Australian Journal of Basic and Applied Sciences, 7(5): 336-341, 2013
ISSN 1991-8178

Estimation of Missing Values in Environmental Data Set using Interpolation Technique: Fitting on Lognormal Distribution

[1]Noor M.N., [2]A.M. Mustafa Al Bakri, [3]Yahaya A.S., [3]Ramli N.A., [3]Fitri N.F.M.Y.

[1]School of Environmental Engineering, Universiti Malaysia Perlis (UniMAP), P.O. Box 77, d/a Pejabat Pos Besar, 01000 Kangar, Perlis, Malaysia.
[2]Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), School of Materials Engineering, Universiti Malaysia Perlis (UniMAP), P.O. Box 77, d/a Pejabat Pos Besar, 01000 Kangar, Perlis, Malaysia.
[3]Clean Air Research Group. Environmental and Sustainable Development Section, School of Civil Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal, 14300, Pulau Pinang, Malaysia.

Abstract: The problem of imputation of missing data emerges in many areas especially environment field. These data usually contained missing values due to many factors such as machine failures, changes in the siting monitors, routine maintenance and human error. Incomplete data set usually cause bias due to differences between observed and unobserved data. One approach that commonly used for handling missing data is imputation technique. This paper discusses three interpolation methods that are linear, quadratic and cubic. A total of 8567 observations of particulate matter (PM_{10}) data for a year were used to compare between the three methods when fitting the lognormal distribution. The goodness-of-fit were obtained using three performance indicators that are mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination (R^2). It was found that linear interpolation method give the best fit with smallest error (value for *MAE* is 1.99) and highest R^2 that is 0.9889.

Key words: Missing values, PM_{10}, performance indicators, lognormal distribution

INTRODUCTION

The presence of missing values in statistical survey data is an important issue to deal with (Little, R.J.A. and B.B. Rubin, 1987). These missing items are usually due to machine failure, routine maintenance, changes in siting monitors, human errors and other factors. Incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset (Hawthorne, G. and P. Elliot, 2005). 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, there maybe bias due to systematic differences between observed and unobserved data.

In the last decades, a number approaches have been proposed and utilised for filling missing values. The most straightforward idea of filling missing data is by substituting average or mean values into the missing entries (Wasito, I. and W. Mirkin, 2005). However, mean substitution lead to large errors in the correlation of matrix and therefore degrading the performance of statistical modelling (Junninen, H., 2004). One approach to solve incomplete data problems is the adoption of imputation techniques.

This paper concentrate on replacing the real missing item of annual PM_{10} monitoring data using three methods that are linear, quadratic and cubic interpolation. The main objective of the study is to observe the performance of each method when fitted to lognormal distribution. Three performance indicators were calculated in order to determine the goodness-of-fit for each method that are mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination (R^2).

MATERIAL AND METHODS

The Data:

The data used for this analysis is particulate matter (PM_{10}) data (measured in $\mu g/m^3$) in Kuala Lumpur. Kuala Lumpur was chosen because it has experienced a rapid growth of population due to the presence of large development area. This is accompanied by a growing number of vehicles that contribute to air pollution. Table 1 gives the summary of particulate matter (PM_{10}). 8567 hourly concentrations are available that is 2.2 percent (193 observations) of missing values. The mean value (77.2) is higher than median (74.0), which indicates that

Corresponding Author: Noor M.N., School of Environmental Engineering, Universiti Malaysia Perlis (UniMAP), P.O. Box 77, d/a Pejabat Pos Besar, 01000 Kangar, Perlis, Malaysia.
E-mail: norazian@unimap.edu.my

the pollutants distributions are skewed to the right. This is confirmed with the range of values between 9 μg/m^3 to 314 μg/m^3.

Table 1: Descriptive statistic of PM_{10} data

Valid data	8567
Missing data	193
Mode	59.0
Standard Deviation	0.31
Minimum Value	9.0
Maximum Value	314.0

The Interpolation Techniques:

There are three interpolation techniques used to estimate the missing values that are linear, quadratic and cubic.

Linear Interpolation:

The simplest form of interpolation is to connect two data points with a straight line. This technique is called linear interpolation. The equation of the linear interpolation function is (Chapra, S.C. and R.P. Canale, 1998):

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

where x is the independent variable, x_0 is a known value of the independent variable and $f_1(x)$ is the value of the dependent variable for a value x of the independent variable. Then from Equation (1),

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

and

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

Quadratic Interpolation:

If three data points are available, estimation are carried out using a second-order polynomial. A particularly convenient form for this estimation is given by (Chapra, S.C. and R.P. Canale, 1998):

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

where x is the independent variable, x_0 and x_1 are known values of the independent variable, b_0 and b_1 are the unknown coefficients and $f_2(x)$ is the second-order interpolating polynomial. The procedure to determine the coefficients of, b_0 and b_1 are the same as in Equations (2) and (3). The coefficients for b_2 is obtained below:

