RESEARCH ARTICLE

Prediction of Multivariate Air Quality Time Series Data using Long Short-Term Memory Network

Mohd Aftar Abu Bakar[a], Noratiqah Mohd Ariff[a,*], Mohd Shahrul Mohd Nadzir[b], Ong Li Wen[a], Fatin Nur Afiqah Suris[a]

[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

Abstract Malaysia often suffers from haze problems almost every year. Therefore, there is a need for good air quality forecasting model for monitoring and management purposes. In this study, the air quality model based on the Long Short-Term Memory Network (LSTM) and Auto-Regressive Integrated Moving Average (ARIMA) was developed. The prediction of the particulate matter 10 micrometers or less in diameter (PM10) in Malaysia could be made from both models, and their performance was compared. The purpose of comparison between the two models was to determine the most suitable model to use in predicting PM10 since it is the dominant pollutant in Malaysia most of the time, especially during the haze period. This study used air quality data obtained from the Department of Environment Malaysia from July 2017 to June 2019. The results showed that forecasting for PM10 using multivariate LSTM model was better than the univariate LSTM model and univariate ARIMA model with the lowest root mean square error (RMSE) for those selected stations. The model with a lower RMSE value means better models and provide higher accuracy in forecasting for PM10.

Keywords: air quality, Long Short-Term Memory Network (LSTM), Auto-Regressive Integrated Moving Average (ARIMA), forecasting model, multivariate.

Introduction

The rapid process of urbanization and industrialization is one of the main factors that cause air pollution. According to Gurjar et al. [1], air pollution affects human health and the environment. Currently, studies on air pollution and air quality prediction are a field of interest [2, 3]. However, most studies depend on mathematical functions or simulation techniques in order to represent the air pollution scenario [4]. For example, Dong et al. [5] introduce the use of hidden semi-Markov models (HSMMs) to forecast the density of $PM_{2.5}$. Meanwhile, a study by Donnelly et al. [6] suggests the use of integrated parametric and non-parametric regression methods to predict air quality with high accuracy and computational efficiency. These techniques use classic machine learning algorithm. However, statistical methods and classical machine learning models are deemed unable to represent or forecast air pollution accurately because air pollution are usually influenced by the weather, transportations, and several other factors [7].

In order to overcome this, hybrid models and deep learning techniques are suggested to develop new models for air quality prediction. For example, Wang et al. [8] uses a hybrid model comprised of a

decomposition method, a modified optimization algorithm and the least squares support vector machine (LSSVM) to forecast air pollutions at points or intervals. Meanwhile, deep learning techniques are methods based on Artificial Neural Networks (ANN). ANN have been used in many fields for various applications such as to predict the density of hourly PM_{10} by combining meteorological variables and time scales as input that are chosen through the optimization of genetic algorithm [9]. Though ANN have been shown to perform better than conventional methods [10], it is unable to keep the output of previous layer and send them as information for the input in the next layer to predict the next layer's output [11]. This is improvised by putting a hidden state vector in the network which is then known as the Recurrent Neural Network (RNN).

Long Short-Term Memory (LSTM) uses the architecture of artificial recurrent neural network (RNN) for sequence prediction problems. LSTM was introduced by Hochreiter and Schmidhuber [12]. LSTM is designed to avoid long-term dependency problem by remembering information for long periods of time and handling the vanishing gradient problem of RNN. According to Salman et al. [13], the LSTM model differs from the RNN model because it contains a cell state in its recurrent hidden layer. Since LSTM is capable of learning long-term dependencies, this approach has a better accuracy compared to the RNN. LSTM have been used in many areas of study including acoustic for voice recognition [14], meteorology for weather prediction [13] and hydrology for rainfall modelling [15].

This study aims to build a multivariate air quality time series data model in Malaysia based on Long Short-Term Memory Network (LSTM). First, the common Autoregressive Integrated Moving Average (ARIMA) model is compared to the univariate air quality time series data model based on Long Short-Term Memory Network (LSTM) in predicting PM_{10}. Then, the univariate air quality time series data model based on Long Short-Term Memory Network (LSTM) is compared to the multivariate air quality time series data model based on Long Short-Term Memory Network (LSTM) in predicting PM_{10}.

