Journal of Engineering Research & Education
Vol. 3, 2006 (96-105)

The Replacement of Missing Values of Continuous Air Pollution Monitoring Data Using Mean Top Bottom Imputation Technique

NORAZIAN MOHAMED NOOR[1], AHMAD SHUKRI YAHAYA[2], NOR AZAM RAMLI[2], MOHD MUSTAFA AL BAKRI ABDULLAH[3]

[1] School of Environmental Engineering
[1]norazian@kukum.edu.my

[3]School of Material Engineering,
Kolej Universiti Kejuruteraan Utara Malaysia,
P.O Box 77, d/a Pejabat Pos Besar,01007 Kangar, Perlis.
[3]mustafa_albakri@kukum.edu.my

[2]School of Civil Engineering,Universiti Sains Malaysia,
Engineering Campus,14300 Nibong Tebal, Pulau Pinang.
[2]shukri@eng.usm.my ceazam@eng.usm.my

Received : 14 January 2006 / Accepted : 22 November 2006

ABSTRACT

Air pollutants data such as PM_{10}, carbon monoxide, sulphur dioxide and ozone concentration were obtained from automated monitoring stations. These data usually contain missing values that can cause bias due to systematic differences between observed and unobserved data. Therefore, it is important to find the best way to estimate these missing values to ensure that the data analyzed are of high precision. This paper focuses on the usage of mean top bottom imputation technique to replace the missing values. Three performance indicators were calculated in order to describe the goodness of fit of this technique. In order to test the efficiency of the method applied, PM_{10} monitoring dataset for Kuala Lumpur was used as case study. Three distributions that are Weibull, gamma and lognormal were fitted to the datasets after replacement of missing values using mean top bottom method and performance indicators were calculated to describe the qualities of the distributions. The results show that mean top bottom method gives very good performances at low percentage of missing data but the performances slightly decreased at higher degree of complexity. It was found that gamma distribution is the most appropriate distribution representing PM_{10} emissions in Kuala Lumpur.

INTRODUCTION

Air quality monitoring is carried out to detect any significant pollutant concentrations, which have possible adverse effects on human health. However, such analysis is frequently interrupted by the large proportions of missing data. Missing data is incomplete data matrices that cause wrong interpretation of monitoring activities. This might be due to machine failure, routine maintenance, changes in the siting of monitors, human error and other factors (Hawthorne and Elliott, 2004).

The easiest and most common approach to deal with missing values is to completely ignore the missing values and continue with the complete datasets. However, this method is only applicable when the percentage of missing values is low (Little and Rubin, 2002). When there are large proportion of missing values, ignoring the missing observations will cause biased estimations since it assumes that the loss of data takes place in a completely random way (Yahaya *et al.,* 2005). The method of estimating missing values is called imputation technique. Hawthorne and Elliot (2005) had conducted the comparison study of common techniques in imputing cross-sectional missing data. In this study, six procedures for handling missing data were considered viz., listwise deletion; item (or vertical) mean substitution, two levels of person mean substitution, regression imputation and 'hot deck'. Twisk and Vente (2001) performed the study on how to deal with missing data in longitudinal studies. They applied cross sectional and longitudinal imputation methods to replace the missing data. Junninen *et al.* (2004) provides a comprehensive study of imputation techniques that can be used in air quality datasets.

This paper discusses the use of mean top bottom method to substitute the missing values for the data of PM_{10} monitoring datasets. Three performance indicators were calculated to determine the quality of the method used. For case study analysis, PM_{10} monitoring dataset for Kuala Lumpur was used. Weibull, gamma and lognormal distributions were fitted to the datasets after replacement of missing values by using mean top bottom method. Finally, the goodness of fit of the distributions was determined using performance indicators.

EXPERIMENTAL PROCEDURES

Data

*F*or the first stage of analysis that is simulation of missing data, the annual hourly monitoring datasets for PM_{10} in Seberang Perai, Penang was selected. The test dataset consisted of particulate matter (PM_{10}) concentration on a time-scale of one per hour (hourly averaged). For case study analysis, dataset for Kuala Lumpur monitoring stations was chosen. The proposed dataset consisted of hourly particulate matter (PM_{10}) concentration. PM_{10} was selected because it is the most prevailing pollutant recorded in many areas in Malaysia (Department of Environment, 2004). The main contributors for PM_{10} are motor vehicles exhaust, power generation and industrial processes. Table 1 below describes the datasets.