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

Cubic Interpolation:

When four data points are available, a third-order polynomial (also called a cubic polynomial) can be applied. The cubic interpolation formula has the form (Ayyub, B.M. and R.H. McCuen, 1996):

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

The procedure to determine the coefficients of, b_0, b_1 and b_2 are the same as in Equations (3) to (5). b_3 is given by

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

Lognormal Distribution:

The PM_{10} data, which is used in this analysis, is a set of real data where missing values are available. Thus, to compare the performance of the interpolation technique, the data is fitted to lognormal distribution. There are three distributions that are widely used in fitting data that are Weibull, Lognormal and Gamma distribution. However, for this paper, only lognormal distribution is used.

The probability density function (pdf) for the two parameter lognormal distribution is given as (Evans, M., 2000):

$$f(x,\alpha,\beta) = \frac{1}{x\alpha\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\beta}{\alpha}\right)^2\right], x>0, \alpha, \beta > 0 \tag{8}$$

where x is the variable, α is the shape parameter and β is the scale parameter.

The cumulative distribution function (cdf) equation for lognormal distribution is as follows (Evans, M., 2000):

$$F(x,\alpha,\beta) = \frac{1}{2\pi} \int_{-\infty}^{\frac{\ln x-\beta}{\alpha}} e^{\frac{-x^2}{2}} dt, x>0, \alpha, \beta > 0 \tag{9}$$

where x is the variable, α is the shape parameter and β is the scale parameter.

The values of α and β for lognormal distribution were estimated using maximum likelihood estimators (MLE) method. In this method, the equation applied is as follow (Evans, M., 2000):

$$L(x,\alpha,\beta) = \alpha^{-n}(2\pi)^{-\frac{n}{2}} \prod_{i=1}^{n} x_i^{-1} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln(x_i-\beta)}{\alpha}\right)^2\right] \tag{10}$$

To obtain α and β values, the following equation were applied (Evans, M., 2000):

$$\alpha = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[\ln(x_i)-\beta]^2} \tag{11}$$

$$\beta = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{12}$$

where x is the variable, n is the numbers of observation, α is the shape parameter and β is the scale parameter.

Performance Indicators:

Three performance indicators were used to describe the goodness-of-fit or the lognormal distribution when missing values are estimated using interpolation techniques.

Mean Absolute Error (MAE):

Mean absolute error is the average of the difference between predicted and actual values of the data. The mean absolute error (*MAE*) is evaluated by the equation (Junninen, H., 2004):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - O_i\right| \tag{13}$$

where N is the number of imputations, O_i is the observed data points and P_i is the imputed data point.

Mean absolute error *(MAE)* ranges from 0 to infinity and a perfect fit is obtained when *MAE* equals to 0.

Root Mean Squared Error (RMSE):

The mean-squared error is computed by (Junninen, H., 2004):

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} [P_i - O_i]^2 \right)^{\frac{1}{2}} \tag{14}$$

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.

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, H., 2004):

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

where N is the number of imputations, O_i is the observed data points, P_i is 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

The descriptive statistic using interpolation techniques are presented in Table 2. From Table 2, it can be seen that very small differences between the values of the descriptive statistics. The mean values are in the range of 74.9 to 75.0. There is small variability on the data for the three methods used. For each method, the data is skewed to the right.

Table 2: Descriptive Statistics for the three methods.

	Linear	Quadratic	Cubic
Valid	8760	8760	8760
Missing	0	0	0
Mean	74.98	75.00	74.93
Std. Deviation	30.743	30.797	30.770
Skewness	1.407	1.403	1.402
Kurtosis	7.874	7.840	7.822
Percentile			
25^{th}	54	54	54
50^{th}	70	70	70
75^{th}	90	90	91
95^{th}	132	132	132

Table 3 below indicates the values of the shape parameter, α, and the scale parameter, β when the lognormal distribution was fitted to the three data sets and when the maximum likelihood method was used to estimate the parameters.

Table 3: Parameter values for the lognormal distribution.

Data	α	β
Linear	4.2349	0.4166
Quadratic	4.2340	0.4177
Cubic	4.2340	0.4177

From Table 3, it is observed that quadratic and cubic interpolation method give the same value for the shape and scale parameters. The slightly different values of α and β are probably because there are small percentages of missing values in the tested PM_{10} data.

Figure 1 shows the probability distribution function (pdf) plots for the three methods. From the plots, the modes for PM_{10} data for linear, quadratic and cubic interpolation are 62.2, 63.0 and 63.0 respectively. The modes changes slightly according to the methods used for replacing missing data.

