Predicting Particulate Matter (PM$_{2.5}$) in Malaysia using Multiple Linear Regression and Artificial Neural Network

Norafefah Mohamad Sobri$^1$, Wan Fairus Wan Yaacob$^2$, Nor Azima Ismail$^1$, Mohd Azry Abdul Malik$^1$, Raudhah Ab. Rahman$^2$, Nur Aqilah Basir$^2$, Suwaibah Afiqah Mohd Sukhairi$^2$

$^1$Faculty of Science and Mathematics, University Teknologi Mara Cawangan Kelantan Campus Machang, Bukit Ilmu, 18500 Machang, Kelantan

$^2$Faculty of Science and Mathematics, Universiti Teknologi Mara Cawangan Kelantan Campus Kota Bharu, Lembah Sireh, 15050 Kota Bharu, Kelantan

*noraf378@uitm.edu.my

Abstract. Air pollution is a well-known issue for all countries, including Malaysia. It has been stated that particulate matter that less than 2.5mm known as PM$_{2.5}$ has a greater effect on health as the smaller particulate size can penetrate deep into the respiratory system and affect the cardiovascular system significantly. Therefore, it is necessary to estimate the concentration of PM$_{2.5}$ for haze precautions. This study characterizes the pattern of PM$_{2.5}$ concentrations involving seven stations including Alor Setar, Shah Alam, Pasir Gudang, Ipoh, Kuantan, Kuala Terengganu and Miri with seven indicator parameters (Carbon Monoxide, Ozone, Sulphur Dioxide, Nitrogen Dioxide, Humidity, Temperature and Wind Speed). PM$_{2.5}$ concentrations were predicted for each station using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). Descriptive and trend analysis using Mann-Kandell Trend analysis was used to describe the haze characteristics and identify significant trends in the haze selected locations in Malaysia. MLR and ANN were fitted for the data. The performance of both prediction models was compared based on $R^2$ and Mean Square Error (MSE). The results show ANN performed better than MLR with a high value of coefficient determination ($R^2$) and low error measure. The ANN model was used to predict the occurrence of haze for the next day in the Air Quality Index (API).

Keywords: Artificial Neural Network, Multiple Linear Regression, particulate size (PM$_{2.5}$)

1. Introduction

Haze is characterized by the presence of small particles (0.1-1.0 in diameter) spread across a part of the atmosphere, which reduces the level of clarity and gives the atmosphere a characteristic cloudy appearance. Haze occurred annually in Malaysia from January to February, and from June to August [1]. The incidence of haze severely affected the daily livelihoods of the people in and around the countries [1]. Haze can affect human health by striking the lung, heart, and circulatory systems and create gaze [2]. The Air pollution index (API) readings also went far beyond the hazardous range in some areas in Malaysia [1]. The Malaysian Air Pollution Index Standard is classified into five levels according to the Department of the Environment (DOE) as in table 1.
Table 1. Category of Air Pollution Index (API)

| Air Pollution Index (API) (unit) | Air Pollution Level |
|---------------------------------|---------------------|
| 0 – 50                           | Good                |
| 51 – 100                         | Moderate            |
| 101 – 200                        | Unhealthy           |
| 201 – 300                        | Very unhealthy      |
| 301 +                            | Hazardous           |

The API Index is classified from the average concentration of air pollutants namely air pollutants particulate matter of fewer than 10 micrometers in diameter (PM$_{10}$) or less than 2.5 micrometers in diameter (PM$_{2.5}$), Ozone (O$_3$), Carbon Monoxide (CO), Sulphur Dioxide (SO$_2$) and Nitrogen Dioxide (NO$_2$) over a given time [3, 4]. In Peninsular Malaysia, Particulate matter is the most prevailing pollutant having the highest API value compared to the other criteria pollutant [5]. Prediction of particulate matter is one of the major environmental contaminants. The particulate matter is a combination of solid particles and liquid droplets that can be found in the air. PM$_{10}$ is a particulate matter with a diameter of 10 microns or less, while PM$_{2.5}$ is a particulate matter with a diameter of 2.5 microns or less. Small particulate matter is generally referred to as PM$_{2.5}$ [6]. Particles in this range of sizes constitute a large proportion of particles that can draw deep into the lungs. Larger particles seem to have been stuck in the ear, mouth, or nose. PM$_{2.5}$ has a greater effect on health as the smaller particulate size can penetrate deep into the respiratory system and affect the cardiovascular system significantly. About 91% of the world’s population live in an area with PM$_{2.5}$ concentrations above the allowable limits of 10-20 μgm$^{-3}$ and it is estimated that about 7 million people die every year worldwide because of exposure to fine particulate < 2.5 μm (PM$_{2.5}$) [7].

Due to the frequent incidence of haze events in Malaysia every year over the past few, this issue has gained greater attention due to its adverse impact on the environment, air quality, and human health. Implementation of an effective model for predicting API is crucial. There were several models used to analyzed PM$_{10}$ and PM$_{2.5}$ such as time series regression [8], multiple linear regression (MLR) [5], artificial neural network (ANN), autoregressive integrated moving average model (ARIMA), and exponential smoothing method (ESM). However, a few studies used Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) to predict PM$_{10}$ in Malaysia [9 - 12]. PM$_{2.5}$ studies are very few in Malaysia. The first study in Malaysia for PM$_{2.5}$ is conducted by Zaman et. all [7] combining satellite aerosol retrievals with ground-based pollutants, metrological factors and Machine Learning techniques. ANN is an alternative to traditional statistical prediction methods such as linear regression [13]. The algorithm developed is suitable to model complex patterns and problem prediction. ANN has been proven to be reliable and performed well with an accurate prediction model [14 - 20]. Thus, this research aimed to compare the MLR model and ANN model in predicting particulate matter (PM$_{2.5}$) for seven stations including Alor Setar, Shah Alam, Pasir Gudang, Ipoh, Kuantan, Kuala Terengganu and Miri.

2. Materials and Methods

2.1 Study Location

This study used the daily data of the concentrations of PM$_{2.5}$ in Malaysia involving seven stations which are Alor Setar, Shah Alam, Pasir Gudang, Ipoh, Kuantan, Kuala Terengganu and Miri from January 2018 to December 2019. Pasir Gudang is categorized as an industrial area with a high concentration of PM$_{10}$ [21]. Klang and Shah Alam are categorized in the urban area and showed a high concentration of PM$_{2.5}$ [22]. Ipoh is located in the city center, sub-cities and sub-urban areas, and is also known as the rapidly growing city of development [23]. Alor Setar and Kuantan are located within an urban-industrial environment. Meanwhile, Miri is a coastal city located on the island of Borneo and near the border of
Brunei. The air quality status over seven locations is varied. The different area has a different factor variable due to its characteristics.

2.2 Data Description
The daily data on air pollutants and meteorological factors as the input variables from 2018 to 2019 were collected from the Department of Environment (DOE). The data were recorded on a daily basis and each data have 5,040 observations that represent the value of daily concentrations value for a day. The details of the data collected were display in table 2.

| Type                  | Variable                                      | Unit                |
|-----------------------|-----------------------------------------------|---------------------|
| Dependent Variable, Y | Prediction Particulate Matter (PM$_{2.5,t+1}$) | Micrograms/cubic meter(μgm$^3$) |
| Independent variables, X | Carbon Monoxide (CO) | Parts per million (ppm) |
|                       | Ozone(O$_3$)                                  | Parts per million(ppm) |
|                       | Sulphur Dioxide (SO$_2$)                     | Parts per million(ppm) |
|                       | Nitrogen Dioxide (NO$_2$)                    | Parts per million(ppm) |
|                       | Humidity (RH)                                | Vapour/kilogram (kg$^{-1}$) |
|                       | Temperature (T)                              | Celsius (°C)        |
|                       | Wind Speed (WS)                              | Meter per second(m/s) |
|                       | Particulate Matter (PM$_{2.5}$)              | Micrograms/cubic meter(μgm$^3$) |

2.3 Imputation Missing Value
Missing value in dataset might be due to the changing of sitting monitor, instruments problem, maintenance and weather. In the data cleaning process, the imputation process was done to manage any missing values and filter the process for the outlier. Linear interpolation technique was applied in this study to overcome the missing data [24]. In the linear interpolation imputation process, the series means technique was applied. The formulation is shown below:

$$f(x) = f(x_i) + \frac{f(x_i) - f(x_0)}{x_i - x_0} \times (x - x_0)$$  \hspace{1cm} (1)

where $x$ is an independent value, $x_i$ and $x_0$ is known values of the independent variable and $f(x)$ is value of dependent variable for a value of the independent variable.

