COMPARISON OF LINEAR INTERPOLATION METHOD AND MEAN METHOD TO REPLACE THE MISSING VALUES IN ENVIRONMENTAL DATA SET

[1]Norazian Mohamed Noor, [2]Mohd Mustafa Al Bakri Abdullah, [3]Ahmad Shukri Yahaya, [3]Nor Azam Ramli

[1]School of Environmental Engineering, [2]School of Material Engineering, Universiti Malaysia Perlis, P.O Box 77, d/a Pejabat Pos Besar, 01007 Kangar. Perlis, Malaysia.

[3]School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia.

Abstract

Missing data is a very frequent problem in many scientific field including environmental research. These are usually due to machine failure, routine maintenance, changes in siting monitors and human error. Incomplete datasets can cause bias due to systematic differences between observed and unobserved data. Therefore, the need to find the best way in estimating missing values is very important so that the data analysed is ensured of high quality. In this study, two methods were used to estimate the missing values in environmental data set and the performances of these methods were compared. The two methods are linear interpolation method and mean method. Annual hourly monitoring data for PM_{10} were used to generate simulated missing values. Four randomly simulated missing data patterns were generated for evaluating the accuracy of imputation techniques in different missing data conditions. They are 10%, 15%, 25% and 40%. Three types of performance indicators that are mean absolute error (*MAE*), root mean squared error (*RMSE*) and coefficient of determination (R^2) were calculated in order to describe the goodness of fit for the two methods. From the two methods applied, it was found that linear interpolation method gave better results compared to mean method in substituting data for all percentage of missing data considered.

Introduction

Data collected in air pollution monitoring such as PM_{10}, sulphur dioxide, ozone and carbon monoxide are obtained from automated monitoring stations. These data usually contained missing values due to machine failure, routine maintenance, changes in the siting of monitors and human error. Incomplete datasets 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 the quality of data analysed are of high quality. Incomplete data matrices are problematic: incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset (Hawthorne and Elliott, 2004). 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 irregulaties in data structure and the impossibility of using standard software. Third, and more important, there maybe bias due to systematic differences between observed and unobserved data. One approach to solve incomplete data problems is the adoption of imputation techniques (Junninen *et al.*, 2004). Thus, this study compared the performance between linear interpolation method (imputation technique) and substitution of mean value for replacement of missing values in environmental data set.

Material and Methods

Data

Annual hourly monitoring records for PM_{10} in Seberang Perai, Penang were selected to carry out the simulation of missing data. The test dataset consisted of particulate matter (PM_{10}) concentration on a time-scale of one per hour (hourly averaged) for one year**.** Table 1 gives the summary of particulate matter (PM_{10}).

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 were divided into three degree of complexity that are small, medium and large. The patterns of missing data simulation are represented in Table 2.

Table 1 Descriptive statistic of PM_{10} data

Valid data	8757
Missing data	3
Mode	45.0
Standard Deviation	58.5
Minimum Value	8.0
Maximum Value	718.0

Table 2 The patterns of missing data simulation

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

Computational Methods

a) Linear Interpolation Method

The equation of the linear interpolation function is (Chapra and Canale, 1998):

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \tag{1}$$

where x is the independent variable, x_1 and x_o are known values of the independent variable and $f(x)$ is the value of the dependent variable for a value x of the independent variable.

b) Mean Method

This method replaces all missing values with the mean of all available data. Thus the equation is (Yahaya *et al.*, 2005) :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i \tag{2}$$

where n is the number of available data and y_i is the data points.

Performance Indicators

Several performance indicators were used to describe the goodness of the imputation methods used in this research. The theoretical data and observed data were compared to select the best method for estimating missing values. Three performance indicators were used that are mean absolute error (*MAE*), root mean squared error (*RMSE*) and coefficient of determination (R^2).

a) Mean Absolute Error (*MAE*)

The mean absolute error (*MAE*) is evaluated by the equation (Junninen *et al.*, 2004):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left| P_i - O_i \right| \tag{3}$$

where N is the number of imputations, O_i the observed data points and P_i the imputed data point. Mean absolute error *(MAE)* ranges from 0 to infinity and a perfect fit is obtained when *MAE* equals to 0.

b) Root Mean Squared Error (*RMSE*)

The mean-squared error is computed by (Junninen *et.al.*, 2004):

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

where N is the number of imputations, O_i the observed data points and P_i the imputed data point. The *RMSE* gives the error value the same dimensionality as the actual and predicted values. The smaller value of *RMSE* indicates the better performance of the model.

c) Coefficient of Determination (R^2)
The coefficient of determination (R^2) takes on values between 0 and 1, with values closer to 1 implying a better fit. The equation of coefficient of determination (R^2) is given as follows (Junninen *et al.*, 2004):

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right) \right]}{\sigma_P \sigma_O} \right]^2 \tag{5}$$

where N is the number of imputations, O_i the observed data points, P_i the imputed data point, $\overline{P}$ is the average of imputed data, $\overline{O}$ is the average of observed data, σ_P is the standard deviation of the imputed data and σ_O is the standard deviation of the observed data.