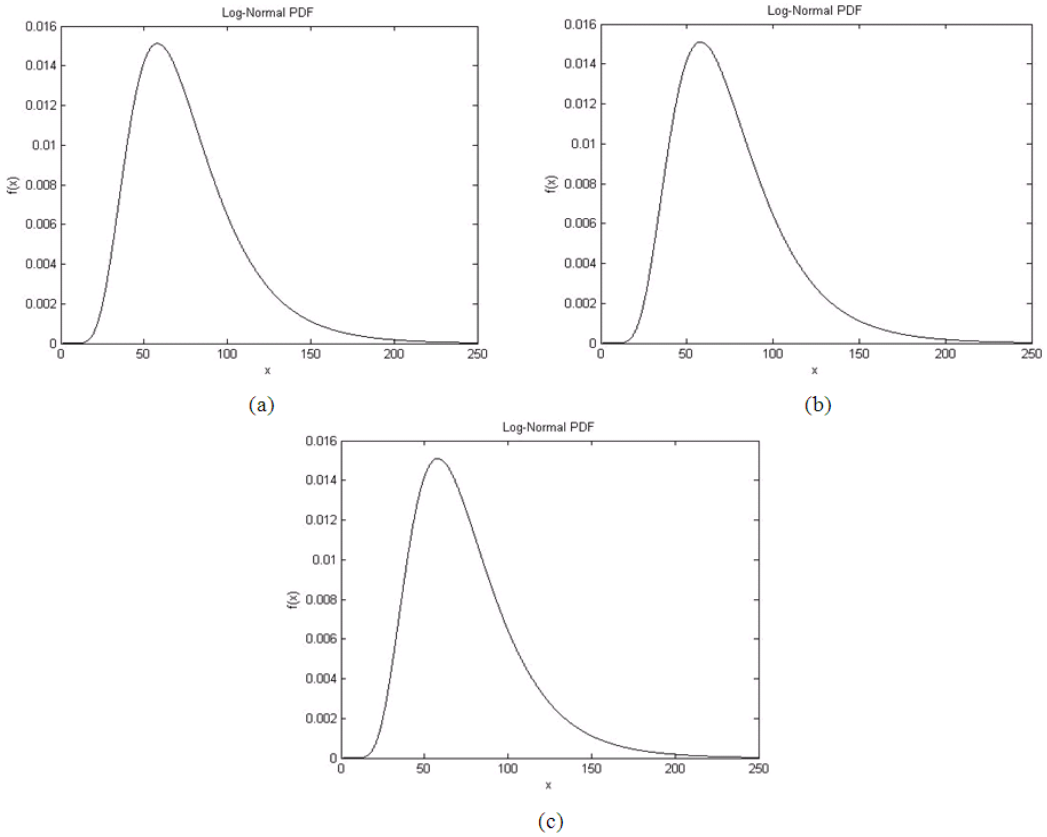

Fig. 1: Probability distribution function (pdf) of the lognormal distribution using (a) Linear Interpolation (b) Quadratic Interpolation (c) Cubic Interpolation. The dotted straight line (-) indicates the theoretical data whereas the dotted line (----) contained the predicted data.

Figure 2 shows the cumulative distribution function (cdf) plots for three methods. The cdf plots show that the lognormal distribution fit the data very well at low concentration but starts to under estimate the data when the concentration reaches 94 $\mu g/m^3$. Calculating the performance indicators does confirmation of the results from Figure 2. Three types of performance indicators were calculated that are mean absolute error (*MAE*), root mean squared error (*RMSE*) and coefficient of determination (R^2). The result is given in Table 4.

Table 4: Performance Indicators.

Methods	*MAE*	*RMSE*	R^2
Linear	1.9878	3.8485	0.9889
Quadratic	2.0194	3.8937	0.9888
Cubic	2.0194	3.8937	0.9888

From Table 4, it can be confirmed that linear interpolation methods gives the best result with the lowest values of error and highest value of R^2 compared to quadratic and cubic interpolation methods. The values of performance indicators for quadratic and cubic interpolation methods are the same because of the value of shape parameter, α, and the scale parameter, β are the same. Overall, the replacement of missing values using linear interpolation fits the data very well. There are only slight variations of the values for all performance indicators measured because the amount of missing item for this data is small.

Conclusions:

Three methods of interpolation techniques were used to estimate the missing values. Since the data used is a real data set, fitting the lognormal distribution to the data made comparison. Estimates of the lognormal parameters were obtained by using the maximum likelihood estimator (MLE). From observing the probability density function (pdf) and cumulative distribution function (cdf), it is observed that all the three methods fit the data well. This might be due to small number of missing observations in the data. Three types of performance indicators that are *MAE*, *RMSE* and R^2 were calculated in order to reconfirm the results. It was found that linear interpolation method give the best fit with smallest error (value for *MAE* is 1.99) and highest R^2 that is 0.9889.

