Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

PREDICTION MODEL OF MISSING DATA: A CASE STUDY OF PM_{10} ACROSS MALAYSIA REGION

N. L. Abd Rani[1], A. Azid[1,2,*], S. I. Khalit[1,2] and H. Juahir[3]

[1]Faculty Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia

[2]UniSZA Science and Medicine Foundation centre, Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu

[3]East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia

Published online: 15 January 2018

ABSTRACT

PM_{10} is one of the major concerns that have high potential for harmful effects on human health. Thus, prediction of PM_{10} was performed with the objectives to model suitable PM_{10} prediction formula to predict the concentration of PM_{10}. Imputation methods of EMB-algorithm and nearest neighbor were applied to treat missing data before analyzed by Fit model, MLR and ANN. R^2 obtained for Fit-model, MLR and ANN using imputation method of EMB-algorithm and nearest neighbor are (0.9975, 0.3858), (0.9623, 0.3857) and (0.9975, 0.4025) respectively. Sensitivity analysis (SA) shows humidity, temperature, CO, UVB and O_3 out of fifteen parameters contribute the most to the present of PM_{10} concentration. In conclusion, formula for the best PM_{10} prediction can be modeled by using ANN or Fit model together with the imputation method of EMB-algorithm.

Keywords: PM_{10} prediction; fit-model; MLR, ANN; imputation method.

Author Correspondence, e-mail: azmanazid@unisza.edu.my

doi: http://dx.doi.org/10.4314/jfas.v10i1s.13

1. INTRODUCTION

1.1. Background

Air pollution give a severe risk to the environment as well as to the human, especially for the elderly and children as it can causes serious respiratory and skin diseases as well as increases respiratory and heart illnesses. According to [1], human might experience health problems such as respiratory system, asthma and deaths due to the PM and it has been reported that exposed to that particles concentrations give PM detrimental effects on human health.

Besides, industrialization development and increased cities number also contributed to the air pollution problems [1]. Growth of population, traffic density, rapid urbanization and industrialization become worrying to the authority according to their contribution to the atmospheric pollution [2]. Main pollutants monitored in Malaysia's air are ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2) and solid such as PM_{10} (particulate matter with an aerodynamic diameter less than 10 μm) which are considered has significant impacts on both human and the environment.

Studies done by [3-4] found the significant relation between health effects and increased particulate air pollution concentrations. Air pollution can be controlled and reduce by estimate the pollutants density and to define the air quality state in comparison with the standard conditions [5]. Air pollutants concentration levels guidelines and limitations have been set by many environment agencies such as the European Union (EU), the U. S. Environmental Protection Agency (USEPA) and the World Health Organization (WHO) [6]. Air quality forecasting systems are necessary to provide good policies development and also give warning when the air pollutants surpass maximum limit values [7].

Particular guidelines in air pollution management should be implemented and be part of the air pollution management policy due to the accelerated growth of air pollutants emission sources in residential megacities [8]. According to [9], total of Premature deaths attributable to exposure of $PM_{2.5}$, NO_2 and O_3 in 41 European countries in 2013 are 467 000, 71 000 and 17 000 respectively.

Risk of human health especially with special health conditions such as asthma patients can be reduce by providing useful information to the public through the early and precise prediction of air pollutants that have significant impacts on both humans and the environment. Air quality forecasted for particular times and locations especially as well as air pollutants that exceed the permitted values [8].

Air pollutants and weather conditions are related to each other. Thus, processes become more complex in air quality modeling or prediction due to the air pollutants concentration levels influenced by the daily climate inconsistency. For instance, high wind speed and varied wind direction indications to increase the particulate matter concentrations as well as reducing visibility [10]. According to [11], higher PM concentration level due to the low relative humidity. Thus, rainfall reducing PM_{10} concentration levels along with cleaning the atmosphere. Besides, study done by [12] related the temperature with air pollutants concentrations where air pollutants concentrations become increase due to the air temperature variations between daytime and night time which contribute in radiological implications.

1.2. Missing Data

Multiple imputations can be applied on missing at random (MAR) data set as it is the greater method to do so [13]. The data analysis becomes problematic due to the missing observations. The missing data usually occurs due to loss of efficiency and complication in handling and analyzing data [13]. In air pollution studies, the missing data might happened due to the malfunctioned of equipment or errors in measurements [14]. Various techniques of imputation method were applied in other studies to impute missing data. One of them is mean top bottom technique to replace missing data of PM_{10} concentrations that had been applied by [15] and [16]. Imputation method of nearest neighbor was applied by [17] to complete the concentration of PM_{10} data. According to [18], nearest neighbor, linear interpolation and multilayer perceptron methods are advanced model-based imputation methods. These methods usually are more accurate when the missing data amount is small. However, long missing data gaps can cause loss of their advantage [19]. Accountability of covering the uncertainty adjacent the real data in multiple imputation make it becomes favorable method as bias between observed and unobserved data can be reduced [13].

1.3. Expectation Maximization Based Algorithms (EMB-Algorithm)

Multiple imputations which a common purpose approach to missing values of data can be performed by using Amelia II, whereby generates multiple incomplete data set versions. Complete observations of the analyses can be appropriately using all the missingness of information present as this method creates multiple incomplete data set versions. Compared to listwise deletion, it has been shown to decrease bias and increase efficiency. Imputation methods of ad hoc such as mean imputation can cause serious biases in variance and covariance. Due to the algorithms technical nature, multiple imputations can be a burdensome process. However, Amelia II offers user with a simple way to create and implement an imputation method, produce imputed datasets, and check it using diagnostics.

