Malaysian Journal of Mathematical Sciences 13(S) December: 85–95 (2019)
Special Issue: Conference on Mathematics, Informatics and Statistics (CMIS2018)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Multi-Layer Perceptron Model for Air Quality Prediction

Abdullah, S. *[1], Ismail, M.[2], and Ahmed, A. N.[3]

[1]*School of Ocean Engineering, Universiti Malaysia Terengganu, Malaysia*
[2]*School of Marine and Environmental Sciences, Universiti Malaysia Terengganu, Malaysia*
[3]*College of Engineering, Universiti Tenaga Nasional, Malaysia*

E-mail: samsuri@umt.edu.my
* *Corresponding author*

ABSTRACT

This study trained two MLP models with different activation functions in assessing the capability of the model for the prediction of air quality. The daily air quality data and meteorological variables from year 2010-2014 were assembled in training and testing the models. The MLP model with the combination of tansig and purelin activation function revealed 69.0% of variance in data with 5.58 $\eta g/m^3$ (RMSE) and 80.0% of variance in data with 8.14 $\eta g/m^3$ (RMSE), during training and testing phase, respectively. This model is appropriate for operational used by respected authorities in managing air quality and as early warning during unhealthy level of air quality.

Keywords: Air quality, Meteorological and Prediction.

1. Introduction

The triggering factors in diminishing air quality in Malaysia are including the rapid development of population, urbanization and industrialization (Latif et al. (2011)). Among the criteria pollutants, particulate matter or known as PM_{10} that having size of 10 micron and less is responsible for the deterioration of air quality in Malaysia (Department of Environment (2015)). Emission from motor vehicles, open combustion and industrial activities are listed as the sources of PM_{10} (Department of Environment (2015)). Due to its small size, it can be inhaled and deposited in our respiratory system, which then in long-term can caused several respiratory illness (Buteau and Goldberg (2016)). It is understood that, PM_{10} once it is emitted into the atmosphere by several sources are subjected by numerous meteorological factors, which is then made it nonlinear in nature. The influence of gaseous pollutants also increased the complexity of PM_{10} in atmosphere. These nonlinearity and complexity of PM_{10} data should best be captured by nonlinear model. The emergence of several computational techniques in machine learning such as Artificial Neural Network (ANN) seems best to be used for the purpose of PM_{10} prediction in dealing the nonlinearity and complexity. Thus, this study proceeds with the prediction of PM_{10} concentration using one of the ANN function known as MLP for the suburban area in Kuantan, Pahang.

2. Materials and Methods

2.1 Study Area

The Air Quality Monitoring Station (AQMS) is located at the SK Indera Mahkota, Kuantan (N03^{o}49.138; E103^{o}17.817) as shown in Figure 1. It is classified as the suburban background by the Department of Environment, Malaysia. Kuantan is monopolized by the tourism industry and petrochemical industry (Department of Environment (2015)).

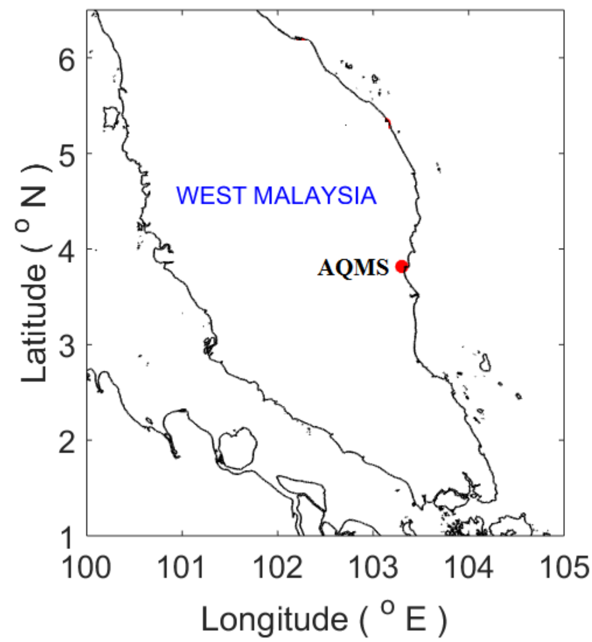

Figure 1: The location of the Air Quality Monitoring Station (AQMS) in Kuantan, Malaysia.

2.2 Monitoring Records

The 5 years (2010-2014) monitoring records including particulate matter having aerodynamic diameter of below 10 ηm (PM_{10}, $\eta g/m^3$), ambient temperature (^{o}C), relative humidity (%), wind speed (m/s), carbon monoxide (CO, ppm), nitrogen dioxide (NO_2, ppm) and sulfur dioxide (SO_2, ppm) were utilized in this study. The observing records were acquired from the Department of Environment, Malaysia. The data was separated into two sections; training data set (Year 2010-Year 2012) and testing data set (Year 2013-Year 2014), respectively.