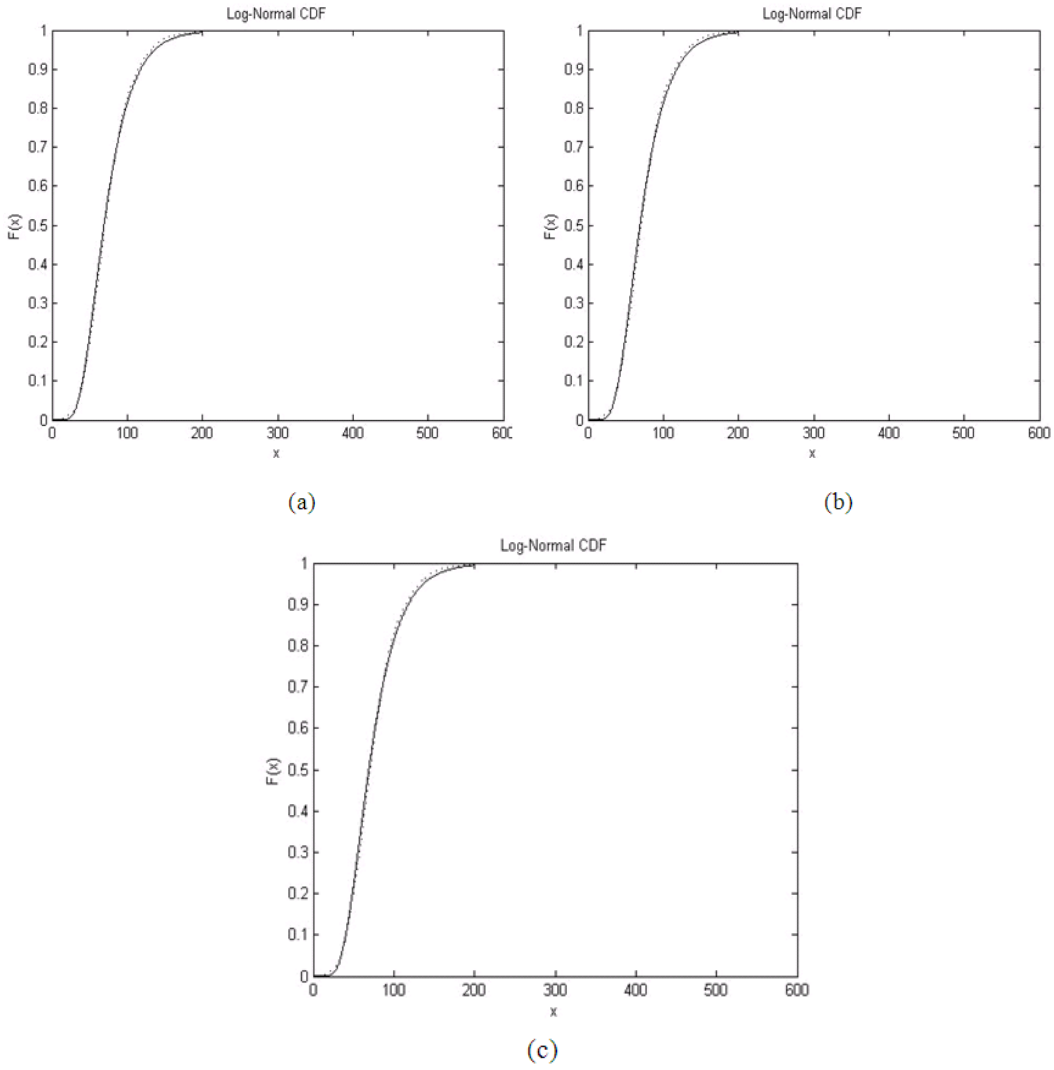

Fig. 2: Cumulative distribution function (cdf) of the lognormal distribution using (a) Linear Interpolation (b) Quadratic Interpolation (c) Cubic Interpolation. The dotted straight line (-) indicate the theoretical data whereas the dotted line (----) contained the predicted data.

REFERENCES

Ayyub, B.M. and R.H. McCuen, 1996. *Numerical Methods for Engineers*. New Jersey: Prentice-Hall.

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

Evans, M., N. Hastings, B. Peacock, 2000. *Statistical Distribution*. United States of America: Wiley Series.

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

Junninen, H., H. Niska, K. Tuppurrainen, J. Ruuskanen, M. Kolehmainen, 2004. Methods for Imputation of Missing Values in Air Quality Data Sets. *Journal of Atmospheric Environment,* 38: 2895-2907.

Little, R.J.A. and B.B. Rubin, 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Wasito, I. and W. Mirkin, 2005. Nearest neighbour approach in the least-squares data imputation algorithms. *Journal of Information Sciences*, 169: 1-25.