Environmental Health Engineering and Management Journal 2021, 8(3), 215–226
http://ehemj.com
doi 10.34172/EHEM.2021.25

Environmental Health Engineering and Management Journal

Original Article

Environmental Health Engineering and Management

Open Access Publish Free

Application of imputation methods for missing values of PM_{10} and O_3 data: Interpolation, moving average and K-nearest neighbor methods

Parisa Saeipourdizaj[1], Parvin Sarbakhsh[2*], Akbar Gholampour[3]

[1]Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran
[2]Health and Environment Research Center, Tabriz University of Medical Sciences, Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran
[3]Health and Environment Research Center, Tabriz University of Medical Sciences, Department of Environmental Health Engineering, School of Public Health, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract

Background: PIn air quality studies, it is very often to have missing data due to reasons such as machine failure or human error. The approach used in dealing with such missing data can affect the results of the analysis. The main aim of this study was to review the types of missing mechanism, imputation methods, application of some of them in imputation of missing of PM_{10} and O_3 in Tabriz, and compare their efficiency.

Methods: Methods of mean, EM algorithm, regression, classification and regression tree, predictive mean matching (PMM), interpolation, moving average, and K-nearest neighbor (KNN) were used. PMM was investigated by considering the spatial and temporal dependencies in the model. Missing data were randomly simulated with 10, 20, and 30% missing values. The efficiency of methods was compared using coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE).

Results: Based on the results for all indicators, interpolation, moving average, and KNN had the best performance, respectively. PMM did not perform well with and without spatio-temporal information. Conclusion: Given that the nature of pollution data always depends on next and previous information, methods that their computational nature is based on before and after information indicated better performance than others, so in the case of pollutant data, it is recommended to use these methods.

Keywords: Air pollution, Algorithms, Environmental pollutants, Spatio-temporal analysis, Humans

Citation: Saeipourdizaj P, Sarbakhsh P, Gholampour A. Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbor methods. Environmental Health Engineering and Management Journal 2021; 8(3): 215–226. doi: 10.34172/EHEM.2020.25.

Article History:
Received: 27 March 2021
Accepted: 19 May 2021
ePublished: 16 September 2021

***Correspondence to:**
Parvin Sarbakhsh
Email: p.sarbakhsh@gmail.com

Introduction

Air pollution as one of the most important issues has major environmental risks to humans, animal health, and other living organisms (1,2). Several natural and human factors produce pollutants such as particulate matter (PM), ozone (O_3), carbone monoxide, etc. in the atmosphere (2,3). In recent decades, urbanization and development of industrial towns and factories have been the main sources of increasing pollutants' production (2,4,5).

Particulate matter (PM_{10}) is one of the most important particles in the atmosphere due to its physical and chemical structural properties (6). Ozone is a secondary pollutant formed in photochemical reactions with precursors, which is produced by humans (7). According to scientific studies, exposure to PM_{10} and O_3 increases the risk of asthma, cardiovascular, pulmonary diseases, and depression (7-9).

In recent years, many epidemiological studies and systematic analysis have been conducted on pollutants and their relationships with various diseases and premature death (2,10). These studies require detailed data of the pollutant concentrations, so the Air Quality Monitoring Organization always measures the concentration of pollutants in the air. Due to the high volume of information, it is always possible that some parts of observations would not be measured when machines fail, position of monitors change, filters are changed, the level of pollution is reduced from the specified range, and human error occurs (11,12).

Therefore, we are always faced with incomplete dataset in air quality data, which in the analysis leads to different conclusions from the results of the complete dataset (13).

There are three major problems in dealing with missing data. First, the loss of information decreases the efficiency and power of the analyses. Second, irregularities in the data structure and the impossibility of using standard software reflect in complexities related to data management and analysis, especially in time series analysis which we need sequential data to make predictions. And thirdly, systematic differences between observed and unobserved data are among the most important problems of missing data, which change the obtained results (12).

There is no conducted study on the imputation of missing values of pollutant concentration in Tabriz. Due to the importance of missing issue and the consequences of arbitrary removal of missing data, which causes bias in the results, the aim of this study was to evaluate and compare the efficiency of imputation methods for missing values of PM_{10} and O_3 concentrations in Tabriz in 2017. These methods include single univariate methods of mean, K-nearest neighbor, linear interpolation, moving average (simple, linear, and exponential), EM algorithm, linear regression and univariate multiple classification, and regression tree method, as well as multivariate multiple predictive mean matching (PMM) method by considering the spatial and temporal dimensions.

Materials and Methods

Data

In this study, the hourly mean concentrations of PM_{10} and O_3 recorded in Tabriz air quality monitoring stations in 2017 were used. For both pollutants, no data were recorded from January 1 to February 22, June 22 to July 22, and August 23 to September 22. Therefore, the mentioned periods were excluded, and approximately, data of 8 months of 2017 were analyzed. Concentrations of PM_{10} and O_3 were measured by beta attenuation and UV-spectrophotometry methods at each station, respectively. After removing the outliers using the Z-score method, only 24000 cases of hourly PM_{10} and O_3 concentrations for each pollutant were left, of which 5% and 2% (1187 and 425 observations) were missing values, respectively. Some variability in the concentration of PM_{10} and O_3 are as follows (for PM_{10} range: 236.28 μg/m^3, standard deviation: 36.16 μg/m^3; for O_3 range: 149.61 μg/m^3, standard deviation: 32.46 μg/m^3).

Identifying the missing mechanism

To deal with the missing problem and have accurate statistical analysis, it is necessary to identify the pattern and mechanism of missing data. First, the pattern of missing must be determined. Then, in accordance with the pattern of missing, we must adopt an appropriate approach to deal with the missing data. To identify the mechanism of missing data, the missing data classification system was used according to the Rubin's theory (14). This system actually describes the relationship between the data and the probability of missing values. To better understand and describe the distribution mechanism of the missing data: Suppose the vector $X = (X_1, X_2, \ldots, X_n)^T$ represents a random variable of complete data that includes the observed values X_o and the missing values X_m with the probability density function f_θ. The goal is to estimate the unknown parameter vector θ. The missing data indicator $M = (M_1, M_2, \ldots, M_n)^T$ is a binary variable that identifies the observed or missing state of the variable (if the value x_i is observed, $M_i = 0$, and if the value is missing, $M_i = 1$), in fact, the missing data indicator defines the missing pattern. Representing missing data as a variable indicates that a probabilistic distribution manages the value of the missing data indicator. In practice, it is impossible to understand the exact distribution of M. However, the nature of the relationship between the indicator M and the data reveals the mechanism of the missing data as defined by the conditional distribution $f(M \mid X, \varphi)$ of M over the complete data X as the vector φ is an unknown parameter that indicates the probability of missing data.

The missing completely at random (MCAR) mechanism requires that the probability of missing data in one variable X to be unrelated to the other measured variables, as well as the X values themselves; in this case, a random sample of complete data can be considered. In cases where there is a univariate time series, time is considered as an implicit variable, the probability of missing an item is independent of the observed time. Given the above-mentioned information, the distribution that controls the MCAR mechanism is as follows:

$$P(M \mid X_o, X_m, \varphi) = P(M \mid \varphi). \text{ for all } X. \varphi \quad (1)$$

A more limited assumption than MCAR is that the missing values depend only on the observed variables, not on the variables that have the missing values, so the mechanism of missing data is called missing at random (MAR). Since there is no variable other than time in univariate time series, it is assumed that the probability of missing data depends on the point in time at which it is observed. The distribution of the MAR mechanism is as follows:

$$P(M \mid X_o, X_m, \varphi) = P(M \mid X_o, \varphi). \text{ for all } X_m \quad (2)$$

Finally, when the probability of missing data in X is dependent on X values that are observed or missed, the data are missing not at random (MNAR). In the case of univariate time series, the probability of missing data may depend, but not necessarily, on the point in time at which it is observed. The distribution of the probability of MNAR mechanism is as follows:

$$P(M \mid X, \varphi) = P(M \mid X_o, X_m, \varphi) \quad (3)$$

The concentration of pollutants is recorded at a specific time and location. In this dataset, time and location are as observed variables. The probability of missing an observation in pollutant concentration variable is independent of other observations but dependent on time and location variables (15-17). Thus, we can conclude that the missing mechanism of PM_{10} and O_3 concentrations would be MAR.

In addition, a number of studies have been conducted on the mechanism of missing that have considered the missing mechanism in air pollution data as MAR. Therefore, according to the available evidences and reviewing the literature (16-24), the mechanism of missing in PM_{10} and O_3 data was considered as MAR.

Imputation methods

In recent decades, various techniques have been introduced to solve the problem of missing data (14). One of the most popular and simplest methods is to delete the missing data, which is done in two ways: Pairwise deletion and listwise deletion. In the pairwise deletion, only the missing observations are deleted, which is the reason why the number of observations for analysis varies from one variable to another. In listwise deletion (also known as case deletion or complete analysis), all observations that have missing data in one or more variables are deleted. If the missing data mechanism is MCAR and the missingness rate is less than 5%, the complete data series can be obtained by deleting the missing values. Otherwise, if the missingness rate is high or the mechanism is MNAR or both of them, by eliminating information, a reduction in power or bias of results will occur (25,26).

On the other side of the deletion methods is the imputation approach, in which an estimate for the missing values is obtained and used. Imputation can be done with different techniques. These techniques can be categorized based on the number of imputed values (single and multiple) generated in the presence or absence of other variables (univariate or multivariate). Imputed values are replaced for each of the missing values (14).

In the single imputation method, the missing values are filled by only one amount and the imputation process is performed only once. The multiple imputation method for maintaining uncertainty in missing data (14) generates several simulated values for each missing value.

In a univariate imputation, the missing values of a variable are estimated as a function of the observed values of the same variable. In multivariate imputation, the missing values in one variable are estimated with other variables that are recorded simultaneously, in which multivariate imputation performance may be better than that of the univariate imputation (27). However, when a number of variables that are simultaneously recorded are missing, it is difficult to access the original data pattern (28).

The univariate single imputation generally works using the mean or median of the measured values, moving the previous observations forward, the next observations backwards, or the average of the before and after observations. The single multivariate imputation also proceeds to use a function of the mean or median of the values measured simultaneously (29-31).

In this study, 8 imputation methods including univariate single methods such as mean, K-nearest neighbor, linear interpolation, moving average (simple, linear, and exponential), EM algorithm, linear regression, and multiple univariate method such as classification and regression tree, as well as multivariate multiple PMM method were examined to select a better method for air pollution data analysis. R (4.0.2) (packages: mice (3.9.0), imputeTS (3.1), VIM (6.0.0)), SPSS version 25, and Microsoft Excel software were used to perform the imputation methods and analyze the obtained information.

To compare the efficiency of the imputation methods in this study, 5 performance characteristics including coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), index of agreement (d_2), and coefficient of efficiency (E_2) were used. The values of each of these characteristics were compared using the original and estimated values in the test group to select the best method for estimating the missing values. A brief description of how each of the methods used works, is presented as follows.

Mean

Mean imputation technique is one of the simplest methods for imputing the missing values. In this method, the total missing values in the dataset are filled by the average of the available values (32). Mean imputation method has advantages and disadvantages. One of its advantages is that it is comprehensible and applicable in most statistical softwares. In this case, the sample size also does not decrease due to the fact that all the missing values are placed with the average. One of its disadvantages is that the mean imputation method leads to the bias of multivariate estimates such as correlation or regression coefficients. In general, the values imputed by the mean of the variables have no correlation with the other variables. Thus, the relationship between the variables is skewed to zero, and the standard error of the imputed variables are biased (33).

Moving average (MA)

In this function, the missing values are replaced by the moving average values. In this method, the mean is taken from an equal number of observations on either side of a central value, that is, for the missing value in position *i* of a time series, observations *i*-1, *i*+1, *i*+1, *i*+2, and so on

(assuming window size of k = 2) are used to calculate the mean.

Since long gaps of missing values may occur and all values next to the central value are also missing, the algorithm has a semi-adaptive window size. When there are less than 2 non-missing values in the full available window, the window size will gradually increase to at least 2 non-missing values. In all other cases, the algorithm returns to the size of the preset window.

In simple moving average (SMA), all observations in the window have the same weight to calculate the average. In linear weighted moving average (LWMA), the weight of observations decreases in the algorithm process. Observations that are exactly next to the central value *i*, have a weight of 1/2 ; observations one farther away (*i*+2.*i*-2) have a weight of 1/3; the next (*i*+3.*i*-3) have a weight of 1/4, and so on. The exponential weighted moving average (EWMA) also uses weighting factors that decrease exponentially. Observations that are exactly next to the central value of *i* have a weight of $(\frac{1}{2})^1$; observations one farther away (*i*+2.*i*-2) have a weight of $(\frac{1}{2})^2$; the next (*i*+3.*i*-3) have a weight of $(\frac{1}{2})^3$, and so on.

Linear interpolation

In linear interpolation method, two data points are connected by a line and the interpolation function as Eq. (4).

$$f_1(x) = b_0 + b_1(x - x_0) \quad (4)$$

So the independent variable, x_i (i = 0.1.…), is a known value and the coefficient b_0 is unknown. So in Eq. (4), we have: $b_0 = f(x_0)$; $x_0 < x < x_1$ and $b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$, in this case, $f = f_1$ have the same distribution (34).

K-Nearest Neighbor

The k-nearest neighbor (KNN) imputation method is the simplest strategy because the endpoints of the gaps (missing parts) are used as estimates for all missing values (28). The KNN method is presented in Eq. (5).

$$y = \begin{cases} y_1 & if \quad x \le x_1 + \frac{(x_2 - x_1)}{2} \\ y_2 & if \quad x > x_1 + \frac{(x_2 - x_1)}{2} \end{cases} \quad (5)$$

where, y is interpolated, x is the time points of interpolation, y_1 and x_1 are the coordinates of the starting points of the gap, y_2 and x_2 are the coordinates of the end points of the gap.

EM algorithm

The EM algorithm can be used when the joint distribution of missing data (X_m) and observed data (X_o) is known (14,35). If for $\theta \in \mathbb{R}^d, f(X;\theta)$ is a probability density function of $X = (X_o.X_m)$, the goal of the EM algorithm is to find an estimate of θ that maximizes the accuracy of the observed data. This quantity cannot be explicitly calculated in general cases; the EM algorithm finds the expected MLE value by repeating it in the complete data likelihood maximization. Then, by starting with the initial value of $\theta^{(o)}$, and assuming that $\theta^{(t)}$ is an estimate of θ in t, the algorithm is performed in two steps:

E-Step: Calculation of the expected value of complete data likelihood according to the conditional distribution of the parameter value of the missing variable of $\theta^{(t)}$.

M-Step: Maximizing the Q function and determining the value of $\theta^{(+1t)}$.

Multiple imputation by chained equation

Assume that X is the data matrix ($n \times p$) and $X = (X_p.X_c)$ so that X_p consists of p_1 columns of X which are almost as observed and X_c contains the rest of the columns that are completely observed. X_o is a set of elements observed in X, and X_m is a set of missing observations in X. For a multiple imputation based on chain equations, the equation specifies a set of conditional distributions $P(X_i|X_{-i})$, where X_i is the i-th column of X_p and X_{-i} is the matrix of X whose i-th column is omitted. Imputed values are generated in 4 steps:

Step 1: The initial values for the missing values are completed as follows: The matrix Z is defined as X_c. Then, for each i = 1.….p_1, the missing values in X_i are imputed using the conditional posterior distribution on Z, and the full version of X_i is added to Z before the value of i is increased.

Step 2: For each i = 1.….p_1, the missing values in X_i are replaced on X_{-i} using the conditional prior distribution.

Step 3: The second step is implemented l times.

Step 4: The first to third steps are repeated until m imputation sets are obtained.

The X_p columns are adjusted to increase the number of missing values until more information is available in the second part of the first step. Although random convergence can be conventionally investigated using a diagnostic tool such as scale reduction coefficient, satisfactory results are usually obtained using l = 10 (36). In the first and second steps, the basis of prediction is the use of generalized linear models as a criterion. In this study, classification and regression tree, linear regression and PMM were performed based on the multiple imputation by chained equation method.

Classification and regression tree

Classification and regression tree (CART) models seek to approximate the conditional distribution of a response variable on several predictor variables. The CART algorithm divides the space of the predictors so that the subset of the units formed by the partitions have relatively homogeneous results. Partitions are formed by the recursive binary division of the prophets. A set of partitions can be effectively represented by a tree structure with its leaves corresponding to a single subset.

The values in each leaf represent the conditional

distribution of the response variable for the units in the data with the predictors that define the leaf separation criteria.

If the parametric models are equal and there is no discontinuity in the separation boundaries (37), the performance of CART method decreases, which is one of the main disadvantages of CART in comparison to parametric models.

Once a tree has grown, it can be trimmed by removing the branches. When trees are used as an analytical tool, it is better to modify them because smaller trees are easier to interpret. The trees are not modified when using multiple imputation methods. Instead, by adjusting the minimum number of observations and minimizing the heterogeneity in the values per leaf, the size of the trees is adjusted to allow for more division. More information on the CART method is presented in previous studies (38,39).

Regression (stochastic vs deterministic)

This method fits a statistical model on a variable with missing values. The predictions of this regression model are used to replace the missing values of this variable (36). Regression imputation has 2 steps:

1. A linear regression model is estimated based on the variables observed in the objective variable *Y* and some explanatory variables *X*.
2. This model is used to predict the missing values in *Y*. Then, missing values *Y* are replaced based on these predictions.

Regression imputation is classified into two different types: Stochastic and deterministic regression imputation. Deterministic regression imputation replaces missing values with exact predictions from the regression model. Therefore, the imputed values are too accurate and lead to the overestimation of the correlations between *X* and *Y*. To solve this problem, stochastic regression imputation is used instead of deterministic regression imputation. Stochastic regression imputation adds a random error sentence to the predicted value, so it can reproduce the correlation between *X* and *Y* more appropriately.

One of the advantages of the above-mentioned method is that the relationships between *X* and *Y* (correlation, regression coefficients, etc.) are preserved because the imputed values are based on regression models. And the disadvantage of this method is that it may lead to impossible values. There are some limitations in this method and variables are often limited to a certain range (e.g., income must always be positive), so that the regression imputation is not able to operate under such restrictions (40,41).

Predictive mean matching

The PMM method is a new method in imputation methodology (42,43). The PMM algorithm can be divided into 6 steps:

1. Estimation of a linear regression model, in which *Y* is selected as the variable to be imputed with an appropriate predictive set such as *X*. Only the observed *X* and *Y* values are used to estimate the model.
2. $\bar{\beta}$ is randomly selected from the posterior predictor distribution and generates a set of new coefficients $\bar{\beta}$ (this Bayesian step requires all multiple imputation methods to generate some random variability in the imputed values).
3. $\bar{\beta}$ values are used to calculate the predicted values for the observed values *Y*, and β* values are used to calculate the predicted values for the missing values *Y*.
4. 4. For each case where *Y* is missing, the nearest predicted value is selected from the items where *Y* is observed (the PMM algorithm selects the nearest observed value for missing Y_i (usually 3 items)).
5. Then, it randomly selects one of these three items and places it with the corresponding value for the missing value.
6. In the multiple imputation, steps 1 to 5 are repeated several times. Each iteration of steps 1 to 5 creates a new imputed data set. In a multiple imputation, missing data is usually imputed 5 times. In order to choose which of the 5 times is the final imputed values, it is better to average the obtained values and analyze the obtained average as the final imputation values.

The advantage of the PMM method is that it only operates based on the values observed for other units, so that the range of imputed values is always between the lowest and highest observed values. In addition, unlike other methods such as regression imputation overestimating the variance of small values *X* and underestimating the variance of large values *X*, the PMM method reflects the structure of the observed values well (44-47).

Evaluation the performance of imputation methods

In this study, the training-testing validation approach was used to evaluate the performance of imputation methods. In this approach, a number of data were randomly deleted from the existing main dataset. Then, the deleted data were replaced by the estimated values obtained from different imputation methods to compare with the original data. To perform the above-mentioned approach, complete data without any missing data were selected from the original dataset. In the next step, from 22 813 and 18 883 complete and available observations for PM_{10} and O_3, respectively, 10, 20, and 30% were randomly selected and deleted (48,49). These deleted data were treated like the missing data and considered as test data. Then, missing values were imputed by various imputation techniques to recover the deleted values. Finally, the imputed values were compared with the observed values.

Performance indicators

To evaluate and compare the methods, the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), as well as two measurement criteria without dimension of agreement index (d_2) and efficiency coefficient (E_2) were used.

Coefficient of determination (R^2)

The value of R^2 indicates how much of the changes in the imputed data can be described by the observed data or points that are close to the regression line (28), so we have:

$$R^2 = [\frac{1}{N}\sum_{i=1}^{N}\frac{(X_{m_i}-\bar{X}_m)(X_{o_i}-\bar{X}_o)}{\sigma_m \sigma_o}]^2 \quad (6)$$

The value of R^2 is between 0 and 1, with values closer to 1 implying a better fit.

Mean absolute error

The average difference between imputed and observed data is shown by the following equation (28):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_{m_i}-X_{o_i}| \quad (7)$$

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

Root mean square error

RMSE is one of the most common methods for evaluating numerical prediction (28). Its value is calculated by Eq. (8).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{m_i}-X_{o_i})^2} \quad (8)$$

A smaller value of RMSE indicates better performance of the model.

Index of agreement (d_2)

d_2 is a measure of relative error between imputed and observed data, which is given by Eq. (9) (50).

$$d_2 = 1-\left[\frac{\sum(X_o-X_m)^2}{\sum(|X_m-\bar{X}_o|+|X_o-\bar{X}_o|)^2}\right] \quad (9)$$

The value of d_2 is always between 0 and 1, and higher values indicating better agreement.

Coefficient of efficiency (E_2)

The value of E_2 is calculated as follows (51):

$$E_2 = 1-\left[\frac{\sum(X_o-X_m)^2}{\sum(|X_o-\bar{X}_o|)^2}\right] \quad (10)$$

E_2 is always between infinity to 1, and higher values indicating better agreement (52).

Results

Table 1 shows the descriptive statistics for 10, 20, and 30% missing for PM_{10} and O_3.

By changing the percentage of missing, the average value has changed very little and is always higher than the median value. As shown in Table 1, there is very little variation in percentiles of various missing rate. This is due to the random generation of missing values and the large number of observations in the same range. After imputation with different methods, the efficiency of each method was calculated and compared.

Table 2 shows the performance indicators values for various methods with missing of 10, 20, and 30% for both pollutants.

First, the results obtained for PM_{10} are discussed in detail. According to the results for 20% missing as the medium rate, values of R^2, RMSE, and MAE for the linear interpolation method were 0.822, 15.14, and 8.33, respectively. After linear interpolation, the moving average and the nearest neighbor had the best performance and PMM and linear regression showed the worst fit.

PMM method has been introduced as one of the

Table 1. Descriptive statistics for PM_{10} and O_3 at various missingness rates

Descriptive Statistics		PM_{10}			O_3		
Missingness rate (%)		10	20	30	10	20	30
Number of valid data point		20523	18290	15940	21350	18883	16516
Number of missing data point		3477	5710	8060	2650	5117	7484
Mean (μg/m³)		57.2	57.28	57.11	45.28	45.26	45.27
Standard deviation (μg/m³)		36.13	36.27	36.03	32.48	32.42	32.34
Skewness (μg/m³)		1.44	1.44	1.43	0.37	0.37	0.36
Kurtosis (μg/m³)		2.49	2.47	2.48	-0.73	-0.73	0.02
Range (μg/m³)		236.28	236.28	236.28	149.61	149.61	149.61
Minimum value (μg/m³)		0.24	0.24	0.24	0.02	0.02	0.02
Maximum value (μg/m³)		236.52	236.52	236.52	149.63	149.63	149.63
Percentile (μg/m³)	25	32.14	32.1	32.06	15.46	15.53	15.61
	50	47.82	47.9	47.83	43.53	43.50	43.68
	75	68.79	72.86	72.71	70.02	69.94	69.86

Table 2. Performance of the investigated imputation methods for various rate of missing values for PM_{10} and O_3

Missingness Rate	Method	PM_{10}					O_3				
		MAE	RMSE	R^2	E_2	d_2	MAE	RMSE	R^2	E_2	d_2
10%	PMMST	29.18	38.56	0.015	-0.122	0.391	28.84	34.94	0.003	-0.169	0.372
	PMMT	28.99	38.38	0.02	-0.11	0.402	28.93	34.74	0.003	-0.155	0.369
	PMMS	29.22	38.56	0.011	-0.124	0.377	29.14	35.16	0.001	-0.183	0.361
	PMM	29	38.70	0.038	-0.13	0.493	22.20	28.10	0.270	0.244	0.698
	CART	14.35	20.75	0.684	0.675	0.906	16.12	21.36	0.580	0.563	0.869
	LRNP	27.36	36.28	0.007	0.007	0.119	26.64	31.70	0.038	0.038	0.285
	KNN	9.84	15.92	0.808	0.809	0.944	10.92	15.23	0.779	0.778	0.933
	LINT	7.82	14.15	0.848	0.849	0.959	7.04	11.29	0.877	0.878	0.967
	SMA	10.83	16.93	0.783	0.784	0.935	13.08	17.49	0.708	0.707	0.905
	LWMA	9.81	15.74	0.813	0.813	0.945	11.21	15.28	0.779	0.777	0.930
	EWMA	8.97	14.96	0.83	0.831	0.951	9.54	13.45	0.829	0.827	0.948
	EM	27.36	36.28	0.007	0.007	0.119	11.05	16.25	0.747	0.747	0.923
	MEAN	27.56	36.41	-	0	0.006	27.58	32.33	-	0.000	0.027
20%	PMMST	29.16	38.39	0.007	-0.154	0.361	29.32	35.35	0.003	-0.170	0.376
	PMMT	28.90	38.47	0.006	-0.16	0.358	29.50	36.25	0.000	-0.187	0.349
	PMMS	28.74	38.19	0.009	-0.142	0.366	29.27	35.78	0.002	-0.169	0.367
	PMM	28.27	37.71	0.052	0.114	0.494	23.27	30.40	0.557	0.194	0.666
	CART	14.46	21.37	0.655	0.642	0.896	16.35	23.20	0.571	0.557	0.864
	LRNP	26.98	35.61	0.000	0.007	0.124	26.88	32.50	0.042	0.042	0.282
	KNN	10.38	16.85	0.777	0.778	0.935	11.95	17.55	0.750	0.749	0.921
	LINT	8.33	15.14	0.822	0.82	0.951	7.68	12.15	0.867	0.867	0.963
	SMA	11.03	17.44	0.762	0.762	0.928	13.31	18.90	0.705	0.704	0.904
	LWMA	10.05	16.33	0.791	0.791	0.938	11.51	16.53	0.774	0.771	0.928
	EWMA	9.27	15.64	0.808	0.808	0.945	9.94	14.62	0.820	0.817	0.945
	EM	26.27	34.88	0.047	0.047	0.299	11.94	21.07	0.716	0.716	0.911
	MEAN	27.12	35.73	-	0.000	0.019	27.80	33.39	-	0.000	0.007
30%	PMMST	28.71	38.39	0.022	-0.109	0.399	29.39	35.45	0.002	-0.177	0.362
	PMMT	28.63	38.37	0.016	-0.107	0.387	29.85	35.76	0.000	-0.187	0.346
	PMMS	29.24	38.84	0.009	-0.135	0.372	29.60	35.48	0.001	-0.166	0.361
	PMM	29.08	38.54	0.053	-0.117	0.493	22.77	28.87	0.252	0.190	0.667
	CART	15.28	22.37	0.635	0.624	0.887	16.34	21.76	0.573	0.550	0.862
	LRNP	27.28	36.29	0.01	0.009	0.113	26.98	32.02	0.044	0.044	0.244
	KNN	11.37	17.85	0.76	0.76	0.927	12.93	17.59	0.712	0.706	0.906
	LINT	8.74	15.24	0.825	0.825	0.951	8.24	12.68	0.850	0.854	0.959
	SMA	11.47	18.07	0.755	0.755	0.925	13.42	18.05	0.696	0.693	0.901
	LWMA	10.47	16.88	0.786	0.786	0.936	11.72	16.07	0.761	0.758	0.925
	EWMA	9.7	16.14	0.904	0.804	0.943	10.27	14.56	0.803	0.803	0.941
	EM	27.27	36.29	0.1	0.009	0.113	12.95	18.66	0.675	0.618	0.869
	MEAN	27.45	36.47	-	0	0.011	27.93	32.75	-	-0.001	0.006

PMMST: Predictive mean matching spatial and temporal dependencies; PMMT: Predictive mean matching temporal dependencies; PMMS: Predictive mean matching spatial dependencies; PMM: Predictive mean matching; CART: Classification and regression tree, LRNP: Linear regression with predictive variable; KNN: K-nearest neighbor, LINT: Linear interpolation; SMA: Simple moving average; LWMA: Linear weighted moving average; EWMA: Exponential weighted moving average; EM: Expectation-Maximization algorithm.

efficient methods in the field of missing data. Thus, spatial and temporal information of the data were entered into the model as an independent variable and examined. First of all, the model was fitted by considering spatial and temporal information of the data (R^2 = 0.007, RMSE = 38.39), temporal information (R^2 = 0.006, RMSE = 38.47), and spatial information (R^2 = 0.009, RMSE = 37.71). Finally, the model was fitted with none of the spatial and

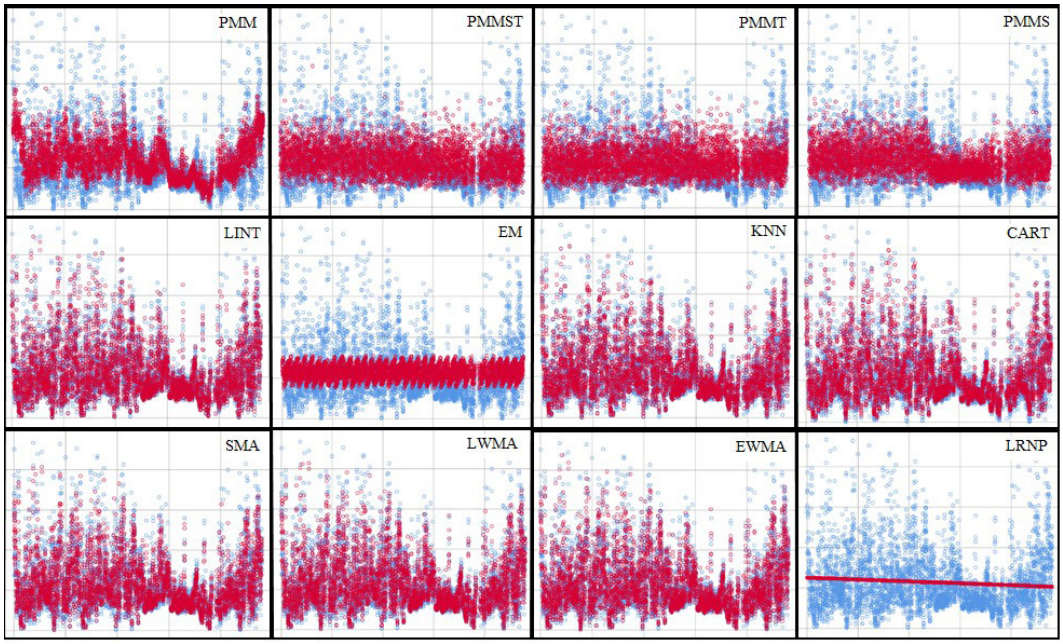

Figure 1. Scattering of imputed values versus original values at 20% missing rate for PM_{10}. Blue dots present original values and red dots present imputed values.

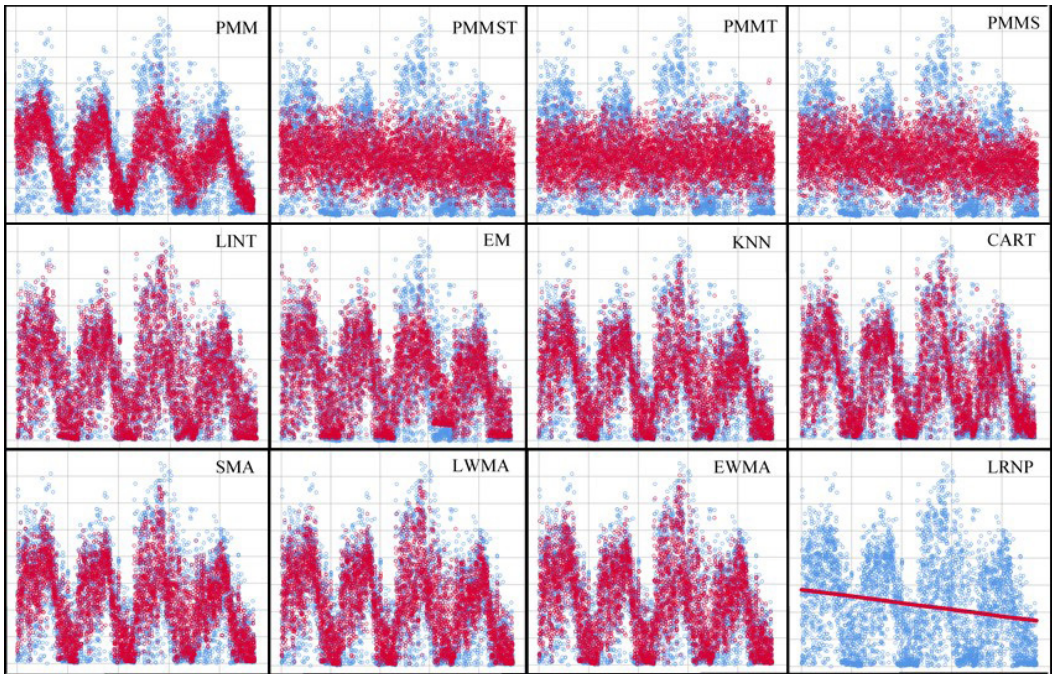

Figure 2. Scattering of imputed values versus original values at 20% missing rate for O_3. Blue dots present original values and red dots present imputed values.

temporal information ($R^2 = 0.52$, RMSE = 37.71).

The PMM method without spatial and temporal dependencies in imputation has shown good performance in 20% missing in all validation methods. Although the results for 10% missing were consistent with the results of 20%, in 30% missing, considering the time-dependent variable in imputation, it performed well in most validation methods.

Based on the results, when spatial and temporal information was not entered into the model, PMM method showed better performance, which indicates that even with fitting the model using spatial and temporal information as an independent variable, the PMM method does not have a good ability to impute air pollution data which are time series.

CART method, which is based on regression and decision tree, and the EM algorithm method, which is based on iteration in estimating parameters and convergence, both had better performance than the PMM and regression methods in evaluation and validation.

The moving average method was also evaluated in three models: SMA, LWMA, and EWMA. According to the results, EWMA model (R^2 = 0.808 and RMSE = 15.64) had the best performance, followed by LWMA model, and, SMA model had the weakest performance.

According to the obtained results for PM_{10} data, mean imputation (RMSE = 35.73 and d_2 = 0.019) had poor performance, so that it is not a suitable method for nesting missing values. The results of evaluating methods in 10% and 30% missing are the same as 20% missing. In all scenarios, the linear interpolation, moving average, and nearest neighbor methods had the best performance, respectively.

Regarding O_3, according to the obtained results for all missing rates in the imputation of O_3 missing values, linear interpolation had the lowest RMSE and MAE whereas these values were the highest for PMM and MEAN. It should be noted that the results obtained for O_3 were similar to those obtained for PM_{10}, and the linear interpolation, moving average, and nearest neighbor in terms of performance indicators were the best methods for imputing the missing values, respectively.

Figure 1 and Figure 2 show the scatter plot of imputed values on the original dataset for each imputation method for PM_{10} and O_3, separately.

The obtained plots for PM_{10} and O_3 also show the highly overlap of the original and missing values in the LINT, KNN, LWMA, and EWMA methods compared to the other methods. For PM_{10}, PMM plots had the least overlap in all cases and the plots of the EM and LRNP also did not show a significant overlap. In the case of O_3, the plot of EM had a better overlap, but the results obtained in terms of performance indicators are not acceptable to EM.

Discussion

The aim of this study was to evaluate and compare several imputation methods in the various missing rates (10%, 20%, and 30%) for air quality dataset of Tabriz. Due to the importance of missing issue and the consequences of arbitrary removal of missing data, which causes bias in the results, the mechanism of missingness was assessed and several imputation methods were performed. Moreover, performance indicators for evaluation and validation of the underlying methods along with statistical concepts and models were reported.

In this study, the nature of air pollution data in the PMM method was completely considered. That is, in this method, the spatial information as air quality monitoring stations, and temporal information such as diurnal and hourly recordings of the air pollution data were considered. The results obtained from the PMM method were compared with those from other methods whose efficiency and validity have been determined in other studies (36-38,42-46).

It is essential to notice that every method can not be used to fill the unobserved values for imputation, even if the introduced method is one of the best imputation methods. Each method in each situation works differently depending on the nature of the data. In the present study, several methods were investigated for imputing the missing values of PM_{10} and O_3. The results of both pollutants were similar, so the results obtained in this study can be generalized to other pollutants. Due to the nature of air pollution data in Iran, which is almost similar to the data of Tabriz, the proposed methods in this study with acceptable accuracy can be used to replace the missing values in similar air pollution dataset.

The results showed poor performance for mean imputation method. According to a study by Junninen et al mean imputation always disrupts the intrinsic structure of the data and causes a high bias in correlation, so it is not a good method for imputation, especially if there is a high percentage of missing (28).

The nature of air pollution data is a type of time series and always depends on the previous and next information. Methods as linear interpolation, moving average, and nearest neighbor have computational nature based on before and after information. Thus, in comparison to other methods, they showed better agreement and fit, according to the all performance criteria.

Engels and Diehr compared several regression methods based on predictor variables such as hot deck, mean and median of a column, row, previous rows, before and after rows (30). They concluded that the methods based on previous observations, and before and after methods are better methods to place missing values in longitudinal data.

