Journal of Sustainability Science and Management *eISSN: 2672-7226*
Volume 14 Number 3, June 2019: 7-11 *©Penerbit UMT*

DAYTIME OZONE CONCENTRATION PREDICTION USING STATISTICAL MODELS

NORHAZLINA SUHAIMI[1], NURUL ADYANI GHAZALI[1*], MUHAMMAD YAZID NASIR[1], MUHAMMAD IZWAN ZARIQ MOKHTAR[1], NOR AZAM RAMLI[2], NOOR FAIZAH FITRI MD YUSOF[2] AND AHMAD ZIA UL-SAUFIE[3]

[1]School of Ocean Engineering, Universiti Malaysia Terengganu,
21300 Kuala Nerus, Terengganu, Malaysia
[2]Clean Air Research Group, School of Civil Engineering, Universiti Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang Malaysia
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara,
13500 Permatang Pauh, Pulau Pinang, Malaysia

**Corresponding author email: nurul.adyani@umt.edu.my*

Abstract: Ground-level ozone (O_3) has a significant effect on human health when the concentration level exceeds Malaysia Ambient Air Quality Guidelines (MAAQG). This research focuses on daytime ground-level O_3 concentrations in Kemaman, Terengganu. The aim of this study is to compare the performance of the multiple linear regression model and the feed forward backpropagation neural network model for predicting daytime O_3 concentrations. This study used hourly average monitoring records from 2009 to 2012. Five performance indicators that are normalized absolute error (NAE), root mean squared error (RMSE), index of agreement (IA), prediction accuracy (PA) and coefficient of determination (R^2) were used to evaluate the models performances. The feed forward backpropagation neural network model shows better performances with smaller calculated errors (NAE = 0.1729, RMSE = 6.7906) and high accuracy (IA = 0.9427, PA = 0.8054, R^2 = 0.8022) than the multiple linear rregression. The performances of feed forward backpropagation neural network model can be used for O_3 concentration prediction in the future.

Keywords: Ground-level ozone, daytime, performance indicators

Introduction

Ground-level ozone (O_3) is a well-known secondary pollutant that is regulated under the National Ambient Air Quality Standards (NAAQS) and it is formed through photochemical reaction. Hydrocarbons and NO_x are the two main chemical precursors for O_3 formation. O_3 is one of major troposphere photochemical oxidants that is formed by a series of complex reactions between nitrogen oxides ($NO_x = NO + NO_2$) and volatile organic compounds (VOC_s) in the presence of sunlight and meteorological conditions (Ghazali *et al.*, 2010). Photochemical interactions between emitted pollutants (NO_x and VOCs) and favorable meteorological conditions (high temperatures and strong solar radiation) can lead to high O_3 concentrations (Sousa *et al.*, 2007).

High levels of O_3 concentrations cause damage to the plant species, various natural materials and manufactured goods, but can also lead to the damage of lung tissues in human (Wang and Georgopoulos, 2001), and increase respiratory symptoms such as chest pains and coughing (Soni and Shukla, 2012).

O_3 concentration prediction models are very difficult to develop because of the different interactions between pollutants and meteorological variables (Borrego *et al.*, 2003). However, statistical O_3 modelling to identify the relationship between primary pollutants, meteorological conditions and O_3 concentrations have been largely studied. Multiple linear regression (MLR) analysis is one of the most widely used for expressing the relationship on several independent (predictor) variables especially in environmental study (Ghazali *et al.*, 2010; Awang *et al.*, 2015). Recently, Neural Networks (NNs) have been developed in the prediction of O_3 concentrations (Banan *et al.*, 2014) and demonstrated greater efficiency to deal with the O_3 prediction compared to statistical linear method (Sousa *et al.*, 2007).

Therefore, the primary goal of this study was to build an accurate statistical model to predict daytime O_3 concentrations using linear and non-linear statistical models.

Materials and Methods

The hourly concentrations of air pollutants and hourly meteorological parameters were measured simultaneously at Kemaman stations and were are taken between 2009 to 2012 for this study from the continuous air quality monitoring stations by the Department of Environment (DoE) in Malaysia. Kemaman is a developing Malaysian town located ~~at~~ in an area where the industrial Kertih Petrochemical Industrial Area is in the North and the industrializing and urbanizing

Gebeng Industrial Area in the South. The monitoring station was located at Sekolah Rendah Bukit Kuang with coordinates (4°14'21.9"N 103°11'31.8"E). The location of the continuous air monitoring station for this study is shown in Figure 1.