2.4 Preliminary Analysis
The mean analysis was used to analyze the haze characteristics and identify the highest and lowest value of the concentration of PM$_{2.5}$ for each station. Graphical analysis was used to illustrate the trend, and classify it according to year. The time series plots were conducted to monitor records in the daily concentration series PM$_{2.5}$. The Man Kendall test had been used to test the trend lines significantly. The formula used for Mann-Kendall Test is given as follows:

$$\text{Tau} = RC = \frac{1}{2\pi f_c}$$  \hspace{1cm} (2)

where, $R$= resistance, $C$= capacitance, $f_c$ = cut off frequency.
2.5 **Multiple Linear Regression**

MLR is used to explain a constant relationship between dependent and two or more independent variables. It assumes the residuals are normally distributed and the variance is constant. The theoretical equation of MLR can be expressed as follows [22]:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i \quad \text{with} \quad i = 1, 2, \ldots, n \]  

(3)

where, \( y_i \) and \( x_j \) is dependent and independent variable, \( \beta_0 \) is the value \( y \)-intercept, \( \beta_k \) is the regression coefficients and \( \epsilon_i \) is the value of error. \( x_1 \) is Carbon Monoxide (CO), \( x_2 \) is Nitrogen Dioxide (NO2), \( x_3 \) is Sulphur Dioxide (SO2), \( x_4 \) is Ozone (O3), \( x_5 \) is Humidity (H), \( x_6 \) is Temperature (T) and \( x_7 \) is Wind Speed (WS).

Model adequacy checking was conducted to assess the assumptions of MLR. These include checking on normality assumption of residuals, homoscedasticity where the variation of error terms should be identical across independent variable values and multicollinearity. The plot of standardized residuals vs the predicted value should indicate how many points are spread uniformly over all independent variable values. Multicollinearity inspection should suggest that independent variables are not strongly correlated by looking at the inflation factor variance (VIF) along with the regression output, where the VIF under 10 is assumed to be perfect, where there is no multicollinearity between the independent variable.

2.6 **Artificial Neural Network**

ANN is a class of feed-forward neural network system where a group of simple units called neurons that calculate their input and perform patterns and non-linear classification. The structure of a typical neural network consists of an input layer where data enters the networks, a hidden layer and an output layer that combines the artificial neuron results. The hidden layer includes the non-linear activation functions, mostly preferred by hyperbolic tangents because it covers both positive and negative values. The thicker the layer, the greater the network’s ability to identify trends. The scaled input vector, which introduces neurons to the input layer is multiplied by a weight, which is a real number quantity. The neuron in the hidden layer sums up this information including bias [10,24].

\[ f(x) = \frac{1}{1 + e^{-x}} \]  

(4)

then,

\[ y_0 = f \left[ \sum_{i=1}^{n} w_i x_i + b \right] \]  

(5)

where; \( y_0 \) is the calculation output, \( i \) \( w \) is weight of vector, the scaled input the vector is \( x_i \), the bias is \( b \), the transfer function is \( f \), and \( x \) are the total sum of weighted inputs.

Weighted sum information remains in linear model terms. Information of model non-linearity occurs because it is passing through the transfer function or the activation function. The activation function is a mathematical representation of the input-output relationship. The activation function introduces non-linearity in the ANN model and then compares it to the linear model.

The structure of the artificial neural network with one hidden layer used in this study is displayed in figure 1.
2.7 Model Evaluation

Model accuracies for MLR and ANN were assessed using the coefficient of determination ($R^2$). In addition, the errors were evaluated based on the sum of square error (SSE) and mean square error (MSE).

\[
R^2 = \left( \frac{\sum (P_i - \bar{P})(O_i - \bar{O})}{n \cdot S_{pred \cdot obs}} \right)^2
\]

(6)

\[
SSE = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (k - 1)}}
\]

(7)

\[
MSE = \frac{SSE}{df_{mse}}
\]

(8)

where, the number of independent variables is $k$, $n$ represent the sample size, $P_i$ is the forecasted, $O_i$ is the observed values, $\bar{P}$ is the mean of forecasted values, $\bar{O}$ is the mean of observed values, $y$ is the observe values and the predict values is $\hat{y}$ and $df_{mse}$ is the degree freedom of mean square error.

