Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 88 (2015) 36261-36268

ISSN: 2229-712X

Modelling Imputation Techniques for Prediction Ozone Concentration

N. Suhaimi, N.A. Ghazali, M.Y. Nasir and M.I.Z. Mokhtar

School of Ocean Engineering, University Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia.

ARTICLE INFO

Article history:

Received: 1October 2015;
Received in revised form:
28 October 2015;
Accepted: 05 November 2015;

Keywords

Ozone,
Mean Top Bottom,
Linear Interpolation,
Multiple Imputation,
Multiple Linear Regression.

ABSTRACT

The presence missing values in air quality data is a common issue. This paper considers hourly air pollutant concentrations such as NO, NO_2, CO, SO_2 and meteorological variables such as AT, RH, WS at Kemaman over a four-year period. This paper discusses the performance of Multiple Linear Regression (MLR) model using different imputation techniques such as mean top bottom, linear imputation and Markov Chain Monte Carlo (MCMC). The result shows that MCMC is the best method to replace missing values.

Introduction

Air quality study was carried out to detect any pollutant concentrations which could have possible adverse effects on human health especially ozone (O_3) concentrations. The effect of O_3 can damage the growth of plant, various natural materials and manufactured goods and lead to the damage of human lung tissues (Wang and Georgopoulos, 2001). O_3 is a combination of three atoms of oxygen and occurs both in Earth's upper atmosphere and at the ground level (Ramli *et al*., 2010). It is a secondary pollutant regulated under the National Ambient Air Quality Standards (NAAQS) and formed through photochemical reaction of a variety of natural and anthropogenic precursors. It is produced from the interaction of meteorology, nitrogen oxides (NO_x), and anthropogenic precursors such as industrial sources and automobiles of volatile organic compounds (VOC_S) (Finlayson-Pitts and Pitts, 1986; Saunders *et al*., 1997, Abdul-Wahab *et al*., 2005). According to Tiwary and Colls (2010), the formation of O_3 in the troposphere was described from the chemical process of nitrogen dioxides (NO_2) photolysis that involving radiant energy (*hv*) from the sun. Photons with a wavelength less than 400 nm are able to break the NO_2 molecule to NO and O atom as shown in equation (1). Then, free O atom will combine with O molecule and 'third body' (M) molecule to generate O_3 (equation 2). The O_3 can react quickly with NO to produce NO_2 and O_2 as shown in equation (3).

$$NO_2 + hv\,(\lambda < 400\text{ nm}) \longrightarrow NO + O \tag{1}$$

$$O + O_2 + M \longrightarrow O_3 + M \tag{2}$$

* M is any "body" with mass

$$NO + O_3 \longrightarrow NO_2 + O_2 \tag{3}$$

The air quality data was collected from the continuous air quality monitoring station managed by Alam Sekitar Malaysia Sdn. Bhd. (ASMA). Missing data was clearly visible in real datasets especially in air quality dataset. The missing data will give a large problem for the real dataset because statistical analysis will become complicated. The effect of missing value is inability data to make accurate result, loss of efficiency, bias estimates and reduction of statistical power (Hawthorne and Elliot, 2005). The missing data in air quality monitoring might be caused by machine failure, routine maintenance, changes the place of air monitors, and human error (Hawthorne and Elliott, 2005). In statistics, there are three types of missing data. The first form is missing completely at random (MCAR) where the missing value occurs at random for the whole dataset (Rubin, 1976; Little and Rubin, 2004). The second form is missing at random (MAR) where data is missing independently. The third form for missing data is missing not at random (MNAR) where the missing value is related to unobserved data. Normally, air quality data is classified as MAR form because missing data is generally random and due to the monitoring site down (Junninen, 2004; Noor *et al*., 2006; Plaia & Bondí, 2006).

Selection for an appropriate method for handling missing values depends on the missing data pattern and missing data mechanism (Plaia and Bondi, 2006). There are various techniques such as single imputation and multiple imputation. The selection of imputation method depends on the pattern of missingness in the data and the type of the imputed variable. There are two patterns of missingness in the data such as monotone and arbitrary (Yuan, 2011). The monotone missing pattern for continuous variable consist of three different patterns such as regression method (Rubin, 1987), a predictive mean matching method (Heitjan and Litte, 1991; Schenker and Taylor, 1996) and propensity score method (Lavori, Dawson, and Shera, 1995). Arbitrary missing data pattern occurred when data missing in random (Dong and Peng, 2013). For arbitrary missing pattern, Markov Chain Monte Carlo is one of the example (Yuan, 2011). From the previous study, various imputation techniques have been proposed in environmental study such as single and multiple techniques (Little and Rubin, 2002; Yahaya *et al*., 2005; Noor *et al*., 2006; Noor *et al*., 2014). . The examples for single techniques are interpolation techniques, mean imputation and hot-deck (Noor *et al*., 2006; Plaia & Bondí, 2006). The example for multiple imputation is Markov Chain Monte Carlo method (Lu & Wang, 2008; Jerez *et al*., 2010).

