MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Development of missing data prediction model for carbon monoxide

.X 0RKG .DONDXVDU[a] 6LWL 1RRU 6\XKDGD 0XKDPPDG[b] ,VPDLO $]PDQ $]LG[c]

[a] Faculty Bioresources and Food Industry, Universiti Sultan Zainal Abidin (UniSZA), Besut Campus, 22200 Besut, Terengganu, Malaysia
[b] International Institute for Halal Research and Training, International Islamic University Malaysia, Selangor, Malaysia
[c] Faculty of Health Sciences, Universiti Sultan Zainal Abidin (UniSZA), Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia

Corresponding author: azmanazid@unisza.edu.my

Article history
5HFHLYHG -DQXDU\
Revised 22 March 2018
FF 20 May 2018
Published Online 4 February 2018

Abstract

&DUERQ PRQR[LGH LV RQH RI WKH PRVW LPSRUWDQW SROOXWDQWV LQ HWLVVHOHFWHG IRU $,FDOFXODWLRQ 7KHUHIRUH LW LV SDUDPRXQW WR HQVXUH WKDW WKH HUURU LQ PLVVLQJ GDWD RI &2 GXULQJ WKH DQDO\VLV 7KHUH DUH QXPEHUV RI RFFXUUHQFH WKDW PD\ FRQWULEXWH WR WKH PLVVLQJ GDWD SUREOHPV VXFK DV LQDELOLW\ RI WKH LQVWUXPHQW WR UHFRUG FHUWDLQ SDUDPHWHUV ,Q OLHX RI WKLV IDFW D &2 SUHGLFWLRQ PRGHO QHHGV WR EH GHYHORSHG WR DGGUHVV WKLV SUREOHP GDWD VHW RI PHWHRURORJLFDO DQG DLU SROOXWDQWV YDOXH V REWDLQHG IURP WKH $LU 4XDOLW\ 'LYLVLRQ 'HSDUWPHQW RI (QYLURQPHQW 0DOD\VLD 2 WRWDO RI GDWD VHWV ZHUH XVHG WR GHYHORS WKH PRGHO XVLQJ VWDWLVWLFDO DQDO\VLV 6$ 6KRUWV $UWLILFLDO 1HXUDO 1HWZRUN $11 6$ VKRZ SDUWLFXODWH PDWWHU 30 DQ R]RQH ZHUH WKH PRVW VLJQLILFDQW LQSXW YDULDEOH IRU PLVVLQJ GDWD SUHGLFWLRQ PRGHO RI &2 7KH WKLUG DQG IRXUWK WR EH WKH RSWLPXP QXPEHU WR GHYHORS WKH $11 PRGHO ZLWK WKH YDOXH RI 5 HTXDO WR %RWK PRGHOV DUWLILFLDO QHXUDO QHWZRUN WR FDUERQ PRQR[LGH DOO SDUDPHWHUV $11& DQ DUWLILFLDO QHXUDO QHWZRUN WR FDUERQ PRQR[LGH ZLWKRXW $11& 2 VKRZ KLJK YDOXH RI 5 DQ DQ ORZ YDOXH RI 506(DQ UHVSHFWLYHO\ 7KH UHVXOW LQGLFDWHG WKDW WKH PRGHOV PLJKW QRW FRPSOHWH WKH PRVW VLJQLILFDQW LQSXW YDULDEOH WR UHSUHVHQW WKH SDUDPHWHUV WKDQ XVLQJ DOO LQSXW YDULDEOH

Keywords: 3UHGLFWLRQ PRGHO FDUERQ PRQR[LGH DUWLILFLDO QHXUDO QHWZRUN PLVVLQJ GDWD PRGHO

3HQHUELW 870 3UHVV $OO ULJKWV UHVHUYHG

INTRODUCTION

Air pollution imposes severe environmental challenges as well as prominent health risks to human (Najafpoor *et al*., 2014). The worst air pollution problem usually takes place at urban areas of both developed and developing countries (Hassanzadeh *et al*., 2009) which indirectly affects the quality of life as well as public health. Air pollutions are mostly produced by natural activities such as volcano eruptions and human activities, owing to the industrial processes, production of energy from power plants, residential heating and open burning (Najafpoor *et al*., 2014; Afroz *et al*., 2003). In addition, the fuel burning vehicles in urban area have worsened the air quality (Wang & Lu, 2006). Consequently, the human health and environment may be affected by the air pollution and in the long term, air pollution has a tendency to intensify the threats to earth.

