




Article

Prediction of PM₁₀ Concentration in Malaysia Using K-Means Clustering and LSTM Hybrid Model

Noratqah Mohd Ariff *, Mohd Aftar Abu Bakar  and Han Ying Lim 

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; aftar@ukm.edu.my (M.A.A.B.); a176483@siswa.ukm.edu.my (H.Y.L.)

* Correspondence: tqah@ukm.edu.my

Abstract: Following the rapid development of various industrial sectors, air pollution frequently occurs in every corner of the world. As a dominant pollutant in Malaysia, particulate matter PM₁₀ can cause highly detrimental effects on human health. This study aims to predict the daily average concentration of PM₁₀ based on the data collected from 60 air quality monitoring stations in Malaysia. Building a forecasting model for each station is time-consuming and unrealistic; therefore, a hybrid model that combines the k-means clustering technique and the long short-term memory (LSTM) model is proposed to reduce the number of models and the overall model training time. Based on the training set, the stations were clustered using the k-means algorithm and an LSTM model was built for each cluster. Then, the prediction performance of the hybrid model was compared with the univariate LSTM model built independently for each station. The results show that the hybrid model has a comparable prediction performance to the univariate LSTM model, as it gives the relative percentage difference (RPD) less than or equal to 50% based on at least two accuracy metrics for 43 stations. The hybrid model can also fit the actual data trend well with a much shorter training time. Hence, the hybrid model is more competitive and suitable for real applications to forecast air quality.

Keywords: air quality; forecasting; hybrid model; PM₁₀; time series clustering; k-means; LSTM



Citation: Ariff, N.M.; Bakar, M.A.A.; Lim, H.Y. Prediction of PM₁₀ Concentration in Malaysia Using K-Means Clustering and LSTM Hybrid Model. *Atmosphere* **2023**, *14*, 853. <https://doi.org/10.3390/atmos14050853>

Academic Editor: Célia Alves

Received: 29 March 2023

Revised: 19 April 2023

Accepted: 26 April 2023

Published: 11 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In line with the rapid development of various industrial sectors, air pollution frequently occurs worldwide, including in Malaysia. According to the World Health Organization (WHO) [1], air pollution is defined as the contamination of indoor and outdoor environments by impurities that modify the natural features of the environment. Data collected by WHO reveal that most of the global population breathes highly contaminated air that exceeds WHO guidelines. Air pollution can cause detrimental effects on human health, especially the respiratory system, and becomes one of the fundamental sources of morbidity and mortality [1].

In Malaysia, the air pollutant index (API) adopts six main air pollutants and serves as an indicator to deliver accurate and insightful information on air quality status in any area to the public [2]. Rani et al. [3] analyzed the trend of the API in Malaysia from years 2010 to 2015 based on various categories by using XLSTAT. In October 2010, the concentration of particulate matter 10 µm or less in diameter, better known as PM₁₀, was extremely high in some areas in Johor following the occurrence of forest fires in Indonesia, which led to high API values as the highest relative subindex of monitored pollutants that account for the API readings [3,4]. This suggests that such fine dust often found in polluted air contributes greatly to the variability of the API [3].

Particulate matter is not just the main air pollutant in the Southeast Asia region, but is also identified as the most severe city pollutant around the globe [5,6]. For instance, most of the daily average PM₁₀ concentration at three monitoring stations in Buenos Aires from the

years 2010 to 2018 exceeded the standard limit of WHO guidelines, that is, $50 \mu\text{g}/\text{m}^3$ [7]. Some research findings highlight that the particulate matter concentrations have certain correlations with the weather conditions, four seasons and monsoons [8–10].

Due to the increasing public awareness of the dangers of air pollution, numerous air quality-related studies have been performed using various statistical and deep learning models, including forecasting and clustering. Clustering is an exploratory data analysis technique that investigates the fundamental structure of data [11]. By adopting the clustering technique, the data are assigned into several distinct groups based on their degree of similarity before any further analysis or modeling can be performed. As the data within the cluster can be treated using the same analysis technique, it can save costs and computation time. There are several types of clustering methods, such as partitionial clustering, hierarchical clustering and fuzzy clustering. Hierarchical clustering groups similar objects into clusters that eventually merge into a single cluster, whereas fuzzy clustering is a soft-clustering technique in which the objects can be clustered into more than one cluster. As a partitionial clustering method, the k-means algorithm is one of the most common and popular techniques since it can be implemented easily [12]. It classifies data with closer centroid values into the same cluster such that the differences between the clusters are maximized. For instance, k-means clustering was used to analyze the significant changes in air quality in Southampton [13]. While Kim et al. [14] applied this algorithm to cluster monitoring stations in the United States based on different temporal patterns of $\text{PM}_{2.5}$, Beaver and Palazoglu [15] adopted it to classify classes of ozone episodes in San Francisco.

Air quality time series clustering in Malaysia is often utilized to identify the pattern between the clusters and categorize the area into zone based on the pollution level so that government policies can be executed accurately [16]. In this context, Suris et al. [17] clustered the PM_{10} data in Malaysia using dynamic time warping (DTW) as the dissimilarities measure. Adopting four clustering techniques, that is, k-means, partitioning around medoid (PAM), agglomerative hierarchical clustering (AHC) and fuzzy k-means (FKM), the results show that the clusters were formed mainly on the basis of the region and geographical location of the stations instead of the station category and local economic activities. A similar result was obtained by Rahman et al. [11], whereby the stations were classified into high, medium and low pollution regions, respectively, using the AHC technique based on the daily average $\text{PM}_{2.5}$ concentration.

As climatic and environmental issues concern society, air quality forecasting has become the focus among researchers as an accurate prediction that can reduce the effect of pollution on humans and the biosphere [18]. Therefore, various types of prediction models have been applied in previous studies. For instance, Aditya et al. [19] used the logistic regression and autoregression (AR) models to detect air quality and predict the concentration of $\text{PM}_{2.5}$. A similar approach is shown in the research by Bhalgat et al. [18], which adopted AR and autoregression integrated moving average (ARIMA) models to predict the concentration of sulfur dioxide (SO_2). Meanwhile, Guo et al. [20] used a geographically and temporally weighted regression model to calibrate the spatiotemporal dynamic $\text{PM}_{2.5}$ concentrations to manage haze pollution in China. The random forest method is also deemed capable of modelling various concentrations of air pollutants, such as $\text{PM}_{2.5}$ and ozone [21,22]. In fact, random forest regression is believed to predict air pollutant concentrations more accurately than linear regression and decision trees [23].

In recent years, neural networks have been preferred by researchers rather than the abovementioned traditional models due to their ability to fit non-linear data with higher accuracy [10]. The long short-term memory (LSTM) model is a deep learning method modified based on the concept of the recurrent neural network (RNN). Given its strength in solving the shortcomings of the RNN model, such as poor performance with tasks that involve long-term dependency and a vanishing and exploding gradient, the LSTM is found to be suitable to predict sequential data, including time series data. The outstanding performance of the LSTM model is observed through a lower root mean squared error

(RMSE) in predicting the prices of gold [24] and Bitcoin [25], as well as influenza-like illnesses and respiratory diseases [26].

In terms of air quality prediction, the LSTM model also possesses great potential to give an accurate result [27]. The findings obtained by Bakar et al. [28] show that the multivariate LSTM model predicted the PM₁₀ concentration at five selected monitoring stations most accurately with the lowest RMSE values, followed by the univariate LSTM model and the univariate ARIMA model. Aiming to increase prediction accuracy, hybrid models that involve a combination of techniques are gaining popularity in the research field. Zhang et al. [29] discovered that the combination of principal component analysis (PCA) and least squares support vector machine (LSSVM) can reduce the noise in meteorological data, hence giving more accurate predictions in API than the ARIMA model. The PCA-ANN model that uses only the significant parameters also seems competitive in giving a better prediction than the standalone artificial neural network (ANN) model [30].

For the case of clustering-based LSTM model, it considers the changes in features that are more specific in each cluster, making it an ideal choice to improve prediction accuracy. Yulita et al. [31] utilized fuzzy clustering and bidirectional LSTM (Bi-LSTM) to obtain higher accuracy and precision in classifying sleep stages. In accordance with the findings obtained in the study on the load prediction for dynamic spectrum allocation performed by Liu et al. [32] using AHC-LSTM, Li et al. [33] also found that type-2 fuzzy clustering-based LSTM can increase the accuracy with a much shorter model training time in long-term traffic volume prediction than the LSTM, random forest, back propagation network (BPN) and deep neural network (DNN).

Besides the abovementioned combinations, k-means clustering is also one of the widely used techniques in hybrid models. Ao et al. [10] first clustered meteorological data according to seasons using the k-means algorithm, then combined the clustering results with the air pollutant concentrations to be input into the Bi-LSTM model. It was found that the proposed model outperforms the other models as it can overcome the continuous fluctuation in meteorological conditions. Using the k-means-LSTM model, Baca et al. [34] also obtained a better air quality prediction in Andahuaylas, Peru.

Air quality prediction is indeed important for society to take preliminary preparations and preventive measures against poor air conditions. In order to figure out the potential of the hybrid model in predicting the daily average PM₁₀ concentration in Malaysia, this study proposes a clustering-based LSTM model and compares its performance with the univariate LSTM model without clustering. Being a state-of-the-art deep learning method, the LSTM model usually outperforms conventional forecasting models in prediction accuracy. However, it is too time-consuming and unrealistic to construct the model individually for each station, especially in real-life applications. If the model is trained based on a few samples and generalizes its finding to all stations, it might cause an undesirably low accuracy at some stations outside the sampling. Therefore, such a combination of techniques is deemed capable of increasing the prediction accuracy with much less computation time, thus proving to be more efficient than the classical forecasting technique.

2. Materials and Methods

2.1. Data Preprocessing

The data used in this study are the daily average PM₁₀ concentrations monitored at 60 air quality monitoring stations in Malaysia from 5 July 2017 to 31 January 2019, provided by the Malaysian Department of Environment (DOE). The dataset, with a length of 576 days for each time series, was divided into the training set and test set based on a ratio of 8:2 [18,26,35]. Data normalization was carried out in order to eliminate the effect of a wide range observed in the PM₁₀ concentration, to speed up the training process and to increase prediction accuracy [35]. The training data was scaled into a range of [0, 1] using the min-max scaler as follows:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (1)$$

where x_{scaled} and x refer to the scaled data and the original data, respectively, whereas x_{min} and x_{max} represent the minimum and maximum values of the data, respectively.