Data were partitioned into two parts: 60% for training and 40% for testing for both MLR and ANN models. The predictive accuracy for both models was measured using the coefficient of determination ($R^2$), while the errors were calculated using the sum of square error (SSE) and mean square error measure as in Equations 6, 7 and 8.

3. Result and Discussion

3.1 Haze Characteristics

Haze has become a serious environmental and human health concern across the country especially in urban, industrialized and congested traffic areas such as Klang Valley, Kuantan and Miri. Table 3
analyses and tabulates the PM$_{2.5}$ data using the mean approach. The highest and lowest value of the concentration of PM$_{2.5}$ is determined for each station each year. For hazy days, PM$_{2.5}$ is equivalent to or greater than 101μg/m$^3$ per year. This fits the limit set in Recommended Malaysia Air Quality (RMCQA) by the Department of the Environment (DOE).

### Table 3. Haze characteristics in 2018 and 2019

| Station       | Year | PM$_{2.5}$ concentration (μg/m$^3$) | Hazy Days |
|---------------|------|-----------------------------------|-----------|
|               |      | Mean | Highest (status) | Lowest |         |
| Ipoh          | 2018 | 20.16 | 73.77(Moderate) | 7.64  | 0       |
|               | 2019 | 22.32 | 107.59(Unhealthy) | 8.17  | 1       |
| Shah Alam     | 2018 | 25.20 | 62.64(Moderate) | 8.15  | 0       |
|               | 2019 | 32.10 | 163.06(Unhealthy) | 10.07 | 9       |
| Pasir Gudang  | 2018 | 18.61 | 39.92(Good) | 4.68  | 0       |
|               | 2019 | 21.07 | 83.96(Moderate) | 5.38  | 0       |
| Alor Setar    | 2018 | 14.66 | 84.18(Moderate) | 3.34  | 0       |
|               | 2019 | 16.11 | 73.50(Moderate) | 4.08  | 0       |
| Kuantan       | 2018 | 14.43 | 51.35(Moderate) | 2.90  | 0       |
|               | 2019 | 18.21 | 114.99(Unhealthy) | 4.04  | 3       |
| Kuala Terengganu | 2018 | 17.34 | 53.17(Moderate) | 3.01  | 0       |
|               | 2019 | 19.74 | 99.94(Moderate) | 3.70  | 0       |
| Miri          | 2018 | 14.39 | 48.55(Good) | 3.57  | 0       |
|               | 2019 | 21.41 | 126.45(Unhealthy) | 4.97  | 3       |

Shah Alam has the largest cumulative number of hazy days in the year 2019, which is 9 days. It also had the highest total concentration of PM$_{2.5}$ relative to other concentrations. Along with the level of air quality, most of the stations reached unhealthy levels in 2019. This is because, in 2019, transboundary haze from fires, especially from sources in Sumatra and Kalimantan, entered Singapore and Malaysia and dramatically deteriorated the measured air quality (Greenpeace Southeast Asia, 2019). As a result, Ipoh, Shah Alam, Kuantan and Miri have reached an unhealthful amount where their PM$_{2.5}$ levels are more and equivalent to 101μg/m$^3$. Generally, if air pollution reaches an unhealthy level the risk of a health condition for elderly people, pregnant women, children getting worse is high, including those with heart and lung problems.

