Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Comparison of Performance of Data Imputation Methods for Numeric Dataset

Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan

To cite this article: Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan (2019) Comparison of Performance of Data Imputation Methods for Numeric Dataset, Applied Artificial Intelligence, 33:10, 913-933, DOI: 10.1080/08839514.2019.1637138

To link to this article: https://doi.org/10.1080/08839514.2019.1637138

Published online: 04 Jul 2019.

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 10, 913–933
https://doi.org/10.1080/08839514.2019.1637138

Taylor & Francis
Taylor & Francis Group

Comparison of Performance of Data Imputation Methods for Numeric Dataset

Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan

Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India

ABSTRACT

Missing data is common problem faced by researchers and data scientists. Therefore, it is required to handle them appropriately in order to get better and accurate results of data analysis. Objective of this research paper is to provide better understanding of data missingness mechanism, data imputation methods, and to assess performance of the widely used data imputation methods for numeric dataset. It will help practitioners and data scientists to select appropriate method of data imputation for numeric dataset while performing data mining task. In this paper, we comprehensively compare seven data imputation methods namely mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob), and random sample. We have used five different numeric datasets obtained from UCI machine learning repository for analyzing and comparing performance of the data imputation methods. Performance of the data imputation methods is assessed using Normalized Root Mean Square Error (RMSE) method. The results of analysis show that kNN imputation method outperforms the other methods. It has also been found that performance of the data imputation method is independent of the dataset and percentage of missing values in the dataset.

Introduction

Quality of data is main concern of data scientists and researchers working in the field of data science and data analytics. Although quality of output of the machine learning algorithm depends on several factors such as feature selection, selection of algorithm, sampling techniques, training, test, and validation datasets, one of the main concerns of the data scientists is how to handle missing data (Brown and John, 2003). Most statistical and machine learning algorithms are not robust enough to handle missing values. They get affected by missing data. Missing data introduces an element of ambiguity while analyzing data and that can affect properties of statistical estimators and results in loss of power and misleading conclusions (Schmitt, Mandel, and Guedj 2015; Somasundaram and Nedunchezhian 2011). Appropriately dealing with missing values is important

CONTACT Anil Jadhav a_s_jadhav74@yahoo.co.in Symbiosis Centre for Information Technology, Symbiosis International (Deemed University)

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

and challenging task because it requires i) careful examination of all instances of data to identify pattern of missingness in the data and ii) clear understanding of different imputation techniques.

Researchers and data scientist community is continuously working on problem of dealing with missing values (Little and Rubin 2002; Rubin 1987; Schafer and Graham 2002). The most accepted way to handle this problem is missing data imputation which is nothing but estimation of plausible values to substitute the missing ones. The main reason of imputation of missing data is to reduce the bias due to missingness rather than deleting incomplete cases. A variety of techniques have been developed for substituting missing values using statistical prediction and this process is generally referred as missing data imputation (Little and Rubin 1989, 2002; Rubin 1987; Schafer 1999; van Buuren and Groothuis-Oudshoorn 2011)

Even though problems associated with missing data are well documented and addressed, it is common practice to ignore missing data and employ analytical techniques that simply deletes all cases that have some values missing in the dataset on any of the variable considered for analysis (Horton and Kleinman 2007). King et al. (2001) in their paper on alternative algorithm for multiple imputation states that approximately 94% use listwise deletion to eliminate entire observation which results in loss of valuable information. It is also stated that multiple imputation will normally better than, and almost always not worse than listwise deletion approach. Repeated Multiple Imputation (MI) method is becoming popular method of data imputation for handling missing data.

Objective of this paper is to provide better understanding of data missingness mechanism, data imputation methods, and to assess performance of most common and widely used data imputation techniques namely mean imputation, median imputation, kNN imputation, predictive mean matching (pmm) imputation, Bayesian Linear regression, Linear regression non Bayesian, and Sample imputation methods. This will help practitioners and data scientists to select appropriate data imputation method while carrying data mining task. Most published articles in this field deals with imputation techniques but there are few studies that reports evaluation of existing methods to provide guidelines to make more appropriate decision during missing data imputation (Schmitt, Mandel, and Guedj 2015; Somasundaram and Nedunchezhian 2011; Zhang and Aytug 2016). Focus of this study is to analyze and compare performance of imputation methods for numeric dataset.

Rest of the paper is organized as follows: Section 2 describes theoretical background and related work. Section 3 describes research methodology for comparison of imputation methods. Section 4 presents results and analyses of performance of imputation methods. The paper is concluded in section 5.

Theoretical Background and Related Work

One of the main concerns in data analysis is to appropriately incorporate missing information. It is very important to note that there is difference in empty and missing value. Empty value means no value can be assigned whereas missing value means actual value for that variable exist but not available or captured in dataset due to some reasons. The data miner should differentiate between empty value and missing value. If not, both the values will be treated as missing value. Missing data may be due to equipment malfunction, inconsistent with other data so deleted, data not entered due to misunderstanding, certain data may not be considered important at the time of data collection. Some data mining algorithms do not require replacement of missing values as they are designed and developed to handle missing values but some data mining algorithms can't deal with missing values.

Before using any method of dealing with missing values it is important to understand why data is missing. Little and Rubin (2002) and Rubin (1976) formulated three possible missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR).