Tele:
E-mail addresses: nurul.adyani@umt.edu.my

Single imputation technique explained each missing item was imputed by only one estimated value. According to Little and Rubin (2002) when the percentage of missing value is low, the common approach to solve the missing value problem is ignoring the missing value. However, Yahaya *et al.* (2005) explained ignoring the missing value will caused biased estimations when there are large percentages of missing value. Noor *et al.* (2006) has conducted a study using mean top bottom technique to replace missing values for the PM_{10} data. This study concluded that mean top bottom method gave good performances but the performances decreased slightly at higher percentage of missing value where coefficient of determination (R^2) for 5% missing value is 0.87, for 15% of missing values is 0.86 and for 40% missing values is 0.77. Another study conducted by Noor *et al.* (2014) has explained the interpolation methods to solve the missing value problem. In this study, they compared several interpolation methods that are linear, quadratic and cubic interpolation. From the study, they concluded that all three methods can be used to replace the missing value for PM_{10} data because of high value of R^2 is 0.98.

Meanwhile, multiple imputation technique was defined as replaced missing value with a set of plausible values (Clark *et al.* 2003). Multiple imputation is a sophisticated technique to handle missing value that give much better results (Rubin, 1987; Little, 1992; Greenland *et al*., 1995; Schafer, 1997; Vach, 2004). The study about simulation for the multiple imputation (MI) has shown that for missing data less than 30% required five to ten replications are required to provide reasonable estimate of missing data for the parameters of interest (ESI, 2005). According to Yuan (2011), MI inference involves three different steps; firstly, the missing data are filled in *m* times to generate *m* complete data sets, secondly, the *m* complete data sets are analyzed using standard testing such as regression modelling and thirdly, the result from the *m* complete data sets are combined for the inference. The Markov Chain Monte Carlo (MCMC) method is one of the multiple imputation technique based on the imputation algorithm that assumed the data is multivariate normal distribution and generated imputation from the Bayesian distribution (Schafer, 1997). Markov chain is a sequence of random variables in the distribution of each variable depending on the value of the previous variable. The expectation-maximization (EM) algorithm is a technique that estimates maximum likelihood for MCMC method (Little and Rubin, 2002). A study from Gómez-Carracedo *et al.* (2014) reported that MI is a suitable method to solve the high ratio of missing values. Another study from Ingsrisawang *et al.* (2012) shows that MCMC method was more effective than simple mean for the monthly rainfall data in the northern region of Thailand.

In addition, after replacing the missing value in dataset, such relationships between precursors have been examined. The development of model to predict O_3 concentrations can improve public health strategies. Concerning statistical approaches, linear and non-linear models have been widely applied to predict O_3 concentrations. Several studies have been reported in the environment field regarding different kind of models. In recent years, many statistical analyses have been used to study the air pollution pattern especially in urban areas. According to Borrego *et al.* (2003), O_3 concentration is very difficult to predict because of the different interactions between primary air pollutants and meteorological variables. However, empirical O_3 modelling and regression models to identify the relationship between primary air pollutants, meteorological conditions and O_3 concentrations have been largely studied which have used a combination of statistical regression, graphical analysis, fuzzy logic based method, artificial neural networks, and cluster analysis (Abdul-Wahab *et al*., 2000; Abdul-Wahab, 2001; Abdul-Wahab and Al-Alawi, 2002; Go'mez-Sanchis *et al*., 2006; Özbay *et al*., 2011). Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several independent (predictor) variables especially in O_3 prediction (Abdul-Wahab, 2005; Ghazali *et al*., 2010; Awang *et al*., 2015). MLR is a statistical tool for understanding the relationship between two or more variables. MLR is the most widely used multivariate technique for the primary purpose of prediction. The goal in MLR is to develop a statistical model that can be used to predict the values of a dependent variable based on the values of at least one independent variable. MLR also enables to determine the effect of predictors on the dependent variable. However, multicollinearity can occur when two or more independent variables are highly correlated. Thus, can make it difficult to identify correctly the most important contributors to a response variable. These can be detected by examining the variance inflation factor (VIF) and Tolerance values (TOL). There is no multicollinearity problem when TOL values are greater than 0.1 and VIF values less than 10 (Neter *et al*., 1983).

Thus, in this study, the main objective is to replace the missing value in order to fit multiple linear regression model for predicting O_3 concentrations. This paper focuses on the several imputation techniques to replace the missing values of environmental pollutants data and meteorological variables data. Then, the most precursors that affecting the ozone concentrations will be identified via multiple linear regression model.

Methodology

Description of air quality data

This approach is developed to replace the missing value for measuring the air quality data over the period of 2009 to 2012 at Kemaman, Terengganu air monitoring site. Annual hourly monitoring records for air quality concentrations were selected to carry out the simulation of missing values. The air quality concentration that has calibration was considered missing. There are five selected air pollutants such as ozone (O_3), nitrogen oxide (NO) and nitrogen dioxide (NO_2), carbon monoxide (CO), sulphur dioxide (SO_2) and three meteorological variables such as ambient temperature (AT), relative humidity (RH) and wind speed (WS) were used in this study.