Table 1 Descriptive statistics of PM_{10} datasets.

	Seberang Perai, Penang	**Kuala Lumpur**
Number of observations	8757	8567
Number of missing observations	3	193
Mean	77.0	77.2
Standard deviation	58.5	0.31
Minimum	8	9
Maximum	718	314

Monitoring record for Seberang Perai station contained only three missing observations (0.03%). Hence, the missing values were omitted in order to get the complete dataset (Olinsky *et al.,* 2002). PM_{10} monitoring dataset for Kuala Lumpur station consisted 2% of missing values. These missing values will be replaced by using mean top bottom method.

Simulation of missing data

Five randomly simulated missing data patterns were used for evaluating the accuracy of imputation techniques in different missing data conditions. The simulated data patterns are divided into three degree of complexities that are small, medium and large. The small degree of complexity consists of 5% and 10% missing data, for medium complexity the percentages of missing values are 15% and 25% whereas large complexity consists of 40% missing data. The patterns of missing data simulation are represented in Table 2. A random sample of approximately the specified percentage of missing data conditions was generated using SPSS 11.5 for Windows.

Table 2 The patterns of missing data simulation.

Degree of Complexities	**Percentage of Missing Data (%)**
Small	5 10
Medium	15 25
Large	40

Computational Method

Assume that y_1, y_2,…,y_n be a times series data with n observations and there are k missing values denoted by $y_1^*, y_2^*, ..., y_k^*$ where k < n. Thus, the observed data with missing values can be represented as follows (Yahaya *et al.,* 2005):

$$y_1, y_2, ..., y_{n1}, y_1^*, y_{n1+1}, y_{n1+2}, ..., y_{n2}, y_2^*, y_{n2+1}, y_{n2+2}, ..., y_k^*, y_n \quad (1)$$

Therefore, the first missing value occurs after n_1 observations, the second missing value occur after n_2 observations and so on. Note that there might be more than one consecutive missing observation. The method is described below:

(i) Mean Top Bottom Method

This method replaces all missing values with the mean of the datum above the missing value and one datum below the missing value. Thus for the data in equation (1), y_1^* will be replaced by (Yahaya *et al.,* 2005):

$$\bar{y} = \frac{y_{n1} + y_{n1=1}}{2} \quad (2)$$

and y_2^* will be replaced by

$$\bar{y} = \frac{y_{n2} + y_{n2+1}}{2} \quad (3)$$

Performance Indicators

Several performance indicators were used to describe the goodness of fit of the mean top bottom methods used in this research. Three performance indicators were used that are Root Mean Square Error (RMSE), Prediction Accuracy (PA) and coefficient of determination (R^2).

Assume that N is the number of imputations, O_i the observed data points, P_i the imputed data point, $\bar{P}$ is the average of imputed data, $\bar{O}$ is the average of observed data, s_P is the standard deviation of the imputed data and s_O is the standard deviation of the observed data. Then, the six performance indicators are given in Table 3.

Table 3 Performance indicators.

Performance Indicators	Formula
Root mean square error *(RMSE)*	$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} [P_i - O_i]^2 \right)^2$
Prediction Accuracy *(PA)*	$PA = \sum_{i=1}^{N} \frac{\left[\left(P_i - P \right) \left(O_i - O \right) \right]}{(N-1)\, \sigma_P \sigma_O}$
Coefficient of determination (R^2)	$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[\left(P_i - \overline{P} \right) \left(O_i - \overline{O} \right) \right]}{\sigma_P \sigma_O} \right]^2$

Distributions

In order to model the dataset from Kuala Lumpur station, three common distributions were used. The distributions are Weibull, lognormal and gamma distributions. These three distributions are explained below.