2.2. Time Series K-Means Clustering

The k-means approach is a partitional clustering technique that decomposes the data into a set of disjointed clusters based on the nearest centroids.

Let $X = \{x_{ij} : 1, \dots, I; j = 1, \dots, J\}$ as a data matrix, where x_{ij} represents the j -th variable observed for the i -th object. According to Kobylin and Lyashenko [36], the k-means algorithm usually adopts the Euclidean distance as the proximity measure:

$$d_{il} = \sqrt{\sum_{j=1}^J (x_{ij} - x_{lj})^2}. \tag{2}$$

This distance measure has been proven competitive in terms of time series classification accuracy [37].

Additionally, the shape-based DTW distance can also be implemented to measure the proximity in time series clustering. Despite being a good similarity and dissimilarity measure [17], this approach typically consumes more computation time due to its dynamic and complicated calculations [38]. Since the time series data are of the same length, the Euclidean distance has been chosen as the proximity measure [39].

The procedure for time series k-means clustering is as follows:

- (i) Initiate the k -cluster based on the randomly chosen cluster centroids;
- (ii) Allocate each datapoint into the nearest cluster by employing the Euclidean distance;
- (iii) Recompute the cluster centroids based on the current cluster members;
- (iv) Repeat steps (ii) and (iii) until no there are changes in the cluster membership.

The k-means algorithm classifies a time series into k clusters in such a way that the within-group sum of squares (WGSS) is minimized. According to Maharaj et al. [40], the objective function of the k-means clustering is as follows:

$$\min \left[\sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J u_{ik} \|x_{ij} - k_{ij}\|^2 \right], \tag{3}$$

where u_{ik} is the degree of membership of the i -th object in the k -th cluster that takes the value of $\{0, 1\}$. If $u_{ik} = 1$, it indicates that the i -th object is in the k -th cluster. On the contrary, $u_{ik} = 0$ shows that the i -th object is not in the k -th cluster.

Choosing an optimum number of k clusters could be a challenging task. In this study, the optimal k is chosen based on the internal index, that is, the WGSS visualized on the elbow plot and the silhouette index. For each time series, the error is defined as the distance to the nearest cluster [41].

The k that gives the highest gradient and the sharpest elbow curve is chosen as the candidate before it is evaluated by the silhouette index, as shown below:

$$s = \frac{b - a}{\max(a, b)}, \tag{4}$$

where a is the average distance within the cluster and b represents the average distance between the clusters. This index is a metric that evaluates the accuracy of a clustering technique based on scores between -1 and 1 . A coefficient of 1 indicates that the clusters are well separated and clearly distinguished, whereas a score of -1 means that the clusters are not appropriately partitioned. If the silhouette index has a value of 0 , it shows that the distance between the clusters is insignificant. Therefore, a higher index score indicates a better separation of the clusters [42,43].

2.3. Model LSTM

2.3.1. Introduction

An LSTM model is the extension of RNN and is capable of learning long-term dependency and storing the information for a long period. These characteristics of LSTM make it a state-of-the-art model, especially in time series prediction, which highly depends on the changing patterns of previous values.

Generally, the chain-like LSTM structure consists of three gates that control the flow of information in the memory cell, namely, the forget gate, input gate and output gate. In every cell, there are two types of non-linear activation functions, that is, the sigmoid function and the hyperbolic tangent (tanh) function. The other components of the LSTM cell include the cell state and hidden state. At each gate, there exist weights, W , and biases, e .

According to Colah [44], the key to LSTM is the cell state, which is the horizontal line running through the top of the diagram shown in Figure 1.

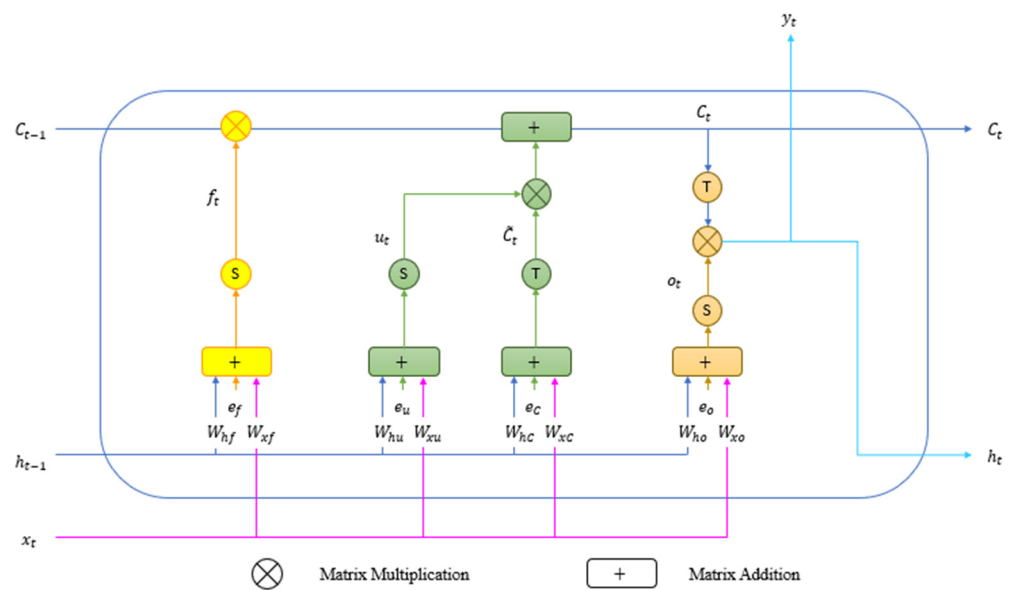


Figure 1. LSTM cell structure.

The cell state runs straight down the entire chain with a few minor linear interactions. Information can flow along the cell state under the control of three gates that are composed of a sigmoid neural net layer and a pointwise multiplication operation. The sigmoid layer gives an output between 0 and 1 to indicate how much of each component should be let through. None of the information can flow through the gates when a value of 0 is output. On the other hand, a value of 1 indicates all the information can be let through.

The process in the LSTM cell begins at the forget cell, whereby the sigmoid layer determines what information needs to be removed from the cell state. Looking at the former hidden state h_{t-1} and input data x_t , it outputs a value between 0 and 1 for each number in the former cell state C_{t-1} . This process can be described by the following equation:

$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + e_f). \tag{5}$$

Next, the new information to be stored in the cell state will be determined in two steps. Firstly, the sigmoid layer at the input gate will determine which values are to be updated. Secondly, the tanh layer will produce a vector of new candidate values \tilde{C}_t that could be added to the cell state. These processes can be expressed as follows:

$$u_t = \sigma(W_{hu}h_{t-1} + W_{xu}x_t + e_u), \tag{6}$$

$$\tilde{C}_t = \tanh(W_{hu}h_{t-1} + W_{xu}x_t + e_u). \quad (7)$$

Then, a combination of the outputs will be used to update the former cell state C_{t-1} into the new cell state C_t . The former cell state is multiplied by f_t to lose the decided information before it is added to the product of $u_t \cdot \tilde{C}_t$. These are the new candidate values that have been scaled by how much each cell state value should be updated. The process is described by the following equation:

$$C_t = f_t \cdot C_{t-1} + u_t \cdot \tilde{C}_t. \quad (8)$$

Finally, the output gate decides what information should be output based on the filtered cell state. Firstly, the former hidden state and the input data will be run through the sigmoid layer to decide which part is to be eliminated. Then, the cell state will be put through the tanh layer to generate the values between -1 and 1 before multiplying by the output from the sigmoid layer. Eventually, only the decided portion will be output. The following equation summarizes the processes that occur at the output gate:

$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t + e_o), \quad (9)$$

$$h_t = y_t = o_t \cdot \tanh(C_t). \quad (10)$$

2.3.2. Multivariate LSTM Model

As more than one feature are considered when constructing the hybrid model for each cluster, the LSTM model is said to be multivariate. In this study, the mean squared error (MSE) was adopted as the loss function.

Adaptive moment estimation (Adam) was employed to update the weights in the neural network based on the training data. The number of epochs was set as 100. Aiming to avoid overfitting, early stopping was employed to stop the training whenever there was no improvement in the model performance for 15 consecutive epochs [45].

The optimum values for other hyperparameters, such as the dropout rate, hidden neuron, timestep, batch size and hidden layer, were determined by using the manual tuning approach to obtain the best model performance at the training stage.

2.3.3. Univariate LSTM Model

A univariate LSTM model is a model that is trained based on one feature only, that is, it only involves one time series. The model construction process is the same as in the multivariate LSTM model, except for the number of input features.

2.4. Comparison of Model Prediction Performance

There are three accuracy metrics adopted as the prediction performance indicators for the constructed models in this study, namely RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE).

Then, the relative percentage difference (RPD) was calculated for each accuracy metric to compare the prediction performance between both models. Generally, the RPD is computed using the following formula:

$$RPD = \frac{|D_1 - D_2|}{\left(\frac{D_1 + D_2}{2}\right)} \times 100\%, \quad (11)$$

where D_1 and D_2 are the values measured by the first and second methods, respectively, which are the values obtained from the proposed hybrid model and univariate LSTM model in this case. The RPD is a common method to compare two experimental values when

there is no theoretical value as a reference [46]. A good RPD value can be defined based on the types of experiments. In general, an acceptable RPD value ranges from 0% to 50% [47].

2.5. Framework

This study involves three main components, namely, the time series clustering phase, the modeling phase and a comparison of the model prediction performance, as summarized in Figure 2.

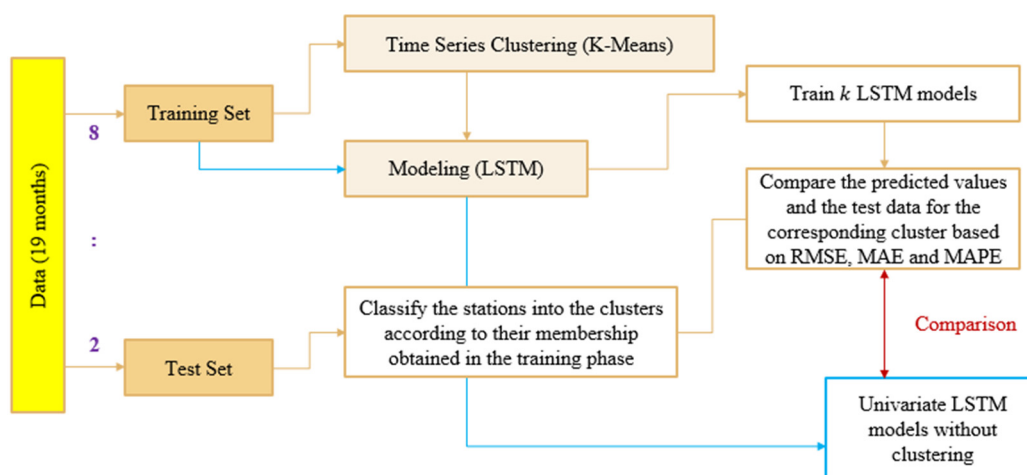


Figure 2. Flow chart of the framework.

As the first step to constructing the proposed model, the air quality monitoring stations were grouped into k clusters by utilizing the time series k-means clustering approach based on the training set. Then, a multivariate LSTM model was trained for each cluster. Combined with the clustering results, the observed values in the test set were compared with the corresponding predicted values based on RMSE, MAE and MAPE.

After that, a univariate LSTM model was constructed independently for each station by using the same hyperparameter settings with its corresponding hybrid model. Hence, a total of 60 univariate LSTM models were built. Similar to the proposed model, the prediction performance for each univariate model was measured based on three accuracy metrics. Lastly, the prediction accuracy was compared between both models by using RPD.

3. Results and Discussion

3.1. Descriptive Analysis

The dataset was split into a training set and a test set by a ratio of 8:2, where the training set consists of data ranging from 5 July 2017 to 30 September 2018 and the test set comprises the last four months, that is, from 1 October 2018 to 31 January 2019.

Table 1 shows the minimum value, maximum value and quartiles for the whole dataset.

Table 1. Minimum value, maximum value and quartiles for the whole dataset ($\mu\text{g}/\text{m}^3$).

Minimum	1st Quartile	Median	3rd Quartile	Maximum
4.37	16.29	21.87	29.38	235.72

3.2. Time Series K-Means Clustering

Before the clustering and modeling phases were carried out, the training set was scaled into a range of [0, 1] by adopting min–max normalization. Then, the 60 monitoring stations were clustered based on the k-means algorithm. To identify the optimum k clusters, the values of WGSS were calculated and visualized in Figure 3 for $k = 1, 2, \dots, 10$.

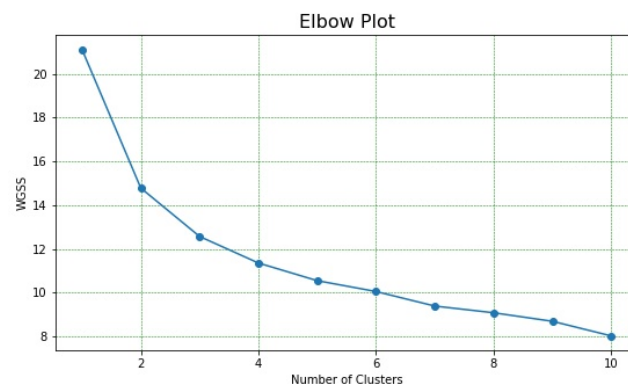


Figure 3. Elbow plot.

By using the elbow method, the optimum number of clusters was estimated to be between $k = 2, 3$ and 4 . To further validate the goodness of separation, the silhouette index was applied to the identified candidates. Table 2 shows the silhouette scores for each number of clusters.

Table 2. Silhouette scores for each number of clusters.

k	2	3	4
Silhouette Score	0.2628	0.1742	0.1532

Based on the table above, $k = 2$ has the highest silhouette score, while $k = 4$ has the lowest index. A higher index score indicates a better partitioning of the data, hence $k = 2$ is said to be the optimum number of clusters.

The clustering results show that Cluster 1 consists of 19 stations, whereas Cluster 2 comprises 41 stations. Table 3 lists the cluster membership for the daily average PM_{10} concentration according to the stations.

Table 3. Cluster membership for daily average PM_{10} concentration according to stations.

Station	Station Location	Station Category	Longitude	Latitude	Cluster
CA01R	Kangar, Perlis	Suburban	100.2111	6.429922	2
CA02K	Langkawi, Kedah	Suburban	99.85846	6.331539	2
CA03K	Alor Setar, Kedah	Suburban	100.3468	6.137244	2
CA04K	Sungai Petani, Kedah	Suburban	100.4678	5.629631	2
CA05K	Kulim Hi-Tech, Kedah	Industry	100.5903	5.424147	2
CA06P	Seberang Jaya, Pulau Pinang	Urban	100.4039	5.39817	1
CA07P	Seberang Perai, Pulau Pinang	Suburban	100.4435	5.329358	1
CA09P	Balik Pulau, Pulau Pinang	Suburban	100.2147	5.337598	2
CA10A	Taiping, Perak	Suburban	100.6791	4.89885	1
CA11A	Tasek Ipoh, Perak	Urban	101.1167	4.629444	1
CA12A	Pegoh Ipoh, Perak	Suburban	101.0802	4.553336	1
CA13A	Seri Manjung, Perak	Rural	100.6634	4.200344	1
CA14A	Tanjung Malim, Perak	Suburban	101.5245	3.687758	2
CA15W	Batu Muda, Kuala Lumpur	Suburban	101.6822	3.212439	1
CA16W	Cheras, Kuala Lumpur	Urban	101.7179	3.106236	1
CA17W	Putrajaya	Suburban	101.6901	2.914816	1
CA18B	Kuala Selangor, Selangor	Rural	101.2562	3.321308	2
CA19B	Petaling Jaya, Selangor	Suburban	101.608	3.133169	1
CA20B	Shah Alam, Selangor	Urban	101.5562	3.104717	1
CA21B	Klang, Selangor	Suburban	101.4131	3.014889	1

Table 3. Cont.

Station	Station Location	Station Category	Longitude	Latitude	Cluster
CA22B	Banting, Selangor	Suburban	101.6232	2.816689	1
CA23N	Nilai, Negeri Sembilan	Suburban	101.8115	2.821692	1
CA24N	Seremban, Negeri Sembilan	Urban	101.9685	2.723381	2
CA25N	Port Dickson, Negeri Sembilan	Suburban	101.8669	2.441383	2
CA26M	Alor Gajah, Melaka	Rural	102.2246	2.370925	2
CA27M	Bukit Rambai, Melaka	Suburban	102.1727	2.258519	1
CA28M	Bandaraya Melaka, Melaka	Urban	102.2571	2.190936	2
CA29J	Segamat, Johor	Suburban	102.8627	2.493914	2
CA31J	Batu Pahat, Johor	Suburban	102.8666	1.919323	2
CA32J	Kluang, Johor	Rural	103.3121	2.037882	2
CA33J	Larkin, Johor	Urban	103.736	1.494625	1
CA34J	Pasir Gudang, Johor	Urban	103.8935	1.470122	1
CA35J	Pengerang, Johor	Industry	104.1496	1.389489	2
CA36J	Kota Tinggi, Johor	Suburban	104.2253	1.564056	2
CA37C	Rompin, Pahang	Rural	103.4192	2.926645	2
CA38C	Temerloh, Pahang	Suburban	102.3764	3.471603	1
CA39C	Jerantut, Pahang	Suburban	102.3666	3.94836	2
CA40C	Indera Mahkota, Kuantan, Pahang	Suburban	101.9197	3.276529	2
CA41C	Balok Baru, Kuantan, Pahang	Industry	103.3622	3.951842	1
CA42T	Kemaman, Terengganu	Industry	103.4258	4.262121	2
CA43T	Paka, Terengganu	Industry	103.4348	4.598064	2
CA44T	Kuala Terengganu, Terengganu	Rural	103.1204	5.308094	2
CA45T	Besut, Terengganu	Suburban	102.5156	5.748449	2
CA46D	Tanah Merah, Kelantan	Suburban	102.1345	5.811172	2
CA47D	Kota Bahru, Kelantan	Suburban	102.2492	6.147431	2
CA48S	Tawau, Sabah	Suburban	117.9359	4.249786	2
CA49S	Sandakan, Sabah	Suburban	118.0911	5.864467	2
CA50S	Kota Kinabalu, Sabah	Suburban	116.0433	5.89372	2
CA51S	Kimanis, Sabah	Industry	115.8506	5.538225	2
CA54Q	Limbang, Sarawak	Rural	115.0137	4.758891	2
CA55Q	Permyjaya, Miri, Sarawak	Rural	114.0434	4.494791	2
CA56Q	Miri, Sarawak	Suburban	114.0124	4.424679	2
CA57Q	Samalaju, Sarawak	Industry	113.2952	3.537059	2
CA58Q	Bintulu, Sarawak	Suburban	113.0411	3.177084	2
CA59Q	Mukah, Sarawak	Rural	112.0197	2.883238	2
CA61Q	Sibu, Sarawak	Suburban	111.8319	2.314408	2
CA62Q	Sarikei, Sarawak	Rural	111.5229	2.132809	2
CA63Q	Sri Aman, Sarawak	Rural	111.4648	1.219656	2
CA64Q	Samarahan, Sarawak	Rural	110.4915	1.454853	2
CA65Q	Kuching, Sarawak	Urban	110.389	1.562229	2

Figure 4 shows the distribution of stations according to clusters.

It was found that most stations in Cluster 1 are in the more developed states along the west coast of Peninsular Malaysia, such as Selangor, Perak, Pulau Pinang and Kuala Lumpur. On the other hand, Cluster 2 is mainly made up of stations that are widely distributed in the less developed states around the east coast of Peninsular Malaysia and east Malaysia, including Terengganu, Kelantan, Sabah and Sarawak.

Moreover, the number of stations based on categories according to the clusters is shown in Figure 5.

The figure above demonstrates that most stations in Cluster 1 are located in suburban and urban areas in Klang Valley with only one station falling in the rural and industrial areas, respectively. In addition, the majority of the stations in Cluster 2 are categorized as suburban, followed by rural, industrial and urban. On top of that, it was observed that there are more stations located in suburban, rural and industrial areas in Cluster 2 as compared to Cluster 1, which has more urban stations.

After classifying the test set into the clusters, the minimum values, maximum values and quartiles according to the clusters are tabulated in Table 4.

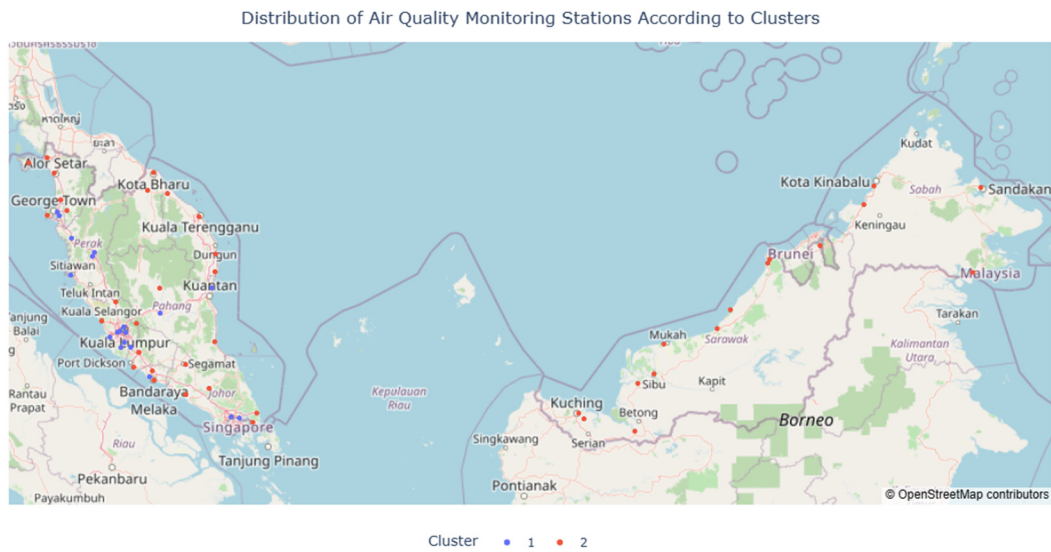


Figure 4. Distribution of air quality monitoring stations according to clusters.

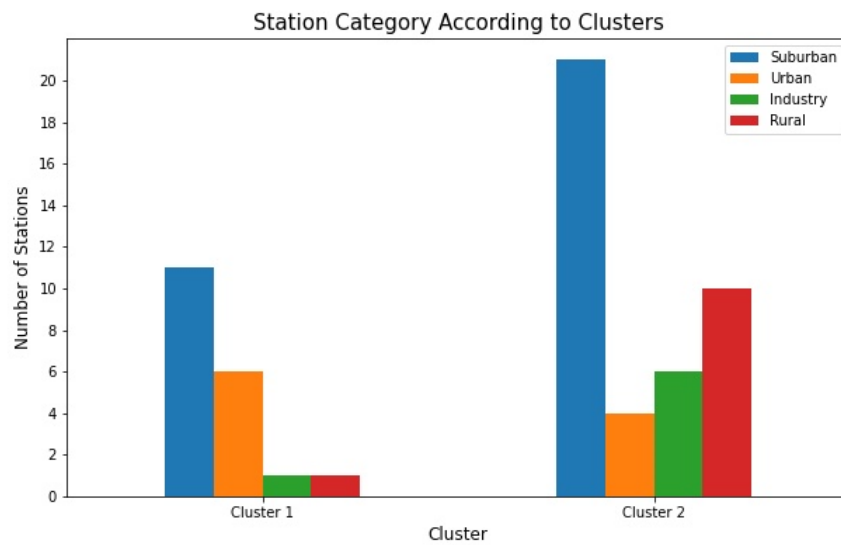


Figure 5. Bar chart for the number of stations based on categories according to clusters.

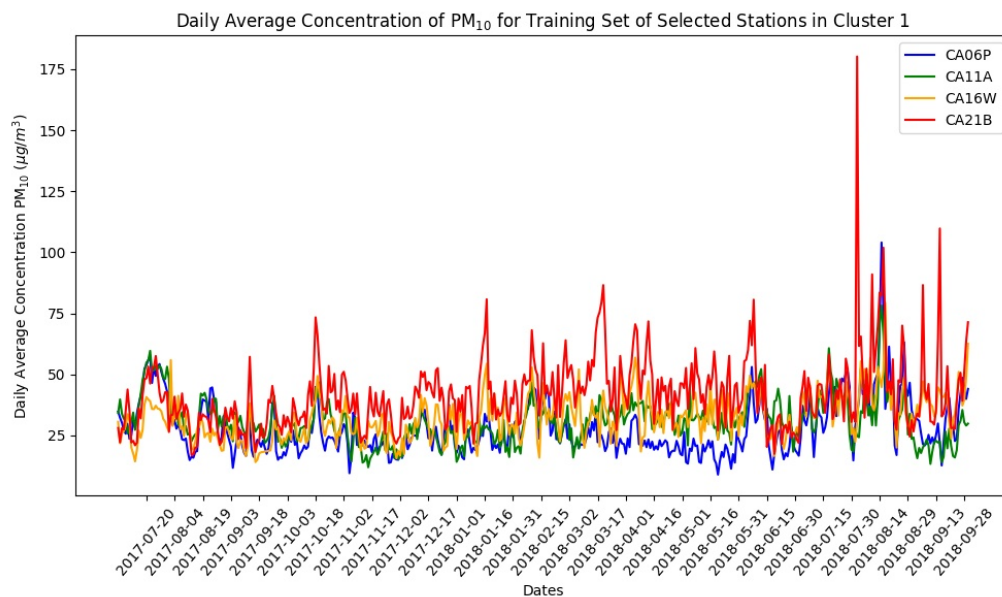
Table 4. Minimum values, maximum values and quartiles according to clusters ($\mu\text{g}/\text{m}^3$).

Element	Whole Dataset		Training Set		Test Set	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Minimum	6.57	4.37	6.57	4.37	7.32	5.92
1st Quartile	22.38	14.83	23.13	15.32	20.35	13.58
Median	28.25	19.17	29.45	20.06	24.90	16.75
3rd Quartile	37.76	25.33	37.17	26.52	30.29	21.00
Maximum	180.23	235.72	180.23	235.72	70.77	72.78

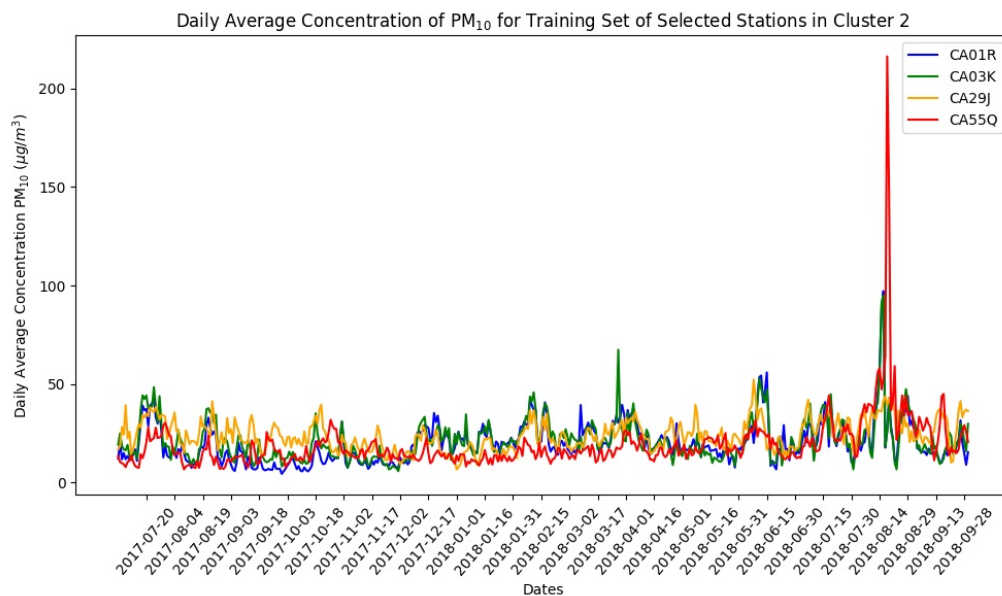
Table 4 highlights that the range of the daily average PM_{10} concentration for the whole dataset in Cluster 2, that is, $231.45 \mu\text{g}/\text{m}^3$, is much higher than the range of $173.66 \mu\text{g}/\text{m}^3$

in Cluster 1. The station locations that mainly spread in the neighboring states might give rise to this situation in accordance with a similar level of haze pollution carried by the monsoon winds [8,9]. On the other hand, the median of the daily average concentration of PM₁₀ of the whole dataset in Cluster 1 is higher than Cluster 2 by 9.08 μg/m³. Such a circumstance is believed to be closely related to the fact that most stations in Cluster 1 are in highly developed areas, including Klang Valley and Pulau Pinang [11].

The time plots of the daily average concentration of PM₁₀ for the training set and test set of the selected stations in each cluster are extracted and visualized in Figures 6 and 7, respectively.

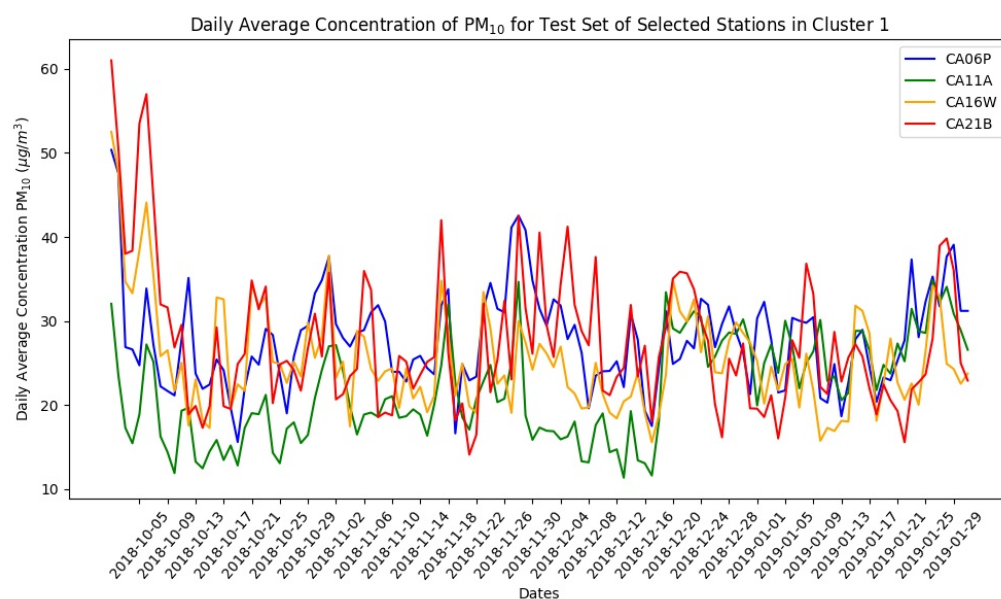


(a)

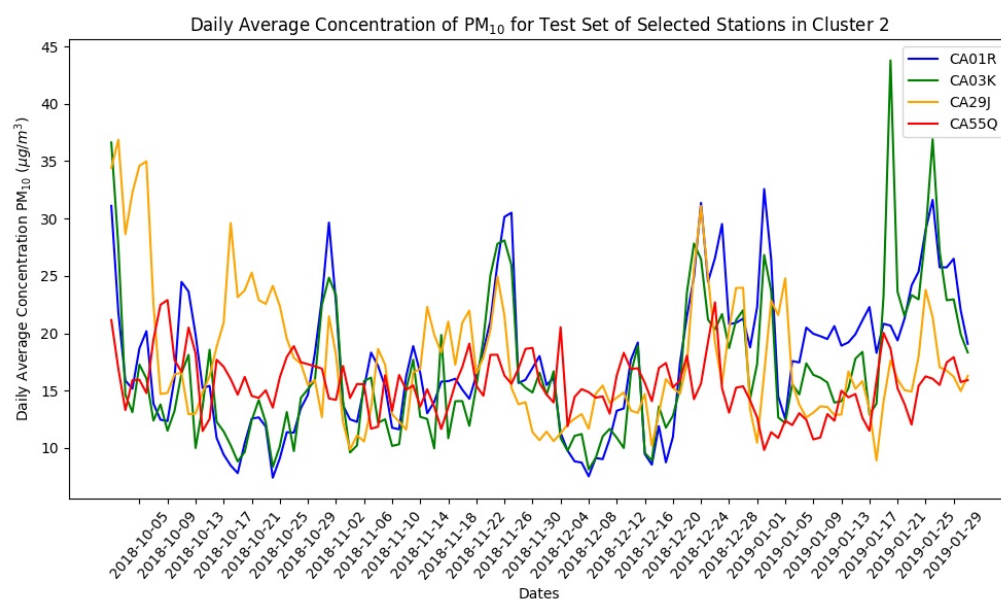


(b)

Figure 6. Time plots of daily average concentration of PM₁₀ for training set of selected stations in each cluster: (a) Cluster 1; (b) Cluster 2.



(a)



(b)

Figure 7. Time plots of daily average concentration of PM_{10} for test set of selected stations in each cluster: (a) Cluster 1; (b) Cluster 2.

From Figure 6, it can be seen that the stations within each cluster have a similar and stable time series pattern across the time range, except for a few spikes observed during a certain period. The drastic increase in the concentration of PM_{10} for both clusters around August until mid-September 2018 seems to be closely associated with the transboundary haze that affected most areas of Malaysia at that point.

According to Yusof [48], the unhealthy API readings were recorded in some states due to haze originating from North Sumatra and West Kalimantan at the time. The situation became worse and lasted until September as the southwest monsoon wind blew toward Peninsular Malaysia. Some states also experienced hot and dry climates with less rainfall, giving rise to the increase in the daytime temperature. Such weather caused wildfires in certain locations, for instance, the occurrence of peatland fires in Klang, Selangor [49]. As a

result, the air quality decreased at station CA21B in Klang, followed by an increase in the daily average concentration of PM₁₀ to the maximum value of 180.23 µg/m³ in Cluster 1.

Referring to the time plots in Cluster 2, the highest daily average concentration of PM₁₀ during the hazy period was recorded by station CA55Q, which is located in Permyjaya, Miri, Sarawak. This situation was deemed to be primarily driven by the forest fires at the nearby Industrial Training Institute, Permyjaya, which reduced the air quality in Miri and worsened the hazy conditions. According to Kawi [50], the API reading in Miri reached an unhealthy level of 130 in the morning on 19 August 2018. In conjunction with the nearly unhealthy API readings caused by the wildfire smoke from West Kalimantan, Indonesia, the PM₁₀ concentration at other stations in Sarawak, such as Bintulu, Mukah, Sibul and Sarikei, also reported an increase during the hazy period.

Generally, the values of the test set data are at a lower level compared to the training set, that is, not exceeding 75 µg/m³ in both clusters, as shown in Figure 7. It then leads to a small difference of 3.41 µg/m³ in the data range between both clusters based on Table 4.

In a nutshell, the time series k-means clustering has assigned the stations into two clusters with a size of 19 and 41 stations, respectively. This result forms the basis of the proposed model.

3.3. Construction of Hybrid Models

A multivariate LSTM model was trained based on the training set for each cluster. An optimum setting of the values of the hyperparameters was tuned manually to achieve the best model performance in the training phase. After a few trials, it was found that the models for both clusters perform well under the same hyperparameter settings as tabulated in Table 5.

Table 5. Optimum hyperparameter settings according to clusters.

Hyperparameter	Setting
Hidden layer	1
Hidden neuron	400
Dropout rate	0.1
Timestep	7
Batch size	32
Epochs	100
Activation function	Tanh
Recurrent activation	Sigmoid
Loss function	MSE
Optimizer	Adam

By applying the settings above, the MSE and RMSE, as well as the computation time were computed to evaluate the fitness of the hybrid models to the training set, as shown in Table 6.

Table 6. Model performance of hybrid models and computation time in training phase.

Hybrid Model	MSE	RMSE	Computation Time (Seconds)
Cluster 1	0.0030	0.0551	83.165
Cluster 2	0.0023	0.0481	85.072

As depicted in the table above, the RMSE values for both of the hybrid models are significantly low in the training phase, indicating that the constructed models can learn the trend of the training set well. In terms of the training time, both models required a similar duration, between 83 s and 85 s.

3.4. Construction of Univariate LSTM Models

By using the same hyperparameter settings with the corresponding hybrid models as shown in Table 5, a univariate LSTM model was constructed independently for each station. The model performance and computation time were recorded in Table 7 to assess the degree of fitness of each model to the training set.

Table 7. Model performance of univariate LSTM models and computation time in training phase.

Station	MSE	RMSE	Computation Time (Seconds)
CA01R	0.0044	0.0661	75.517
CA02K	0.0116	0.1078	80.691
CA03K	0.0059	0.0770	75.653
CA04K	0.0068	0.0822	78.791
CA05K	0.0054	0.0738	78.364
CA06P	0.0045	0.0668	83.509
CA07P	0.0056	0.0745	83.682
CA09P	0.0038	0.0616	74.093
CA10A	0.0084	0.0914	88.693
CA11A	0.0092	0.0958	91.114
CA12A	0.0056	0.0748	96.784
CA13A	0.0071	0.0841	84.607
CA14A	0.0061	0.0781	76.651
CA15W	0.0079	0.0890	84.997
CA16W	0.0115	0.1073	83.640
CA17W	0.0083	0.0910	77.432
CA18B	0.0057	0.0752	74.904
CA19B	0.0143	0.1194	87.481
CA20B	0.0138	0.1176	84.607
CA21B	0.0060	0.0777	92.034
CA22B	0.0042	0.0648	86.087
CA23N	0.0103	0.1013	85.621
CA24N	0.0121	0.1100	77.689
CA25N	0.0160	0.1266	80.627
CA26M	0.0138	0.1172	76.856
CA27M	0.0109	0.1045	89.117
CA28M	0.0098	0.0991	86.062
CA29J	0.0104	0.1019	80.175
CA31J	0.0161	0.1271	81.089
CA32J	0.0172	0.1312	79.806
CA33J	0.0159	0.1260	87.609
CA34J	0.0154	0.1240	90.253
CA35J	0.0109	0.1045	80.129
CA36J	0.0154	0.1240	80.523
CA37C	0.0071	0.0843	78.975
CA38C	0.0097	0.0986	99.205
CA39C	0.0100	0.1001	78.328
CA40C	0.0070	0.0837	77.969
CA41C	0.0075	0.0867	92.146
CA42T	0.0066	0.0812	87.247
CA43T	0.0124	0.1114	78.129
CA44T	0.0029	0.0534	78.864
CA45T	0.0083	0.0910	83.894
CA46D	0.0105	0.1027	78.864
CA47D	0.0124	0.1111	82.339
CA48S	0.0159	0.1262	83.257
CA49S	0.0129	0.1137	77.018
CA50S	0.0053	0.0730	81.258
CA51S	0.0081	0.0903	84.210

Table 7. Cont.

Station	MSE	RMSE	Computation Time (Seconds)
CA54Q	0.0063	0.0792	81.136
CA55Q	0.0019	0.0434	83.684
CA56Q	0.0089	0.0945	78.898
CA57Q	0.0093	0.0963	80.553
CA58Q	0.0083	0.0910	81.694
CA59Q	0.0045	0.0667	80.901
CA61Q	0.0053	0.0728	76.724
CA62Q	0.0067	0.0816	84.564
CA63Q	0.0056	0.0750	82.353
CA64Q	0.0053	0.0725	81.946
CA65Q	0.0063	0.0796	82.799

Overall, the RMSE values for the univariate LSTM models during the training phase are comparatively higher than the hybrid models, indicating a more unsatisfied fitness to the training set. Nevertheless, there are 38 stations with RMSE values lower than 0.1 in the training phase. In addition, about 74 s to 99 s were needed to train the univariate models.

3.5. Comparison of Prediction Performance between Hybrid Models and Univariate LSTM Models

The prediction performance was computed by comparing the predicted values and the actual test data based on three accuracy metrics, namely RMSE, MAE and MAPE. Then, the difference in prediction performance between the two models was measured based on RPD for each metric. If a model has a smaller value than another for at least two metrics, then it is said to have a better prediction performance. Moreover, a hybrid model is said to have comparable prediction accuracy to the univariate model if the RPD values are less than or equal to 50%. Table 8 displays the abovementioned values for all the stations; the smaller values of accuracy metrics and RPD values below or equal to 50% are listed in bold.

Table 8. Comparison of prediction performance between hybrid models and univariate LSTM models.

Station	Cluster	RMSE			MAE			MAPE		
		Hybrid Model	Univariate Model	RPD (%)	Hybrid Model	Univariate Model	RPD (%)	Hybrid Model	Univariate Model	RPD (%)
CA01R	2	4.5198	4.3701	3.37	3.4995	3.3726	3.69	21.5784	21.3973	0.84
CA02K	2	4.5505	4.9538	8.49	3.2950	4.1793	23.66	18.1676	29.0792	46.19
CA03K	2	4.9896	4.8972	1.87	3.6391	3.5968	1.17	23.0330	23.2614	0.99
CA04K	2	6.0298	5.3184	12.54	4.5432	3.8992	15.26	19.9591	20.2092	1.25
CA05K	2	5.0714	4.1359	20.32	3.8023	3.2244	16.45	17.0136	15.3374	10.36
CA06P	1	4.9699	4.6897	5.80	3.9496	3.5980	9.32	14.3867	13.3902	7.18
CA07P	1	5.3541	5.4275	1.36	4.3668	4.2177	3.47	18.5370	18.0430	2.70
CA09P	2	4.7474	4.2784	10.39	3.8791	3.3996	13.17	23.3050	21.9606	5.94
CA10A	1	7.9905	7.6230	4.71	6.4603	6.4188	0.65	26.4474	28.4973	7.46
CA11A	1	6.0224	5.0272	18.01	4.9381	4.0166	20.58	26.0518	21.6465	18.47
CA12A	1	5.4008	5.1911	3.96	4.1150	4.1098	0.13	15.1381	15.0629	0.50
CA13A	1	5.7971	5.4431	6.30	4.3236	4.0655	6.15	22.2513	21.1520	5.07
CA14A	2	3.0089	2.0047	40.06	2.1790	1.5337	34.76	19.3104	14.4543	28.76
CA15W	1	9.0777	5.5340	48.51	6.2347	4.2351	38.20	21.7435	18.5700	15.74
CA16W	1	5.1493	5.1184	0.60	3.9703	4.0972	3.15	16.1850	17.2971	6.64
CA17W	1	6.2387	6.3347	1.53	4.7604	4.7163	0.93	18.1534	19.0525	4.83
CA18B	2	6.7243	4.8416	32.56	5.3556	3.7750	34.62	26.6917	19.6442	30.42
CA19B	1	6.9502	6.4325	7.74	5.3705	5.1882	3.45	16.9225	18.8457	10.75
CA20B	1	11.3570	7.2624	43.98	8.7543	5.6374	43.32	24.8483	18.9975	26.69
CA21B	1	15.9908	8.9914	56.04	13.7508	7.8400	54.75	57.6036	33.6261	52.57
CA22B	1	10.5627	5.7734	58.63	8.8409	4.6535	62.06	36.0264	19.0299	61.74

Table 8. Cont.

Station	Cluster	RMSE			MAE			MAPE		
		Hybrid Model	Univariate Model	RPD (%)	Hybrid Model	Univariate Model	RPD (%)	Hybrid Model	Univariate Model	RPD (%)
CA23N	1	12.5397	7.7725	46.94	9.2090	6.2618	38.10	23.7006	19.0542	21.73
CA24N	2	8.4372	5.0106	50.96	6.0306	4.0254	39.88	26.0021	21.2193	20.26
CA25N	2	7.0380	4.2323	49.79	5.1451	3.3116	43.36	22.6067	17.5305	25.29
CA26M	2	7.5355	4.5812	48.76	5.2215	3.6684	34.94	24.0427	21.2575	12.30
CA27M	1	6.2810	4.7058	28.67	4.7925	3.8138	22.74	20.2316	17.6066	13.87
CA28M	2	7.2656	5.8701	21.25	5.9451	4.9969	17.33	39.1783	35.0929	11.00
CA29J	2	7.0781	3.6154	64.76	5.1203	2.9254	54.56	25.1293	18.1998	31.99
CA31J	2	8.6128	4.6720	59.33	6.0471	3.8098	45.40	25.4477	21.0194	19.06
CA32J	2	9.6117	4.7057	68.53	7.6127	3.6751	69.77	34.5874	21.6189	46.15
CA33J	1	14.0436	6.2091	77.37	11.4893	4.7969	82.18	36.9123	17.9316	69.22
CA34J	1	13.2536	6.1375	73.40	10.4014	4.6397	76.61	35.8422	19.7283	57.99
CA35J	2	7.4453	4.2220	55.25	5.3145	3.4483	42.59	25.3247	19.6083	25.44
CA36J	2	10.0448	3.6776	92.80	8.2307	2.7321	100.31	41.9039	15.6868	91.05
CA37C	2	7.9762	4.1406	63.31	5.7866	3.2104	57.27	26.7943	17.4523	42.23
CA38C	1	6.6168	5.6409	15.92	4.5975	4.3858	4.71	19.0795	20.4252	6.81
CA39C	2	5.9645	3.9025	41.79	3.9081	2.9436	28.15	22.9331	20.4443	11.47
CA40C	2	5.3055	3.9530	29.22	3.7472	3.0742	19.73	23.6457	21.7060	8.55
CA41C	1	6.9434	5.6108	21.23	5.7482	4.3989	26.59	27.0432	22.8814	16.67
CA42T	2	5.2559	4.4650	16.27	4.0192	3.4877	14.16	21.5831	21.0374	2.56
CA43T	2	7.3848	4.0608	58.08	5.8110	2.9586	65.05	30.1447	19.8797	41.04
CA44T	2	19.8661	7.4887	90.50	17.1922	6.1621	94.46	106.2063	41.5364	87.54
CA45T	2	6.6239	4.8986	29.95	5.0560	3.6732	31.68	29.2778	24.7883	16.61
CA46D	2	10.2958	6.9140	39.30	7.7256	5.4100	35.26	31.8290	30.6160	3.88
CA47D	2	7.7550	6.6434	15.44	5.7597	5.1561	11.06	27.9536	29.5172	5.44
CA48S	2	4.0429	2.1662	60.45	3.4659	1.6946	68.65	25.9122	15.2541	51.78
CA49S	2	5.5137	2.5615	73.12	4.7596	1.8774	86.85	27.3653	11.7014	80.19
CA50S	2	9.1110	7.0216	25.90	6.1918	5.0411	20.49	25.0603	22.1106	12.51
CA51S	2	3.8643	2.1653	56.36	3.2355	1.6343	65.76	22.3544	12.4954	56.58
CA54Q	2	4.4997	3.4582	26.17	3.4441	2.6571	25.80	22.9064	18.6653	20.40
CA55Q	2	23.3102	2.4208	162.37	20.2460	1.8668	166.23	136.0466	11.8319	168.00
CA56Q	2	7.7953	4.1198	61.69	6.6007	3.4160	63.59	29.3824	17.4739	50.83
CA57Q	2	5.5144	5.0829	8.14	3.7241	3.4618	7.30	19.9275	19.6182	1.56
CA58Q	2	8.3022	6.7478	20.66	6.4231	5.3905	17.48	31.0169	29.7875	4.04
CA59Q	2	9.3398	4.6195	67.63	7.4647	3.9870	60.74	46.9930	25.8482	58.06
CA61Q	2	6.5227	3.7648	53.62	5.1163	3.0902	49.38	29.0637	18.1786	46.08
CA62Q	2	5.2195	3.1495	49.47	4.2292	2.5453	49.71	29.8811	18.0850	49.19
CA63Q	2	4.6250	3.1070	39.26	3.6524	2.4379	39.88	23.6166	16.5186	35.37
CA64Q	2	5.3058	2.6717	66.04	4.1988	2.1914	62.83	28.9101	15.0593	63.00
CA65Q	2	6.2138	4.7521	26.66	4.6336	3.4518	29.23	26.0402	18.6347	33.15

Based on Table 8, the hybrid model has recorded a lower value for at least two accuracy metrics at two stations in Cluster 1, which are CA16W and CA17W. Despite having a better prediction performance for most stations, the univariate model does not significantly outperform the hybrid model based on RPD values. This is because the RPD values are more than 50% for at least two accuracy metrics at only four stations, which are CA21B, CA22B, CA33J and CA34J. Hence, a conclusion stating that the proposed model has a competitive prediction performance in Cluster 1 can be drawn.

On the other hand, it is highlighted that the proposed model is capable of giving a more accurate prediction for station CA02K based on much lower RMSE, MAE and MAPE values compared to the univariate model. Focusing on the RPD values, the prediction performance of the proposed model only varies significantly from the univariate model at 13 stations in Cluster 2.

There are 39 stations with an RPD less than or equal to 50% for RMSE. Among these stations, 12 of them have RPD values within 0–10%, 6 stations have RPD around 10–20%,

10 stations and 3 stations have a range of 20–30% and 30–40%, respectively, while the rest have RPD values within 40–50%. Meanwhile, most of the satisfactory RPD values based on MAE fall in the range of 0–10% (12 stations), followed by the range of 30–40% (9 stations), 10–20% and 20–30% (8 stations, respectively) and 40–50% (6 stations). Lastly, 47 stations have an RPD less than or equal to 50% for MAPE. It is observed that most of the RPD values based on MAPE fall in the range of 0–10% (18 stations), followed by 10–20% (12 stations), 20–30% (7 stations), 40–50% (6 stations) and 30–40% (4 stations). In short, the hybrid model can output a competitive prediction performance compared to the univariate model, as it records an acceptable range of RPD values based on all three metrics.

If the prediction performance of the hybrid model does not significantly vary from the univariate model based on RPD for at least two accuracy metrics at each station, then it can be concluded that the proposed model is suitable to forecast the PM_{10} concentration at that station. From Table 8, the hybrid model seems to be potentially adopted as the PM_{10} prediction model for 43 stations (71.67%), whereas the univariate LSTM model is more suitable to be employed for the stations in Johor, Terengganu and Sabah.

Figure 8 shows the actual and predicted values for selected stations from both clusters. Both models can fit the actual data trend well for stations CA10A (Cluster 1) and CA01R (Cluster 2). Plots from other stations were also investigated and similar results were observed.

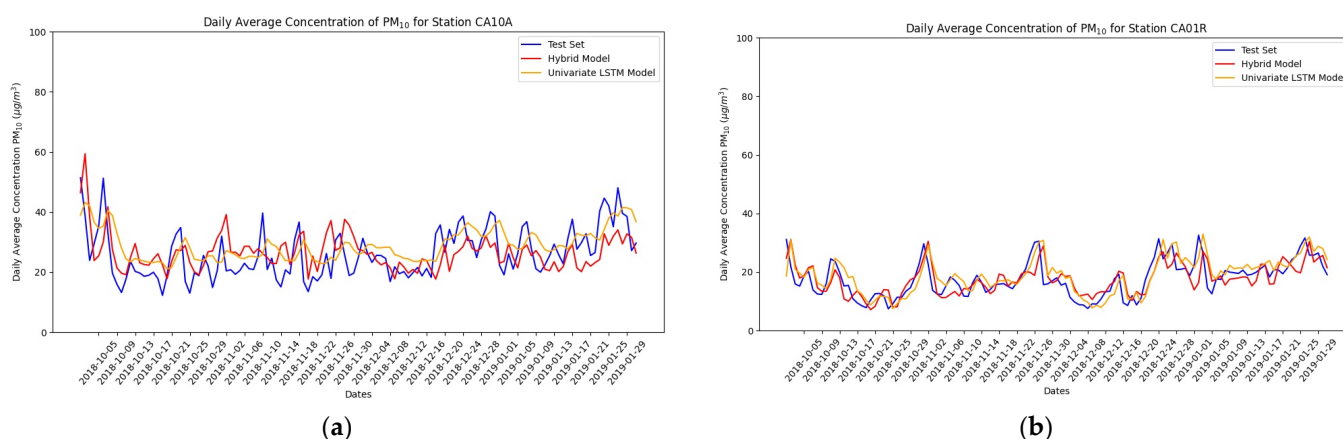


Figure 8. Actual and predicted values for selected stations from each cluster: (a) CA10A (Cluster 1); (b) CA01R (Cluster 2).

To summarize, the prediction accuracy of the hybrid model does not significantly deviate from the univariate model, as the RPD values are within the 50% acceptable range at 43 stations for 71.67% of the stations. This has proven the capability of the hybrid model to predict the PM_{10} concentration at a similar accuracy level to the univariate model. Furthermore, the hybrid model can capture and fit the actual data trend quite well for most stations with a rather shorter computation time than the univariate LSTM model. This is closely related to the fact that only one hybrid model is constructed for each cluster, whereas the univariate model is individually constructed for each station, leading to a total model training time of 4951.842 s for 60 univariate models and just 168.237 s for two hybrid models. Such a rather shorter computation time without any drawback on prediction performance or trend fitness has made the hybrid model a more ideal forecasting model.

Nevertheless, the occurrence of hazy conditions at certain periods in the training set that negatively affected the air quality of each location at different levels is one of the factors that leads to a better prediction accuracy of the univariate LSTM model for some stations. The PM_{10} concentration increases drastically during hazy days in conjunction with the high emissions of particulate matter and greenhouse gases. On the other hand, PM_{10} is at a low concentration during normal days as the aerosol particles are released by mobile sources, including motor vehicles, and stationary sources, such as factories [6]. Due to the

nature of the hybrid model that uses the data from all the stations within the same clusters to predict the PM_{10} values without considering much about the localized pollution level as in the univariate model, this might cause the tendency to overestimate PM_{10} for some stations that are less affected by the transboundary haze.

In addition, the concentration of PM_{10} is mainly influenced by other meteorological factors, such as wind speed, temperature and relative humidity [6]. The concentration of particulates is found to have a correlation with the temperature, wind speed, dew point and air pressure [6,19]. In accordance with this, Zhang et al. [51] found that there is a significant correlation between particulates and relative humidity during the winter season in Nanyang. Meanwhile, Pineda Rojas et al. [7] also revealed that the high daily average PM_{10} concentration is often recorded when the sky cover and relative humidity are low. Similar to the finding that the PM_{10} concentration is high during the southwest monsoon season [9], Yassen and Jahi [8] discovered that the TSP concentration in Klang Valley is higher during that season as compared to the rainy season. Thus, it can be concluded that different real-time meteorological conditions at each station will influence the concentration of particulate matter and lead to a slightly lower prediction accuracy of the hybrid model for some stations.

4. Conclusions

In brief, this study proposed a novel hybrid model that combines both the k-means clustering technique and the state-of-the-art LSTM model in predicting the daily average PM_{10} concentration in Malaysia. Throughout the study, comparisons were made between the hybrid model and the univariate LSTM model in terms of prediction performance, trend fitting and computation time.

In this study, 60 air quality monitoring stations were divided into two distinct clusters by adopting the time series k-means clustering method. Cluster 1 consists of 19 stations that are mainly distributed in highly developed areas, such as Klang Valley and Pulau Pinang, such that most of them fall under the urban and suburban categories. On the other hand, Cluster 2 comprises 41 suburban and rural stations that are located mainly on the east coast of Peninsular Malaysia, Sabah and Sarawak. The within-cluster time series patterns are quite similar and relatively stable with a few unexpected spikes, especially during the transboundary hazy period.

The results show that the hybrid model can give a comparable prediction performance to the univariate LSTM model based on the RPD values for three accuracy metrics. In terms of fitting the actual trend, the hybrid model can capture the patterns of daily average PM_{10} concentration, although it gives a poorer result compared to the univariate model for some stations due to several factors, such as the hazy period in the training set that contaminated the air quality at a different level and the varying meteorological conditions at each location. In addition, the hybrid model significantly outperforms the univariate LSTM model based on its much shorter training time, suggesting the capability of the proposed model to effectively increase the prediction efficiency in real-life applications.

As for the future research direction, it is suggested to consider the other meteorological factors, especially wind speed, during the clustering phase to reduce their impacts on the PM_{10} concentration. Moreover, the hourly PM_{10} concentration also warrants further study so that the public can better plan their daily activities beforehand. In such a context, two-step k-means clustering could be implemented to better capture the variation in the PM_{10} concentration before constructing the forecasting model for each subclass of the main clusters. Last but not least, a comparison between hybrid models that employ different forecasting methods, such as ARIMA, gated recurrent unit (GRU) and LSSVM models, can be carried out to identify which combination of techniques can predict the PM_{10} concentration better.

Author Contributions: Conceptualization, N.M.A. and H.Y.L.; methodology, N.M.A. and H.Y.L.; software, H.Y.L.; validation, N.M.A. and M.A.A.B.; formal analysis, H.Y.L.; investigation, H.Y.L.; resources, N.M.A. and M.A.A.B.; data curation, N.M.A. and H.Y.L.; writing—original draft preparation,

H.Y.L.; writing—review and editing, N.M.A. and M.A.A.B.; visualization, H.Y.L.; supervision, N.M.A. and M.A.A.B.; project administration, N.M.A.; funding acquisition, M.A.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Kebangsaan Malaysia with the grant number GP-K017073.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data were obtained from the Malaysian Department of Environment (DOE) and are available from DOE upon request.

Acknowledgments: The authors would like to express their utmost gratitude to the Malaysian Department of Environment (DOE) for providing the air quality data used in this study. In addition, the authors would also like to thank Universiti Kebangsaan Malaysia for the allocation of the research grant, GP-K017073.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. Air Pollution. Available online: <https://www.who.int/health-topics/air-pollution> (accessed on 15 May 2022).
2. Kamaruddin, S.B. UKM Pakarunding Kaji Semula Cara Nilai Kualiti Udara. Available online: https://www.ukm.my/news/Latest_News/ukm-pakarunding-kajli-semula-cara-nilai-kualiti-udara/ (accessed on 15 May 2022).
3. Rani, N.L.A.; Azid, A.; Khalit, S.I.; Juahir, H.; Samsuding, M.S. Air Pollution Index Trend Analysis in Malaysia, 2010–2015. *Pol. J. Environ. Stud.* **2018**, *27*, 801–807. [CrossRef]
4. Malaysian Department of Environment (DOE). Pengiraan Indeks Pencemar Udara (IPU). Available online: http://apims.doe.gov.my/pdf/API_Calculation.pdf (accessed on 20 January 2023).
5. Al Jallad, F.; Al Katheeri, E.; Al Omar, M. Concentrations of Particulate Matter and Their Relationships with Meteorological Variables. *Sustain. Environ. Res.* **2013**, *23*, 191–198.
6. Chooi, Y.H.; Yong, E.L. The Influence of PM_{2.5} and PM₁₀ on Air Pollution Index (API). In Proceedings of the Civil Engineering Research Work: Environmental Engineering, Hydraulics & Hydrology, UTM, Johor Bahru, Malaysia, 7–8 June 2016; pp. 132–143.
7. Pineda Rojas, A.L.; Borge, R.; Mazzeo, N.A.; Saurral, R.I.; Matarazzo, B.N.; Cordero, J.M.; Kropff, E. High PM₁₀ Concentrations in the City of Buenos Aires and Their Relationship with Meteorological Conditions. *Atmos. Environ.* **2020**, *241*, 117773. [CrossRef]
8. Yassen, M.E.; Jahi, J.M. Investigation of Variations and Trends in TSP Concentrations in the Klang Valley Region, Malaysia. *Malays. J. Environ. Manag.* **2007**, *8*, 57–68.
9. Rahman, S.R.A.; Ismail, S.N.S.; Raml, M.F.; Latif, M.T.; Abidin, E.Z.; Praveena, S.M. The Assessment of the Ambient Air Pollution Trend in Klang Valley, Malaysia. *World Environ.* **2015**, *5*, 1–11.
10. Ao, D.; Cui, Z.; Gu, D. Hybrid Model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network. In Proceedings of the 38th Chinese Control Conference, Guangzhou, China, 27–30 July 2019; pp. 8416–8421.
11. Rahman, E.; Hamzah, F.M.; Latif, M.T.; Dominick, D. Assessment of PM_{2.5} Patterns in Malaysia Using the Clustering Method. *Aerosol Air Qual. Res.* **2022**, *22*, 210161. [CrossRef]
12. Ariff, N.M.; Bakar, M.A.A.; Zamzuri, Z.H. Academic Preference Based on Students' Personality Analysis through K-Means Clustering. *Malays. J. Fund. Appl. Sci.* **2020**, *16*, 328–333. [CrossRef]
13. Shafi, J.; Waheed, A. K-Means Clustering Analysing Abrupt Changes in Air Quality. In Proceedings of the Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; pp. 26–30.
14. Kim, S.B.; Park, S.K.; Sattler, M.; Russell, A.G. Characterization of Spatially Homogeneous Regions Based on Temporal Patterns of Fine Particulate Matter in the Continental United States. *J. Air Waste Manag. Assoc.* **2008**, *58*, 965–975. [CrossRef] [PubMed]
15. Beaver, S.; Palazoglu, A. A Cluster Aggregation Scheme for Ozone Episode Selection in the San Francisco, CA Bay Area. *Atmos. Environ.* **2006**, *40*, 713–725. [CrossRef]
16. Aghabozorgi, S.; Shirkhorshidi, A.S.; Teh, Y.W.; Soltanian, H.; Herawan, T. Spatial and Temporal Clustering of Air Pollution in Malaysia: A Review. In Proceedings of the International Conference on Agriculture, Environment and Biological Sciences (ICFAE'14), Antalya, Turkey, 4–5 June 2014; pp. 67–72.
17. Suris, F.N.A.; Bakar, M.A.A.; Ariff, N.M.; Mohd Nadzir, M.S.; Ibrahim, K. Malaysia PM₁₀ Air Quality Time Series Clustering Based on Dynamic Time Warping. *Atmosphere* **2022**, *13*, 503. [CrossRef]
18. Bhalgat, P.; Pitale, S.; Bhoite, S. Air Quality Prediction Using Machine Learning Algorithms. *Int. J. Comput. Appl. Technol. Res.* **2019**, *8*, 367–370. [CrossRef]
19. Aditya, C.R.; Chandana, R.D.; Nayana, D.K.; Praveen, G.V. Detection and Prediction of Air Pollution Using Machine Learning Models. *Int. J. Eng. Trends Technol.* **2018**, *59*, 204–207.

20. Guo, B.; Wang, X.; Pei, L.; Su, Y.; Zhang, D.; Wang, Y. Identifying the spatiotemporal dynamic of PM_{2.5} concentrations at multiple scales using geographically and temporally weighted regression model across China during 2015–2018. *Sci. Total Environ.* **2021**, *751*, 141765. [[CrossRef](#)] [[PubMed](#)]
21. Guo, B.; Zhang, D.; Pei, L.; Su, Y.; Wang, X.; Bian, Y.; Zhang, D.; Yao, W.; Zhou, Z.; Guo, L. Estimating PM_{2.5} concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017. *Sci. Total Environ.* **2021**, *778*, 146288. [[CrossRef](#)]
22. Guo, B.; Wu, H.; Pei, L.; Zhu, X.; Zhang, D.; Wang, Y.; Luo, P. Study on the spatiotemporal dynamic of ground-level ozone concentrations on multiple scales across China during the blue sky protection campaign. *Environ. Int.* **2022**, *170*, 107606. [[CrossRef](#)]
23. Sharma, R.; Shilimkar, G.; Pisal, S. Air Quality Prediction by Machine Learning. *Int. J. Sci. Res. Sci. Technol.* **2021**, *8*, 486–492. [[CrossRef](#)]
24. Uh, B.H.; Majid, N. Comparison of ARIMA Model and Artificial Neural Network in Forecasting Gold Price. *J. Qual. Meas. Anal.* **2021**, *17*, 31–39.
25. Chee, K.C.; Omar, N. Bitcoin Price Prediction Based on Sentiment of News Article and Market Data with LSTM Model. *Asia-Pac. J. Inf. Technol. Multimed.* **2020**, *9*, 1–16.
26. Tsan, Y.T.; Chen, D.Y.; Liu, P.Y.; Kristiani, E.; Nguyen, K.L.P.; Yang, C.T. The Prediction of Influenza-Like Illness and Respiratory Disease Using LSTM and ARIMA. *Int. J. Environ. Res. Public Health* **2022**, *19*, 1858. [[CrossRef](#)]
27. Khumaidi, A.; Raafi'udin, R.; Solihin, I.P. Pengujian Algoritma Long Short Term Memory untuk Predikasi Kualitas Udara dan Suhu Kota Bandung. *J. Telematika* **2020**, *15*, 13–18.
28. Bakar, M.A.A.; Ariff, N.M.; Mohd Nadzir, M.S.; Ong, L.W.; Suris, F.N.A. Prediction of Multivariate Air Quality Time Series Data Using Long Short-Term Memory Network. *Mal. J. Fund. Appl. Sci.* **2022**, *18*, 52–59. [[CrossRef](#)]
29. Zhang, Y.; Yang, M.; Yang, F.; Dong, N. A Multi-Step Prediction Method of Urban Air Quality Index Based on Meteorological Factors Analysis. In Proceedings of the International Conference on Environment, Renewable Energy and Green Engineering (EREGCE 2022), Online, China, 22–24 April 2022; p. 01010.
30. Azid, A.; Juahir, H.; Toriman, M.E.; Kamarudin, M.K.A.; Saudi, A.S.M.; Hasnam, C.N.C.; Aziz, N.A.A.; Azaman, F.; Latif, M.T.; Zainuddin, S.F.M.; et al. Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: A Case Study in Malaysia. *Water Air Soil Pollut.* **2014**, *225*, 2063. [[CrossRef](#)]
31. Yulita, I.N.; Fanany, M.I.; Arymurthy, A.M. Fuzzy Clustering and Bidirectional Long Short-Term Memory for Sleep Stages Classification. In Proceedings of the 2017 International Conference on Soft Computing, Intelligent System and Information Technology, Denpasar, Bali, Indonesia, 26–29 September 2017; pp. 11–16.
32. Liu, L.; Jahromi, H.M.; Cai, L.; Kidston, D. Hierarchical Agglomerative Clustering and LSTM-Based Load Prediction for Dynamic Spectrum Allocation. In Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2021; pp. 1–6.
33. Li, R.; Hu, Y.; Liang, Q. T2F-LSTM Method for Long-Term Traffic Volume Prediction. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 3256–3264. [[CrossRef](#)]
34. Baca, H.A.H.; Valdivia, F.d.L.P.; Ibarra, M.J.; Cruz, M.A.; Baca, M.E.H. Air Quality Prediction Based on Long Short-Term Memory (LSTM) and Clustering K-Means in Andahuaylas, Peru. In Proceedings of the 2021 Future of Information and Communication Conference (FICC): Advances in Information and Communication, Vancouver, Canada, 29–30 April 2021; pp. 179–191.
35. Chen, H.; Guan, M.; Li, H. Air Quality Prediction Based on Integrated Dual LSTM Model. *IEEE Access* **2021**, *9*, 93285–93297. [[CrossRef](#)]
36. Kobylin, O.; Lyashenko, V. Time Series Clustering Based on the K-Means Algorithm. *J. La Multiapp* **2020**, *1*, 1–7. [[CrossRef](#)]
37. Lkhagva, B.; Suzuki, Y.; Kawagoe, K. New Time Series Data Representation ESAX for Financial Applications. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; pp. 17–22.
38. Sardá-Espinosa, A. Time-Series Clustering in R Using the dtwclust Package. *R. J.* **2019**, *11*, 22–43. [[CrossRef](#)]
39. Hautamaki, V.; Nykanen, P.; Franti, P. Time-Series Clustering by Approximate Prototypes. In Proceedings of the 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
40. Maharaj, E.A.; D'Urso, P.; Caiado, J. *Time Series Clustering and Classification*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2019.
41. Aghabozorgi, S.; Shirkhorshidi, A.S.; Teh, Y.W. Time-Series Clustering—A Decade Review. *Inf. Syst.* **2015**, *53*, 16–38. [[CrossRef](#)]
42. Bhardwaj, A. Silhouette Coefficient. Available online: <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c> (accessed on 31 May 2022).
43. Denyse. Time Series Clustering—Deriving Trends and Archetypes from Sequential Data. Available online: <https://towardsdatascience.com/time-series-clustering-deriving-trends-and-archetypes-from-sequential-data-bb87783312b4> (accessed on 31 May 2022).
44. Colah. Understanding LSTM Networks. Available online: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (accessed on 31 May 2022).
45. Vijay, U. Early Stopping to Avoid Overfitting in Neural Network—Keras. Available online: <https://medium.com/zero-equals-false/early-stopping-to-avoid-overfitting-in-neural-network-keras-b68c96ed05d9> (accessed on 10 January 2023).
46. NC State University Physics Department. Percent Error and Percent Difference. Available online: https://www.webassign.net/question_assets/ncsucalcpphysmechl3/percent_error/manual.html (accessed on 10 January 2023).

47. Northern Territory Department of Lands, Planning and the Environment (DLPE). Appendix D—Data Quality Objectives, Quality Assurance, Quality Control. Available online: https://ntepa.nt.gov.au/__data/assets/pdf_file/0003/286149/Edith-River-Investigation-Report (accessed on 10 January 2023).
48. Yusof, N.A.M. Jerebu Akibat Kebakaran di Sumatera dan Kalimantan. Available online: <https://www.bharian.com.my/berita/nasional/2018/08/463184/jerebu-akibat-kebakaran-di-sumatera-dan-kalimantan> (accessed on 10 January 2023).
49. Nufael, A. Malaysia Alami Jerebu Akibat Pembakaran Terbuka di Kalimantan. Available online: <https://www.benarnews.org/malay/berita/my-jerebu-180817-08172018183152.html> (accessed on 10 January 2023).
50. Kawi, M.R. IPU Sarawak Naik, Miri Catat Bacaan Tidak Sihat. Available online: <https://www.bharian.com.my/berita/wilayah/2018/08/463688/ipu-sarawak-naik-miri-catat-bacaan-tidak-sihat> (accessed on 10 January 2023).
51. Zhang, M.; Chen, S.; Zhang, X.; Guo, S.; Wang, Y.; Zhao, F.; Chen, J.; Qi, P.; Lu, F.; Chen, M. Characters of Particulate Matter and Their Relationship with Meteorological Factors during Winter Nanyang 2021–2022. *Atmosphere* **2023**, *14*, 137. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.