Materials and methods

Data

The data used is air quality data in Malaysia with variables comprising of particles with a diameter of 10 micrometers or less (PM_{10}), wind direction, wind speed, ambient temperature, and relative humidity. These five variables were observed because PM_{10} is a small particle and is very sensitive to the surrounding conditions. Furthermore, the concentration of fine dust such as PM_{10} is the dominant pollutant, especially during the haze in Malaysia [16]. Based on Chooi and Yong [17], air pollution is greatly influenced by meteorological factors such as temperature, humidity, and wind speed. Therefore, in this study, these factors were considered in modelling PM_{10}.

There were 65 air quality monitoring stations obtained from the Department of Environment Malaysia. However, only a few stations had been selected based on data adequacy. Stations were selected based on industrial, residential, and commercial areas [18]. In this study, five stations which are station Kulim Hi-Tech, Kedah; station Kuala Selangor, Selangor; station Segamat, Johor; station Kuala Terengganu, Terengganu; and station Tanah Merah, Kelantan were selected. The duration of the data used in this study is two years starting from July 2017 until June 2019.

Before performing the analysis, the data set at each station is normalized using the min-max normalization technique to rescale the data into a close interval of [0, 1]. Then the data set is divided such that the first 21 months are used as the training set and the last 3 months are taken as the testing set.

Long Short-Term Memory Network (LSTM)

LSTM consists of hidden states like RNN and additional cell states. The hidden state works as short-term memory while the cell state works as long-term memory of the network. The cell state or also known as LSTM cell comprises of three parts referred to as gates; Forget gate, Input gate and Output gate. These gates control the flow of information into and out of the cell at each time step t.

The Forget gate determine whether the information from the previous time step should be kept or forgotten. For time step and cell state, the Forget gate can be found with the following equation [19]:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \tag{1}$$

with σ is a sigmoid function, $\boldsymbol{x}^{(t)}$ and $\boldsymbol{h}^{(t)}$ is the input and hidden layer vector at the current time step t respectively, while $\boldsymbol{b}^f$, $\boldsymbol{U}^f$ and $\boldsymbol{W}^f$ are biases, input weights and recurrent weights for the Forget gates respectively.

The Input gate controls the flow of input into the cell and is obtained similarly to the Forget gate with [19]

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right) \tag{2}$$

where $\boldsymbol{b}^g$, $\boldsymbol{U}^g$ and $\boldsymbol{W}^g$ are biases, input weights and recurrent weights for the Input gates respectively while the others are as in equation (1). Then, the cell state at current time step is updated with information from the Forget and Input gate such that

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \tag{3}$$

with $\boldsymbol{b}$, $\boldsymbol{U}$ and $\boldsymbol{W}$ denote biases, input weights and recurrent weights into the LSTM cell.

Finally, the Output gate determines the flow of output to another cell with similar equation as the two previous gates [19]

$$o_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \tag{4}$$

where biases, input weights and recurrent weights for the Output gates are denoted by $\boldsymbol{b}^o$, $\boldsymbol{U}^o$ and $\boldsymbol{W}^o$ respectively. Then, the hidden state for the current time step is calculated as

$$h_i^{(t)} = o_i^{(t)} \tanh\left(s_i^{(t)}\right) \tag{5}$$

In this study, both univariate and multivariate LSTM are considered. In building the univariate LSTM model, only variable PM_{10} was considered as the input. Meanwhile, the multivariate LSTM model, PM_{10}, wind direction, wind speed, ambient temperature and relative humidity were considered as the input variables. This is because PM_{10} may be influenced by meteorological factors such as temperature, wind speed and humidity.

For both univariate and multivariate LSTM, the architecture and parameters used are the same. A dropout with probability of 0.3 was applied to avoid overfitting of the deep learning model [7]. Both the proposed univariate and multivariate LSTM model have two hidden layers of 128 and 64 cells respectively. The two hidden layers were used to learn the temporal characteristics in building the LSTM models for each station, and the Adaptive Movement Estimation algorithm (Adam) was used to update the weightage of the neural network based on the training data. The time step used in all five stations were the same, which is 24 hours. However, the batch size used to build the LSTM models for each station was different. The univariate and multivariate LSTM model for Kulim Hi-Tech station and Kuala Selangor station were built with a batch size of 64. In contrast, the LSTM models for Segamat station, Kuala Terengganu station and Tanah Merah station were built with a batch size of 128. These models were trained for 100 epochs because too many epochs would cause overfitting. After building the model using training data, the prediction of PM_{10} was carried out. A set of predicted values was generated and compared with test data. After that, the root mean square error (RMSE) between the predicted values and test data was calculated as follows

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

with n = number of values, $\hat{y}_i$ = predicted value and y_i = test data value. The model with lower RMSE value is taken as the better model which gives higher accuracy in predicting PM_{10}.

