SHORT REPORT
ScienceAsia 34 (2008): **341-345**
doi: 10.2306/scienceasia1513-1874.2008.34.341

Estimation of missing values in air pollution data using single imputation techniques

Mohamed Noor Norazian[a,*], Yahaya Ahmad Shukri[c], Ramli Nor Azam[c], Abdullah Mohd Mustafa Al Bakri[b]

[a] School of Environmental Engineering, Malaysia University of Perlis, Pejabat Pos Besar, 01007 Kangar, Perlis, Malaysia
[b] School of Material Engineering, Malaysia University of Perlis, Pejabat Pos Besar, 01007 Kangar, Perlis, Malaysia
[c] School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia

* Corresponding author, e-mail: norazian@unimap.edu.my

Received 29 Oct 2007
Accepted 17 Jun 2008

ABSTRACT: Air pollution data obtained using automated machines often contain missing values which can cause bias due to systematic differences between observed and unobserved data. We used interpolation and mean imputation techniques to replace simulated missing values from annual hourly monitoring data for PM10. The most effective method for generating the missing data points was to replace each missing value with the mean of the two data points before and after the missing value. This approach was referred to as the mean-before-after method.

KEYWORDS: air pollution, imputation, performance indicators

INTRODUCTION

Air quality monitoring is carried out to detect any significant pollutant concentrations which may have possible adverse effects to human health. However, such analysis is complicated by the frequently large proportions of observations missing from the data due to machine failure, routine maintenance, changes in the siting of monitors, human error, or other factors. Incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset[1]. There are three major problems that may arise when dealing with incomplete data. First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to the irregularities in data structure and the impossibility of using standard software. Third, and most important, the results may be biased due to systematic differences between observed and unobserved data. At present, there are certain statistical software packages such as SPSS[2] that can perform limited replacement of missing values.

One approach to solve incomplete data problems is the adoption of imputation techniques[3]. Therefore, this research focuses on several single imputation techniques to determine the best technique to replace missing values.

Generally, there are two important types of missing data[3]. Non-ignorable is where the probability of missing a datum is dependent upon its value and ignorable missing data is where the probability of missing a datum is not dependent upon its value. There are three forms of ignorable missing data. The first is associated with sampling. In most situations it is neither efficient nor possible to obtain data from a whole population. Probability sampling is widely used to obtain a representative population sample[1]. The second form of ignorable missing data is missing at random (MAR)[3]. It occurs where the pattern of missingness for a particular variable (Y) may vary for subsets. In this research, the MAR form of ignorable missing data is used because the missing data mechanism of air quality data is generally random. A third form of ignorable missing data is missing completely at random (MCAR), where the missingness occurs at random across the whole data set[3].

From a complete dataset, incomplete datasets need to be generated in order to test the methods. In a study of methods for imputation of missing values in air quality datasets, Junninen et al[4] generated three randomly simulated missing data patterns for evaluating the methods in different missing data conditions. Blended data patterns in the proportions ¼, ½, and ¼ were constructed for examining the methods in a way that reflected the heterogeneity of

the air quality datasets. The patterns were simulated with 10% and 25% missing data. Twisk and Vente[5] have carried out similar work (using 10% or 25% missing data of types MCAR, MAR, and MNAR) on generated incomplete data sets from longitudinal studies.

Most studies of single imputation techniques have been done in areas other than engineering. Engels and Diehr[6] compared four methods and found that the 'last and next' method and last observation carried forward are the best methods to replace missing values. Perneger and Burnand[7] considered a population-based survey to compare the performance of several single imputation techniques. They recommended an imputation algorithm based on the number of key missing items.

MATERIALS AND METHODS

Data

Annual hourly monitoring records for PM10 in Seberang Perai, Penang, Malaysia were selected to carry out the simulation of missing data. The test dataset consisted of particulate matter (PM10) concentrations on a time-scale of one per hour (hourly averaged) for one year**.** A total of 8,757 hourly concentrations are available of which 0.03% (3 observations) are missing (Table 1). The data shows some variability in the PM10 concentration (range: 8–718 μg/m^3, standard deviation: 58.5 μg/m^3). The data is skewed to the right showing that high concentrations of PM10 sometimes occur.

From the complete PM10 dataset, randomly simulated missing data patterns with 5%, 10%, 15%, 25%, and 40% of the data missing were produced for evaluating the accuracy of imputation techniques.

Single imputation techniques

Imputations are means of drawing from a predictive distribution of the missing values, and therefore require a method of creating such a predictive distribution based on the observed data. Complete data matrices can be created using either single imputation or multiple imputation methods[3]. With single imputation, one value is estimated for each missing datum. It has appealing features; for example, the standard complete-data method can be applied directly, and the substantial effort required to create imputations is only needed once. Multiple-imputation is a method of generating multiple simulated values for each missing item in order to properly reflect the uncertainty attached to missing data[3].

In this analysis, six single imputation techniques were applied to estimate the simulated missing values. Four of these were interpolation techniques (linear, quadratic, cubic, and nearest neighbour interpolation). The remaining two were the mean imputation techniques which we will refer to as the mean-before-after and mean-before methods.