The O_3 concentration at Alam Sekitar Malaysia Sdn. Bhd. (ASMA) stations was measured using Teledyne Ozone Analyzer Model 400E UV Absorption. The analyzer uses Beer-Lambert law for measuring low ranges of O_3 in ambient air. The concentration of nitrogen oxides (NO_x) was determined using chemiluminescence measurement principle, coupled with state-of-the-art microprocessor technology for monitoring high and medium levels of NO_x (Teledyne Models 200EH and 200EM).

Description of sampling site

The air quality data was collected from the air quality monitoring sites owned by Department of Environment (DoE) in Malaysia and managed by a private company ASMA. Kemaman is a developing Malaysian city located at the industrial Kertih Petrochemical Industrial Area in the North and the industrializing and urbanizing Gebeng Industrial Area in the South. The monitoring station is located at Sekolah Rendah Bukit Kuang with coordinate (4°14'21.9" N, 103°11'31.8" E). Total population for the year 2010 is 167, 824 and width is 253,599.9 hectare (MPK, 2015). The description location of the continuous air monitoring station is a summarized in Fig. 1.

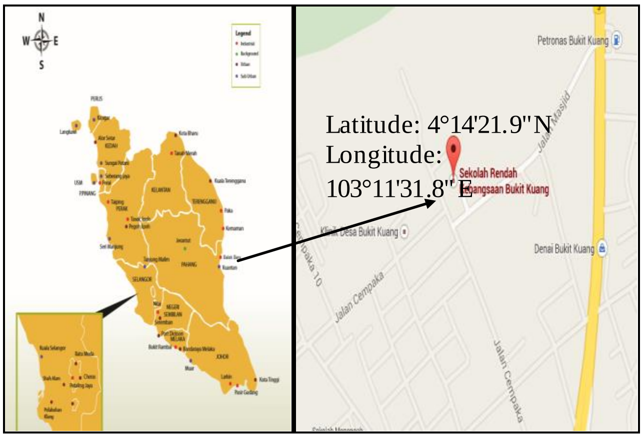

Figure 1. Location of Air Quality Monitoring Stations in Peninsular Malaysia, 2012

Imputation techniques

Missing data can be solved by either single or multiple imputation technique. This study compared the missing data imputation techniques between single and multiple imputations. The new dataset obtained will be used to fit MLR model with the different imputation techniques. Given the dependent variable represents O_3 concentrations and the independent variables represent primary pollutant and meteorological variables. The statistical analysis of imputation techniques was conducted by using SPSS software.

Single imputation

Linear interpolation

The linear interpolation is connecting two data points with a straight line. The linear interpolation is evaluated by the equation (Chapra and Canale, 1998):

$$f_1(x) = b_0 + b_1(x - x_0) \tag{4}$$

where x is the independent variable, x_0 is a known value of the independent variable, $f_1(x)$ is the value of the dependent variable for a value x of the dependent variable, and b_1 is unknown coefficient. Then from equation (4),

$$b_0 = f(x_0) \tag{5}$$

and

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{6}$$

Mean top bottom

Let $x_1, x_2, \ldots, x_n$ be a time series data with denoted by $x_1^*, x_2^*, \ldots, x_k^*$ where n observations and k is the missing values. The first missing value occurs after n_1 observations, the second missing value occur after n_2 observations and so on. Thus, the observed data with missing values can be expressed as follows (Yahaya et al., 2005; Noor et al., 2006; Noor et al., 2008):

$$x_1, x_2, \ldots, x_{n1}, x_1^*, x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_2}, x_2^*, x_{n_2+1}, x_{n_2+2}, \ldots, x_k^*, x_n \tag{7}$$

The mean top bottom method replaced all missing values with the mean of valid surrounding values. Then the nearby point is the valid values number above and below the missing value used to compute the mean. Thus x_1^* in equation (7) will be replaced using (Yahaya et al., 2005):

$$\overline{x_1} = \frac{x_{n_1} + x_{n_1+1}}{2} \tag{8}$$

and x_2^* can be replaced by

$$\overline{x_2} = \frac{x_{n_2} + x_{n_2+1}}{2} \tag{9}$$

Multiple imputation

Markov Chain Monte Carlo

The steps of MCMC method was applied the Bayesian inference to replace the missing value by repeating the following steps:

1. The imputation (I-step): the I-step simulates the missing values for each element independently by estimating mean vector and covariance matrix. If the variables with missing values for observation i are denoted by $Y_{i(mis)}$ and the variables with observed values are denoted by $Y_{i(obs)}$, then the I-step estimates values for $Y_{i(mis)}$ from a conditional distribution $Y_{i(mis)}$ given $Y_{i(obs)}$.
2. The posterior (P-step): the P-step simulates the posterior population with mean vector and covariance matrix from the complete sample estimates. Then new estimates used in the I-step.

The current parameter estimate θ(t) at *t*-th iteration then I-step estimates $Y_{mis}^{(t+1)}$ from $p(Y_{mis}|Y_{obs}, \theta(t))$ and the P-step estimates θ(t+1) from $p(\theta|Y_{obs}, Y^{(t+1)}_{mis})$.

Exploratory analysis

Scatter plot

Plot of residuals versus predicted values to check if residuals have a constant variance:

If all points are randomly scattered showing no systematic pattern, then the distribution of residuals has a constant variance.