Autoregressive Integrated Moving Average (ARIMA) models

The univariate LSTM models in this study are compared to the common univariate Autoregressive Integrated Moving Average (ARIMA) models. ARIMA models which is based on the Box-Jenkins method [20, 21] is among the most used method in modeling and forecasting time series data. Transformation by using differentiation needs to be done if the data is nonstationary before fitting the ARIMA model to the data set under study. The ARIMA(p, d, q) model can be written as

$$\theta_p(B)Y_t = \phi_q(B)\omega_t \tag{7}$$

with

$$Y_t = (1 - B)^d X_t,$$
$$\theta_p(B)X_t = \left(1 - a_1 B - a_2 B^2 - \cdots - a_p B^p\right)X_t,$$
and
$$\phi_q(B)\omega_t = \left(1 + b_1 B + b_2 B^2 + \cdots + b_q B^q\right)\omega_t$$

where p and q refer to the order of the autoregressive (AR) and moving average (MA) model respectively. Meanwhile, the degree of differencing is denoted as d.

A suitable ARIMA model for each of the station is identified and fitted to the training data. Then, a set of predicted values is calculated to be compared to the test data of PM$_{10}$. The comparison is done by finding the RMSE value between the predicted and actual values of PM$_{10}$ as in equation (6).

Results and discussion

The analysis of this study comprises of two parts. The first part is the comparison between the common univariate Autoregressive Integrated Moving Average (ARIMA) model to the univariate LSTM model for predicting PM$_{10}$. The second part is the comparison between the univariate LSTM model with the multivariate LSTM model for PM$_{10}$ prediction. Both comparisons are done on all five stations under consideration.

Comparison between univariate ARIMA and univariate LSTM model

In order to find out the potential use of LSTM model in fitting and forecasting PM$_{10}$, the univariate LSTM models at all five stations are compared to the well-known univariate ARIMA models. A one-time difference was required for all the five stations to achieve stationary for the univariate ARIMA model. In this study, it is found that the ARIMA (1, 1, 2), ARIMA (1, 1, 5), ARIMA (1, 1, 5), ARIMA (1, 1, 2) and ARIMA (5, 1, 1) model are suitable to fit the Kulim Hi-Tech station, Kuala Selangor station, Segamat station, Kuala Terengganu station and Tanah Merah station respectively. The RMSE calculated for the differences between the predicted values from the univariate ARIMA model at each station and the actual values of PM$_{10}$ for the test data was calculated. The value of RMSE for each station was 23.4223, 30.9408, 35.7524, 27.6501 and 23.9955, respectively.

The univariate LSTM model for each station was built. For each model, values for mean squared error (MSE) and mean absolute error (MAE) of the training data set after 100 epochs are shown in Table 1. All the values of MSE and MAE are small which are less than 0.05 at all five stations. Hence, these univariate LSTM models are deemed suitable to fit PM$_{10}$ values at the five stations.

Next, RMSE between actual values for variable PM$_{10}$ in test data and forecast values found from the univariate LSTM model was calculated. RMSE for Kulim Hi-Tech station, Kuala Selangor station, Segamat station, Kuala Terengganu station and Tanah Merah station are 7.049, 7.833, 8.306, 7.398 and 9.753, respectively. The RMSE values for both the univariate ARIMA and univariate LSTM models are shown in Table 2.

Table 1. Mean square error (MSE) and mean absolute error (MAE) of the univariate LSTM model for training data set.

Station	MSE	MAE
Kulim Hi-Tech, Kedah	0.0019	0.0300
Kuala Selangor, Selangor	0.0006	0.0179
Segamat, Johor	0.0015	0.0245
Kuala Terengganu, Terengganu	0.0002	0.0066
Tanah Merah, Kelantan	0.0020	0.0309

Table 2. Root mean square error (RMSE) for univariate ARIMA and univariate LSTM model.

Station	Univariate ARIMA model	Univariate LSTM model
Kulim Hi-Tech, Kedah	23.4223	7.049
Kuala Selangor, Selangor	30.9407	7.833
Segamat, Johor	35.7524	8.306
Kuala Terengganu, Terengganu	27.6501	7.398
Tanah Merah, Kelantan	23.9955	9.753

A comparison between the univariate LSTM model and the univariate ARIMA model for the five stations was made. It was found that RMSE values for the univariate LSTM models for all five stations were lower than the corresponding univariate ARIMA models. Therefore, it could be concluded that the univariate LSTM model can predict PM_{10} better than the univariate ARIMA model.

Comparison between univariate LSTM and multivariate LSTM model

The second part of this analysis is to look at whether the predicted values of PM_{10} based on the LSTM model could be improved if we considered meteorological factors such as wind direction, wind speed, ambient temperature, and relative humidity. Hence, the multivariate LSTM model was built for each station under study. Table 3 shows the MSE and MAE values for the multivariate LSTM models after training 100 epochs on each station's train data set. Similar to the univariate LSTM model, the MSE and MAE values obtained are small with all values less than 0.05 at each station. In fact, all values are almost the same if not smaller than the ones found from their univariate LSTM counterpart. Hence, these multivariate LSTM models are also deemed suitable models to fit PM_{10} values at the five stations.

Prediction on variable PM_{10} was performed for all five stations based on the multivariate LSTM model. RMSE was calculated between the actual value from the test data and the forecast value from the model for each station. RMSE for Kulim Hi-Tech station, Kuala Selangor station, Segamat station, Kuala Terengganu station and Tanah Merah station are 7.040, 7.491, 8.180, 7.041 and 9.561, respectively. These values are shown in Table 4.

Table 3. Mean square error (MSE) and mean absolute error (MAE) of the multivariate LSTM model for training data set.

Station	MSE	MAE
Kulim Hi-Tech, Kedah	0.0019	0.0297
Kuala Selangor, Selangor	0.0006	0.0177
Segamat, Johor	0.0015	0.0244
Kuala Terengganu, Terengganu	0.0002	0.0066
Tanah Merah, Kelantan	0.0017	0.0292

Table 4. Root mean square error (RMSE) for multivariate and univariate LSTM model.