Incomplete combustion of hydrocarbons contributes to the presence of carbon monoxide (CO) (Levy, 2015), which associated to cardiovascular diseases, daily mortality and morbidity (Chen *et al*., 2011). In a homogeneous environment, air pollutant levels including CO at each fixed monitoring stations are indicated by the applied measurements. Nevertheless, the dispersion of the actual pollutant contents is still unknown, owing to substantial influences of the prevailing conditions of dispersion, emission sources distribution and the region topography (Zoroufchi & Fatehifar, 2015) and thus, affecting the exact concentrations of the measured air pollutants.

Besides PM_{10}, the development of CO missing data prediction model is important due to the fact that this pollutant is one of the leading contributors to the air pollutants and produced from most of the sites (Azid *et al*., 2016 & Mohamad *et al*., 2015). According to Awang *et al*., (2015) PM_{10} and CO are being grouped into the same component, indicating that CO is as important as PM_{10} and may come from the same sources since it is being grouped together with PM_{10}.

Missing data is a common occurrence in air pollution studies. Failure of equipment and anomalous measurement are the reasons for this missing data (Chen *et al*., 2016). Researchers are opted to remove the missing data prior to statistical analysis. However, this action can reduce the data size. The data becomes unrepresentative with the removal of massive missing data and subsequently, unreliable result will be produced. Hence, imputing the missing values can be the alternative way to evade unreliable or ravage result on the statistical interpretation.

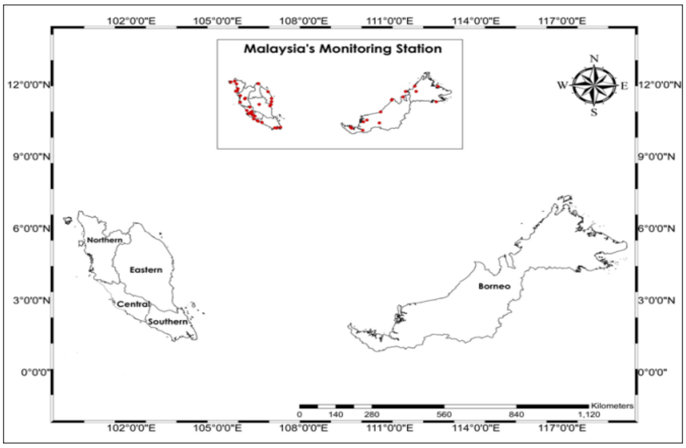
In general, single and multiple imputation methods are the acceptable approaches to generate complete information matrices (Little & Rubin, 1987). One value for each missing one is specifically filled in the former method. For the multiple-imputation, simulated values for each missing data are generated to appropriately determine the uncertainty of the missing data (Junninen *et al*., 2004) using expectation maximization based (EMB) algorithm. Other algorithms are also applicable; however, EMB algorithm possesses unique features such as simplicity, power, rapidity (Honaker et al., 2011) and practicality in forecasting the missing air quality data using ANN model.

ANN is a non-linear model (Kukkonen *et al*., 2003) and has been acknowledged as a cost-effective model (Azid *et al*., 2014) since it is able to discover and ascertain patterns (Zare, 2014), solve complex functions and produce reliable air pollutants prediction. This model can be utilized to assist forthcoming planning with the presence of missing data during the air monitoring, mend air quality management system (Kumar & Goyal, 2011). Besides air quality, ANN model has also been applied for the water quality studies. (Zali *et al*., 2011) used ANN for the water quality index (WQI) prediction for Kinta River, Malaysia. Besides that, (Nasir *et al*., 2011) also applied ANN in their study for WQI prediction model in Juru River, Malaysia.

EXPERIMENTAL

Study area

There are 52 continuous monitoring stations (N06˚ 25.424' E100˚ 20.880' to N04˚ 15.016' E117˚ 56.166') for ambient air quality throughout Malaysia (Fig. 1). These stations were chosen due to the desired locations in urban, suburban, and industrial area (Azid *et al*., 2015).



Source: (Kanniah et al., 2016; DOE, 2004)

Fig. 1RQWLQXRXVDPELHQWDLUTXDOLWPRQLWRULQJVWDWLRQVLQ0DODVLD

Data collection

The air pollutants and meteorological data were obtained from the Air Quality Division, Department of Environment (DOE) Malaysia. The obtained data in this study was entailed daily average observations for most of air pollutants (PM_{10}, NO_2, CO, O_3, NO, NO_x, THC, CH_4, NmHC, SO_2) and meteorological conditions (wind direction, wind speed, temperature, humidity, UVB) from 2010 until 2015. These variables were independently analysed at each monitoring stations. The model equipments for continuous~~ly~~ monitoring program (CAQM) on each atmospheric pollutants and meteorological parameters were listed in Table 1.

Variables selection

The best input variables for the CO missing data modelling were selected prior to the designing of the model. Input nodes in ANN were based on the selected input variables, which significantly contributing to the process of forecasting. In addition, high numbers of input could cause reduction of training speed, over-fitting, redundancy and noise variables (Ababneh *et al*., 2014). Thus, only selected input variables were used for ANN analysis.

One of the methods to select the best input variables for the prediction of CO missing data modelling using ANN model was by applying sensitivity analysis (SA). This method used "leave-one-out" technique to rank the importances of the model input variables by considering their influences on the unpredictability of the model output (Manache & Melching, 2008). This method was carried out manually whereby one-by-one parameters were removed. The R^2 values of each leave one out parameters were applied to show the differences of R^2. The percentage (%) contribution of the parameters was determined by applying Equation 1:

$$\%\,\boldsymbol{contribution} = \left(\frac{b_i - a_i}{z_i}\right)\mathbf{100} \qquad (1)$$

where:
ai = the value of R^2 after a leave-one-out parameter for each model
bi = the reference value of R^2 from ANN-CO-AP
zi = the sum value of difference R^2

SA has been used in various fields of environmental studies. For instance, it has been used by (Latif *et a*l., 2014) in their study to investigate the important level of each inputs (CO, NO, NO_2, NO_x, O_3, PM_{10}, SO_2, THC, CH_4 and NmHC) on the output (vehicles). Meanwhile, (Asadollahfardi et al., 2016) applied SA to determine which of the input variables have a great role in predicting ground layer of O_3 where in their study, the maximum and minimum roles for ground-level O_3 concentration prediction are $PM_{2.5}$ and benzene, respectively. On top of that, research done by (Rahimi, 2017) also applied SA in prediction of NO_2 and NO_X by calculating important level of input variables in prediction.

Table 17KH0PRGHOHTXLSPHQWIRUHDFKSDUDPHWHU

Parameter	Model equipment
3DUWLFXODWH0DWWHU30	%0%HWD$WWHQDWLRQ
:LQG6SHHG:6NPKU	0HW2Q
:LQLUHFWLRQÛ	0HW2Q
$LU7HPSHUDWXUHÛ&	0HW2Q
5HODWLYHXPLGLW5+	0HW2Q
2[LGHVRI1LWURJHQ2[SSP	7HOHGQH0RGHO$
1LWURJHQRQR[LGH2SSP	7HOHGQH0RGHO$
8OWUDYLROHWE8 KU	
0HWKDQH SSP	7HOHGQH0
1RQ0HWKDQHURFDUERQ 1P&SP	7HOHGQH0
7RWDO+GURFDUERQ7&SSP	7HOHGQH0
6XOSKXUGLR[LGH2 SSP	7HOHGQH0RGHO$
1LWURJHQGLR[LGH2 SSP	7HOHGQH0RGHO$
2]RQH SSP	7HOHGQH0RGHO(
&DUERQPRQR[LGH&2SSP	7HOHGQH0RGHO(

Data pre-processing

Analysis, filtration and transformation were implemented to organize the obtained data set for the model development. There were ~~a~~ few missing data and outliers observed which might be resulted due to technical failure (Zakaria & Noor, 2018) and incorrect recorded results during the data collections. The outliers should not be deleted because they might give true measurements (Burke, 1999). Moreover, in air pollution modelling, it was important to use~~d~~ whole year data by considering the full coverage variation of the seasonal pollutant levels and meteorological parameters (Arhami *et al*., 2013).