2.3 Data Pre-Processing

The missing values in the data set was inserted by a method of linear interpolation that is proven suitable previously by Mohamed et al. (2008). The linear interpolation imputed missing data in a straight line and is given as:

$$f(x) = f(x) + \frac{f(x_1) - f(x_o)}{x_1 - x_o}(x - x_o) \tag{1}$$

where x = independent value, x_1 and x_o = known values of the independent variable, and $f(x)$ = value of dependent variable for a value of the independent variable.

The data is normalized due to the occurence of different units of measurement. The suitable min-max technique is used in this study in generating new set of normalized data in the range of [0 1] (de Mattos Neto et al. (2014)). The data was normalized by using the following formula(Zhao et al. (2015)):

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{2}$$

where $x = (x_1, ..., x_n)$ and z_i is the ith normalized data.

2.4 Multi-Layer Perceptron Model (MLP)

Multilayer Perceptron (MLP) is a type of feed forward neural network which has the ability in developing the nonlinear models with high complexity, make it preferred in air pollution prediction. The MLP starts when the interested input parameters are feed into the network. These input parameters provide input signals and these signals are sent in the network starting from input layer to hidden layer and hidden layer to output layer. The scaled input vector which introducing by neurons in the input layer is multiplied with weights, which a real number quantity. The neuron in the hidden layer sums up this information, including bias.

$$y_o = \sum_{i=1}^{n} w_i x_i + b \tag{3}$$

This weighted sum information is still in its linear model. The non-linearity of information or model occurs when it is passing through the activation or transfer function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

Then,

$$y_o = f(x)[\sum_{i=1}^{n} w_i x_i + b] \tag{5}$$

where y_o = output, w_i = weight vector, x_i = scaled input vector, b = bias, f = transfer function and x = total sum of weighted inputs.

Once the error signal is computed, the process of model fitting is ended. The difference between target and output is used to compute the error signal in the model which corresponds to the input.

Mathematically, the equation of MLP with several numbers of neurons is given as;

$$y_o = f[\sum WO_{kj}(\sum_{i=1}^{n} WI_{ij} x_i + b_1) + b_2] \tag{6}$$

where WI_{ij} = weight of input layer, WO_{kj} = weight of output layer, b_1 = biased in input layer and b_2 = biased in output layer.

The determination of neuron numbers in the hidden layer is the most critical part as there is a strong impact on the output in which determines by the neuron numbers in the hidden layer. Under-fitting will occur if the number of the neuron is too few and too many of neurons contribute to over-fitting, thus the determination of an optimum number of the neuron is really important. Trial and error method was employed for the determination of neuron numbers in the hidden layer. There are no standard rules on the minimum and a maximum number of neurons in the hidden layer (Sun et al. (2008)). Ul-Saufie et al. (2013) propose and effectively tried that the number of neurons in the hidden layer isn't bigger than twice of the number of inputs.

The activation or transfer function played an important role in ANN by producing non-linear decision through non-linear blends of weighted inputs. This activation function is then transform input to output signals. This activation function that presenting the nonlinearity in the MLP model and subsequently make it contrasts from the linear model. Sigmoid units of activation functions map input to a value in the range of 0 to 1. This study utilizes three sigmoid functions which are generally utilized as a part of MLP, for example, purelin (linear), log-sigmoid (logsig), and hyperbolic tangent sigmoid (tansig) (Arjun and Aneesh (2015) and Wong et al. (2015)). The initialization of the model was trained with 5000 epochs and the model was run 3 times to get the average of the value. The learning rate is set as 0.05 to train the optimum configuration of MLP network and to improve the performance of the model. MATLAB R2015a was used for models training and testing.

2.5 Performance Indicators

Performance Indicators (PI) were used in determining the best-fitted models in this study. Based on PI, there are two things that are taken into account, namely accuracy measure and error measure. The accuracy measure indicates that the best-fitted model is when the value is closer to 1 while error measure indicates the best model when the evaluated value is closer to 0.