Normal Q-Q Plot (or P-P Plot)

If all the points fall along a straight line, then we can conclude that the residuals have a normal distribution.

Summary Statistic

Summary statistic only focused on mean, median, variance and skewness. Skewness measures to what extent a distribution of values deviates from symmetry around the mean.

- A value of zero (0) represents a symmetric or evenly balanced distribution or normal distribution.
- A positive skewness indicates a greater number of smaller values.
- A negative skewness indicates a greater number of larger values.

Multiple linear regression

MLR uses a number of independent variables to predict the dependent variable. Dependent variable must be a continuous variable. Meanwhile, independent variable can be continuous or categorical variable.

The MLR model is as stated below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \tag{10}$$

where;

Y is the dependent variable representing ozone concentrations,

X is the independent variables representing primary pollutant and meteorological variables,

Ɛ is the error term. It is assumed to be independent and have a normal distribution with zero mean and constant variance,

β_0 and $\beta_{1,\ldots,k}$ are the regression coefficient where β_0 is the constant (intercept on the Y-axis) and $\beta_{1,\ldots,k}$ are the slope of the regression line.

The estimated MLR model is as stated below:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \tag{11}$$

where;

$\hat{Y}$ is the estimated value of Y given a value of X,

X_1, X_2,…, X_k are the independent variables,

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

The goal of the MLR analysis is to determine the values of the parameters of the regression equation and then to quantify the goodness of the fit of the dependent variable Y.

Performance indicator

The performance indicator was used to identify the performance of replacing missing values techniques and to describe the performance of MLR model using coefficient of determination (R^2). The predicted and actual data were compared to select the best method for replacing missing values to model MLR.

The R^2 is described on how much of the variability in the imputed data can be explained that related to observed values or how close the points are to the line. R^2 value between 0 and 1 implying a better fit when the value is closer to 1. The equation of R^2 is given as follows [Junninen, 2004]:

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right) \right]}{\sigma_P \sigma_O} \right]^2 \qquad (12)$$

Results and Discussion

Descriptive Analysis

Table 1 shows the summary of descriptive statistics for O_3 concentration and its precursors. The annual average concentration of O_3 concentration for the four-year period is 0.021 ppm, for NO is 0.002 ppm, NO_2 is 0.0035 ppm, CO is 0.322 ppm and SO_2 is 0.001 ppm are found below than Malaysia Ambient Air Quality Guideline (MAAQG) (DOE, 2012). Meanwhile, the annual average of meteorological variables for ambient temperature is 27 ^{0}C, wind speed is 79.5 m/s and relative humidity is 5.38%. The minimum value for O_3 concentration, NO, NO_2 and SO_2 is 0.001 ppm and for CO is 0.01 ppm. The minimum temperature for this study is 20 ^{0}C which is quite low for this country.

However, the maximum value for O_3 concentration (0.095), NO (0.062), NO_2 (0.03), CO (1.940) and SO_2 (0.055) were recorded below than MAAQG. Based on the skewness, the distribution of O_3 (sk=0.889), T (sk=0.671), and WS (sk=-0.316) are relatively normal distributed and no violation of normality assumption because skewness value is between positive one and negative one. The kurtosis value for O_3 (k= 0.431), T (k=-0.653) and WS (k=0.143) are small and greater tendency to be normal distributed. Meanwhile, the distribution for NO (sk=7.945), NO_2 (sk=2.155), CO (sk=1.30), SO_2 (sk=8.463) and RH (sk=1.723) are highly skewed to the right. The kurtosis value for NO (k=135.286), NO_2 (k=7.166), CO (k=4.438), SO_2 (k=107.75) and RH (k=3.986) are also large and have the tendency for not normal distribution. According to missing data, all variables have missing data. The highest percentage of missing data is 44.1% for NO parameter.

Multiple regression analysis

Multiple linear regression modelling of ozone concentrations was developed based on different imputation techniques for replacing missing values. There is no multicollinearity problem for the three imputation techniques because TOL values are greater than 0.1 and VIF values less than 10 as shown in Table 2.

Since distribution for independent variables shows a tendency of not normally distributed, then data transformation should be applied to solve the nonnormality problem and to improve the accuracy performance of model. Transformation may be applied to either the dependent or independent variables, or both such as using the logarithm or square root of the variable (Abdul-Wahab et al., 2005; Hair et al., 2009). Hence, this study used the logarithm for transformation of independent variables. Table 2 presents the result of the analysis for the logarithmic transformation of some independent variables for each imputation technique. For liner interpolation technique, R^2 is 0.64. Mean top bottom technique is 0.636. Meanwhile, for MCMC technique the R^2 was presented the highest is 0.646 compared to other imputation technique. It can be concluded that MCMC method is the best method to replace missing value for fitting MLR model because of highest R^2 (0.649) compared to linear interpolation and mean top bottom techniques. The highest R^2 represent the best model. The best fitting regression model can make better decision and prediction about the O_3 concentrations.

The coefficient of determination, R^2, gives the proportion of the variance in the O_3 concentrations that is explained by the independent variables in the model. A significantly medium coefficient of determination for linear interpolation (R^2 = 0.640), mean top bottom (R^2 = 0.636) and multiple imputation techniques (R^2 = 0.646) explained that a few possibilities of O_3 variations were represented by the selected variables. The performance of regression model differs from the other study because of different in climatic conditions, selected independent variables and main economic activities (Dominick et al., 2012). Some previous studies shows the different of R^2 for prediction of O_3 concentration. A study from Chicago, United States using the precursors such as T, WS, $O_{3,t-1}$ and presents the R^2 is 0.600 (Comrie, 1997). Another study from Khaldiya, Kuwait used precursors for O_3 prediction such as NO, RH, SO_2, Solar, NMHC, CH_4, CO represents the R^2 is 0.686 (Abdul-Wahab et al., 2005). A study was conducted by Ghazali et al., (2010) at Shah Alam, Malaysia shows R^2 is 0.899 using the precursors such as $O_{3,t-1}$, NO_2, NMHC, T for prediction of O_3. Therefore, it can be concluded, by using different precursors in prediction of O_3 will affect the accuracy of performance model.

Table 1. Summary of descriptive statistics of O_3 concentration and its precursors

DA	O_3 (ppm)	NO (ppm)	NO_2 (ppm)	CO (ppm)	SO_2 (ppm)	T (^{0}C)	WS (m/s)	RH (%)
M	0.02	0.002	0.003	0.32	0.001	27.2	79.5	5.3
ME	0.01	0.001	0.003	0.31	0.001	25.9	81.0	4.30
MI	0.001	0.001	0.001	0.01	0.001	20.0	16.0	0.90
MA	0.09	0.062	0.030	1.94	0.055	39.5	100.0	34.8
SD	0.01	0.001	0.002	0.15	0.002	4.57	11.7	3.7
K	0.43	135.2	7.16	4.43	107.7	-0.65	0.14	3.9
SK	0.88	7.94	2.15	1.30	8.4	0.67	-0.31	1.7
MD(%)	5.6	44.1	12.7	7.2	33.4	10.1	18.6	2.0

*DA=Descriptive Analysis, M=Mean, ME=Median, MI= Minimum, MA= Maximum, SD=Standard Deviation, K=Kurtosis, SK=Skewness, MD=Missing Data

Based on the Tab. 2, MCMC technique shows that all independent variables were significant variables and ability to predict O_3 concentrations. The most precursors that contribute to predict O_3 concentration are WS and RH. The coefficients of the parameters were all statistically highly significant ($P<0.05$). These five variables used as predictor variables in modelling multiple linear regression analysis, then the model was derived as follows;

$$O_3 = 0.04 - 0.01 \log(NO) + 0.005 \log(NO_2) + 0.01 \log(CO) + 0.008 \log(SO_2) + 0.02 \log(WS) - 0.027 \log(RH) + 0.001 AT \qquad (13)$$

From Eq. 13, the highest estimated regression coefficient shows the greater effect to O_3 model is RH which is –0.027 followed by WS (0.02), NO (-0.01), CO (0.01), SO_2 (0.008), NO_2 (0.005) and AT (0.001). From the equation 13, the primary pollutants such as NO, NO_2, CO and SO_2 were influenced the O_3 values. The model also suggested the influence from meteorological parameters such as RH, WS and AT. There is positive correlation between O_3 and NO_2 indicating that the NO_2 has the ability to produce O_3 in the atmosphere (Ismail et al., 2010). The O_3 concentration also showed positive correlation with wind speed in the atmosphere. The function of wind is to transport the dispersion of both O_3 and its precursors (Kim and Guldman, 2011; Toh et al., 2013). This is supported from the previous studies that strong winds during the monsoon season are capable of transporting O_3 from the long distance to the monitoring station (Pochanart and Kreasuwun, 2001; Akimoto, 2006; Lu and Wang, 2006; Ishii et al., 2007; Al-Jeran and Khan, 2009; Shan et al., 2009).

Figure 2 shows the scatter plots of the predicted O_3 concentration against observed O_3 concentration. It was found that 64.9% of the point falls along the line suggesting the accuracy of the model developed at $R^2 = 0.649$.

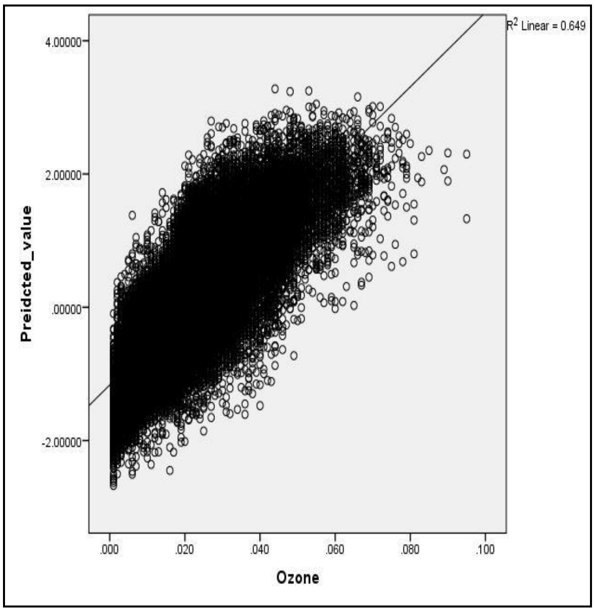

Figure 2. Scatter plot of predicted O_3 versus observed O_3 for MCMC technique

Diagnostic plots were used to check the adequacy of regression model. Figure 3 shows the histogram standardized residual and normal probability plot that indicates a relatively normal distribution. P-P plot also concludes that the residuals have a normal distribution because all of the points fall along a straight line as shown in Fig.4. Figure 5 shows the scatterplot of residuals against predicted values shows the right side pattern because the characteristic of ozone concentration is positive values. Thus, this plot indicates that residuals have a constant variance.

Table 2. MLR model for prediction O_3 using different imputation techniques

P IT		C	Log NO	Log NO_2	Log CO	Log SO_2	Log RH	AT	Log WS
LIN									
R^2	0.640								
ERC		0.045	-0.009	0.004	0.01	0.008	-0.026	0.001	0.021
SE		0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000
TOL	0.564-0.907								
VIF	1.103-1.774								
MTB									
R^2	0.636								
ERC		0.046	-0.008	0.005	0.01	0.007	-0.028	0.001	0.021
SE		0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000
TOL	0.554-0.906								
VIF	1.103-1.805								
MCMC									
R^2	0.649								
ERC		0.04	-0.01	0.005	0.01	0.008	-0.027	0.001	0.02
SE		0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000
TOL	0.517-0.820								
VIF	1.219-1.933								

*P=Pollutants, IT= Imputation Techniques, ERC= Estimated regression coefficient, SE= Standard error, C=Constant, LIN=Linear interpolation, MTB=Mean top bottom, MCMC=Markov Chain Monte Carlo

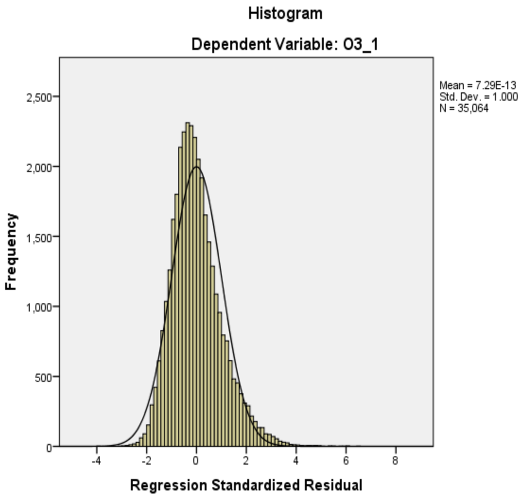

Figure 3. Histogram standardized residual

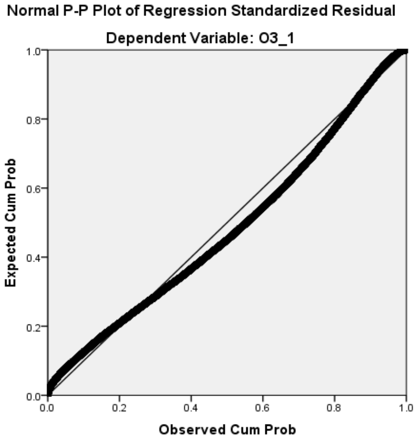

Figure 4. Normal P-P plot of standardized residual

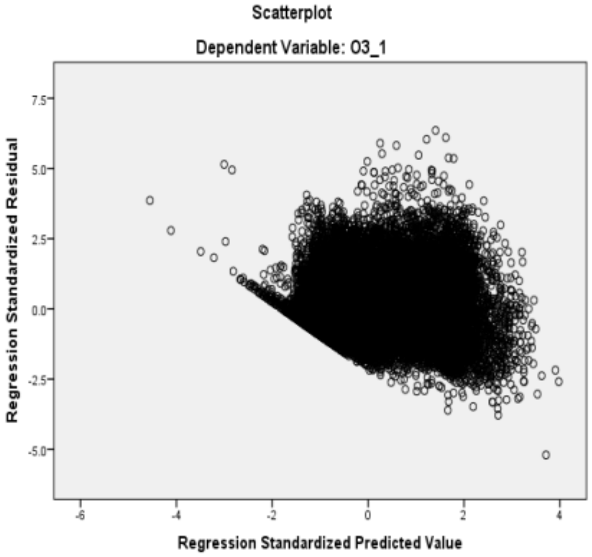

Figure 5. Scatterplot of residuals versus predicted value

Conclusions

This study discussed the comparison of single imputation and multiple imputation technique to replace missing value. Three techniques were used to replace the missing value namely, mean top bottom, linear interpolation and Markov Chain Monte Carlo. The hourly O_3 concentrations data for the duration of four years was used to compare the performance of the imputation techniques then fitted to multiple linear regression model for identifying the possible source of precursors for O_3 prediction. Based on this study, MCMC method is the best approach to replace missing values then led to statistically significant improvement in prediction accuracy. It can be concluded that Markov Chain Monte Carlo method may be the best approach to replace the missing value due to number of missing value in the data and the pattern of missing data. Imputation techniques depend on the available data and the prediction model used. The sample size and number of missing values influencing the estimation of the parameters for regression modelling.

The result of fitting the best multiple linear regression model (R^2 = 0.649) on the O_3 concentration showed that O_3 concentrations were negatively correlated with nitrogen oxide and relative humidity. Meanwhile, O_3 concentration also presented positive correlation with nitrogen dioxide, carbon monoxide, sulphur dioxide, wind speed and ambient temperature. The result of this study might be used to predict O_3 concentration at this air monitoring station. The prediction of tropospheric ozone concentrations is very important due to negative impacts of ozone on human health, climate and vegetation.

However, for further study, this study recommends to identify the effect of other precursors or primary pollutant and other meteorological variables on the O_3 concentrations, such as UV-B, sulphur dioxide, carbon monoxide, volatile organic compounds, non-methane hydrocarbons, the previous hour of O_3 concentration, and solar radiation. Another factor also recommended for investigation such as number of vehicles. Another statistical model also can be applied to improve the accuracy of the model like artificial neural network, fuzzy logic or support vector machine.

Acknowledgments

This study was funded by Ministry of Higher Education Malaysia under 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 program and Department of Environment Malaysia for their permission to utilize air quality data for this study.

References

1. Abdul-Wahab SA, Bouhamra W, Ettouney H, et al. Analysis of air pollution at Suhaiba industrial area in Kuwait. Toxicol. Environ. Chem. 2000; 78:213-232.
2. Abdul-Wahab SA. IER Photochemical smog evaluation and forecasting of short-term ozone pollution levels with artificial neural networks. Trans ICHEME Process Saf. Environ. Prot. 2001; 79: 117-128.
3. Abdul-Wahab SA, Al-Alawi SM. Assessment and prediction of tropospheric ozone concentrations levels using artificial neural networks. Environ. Modell. Software 2002; 17: 219-228.
4. Abdul-Wahab SA, Bakheit CS, Al-Alawi SM. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Modell. Software 2005; 20: 1263-1271.
5. Akimoto H. Long-range transport of ozone in the East Asian Pacific Rim Region. J. Geophys. Res. 2006; 101: 1999-2010.
6. Al-Jeran HO, Khan AR. The effect of air pollution on ozone layer thickness in troposphere over the state of Kuwait. Am. J. Environ. Sci. 2009; 5: 230-237.
7. Awang NR, Ramli NA, Yahaya AS, Elbayoumi M. Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas. Atmos. Pollut. Res. 2015; 6: xx-xx.

8. Borrego C, Tchepel O, Costa AM, et al. Emission and dispersion modelling of lisbon air quality at local scale. Atmos. Environ. 2003; 37: 5197-5205.

9. Chapra SC and Canale RP. Numer. Methods Eng. Singapore:McGraw-Hill; 1998.

10. Chelani AB. Study of extreme CO, NO_2 and O_3 concentrations at traffic site in Delhi: Statistical persistence analysis and source identification. Aerosol Air Qual. Res. 2012; 13: 377-384.

11. Clark TG, Bradburn MJ, Love SB and Altman DG. Survival analysis part IV: Further concepts and methods in survival analysis. Br. J. Cancer 2003; 89: 781-786.

12. Comrie AC. Comparing neural networks and regression models for ozone forecasting. J. Air Waste Manage. Assoc. 1997; 47(6): 653-663.

13. Department of Environment (DoE), Malaysia. Malaysia Environmental Quality Report 2012. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia; 2012.

14. Dominick D, Latif MT, Juahir H, et al. An assessment of influence of meteorological factors on PM_{10} and NO_2 at selected stations in Malaysia. Sustain. Environ. Res. 2012; 22(5): 305-315.

15. Dong Y, and Peng C-YJ. Principled missing data methods for researcher. SpringerPlus 2013; 2: 222.

16. Environmental Sustainability Index (ESI), Benchmarking National Environmental Stewardship, 2005.

17. Finlayson-Pitts BJ, Pitts JN. Atmospheric Chemistry Fundamentals and Experimental Techniques. Wiley, New York; 1986.

18. Ghazali NA, Ramli NA, Yahaya AS, et al. Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. Environ. Monit. Assess. 2010; 165: 475-489.

19. Gómez-Carracedo MP, Andrade JM, López-Mahlía P, et al. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom. Intell. Lab. Syst. 2014; 134: 23–33.

20. Go'mez-Sanchis J, Martin-Guerrero JD, Soria-Olivas E, et al. Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. Atmos. Environ. 2006; 40: 6173-6180.

21. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am. J. Epidemiol. 1995; 142: 1255-64.

22. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis a global perspective 7th edn, Prentice Hall, 2009; pp 34.

23. Hawthorne G, and Elliot P. Imputing cross-sectional missing data: comparison of common techniques. Aust. NZ. J. Psychiat. 2005; 39: 583-590.

24. Heitjan F, Little RJA. Multiple imputation for the fatal accident reporting system. Appl. Stat. 1991; 40: 13-29.

25. Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. J. Am. Stat. Assoc. 2011; 55: 244-254.

26. Ingsrisawang L, Potawee D. Multiple imputation for missing data in repeated measurements using MCMC and copulas. Proceedings of the International Multiconference of Engineers and Computers Scientists Vol II, March 14-16 2012, Hong Kong.

27. Ishii S, Bell JNB, Marshall FM. Phytotoxic risk assessment of ambient air pollution on agricultural crops in Selangor state, Malaysia. Environ. Pollut. 2007; 150: 267-279.

28. Ismail AS, Latif MT, Azmi SZ, Juneng L, Jemain AA. Variation of surface ozone recorded at the eastern coastal region of the Malaysian Peninsula. Am. J. Environ. Sci. 2010; 6(6): 560-569.

29. Jerez JM, Molina I, GarcÍa-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in real breast cancer problem. Artif. Intell. Med. 2010; 50: 105-115.

30. Junninen H, Niska H, Tuppurrainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality dataset. Atmos. Environ. 2004; 38: 2895-907.

31. Kim Y, and Guldman JM. Impact of traffic flows and wind directions on air pollution concentrations in Seoul, Korea. Atmos. Environ. 2011; 45: 2803-2810.

32. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. Stat. Med. 1995; 14: 1913-1925.

33. Little RA. Regression with missing x's; A review. J. Am. Stat. Assoc. 1992; 87: 1227-37.

34. Little RA, Rubin BB. Statistical Analysis with Missing Data 2nd edn, Wiley, New York, 2002; pp 4-22

35. Lu WZ, and Wang XK. Evolving trend and self-similarity of ozone pollution in central Hong Kong ambient during 1984-2002. Sci. Total Environ. 2006; 357: 160-168.

36. Majlis Perbandaran Kemaman (MPK). Retrieved on 30th April 2015 from http://mpk.terengganu.gov.my/ms/web/guest/geografi-mpk

37. Neter J, Wasserman W, Kutner MH. Applied Linear Regression Models. Richard D. Irwin Inc., Homewood, 1983; pp. 547.

38. Noor MN, Yahaya AS, Ramli NA, Mustafa Al Bakri AM. Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 2008; 34: 341-345.

39. Noor MN, Yahaya AS, Ramli NA, Mustafa Al Bakri AM. The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique. J. Eng. Res. Educ. 2006; 3: 96-105.

40. Noor MN, Yahaya AS, Ramli NA, Mustafa Al Bakri AM. Filling missing data using interpolation methods: study on the effect of fitting distribution. Key Eng. Mater. 2014; 5: 889-895.

41. Özbay B, Keskin GA, et al. Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models. Ecol. Inf. 2011; 6: 242-247.

42. Plaia A, and Bondi AL. Single imputation method of missing values in environmental pollution data sets. Atmos. Environ. 2006; 40: 7316-7330.

43. Pochanart P, and Kreasuwun J. Tropical tropospheric ozone observed in Thailand. Atmos. Environ. 2001; 35: 2657-2668.

44. Ramli NA, Ghazali NA, Yahaya AS. Diurnal fluctuations of ozone concentrations and its precursors and prediction of ozone using multiple linear regression. Malaysian J. Environ. Manage. 2010; 11(2): 57-69.

45. Saunders SM, Jenkin ME, Derwent RG, et al. Site of a master chemical mechanism for use in tropospheric chemistry models. Atmos. Environ. 1997; 31: 12-49.

46. Schafer JL. Analysis of Incomplete Multivariate Data. Chapman and Hall, New York, 1997.

47. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. Comput. Stat. Data Anal. 1996; 22: 425-446.

48. Shan W, Yin Y, Lu H, Liang S. A meteorological analysis of ozone episodes using HYSPLIT model and surface data. Atmos. Res. 2009; 93: 767-776.

49. Song F, Shin JY, Atresino RJ, and Gao Y. Relationship among the springtime ground-level No_x, O_3, and NO_2 in the vicinity of highways in the US East Coast. Atmos. Pollut. Res. 2011; 2: 374-383.

50. Sousa SIV, Martins FG, Alvin-Ferraz MCM, Pereira MC. Multiple linear regression and artificial neural networks based on principal component to predict ozone concentration. Environ. Modell. Software 2007; 22: 97-103.

51. Tiwary A, Colls J. Air Pollution: Measurement, Modelling and Mitigation, Routledge, London, 2010.

52. Toh YY, Fook LS, and Von Glasow R. The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah Rata, Malaysia. Atmos. Environ. 2013; 70: 435-446.

53. Vach W. Logistic Regression with Missing Values in the Covariates. New York: Springer, 2004.

54. Wang SW, Georgopoulus PG. Observational and mechanistic studies of tropospheric studies of tropospheric ozone precursors relations: photochemical models performance evaluation with case study. Technical Report ORC-TR99-03, 2001.

55. Yahaya AS, Ramli NA, Yusof NF. Effects of estimating missing values on fitting distributions: International Conference on Quantitative Sciences and its Applications, 6-8 December 2005, Penang, Malaysia: Universiti Utara Malaysia.

56. Yuan Y. Multiple imputation using SAS software. J Stat Software 2011; 45(6): 1-25.