In this study, the missing data was replaced by using the EMB (expectation-maximization with bootstrapping) algorithm. Likelihood observed data was represented as shown in Equation 2:

$$\boldsymbol{p}(\boldsymbol{D}^{obs}, \boldsymbol{M}|\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{M}|\boldsymbol{D}^{obs})\boldsymbol{p}(\boldsymbol{D}^{obs}|\boldsymbol{\theta}) \qquad (2)$$

With D^{obs} and M were known as observed data and missingness matrix, respectively. While, likelihood was written as Equation 3 if only complete data parameters were concerned:

$$\boldsymbol{L} = (\boldsymbol{\theta}|\boldsymbol{D}^{obs}) \propto \boldsymbol{p}(\boldsymbol{D}^{obs}|\boldsymbol{\theta}) \qquad (3)$$

While, equation could be rewrite as Equation 4 based on the iterated expectations law:

$$p\left(D^{obs}\middle|\theta\right) = \int p\left(D|\theta\right) dD^{mis} \qquad (4)$$

The posterior with this likelihood and a flat prior on θ as shown in Equation 5:

$$p\left(\theta\middle|D^{obs}\right) \propto p\left(D^{obs}\middle|\theta\right) = \int p(D|\theta)\, dD^{mis} \qquad (5)$$

Table 2 illustrates the statistical measurements for selected meteorological and air pollutant variables (after SA) for the modelling of CO missing data using ANN.

Table 2HFULSWLMQRUPDWLRQRUVHFWVDULDEOHIRUPRGHOLQI &PLVVLQDWD

Variables	Min	Max	Mean
2 SSP			
30 µg/m³			

The best input variables were used in prediction, evaluation and validation processes of the ANN technique. However, prior to these processes, the data variables were normalized by employing scaling range of (-1,1) to increase training speed, minimize variable values differences and reduce computational problems (Srinivasan et al., 1994). The hyperbolic tangent function was used to transform value to be between -1 and 1, for normalization (JMP, 2012) with the formula of the hyperbolic tangent function as shown in equation 6:

$$\boldsymbol{Hyperbolic\ tangent\ function} = \frac{e^{2x}-1}{e^{2x}+1} \qquad (6)$$

In addition, data normalization within range output of activation function of ANN output layer was necessary when using non-linear activation function (Zhang *et al*., 1998). Studies revealed that only small errors of prediction were produced when applying the hyperbolic tangent function as compared to the sigmoid (logistic) transfer function (Chaloulakou *et al*., 2003).

It was crucial to select the optimum number of nodes in hidden layer since high number of nodes might result in overfitting (He et al., 2014) while less number of nodes might not adequately capture the information. The proposed equation was to determine the appropriate number of nodes ranges (Fletcher & Goss, 1993) as shown Equation 7:

$$Number\ of\ nodes\ ranges = 2S^{\frac{1}{2}} - O\ \ to\ \ 2S - 1 \qquad (7)$$

where, S is the number of input nodes and O is the number of output nodes.

Prediction, evaluation and validation of ANN model

The system predicted missing data of CO with the input selection used in ANN was based on the results of SA. To determine the error in predicting the concentration of CO and models performance evaluation, a Root Mean Squared Error (RMSE) which indicated in Equation 8 was applied:

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(O_t - P_t)^2} \qquad (8)$$

Meanwhile, model validity was illustrated by applying coefficient of determination (R^2) between observations and predicted as shown in Equation 9:

$$\mathbf{R^2 = 1} - \left(\sum_{t=1}^{n}\frac{(O_t-P_t)^2}{(O_t-\bar{o})^2}\right) \qquad (9)$$

where, O_t, P_t and ō represent observed value, predicted value and observed mean value of CO concentrations at time, t respectively, a n is the number of data (Ahmat *et al*., 2015).

RESULTS AND DISCUSSION

Table 3 shows the SA results for the prediction of CO. From the result, different values R^2 exhibited different parameter values that affected the prediction of CO. The % contribution > 10% indicated strong contribution towards the CO presence (Azid *et al*., 2016). The highest and lowest influences of input variable were PM_{10} and WD with % contribution of 35.95% and 0.92%, respectively.

Based on the R^2 and % contribution in Table 3, the most important input variables were ranked as $PM_{10} > O_3 > NO_2 > SO_2 > NO_x > CH_4 >$ Temp > NO > Humidity > THC > WS > UVB > NmHC > WD with PM_{10} and O_3 became the main contributors to the CO presence in this study.

Studies on SA shown that air pollutants rendered a strong influence on the climate daily variability (Ababneh *et al*., 2014). This finding was supported by SA study on the air pollution index (API) which was found that PM_{10} and CO were the main contributors to the air pollutants (Azid *et al*., 2016). In the Southeast Asia including Malaysia, these pollutants have been acknowledged as significant atmospheric pollutants in major cities and were produced by complete combustion of motor vehicles as well as various industrial practices (Latif et al., 2011; Mustafa et al., 2012). Beside of PM_{10}, O_3 was also correlated to CO since the letter was originated from diesel fuel (Rani *et al*., 2017) and mobile sources, causing it to become secondary contributor to ozone depletion (Kumar *et al*., 2017).

Table 36QLWLYWDQOVLVIRUVHFWLQXWDULDEOHIRU& SUHLFWHPRGH

Model	R^2	Difference, R^2	% Contribution
$12&$	0.7639		
$1/30	0.6618	0.1021	35.95
$1/2	0.6784	0.0855	30.11
$1/12			
$1/62			
$1/12 ;			
$1/&			
$1/7(03			
$1/12			
$15+			
$1/7&			
$1/:6			
$1/89%			
$1/10&			
$1/:'			
7RWDO		0.2839	100.00

Table 4 shows the R^2 and RMSE values for ANN with a hidden layer and different number of hidden nodes. Based on two main input variables obtained from SA, the most appropriate number of hidden nodes was between 1 to 3. Among these hidden nodes, three hidden nodes exhibited the highest R^2 and the lowest RMSE. Higher value of R^2 indicated a closer relation between predicted output value and the exact output value. Moreover, the three hidden nodes have lower RMSE value compared to other hidden nodes where the nearest RMSE value to 0 was indicated to the best ANN model performance (Ababneh *et al*., 2014). Thus, three hidden nodes were considered as the optimum number for the ANN model.

Post ANN-model prediction step, model evaluation and validation were carried out to evaluate the ability of ANN in the prediction process. In Table 5, ANN-CO-AP (14 variables) and ANN-CO-LO (using PM_{10} and O_3) prediction models showed RMSE (0.2482 and 0.3506) and R^2 (0.7639 and 0.5311) respectively. The nearest RMSE value to 0 and R2 to 1 were indicated to the best ANN model performance and model robustness (Zali *et al*., 2011), as well as provided the highest accuracy between predicted and actual output values (Ababneh et al., 2014). ANN-CO-LO was considered as the optimum model as it used fewer input variables (PM_{10} and O_3) although the model exhibited lower R^2 (0.5311) and higher RMSE (0.3506). Moreover, more input data would lead to better predictions (Esfandani & Nematzadeh, 2016). Thus, by reducing the input variables, R^2 value would become lower.

Table 4 7KH HIIHFW RI GHWHUPLQDWLRQ DQG HUURU VOXH 506(RU WKH $11 ZLWK KLGGHQ DQG GLIIHUHQW QXPEHU RI KLGGHQ QRGHV

No. of hidden nodes	R^2	RMSE

Table 5 0RGHO HYDOXDWLRQ DQG PRGHO YDOLGDWLRQ IRU WKH GLIIHUHQW 2 SUHGLFWLRQ PRGHOV

Models	Model evaluation	Model validation
	RMSE	R^2

[1]All input variables were employed in ANN model.
[2]PM_{10} and O_3 input variables were employed in ANN model.

From the developed model, the ANN-CO-LO model equation could be interpreted as in Equation 10:

Predicted CO= -0.23 + (-1.85 x H1) + (-0.43 x H2)+ (9.19 x H3) (10)

CONCLUSION

In conclusion, only selected input variables were used to generate a model for missing CO data. By using SA, the number of input variables could be reduced based on their significant values to the presence of CO. In this study, the most significant input variables were PM_{10} and O_3. In addition to the selection of input variables, the selection number of hidden nodes was also vital to ensure the obtained results were not over fitting and only captured sufficient information. The optimum number of hidden nodes for this study was three, which exhibited ~~had~~ the highest R^2 (0.5311) and the lowest RMSE (0.3506) values as compared to other hidden nodes. Research done by Esfandani & Nematzadeh (2016) showed that greater input number would give better prediction. Thus, ~~by~~ decreasing the input number by using less number of input variables, ANN-CO-LO model (used input variables PM_{10} and O_3) gave lower R^2 value compared to ANN-CO-AP which used all input variables (14 parameters). Although ANN-CO-LO model gave lower R^2 value compared to ANN-CO-AP R^2 value, ANN-CO-LO model still possessed better performance with good R^2 value (R^2 = 0.5311). On top of that, the reductions of air quality parameters was much applicable for air resource management because of its time and cost of operation. Hence, the ANN-CO-LO model could be utilized in predicting missing CO data.