Assume that for the three distributions, α is the shape parameter and β is the scale parameter. Then the Weibull probability density function (pdf) with two parameters is given by (Evans *et al.,* 2000):

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} [P_i - O_i]^2 \right)^{\frac{1}{2}} \tag{4}$$

and the cumulative distribution function (cdf) takes the form

$$PA = \sum_{i=1}^{N} \frac{\left[\left(P_i - \overline{P} \right) \left(O_i - \overline{O} \right) \right]}{(N-1)\, \sigma_P \sigma_O} \tag{5}$$

The probability density function (pdf) for the two parameter lognormal distribution is given as:

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[\left(P_i - \overline{P} \right) \left(O_i - \overline{O} \right) \right]}{\sigma_P \sigma_O} \right]^2 \quad (6)$$

The cumulative distribution function (cdf) equation for lognormal distribution is as follows:

$$f(x,\alpha,\beta) = \frac{\alpha}{\beta} \left(\frac{x}{\beta} \right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta} \right)^{\alpha} \right], x > 0, \alpha > 0, \beta > 0 \quad (7)$$

The probability density function (pdf) for the two parameter gamma distribution is as follows:

$$F(x,\alpha,\beta) = 1 - \exp\left[-\left(\frac{x}{\beta} \right)^{\alpha} \right], x > 0, \alpha > 0, \beta > 0 \quad (8)$$

The cdf for the gamma distribution is as follows:

$$f(x,\alpha,\beta) = \frac{1}{x\alpha\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{\ln(x) - \beta}{\alpha} \right)^2 \right], x > 0, \alpha, \beta > 0 \quad (9)$$

Three performance indicators were calculated to describe the goodness of fit for each selected distribution; the performance indicators applied are Root Mean Square Error (RMSE), Prediction Accuracy (PA) and coefficient of determination (R^2). The equations for these performance indicators selected are shown in Table 3.

RESULTS AND DISCUSSION

Analysis

The descriptive statistics for all complexity of missing values were shown in Table 4.

Table 4 Descriptive statistics for simulated missing data.

Percentage of missing data		5 %	10%	15%	25%	40%
Valid data		8275	7886	7425	6547	5233
Missing data		479	871	1332	2210	3524
Mean		76.9	76.87	77.14	77.4	77.2
Standard Deviation		58.0	57.8	57.5	57.9	58.7
Skewness		3.55	3.54	3.54	3.51	3.57
Kurtosis		22.2	22.2	21.9	21.4	22.6
Percentiles	20	38.0	38.0	38.0	38.0	38.0
	40	55.0	55.0	56.0	56.0	55.0
	60	75.0	75.0	75.0	75.0	74.0
	80	104.0	104.0	104.0	104.0	104.0
	100	718.0	718.0	715.0	715.0	718.0

From the table, it can be seen that, although there are differences in the amount of data, the analyses are producing almost similar results for all percentage of missing values. For every percentile, there are not many differences even though the percentage of missing values increases. This occurrence is due to three causes that are (1) the random number generated in producing the simulated missing values patterns, (2) the availability of large number of observation with the same range and (3) the largest observation was not omitted from all percentage of simulated missing data.

Table 5 below, presents the values of every performance indicators applied to describe the goodness of fit for mean top bottom method in estimating simulated missing values. From the table, it shows that the highest error is obtained at 25% simulated missing values whereas the smallest error is at 15%. This is maybe due to the random number deletion during producing the simulated missing values. For describing the qualities of prediction made, PA and R^2 were calculated. The PA and R^2 values ranges from 0 to 1, with higher values indicate the better fit. The results show that all the values of PA and R^2 decreased with the increased of percentage of simulated missing values.

Table 5 The values of performance indicators for mean top bottom method.

Degree of Complexity (%)	Performance indicators		
	Root Mean Square Error *(RMSE)*	Prediction Accuracy *(PA)*	Coefficient of Determination *(R^2)*
5	23.71	0.93	0.87
10	24.35	0.93	0.86
15	23.27	0.93	0.86
25	29.12	0.87	0.77
40	27.79	0.88	0.77

Case Study Analysis

Table 6 presents the values of α (shape parameter) and β (scale parameter) of Weibull, lognormal and gamma distributions for data in Kuala Lumpur after replacement of missing values using mean top bottom method.

Table 6 Parameters values for distribution applied.

