10} and $PM_{2.5}$ variables, plus the ambient temperature, wind speed and humidity time series data.

Table 1. Cluster membership for PM_{10} and $PM_{2.5}$ according to stations

STATION ID	CLUSTER_PM10	CLUSTER_PM2.5	STATION ID	CLUSTER_PM10	CLUSTER_PM2.5
CA01R	0	2	CA34J	1	1
CA02K	0	2	CA35J	0	2
CA03K	0	2	CA36J	0	2
CA04K	0	1	CA37C	1	3
CA05K	0	1	CA38C	1	1
CA06P	1	1	CA39C	0	2
CA07P	1	1	CA40C	0	2
CA09P	0	2	CA41C	1	1
CA10A	1	1	CA42T	0	1
CA11A	1	1	CA43T	0	2
CA12A	1	1	CA44T	0	1
CA13A	1	1	CA45T	0	2
CA14A	0	2	CA46D	0	1
CA15W	1	1	CA47D	0	1
CA16W	1	1	CA48S	0	2
CA17W	1	1	CA49S	0	2
CA18B	1	1	CA50S	3	0
CA19B	1	1	CA51S	0	2
CA20B	1	1	CA52S	0	2
CA21B	2	1	CA54Q	0	2
CA22B	1	1	CA55Q	0	2
CA23N	1	1	CA56Q	0	2
CA24N	0	1	CA57Q	0	2
CA25N	0	1	CA58Q	0	2
CA26M	0	1	CA59Q	0	2
CA27M	1	1	CA60Q	0	2
CA28M	0	1	CA61Q	0	2
CA29J	0	1	CA62Q	0	2
CA31J	0	1	CA63Q	0	2
CA32J	0	1	CA64Q	0	2
CA33J	1	1	CA65Q	0	2

The initial analysis was performed for univariate time series data, PM_{10} and $PM_{2.5}$, from each station. Using the elbow method [13], it is determined that there are four clusters determined for each PM_{10} and $PM_{2.5}$ named. Here, those clusters were set according to the number 0 to 3, without any preference for the order of those clusters. After examining the pattern of each cluster in PM_{10} and $PM_{2.5}$, the station membership in each cluster was indicated. The station membership for each cluster is shown in Table 1.

The table shows that the stations grouped to cluster 0 based on PM_{10} time series data were either in cluster 1 or 2 based on $PM_{2.5}$ time series data. The stations assigned to cluster 1 based on PM_{10} time series data are in cluster 1 based on $PM_{2.5}$ time series data. Only stations CA21B (Klang, Selangor) and CA50S (Kota Kinabalu, Sabah) are in the different clusters for clusterisation based on PM_{10}, while CA37C (Rompin, Pahang) and CA50S (Kota Kinabalu, Sabah) are in their own clusters if based on $PM_{2.5}$ time series data. Considering both variables, it was evident that station CA50S somehow have very different patterns compared to the other stations. It is noticeable that clusters obtained from both analyses show almost similar clustering patterns.

Table 2 shows the crosstable between the cluster generated from the univariate time series clustering of the PM_{10} with the stations' location categories. From the table, it is shown that stations in cluster 0 consists of various station's categories. However, the cluster is dominated by the suburb stations followed by the rural and industrial stations. Cluster 1 is also dominated by the suburb stations. However, the city stations are more dominant in cluster 1 compared to the rural stations. Moreover, stations CA21B and CA50S, which are in cluster 2 and 3, respectively, are located in suburb areas. Therefore, it can be deduced that cluster 0 from PM_{10} univariate time series clustering represents suburb, rural and industrial areas while cluster 1 represents suburb and city areas.

Table 3 shows the crosstable between the clusters produced by the univariate cluster analysis of the $PM_{2.5}$ time series with the stations' location categories. The majority of stations in cluster 1 are located in the suburb areas, followed by stations in city areas. Meanwhile, stations located in suburb areas were dominant in cluster 2, followed by those in rural areas. However, those stations in industrial areas were divided evenly on cluster 1 and 2. Station CA50S, the only station in cluster 0, is located in the suburb area, while station CA37C, also the only station in cluster 3, is located in the rural area. Hence, it can be inferred that cluster 1 from the $PM_{2.5}$ univariate cluster analysis represents suburb and city areas while cluster 2 represents suburb and rural areas.