(a) Root Mean Square Error

$$RMSE = (\frac{1}{n}\sum_{i=1}^{n}[P_i - O_i]^2)^{1/2} \tag{7}$$

(b) Correlation Coefficient

$$R^2 = [(\frac{1}{n.S_{pred}.S_{obs}})\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})]^2 \tag{8}$$

where n = total number of data, P_i = predicted values, O_i = observed values, $\bar{P}$ = mean of predicted values, $\bar{O}$ = mean of observed values, S_{pred} = standard deviation of predicted values and S_{obs} = standard deviation of observed values

3. Results and Discussion

The boxplots showing the temporal variation of PM_{10} concentrations during the study period (2010-2014) are summarized in Figure 2. PM_{10} concentration has a steady fluctuations trend from year to year as shown in Figure 2. None of the yearly average reading for Kuantan site exceed the New Ambient Air Quality Standard (NAAQS). The daily PM_{10} concentration also did not exceeded the NAAQS, except for year 2013 which Kuantan site recorded maximum concentration of 265.13 $\eta g/m^3$ on 23 June 2013. On 24 June 2013, this Kuantan site also records high PM_{10} concentration which exceeded the limit with 167.75 $\eta g/m^3$. Overall, it was noted that the exceeding daily concentrations of PM_{10} occurred during month of June. This month was noted as in the dry season months (May to September) in Malaysia. Generally, during this dry season, Malaysia experiences high PM_{10} concentrations due to the transboundary smoke which resulted from the forest fire in Sumatera region (Ahmat et al. (2015) and Amanollahi et al. (2011)).

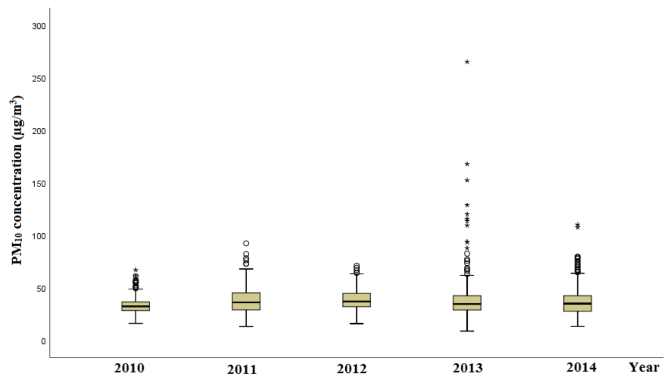

Figure 2: Temporal variation of PM_{10} concentration

The number of neurons in the hidden layer are assessed through trial and error method with respect to early stopping technique by dividing the data into training and testing set. Table 1 shows outcome utilizing distinctive activation functions for MLP model, together with the range of the tried number of neurons in the hidden layer. The selection of best model during training phase is strictly based on the RMSE and R^2. It should be noted that the best results are marked with bold. It was found that, the best RMSE and R^2 for Logsig-Purelin and Tansig-Purelin activation function were 5.70 $\eta g/m^3$ and 0.68, and 5.58 $\eta g/m^3$ and 0.69, respectively. The optimum number of neurons in hidden layer for the Logsig-Purelin combination is 14, while for Tansig-Purelin is 13. Thus, comparing both models, the MLP with activation function of Tansig-Purelin is chosen in representing the best prediction model for this suburban area as it has low RMSE and high R^2 as compared to Logsig-Purelin of MLP model. The testing of Tansig-Purelin MLP model was 8.14 $\eta g/m^3$ (RMSE) and 0.80 (R^2) as shown in Figure 3. The combination of Tansig-Purelin activation function for MLP model is similar with the results deliberated by Voukantsis et al. (2011), Singh et al. (2012) and Feng et al. (2015).

Great performance of nonlinear models were because of the way that they can display very nonlinear functions and can be prepared to precisely sum up when given new, concealed data. Moreover, ANN don't require any earlier assumptions with respect to the conveyance of training data and no decision in regards to the relative significance of the different info estimations should be made (Chen and Kim (2006)). These highlights of the neural networks make them alluring contrasting option to creating numerical models and furthermore while picking between statistical methodologies.

Table 1: Result using Logsig-Purelin activation functions

Number of Neurons in Hidden Layer	RMSE (ηg/m^3)	R^2
1	6.86	0.53
2	6.77	0.55
3	6.63	0.56
4	6.49	0.58
5	6.49	0.58
6	6.35	0.60
7	6.17	0.62
8	6.17	0.62
9	6.16	0.62
10	5.98	0.65
11	5.91	0.65
12	5.89	0.66
13	5.71	0.68
14	**5.70**	**0.68**

Table 2: Result using Tansig-Purelin activation functions

Number of Neurons in Hidden Layer	RMSE (ηg/m^3)	R^2
1	6.86	0.53
2	6.83	0.54
3	6.71	0.55
4	6.50	0.58
5	6.40	0.59
6	6.45	0.59
7	6.38	0.60
8	6.40	0.59
9	6.19	0.62
10	6.03	0.64
11	6.10	0.63
12	5.91	0.65
13	**5.58**	**0.69**
14	5.59	0.69