Station	Multivariate LSTM model	Univariate LSTM model
Kulim Hi-Tech, Kedah	7.040	7.049
Kuala Selangor, Selangor	7.491	7.833
Segamat, Johor	8.180	8.306
Kuala Terengganu, Terengganu	7.041	7.398
Tanah Merah, Kelantan	9.561	9.753
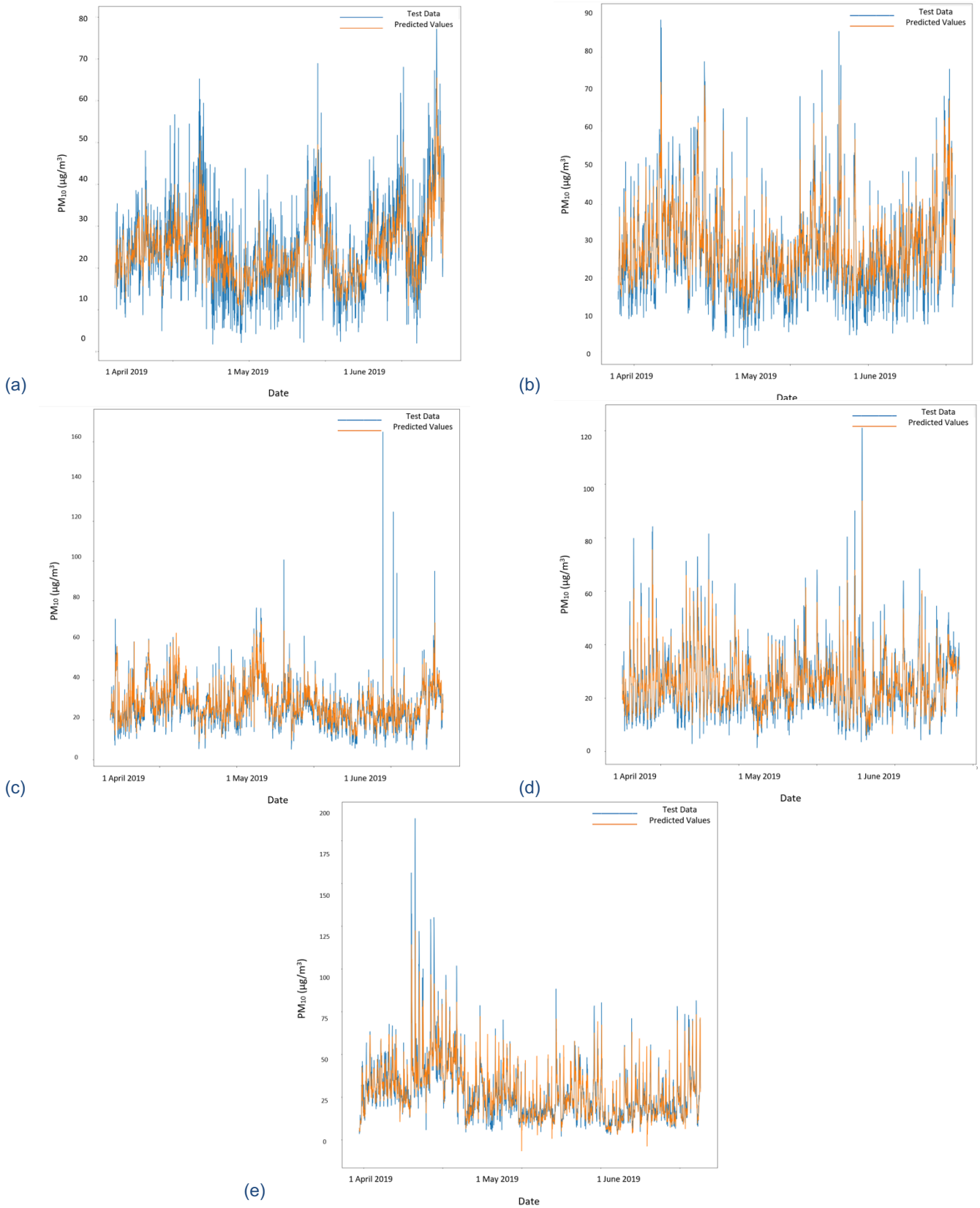
Figure 1. Plots for test data and predicted values based on multivariate LSTM model. (a) Kulim Hi-Tech, Kedah. (b) Kuala Selangor, Selangor. (c) Segamat, Johor. (d) Kuala Terengganu, Terengganu. (e) Tanah Merah, Kelantan.

The RMSE values for the multivariate LSTM models of the five stations were slightly lower compared to the univarite LSTM models. Therefore, it could be concluded that the multivariate LSTM model can predict variable PM_{10} better than the univariate LSTM model. Hence, the additional meteorological variables as input to the LSTM model do improve the prediction for the values of PM_{10}.

Figure 1 shows the plots for time series from the test data set as well as the time series predicted based on the multivariate LSTM model at all five stations. The two time series plots at each station overlapped with each other majority of the times for the duration of the test data. Furthermore, it can be seen that the predicted values from the multivariate LSTM models are able to capture the patterns of the actual time series as given by the test data. Hence, the multivariate LSTM models are able to forecast values close to the actual values of the test data.

Conclusions

In this study, the potential use of LSTM models in predicting PM_{10} values in Peninsular Malaysia are explored. Both the univariate and multivariate LSTM models were considered where the multivariate LSTM model attempts to improve the univariate LSTM model by adding meteorological variables as input into the network. Both models show suitability in fitting PM_{10} values with small MSE and MAE values at all five stations.

The univariate LSTM model was compared to the univariate ARIMA model at each of the station. The univariate LSTM model was deemed as a better model compared to the univariate ARIMA model at all five stations with smaller RMSE values. This might be due to limitations of ARIMA model which transformed factors into some time variables and hence only analyze the linear part of the time series as stated by Zhang et al. [22] where in their study, LSTM models also perform better than ARIMA models in predicting the incidence of hand, foot, and mouth disease (HFMD) due to the nature and design of LSTM model which learns the temporal patterns, captures non-linear dependencies and stores useful memory for a longer duration. Furthermore, the result of this study is in accordance with those obtained by a recent previous study [23] which concludes that LSTM models efficiently deals with complexities of modelling and forecasting air quality in Indian National Capital Territory of Delhi.

Next, the univariate LSTM model at each station was compared to their corresponding multivariate LSTM model. Based on the RMSE values, the multivariate LSTM models which also considers ambient temperature, wind speed, wind direction and relative humidity proved to be better models at all five stations in predicting PM_{10} values in Peninsular Malaysia. This might be due to the fact that the multivariate time series models do not limit themselves to the past information of the predicted variable but also integrate the past values of other variables of influence.