### 3.2 Haze Trend Analysis

Table 4 shows the analysis from the Mann-Kendall Trend test that has resulted in the trend of PM$_{2.5}$ in several Malaysia from 2018 to 2019. The location with a significant trend was Shah Alam, Pasir Gudang and Miri as the p-value computed are lower than significant level alpha 0.05, which were 0.009, <0.0001, <0.0001 respectively. Ipoh, Alor Setar, Kuantan and Kuala Terengganu, have resulted in the p-value was above the significant level alpha, so the null hypothesis was accepted indicates no trend in those stations. Therefore, there is enough evidence to conclude that the trend exists on Shah Alam, Pasir Gudang and Miri stations and also that there is a significant increasing flow trend.
Table 4. Concentration trend of PM<sub>2.5</sub> on daily data

| Station          | P-value | Tau’s Value | Trend       |
|------------------|---------|-------------|-------------|
| Ipoh             | 0.112   | -0.039      | Decreasing  |
| Shah Alam        | <0.0001 | 0.065       | Increasing  |
| Pasir Gudang     | 0.188   | -0.033      | Decreasing  |
| Alor Setar       | 0.246   | 0.029       | Decreasing  |
| Kuantan          | 0.202   | -0.032      | Decreasing  |
| Kuala Terengganu | <0.0001 | 0.187       | Increasing  |

3.3 Model Comparison of MLR and ANN

R-square (R<sup>2</sup>), Sum of square error (SSE), and mean square error (MSE) values of the Multiple Linear Regression Model and the Artificial Neural Network Model were evaluated for each station to choose the best model for predicting the future value of the PM<sub>2.5</sub> concentrations. The higher R<sup>2</sup> value implies that PM<sub>2.5</sub> is very well explained by the input variable in the model being developed. The lower the value of sum square error (SSE) and the mean square error (MSE) would lead to a better model for prediction.

Model adequacy checking has been conducted for the MLR model. All assumptions pertaining to residuals for all models that represent data from seven stations were found to be satisfied. Table 5 summarizes the MLR model performance. The coefficient of determination (R<sup>2</sup>) measures the percentage of change in PM<sub>2.5</sub> concentration that can be explained by air pollutants and meteorological factors in the regression model. Comparing these 7 stations, Kuantan has the highest value of R<sup>2</sup> which is 0.775. 77.5% of the variation in the PM<sub>2.5</sub> in Kuantan can be explained by the independent variables in MLR. Followed by Shah Alam and Pasir Gudang with 72.8% and 69.4% of the variation in PM<sub>2.5</sub> were explained by the independent variables in MLR. About 72.1% of the variation in the PM<sub>2.5</sub> can be explained by the independent variables for Miri. For Kuala Terengganu, about 65.1% of the variation in the PM<sub>2.5</sub> can be explained by the independent variables. The R<sup>2</sup> value for Ipoh is 0.649. Thus about 64.9% of the variation in PM<sub>2.5</sub> can be explained by the independent variables. Alor Setar, has the lowest value in R<sup>2</sup> compared to the other 7 stations. About 59.6% of the variation in the PM<sub>2.5</sub> was explained by the independent variables for Alor Setar.

Table 5 also shows the summary result obtained from the ANN model. The R<sup>2</sup> values are 0.691 for Alor Setar station, 0.733 for Shah Alam station, 0.770 for Pasir Gudang Station, 0.795 for Ipoh station, 0.867 for Kuantan station, 0.756 for Kuala Terengganu station and lastly 0.749 for Miri station. Other than that, the sum of square error (SSE) for the stations are 91.403, 16.831, 39.815, 62.121, 10.496, 48.238 and 27.590 for Alor Setar, Shah Alam, Pasir Gudang, Ipoh, Kuantan, Kuala Terengganu and Miri respectively while for their mean square error (MSE) are 26.142 for Alor Setar station, 35.462 for Shah Alam station, 17.999 for Pasir Gudang station, 20.088 for Ipoh station, 12.929 for Kuantan station, 21.683 for Kuala Terengganu station and lastly 21.5825 for Miri station. As we have all the performance indicator values such as R-square, Sum Square Error (SSE) and Mean Square Error (MSE), a comparison was made between Artificial Neural Network (ANN) and Multiple Linear Regression (MLR).

Comparison results in table 5 prove that Artificial Neural Network is better in the prediction of PM<sub>2.5</sub> concentrations than Multiple Linear Regression with the higher value of R<sup>2</sup> and small value of SSE and MSE for all locations. The Artificial Neural Network (ANN) outperformed the MLR model for all locations in estimating the potential value of the PM<sub>2.5</sub> concentrations. This research was able to show
that the ANN model performs better and is able to reduce the deviation of the model and increase the precision of the PM$_{2.5}$ model forecast.