Results and Discussion
Figure 1 below plots the performance of linear interpolation methods and mean methods for replacing the simulated PM_{10} data. From Figure 1, obviously, linear interpolation method gives the best results for all percentage of missing values compared to mean method. The mean method contributes to very large errors compared to linear interpolation method. The R^2 values of linear interpolation method for all percentages of missing values are from 0.69 to 0.86 whereas mean method is 0.00 for all percentage of missing values. This is consistent with that reported by Junninen et al. (2004) which stated that the substitution of mean values for missing data disrupt the inherent sructure of the data and lead to large error in the matrix correlation thus degrading the performance of the statistical modelling.

Conclusions
This paper discusses the comparison of linear interpolation method and mean method to estimate missing values. This study is carried out to prove that substitution of mean values will degrade the statistical performance of the data. The PM_{10} hourly data for a year was used to compare the performance of the methods. Simulated missing values which were categorised as small, medium and large complexities were used. The best imputation techniques for all percentages of the simulated missing data were obtained. Three performance indicators were calculated in order to select the best method replacing the missing values. They are mean absolute error (*MAE*), root mean square error (*RMSE*) and coefficient of determination (R^2),. From these performance indicators, for all degree of complexities the best method was found to be the linear interpolation method.

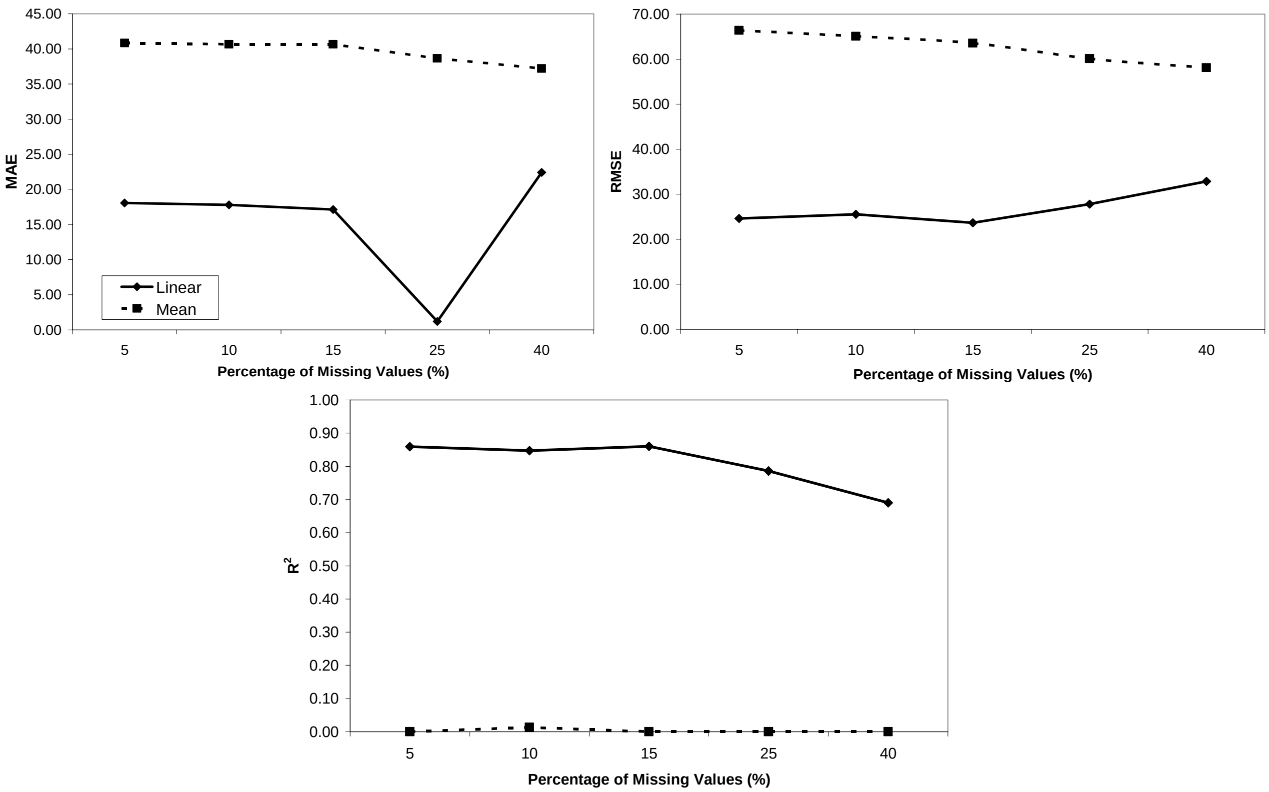

Figure 1: Performance indicators for two methods

References

[1] Chapra, S.C. and Canale, R.P., (1998) *Numerical Methods for Engineers.* Singapore:McGraw-Hill.

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

[3] 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, p. 2895-2907.

[4] Yahaya, A.S, Ramli, N.A. and Yusof, N.F., (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.