Since the issue of missing data is one of the main problems in environmental data, many studies have been conducted and published in this regard, but no paper was found to use the PMM imputation method considering spatial and temporal dependencies in air pollution data.

Several studies have extensively discussed the missing issues in environmental and air pollution data, and examined appropriate models for them (12,28,36,53-59).

According to the study of Norazian et al (12) who examined the three models of linear, quadratic, and cubic interpolation methods, the linear interpolation method was better than the quadratic and cubic method.

According to a review of the literature on air quality studies, several studies have been conducted on imputation methods. Junninen et al collected air quality data along with meteorological data collected simultaneously at two stations to evaluate several imputation methods, including univariate single imputation (linear, spline, and nearest neighbor), multivariate single imputation (regression-based nest, nearest neighbor), a combination of univariate and multivariate imputation and multiple imputation methods (calculation of average methods used multivariate and hybrid methods) (28).

In another study, Plaia and Bondì (54) used air quality

data collected at eight nearby stations to evaluate several methods including univariate single imputation (hourly average, mean before and after), multivariate single imputation (mean of simultaneous values measured at close stations), and multivariate multiple imputation methods (model-based multiple imputation).

Troyanskaya et al used the mean, median, EM algorithm, nearest neighbor, sequential nearest neighbor, and singular value decomposition methods to impute the missing values on gene microarray data, so that the EM algorithm, nearest neighbor, and sequential nearest neighbor appeared to perform well (60).

Conclusion

Imputation methods were used to estimate 3 randomly simulated missing data pattern in PM_{10} and O_3. Due to the nature of the air pollution data, which is time series, methods that depended on before and after information showed good performance. By considering spatial and temporal dependent information in imputation, it was found that PMM technique did not perform well at all percentages of missing. Therefore, in order to choose the appropriate imputation method with the best performance, it is very important to pay attention to the type of data examined.

Limitations

The main limitation of this study was the lack of data in summer and winter in all stations.

Acknowledgments

The authors appreciate the Health and Environment Research Center because of financial support. This article has been extracted from the thesis submitted for MSc degree in Biostatistics which has been approved by the ethics committee of Tabriz University of Medical Sciences, Tabriz, Iran.

Ethics issues

This article has no ethical issues (Ethical code: IR.TBZMED.REC.1398.352).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PS (First author): Formulation and evaluation of overarching research goals and aims; setting the data in software package format; application of statistical, computational, and other formal techniques to analyze; application of available software codes; preparation (drafting, reviewing, translating, and revising the paper), and presentation of the manuscript. PS (Corresponding author): Formulation and evaluation of overarching research goals and aims; statistical analysis; preparation (drafting, reviewing, translating, and revising the paper), and presentation of the manuscript. Akbar Gholampour: Reviewing the paper. All authors have read and approved the final manuscript.

References

1. Kamarehie B, Ghaderpoori M, Jafari A, Karami M, Mohammadi A, Azarshab K, et al. Estimation of health effects (morbidity and mortality) attributed to PM10 and PM2.5 exposure using an Air Quality model in Bukan city, from 2015-2016 exposure using air quality model. Environ Health Eng Manag 2017; 4(3): 137-42. doi: 10.15171/ehem.2017.19.
2. Kim KH, Kabir E, Kabir S. A review on the human health impact of airborne particulate matter. Environ Int 2015; 74: 136-43. doi: 10.1016/j.envint.2014.10.005.
3. Atkinson RW, Fuller GW, Anderson HR, Harrison RM, Armstrong B. Urban ambient particle metrics and health: a time-series analysis. Epidemiology 2010; 21(4): 501-11. doi: 10.1097/EDE.0b013e3181debc88.
4. Kushkbaghi S, Ehrampoush MH, Mirhosseinidehabadi SA. Assessment of role of concrete factories in particulate matter emissions, 2015-2016, using the AQI index and zoning by GIS software (Case study: Nasr Kashan Concrete Factory). Environ Health Eng Manag 2017; 4(3): 149-55. doi: 10.15171/EHEM.2017.21.
5. Chock DP, Winkler SL, Chen C. A study of the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. Journal of the Air & Waste Management Association 2000; 50(8): 1481-500. doi: 10.1080/10473289.2000.10464170.
6. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: a review of the effects of particulate matter air pollution on human health. J Med Toxicol 2012; 8(2): 166-75. doi: 10.1007/s13181-011-0203-1.
7. Amann M, Derwent D, Forsberg B, Hanninen O, Hurley F, Krzyzanowski M, et al. Health Risks of Ozone from Long-range Transboundary Air Pollution. Denmark: World Health Organisation; 2008.
8. Keshtgar L, Shahsavani S, Maghsoudi A, Anushiravani A, Zaravar F, Shamsedini N, et al. Investigating the relationship between the long-term exposure to air pollution and the frequency of depression in Shiraz during 2010-2017. Environ Health Eng Manag 2021; 8(1): 9-14. doi: 10.34172/EHEM.2021.02.
9. Cadelis G, Tourres R, Molinie J. Short-term effects of the particulate pollutants contained in Saharan dust on the visits of children to the emergency department due to asthmatic conditions in Guadeloupe (French Archipelago of the Caribbean). PLoS One 2014; 9(3): e91136. doi: 10.1371/journal.pone.0091136.
10. Shah AS, Langrish JP, Nair H, McAllister DA, Hunter AL, Donaldson K, et al. Global association of air pollution and heart failure: a systematic review and meta-analysis. Lancet 2013; 382(9897): 1039-48. doi: 10.1016/S0140-6736(13)60898-3.
11. Imtiaz SA, Shah SL. Treatment of missing values in process data analysis. Canadian Journal of Chemical Engineering 2008; 86(5): 838-58. doi: 10.1002/cjce.20099.
12. Norazian MN, Shukri YA, Azam RN, Al Bakri AM.

Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 2008; 34(3): 341-5. doi: 10.2306/scienceasia1513-1874.2008.34.341.
13. Hawthorne G, Hawthorne G, Elliott P. Imputing cross-sectional missing data: Comparison of common techniques. Australian and New Zealand Journal of Psychiatry 2005; 39(7): 583-90. doi: 10.1111/j.1440-1614.2005.01630.x.
14. Chatfield C. 19. Statistical Analysis with Missing Data. Journal of the Royal Statistical Society. Series A 1988; 151(2): 375-76. doi: 10.2307/2982783.
15. MSC EPIDEMIOLOGY. Modern methods of data analysis. [cited 2021 Jun] Available from: https://msc-epidemiology.online/courses/modern-methods-in-data-analysis/.
16. Hamzah FB, Hamzah FM, Razali SF, Jaafar O, Abdul Jamil N. Imputation methods for recovering streamflow observation: A methodological review. Cogent Environmental Science 2020; 6(1): 1745133. doi: 10.1080/23311843.2020.1745133.
17. Alsaber AR, Pan J, Al-Hurban A. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). Int J Environ Res Public Health 2021; 18(3): 1333. doi: 10.3390/ijerph18031333.
18. Gill MK, Asefa T, Kaheil Y, McKee M. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. Water Resources Research 2007; 43(7). doi: 10.1029/2006WR005298.
19. Moritz S, Bartz-Beielstein T. imputeTS: time series missing valueImputation in R. R J 2017; 9(1): 207-18.
20. Ishak AB, Daoud MB, Trabelsi A. Ozone concentration forecasting using statistical learning approaches. Journal of Materials and Environmental Science 2017; 8(12): 4532-43. doi: 10.26872/jmes.2017.8.12.478.
21. Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmospheric Environment 2015; 102: 96-104. doi: 10.1016/j.atmosenv.2014.11.049.
22. Valdiviezo HC, Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sciences 2015; 311: 163-81. doi: 10.1016/j.ins.2015.03.018.
23. Kotsiantis S, Kostoulas A, Lykoudis S, Argiriou A, Menagias K. Filling missing temperature values in weather data banks. 2nd IET International Conference on; 2006 Jul 5-6; Athens: IET; 2006.
24. Allen RJ, DeGaetano AT. Estimating missing daily temperature extremes using an optimized regression approach. International Journal of Climatology 2001; 21(11): 1305-19. doi.org/10.1002/joc.679.
25. McLachlan GJ, Peel D. Finite Mixture Models. USA: John Wiley & Sons; 2000.
26. Rubin DB. Multiple Imputation for Nonresponse in Surveys. USA: John Wiley & Sons; 1987.
27. Schafer JL. Analysis of Incomplete Multivariate Data. USA: CRC Press; 1999.
28. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 2004; 38(18): 2895-907. doi: 10.1016/j.atmosenv.2004.02.026.
29. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. UK: Cambridge University Press; 2006. doi: 10.1017/CBO9780511790942.
30. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology 2003; 56(10): 968-76. doi: 10.1016/s0895-4356(03)00170-7.
31. Marwala T. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. New York: Information Science Reference; 2009.
32. Allison PD. Missing Data. USA: SAGE Publications; 2002. doi: 10.4135/9781412985079.
33. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10): 1087-91. doi: 10.1016/j.jclinepi.2006.01.014.
34. Chapra SC, Canale RP. Numerical Methods for Engineers. 6th ed. USA: McGraw-Hill Higher Education; 2010.
35. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 1977; 39(1): 1-22. doi: 10.1111/j.2517-6161.1977.tb01600.x.
36. Lepkowski JM, Raghunathan TE, Solenberger P, Van Hoewyk J. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001; 27(1): 85-96. doi: 12-001-x/2001001.
37. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol 2010; 172(9): 1070-6. doi: 10.1093/aje/kwq260.
38. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton: Routledge; 1984.
39. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. USA: Springer; 2009.
40. Gold MS, Bentler PM. Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Struct Equ Modeling 2000; 7(3): 319-55. doi: 10.1207/S15328007SEM0703_1.
41. Scheffer J. Dealing with missing data. Research Letters in the Information and Mathematical Sciences 2002; 3:153-60.
42. Landerman LR, Land KC, Pieper CF. An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research 1997; 26(1): 3-33. doi: 10.1177/0049124197026001001.
43. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol 2014; 14: 75. doi: 10.1186/1471-2288-14-75.
44. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica 2014; 68(1): 61-90. doi: 10.1111/stan.12023.

45. Yuan YC. Multiple imputation for missing data: Concepts and new development (Version 9.0). [cited 2021 Apr] Available from: https://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf.
46. van Buuren S, Groothuis-Oudshoorn K, Vink G, Schouten R, Robitzsch A, Rockenschaub P. Package 'mice':Multivariate Imputation by Chained Equations. [cited 2021 Jan 30] Available from: https://cran.r-project.org/web/packages/mice/mice.pdf.
47. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. USA: Chapman and Hall; 2018.
48. Widaman KF. Missing data: What to do with or without them. Monographs of the Society for Research in Child Development 2006; 71(3): 42-64. doi: 10.1111/j.1540-5834.2006.00404.x.
49. Dong Y, Peng CY. Principled missing data method for researchers. Springerplus 2013; 2(1): 222. doi: 10.1186/2193-1801-2-222.
50. Willmott CJ. On the Evaluation of Model Performance in Physical Geography. In: Gaile GL, Willmott CJ. Spatial Statistics and Models. Dordrecht: Springer; 1984. p. 443-60.
51. Legates DR, McCabe Jr GJ. Evaluating the use of goodness of fit measures in hydrologic and hydroclimatic model validation. Water Resources Research 1999; 35(1): 233-41. doi: 10.1029/1998WR900018.
52. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology 1970; 10(3): 282-90. doi: 10.1016/0022-1694(70)90255-6.
53. Schneider T. Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. Journal of Climate 2001; 14(5): 853-71. doi: 10.1175/1520-0442(2001)014<0853:aoicde>2.0.CO;2.
54. Plaia A, Bondì AL. Single imputation method of missing values in environmental pollution data sets. Atmos Environ 2006; 40(38): 7316-30. doi: 10.1016/j.atmosenv.2006.06.040.
55. Kondrashov D, Ghil M. Spatio-temporal filling of missing points in geophysical data sets. Processes Geophys 2006; 13(2): 151-9. doi: 10.5194/npg-13-151-2006.
56. Marlinda AM. Rainfall data in-filling model with expectation maximization and artificial neural network [dissertation]. Malaysia: Universiti Teknologi Malaysia; 2008.
57. Pollice A, Lasinio GJ. Two approaches to imputation and adjustment of air quality data from a composite monitoring network. J Data Sci 2009; 7(1): 43-59.
58. Zainuri NA, Jemain AA, Muda N. A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana 2015; 44(3): 449-56.
59. Kamaruzaman IF, Zin WZ, Ariff NM. A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. Malaysian Journal of Fundamental and Applied Sciences 2017; 13(4-1): 375-80. doi: 10.11113/mjfas.v13n4-1.781.
60. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17(6): 520-5. doi: 10.1093/bioinformatics/17.6.520.