

Neural Network Fitting using Levenberg-Marquardt Training Algorithm for PM₁₀ Concentration Forecasting in Kuala Terengganu

Samsuri Abdullah¹, Marzuki Ismail¹, Si Yuen Fong¹ and Ali Najah Ahmed²

¹*School of Marine and Environmental Sciences, Universiti Malaysia Terengganu, 21030, Kuala Terengganu, Terengganu, Malaysia.*

²*Faculty of Engineering, Universiti Tenaga Nasional, 43650, Bangi, Selangor, Malaysia. marzuki@umt.edu.my*

Abstract—The forecasting of Particulate Matter (PM₁₀) is crucial as the information can be used by local authority in informing community regarding the level air quality at specific location. The non-linearity of PM₁₀ in atmosphere after it was subjected by several meteorological parameters should be treated with powerful statistical models which can provide high accuracy in forecasting the PM₁₀ concentration for instance Neural Network (NN) model. Thus, the aim of this study is establishment of NN model using Levenberg-Marquardt training algorithm with meteorological parameters as predictors. Daily observations of PM₁₀, wind speed, relative humidity, ambient temperature, rainfall, and atmospheric pressure in Kuala Terengganu, Malaysia from January 2009 to December 2014 were selected for predicting PM₁₀ concentration level. Principal Component Analysis (PCA) was applied prior the establishment of NN model with the aim of reducing multi-collinearity among predictors. The three principal components (PC-1, PC-2, PC-3) as the result of PCA was used as the input for the NN model. The NN model with 14 hidden neurons was found as the best model having MSE of 0.00164 and R values of 0.80435(Training stage), 0.85735(Validation stage), and 0.8135(Testing stage). Overall the model performance was achieved as high as 81.1% for PM₁₀ forecasting in Kuala Terengganu.

Index Terms—Kuala Terengganu; Levenberg-Marquardt; Particulate Matter; Principal Component Analysis.

I. INTRODUCTION

Malaysia is a developing country which facing the deterioration of air quality due to rapid urbanization, industrialization and population growth [1]. Air Pollution Index (API) is known to be the indicator of air quality at specific location, and has shown that the particulate matter having aerodynamic size of 10 micrometer and less (PM₁₀) was the dominant pollutant among the rest especially in Peninsular Malaysia. Particulate matter (PM) is a mixture of solid and liquid particles that suspended in air [2]. Combustion products such as smoke, fumes, soot and natural particles such as windblown dust, sea salt, pollen and spores also include as particulate matter [3]. Mobile sources such as the emission from motor vehicles, stationary sources such as the emission from power plants and factories, and open burning are the three listed major sources of air pollution in Malaysia [4]. Recently, the association between airborne particulate matter and wide range of health effects has been

found. It is estimated that approximately 3% of cardiopulmonary and 5% of lung cancer deaths are attributable to particulate matter globally [5].

PM₁₀ after emitted from sources is subjected by several meteorological factors such as wind speed, ambient temperature, relative humidity, rainfall, and atmospheric pressure which made it becomes complex and therefore having nonlinear character in atmosphere [6]. This nonlinearity of PM₁₀ in atmosphere is thus must be deal by nonlinear model rather than traditional model (linear). Artificial Neural Network (ANN) has been known to overcome the linearity problem in statistical modeling especially for air pollution forecasting. The incorporation of ANN utilizing the training algorithm to train the data might increase the accuracy of forecasting models rather than using traditional linear method such as MLR. The Levenberg-Marquardt training algorithm was used as it trains the moderate-sized feed forward neural network in fastest way. Nowadays, there are many studies performed on forecasting by using linear models in Malaysia. Thus, there is a need considering the non-linear model development especially for this Kuala Terengganu station. The establishment of such model by taking into consideration of meteorological factors are very important in forecasting PM₁₀ concentrations as the meteorological factors control the transmission, deposition, and dispersion of PM₁₀ concentration in atmosphere. Therefore, the aim of this study is the establishment of NN model using Levenberg-Marquardt training algorithm in Kuala Terengganu. This model is very useful at the local level to gives information which allows the authority and people within a community to take precautionary measures to avoid or limit their exposure to unhealthy levels of air quality and implement significant actions oriented to improve air quality on specific locations

II. METHODOLOGY

A. Site Description

Terengganu is located along the east coast of Peninsular Malaysia facing South China Sea. Kuala Terengganu air monitoring station (N05°18.455'; E103°07.213') is situated at SK Pusat Chabang Tiga, located near to the Kuala Terengganu city center (Figure 1). This monitoring station is affected by busy traffic, especially during the rush hour in the morning and

late afternoon and several meteorological conditions. The factors influencing of air pollution in this area were associated with local traffic, seasonality and open burning [7].



