IMPUTATION METHODS FOR FILLING MISSING DATA IN URBAN AIR POLLUTION DATA FOR MALAYSIA

Nur Afiqah ZAKARIA
School of Environmental Engineering, University Malaysia Perlis (UniMAP), Kompleks Pusat Pengajian Jejawi 3, 02600 Arau, Perlis, e-mail: nurafiqahzakaria15@gmail.com

Norazian Mohamed NOOR
Senior Lecturer (PhD), School of Environmental Engineering, University Malaysia Perlis (UniMAP), Kompleks Pusat Pengajian Jejawi 3, 02600 Arau, e-mail: norazian@unimap.edu.my

Abstract. The air quality measurement data obtained from the continuous ambient air quality monitoring (CAAQM) station usually contained missing data. The missing observations of the data usually occurred due to machine failure, routine maintenance and human error. In this study, the hourly monitoring data of CO, O3, PM10, SO2, NOx, NO2, ambient temperature and humidity were used to evaluate four imputation methods (Mean Top Bottom, Linear Regression, Multiple Imputation and Nearest Neighbour). The air pollutants observations were simulated into four percentages of simulated missing data i.e. 5%, 10%, 15% and 20%. Performance measures namely the Mean Absolute Error, Root Mean Squared Error, Coefficient of Determination and Index of Agreement were used to describe the goodness of fit of the imputation methods. From the results of the performance measures, Mean Top Bottom method was selected as the most appropriate imputation method for filling in the missing values in air pollutants data.

Key words: air pollution, missing data, imputation methods, multiple imputation.

1. Introduction
Air pollution is the condition where the air is contaminated with foreign substances or the substances themselves. According to Md Razak et al. (2013) air pollution could be aerosols or gases with particles or liquid droplets suspended in the air that might change the natural composition of the atmosphere, would be dangerous to human, animals and plants and also caused destruction to land and water bodies.

In the early times when resources were abundant and development pressures minimal, Malaysia paid little attention to environmental issues (Afroz et al., 2003), with severe consequences including air pollution (Lilieveld et al., 2001; Sastry, 2002), a common problem to developing countries (Smith et al., 2000), with adverse consequences for the human population (Bruce et al., 2000) and agriculture (Ishii et al., 2004, 2007). The major sources of pollution are traffic (Han and Naeher,
2006), industrial and power plants and open burning (Azmi et al., 2010), but the most important drivers are demography (Cole and Neumayer, 2004) and urbanization (Dominick et al., 2012). The effects are aggravated by the tropical environment (Azizi et al., 1995). Nevertheless, legislative measures were able to improve the situation (Awang et al., 2000).

Particulate matter (PM$_{10}$) was recorded as the most prevailing pollutant in Southeast Asia Region. The particulate matter have an important role both in atmosphere transparency and air purity; they reduce the quality of the environmental factors (Tudose et al., 2015). There are three main contributors of PM$_{10}$ in Malaysia i.e. vehicular emissions, power stations and industrial sectors. Seventy six percent (4585 tonnes) of PM$_{10}$ emission in Malaysia is from motor vehicles whereas power plant emission impacted fifteen percent (15 tonnes) and only four percent (4 tonnes) caused by industrial sector. Hence, urban areas with higher amount of vehicles contribute more air pollution compared to rural areas. Furthermore, high PM$_{10}$ concentrations were detected during dry season or also known as summer monsoon (June to September) due to the vast quantities of smoke releases by biomass burning from regional sources (Noor et al., 2015).

Air quality monitoring of air pollution is very important. This is because, the data from the air quality monitoring will show or detect any significant pollutant concentration. In Malaysia, the Department of Environmental (DOE) is responsible for monitoring the status of air quality, however, this operation is privatized to Alam Sekitar Malaysia Sdn. Bhd. (ASMA).

The data of air quality obtained from the CAAQM stations usually contained missing data that caused bias due to systematic error between observed and unobserved (Noor et al., 2008). Missing data was a very frequent problem happened in many scientific fields above all in environmental researches (Xia et al., 1999).

The missing data would give impact to the result of statistical analysis depending on the mechanism that made the data to be missed and or the way the data analyst deal with them (Devere, 2006; Plaia and Bondi, 2006). Furthermore, missing observations hindered the ability to make exact conclusion or interpretations about the observation (Noor et al., 2015).

There are a few ways on treating the missing observations. One of the most efficient way on handling the missing data problem is by using the imputation method (Little and Rubin, 1987; Abd Razak et al., 2014). There are two types of imputation method that are single and multiple imputation. Single imputation is filled in one value for each of missing values. Single imputation methods had many appealing features (Plaia and Bondi, 2006; Junninen et al., 2004). The main advantage of this method is that the standard of complete data methods could be applied directly and the substantial effort required creating imputations needs to be carried out only once. Multiple imputation method was done by replacing the missing values with multiple simulated values to reflect properly the uncertainty that attached to the missing data (Junninen et al., 2004). This methods had been supported as a statistically sound approach, however the use has been limited to social and medical science. (Junninen et al., 2004).
The main objective of this research was to find the most appropriate method in filling the missing observations in air pollutant data. A few single imputation methods and multiple imputation method were adopted and the performances of all methods were compared using performance measures.