Distribution	α	β
Weibull	86.85	2.84
Lognormal	4.28	0.38
Gamma	7.53	10.29

Figure 1 shows the cdf plots of three applied distributions for data in Kuala Lumpur whereas Table 7 shows the performance indicators used to describe the goodness of fit for every distribution. Cdf plots for Kuala Lumpur indicates that gamma distribution fit the observed data very well compared to other types of distribution. From Table 7, it can be seen that cdf plot of gamma distribution for Kuala Lumpur gives the smallest error that is 1.95 and highest values of PA and R^2 0.998 and 0.995 respectively. Lognormal distribution fits the observation data better than Weibull distribution which indicates better value for all performance indicators. Therefore, gamma distribution is the most appropriate distribution representing PM_{10} emission data in Kuala Lumpur after replacement of missing values using linear interpolation technique.

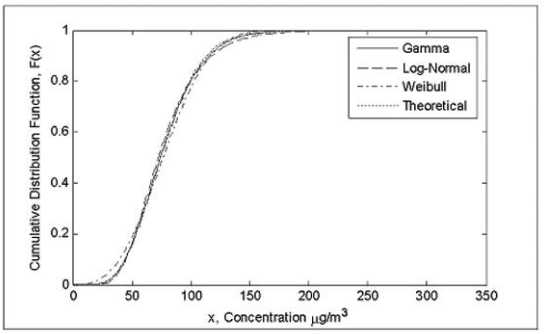

Figure 1. Cdf plots of various distributions for data from Kuala Lumpur after replacement of missing values using mean top bottom method.

Table 7 Performance indicators for Weibull, gamma and lognormal distributions for PM_{10} monitoring data in Kuala Lumpur.

Performance Indicators	***RMSE***	***PA***	R^2
Weibull distribution	4.77	0.987	0.974
Lognormal distribution	3.23	0.996	0.992
Gamma distribution	1.95	0.998	0.995

CONCLUSIONS

Mean top bottom method was used to estimate five randomly simulated missing data patterns. The simulated data patterns are divided into three degree of complexities that are small, medium and large. The small degree of complexity consists of 5% and 10% missing data, for medium complexity the percentages of missing values are 15% and 25% whereas large complexity consists of 40% missing data. Overall, it was found that mean top bottom method gives good performances but the performances decreased slightly at higher degree of complexity. In order to test the efficiencies of the method used, PM_{10} monitoring datasets for Kuala Lumpur station was used as case study. Three distributions that are Weibull, gamma and lognormal were fitted to the dataset after replacement of missing values using mean top bottom methd and performance indicators were calculated to describe the qualities of the distributions. It was found that gamma distribution is the most appropriate distribution representing PM_{10} emissions in Kuala Lumpur.

ACKNOWLEDGEMENT

School of Civil Engineering, Universiti Sains Malaysia, Kampus Kejuruteraan, 14300 Nibong Tebal, Seberang Perai Selatan, Penang.

REFERENCES

1. Chen, J.L., Islam, S. and Biswas, P., (1998). Nonlinear Dynamics of Hourly Ozone Concentrations: Nonparametric Short Term Prediction. *Journal of Atmospheric Environment. 32:* 1839-1848.

2. Department of the Environment, Malaysia (2004) *Malaysia Environmental Quality Report.* Department of the Environment, Minisitry of Science, Technology and Environment, Malaysia.

3. Evans, M., Hastings, N. and Peacock, B. (2000). Statistical Distribution. United States of America: Wiley Series.

4. Hawthorne, G. and Elliot, P. (2005). Imputing Cross-Sectional Missing Data: Comparison of Common Techniques. *Australian and New Zealand Journal of Psychiatry. 39:* 583-590.

5. Junninen, H., Niska, H., Tuppurrainen, K., Ruuskanen, J., Kolehmainen, M., (2002). Methods for Imputation of Missing Values in Air Quality Data Sets. *Journal of Atmospheric Environment: 38:* 2895-2907.

6. Olinsky, A., Chen, S. and Harlow, L. (2002). The Comparative Efficacy of Imputation Methods for Missing Data in Structural Equation Modelling. *Eropean Journal of Operational Research. 151*: 53-79.

7. Engels, M.E. and Diehr, P., (2002). Imputation of Missing Longitudinal Data: A Comparison of Methods. Journal of Clinical Epidemiology. 56: 968-976.

8. Twisk, J.and Vente, W., (2001). Attrition in Longitudinal Studies: How to Deal with Missing Data. *Journal of Clinical Epidemiology. 55:* 329-337.

9. Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* New York: John Wiley.