Figure 1: Study Area

B. Data Acquisition

This study was based on the data measured for the 7 years period, from January 2009 to December 2014. The PM_{10} concentration data used in this study was recorded as part of a Malaysian Continuous Air Quality Monitoring (CAQM) program, using the β -ray attenuation mass monitor (BAM-1020) as manufactured by Met One Instruments Inc. The monitoring network was installed, operated and maintained by Alam Sekitar Malaysia Sdn. Bhd (ASMA) on behalf of the Malaysian Department of Environment [8]. Six daily averaged parameters were used in order to gain a better understanding of PM_{10} variability. The parameters that were used in this study are particulate matter with aerodynamic diameter less than $10\ \mu\text{m}$ of previous day ($PM_{10,t-1}\ \mu\text{g}\cdot\text{m}^{-3}$), ambient temperature ($^{\circ}\text{C}$), relative humidity (%), wind speed (ms^{-1}), Atmospheric Pressure (hPa) and rainfall amount (millimeter). The monitoring records were obtained from the Air Quality Division, Department of Environment (DOE), Ministry of Natural Resources and Environment of Malaysia for $PM_{10,t-1}$, ambient temperature, relative humidity, and wind speed. Atmospheric Pressure and rainfall amount were two parameters of meteorological factors that acquired from MMD due to the limited meteorological data from the DOE. The MMD station selected is from the nearest AQMS of DOE. Several studies performed shows that they are also using the meteorological data from the nearby or nearest meteorological stations. The data from different air quality monitoring stations and meteorological stations are then combined together to develop models for air quality forecasting [9]. In this study, the nearest meteorological station for the Chabang Tiga is at Kuala Terengganu Meteorological Station, which is about 8.715km North-North-West (NNW) of the Chabang Tiga. This station is located at height of 5.0 m above MSL. In literature, it can be up to 20km distance [10] and

therefore the data from these two stations can be combined to forecast PM_{10} concentration.

C. Imputation of Missing Values

In air pollution studies, missing values may occur because of equipment malfunctioned or of errors in measurements [11]. Incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset [12]. In this study, the incomplete data was treated by imputation of missing values of linear interpolation. This imputation technique has widely used by previous researcher [13]. Interpolation is a method of finding new values for any function using the given set of values. In SPSS®, replacing missing values using a linear interpolation meaning that the last valid value before the missing value and the first valid value after the missing value are used for the interpolation. The unknown value at a particular point can be found using many interpolation formula's. Details can be referred in [14].

D. Data Normalization

In this study, the data of dependent and independent variables consist of different units and therefore normalization of data is required. The normalization produce the data are scaled within the range of 0 to 1 [0 1]. This scaling is suitable for improving the accuracy of numeric computation carry out by the NN models for the better outputs. The min-max technique is used. The advantage is preserving exactly all relationship in the data and it does not introduce bias.

E. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variable into a set of values of linearly uncorrelated variables called principal components (PCs) [15]. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under constraint than is orthogonal to (i.e. uncorrelated with) the preceding components. This method has been applied by several researchers such as [16]-[18].

F. Training Algorithm

In conjunction in optimizing the minimum of a multivariate function, the sum of squares of non-linear real valued functions was expressed by an iterative technique known as Levenberg-Marquardt algorithm. This algorithm is known as standard technique whereby it has been adopted in many fields or discipline. Levenberg-Marquardt uses an estimation to Hessian matrix where it is specializing in optimum solution. Details were explained in [19].

G. Neuron Number Selections

The fitting of neurons in hidden layer was performed by trial and error whereby one neuron is try one after another to avoid any under fitting or over fitting problem in developing the NN model. There is no current guidelines in determining the exact number of neuron in hidden layer, but Ul-Saufie et al., 2011

noted that the number of neuron should not be larger two times than the number of input parameters [20].

III. RESULTS AND DISCUSSION

A. Descriptive Statistics of PM₁₀ Concentration

Figure 2 shows the box plot of daily PM₁₀ concentrations in Kuala Terengganu. The box plot is a simple graphical display that is ideal for making comparisons [21]. The highest daily average of PM₁₀ concentration was observed in 2014 that being 208.25 µg/m³, while the lowest was observed in the same year that being 9.56 µg/m³. Generally, Malaysia experienced high PM₁₀ concentrations during the second and third quarter of the year as a result of trans-boundary smoke from the forest fire in Sumatera region during dry season from May to September [22]. East coast of Peninsular Malaysia also affected by this trans-boundary smoke and therefore recorded high concentration of PM₁₀.

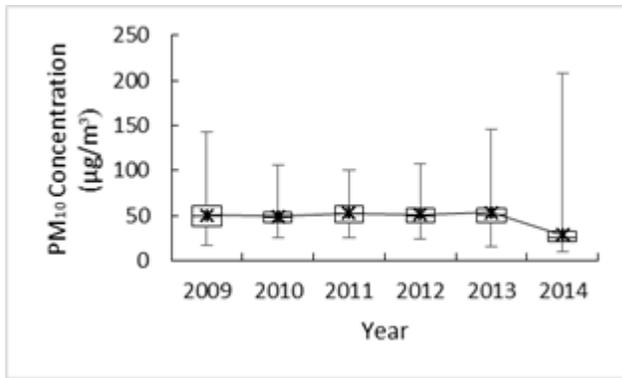


Figure 2: Boxplot of daily PM₁₀ concentration

The descriptive statistics for PM₁₀ concentrations during the study period (2009-2014) is summarized in Table 1. The highest mean concentration of PM₁₀ was recorded as 53.32 µg/m³ (15.17- 146.46 µg/m³) in 2013, while the lowest mean value of PM₁₀ concentration was in 2014 with 28.60 µg/m³ (9.56 – 208.25 µg/m³). Department of Environment Malaysia (DOE) has set the national ambient air quality standards for PM₁₀ concentration in ambient air which the yearly average has been determined as 50 µg/m³ [23]. Most of the average reading exceeded the New Ambient Air Quality Standard (NAAQS) with 50.77 µg/m³, 52.53 µg/m³, 51.25 µg/m³, and 53.32 µg/m³, in year 2009, 2011, 2012, and 2013, respectively. Most of the data also skewed to the right (above 1) especially in year 2010, 2012, 2013, and 2014 indicate of the existence of the extreme concentrations during the particular years, which promote increasing of PM₁₀ concentration. This scenario can also be explained by using coefficient of variation (CV) value. Similar to skewness value, this Kuala Terengganu station recorded the high value for CV most of the time, with highest CV of 0.495 in 2014. The variable with the smaller CV is less dispersed than the variable with the larger CV. This station has urban background, and it is believed that the increasing number of motor vehicles, industries and street dust level are likely to contribute to the total of suspended particulate matter in the atmosphere at this study area, with urban background [24].

Table 1
Descriptive Statistics of PM₁₀ concentration

	2009	2010	2011	2012	2013	2014
Mean	50.77	49.33	52.53	51.25	53.32	28.60
Median	49.33	48.09	51.70	49.29	51.17	25.83
Std. Deviation	16.46	10.46	13.58	12.31	16.91	14.15
Skewness	0.666	1.109	0.539	1.037	1.887	6.281
Coefficient of variation	0.324	0.212	0.258	0.240	0.317	0.495
Min	17.09	25.88	25.58	24.67	15.17	9.56
Max	142.67	105.46	100.18	107.39	146.46	208.25

B. Polar Plot

Polar plot was constructed to visualize the distribution of PM₁₀ concentration over the study period spatially. This polar plot was constructed by using Open Air R Package and the daily average of wind direction, wind speed, and PM₁₀ concentration were used. During study period, the PM₁₀ concentration was heavily distributed from the NE direction particularly affected the area of SW direction (Figure 3). It shows the PM₁₀ concentration is dispersed away heavily in 10-15 km/hr of wind speed. The community staying on that area should be notified to take more precautionary actions in reducing the exposure to PM₁₀ concentration, which therefore concerning their health effects.

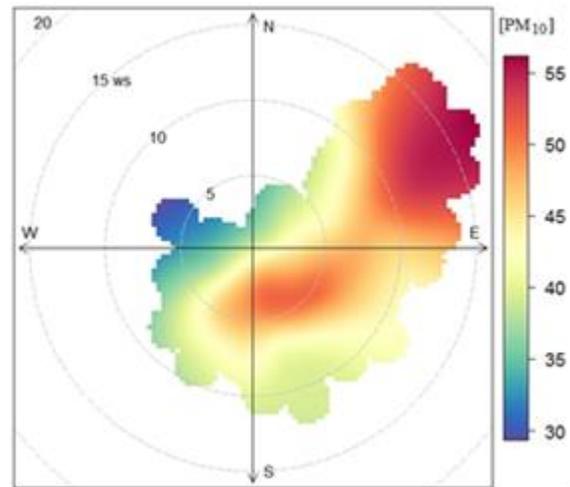


Figure 3: Spatial distribution of PM₁₀ concentration

C. Principal Component Analysis (PCA)

The requirements prior to the PCA were satisfied subjected to the data which the Kaiser-Meyer-Oklin (KMO) of Sampling Adequacy was 0.522 (>0.50) and Bartlett's Test of Sphericity value was 0.000 (<0.001). Six parameters were subjected to this analysis. The selection of Principal Component (PC) was based on the eigenvalues (greater than 1). This study selects three PCs as the third eigenvalues in third PC was near to 1. These three PCs explained about 73.3% of percent reliability on overall data. The rotated Kaiser matrix utilizing Kaiser Normalization was applied on these three factors. The output is suppressed with values less than 0.4. It was found that the PC-1 associated with relative humidity, ambient temperature and rainfall amount, PC-2 composed of atmospheric pressure and wind speed, while PC-3 consists of lag-PM₁₀. The application of PCA prior the establishment of NN model to make sure that the

multi-collinearity between the independent parameters is reduced. IBM SPSS® Statistics version 22 was used in conducting PCA.

D. Neural Network Fitting

The result of PCA was adopted as the input for establishment of NN model. Therefore, the NN has 3 inputs namely PC-1, PC-2, and PC-3 and the network executes one output parameter (PM₁₀ concentration). The data was divided into three parts; 70% for training, 15% for validation, and another 15% for testing of model. The exact number of neuron in hidden layer was identified by trial and error method, whereby the lower mean square error value was selected as the best number of neuron in hidden layer. It was found that, for the data comprises from year 2009-2014 in Kuala Terengganu, the best number of neurons were 14 in hidden layer which having the lowest Mean Square Error (MSE) of 0.00164 among the rest (Table 2)

Table 2
List of neuron numbers and MSE values

No. of neuron	MSE
1	0.00312
2	0.00221
3	0.00181
4	0.00226
5	0.00387
6	0.00209
7	0.00205
8	0.00210
9	0.00322
10	0.00215
11	0.00196
12	0.00238
13	0.00180
14	0.00164
15	0.00284
16	0.00213

Figure 4 shows the architecture of NN model with 3 inputs, 14 neurons in hidden layer, and one output.

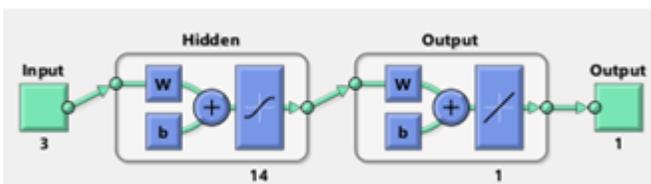


Figure 4: Architecture of NN model

The validation stops when it reaches the best fitted epochs for training and testing. Figure 5 shows the validation process stops at epochs=27 out of 33 epochs. The best validation performance was found as 0.00164 of MSE. Figure 6 shows the regression plot for each stage from training, validation, testing, and overall performance of NN model with 14 neurons in hidden layer. The R values show strong relationship of input-output mapping with 0.80435(Training stage), 0.85753(Validation stage), and 0.8135(Testing stage). Overall the model performance was achieved as high as 81.1%. MATLAB R2015a was utilized in establishment of NN model.

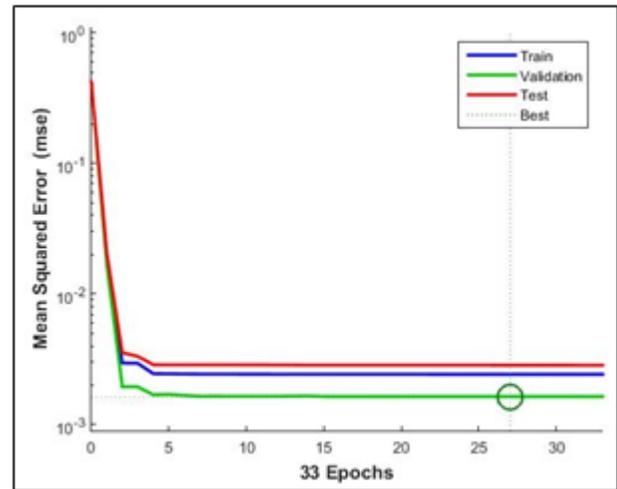


Figure 5: Performance checks of NN model

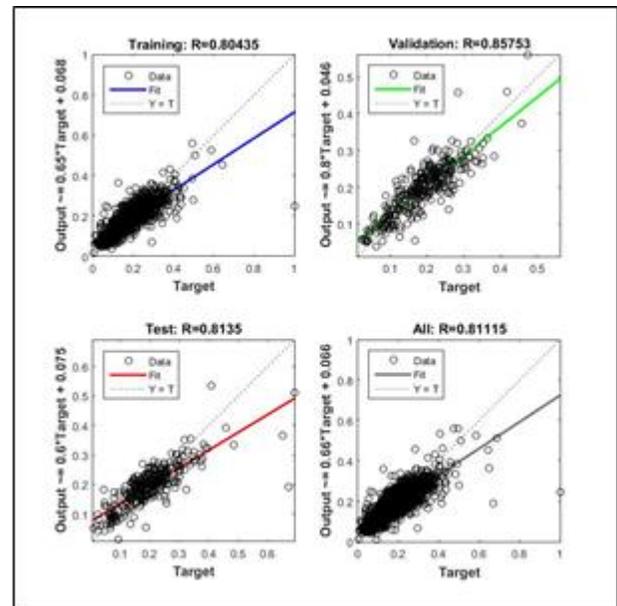


Figure 6: Regression plot

IV. CONCLUSION

The daily data of PM₁₀ concentrations and meteorological factors from year 2009 to 2014 were used to establish NN model. This study showed PM₁₀ concentrations were exceeded the limit set in MAAQG of 50 µg/m³ during the study period, except for year 2010 and 2014. Polar plot shows that PM₁₀ concentration was heavily distributed from the NE direction particularly affected the area of SW direction and was dispersed away heavily in 10-15 km/hr of wind speed. The lowest MSE found was 0.00164 when using Levenberg-Marquardt training algorithm with 14 neurons in hidden layer. The R values show adequate accuracy in forecasting PM₁₀ concentration with 0.80435 (Training stage), 0.85753(Validation stage), and 0.8135(Testing stage). Overall the model performance was achieved as high as 81.1%. The established NN model is appropriate for forecasting PM₁₀ concentrations intended for early warnings system for public health as well as for local

authorities to formulate strategies in improving the air quality at Kuala Terengganu.

ACKNOWLEDGMENT

This study was funded by Fundamental Research Grant Scheme (FRGS) FRGS/2/2013/STWN01/UMT/02/1 (VOT 59375) and Universiti Malaysia Terengganu Scholarship Scheme (BUMT) to Author 1 and Author 3. The authors also would like to thank the Air Quality Division, Malaysian Department of Environment (DOE) and Malaysian Meteorological Department (MMD) for the air quality and meteorological data.