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

Acknowledgments

References

[1]   Gurjar, B.R., Butler, T.M., Lawrence, M.G. and Lelieveld, J. Evaluation of emissions and air quality in megacities. Atmospheric Environment. 2008. 42(7): 1593-1606.
[2]   Zhang, Y., Bouquet, M., Mallet, V., Seigneur, C. and Baklanov, A. Real-time air quality forecasting, part I: history, techniques and current status. Atmospheric Environment. 2012. 60: 632-655.
[3]   Zhang, Y., Bouquet, M., Mallet, V., Seigneur, C. and Baklanov, A. Real-time air quality forecasting, part II: state of the science, current research needs and future prospects. Atmospheric Environment. 2012. 60: 656-676.
[4]   Vardoulakis, S., Fisher, B.E.A., Pericleous, K. and Flesca, N.G. Modelling air quality in street canyons: a review. Atmospheric Environment. 2003. 37(2): 155-182.
[5]   Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S. and Kenski, D. $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. Expert Systems with Applications. 2009. 36(5): 9046-9055
[6]   Donnelly, A., Misstear, B. and Broderick, B. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmospheric Environment. 2015. 103: 53-65.
[7]   Du, S.D., Li, T.R., Yang, Y. and Horng, S.J. Deep air quality forecasting using hybrid deep learning framework. Cornell University: PhD Thesis. 2019.
[8]   Wang, J., Niu, T. & Wang, R. Research and application of an air quality early warning system based on a modified least squares support vector machine and a cloud model. International Journal Environmental Research and Public Health. 2017. 14(3): 249.
[9]   Grivas, G. and Chaloulakou, A. Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment. 2006. 40(7): 1216-1229.
[10]   Zeyhelgil H.L., Demiroren A. and Sengor N.S. The application of ANN technique to automatic generation control for multi-area power system. International Journal of Electrical Power & Energy Systems. 2002. 24(5): 345-354.
[11]   Colah. Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [2020]. 2015.
[12]   Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation. 1997. 9(8): 1735-1780.
[13]   Salman, A.G., Heryadi, Y., Abdurahman, E. and Suparta, W. Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. Procedia Computer Science. 2018. 135: 89-98.
[14]   Sak, H., Senior, A. and Beaufays, F. Long Short-term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modelling. Google, USA. 2014.
[15]   Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 2018. 22: 6005-6022.
[16]   Department of Environment Malaysia. Haze Facts. https://www.doe.gov.my/portalv1/wp-content/uploads/2016/08/Haze-Facts.pdf [2021]. 2016.
[17]   How, C.Y., and Ling, Y.E. The influence of PM2.5 and PM10 on Air Pollution Index (API). Environmental Engineering, Hydraulics and Hydrology: Proceeding of Civil Engineering, Universiti Teknologi Malaysia, Johor, Malaysia. 2016. 3: 132.
[18]   Mutalib, S.N.S.A., Juahir, H., Azid, A., Sharif, S.M., Latif, M. T., Aris, A.Z., Zain, S.M. and Dominick, D. Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. Environmental Science Processes & Impacts. 2013. 15: 1717-1728.
[19]   Goodfellow, I., Bengio, Y. and Courville, A. Deep Learning. Massachusetts: The MIT Press. 2016.
[20]   Ariff, N. M., Zamhawari, N.H. and Bakar, M.A.A. Time series ARIMA models for daily price of palm oil. AIP Conference Proceedings. 2015. 1643: 281-288.
[21]   Singh, S., Parmar, K.S., Kumar, J. and Makkhan, S.J.S. Develoment of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. Chaos, Solitions and Fractals. 2020. 109866.
[22]   Zhang, R., Guo, Z., Meng, Y., Wang, S., Li, S., Niu, R., Wang, Y., Guo, Q. and Li, Y. Comparison of ARIMA and LSTM in forecasting the indices of HFMD combined and uncombined with exogenous meteorological variables in Ningbo, China. International Journal of Environmental Research and Public Health. 2021. 18: 6174.
[23]   Krishan, M., Jha, S., Das, J., Singh, A., Goyal, M.K. and Sekar, C. Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. Air Quality, Atmosphere & Health. 2019. 12: 899-908.