Table 5. Model performance of MLR and ANN

| Station     | Method | $R^2$ | Sum Square of Error | Mean Square Error |
|-------------|--------|-------|---------------------|-------------------|
| Ipoh        | MLR    | 0.649 | 28811.187           | 39.960            |
|             | ANN    | 0.795 | 62.121              | 20.088            |
| Shah Alam   | MLR    | 0.728 | 45151.429           | 62.623            |
|             | ANN    | 0.733 | 16.831              | 35.462            |
| Pasir Gudang| MLR    | 0.694 | 21031.228           | 29.170            |
|             | ANN    | 0.770 | 39.815              | 17.999            |
| Alor Setar  | MLR    | 0.596 | 28806.819           | 39.954            |
|             | ANN    | 0.691 | 27.119              | 23.092            |
| Kuantan     | MLR    | 0.775 | 26600.819           | 36.894            |
|             | ANN    | 0.867 | 10.496              | 12.929            |
| Kuala       | MLR    | 0.651 | 28723.193           | 39.838            |
| Terengganu  | ANN    | 0.756 | 48.238              | 21.683            |
| Miri        | MLR    | 0.721 | 44030.004           | 61.068            |
|             | ANN    | 0.749 | 27.590              | 21.583            |

3.4 Prediction of PM$_{2.5}$ Concentration for Several Locations in Malaysia
The ANN model was chosen as the best model in this analysis and was used to estimate the concentration of particulate matter (PM) at several locations in Malaysia. Microsoft Excel add-in Neuro-XL Predictor was used as a concentration forecasting method. Past data is used for the next day's prediction, which is 1 January 2020. The daily observed and predictions for the next day are shown in the following figures.
Table 6 indicates the predicted value of particulate matter (PM$_{2.5}$) for each station for the next day. Based on the Air Quality Index (API) category, the presence of haze happens when 101μg is surpassed. As a result, the predicted value indicates that there is no haze incidence at 1/1/2020.

| Stations       | Ipoh | Shah Alam | Pasir Gudang | Alor Setar | Kuantan | Kuala Terengganu | Miri |
|----------------|------|-----------|--------------|------------|---------|------------------|------|
| 1/1/2020       | 13.697 | 26.434    | 8.320        | 8.282      | 10.034  | 12.350           | 11.664 |

4. Conclusion
It has been shown that most stations have a high level of PM$_{2.5}$ concentration in 2019. The level of concentration indicates that they had fallen to a hazardous zone as their level was between an unhealthy and a moderate level indicated in the API. This problem occurred because the haze problem happens in 2019 and has affected several sites in Malaysia. It was caused by the Kalimantan fire sources that entered Malaysia and significantly worsened the measured air quality, particularly in Miri. Furthermore, haze has frequently occurred in Shah Alam recorded for 9 days’ period. It is due to haze that occurred in Sumatra, Indonesia which later affected Malaysia. The trend analysis from Mann-Kendall test findings indicates the significant trend of haze in Shah Alam, Pasir Gudang and Miri with a p-value lower than the significant level. It also describes three stations that have an increasing flow pattern. Meanwhile, the results for Ipoh, Alor Setar, Kuantan and Kuala Terengganu show no indication of trend haze occurrence. For model prediction, ANN has proven to outperform the MLR model to forecast air quality. It indicates that the ANN model produces minimum error and achieves a better estimation accuracy in prediction. Lastly, the ANN prediction shows that there were no haze occurrences on the next day. As a result, the ANN model has shown that it could be completely implemented in environmental studies, especially in a haze situation.

Acknowledgment
The authors gratefully acknowledge the use of the facilities of Universiti Teknologi MARA Cawangan Kelantan and the Department of Environment (DOE) for providing the data. The authors also would like
to take the opportunity to thank the staff of Academic Affair and all students at UiTM Kota Bharu, Kelantan.