Table 2. Crosstable between the clusters produced by the univariate cluster analysis of the PM10 time series with the location categories

Location Category	Cluster 0	Cluster 1	Cluster 2	Cluster 3
City	4	6	0	0
Industry	6	1	0	0
Rural	9	3	0	0
Inland	1	[illegible]	0	0
Suburb	20	10	1	1

RETRACTED

Table 3. Crosstable between the clusters produced by the univariate cluster analysis of the PM2.5 time series with the location categories

Location Category	Cluster 0	Cluster 1	Cluster 2	Cluster 3
City	[illegible]	9	1	0
Industry	0	3	4	0
Rural	0	4	7	1
Inland	0	0	1	0
Suburb	1	17	14	0

Table 3 shows the crosstable between the clusters produced by the univariate cluster analysis of the $PM_{2.5}$ time series with the stations' location categories. The majority of stations in cluster 1 are located in the suburb areas, followed by stations in city areas. Meanwhile, stations located in suburb areas were dominant in cluster 2, followed by those in rural areas. However, those stations in industrial areas were divided evenly on cluster 1 and 2. Station CA50S, the only station in cluster 0, is located in the suburb area, while station CA37C, also the only station in cluster 3, is located in the rural area. Hence, it can be inferred that cluster 1 from the $PM_{2.5}$ univariate cluster analysis represents suburb and city areas while cluster 2 represents suburb and rural areas.

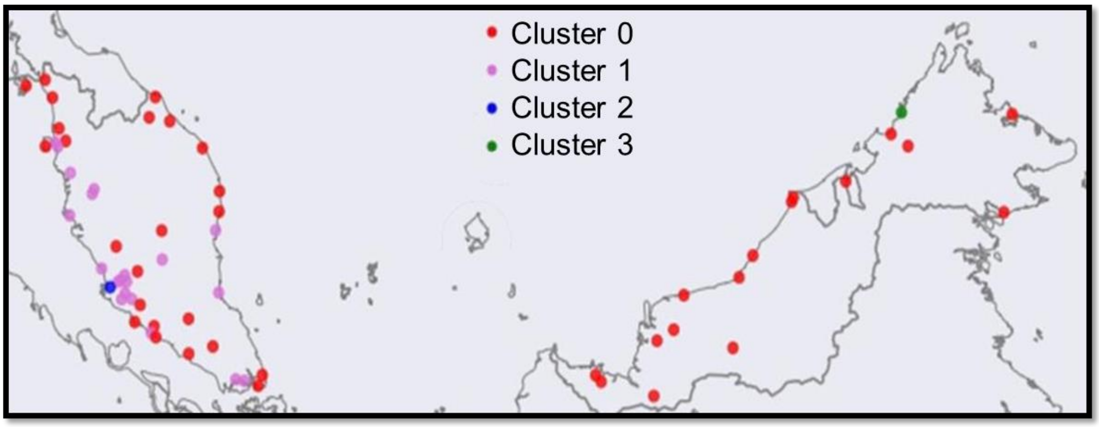

Fig 2. Location of air monitoring stations coloured according to clusters from the univariate cluster time series of PM10.

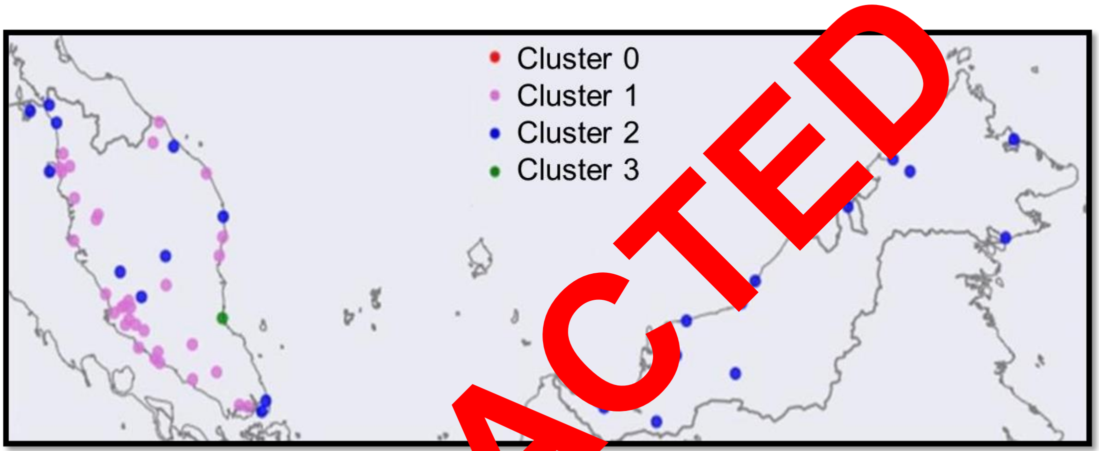

Fig 3. Location of air monitoring stations coloured according to clusters from the univariate cluster time series of PM2.5.