2. Methodology

2.1. Data

In this study, hourly averaged of 5 air pollutants data and 3 meteorological data in Malacca, Malaysia for 2008 were selected. The total observation of these 8 data was 70272 and the total missing data was 2960 (4.212 %). The highest missing observation was found out to be \( O_3 \) concentration with 624 missing observations. Overall, for the ambient air quality, the daily mean concentration of \( CO, O_3, SO_2, NO_2 \) and \( PM_{10} \) were not exceeding the limit stated in the Malaysia Air Quality Guideline (MAAQG).

Table 1 shows the descriptive statistics for all air pollutants in Malacca (2008). All air pollutants concentration except for humidity, the mean was higher than the median. It indicated that the pollutant distributions were skewed to the right and the extreme events occurred. The mean value for humidity parameter was lower than the median value which meant that the pollutant distributions was skewed to the left and the skewness value would be negative.

Table 2 shows the mean percentages of the length of gap (in hour) for all air pollutants data. 1-hour gap of missing observation was recorded highest with the value of 92% whereas for missing gap between 1h and 3h, the value reduced drastically to only 4.7%. The higher percentage of missing data in the length of gap more than 15h was due to the missing observations of three parameters in the gaps of between 51h to 54h.

2.2. Simulation of missing data

The dataset were simulated into four percentages of simulated missing data that were 5%, 10%, 15% and 20%. The purpose of simulation was to evaluate the precision of the imputation technique applied (Noor et al., 2008). This simulation was done by using SPSS version 21 for Windows. The percentages of selected cases were only around the specific percentage because this procedure produced an independent pseudo-random decision.
2.3. Imputation methods

Four imputation methods were used to fill in the simulated missing data. The methods used were Mean Top Bottom, Nearest Neighbour, Linear Regression and Multiple Imputation Method. Multiple Imputation method was carried out to compare the performances of single imputation methods with Multiple Imputation methods.

2.3.1. Mean Top Bottom

Mean Top Bottom or also known as Mean Before After method was the average of one existing observation on the top and the bottom of the missing values (Noor et al., 2015). The equation was written as (Noor et al., 2015):

\[ y^*_i = \frac{1}{2} (y_{n1} + y_{n1+1}) \] (1)

2.3.2. Nearest Neighbour

Nearest Neighbour was the method to replace the missing data with the nearest value to the missing datum (Noor et al., 2015). Nearest Neighbour imputation was the simplest method available, in that the end points of the gaps were used as estimates for all the missing values. The equation is (Junninen et al., 2004);

\[ y = y_1 \text{ if } x \leq x_1 + \frac{(x_2 - x_1)}{2} \]
\[ y = y_2 \text{ if } x > x_1 + \frac{(x_2 - x_1)}{2} \] (2)

2.3.3. Linear Regression

Linear Regression is a model that has relationship between the two variables by fitting a linear equation to the observed data. The missing value of the data will be replaced by regression of the unobserved variables against observed one for that dataset (Noor et al., 2015). The equation is represented as (Noor et al., 2015):

\[ y^*_i = \beta_0 + \beta_1 x_i + \varepsilon_i \] (3)

2.3.4. Multiple Imputation

Multiple Imputation methods is the method that generate multiple simulated values for each of the missing data. Multiple imputation by Markov chain Monte Carlo (MCMC) was used in this study and it was conducted by using SPSS. MCMC is used to generate pseudorandom draws from multidimensional dataset and then, complicated probability distributions were generated via Markov chains (Schafer, 1997).

2.4. Performance indicators

The goodness of fit of each of the imputation methods used in this research were described by using several performance indicators. Four performance indicators were used in this research. These performance measures can be divided into two groups that are the error and the performance measures. For error indicator, the bigger the value, the greater the error. Two tests were carried out namely mean absolute error (MAE) and root mean squared error (RMSE). Performance measures indicate that the closer the value to one, the better the methods. For performance measure, the indicators are coefficient of determination (R²) and index of agreement (d²):
agreement ($d_2$). Table 3 shows the formula for performance indicators.

Where $N$ is the number of imputation, $O_i$ is the observed data points, $P_i$ the imputed data point, $\bar{O}$ is the average of imputed data, $\bar{P}$ is the average of observed data, $\sigma_P$ is the standard deviation of the imputed data and $\sigma_O$ is the standard deviation of the observed data.

4. Results and discussion

4.1. Characteristics of the simulated data

Table 4 shows the percentages of the gap length (in hour) for different percentages of simulated missing observations. The simulated missing data were constructed according to the real missing data trend as shown in Table 2. The maximum number of gaps were limited to 5 hour due to the significant percentages of the missing gap were between 1 h to 5h (Table 2). Hence, the increment of the gap length percentages are gradually increased as the percentages of simulated missing data increases.

Table 4. The percentage of the gap length (hour) for each of the simulated missing data.