ACKNOWLEDGEMENT

This manuscript has been funded through University Research Fund (Grant no.: UniSZA/2016/DPU/01) and Research Initiative Grant (RIGS16-363-0527) of International Islamic University Malaysia, Gombak, Selangor, Malaysia.

REFERENCES

Ababneh, M. F., Al-Manaseer, A. O., Btoush, M. H. 2014. PM_{10} forecasting using soft computing techniques research. *Journal of Applied Sciences, Engineering and Technology,* 7(16): 3253-3265.

Afroz, R., Hassan, M. N., Ibrahim., N. A. 2003. Review of air pollution and health impacts in Malaysia. *Environmental Research,* 92(2): 71-77.

Ahmat, H., Yahaya, A. S., Ramli, N. A, 2016. The Malaysia PM_{10} analysis using extreme value. *Journal of Engineering Science and Technology,*10(12): 1560 – 1574.

Arhami, M., Kamali, N., Rajabi, M. M. 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research,* 20: 4777–4789.

Asadollahfardi, G., Tayebi, J. M., Mehdinejad, M., Rajabipour, M. J. 2016. Short-term prediction of atmospheric concentrations of ground-level ozone in Karaj using artificial neural network. *Pollution*, 2(4): 475-488.

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,* 6:726-734.

Azid, A., Juahir, H., Ezani, E., Toriman, M. E., Endut, A., Rahman, M. N. A., Yunus, K., Kamarudin, M. K. A., Hasnam, C. N. C., Saudi, A. S. M., Umar, R. 2015. Identification source of variation on regional impact of air quality pattern using chemometrics *Aerosol and Air Quality Research,* 15: 1545–1558.

Azid, A., Juahir, H., Toriman, M. K., Endut, A., Rahman, M. N. A., Kamarudin, M. K. A., Latif, M. T., Saudi, A. S. M., Hasnam, C. N. C., Yunus, K. 2016. Selection of the most significant variables of air pollutants using sensitivity analysis. *Journal of Testing and Evaluation*, 44(1): 376-384.

Azid, A., Juahir, H., Toriman, M., Kamarudin, M., Saudi, A., Hasnam, C. 2014. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. *Water, Air, & Soil Pollution,* 225(8): 1-14.

Burke, S. 1999. Missing values, outliers, robust statistics and non-parametric methods. *LC*GC Europe Online Suplement*, 19-24.

Chaloulakou, A., Grivas, G., Spyrellis, N. 2003. Neural network and multiple regression models for PM_{10} prediction in athens: A comparative assessment, *Journal of the Air & Waste Management Association,* 53(10): 1183-1190.

Chen, R.Pan, G., Zhang, Y., Xu, Q., Zeng, G., Xu, X. 2011. Ambient carbon monoxide and daily mortality in three Chinese cities: the China air pollution and health effects study (CAPES). *Science of the Total Environment*, 409(23): 4923-4928.

Chen, W., Tang, H., Zhao, H. 2016. Urban air quality evaluations under two versions of the national ambient air quality standards of China *Atmospheric Pollution Research* 7: 49-57.

DOE, Department Of Environment Malaysia. 2004. Malaysian Environmental Quality Report.

Esfandani, M. A., Nematzadeh, H. 2016. Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network. *Journal of AI and Data Mining,* 4(1): 49-54.

Fletcher, D., Goss, E. 1993.Forecasting with neural networks: An application using bankruptcy data. *Information and Management*, 24:159-167.

Hassanzadeh, S., Hosseinibalam, F. Alizadeh, 2009. R. Statistical models and time series forecasting of sulfur dioxide: A case study Tehran. *Environmental Monitoring and Assessment*, 155(1): 149-155.

He, H., Lu, W. Z., Xue, Y. 2014.Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm. *Building and Environment*, 78: 111-117.

Honaker, J., King, G., Blackwell, M. 2011. Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7): 1-47.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M. 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38: 2895–2907.

Kanniah, K. D., Kaskaoutis, D. G., Lim, H. S., Latif, M. T., Zaman, N. A. F. K., Liew, J. 2016. Overview of atmospheric aerosol studies in Malaysia: Known and unknown. *Atmospheric Research*, 182: 302–318.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G. 2003. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in centralHel sinki. *Atmospheric Environment* 37: 4539–4550.

Kumar, A., Goyal, P. 2011. Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, 409(24): 5517-5523.

Kumar, N., Middey, A., Rao, P. S. 2017. Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, *India Urban Climate*, 20: 148-167.

Latif, M. T., Azmi, S. Z., Noor, A. D. M., Ismail, A. S., Johny, Z., Idrus, S., Mohamad, A. F., And Mokhtar, M., 2011. The impact of urban growth on regional air quality surrounding the Langat River Basin, Malaysia, *Environmentalist*, 31(3): 315–324.

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., Nadzir, M. S. M. 2014. Long term assessment of air quality from a background station on the Malaysian Peninsula. *Science of the Total Environment*, 482–483: 336–348.

Levy, R. J. 2015. Carbon monoxide pollution and neurodevelopment: A public health concern. Neurotoxicology and Teratology, 49: 31-40.

Little, R. J. A., Rubin, D. B. 1987. *Statistical Analysis with Missing Data*, New York: Wiley.

Manache, G., Melching, C. S., 2008. Identification of reliable regression- and correlation-based sensitivity measures for importance ranking of water quality model parameters. *Environmental Modelling & Software*, 23(5): 549–562.

Mohamad, N. D., Ash’aari, Z. H., Othman, M., 2015. Preliminary assessment of air pollutant sources identification at selected monitoring stations in Klang Valley, Malaysia. *Procedia Environmental Sciences*, 30: 121 – 126.

Mustafa, M., Syed Abdul Kader, S. Z., Sufian, A., 2012. Coping with climate change through air pollution control: Some legal initiatives from Malaysia. *2012 International Conference on Environment, Energy and Biotechnology, Kuala Lumpur, Malaysia*. May 5–6, 33,101–105.

Najafpoor, A., Hosseinzadeh, A., Allahyari, S., Javid, A., Esmaily, H. 2014. Modeling of CO and NO_X produced by vehicles in Mashhad, 2012. *Environmental Health Engineering And Management Journal*, 1(1): 45-49.

Nasir, M. F. M., Juahir, H., Roslan, N., Mohd, I., Shafie, N. A., Ramli, N. 2011. Artificial neural networks combined with sensitivity analysis as a prediction model for water quality index in Juru River, Malaysia. *International Journal of Environmental Protection*, 1(3): 1-8.

Rahimi, A. 2017. Short-term prediction of NO_2 and NO_x concentrations using multilayer perceptron neural network: A case study of Tabriz, Iran. *Ecological Processes*, 6:4.

Rani, N. L. A., Azid, A., Khalit, S. I., Gasim, M. B., Juahir, H, 2017. Selected Malaysia air quality pollutants assessment using chemometrics techniques. *Journal of Fundamental and Applied Sciences*, 9(2): 335-351.

Srinivasan, D., Liew, A.C., Chang, C. S. 1994. A neural network short-term load forecaster. *Electric Power Systems Research*, 28: 227-234.

Wang, X. K., Lu, W. Z. 2006. Seasonal variation of air pollution index: Hong Kong case study. *Chemosphere* 63(8): 1261-1272.

Zakaria, N. A., Noor, N. M. 2018.Imputation methods for filling missing data in urban air pollution data for Malaysia. *Urbanism Arhitectură. Construcţii,* 9(2): 159-166.

Zali, M. A., Retnam, A., Juahir, H., Zain, S. M., Kasim, M. F., Abdullah, B., Saadudin, S. B. 2011. Sensitivity analysis for water quality index (WQI) prediction for Kinta River. *Malaysia World Applied Sciences Journal*, 14 (Exploring Pathways to Sustainable Living in Malaysia: Solving the Current Environmental Issues): 60-65.

Zare, A. H. 2014. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, 12: 40.

Zhang, G., Patuwo, E. B., Hu, M. Y. 1998. Forecasting with artificial neural networks: The state of the Art. *International Journal of Forecasting*, 14: 35-62.

Zoroufchi, B. K, Fatehifar, E. 2015. Optimal design of air quality monitoring network around an oil refinery plant: A holistic approach. *International Journal of Environmental Science and Technology*, 12(4): 1331-1342.