References

[1] Hanafi N, Hassim M and Noor Z 2018 Overview of Health Impacts due to Haze Pollution in Johor, Malaysia Journal Of Engineering and Technological Sciences 50 818
[2] Chooi, Y. &. (n.d.). The Influence of PM$_{2.5}$ and PM$_{10}$ on Air Pollution Index (API).
[3] Abd A K 2018 Air Pollution Index Trend Analysis in Malaysia, 2010-15 Pol.J.Environ.Stud. 27 801-7
[4] Zhong S, Zhang L, Jiang X and Gao P 2019 Comparison of chemical composition and airborne bacterial community structure in PM$_{2.5}$ during haze and non-haze days in the winter in Guilin, China. Science of the Total Environment 655 202–10
[5] Abdullah S, Ismail M and Fong S Y 2017 Multiple linear regression (MLR) models for long term PM10 concentration forecasting during different monsoon seasons Journal of Sustainability Science and Management 12 60-69
[6] US EPA, OAR. [Internet]. 2018 [updated 2018 November 14]. Particulate Matter (PM) Basics | US EPA. US EPA. Available from : https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.
[7] Zaman N A F K, Kanniah K D, Kaskaoutis D G and Latif M T 2021 Evaluation of Machine Learning Models for Estimating PM2. 5 Concentrations across Malaysia Applied Sciences 11 7326
[8] Sharma N, Taneja S, Sagar V and Bhatt A 2019 Forecasting air pollution load in Delhi using data analysis tools Procedia Computer Science 132 1077–85
[9] Shahriatini H 2016 Statistical Modeling Approaches for PM10 Prediction in Urban Areas
[10] Ku Yusof K, Azid A, Abdullah Sani M, Samsudin M, Muhammad Amin S, Abd Rani N and Jamalani M 2019 The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models over particulate matter (PM10) variability during haze and non-haze episodes: A decade case study Malaysian Journal of Fundamental and Applied Sciences 15 164-72
[11] Abdullah S, Napi N N L M, Ahmed A N, Mansor W N W, Mansor A A, Ismail M, ... and Ramly Z T A 2020 Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia Atmosphere 11 289
[12] Abdullah S, Ismail M, Ahmed A and Abdullah A 2019 Forecasting Particulate Matter Concentration Using Linear and Non-Linear Approaches for Air Quality Decision Support. Atmosphere 10 667
[13] Mishra D, Goyal P and Upadhyay A 2015 Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India Atmospheric Environment 102 239–48
[14] Assadollahfardi G Z 2016 Predicting PM$_{2.5}$ Concentration using Artificial Neural Networks and Markov Chain, a Case Study Karaj City Asian Journal of Atmospheric Environment 67-79
[15] Liu D J and Li L 2015 Application study of comprehensive forecasting model based on entropy weighting method on trend of PM$_{2.5}$ concentrations in Guangzhou, China International journal of environmental research and public health 12 7085-99
[16] Alimissis A, Philippopoulos K, Tzannis C G and Deligiorgi D 2018 Spatial estimation of urban air pollution with the use of artificial neural network models Atmospheric Environment 191 205–13
[17] Elangasinghe M A, Singhal N, Dirks K N and Salmond J A 2014 Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmospheric Pollution Research 5 696–708
[18] Feng X, Li Q, Zhu Y, Hou J, Jin L and Wang J 2015 Artificial neural networks forecasting of PM$_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment* 107 118–28

[19] Mckendry I 2011 Evaluation of Artificial Neural Network for Fine Particulate Pollution (PM10 and PM2.5) Forecasting *Journal of the Air & Waste Management Association* 1096-101

[20] Lv B, Cobourn W G and Bai Y 2016 Development of nonlinear empirical models to forecast daily PM2.5 and ozone levels in three large Chinese cities *Atmospheric Environment* 147 209–23

[21] Ahmat H, Yahaya A S and Ramli N A 2015 The Malaysia PM10 Analysis Using Extreme Value *Journal of Engineering Science and Technology* 10 1560-74

[22] Department of Environment [Internet]. 2020 [updated 2020 April 25]. Available from : https://www.doe.gov.my/portalv1/en/info-umum/kualiti-udara/114.

[23] Ismail P Report of the Forum on the Impact of Haze on Human Health in Malaysia [Internet]. 2017 [updated 2017 April 17]. Available from : https://www.akademisains.gov.my/report-of-the-forum-on-the-impact-of-haze-on-human-health-in-malaysia/.

[24] Kovač-Andrić E, Brana J and Gvozdić V 2009 Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods *Ecological Informatics* 4 117-22