Novel bootstrapping approach of the EMB (expectation-maximization with bootstrapping) algorithm was used in Amelia II missing values' imputations. Values of the complete data parameters can be drawn from the algorithm that uses the familiar EM (expectation-maximization) algorithm on multiple bootstrapped samples of the original incomplete data. More variable with more observations can be imputed by Amelia II through the bootstrap based EMB algorithm in much less time [20]. EMB algorithm's simplicity and power practically never crashes, makes it unique among present multiple imputation software. Besides, it is also much faster than the alternatives. Parameters with complete data alarmed in multiple imputations, $\theta = (\mu, \Sigma)$. D^{obs} and *M* known as observed data and missingness matrix respectively. Thus likelihood of observed data is:

$$p(D^{obs}, M|\theta) = p(M|D^{obs})p(D^{obs}|\theta) \quad (1)$$

As only complete data parameters were concerned, the likelihood is written as:

$$L = (\theta|D^{obs}) \propto p(D^{obs}|\theta) \quad (2)$$

Based on the law of iterated expectations, the equation can be rewrite as:

$$p(D^{obs}|\theta) = \int p\,(D|\theta)dD^{mis} \quad (3)$$

With this likelihood and a flat prior on θ, the posterior shown as below:

$$p(\theta|D^{obs}) \propto p(D^{obs}|\theta) = \int p(D|\theta)\,dD^{mis} \quad (4)$$

The EM algorithm is a simple computational implementation to find the posterior mode. Fig. 1 shows multiple imputations with EMB algorithm schematic diagram. The diagram shows that the classic EM algorithms were combines with a bootstrap approach to earn draws from posterior. The bootstrap data used to stimulate estimation uncertainty and EM algorithm was run to discover the posterior mode for the bootstrapped data which gives us necessary uncertainty of the EMB algorithm for details.

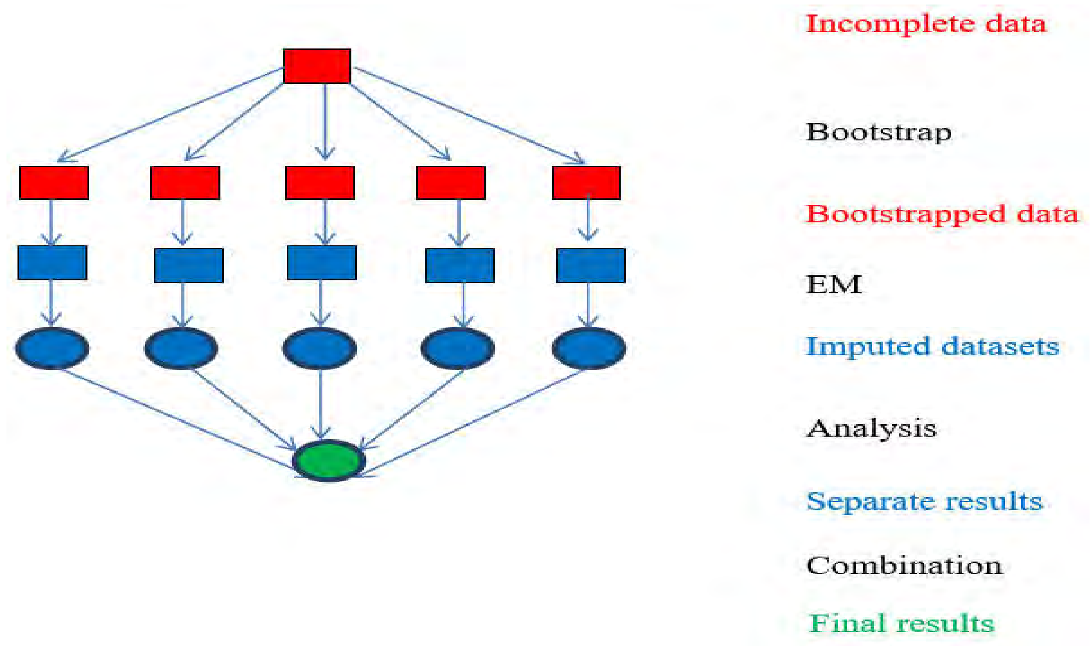


Fig.1. Multiple imputations with EMB algorithm schematic diagram

Imputations were created after draws of the complete data parameters posterior by drawing values of D_{mis} from its distribution conditional on D_{obs} and the draws of θ.

1.4. Nearest Neighbor

Unknown values also can be predicted using nearest neighbor imputation method by using the known values at neighboring locations [21].

3% of the original data sets were recorded as missing data are suitable to implement this imputation method [18]. This method used endpoint of the gaps to estimates all missing values [22]. The equation had shown as below:

$$y = y_1 \text{ if } x \leq x_1 + [(x_2 - x_1) / 2] \quad (5)$$

$$y = y_1 \text{ if } x \geq x_1 + [(x_2 - x_1) / 2] \quad (6)$$

where y represents the interpolate, x is the time point of the interpolate, y_1 and x_1 are the coordinates of the starting point of the gap, y_2 and x_2 are the end points of the gaps.

Some of air quality monitoring stations has incomplete continuous concentration of pollutants recorded. Missing data on that day also might be at the worst condition. Thus, network quality can pose a serious problem when experienced missing or incomplete data whereas the parameters used in the model can be overestimate and underestimate as well as can pose danger to the human beings and environment.

1.5. Artificial Neural Network (ANN)

ANN techniques are able to solve the problem associated with the air pollution modeling and prediction as many researchers have proven it. Besides, this technique also is able to provide advantage over classical statistics approach [23]. It is one of the implemented modeling especially in air pollutants modeling and prediction in soft computing techniques due to its ability to do so. ANN also is a flexibility tool with no former hypothesis and has capability to achieve experience with or without teacher and generalization.

Fig. 2 shows MLP-FF-ANN model network structure where the network structure consists of layers organized with multiple neurons. These allow information to flow via input system known as independent variable. Via a weighted connection system, the signal of input layer is passed to the hidden layer where the actual processing is done. Eventually, the signal then reached the output layer which is also known as dependent variable [24].