From Fig. 2, it is found that cluster 0 stations from the PM_{10} univariate time series clustering are distributed throughout Malaysia. In contrast, cluster 1 stations are primarily located in major urban areas in Peninsular Malaysia. However, Fig 3 shows that the locations of the clusters produced from the univariate time series clustering of $PM_{2.5}$ data are different compared to those produced by the PM_{10} time series. Cluster 2 stations are mainly in East Malaysia, while stations from cluster 1 are concentrated in Peninsular Malaysia.

3.1 Multivariate Time Series Clustering

The multivariate time series clustering in this study consist of two air quality variables: PM_{10} and $PM_{2.5}$, and three surrounding conditions: ambient temperature, wind speed and humidity. The optimal number of clusters based on the elbow method is five. From Table 4, cluster 0 and 3 were very dominated by suburb stations, while cluster 2 is slightly dominated by suburb stations, followed by city stations. Cluster 4 is dominated by rural stations, followed by suburb stations. Only station CA57Q (Samalaju, Bintulu Sarawak), located in the industrial area, is in Cluster 1. Meanwhile, most of the other stations located in the industrial areas were grouped in Cluster 0.

Table 4. Crosstable between the clusters produced by the multivariate time series cluster analysis with the location categories.

Location Category	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Inland	0	0	0	1	0
Industry	4	1	0	1	1
Rural	3	0	1	1	7
Suburb	11	0	9	7	5
City	2	0	5	2	1

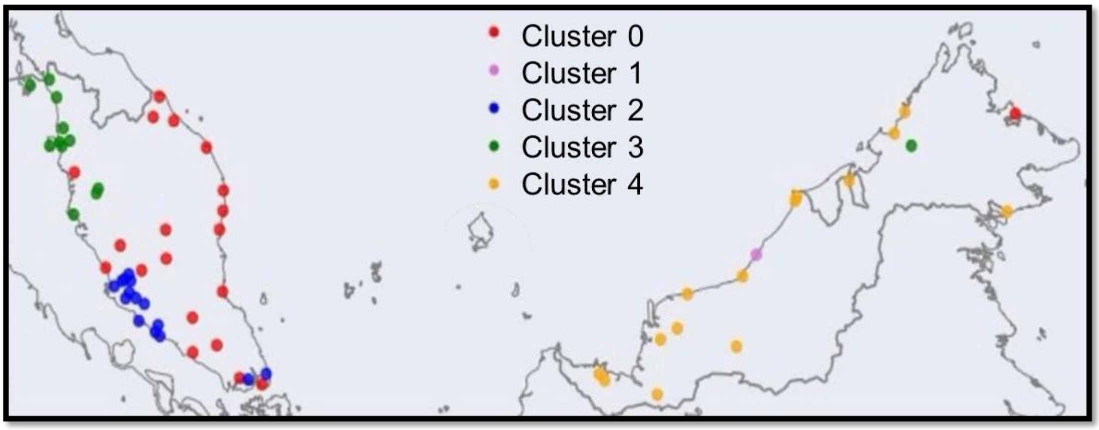

Fig 4. Location of air monitoring stations coloured according to clusters from the multivariate time series clustering.

From Fig 4, it can be inferred that the resulting clusters show more significant differences and dependency on the location of the stations. Cluster 2 stations are scattered in the central region of Peninsular Malaysia, and cluster 3 is scattered in the northern region of Peninsular Malaysia. Meanwhile, cluster 0 represents the other regions in Peninsular Malaysia especially in the southern and eastern regions. Stations in group 4 consist of stations located in Sabah and Sarawak with the exception of only two stations from cluster 0 and 3.