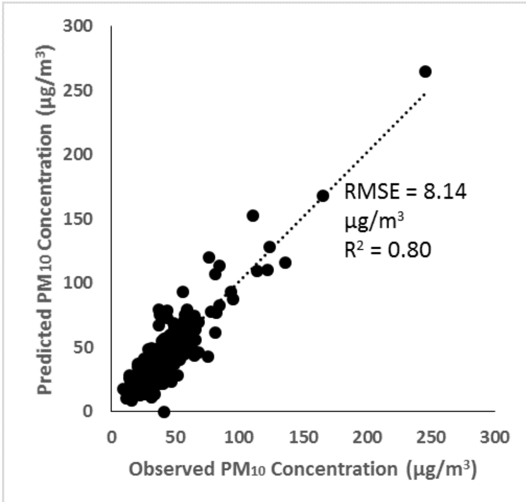

Figure 3: The predicted and observed of PM_{10} concentration during testing of Tansig-Purelin MLP model

4. Conclusions

This study shows that the complexity and nonlinearity of PM_{10} in atmosphere are best captured by MLP with the combination of Tansig-Purelin activation function as showed by strong agreement (R^2=0.69, training) and (R^2 = 0.80, testing) between predicted and observed data with respect to meteorological and gaseous pollutants parameters. The model is robust and is ready for operational used. It is suggested for the local authority and DOE to use the MLP model in predicting PM_{10} concentration for improving air quality at specific location and as an early warning to inform the community for them to reduce the outdoor activities due to the unhealthy level of air quality.

Acknowledgement

The authors would like to thank the Air Quality Division, Malaysian Department of Environment (DOE) for the air quality and meteorological data.

References

Ahmat, H., Yahaya, A. S., and Ramli, N. A. (2015). The Malaysia PM_{10} analysis using extreme value. *Journal of Engineering Science and Technology*, 10(12):1560–1574.

Amanollahi, J., Abdullah, A. M., Ramli, M. F., and Piratesh, S. (2011). Real time assessment of haze and PM_{10} aided by modis aerosol optical thickness over klang valley. *Malaysia World Applied Sciences Journal*, 14:8–13.

Arjun, K. S. and Aneesh, K. (2015). Modelling studies by application of artificial neural network using matlab. *Journal of Engineering Science and Technology*, 10(10):1477–1486.

Buteau, S. and Goldberg, M. S. (2016). A structured review of panel studies used to investigate associations between ambient air pollution and heart rate variability. *Environmental Research*, 148:207–247.

Chen, H. and Kim, A. S. (2006). Forecasting of permeate flux decline in cross flow membrane filtration of colloidal suspension: a radial basis function neural network approach. *Desalination*, 192:415–428.

de Mattos Neto, P. S. G., Madeiro, F., Ferreira, T. A. E., and Cavalcanti, G. D. C. (2014). Hybrid intelligent system for air quality forecasting using phase adjustment. *Engineering Applications of Artificial Intelligence*, 32:185–191.

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*, 107:118–128.

Latif, M. T., Azmi, S. Z., Noor, A. D. M., and Ismail, A. S. (2011). The impact of urban growth on regional air quality surrounding the langat river basin. *Malaysia Environmentalist*, 31:187–196.

Mohamed, N. N., Yahaya, A. S., Ramli, N. A., and Abdullah, M. M. A. L. (2008). Estimation of missing values in air pollution data using singe imputation techniques. *ScienceAsia*, 34:341–345.

Department of Environment. (2015). *Malaysia Environmental Quality Report 2014*. Ministry of Sciences, Technology, and the Environment, Malaysia, Kuala Lumpur, 1st edition.

Singh, K. P., Gupta, S., Kumar, A., and Shukla, S. P. (2012). Linear and nonlinear modeling approaches for urban air quality forecasting. *Science of the Total Environment*, 426:244–255.

Sun, G., Hoff, S. J., Zelle, B. C., and Nelson, M. A. (2008). Development and comparison backpropagation and generalized regression neural network models to predict diurnal and seasonal gas and PM_{10} concentrations and emissions from swine buildings. *American Society of Agricultural and Biological Engineers*, 2(51):685–694.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., Rosaida, N., and Hamid, H. A. (2013). Future daily PM_{10} concentrations forecasting by combining regression models and feedforward backpropagation models with principal component analysis (pca). *Atmospheric Environment*, 77:621–630.

Voukantsis, D., Karatzas, K., Kukkonen, J., Rasanen, T., Karppinen, A., and Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks in thessalloniki and helsinki. *Science of the Total Environment*, 409:1266–1276.

Wong, M. S., Xiao, F., Nichol, J., Fung, J., Kim, J., Campbell, J., and Chan, P. W. (2015). A multi-scale hybrid neural network retrieval model for dust storm detection, a study in Asia. *Atmospheric Research*, 158-159:89–106.

Zhao, N., Wen, X., Yang, J., Li, S., and Wang, Z. (2015). Modelling and forecasting of viscosity of water-based nanofluids by radial basis function neural networks. *Powder Technology*, 281:173–183.