Table 1 Characteristics of PM10 data.

Number of valid data points		8757
Number of missing data points		3
Mode		45.0
Standard deviation		58.5
Skewness		3.6
Kurtosis		21.9
Percentiles	25	42.0
	50	65.0
	75	94.0

Interpolation

In linear interpolation two data points are connected with a straight line and hence the interpolation function is given by[8]

$$f_1(x) = b_0 + b_1(x - x_0) \tag{1}$$

where x is the independent variable, x_i $(i = 0,1,2,\ldots)$ is a known value of the independent variable, and b_i are unknown coefficients. Then from (1),

$$b_0 = f(x_0) \tag{2}$$

and

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{3}$$

in which in this case $f = f_1$.

If three data points are available, interpolation is carried out using a quadratic polynomial. A particularly convenient form for this estimation is[8],

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) \tag{4}$$

The coefficients b_0 and b_1 are obtained from (2) and (3) with $f = f_2$. The coefficient b_2 is obtained using

$$b_2 = \frac{\dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \tag{5}$$

with $f = f_2$

When four data points are available, a cubic polynomial can be applied. The cubic interpolation formula has the form[9]

$$f_3(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + b_3(x - x_0)(x - x_1)(x - x_2) \quad (6)$$

The coefficients b_0, b_1, and b_2 are obtained from (3–5), and b_3 is given by

$$b_3 = \frac{\dfrac{f(x_3) - f(x_2)}{x_3 - x_2} - \dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_3 - x_0} \quad (7)$$

with $f = f_3$.

Univariate nearest neighbour imputation is probably the simplest scheme available in that the endpoints of the gaps are used as estimates for all the missing values[4]. The equation for the nearest neighbour method is given by

$$y = \begin{cases} y_1 \text{ if } x \le x_1 + \dfrac{(x_2 - x_1)}{2} \\ y_2 \text{ if } x > x_1 + \dfrac{(x_2 - x_1)}{2} \end{cases} \quad (8)$$

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

Mean imputation techniques

Let $y_1, y_2, \ldots, y_n$ be a times series with n observations of which k values denoted by $y_1^*, y_2^*, \ldots, y_k^*$ are missing. Thus, the observed data with missing values are[10]

$$y_1, y_2, \ldots, y_{n_1}, y_1^*, y_{n_1+1}, y_{n_1+2}, \ldots, y_{n_2}, y_2^*, y_{n_2+1}, y_{n_2+2}, \ldots, y_k^*, y_n \quad (9)$$

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 mean-before-after method replaces all missing values with the mean of one datum before the missing value and one datum after the missing value. Thus for the data in (9), y_1^* will be replaced by[10]

$$\bar{y}_1 = \frac{y_{n_1} + y_{n_1+1}}{2} \quad (10)$$

and y_2^* will be replaced by

$$\bar{y}_2 = \frac{y_{n_2} + y_{n_2+1}}{2} \quad (11)$$

and so on.

The mean-before method replaces all missing values with the mean of all available data before the missing values. Thus for the data in (9), y_1^* will be replaced by[10]

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i \quad (12)$$

and y_2^* will be replaced by[10]

$$\bar{y}_2 = \frac{1}{(n_2 - n_1 - 1)} \sum_{i=n_1+1}^{n_2} y_i \quad (13)$$

and so on.

Performance indicators

Four performance indicators, namely, prediction accuracy, coefficient of determination, mean absolute error, and root mean square error, were used to assess the imputation methods. The theoretical and observed data were compared to select the best method for estimating missing values.

Prediction accuracy (PA) is computed using[11]

$$\text{PA} = \sum_{i=1}^{N} \frac{[(P_i - \bar{P})(O_i - \bar{O})]}{(N-1)\,\sigma_P\,\sigma_O} \quad (14)$$

where N is the number of imputations, O_i and P_i are the observed and imputed data points, respectively, $\bar{O}$ and $\bar{P}$ are their averages, and σ_O and σ_P their standard deviations. PA values range from 0 to 1, with higher values of PA indicating a better fit.

The coefficient of determination (R^2) explains how much of the variability in the imputed data can be explained by the fact that they are related to the observed values or how close the points are to the line. It is given by[4]

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} [(P_i - \bar{P})(O_i - \bar{O})]}{\sigma_P\,\sigma_O} \right]^2 \quad (15)$$

R^2 takes on values between 0 and 1, with values closer to 1 implying a better fit.

The mean absolute error is the average difference between predicted and actual data values, and is given by[4]

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |P_i - O_i| \quad (16)$$

MAE ranges from 0 to infinity and a perfect fit is obtained when MAE = 0.

The mean-squared error is one of the most commonly used measures of success for numerical prediction. Its value is computed by[4]

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[P_i - O_i\right]^2} \qquad (17)$$

The smaller the RMSE value, the better the performance of the model.

RESULTS AND DISCUSSION

Characteristics of simulated missing data

Table 2 shows the descriptive statistics for all the simulated missing data patterns. The mean value varies very little with the percentage of missing data points, and is consistently higher than the median. Although there are differences in the amount of data, it is interesting that the analysis produces similar results for all percentages of missing values. From Table 3, it can be seen that there is very little variation in the percentiles with the percentage of missing values. This is due to the way in which the missing values were generated, and to the occurrence of a large number of observations within the same range.