<table>
<thead>
<tr>
<th>Length of Gap (h)</th>
<th>Percentage of gap length according to the simulated missing data</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>10%</td>
<td>15%</td>
</tr>
<tr>
<td>1</td>
<td>89.524</td>
<td>81.713</td>
</tr>
<tr>
<td>3</td>
<td>0.683</td>
<td>2.141</td>
</tr>
<tr>
<td>4</td>
<td>0.442</td>
<td>1.098</td>
</tr>
<tr>
<td>5</td>
<td>0.349</td>
<td>0.304</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.342</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

4.2. The best imputation method

Overall, based on the results in Table 5, it shows that the error (MAE and RMSE) would be increased and the measure of performances ($R^2$ and $d_2$) decreased as the percentages of simulated missing data increases. This was consistence with the statement reported by Junger and de Leon (2015) that the validity of the estimates would be decreased when the missing values increased.

The best imputation method for estimating the simulated missing data was Mean Top Bottom (MTB) method. This was because MTB method gave the smallest values of MAE and RMSE and the highest values for $R^2$ and $d_2$ in almost all parameters and percentages of the simulated missing data. This finding was consistent with the study reported by Noor (2006) that MTB was the best imputation method for filling the missing data because this method is able to give the smallest error for all percentages of missing data. The second best imputation method for estimating the simulated missing data was Nearest Neighbor (NN) method. This method also performed better than Multiple Imputation (MI) method for almost all parameters and percentage of missing data. The worst method was Linear Regression (LR) method. This method contributed high error value from the indicators of MAE, NAE and RMSE and failed to fit the simulated missing data with very low values of PA, $R^2$ and $d_2$.

Figure 1 shows the scatter plots of the observed and the predicted data for 5%, 10%, 15% and 20% of the CO observations. The predicted data in this figures was imputed by using MTB methods. $R^2$ in these graphs shows the variability of predicted data (y-axis) that has been clarified by observed data (x-axis). According to Siegel (2012), the larger the value of $R^2$, the better the prediction because it indicated that x and y has stronger relationship. Based on Figure 1 (a), (b), (c) and (d), it shows that the values of $R^2$ for all percentages were large and these proved that MTB method is able to give good estimations for the air pollutants data.
Table 5. The performances of each method for every percentages of simulated missing data.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Performance Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
</tr>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>MTB</td>
<td>1.22</td>
</tr>
<tr>
<td>NN</td>
<td>1.40</td>
</tr>
<tr>
<td>LR</td>
<td>3.18</td>
</tr>
<tr>
<td>MI</td>
<td>2.74</td>
</tr>
</tbody>
</table>

The $R^2$ values indicate that the predicted values were almost close to the observed values. The values of $R^2$ also decreased when the percentages of missing data increased.

5. Conclusion and recommendation

Hourly averaged of 5 air pollutants data and 3 meteorological data in Bachang, Malacca in 2008 was used. The total observation of these 8 data is 70272 and the total missing data is 2960 (4.212 %). The percentage of total missing observation for all data is 21.081% ($O_3$), 18.885% ($SO_2$), 17.635% ($NO_2$), 17.365% ($NO_x$), 14.595% (CO), 0.734 % (ambient temperature) and 0.304 % (humidity). All data had at least of 1 hour of missing observation and calibration is one of the factors that contributed to the incomplete data.

The longest gap of missing observations was monitored in $SO_2$ with 1 occurrence of 55 hours missing values. Overall, for the ambient air quality, the daily mean concentration of CO, $O_3$, $SO_2$, $NO_2$ and $PM_{10}$ was not exceeding the limit that stated in the In Malaysia Air Quality Guideline.

In this study the data set was simulated into four percentages of missing data. The percentages of the simulated missing are 5%, 10%, 15% and 20%. The simulated of missing data were generated by using SPSS software for Windows. The range number of missing data for 5% was 435 to 460, 10% was 870 to 878, 15% were 1414
Urbanism

Imputation methods for filling missing data in urban air pollution data for Malaysia • N. A. Zakaria, N. M. Noor

to 1325 and for 20% was 1740 to 1753. The missing data were simulated until 20% because of the percentages of missing data recorded in Malaysia was not exceeded 20%.

Four imputation methods were used to estimate the all percentages of simulated missing data. The methods used are Mean Top Bottom (MTB), Nearest Neighbor (NN), Linear Regression (LR) and Multiple Imputation (MI). Four performance measures were calculated to determine the goodness of fit for these imputation methods. The best imputation method obtained was Mean Top Bottom method, meanwhile Linear Regression is the worst method that can be used to impute the missing observations in air pollution data. Nearest Neighbour method performed better than Multiple Imputation methods but less efficient compared to MTB.

Acknowledgement

This study is supported by School of Environmental Engineering, UniMAP and their efforts on behalf of the project are greatly appreciated.

REFERENCES


Md Razak M. I., Ahmad I., Bujang I., Talib H., Ibrahim Z. (2013), Economics of Air Pollution

Noor N. M. (2006), The replacement of missing values of continuous air pollution monitoring data using various imputation technique, Universiti Sains Malaysia, Perlis, Malaysia.


Received: 4 March 2016 • Revised: 16 August 2016 • Accepted: 19 November 2016

Article distributed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND)