Indian Journal of Science and Technology, *Vol 10(19), DOI: 10.17485/ijst/2017/v10i19/110646, May 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Comparison of Multiple Imputation Methods for Data with Missing Values

Geeta Chhabra[1]*, Vasudha Vashisht[2] and Jayanthi Ranjan[3]

[1]Amity School of Institute Technology, Amity University, Noida – 201313, Uttar Pradesh, India; geeta_chhabra@rediff.com
[2]Department of Computer Science and Engineering, Amity School of Engineering, Amity University, Noida – 201313, Uttar Pradesh, India; vvashisht@amity.edu
[3]Institute of Management Technology, Ghaziabad – 201001, Uttar Pradesh, India; jranjan@imt.edu

Abstract

Missing data is relatively common in all type of research, which can reduce the statistical power and have biased results if not handled properly. Multivariate Imputation by Chained Equations (MICE) has emerged as one of the principled method of addressing missing data. This paper provides comparison of MICE using various methods to deal with missing values. The chained equations approach is very flexible and can handle various types of data such as continuous or binary as well as various missing data patterns. **Objectives**: To discuss commonly used techniques for handling missing data and common issues that could arise when these techniques are used. In particular, we will focus on different approaches of one of the most popular methods, Multiple Imputation using Chained Equations (MICE). **Methods/Statistical Analysis**: Multivariate Imputation by Chained Equation is a statistical method for addressing missing value imputation. The paper will focus on Multiple Imputation using Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, Multiple Linear Regression using Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART), Multiple Linear Regression with Bootstrap Imputation which provides a general framework for analyzing data with missing values. **Findings:** We have chosen to explore Multiple Imputation using MICE through an examination of sample data set. Our analysis confirms that the power of Multiple Imputations lies in getting smaller standard errors and narrower confidence intervals. The smaller is the standard error and narrower is the confidence interval; the predicted value is more accurate, thus, minimizing the bias and inefficiency considerably. In our results from sample data set, it has been observed that standard error and mean confidence interval length is the least in case of Multiple Imputation combined with Bayesian Regression. Also, it is obvious from the density plot that the imputed values are more close to the observed values in this method than other methods. Even in case of random forest, the results are quite close to Bayesian Regression. **Application/Improvements:** These Multiple Imputation methods can further be combined with machine learning and Genetic Algorithms on real set data to further reduce the bias and inefficiency.

Keywords: Missing Completely at Random, Missing at Random (MAR), Multiple Imputation, Not Missing at Random (NMAR)

1. Introduction

There are different types of data collection techniques available, namely, the Computer Assisted Personal Interview (CAPI), the Computer Assisted Telephonic Interview (CATI) or the Web Assisted Personal Interview (WAPI). However, no single procedure can ensure the complete data sets and there is always a chance of having errors. Therefore, such data sets are likely to have missing values or values which are not consistent. One of the reasons of missing values is due to the refusal of answers to certain questions, while inconsistent values are generated when the answers are not recorded properly. Data cleaning is one of the important processes associated with data knowledge discovery and the treatment of missing values.

*Author for correspondence

The easiest way of dealing with missing data is to delete all incomplete cases and continue the analysis only with the complete cases[1]. Though this technique can be efficient in simplifying the problem in hand, but it can generate serious bias and inefficiency, especially when the number of such cases is large as compared to the sample size and/or when there is a specific reason for missing data relevant to the study; it could be, for example, that wealthier people are more reluctant to share their gross income or vice versa.

One way to deal with this is data imputation using standard procedures in which missing value data set is replaced by computed values. In other words, data imputation is capable of filling in non-response missing values by generating a complete data set with errors. Many traditional methods of imputation approach are the mean/mode, hot-deck imputation and regression models.

One of the statistical techniques is Multiple Imputation (MI), widely used for dealing with missing data. It provides practical ways for dealing with incomplete data. Instead of substituting missing value with a single value, Multiple Imputation procedures substitute it by a set of plausible values that reflects uncertainty about which value is to substitute; that is why these multiple imputed data sets are analyzed using standard techniques so as to combine the results from these analyses. The procedure of combining inference from distinct data sets is essentially the same[2–4], it does not matter which complete-data analysis is used.

In this paper data imputation is formulated as missing values estimation problem using most powerful and flexible imputation methods known as Multivariate Imputation by Chained Equation (MICE). In this method the first step is to produce multiple imputed datasets. The regression equations are used to predict missing values using regression algorithms combined with different methods like Predictive Mean Matching, Multiple Random Forest[5–9], Multiple Bayesian, Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART)[10], Multiple Linear Regression using Bootstrap[11,12] etc.

2. Goals and Criteria

The goal of statistical methods with or without missing values should be to make efficient and valid inferences about the sample data rather than to predict, estimate or retrieve missing data or to get the similar output that is with complete data. The inference may be affected in an attempt to recover missing data. The usual practice of substituting missing value with the average of the observed values can distort estimated variance and correlation though accurately predicts missing data. The behavior of an estimate can be described by mean square error, bias and variance, but it is also desired that the measures of uncertainty may be reported. The estimated standard error and the true standard error should be close. The 95% interval confidence intervals should cover the true population. The probability of Type 1 error is accurate, only if the coverage rate is accurate. Subject to precise coverage, it is desired that the confidence interval should be small, as the smaller confidence intervals will increase the power and reduce the Type 2 error[12,13].

3. Missing Data Patterns and Mechanisms

Survey statisticians have distinguished unit non-responses from item non-response. The Unit non-response arises when the complete data collection for a particular case is non-reported (because the sampled instance is unavailable, decline to take part, etc.) and item non-response arises when some data is unavailable (i.e. the sample instance take part but refuses to answer some questions). The unit non-response values are treated by reweighting and item non-response are treated by single imputation. These traditional methods perform well in some particular situations, but modern techniques (e.g. Machine Learning and Multiple Imputation) outperform the traditional methods.

The types of missing data fit into three categories which are based on the relationship between the missing data mechanism, the observed and missing values[14]. These categories are important to understand because the problems caused by missing values and the solution to these problems are different for these three categories.

Some data sets are organized in a matrix or rectangular form where the columns represent items or variables and the rows represent observational units. The data in rectangular form depicts some important classes of missing data patterns. In Figure 1(a), the missing values appear in an item X_k only, while in other items it is fully observed, is known as univariate pattern. In monotone pattern, Figure 1(b), items or item groups $X_1, \ldots, X_k$ may be organized in such a way that if X_i is missing, then $X_i+1, \ldots, X_k$ are also missing, Figure 1(c) depicts an arbitrary pattern which is random scatter of missing data for any unit[15].

Missing data mechanism can be viewed as[16–19];

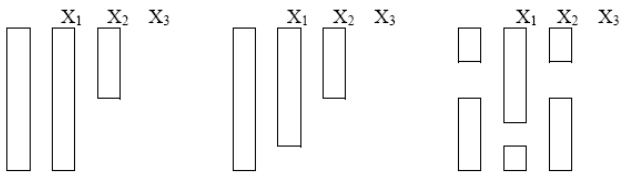


Figure 1. Different missing value patterns. **(a)**. Univariate, f=Figure, 1 **(b)**. Monotonef, Figure 1**(c)**. General.

- **Missing Completely at Random (MCAR)**. In MCAR, missing values for an attribute neither dependent on observed data nor on unobserved data. In this type of randomness, there is no risk of introducing biasness whatever missing value treatment is applied.
- **Missing at Random (MAR)**. In MAR, an instance with missing value is dependent on any of the observed values, but is independent of the unobserved data.
- **Not Missing at Random (NMAR)**. In NMAR, an instance with missing value depends on unobserved data. There are several methods available to treat this type of randomness. One of these methods is instance substitution or substitution with mean or mode. This is naive way but likely to introduce bias, therefore should be carefully handled.

Often MCAR and MAR mechanisms are referred to as ignorable and NMAR as non-ignorable missing value mechanism.

3.1 Methods of Handling Missing Data

3.1.1 Single Imputation

Missing value variables are replaced by one single value within a data set and then analyzed as if all data were originally observed. It generally results in small variance and may produce bias results. Several types of single imputation, which includes[20]:

3.1.2 Mean and Median Imputation

Mean or Median from the observed data is being used to replace the missing values and it generates biased parameter estimates because missing values are substituted with measure of central tendency of the distribution[16,18,20,21].

3.1.3 Regression Imputation

This method assumes the linear relationship between variables. It assumes that the value of one variable changes in linear way with the other variables. The missing values are replaced by a linear regression function instead of replacing all missing data with a statistics. But in most of the cases, the relationship is not linear and using regression imputation in such cases will bias the model[20].

3.1.4 Hot (cold) Deck Imputation

An estimated distribution of the observed data is used to replace a missing item value in hot deck imputation. It is implemented in two stages: a) Data is partitioned into clusters. b) Missing values are replaced within a cluster. The variable mean or mode of a cluster is used to fill the missing values. In Random Hot Deck, an observed value of an attribute is selected randomly to substitute the missing value. Cold deck imputation resembles hot deck but the incomplete value of a variable is substituted by the value from the data source other than the data source in question[5,20].

3.2 Multiple Imputations

Is a statistical method for dealing with missing values. It follows three stages: data imputation, data analysis and data pooling[22]. Figure 2 depicts these stages;

3.2.1 Imputation

In this step m imputed data sets are generated from a distribution which results in m complete data sets. The distribution can be different for each missing entry.

3.2.2. Analysis

The analysis is performed on each m imputed data sets known as complete data analysis.

3.2.3 Pooling

The output obtained after data analysis is pooled to get the final result using simple rules.

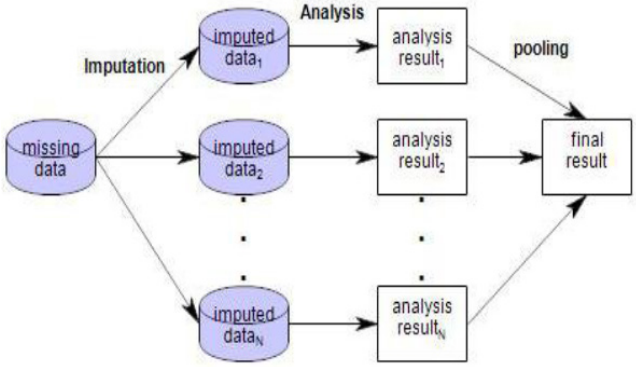


Figure 2. Multiple Imputation method mechanism.

The resulting inferences are statistically valid if the methods to create imputations are 'proper'.

The Multiple Imputation method used to substitute missing values with possible solutions. The incomplete data set is transformed into complete data set by using imputation methods that can then be analyzed by any standard analysis method. Therefore, the multiple imputations have gained popularity to handle missing data.

In Multiple Imputation method, the process is repeated multiple times as the name itself suggests for all variables having missing values and then analyzed to combine m number of imputed data set into one imputed data set. 'R' provides easy to use MICE package for this.

The comparison of six Multiple Imputation methods in MICE has been done in this paper[23,24]:

- Predictive Mean Matching.
- Multiple Random Forest Regression Imputation.
- Multiple Bayesian Regression Imputation.
- Multiple Linear Regression using Non-Bayesian Imputation.
- Multiple Classification and Regression Tree (CART).
- Multiple Linear Regression with Bootstrap Imputation

The predictive mean matching technique is an attractive technique available for missing value substitution in case of quantitative variables. It is somewhat similar to the regression technique which equate observed value with the missing value so that it is close to the predicted mean[25,26]. It uses the linear regression and the nearest-neighbor together to estimate the values.

In random forest, a forest of classification or regression trees is constructed using bootstrap-or subsamples of the original data and the majority vote or overall average of trees[8,9,27] generate the prediction rule for the target variable.

Bayesian Linear Regression is a form of linear regression within the context of Bayesian inference[27–29].

CART is an algorithm for both classification and regression that uses decision trees which are binary to classify new data.

Multiple Linear Regression using Bootstrap Imputation uses any test or metric that relies on random sampling with replacement.

3.3 Experimental Analysis

The paper uses iris dataset from UC Irvine Machine Learning Repository[30]. The iris data has 3 classes, each

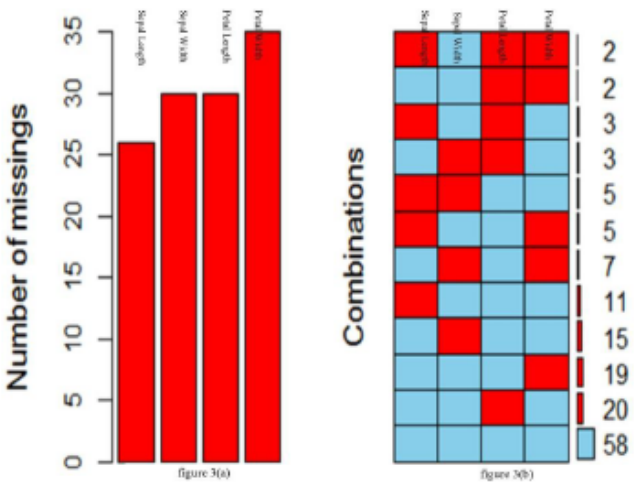
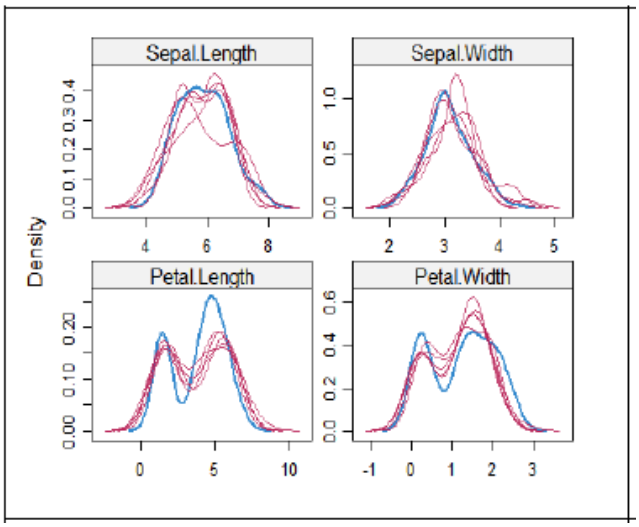
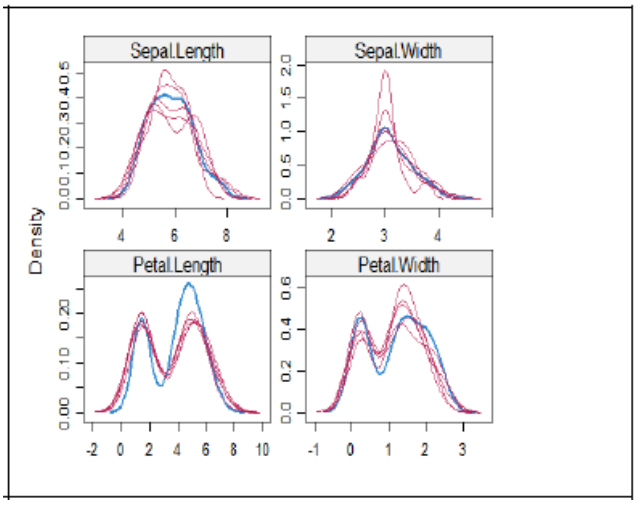


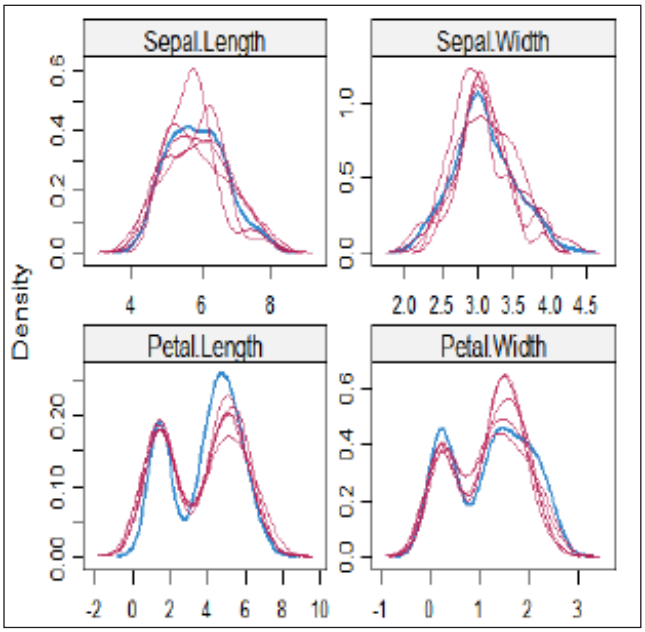
Figure 3. Missing value mechanism in sample data.



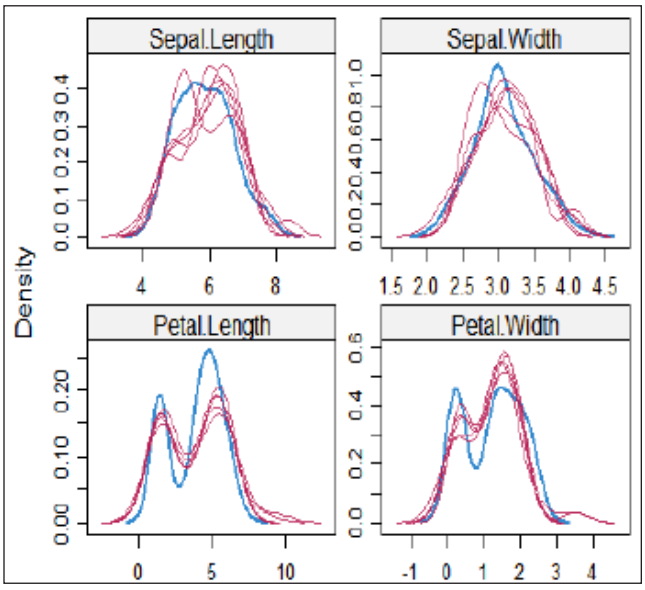
(a)



(b)



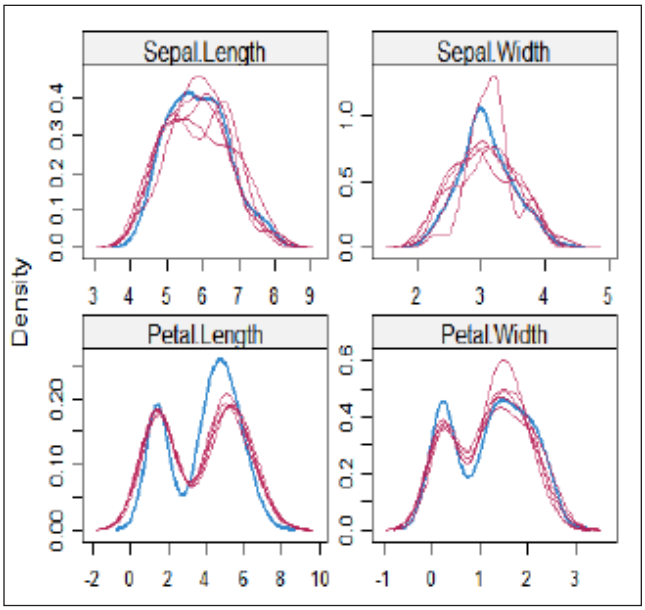
(c)



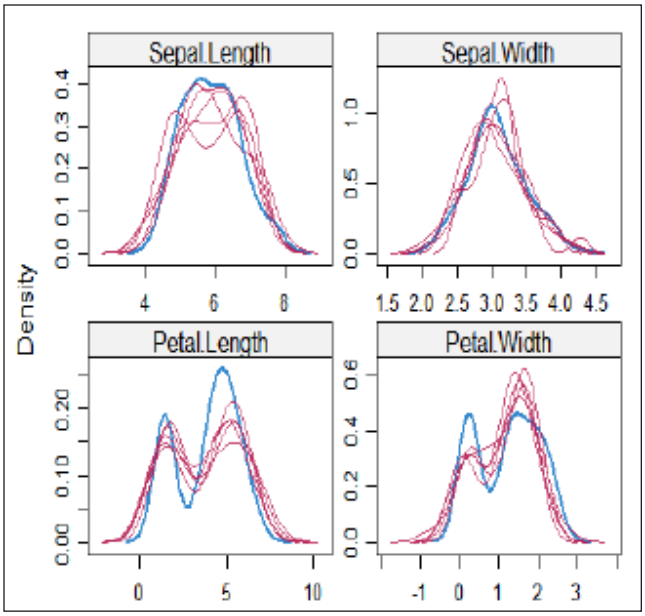
(d)

having 50 cases. Each class is represented by a species of iris plant. The data has four continuous features viz. sepal width, sepal length, petal width, petal length, all measured in cms. These four continuous features have been introduced artificially with about 20% missing values. The petal width has the highest number of missing values in Figure 3(a). There are 58 observations, Figure 3(b) which has no missing value, 20 which has petal length, Figure 3(b) as missing value and so on.

And then, they have been imputed with the multiple imputation combined with PMM, Random Forest, CART, Bayesian, Non-Bayesian and Bootstrapping.



(e)



(f)

Figure 4. Density plot comparison between observed and imputed values. **(a).** Multiple Bayesian regression imputation. **(b).** Multiple random forest regression imputation. **(c).** Predictive mean matching. **(d).** Multiple linear regression using Non-Bayesian. **(e).** Multiple Classification and Regression Tree (CART). **(f).** Multiple linear regression with bootstrap.

Figure 4(a)-4(f) shows that the Density Plot comparison between observed and imputed values.

Table 1. Comparison of different Multiple Imputation methods

S. No.	Method	Mean Standard Error	Mean C.I Length
1	Predictive Mean Matching	0.10608496	0.4533471
2	Multiple Random Forest Regression Imputation	0.09765137	0.4216084
3	Multiple Bayesian Regression Imputation	0.09503033	0.3847437
4	Multiple Linear Regression using Non-Bayesian Imputation	0.11876531	0.5388169
5	Multiple Classification and Regression Tree (CART)	0.10915661	0.4670749
6	Multiple Linear Regression with Bootstrap Imputation	0.11446101	0.4981347

4. Conclusions

Missing data is a part of almost all research and there are various ways to handle the missing data. In the present study, we performed a comparison of different approaches of MICE methods based on iris datasets from UC Irvine Machine Learning Repository, under an MCAR assumption. Validation of imputation results is an important step and we considered two evaluation criteria, namely standard error and mean confidence interval length. Overall, results of performance are summarized in Table 1. Standard error and mean confidence interval length is the least in case of Multiple Imputation combined with Bayesian Regression. Also from the density plot it is obvious that in case of Bayesian the imputed values are close to the observed values. The results of Multiple Random Forest Regression Imputation are also close to Multiple Bayesian Regression Imputation. A possible explanation for the efficiency gain with Multiple Imputation combined with Bayesian Regression is that it is able to make better use of the available information by accommodating nonlinearities among the predictors.

5. Reference

1. Li H. missing values imputation based on iterative learning. International Journal of Intelligence Science. 2013; 3(1):50–5. Crossref
2. Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 1994; 9(4):538–73. Crossref
3. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple Imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research. 2014; 24(4):462–87. PMid: 24525487 PMCid: PMC4513015. Crossref
4. Burgette LF, Reiter JP. Multiple Imputation for missing data via sequential regression trees. American Journal of Epidemiology. 2010; 172:1070–6. PMid: 20841346. Crossref
5. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. International Statistical Review. 2010; 78(1):40–64. PMid: 21743766. PMCid: PMC3130338. Crossref
6. Doove LL, Buuren SV, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics and Data Analysis. 2014; 72:92–104. Crossref
7. Breiman L. Random forests. Machine Learning. 2001Jan; 26(2):123–40.
8. Lin Y, Jeon Y. Random forests and adaptive nearest neighbours. Journal of the American Statistical Association. 2006; 101(474):578–90. Crossref
9. Biau G. Analysis of a random forests model. The Journal of Machine Learning Research. 2012; 13(1):1063–95.
10. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. Computational Statistics and Data Analysis. 1996; 22(4):425–46. Crossref
11. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A Caliber Study. American Journal of Epidemiology. 2014; 179(6):764–74. PMid: 24589914. PMCid: PMC3939843. Crossref
12. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning. 2nd ed. Springer Series in Statistics. 2009. Crossref
13. Mendez G, Lohr S. Estimating residual variance in random forest regression. Computational Statistics and Data Analysis. 2011; 55(11):2937–50. Crossref
14. Zahed H. Bayesian treatment of missing data using multiple imputation. EPPS 7390, Fall; Dallas: University of Texas; 2013. p. 1–243.
15. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods, American Psychological Association. 2002; 7(2):147–77.
16. Gimpy MDRV. Missing value imputation in multi attribute data set. International Journal of Computer Science and Information Technologies. 2014; 5(4):1–7.
17. Dane S, Thool RC. Imputation method for missing value estimation of mixed-attribute data sets. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 May; 3(5):1–6.

18. Kaiser J. Dealing with missing values in data. Journal of Systems Integration. 2014; 5(1):42–51. Crossref
19. Young W, Weckman G, Holland W. A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. Taylor and Francis. 2010 Jun; 12(1):15–43.
20. Pigott TD. A review of missing data treatment methods. Educational Research and Evaluation. 2001; 7(4):353–83. Crossref
21. Zhang S, Zhang J, Zhu XF, Qin YQ, Zhang C. Missing value imputation based on data clustering. Springer; 2008. p. 128–38. Crossref
22. Rezvan PH, Lee KJ, Simpson JA. The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. BMC Medical Research Methodology. 2015. p. 1–67.
23. Nookhong J, Kaewrattanapat N. Efficiency comparison of data mining techniques for missing-value imputation. Journal of Industrial and Intelligent Information. Suan Sunandha Rajabhat University, Bangkok, Thailand. 2015 Dec; 3(4):1–5.
24. Schmitt P, Mandel J, Guedj M. A comparison of six methods for missing data imputation. J Biomet Biostat. 2015; 6(1):1–6.
25. Vink G, Frank LE, Pannekoek J, Buuren SV. Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica. Wiley Publishing. 2014; 68(1):61–90.
26. Towards an MI-proper predictive mean matching. 2016. Crossref
27. Stekhoven DJ, B'uhlmann P. Missforest non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28(1):112–8. PMid: 22039212. Crossref
28. Yu X, Lim ZJS. Replace missing values with EM algorithm based on GMM and Naive Bayesian. International Journal of Software Engineering and its Applications. 2014; 8(5):177–88.
29. LI XB. A Bayesian approach for estimating and replacing missing categorical data. ACM Journal of Data and Information Quality. 2009 Jun; 1(1):1. Crossref
30. Lichman M. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science; 2013. PMid: 24373753.