Replacement of simulated missing values

The mean-before-after method gives the best result (smallest error and highest values of PA and R^2) for each percentage of missing values (Table 4). Among the mean imputation techniques, the mean-before technique gives the worst values for the performance indicators and the mean-before-after method gives the best results for all percentages of missing values. Among the interpolation techniques, the linear interpolation technique gives the best estimates for the 10%, 15%, and 25% missing values and the nearest neighbour method gives the best estimate for the 40% missing values using the R^2, MAE, and RMSE as the performance indicators. Overall, it seems that the mean-before-after method gives the best performance for predicting missing values. This is followed by the linear interpolation technique. The worst estimators are the mean-before method and the quadratic interpolation method.

Table 2 Descriptive statistics for simulated missing data.

Percentage of Missing Data	5%	10%	15%	25%	40%
Number of valid data points	8275	7886	7425	6547	5233
Number of missing data points	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
Range	710.0	710.0	707.0	707.0	710.0
Minimum value	8.0	8.0	8.0	8.0	8.0
Maximum value	718.0	718.0	715.0	715.0	718.0

Table 3 Percentiles of data for simulated missing values.

Percentage of missing values		5%	10%	15%	25%	40%
Valid		8275	7886	7425	6547	5233
Missing		479	871	1332	2210	3524
Percentile	25	42.0	43.0	43.0	43.0	43.0
	50	65.0	65.0	65.0	65.0	64.0
	75	94.0	94.0	95.0	95.0	95.0
	95	171.0	171.0	170.0	171.0	173.0

Table 4 Performance of methods for various percentages of mission values.

P	Method	PA	R^2	MAE	RMSE
	L	0.93	0.86	18.08	24.61
5%	Q	0.12	0.02	43.54	538.0
	C	0.93	0.85	18.56	25.34
	N	0.90	0.80	21.46	29.54
	A	0.93	0.87	17.25	23.71
	B	0.85	0.72	25.90	35.65
10%	L	0.92	0.85	17.80	25.53
	Q	0.92	0.84	18.36	26.55
	C	0.91	0.83	18.40	26.76
	N	0.90	0.80	20.77	29.84
	A	0.93	0.86	16.93	24.35
	B	0.83	0.69	25.32	36.33
15%	L	0.93	0.86	17.15	23.68
	Q	0.92	0.84	18.07	25.09
	C	0.92	0.85	17.55	24.21
	N	0.88	0.78	20.80	30.61
	A	0.93	0.86	16.61	23.27
	B	0.84	0.70	24.11	34.78
25%	L	0.89	0.77	19.21	27.82
	Q	0.87	0.76	20.19	29.44
	C	0.88	0.78	19.71	28.57
	N	0.86	0.74	21.25	31.32
	A	0.88	0.77	18.33	29.12
	B	0.83	0.68	23.74	34.50
40%	L	0.83	0.69	22.42	32.85
	Q	0.31	0.67	23.51	34.37
	C	0.82	0.68	23.11	34.05
	N	0.85	0.73	21.76	31.61
	A	0.88	0.77	19.12	27.79
	B	0.81	0.66	24.61	35.55

P=Percentage of Missing Values
PA = prediction accuracy, R^2 = coefficient of determination, MAE = mean absolute error, RMSE = root mean squared error
L = Linear C = Cubic
Q = Quadratic N = Nearest neighbour
A = Mean-before-after B = Mean-before

REFERENCES

1. Hawthorne G, Elliot P (2005) Imputing cross-sectional missing data: Comparison of common techniques. *Aust N Z J Psychiatry* **39**, 583–90.
2. SPSS Incorporated, (2000). SPSS reference guide, SPSS Inc., Chicago.
3. Little RJA, Rubin BB (2002) Statistical analysis with missing data, 2nd edn, pp 4–22, Wiley, New York.

4. Junninen H, Niska H, Tuppurrainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. *Atmos Environn* **38**, 2895–907.
5. Twisk J, Vente W (2001) Attrition in longitudinal studies: How to deal with missing data. *J Clin Epidemiol* **55**, 329–37.
6. Engels ME, Diehr P (2002) Imputation of missing longitudinal data: A comparison of methods. *J Clin Epidemiol* **56**, 968–76.
7. Perneger TV, Burnand B (2005) A simple imputation algorithm reduced missing data in SF-12 health surveys. *J Clin Epidemiol* **58**, 142–9.
8. Chapra SC, Canale RP (1998) Numerical methods for engineers, 4th edn, pp 124–340, McGraw-Hill, Singapore.
9. Ayyub BM, McCuen RH (1996) Probability, statistics and reliability for engineers and scientists, 2nd edn, pp 137–239, Prentice-Hall, New Jersey.
10. Yahaya AS, Ramli NA, Yusof NF (2005) 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.
11. Chen JL, Islam S, Biswas P (1998) Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction. *Atmos Environ* **32**, 1839–48.