REFERENCES

- [1] M.T. Latif, S.Z. Azmi, A.D.M. Noor, A.S. Ismail, Z. Johny, S. Idrus, A.F. Mohamed, and M. Mohktar. "The impact of urban growth on regional air quality surrounding the Langat river basin. Malaysia", in *Environmentalist*, vol. 31, pp. 315-324, 2010.
- [2] D.M. Markovic, D.A. Markovic, A. Jovanovic, L. Lazic, and Z. Mijic. "Determination of O₃, NO₂, SO₂, CO and PM₁₀ measured in Belgrade urban area", in *Environmental Monitoring and Assessment*, vol. 145(1), pp. 349-359, 2008.
- [3] D.W. Dockery, "Health effects of particulate air pollution", in *Annals of Epidemiology*, vol. 19(4), pp. 257-263, 2009.
- [4] R. Afroz, M.N. Hassan, and N.A. Ibrahim. "Review of air pollution and health in Malaysia", in *Environmental Research*, vol. 92, pp. 71-77, 2003.
- [5] D.C. Shin. "Health effects of ambient particulate matter", in *Journal of the Korean Medical Association*, vol. 50 (2), pp. 175-182, 2007.
- [6] X. Querol, A. Alastuey, C.R. Ruiz, B. Artinano, H.C. Hansson, R.M. Harisson, E. Buringh, H.M. ten Brink, M. Lutz, P. Bruckmann, P. Straehl, and J. Schneider. "Speciation and origin of PM₁₀ and PM_{2.5} in selected European Cities", in *Atmospheric Environment*, vol. 38, pp. 6547-6555, 2004.
- [7] S. Abdullah, M. Ismail, S.Y. Fong, and A.N. Ahmed. "Principal Component Regression (PCR) for PM₁₀ forecasting in Kuala Terengganu, Terengganu", in *Proceedings of National Conference on Wood based Technology, Engineering and Innovation*, pp. 82-88, 2015.
- [8] R. Afroz, M.N. Hassan, and N.A. Ibrahim. "Review of air pollution and health in Malaysia", in *Environmental Research*, vol. 92, pp. 71-77, 2003.
- [9] G.D. Gennaro, L. Trizio, A.D. Gilio, J. Pey, N. Perez, M. Cusack, A. Alastuey, and X. Querol, X. "Neural network model for the prediction of PM₁₀ daily concentrations in two sites in the Western Mediterranean", in *Science of the Total Environment*, vol. 463-464, pp. 875-883, 2013.
- [10] H. Niska, T. Hiltunen, A. Karppinen, J. Ruuskanen, and M. Kolehmainen. "Evolving the neural network model for forecasting air pollution time series", in *Engineering Applications of Artificial Intelligence*, vol. 17, pp. 159-167, 2004.
- [11] M.N. Noor, and M.I. Zainudin, M.L. "A review: Missing values in environmental data sets", in *Proceeding of International Conference on Environment, Malaysia*, 2008.
- [12] G. Hawthorne, and P. Elliot. "Imputing cross-sectional missing data: comparison of common techniques", in *Australian and New Zealand Journal of Psychiatry*, vol. 39, pp. 583-590, 2005.
- [13] R. Yu, X.C. Liu, T. Larson, and Y. Wang. "Coherent approach for modeling and nowcasting hourly near-road black carbon concentrations in Seattle, Washington", in *Transportation Research Part D*, vol. 34, pp. 104-115, 2015.
- [14] Chapra, S. C., and Canale, R.P. *Numerical Methods for Engineers*. Singapore: McGra-Hill, 1998.
- [15] S.A. Abdul-Wahab, C.S. Bakheit, and S.M. Al-Alawi. "Principal component multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations", in *Environmental Modeling and Software*, vol. 20, pp. 1263-1271, 2005.
- [16] S. Abdullah, M. Ismail, M., S.Y. Fong, and A.N. Ahmed. "Evaluation for long term PM₁₀ concentration forecasting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) models.", in *EnvironemntAsia*, vol. 9(2), pp. 101-110, 2016.
- [17] N.R. Awang, N.A. Ramli, A.S. Yahaya, and M. Elbayoumi, "Multivariate methods to predict gound level ozone during daytime, nighttime, and critical conversion time in urban areas", in *Atmospheric Pollution Research*, vol. 6, pp. 726-734, 2015.
- [18] A.Z. Ul-Saufie, A.S. Yahaya, N.A. Ramli, N. Rosida, and H.A. Hamid. "Future daily PM₁₀ concentrations prediction by combining regression models and feedforward back propagation models with Principle Component Analysis (PCA)", in *Atmospheric Environment*, vol. 77, pp. 621-630, 2013.
- [19] I.N. Daliakopoulos, P. Coulibaly, and I.K. Tsanis, I.K. "Groundwater level forecasting using artificial neural network", in *Journal of Hydrology*, vol. 309, pp. 229-240, 2005.
- [20] A.Z. Ul-Saufie, A.S. Yahya, N.A. Ramli, and H.A. Hamid. "Comparison between multiple linear regression and feed forward back propagation neural network models for predicting PM₁₀ concentration level based on gaseous and meteorological parameters", in *International Journal of Applied Science and Technology*, vol. 1(4), pp. 42-49, 2011.
- [21] N.A. Ramli, N.A. Ghazali, and A.S. Yahaya "Diurnal fluctuations of ozone concentrations and its precursors and prediction of ozone using multiple linear regressions", in *Malaysian Journal of Environmental Management*, vol. 11(2), pp. 57-69, 2010.
- [22] H. Ahmat, A.S. Yahaya, and N.A. Ramli, N. A. "The Malaysia PM₁₀ analysis using extreme value", in *Journal of Engineering Science and Technology*, vol. 10(12), pp. 1560-1574, 2015.
- [23] Department of Environment. *Malaysia Environmental Quality Report 2014*. Kuala Lumpur: Department of Environment Malaysia. 2015.
- [24] R. Afroz, M.N. Hassan, and N.A. Ibrahim. "Review of air pollution and health in Malaysia", in *Environmental Research*, vol. 92, pp. 71-77, 2003.