4. Conclusion

In this study, time series clustering has been performed on the air quality time series data from 64 air quality monitoring in Malaysia. The time series of PM_{10} and $PM_{2.5}$ have been clustered according to their similarity using the k-means algorithm. Three different cluster analysis were performed, resulting in different results in terms of clusters produced.

Based on the clusters generated from all the analysis, generally, it was shown that the activities represented by the station's categories do not heavily influence the similarity in terms of air quality patterns. It seems that the regions play more influential roles in determining the patterns of air quality time series of those stations. It is also apparent that Peninsular Malaysia and East Malaysia have different air quality time series patterns.

The results are also shown that the stations in suburb areas dominated most clusters. This mainly because most of the air quality monitoring stations considered in this study were located in suburban areas. From these results, it can be concluded that even though some suburb and rural areas were not in the industrial or city areas, they may have similar air quality characteristics as those industrial or city areas.

Data availability

The data used in this study were obtained from the Department of Environment Malaysia.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding statement

This research was supported by the Ministry of Education, Malaysia [FRGS/1/2019/STG06/UKM/02/4] and the Universiti Kebangsaan Malaysia [GUP-2019-048].

Acknowledgment

The authors would like to express great gratitude to the Universiti Kebangsaan and the Ministry of Education, Malaysia for the allocation of the research grant and the facilities used for this research.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] A. Saas, A. Guitart, and A. Perianez, "Discovering playing patterns: Time series clustering of free-to-play game data," *IEEE Conf. Comput. Intell. Games, CIG*, vol. 0, Jul. 2016, doi: 10.1109/CIG.2016.7860442.

[2] M. A. Elangasinghe, N. Singhal, K. N. Dirks, J. A. Salmond, and S. Samarasinghe, "Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering," *Atmos. Environ.*, vol. 94, pp. 106–116, Sep. 2014, doi: 10.1016/J.ATMOSENV.2014.04.051.

[3] M. A. A. Bakar, N. M. Ariff, A. A. Jemain, and M. S. M. N[illegible]dzir, "Cluster Analysis of Hourly Rainfalls Using Storm Indices in Peninsular Malaysia," *J. Hydrol. Eng.*, vol. 25, no. 7, p. 05020011, Apr. 2020, doi: 10.1061/(ASCE)HE.1943-5584.00019[illegible].

[4] N. M. Ariff, M. A. A. Bakar, and Z. H. Zamzuri, "Academic preference based on students' personality analysis through k-means clustering," *Malaysian J. Fundam. Appl. Sci.*, vol. 16, no. 3, pp. 328–333, Jun. 2020, doi: 10.11113/MJFAS.V16N3.1640.

[5] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao, and W. [illegible], "Visual Analytics for Spatial Clusters of Air-Quality Data," *IEEE Comput. Graph. Appl.*, vol. 37, no. 5, pp. 98–105, 2017, doi: 10.1109/MCG.2017.3621228.

[6] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21, Suppl. 1, no. SUPPL. 1, Jun. 2005, doi: 10.1093/BIOINFORMATICS/BTI1022.

[7] N. Mohd Ariff *et al.*, "CLUSTERING OF RAINFALL DISTRIBUTION PATTERNS IN PENINSULAR MALAYSIA USING TIME SERIES CLUSTERING METHOD," *Malaysian J. Sci.*, vol. 38, no. Sp2, pp. 84–[illegible]9, Sep. 2019, doi: 10.22452/MJS.SP2019NO2.8.

[8] F. Pattarin, S. Paterlini, and T. Minerva, "Clustering financial time series: An application to mutual funds style analysis," *Comput. Stat. Data Anal.*, vol. 47, no. 2 SPEC. ISS., pp. 353–372, Sep. 2004, doi: 10.1016/J.CSDA.2003.11.009.

[9] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015, doi: 10.1016/J.IS.2015.04.007.

[10] N. M. Ariff, M. A. A. Bakar, and M. I. Rahmad, "Comparative study of document clustering algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 246–251, 2018, doi: 10.14419/IJET.V7I4.11.20816.

[11] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.

[12] E. A. Maharaj, P. D'Urso, and J. Caiado, "Time Series Clustering and Classification," *Time Ser. Clust. Classif.*, Mar. 2019, doi: 10.1201/9780429058264.

[13] "An Introduction to Statistical Learning: with Applications in R | SpringerLink." https://link.springer.com/book/10.1007/978-1-4614-7138-7 (accessed Dec. 28, 2022).