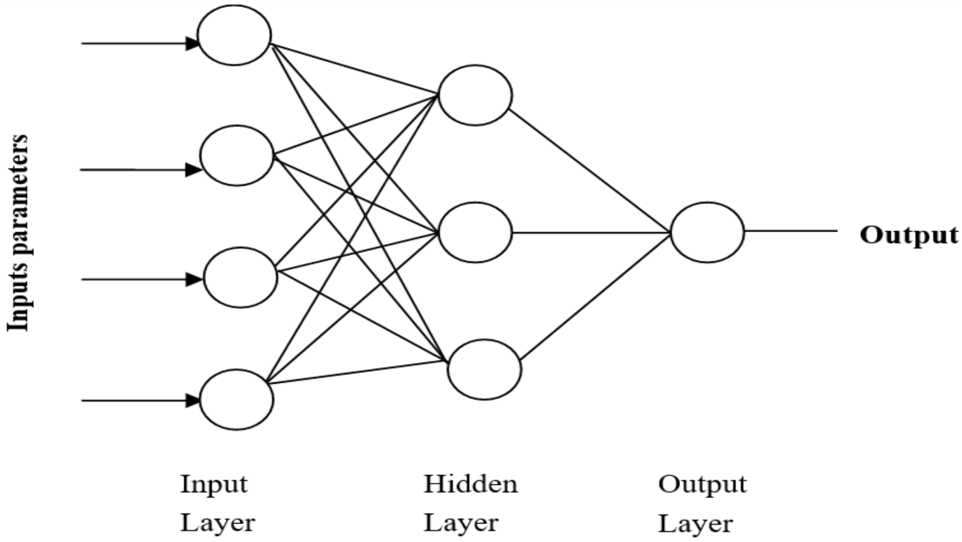


Fig.2. MLP-FF-ANN model network structure

1.6. Multi Linear Regression (MLR)

The variability between the independent variable and the dependent variable also can be predicted by using statistical technique such as multiple linear regressions (MLR).The regression model with k independent variables can be written as below [25]:

$$Y_i = \beta_0 + \beta_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (7)$$

where i = 1,…,n, β_1 = Regression coefficient, X_1= Independent variable and ε = Error associated with the regression.

1.7. Stepwise Linear Fit Model

Stepwise linear fit is a generally traditional linear method. In stepwise regression, a subset of effect was selected for a regression model. It eases searching and selecting among many models [26]. This method is based on successive linear regression where adding and removing operation of candidate sort. The input then attribute to the prediction of linear model. This process includes two types of operation, which are forward selection and backward elimination. For forward selection, beginning with no variables in the model, the addition significance then tested for each variable by adding the most influence variable that improves the model. For backward elimination, beginning with some set of variables, the removal of each variable tested by using a chosen criterion of the model quality. Variables were deleted until improvement is possible [27].

2. RESULTS AND DISCUSSION

In this study, 41.44% of missing data was observed from the analyzed of 15 air pollutants parameters namely Ws (wind speed), Wd (wind direction), Tempt (Temperature), UVB (Ultraviolet B), humidity, NO_x (Nitrogen Oxides), NO (Nitrogen monoxide), CH_4 (Methane), NmHC (Non methane Hydrocarbons), THC (Total Hydrocarbon), SO_2 (Sulphur dioxide), NO_2 (Nitrogen dioxide), O_3 (Ozone), CO (Carbon monoxide) and PM_{10} (Particulate Matter). Dataset gaps may lead to significant difficulty for analyzing the results. Thus, methods of imputation are the most extensively used method for filling missing observations.

According to [28], most omit the missing value from the analyses which may result in biased parameter expected. This support by the study done by [19] where removed parameters may carry significant information about the target, which than would be lost.

Table 1 shows R^2 and RMSE value obtained using different imputation method of linear and non-linear method. An attempt were done to compare imputation method that is EMB-algorithm available on a Amelia package on R software and nearest neighbor method, which is the most method used by the other authors in their research such as [17-18, 29].

Table 1. Value of R^2 and RMSE

	Imputation Method	**R^2**	**RMSE**
Fit-model	Nearest neighbor	0.3858	87.45
	EMB-algorithm	**0.9975**	**5.61**
MLR	Nearest neighbor	0.3857	87.47
	EMB-algorithm	**0.9623**	21.69
ANN	Nearest neighbor	0.4025	85.55
	EMB-algorithm	**0.9975**	**5.60**

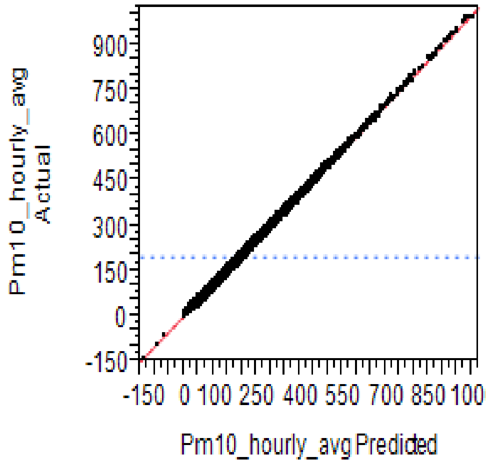
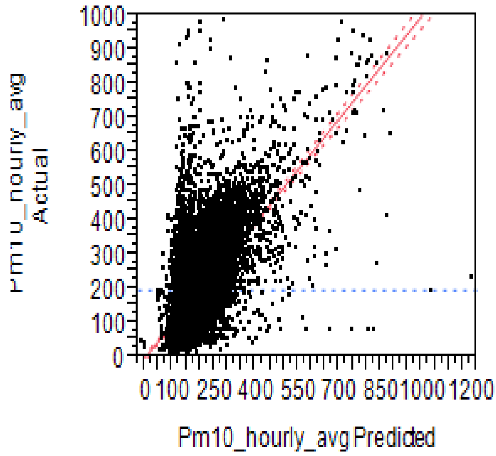
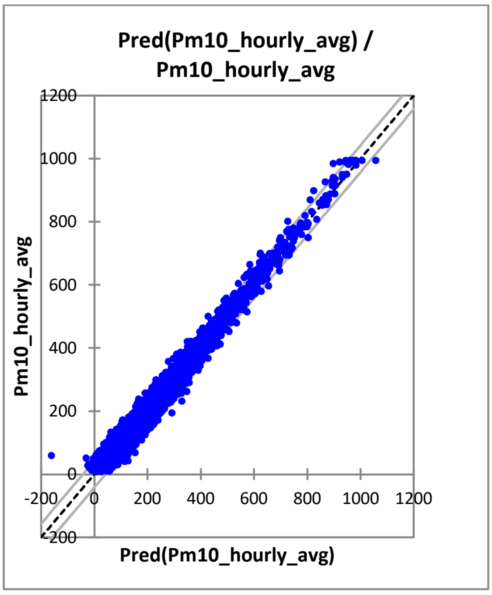
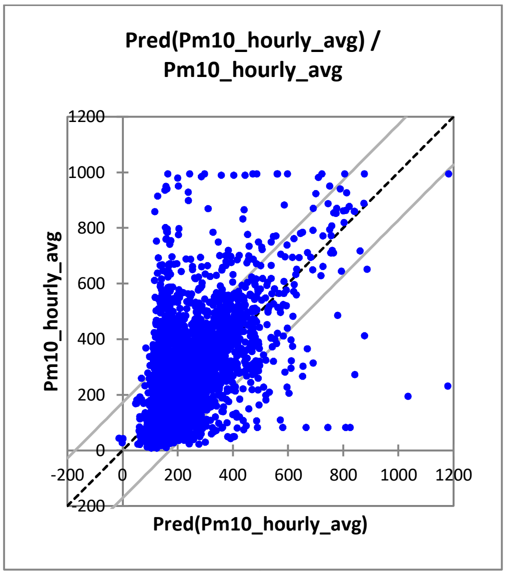
R^2 and RMSE value obtained using Fit-model when the imputation method of nearest neighbor and EMB-algorithm applied were 0.3858, 87.45 and 0.9975, 5.61 respectively. Meanwhile, R^2 and RMSE value obtained using MLR when the imputation method of nearest neighbor and EMB-algorithm applied were 0.3857, 87.47 and 0.9623, 21.69 respectively. While, R^2 and RMSE value obtained using ANN when the imputation method of nearest neighbor and EMB-algorithm applied were 0.4025, 85.55 and 0.9975, 5.62 respectively.

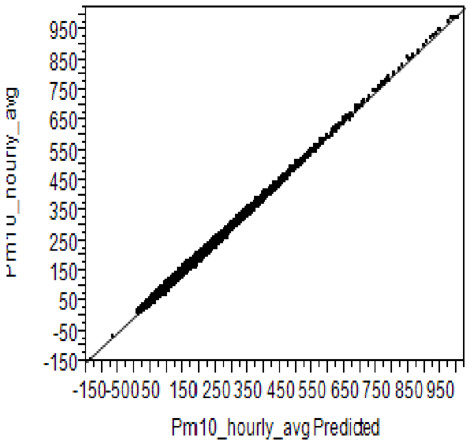
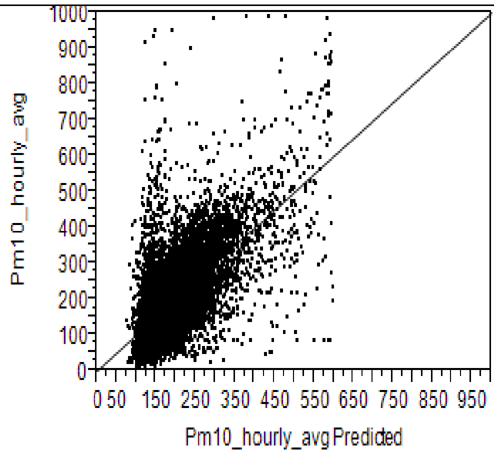
Fit-model and ANN shows better result with high R^2 and low RMSE when analyzed the data using imputation method of EMB-algorithm. According to [28], the EMB-algorithm imputation is to be greater especially when the percentage of missing values is high as it constantly gives low RMSE as compared with other methods. In this study, it shows that EMB-algorithm imputation is greater than nearest neighbor method.

A smaller RMSE is appropriate, since it shows that a forecast value is closer to the exact value, therefore more accurate [30]. Based on the results obtained, EMB-algorithm imputation method is preferable compared with nearest neighbor and suitable to be applied with 41.44% of missing data. While, nearest neighbor imputation method is unsuitable to be applied in this study regarding high percentage of missing data as nearest neighbor suitable be applied with the missing data recorded is 3% of the original data sets [18]. While, EMB-algorithm available on an Amelia package on R software were used for imputation method since it can works faster with larger number of variable, and is far easier to use.

Table 2 shows the actual by predicted plot using imputation method of nearest neighbor and EMB-algorithm. Data with imputation method of EMB-algorithm shows better plotted rather than nearest neighbor. Some of the research by others shows that ANN performed better than MLR in predicting concentration of PM_{10} [31-33]. However, in this study after considering the missing data and applied imputation method of EMB-algorithm into ANN and MLR, MLR also shows good results with R^2 obtained 0.9623. This shows that MLR also has capable as ANN for prediction after apply altogether with imputation method of EMB-algorithm.

Table 2. Actual versus predicted plotted (EMR-algorithm and nearest neighbour imputation methods)

	EMB-Algorithm	Nearest Neighbour
Fit-model		
MLR		

ANN		

The selection of input parameter is very vital in gaining the effective neural network [34-37] modeling. The results of sensitivity analysis (SA) presented in Table 3 show the determination of coefficients for each parameter affecting the prediction concentration of PM_{10}. In this method, one variable omitted at one time (leave-one-out) to determine the contribution percentage poses by the variable that would affect the R^2 values. Fifteen parameters were used to predict concentration of PM_{10}, where the R^2 value of this model was made a reference to other models developed in SA. From the Table 3, we can see that the highest contribution percentage is humidity (20.32%), followed by temperature (19.39%), CO (16.12%), UVB (14.28%), O_3 (11.27%), WD (7.14%), NmHc (3.20%), SO_2 (3.02%), THC(1.97%), NO_2 (1.95%), WS (0.60%), CH_4 (0.46%), NO (0.15%), NO_x (0.12%) in descending order as shown below:

Humidity> TEMP > CO> UVB> O_3>WD>NmHC>SO_2>THC>NO_2>WS>CH_4>NO>NO_x

Table 3. Sensitivity analysis results for PM_{10} prediction

Model	**R^2**		
ANN-PM_{10}-AP	**0.9975**	**Difference R^2**	**% Contribution**
ANN-PM_{10}-LWS	0.9932	0.0043	0.59606321
HM-LWD	0.946	0.0515	7.13889659
HM-LTEMP	0.8576	0.1399	**19.39284724**
HM-LUVB	0.8945	0.103	**14.27779318**
HM-RH	0.8509	0.1466	**20.32159689**
HM-LNO_X	0.9966	0.0009	0.124757416
HM-LNO	0.9964	0.0011	0.152481286
HM-LCH_4	0.9942	0.0033	0.457443859
HM-LNmHC	0.9744	0.0231	3.202107014
HM-LTHC	0.9833	0.0142	1.968394788
HM-LSO_2	0.9757	0.0218	3.021901857
HM-LNO_2	0.9834	0.0141	1.954532853
HM-LO_3	0.9162	0.0813	**11.26975326**
HM-LCO	0.8812	0.1163	**16.12143055**
Total		0.7214	100

The result from the SA method implies that humidity, temperature, CO, UVB and O_3 are the main contributor to the air pollutants in the study area. Table 4 shows predicting performance of the difference ANN models. Two ANN models were done to compare and select the best input selection for PM_{10} prediction in Malaysia. The ANN-PM_{10}-AP model (uses fifteen parameters) was used as a reference model with the R^2 value was 0.9975. While the second ANN model, ANN-PM_{10}-LO (uses five parameters) as inputs shows R^2 value 0.5801. Although the ANN-PM_{10}-LO gives a lower value in R^2 (0.5801) than ANN-PM_{10}-AP (0.9975), but this model was considered as the best model of prediction because it uses fewer variables (only humidity, temperature, CO, UVB and O_3) as input and is far less complex than others.

Table 4. Predicting performance of the difference ANN models

Model	R^2
ANN-PM_{10}-AP	0.9975
ANN-PM_{10}-LO	0.5801

The formula created from the ANN-PM_{10}-LO as mentioned below:

Hidden Layer Code, H1 = tanh [(.5*(0.015*Temp) + (-0.00007*UVB) + (0.00213*Humidity) + (-0.355279*O_3) + (0.17427*CO) + (-0.65745)]

Final Layer Code, PM_{10} = (1126.9226*H1) + 124.24435

3. METHODOLOGY

3.1. Data Collection and Study Areas

Hourly air quality data consist of air pollutants data (NO_x, NO, SO_2, NO_2, O_3, CO, PM_{10}, CH_4, NmHC, THC, Wd, Ws, Temp, UVB, humidity) from 2010 to 2015 were obtained from Air Quality Division, Department of Environment (DOE) Malaysia. Only data with API greater than 100 being analyzed as at this state Air Pollution Index (API) being considered at unhealthy level.

There are 50 reliable and accurate continuous air quality monitoring stations in Malaysia which distributed on regions according to the locations' nature that were to be monitored. These types are: comprehensive, residential, industrial and PM_{10}. Fig. 3 shows the monitoring stations distribution across Malaysia region of air quality. Table 5 (a)-(n) shows the sampling point for air quality monitoring stations in Malaysia.

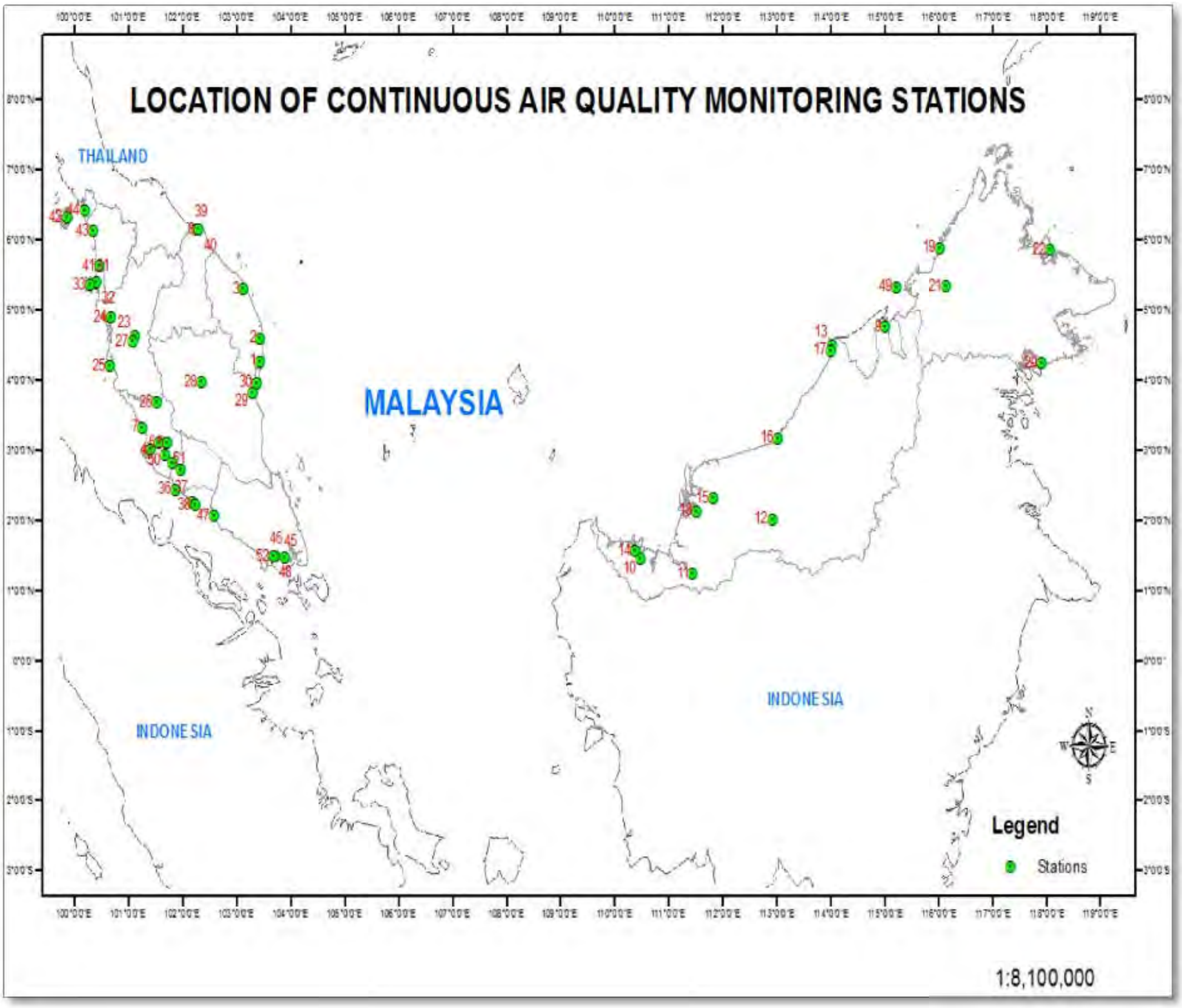


Fig.3. Monitoring stations distribution across Malaysia region of air quality

Table 5 (a). Sampling point for air quality monitoring stations in Terengganu

Loc. No.	Locations	Latitude (N)	Longitude (E)
1	Sek. Ren. Keb. Bukit Kuang, Teluk Kalung, Kemaman	04°16.260’	103°25.826’
2	Kuarters TNB, Paka-Kertih	04°35.880’	103°26.096’
3	Sek. Keb. Chabang Tiga, Kuala Terengganu	05°18.455’	103°07.213’

Table 5 (b). Sampling point for air quality monitoring station in Selangor

Loc. No.	Locations	Latitude (N)	Longitude (E)
4	Sekolah Menengah (P) Raja Zarina, Klang	03°00.602’	101°24.484’
5	Sek. Keb. Bandar Utama, Petaling Jaya	03°06.612’	101°42.274'
6	Sek. Keb. TTDI Jaya, Shah Alam	03°06.286'	101°33.367'
7	Sekolah Menengah Sains, Kuala Selangor	03°19.592’	101°15.532’
8	Kolej MARA Banting	02°49.001’	101°37.381’

Table 5 (c). Sampling point for air quality monitoring stations in Sarawak

Loc. No.	Locations	Latitude (N)	Longitude (E)
9	Dewan Suarah, Limbang	04°45.529’	115°00.813
10	Pej. Daerah, Kota Samarahan, Sarawak	01°27.308’	110°29.498
11	Kompleks Sukan, Sri Aman	01°14.425'	111°27.629'
12	Stadium Tertutup, Kapit	02°00.875’	112°55.640
14	Medical Store, Kuching, Sarawak	01°33.734’	110°23.329
15	Ibu Pej. Polis Sibu, Sarawak	02°18.856’	111°49.906
16	Balai Polis Pusat Bintulu, Sarawak	03°10.587’	113°02.433
17	Sek. Men Dato’ Permaisuri Miri, Sarawak	04°25.456’	114°00.731
18	Balai Polis Pusat Sarikei, Sarawak	02°07.992’	111°31.351

Table 5 (d). Sampling point for air quality monitoring stations in Sabah

Loc. No.	Locations	Latitude (N)	Longitude (E)
19	Sek. Men. Keb. Putatan, Kota Kinabalu	05°53.623’	116°02.596’
20	Pejabat JKR, Tawau, Sabah	04°15.016’	117°56.166’
21	Sek. Men. Keb. Gunsanad, Keningau, Sabah	05°20.313’	116°09.769’
22	Pej. JKR Sandakan, Sandakan	05°51.865'	118°05.479'

Table 5 (e). Sampling point for air quality monitoring stations in Perak

Loc. No.	Locations	Latitude (N)	Longitude (E)
23	Sek. Men Jalan Tasek, Ipoh	04°37.781’	101°06.964’
24	Sek. Men. Keb. Air Puteh, Taiping	04°53.940’	100°40.782’
25	Pejabat Pentadbiran Daerah Manjung, Perak	04°12.038’	100°39.841’
26	UPSI, Tanjung Malim	03°41.267’	101°31.466’
27	Sek. Men. Pagoh, Ipoh, Perak	04°33.155’	101°04.856’

Table 5 (f). Sampling point for air quality monitoring stations in Pahang

Loc. No.	Locations	Latitude (N)	Longitude (E)
28	Pejabat Kaji Cuaca Batu Embun, Jerantut	03°58.238’	102°20.863’
29	Sekolah Kebangsann Indera Mahkota, Kuantan	03°49.138’	103°17.817’
30	Sekolah Kebangsaan Balok Baru, Kuantan	03°57.726’	103°22.955’

Table 5 (g). Sampling point for air quality monitoring stations in Pulau Pinang

Loc. No.	Locations	Latitude (N)	Longitude (E)
31	Sek. Keb. Cenderawasih, Perai	05°23.470'	100°23.213'
32	Sek. Keb. Sebarang Jaya II, Perai	05°23.890’	100°24.194’
33	UniversitiSains Malaysia, Pulau Pinang	05°21.528’	100°17.864’

Table 5 (h). Sampling point for air quality monitoring stations in Negeri Sembilan

Loc. No.	Locations	Latitude (N)	Longitude (E)
35	Sek. Men. Teknik Tuanku Jaafar	02°43.418’	101°58.105’
36	Pusat Sumber Pendidikan Negeri Sembilan	02°26.458’	101°51.956’

Table 5 (i). Sampling point for air quality monitoring stations in Melaka

Loc. No.	Locations	Latitude (N)	Longitude (E)
37	Sek. Men. Keb. Bukit Rambai, Melaka	02°15.510’	102°10.364’
38	Sek. Men. Tinggi, Melaka	02°12.789'	102°14.055'

Table 5 (j). Sampling point for air quality monitoring stations in Kelantan

Loc. No.	Locations	Latitude (N)	Longitude (E)
39	Sek. Men. Keb. Tanjung Chat, Kota Bharu	06°09.520'	102°15.059'
40	SMK Tanah Merah	05°48.671’	102°08.000’

Table 5 (k). Sampling point for air quality monitoring stations in Kedah

Loc. No.	Locations	Latitude (N)	Longitude (E)
41	Sekolah Kebangsaan Bakar Arang, Sg. Petani	05°37.886’	100°28.189’
42	Kompleks Sukan Langkawi, Kedah	06°19.903’	099°51.517’
43	Sek. Men. Agama Mergong, Alor Setar, Kedah	06°08.218’	100°20.880’

Table 5 (l). Sampling point for air quality monitoring stations in Perlis

Loc. No.	Locations	Latitude (N)	Longitude (E)
44	ILP Kangar	06°25.424’	100°11.046’

Table 5 (m). Sampling point for air quality monitoring stations in Johor

Loc. No.	Locations	Latitude (N)	Longitude (E)
45	Sekolah Menengah Pasir Gudang 2	01°28.225’	103°53.637’
46	Institut Perguruan Malaysia, Temenggong Ibrahim	01°28.225'	103°53.637'
47	Sek. Men. Teknik Muar, Muar, Johor	02°03.715'	102°35.587'
48	SMA Bandar Penawar, Kota Tinggi, Johor	01°33.500’	104°13.310’

Table 5 (n). Sampling point for air quality monitoring stations in Wilayah Persekutuan

Loc. No.	Locations	Latitude (N)	Longitude (E)
49	Taman Perumahan Majlis Perbandaran Labuan	05°19.980’	115°14.315’
50	Sek. Keb. Putrajaya 8(2), Jln P8/E2, Presint 8, Putrajaya	02°55.915’	101°40.909’
51	Sek. Men. Keb. Seri Permaisuri, Cheras	03°06.376’	101°43.072’
52	Sek. Keb. Batu Muda, Batu Muda, Kuala Lumpur	03°12.748’	101°40.929’

3.2. Data Availability

Analysis, filtration and transformation of collected data were performed to ensure analysis performs better to diminish noise and focus important relationships during training models in developing models. Only data with API greater than 100 from 50 continuous air monitoring stations were selected to analyze as API at this level considered unhealthy and might give effect to human and environment. Few missing values of variables were.

There are 19,872 data obtained for each parameter. However, there are missing data being observed due to the technical failure that cause monitoring instruments failed to collect data at that time (Table 6).

Table 6. Minimum (min), maximum (max) and mean of air pollutants and meteorological data from 2010 to 2015 in Malaysia

Parameters	Min	Max	Mean
Wind speed (km/hr)	0.7	19.1	4.852
Wind direction	0	360	165.1
Temperature (°C)	19.4	39	28.68
Humidity (%)	20	103	74.55
NO_x (ppm)	0	0.19	0.02286
NO (ppm)	0	0.12	0.006854
SO_2 (ppm)	0	0.05	0.00327
NO_2 (ppm)	0	0.08	0.0158
O_3 (ppm)	0	0.17	0.03542
CO (ppm)	0.05	15.8	1.659
PM_{10} (µg/cu.m))	9	995	196.8
CH_4 (ppm)	1.54	2.63	1.987
NmHC (ppm)	0.03	0.66	0.2564
THC (ppm)	1.6	3.1	2.238
UVB	321	1386	1017

3.3. Variable Selection

Selection of variables is needed in designing forecasting model. Input nodes' number in ANN forecasting model determined from the selected input variables. According to the [8], one experienced slow training speed due to increasing learning time caused by the larger inputs number. Thus, input selection is the vital step in forecasting model to ensure optimal input variables were selected in order to diminish redundant, over-fitting and noise variables. The input variables such as Ws, Wd, temperature, humidity, NO_x, NO, SO_2, NO_2, O_3, CO, CH_4, NmHC, THC and UVB were used to predict PM_{10} at the beginning of the analysis.

3.4. Implementation Model

Stages of implementations were repeated until the ultimate forecasting model with good accuracy obtained. PM_{10} forecasting model were implement using JMP 10.

4. CONCLUSION

As a conclusion, imputation method of EMB-algorithm is better than nearest neighbor since the analysis of imputation data using EMB-algorithm on Fit-model, MLR and ANN shows great value of R^2 which are 0.9975, 0.9623 and 0.9975 respectively. Besides, it can works faster with larger number of variable and is far easier to use. It is applicable on high percentage of missing data whereby in this study missing, data is 41.44% vice versa with imputation method of nearest neighbor. It is prove that it only applicable for the small percentage of missing data as mention by [18], which usually applicable on 3% of missing data. Application of imputation method of EMB-algorithm with Fit-model and ANN shows greater results compared with MLR due to their high percentage of R^2 and low value obtained for RMSE which is 0.9975, 5.61 and 0.9975, 5.60 respectively. From the sensitivity analysis, there are five parameters being main contributor for the prediction of PM_{10} which are humidity, temperature, CO, UVB and O_3. Reduce number of parameters from fifteen to five as input makes it the best model of PM_{10} prediction. A formula created from the ANN-PM_{10}-LO model. From this formula, the missing data of PM_{10} concentration can be predicted.

5. ACKNOWLEDGEMENTS

The authors acknowledge the Air Quality Division of the Department of Environment (DOE) under the Ministry of Natural Resource and Environment, Malaysia who have provided us with valuable data and give permission to utilize air quality data for this study.

6. REFERENCES

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

[2] Kerbachi R, Boughedaoui M, Bounoua L, Keddam M. Ambient air pollution by aromatic hydrocarbons in Algiers. Atmospheric Environment, 2006, 40(21):3995-4003

[3] Pope III C A, Burnett R T, Thun M J, Calle E E, Krewski D, Ito K, Thurston G D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association, 2002, 287(9):1132-1141

[4] Bell M L, McDermott A, Zeger S L, Samet J M, Dominici F. Ozone and short-term mortality in 95 US urban communities, 1987-2000. Journal of the American Medical Association, 2004, 292(19):2372-2378

[5] Nayak P C, Sudheer K P, Rangan D M, Ramasastri K S. Short-term flood forecasting with a Neurofuzzy model. Water Resources Research, 2005, 41(4):2517-253

[6] Brunekreef B. Air pollution and human health: From local to global issues. Procedia-Social and Behavioral Sciences, 2010, 2(5):6661-6669

[7] Kurt A, Gulbagci B, Karaca F, Alagha O. An online air pollution forecasting system using neural networks. Environment International, 2008, 34(5):592-598

[8] Ababneh M F, Ala'a O, Btoush M H. PM10 forecasting using soft computing techniques. Research Journal of Applied Sciences, Engineering and Technology, 2014, 7(16):3253-3265

[9] European Environmental Agency (EEA). Air quality in Europe-2016 report no. 28/2016. Copenhagen: EEA, 2016

[10] Leung Y K, Lam C Y. Visibility impairment in Hong Kong-A wind attribution analysis. Bulletin of Hong Kong Meteorological Society, 2008, 18:33-48

[11] Chan L Y, Kwok W S, Lee S C, Chan C Y. Spatial variation of mass concentration of roadside suspended particulate matter in metropolitan Hong Kong. Atmospheric Environment, 2001, 35(18):3167-3176

[12] Peirce J J, Weiner R F, Vesilind P A. Meteorology and air quality. In P. A. Vesilind, J. J. Peirce, & R. F. Weiner (Eds.), Environmental pollution and control. Massachusetts: Butterworth-Heinemann, 2013, pp. 267-288

[13] Norazian M N, Shukri A, Yahaya P M, Azam N, Fitri N F, Yusof M. Roles of imputation methods for filling the missing values: A review. Advances in Environmental Biology, 7(12):3861-3869

[14] Noor N M, Zainudin M L. A review: Missing values in environmental data sets. In International Conference on Environment, 2008, pp. 1-9

[15] Yusof N F, Ramli N A, Yahaya A S, Sansuddin N, Ghazali N A, Al Madhoun W. Monsoonal differences and probability distribution of PM10 concentration. Environmental Monitoring and Assessment, 2010, 163(1-4):655-667

[16] Noor N M, Yahaya A S, Ramli N A, Abdullah M M. The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique. Journal of Engineering Research and Education, 2006, 3:96-105

[17] Shaadan N, Deni S M, Jemain A A. Assessing and comparing PM10 pollutant behaviour using functional data approach. Sains Malaysiana, 2012, 41(11):1335-1344

[18] Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 2004, 38(18):2895-2907

[19] Žliobaitė I, Hollmén J, Junninen H. Regression models tolerant to massively missing data: A case study in solar-radiation nowcasting. Atmospheric Measurement Techniques, 2014, 7(12):4387-4399

[20] Honaker J, King G, Blackwell M. Amelia II: A program for missing data. Journal of Statistical Software, 2011, 45(7):1-47

[21] Azid A, Juahir H, Toriman M E, Kamarudin M K, Saudi A S, Hasnam C N, Aziz N A, Azaman F, Latif M T, Zainuddin S F, Osman M R. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. Water, Air, and Soil Pollution, 2014, 225(8):2063-2077

[22] Isiyaka H A, Azid A. Air quality pattern assessment in Malaysia using multivariate techniques. Malaysian Journal of Analytical Sciences, 2015, 19(5):966-978

[23] Yildirim Y, Bayramoglu M. Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere, 2006, 63(9):1575-1582

[24] Dongare A D, Kachare A D. Predictive tool: An artificial neural network. International Journal of Engineering and Innovative Technology, 2012, 2(1):209-214

[25] Kovač-Andrić E, Brana J, Gvozdić V. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. Ecological Informatics, 2009, 4(2):117-122

[26] SAS Institute Inc. JMP® 10 modelling and multivariate methods. North Carolina: SAS Institute Inc., 2012

[27] Siwek K, Osowski S. Data mining methods for prediction of air pollution. International

Journal of Applied Mathematics and Computer Science, 2016, 26(2):467-478

[28] Razak N A, Zubairi Y Z, Yunus R M. Imputing missing values in modelling the PM10 concentrations. Sains Malaysiana, 2014, 43(10):1599-1607

[29] Azid A, Juahir H, Latif M T, Zain S M, Osman M R. Feed-forward artificial neural network model for air pollutant index prediction in the southern region of Peninsular Malaysia. Journal of Environmental Protection, 2013, 4(12):1-10

[30] Schafer J L, Graham J W. Missing data: Our view of the state of the art. Psychological Methods, 2002, 7(2):147-177

[31] Afzali A, Rashid M, Sabariah B, Ramli M. PM10 pollution: Its prediction and meteorological influence in Pasir Gudang, Johor. IOP Conference Series: Earth and Environmental Science, 2014, 18(1):1-6

[32] Özdemir U, Taner S. Impacts of meteorological factors on PM10: Artificial neural networks (ANN) and multiple linear regression (MLR) approaches. Environmental Forensics, 2014, 15(4):329-336

[33] Russo A, Lind P G, Raischel F, Trigo R, Mendes M. Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research, 2015, 6(3):540-549

[34] Zabidi A, Yassin I M, Hassan H A, Ismail N, Hamzah M M, Rizman Z I, Abidin H Z. Detection of asphyxia in infants using deep learning convolutional neural network (CNN) trained on Mel frequency cepstrum coefficient (MFCC) features extracted from cry sounds. Journal of Fundamental and Applied Sciences, 2017, 9(3S):768-778

[35] Hashim F R, Daud N N, Ahmad K A, Adnan J, Rizman Z I. Prediction of rainfall based on weather parameter using artificial neural network. Journal of Fundamental and Applied Sciences, 2017, 9(3S):493-502

[36] Hashim F R, Adnan J, Ibrahim M M, Ishak M T, Din M F, Daud N G, Rizman Z I. Heart abnormality detection by using artificial neural network. Journal of Fundamental and Applied Sciences, 2017, 9(3S):1-10

[37] Mohd Yassin I, Jailani R, Ali M, Amin M S, Baharom R, Hassan A, Huzaifah A, Rizman Z I. Comparison between cascade forward and multi-layer perceptron neural networks for

NARX functional electrical stimulation (FES)-based muscle model. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(1):215-221

How to cite this article:
Abd Rani NL, Azid A, Khalit SI, Juahir H. Prediction model of missing data: a case study of PM_{10} across Malaysia region. J. Fundam. Appl. Sci., 2018, 10*(1S), 182-203.*