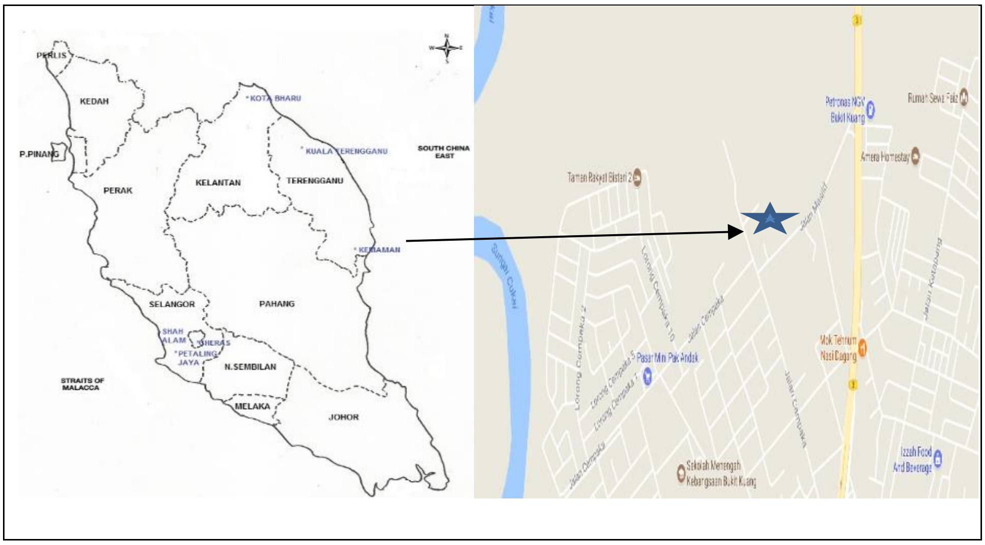

Figure 1: Location of the Air Monitoring Station

Four inputs parameters utilized to be predictors in this study are nitrogen dioxide (NO_2, Ppb), previous hour O_3 ($O_{3,t-1}$, Ppb), previous hour NO_2 ($NO_{2,t-1}$, Ppb) and ambient temperature (T, ^{0}C). Daytime is defined as the complete hour falling between sunrise and sunset and in Malaysia, daytime is defined from 7 AM to 7 PM (12 hour) (Awang *et al.*, 2015).

During the process of screening the air monitoring records, missing values were omitted from the air monitoring records to complete dataset. Randomization raw air monitoring records was applied in the development of the prediction model to reduce bias result.

The air monitoring records were partitioned into two data sets: the training dataset and validation dataset to avoid overfitting (Ul-Saufie *et al.*, 2015). This study randomly partitioned dataset into 80% for training and 20% for validation.

MLR or linear model uses a number of independent variables to predict the dependent variable. The estimated MLR model is as stated:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \qquad (1)$$

where;

$\hat{Y}$ is the estimated value of O_3 concentrations,

X_1, X_2, … , X_k are the independent variables which represent NO_2, $NO_{2,t-1}$, $O_{3,t-1}$, T,

b_1, b_2, … , b_k are the estimated partial regression coefficients,

Feed forward backpropagation neural network (FFBP-NN) or non-linear model is organized in three layers of neurons namely; input, hidden and output layers as shown in Figure 2. The input layer, located at the first layer of neurons, represents input variables. The input layer consists of four input variables namely NO_2, AT, $O_{3,t-1}$ and $NO_{2,t-1}$. The second layer is the hidden layer used to process the input weight from the input layer then transferred to the output layer. The third layer is the output layer which represents the O_3 concentrations.

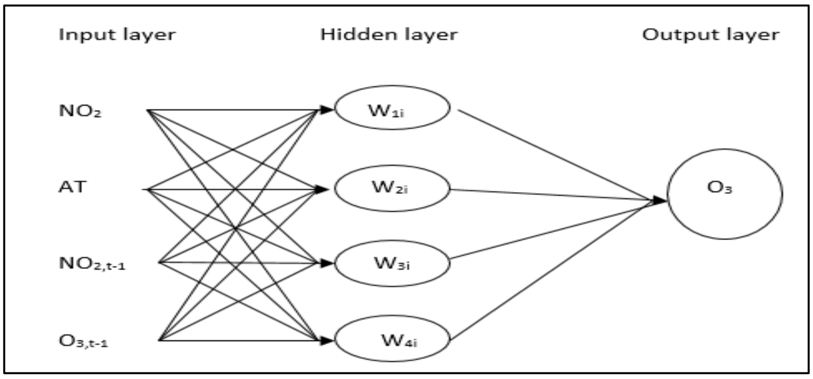

Figure 2: General Structure of One Hidden Layer FFBP-NN Architecture

Determination training algorithm and activation function are important tasks to develop FFBP-NN model. The selection of the activation function has a significant effect on the applicability training algorithm (Ul-Saufie *et al.*, 2015). This study used Lavernberg-Marquardt algorithm (LMA) to calculate weight and bias value because of local error minima and is recommended for developing a model (Ul-Saufie *et al.*, 2015). The activation function applied in this study is to sum up the ~~of~~ weight input for determining the output. This study used logistic sigmoid, linear and tan sigmoid transfer function. The equations for activation function are as follows:

$$\text{Logistic sigmoid}: \ f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

$$\text{Linear}: \ f(x) = x \qquad (3)$$

$$\text{Tangent sigmoid}: \ f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \qquad (4)$$

Performance indicators (PI) were used to show how close the predicted value is to the actual value in analyzing O_3 concentrations. PI have been used in this study and they consist of accuracy measure namely, R^2, PA, IA and error measures namely, RMSE, and NAE as shown in Table 1.

Table 1: Performance Indicators

PI	Equation	Description
NAE	$\text{NAE} = \dfrac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{\sum_{i=1}^{n} O_i}$	Close to 0 is good
RMSE	$\text{RMSE} = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n} (P_i - O_i)^2}$	Close to 0 is good
IA	$\text{IA} = 1 - \left[\dfrac{\sum_{i=1}^{n} (P - O_i)^2}{\sum_{i=1}^{n} \left(\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert \right)^2} \right]$	Close to 1 is good
PA	$\text{PA} = \dfrac{\sum_{i=1}^{n} (P_i - \overline{P})^2}{\sum_{i=1}^{n} (O_i - \overline{O})^2}$	Close to 1 is good
R^2	$R^2 = \left(\dfrac{\sum_{i=1}^{n} (P_i - \overline{P})(O_i - \overline{O})}{n.S_{pred.}S_{obs}} \right)^2$	Close to 1 is good

*n is the total number of hourly monitoring record, P_i is the predicted concentrations of ground–level O_3, O_i is the observed value of ground-level O_3 concentrations, $\overline{O}$ is the mean of the observed value of ground-level O_3 concentrations, $\overline{P}$ is the mean of the predicted value of ground-level O_3 concentrations, S_{pred} is the standard deviation of the predicted value of ground-level O_3 concentrations and S_{obs} is the standard deviation of the observed value of ground-level O_3 concentrations.

Results and Dicsussion

Descriptive statistics for this study was presented in Table 2. The maximum value for O_3 concentrations is 95 ppb indicating that it is below the O_3 Malaysia Ambient Air Quality Guidelines (MAAQG) value of 100 ppb (for 1-h averaging time).

Table 2: Descriptive Statistics

Parameters	Mean	Median	Standard Deviation	Maximum
O_3 (Ppb)	28.9	28.0	15.3	95
NO_2 (Ppb)	3.6	3.0	2.5	27.0
AT (^{0}C)	30.5	31.1	4.4	39.5

The problems in deciding FFBP-NN model network architecture are activation function and the number of neurons in the hidden layer. Table 3 shows that 7 is the best number of neurons in the hidden layer for daytime prediction of O_3 concentrations because of the smallest errors (NAE = 0.1729, RMSE = 6.7905) and the highest accuracy (IA = 0.9427, PA = 0.8053, R^2 = 0.8022).

Table 3: Validation Model using Different Number of Neurons

No. of hidden node	NAE	RMSE	IA	PA	R^2
3	0.1824	7.1029	0.9362	0.7825	0.7836
4	0.1825	7.2169	0.9339	0.7781	0.7766
5	0.1784	6.9953	0.9383	0.7853	0.7899
6	0.1752	6.8420	0.9416	0.7999	0.7992
7	**0.1729**	**6.7905**	**0.9427**	**0.8053**	**0.8022**
8	0.1729	6.8212	0.9418	0.7968	0.8004

After determining the best number of neurons in the hidden layer then the transfer function is obtained. Table 4 shows the best activation function for daytime prediction of O_3 concentrations from the input to the hidden layer is tangent sigmoid (tansig) and activation function from the hidden layer to the output layer is linear transfer function (purelin) based on the performance indicators and it gives the smallest errors (NAE = 0.1729, RMSE = 6.7906) and highest accuracy (IA = 0.9427, PA = 0.8054, R^2 = 0.8022).

Table 4: Results using Different Activation Functions

TF A	TF B	NAE	RMSE	IA	PA	R^2
Tansig	**Purelin**	**0.1729**	**6.7906**	**0.9427**	**0.8054**	**0.8022**
Tansig	Logsig	0.6473	22.3211	0.5028	0.0383	0.2629
Tansig	Tansig	0.1784	6.8892	0.9398	0.7713	0.7969
Logsig	Logsig	0.6471	22.3207	0.5019	0.0369	0.2626
Logsig	Tansig	0.1761	6.8579	0.9408	0.7838	0.7983
Logsig	Purelin	0.1741	6.8181	0.9422	0.8045	0.8006
Purelin	Purelin	0.2245	8.6687	0.8976	0.6792	0.6777
Purelin	Logsig	0.6518	22.3925	0.4887	0.0314	0.1736
Purelin	Tansig	0.2247	8.5772	0.8996	0.6757	0.6844

*TF : Transfer Function

Table 5 shows the comparison of performance indicators between MLR and FFBP-NN models. Table 6 shows the FFBP-NN model perfroms better than MLR model in predicting daytime O_3 concentrations with the smallest error (NAE = 0.1729, RMSE = 6.7906) and highest accuracy (IA = 0.9427, PA = 0.8054, R^2 = 0.8022). Based on the R^2 values 68% to 80% of the variation in O_3 concentration are explained by parameters contributed for prediction.

Table 5: Comparison of Performance Indicators for FFBP-NN and MLR Models

Models	NAE	RMSE	IA	PA	R^2
FFBP-NN	**0.1729**	**6.7906**	**0.9427**	**0.8054**	**0.8022**
MLR	0.2249	8.6678	0.8974	0.6771	0.6780

Conclusion

The FFBP-NN and MLR models approaches are proven to be effective techniques to predict daytime O_3 concentrations. The results show that FFBP-NN performs better than MLR with the smallest errors and highest accuracy that can predict daytime O_3 concentrations accurately. The results of this study can be used as a source of reference data in improving the existing guidelines towards a sustainable environment.

Acknowledgements

This study was funded by the Ministry of Higher Education Malaysia under the Research Acculturation Collaborative Effort (RACE) Grant Scheme Phase 2/2013. The author would like to thank the Ministry of Education Malaysia for its financial support to carry out this study under the MyBrain15 programme and the Department of Environment Malaysia for granting permission to utilize air quality data for this study.

References

Awang, N. R., Ramli, N. A., Yahaya, A. S., & Elbayoumi, M. (2015). Multivariate Methods to Predict Ground Level Ozone during Daytime, Nighttime, and Critical Conversion Time in Urban Areas. *Atmospheric Pollution Research, 6*(5), 726–734.

Banan, N., Latif, M.T., Juneng, L., & Khan, M.F. (2014). An Application of Artificial Neural Networks for the Prediction of Surface Ozone Concentrations in Malaysia. From Sources to Solution, Springer Science Business Media Singapore, A.Z. Aris et al. (eds), chapter *2*, 7-12.

Borrego, C., Tchepel, O., Costa, A.M., Amorim, J.H., & Miranda, A.I. (2003). Emission and Dispersion Modelling of Lisbon Air Quality at Local Scale. *Atmospheric Environment, 37*, 5197-5205.

Ghazali, N.A., Ramli, N.A., Yahaya, A.S., Yusof, N.F., Sansuddin, N. & Al-Madhoun, W.A. (2010). Transformation of Nitrogen Dioxide into Ozone and Prediction of Ozone Concentrations using Multiple Linear Regression Techniques. *Environmental Monitoring Assessment, 165*, 475-489.

Soni, A., & Shukla, S. (2012). Application of Neuro-fuzzy in Prediction of Air Pollution in Urban Areas. *IOSR Journal of Engineering, 2*(5), 1182-1187.

Sousa, S.I.V., Martins, F.G., Alvin-Ferraz, M.C.M., & Pereira, M.C. (2007). Multiple Linear Regression and Artificial Neural Networks Based on Principal Component to Predict Ozone Concentration. *Environtal Modelling Softwware, 22*(1), 97-103.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., & Hamid, H. A. (2015). PM_{10} Concentrations Short Term Prediction using Feedforward Backpropagation and General Regression Neural Network in a Sub-urban Area. *Journal of Environmental Science and Technology, 8*(2), 59–73.

Wang, S.W., Georgopoulus, P.G. (2001). Observational and Mechanistic Studies of Tropospheric Studies of Trophosperic Ozone/Precursors Relations: Photochemical Models Performance Evaluation with Case Study. *Technical Report ORC-TR99-03.*