Missing Data Mechanisms

Missing Completely at Random (MCAR): MCAR is the highest level of randomness and it implies that the pattern of missing value is totally random and does not depend on any variable which may or may not be included in the analysis. Thus, if missingness does not depend on any information in the dataset then it means that data is missing completely at random. The assumption of MCAR is that probability of the missingness depends neither on the observed values in any variable of the dataset nor on unobserved part of dataset.

Missing at Random (MAR): In this case, probability of missing data is dependent on observed information in the dataset. It means that probability of missingness depends on observed information but does not depend on the unobserved part. Missing value of any of the variable in the dataset depends on observed values of other variables in the dataset because some correlation exists between attribute containing missing value and other attributes in the dataset. The pattern of missing data may be traceable from the observed values in the dataset.

Missing Not at Random (MNAR): In this case, missingness is dependent on unobserved data rather than observed data. Missingness depends on missing data or item itself because of response variable is too sensitive to answer. When data are MNAR, the probability of missing data is related to the value of the missing data itself. The pattern of missing data is not random and is non predictable from observed values of the other variables in the dataset.

These different types of missing data are important because they determine which statistical treatment of the missing data can be used effectively. MNAR is often considered as worst missing type as it may lead to biased result whereas MCAR and MAR might lead to loss of statistical power (Graham 2009; Schafer and Graham 2002). It is always recommended to collect as much information as possible about the reasons of missing data. There exist some methods that could be used to distinguish between MCAR and not MCAR. Using a *t*-test to compare characteristics of group with missing values and observed values on certain variable. When missing data are not MCAR the two groups will have different characteristics. This test is only indicative because it always depends on sample size of the data. There is no mechanism to test whether missingness is due to MAR or MNAR (Little and Rubin 2002).

Methods of Handling Missing Data

There are two different strategies for handling missing data (Han and Kamber 2012). The first strategy is simply ignore missing values and second strategy is to consider imputation of missing values.

Ignoring Missing Values: The missing data ignoring technique simply omits the cases that contain missing data. They are widely used and tend to be default method for handling missing data. The serious problem with this method is that it reduces the dataset size. This is appropriate when your dataset has small amount of missing values. There are two general approaches for ignoring missing data: listwise deletion (case deletion or complete case analysis) and pairwise deletion (available case analysis) approach. Complete case analysis approach excludes all observations with missing values for any variable of interest. This approach thus limits the analysis to those observations for which all values are observed which often results in biased estimate and loss of precision (Schafer and Graham 2002). In pairwise deletion, we perform analysis with all cases in which the variables of interest are present. It does not exclude entire unit but uses as much data as possible from every unit. Advantage of this method is it keeps maximum available data for analysis even some of its variables has missing values. Disadvantage of this method is that it uses different sample size for different variables (Schafer and Graham 2002). The sample size for each individual analysis is higher than the complete case analysis.

Imputation of missing values: Missing data imputation is a procedure that replaces missing value with some plausible values (Rubin 1976). The various imputation techniques aim to provide accurate estimation of population parameters so that power of data mining and data analysis techniques is not reduced. Optimal treatment to be given to the missing data depends on amount of missing data. Although there is no thumb rule on what percentage of missing data is bad, it is always better to do comparison of results before and after imputation if more than 25% data is missing.

Missing Data Imputation Methods

The process of estimating missing data of an observation based on valid values of other variables is called as Data Imputation (Rubin 1976). Much has been published in the statistical literature on missing data (Little & Rubin 1987; Schafer 1997). Data imputation methods are broadly classified into two types: Single Imputation Method and Multiple Imputation Method.

Single Imputation

It refers to imputing one plausible value for each missing value of a particular variable in the dataset and then performing analysis as if all data were originally observed. Some popular single data imputation methods are as follows:

Imputation with the constant: In this method, the missing values are replaced with the constant. In case of categorical variable it could replace all missing values with "Missing" or "0" or "999" value.

Mean Imputation: This is most common method of missing data replacement. It replaces missing value with sample mean or median or mode depending or distribution of the data. This method is easily implementable and simple but this method has drawbacks also. If missing values are large in number then all those values are replaced by same imputation value, that is, mean, which leads to change in shape of the distribution. Standard deviation becomes smaller when you compare it before and after imputation. More the missing values more will be shrinkage in the standard deviation. This method can be slightly improved by stratifying data into subgroups.

Simulation studies show that mean imputation indeed yields highly biased parameter estimates (Graham et al. 1997; Graham, Hofer, and MacKinnon 1996; Graham, Hofer, and Piccinin 1994). However, some studies point out that the limitations of mean imputation are almost absent if less than 10% of the data is missing and when the correlations between the variables are low (Raymond 1986; Tsikriktsis 2005). Two techniques similar to mean imputation are median and modus imputation. Those methods were invented to account for imputation of not normally distributed data, but they suffer from the same limitations as mean imputation and are therefore not very popular methods of data imputation.

Imputation with distributions: In this approach, missing values are replaced by random values from known distribution. The imputed value does not change shape of the distribution.

Regression Imputation: This is somewhat more sophisticated single imputation technique. In this method, missing value is replaced by predicted data using regression based on non missing data of other variables. This method is based on assumption of linear relationship between the attributes. But most of the time relationship is not linear so replacing missing value using regression will bias the model. Advantage of this method over mean imputation method is that regression imputation is able to preserve the distribution

shape. This method may produce biased results especially with MNAR and MAR (Schafer and Graham 2002).

kNN Imputation: In this method, missing values are imputed by copying values from similar records in the same dataset. The similarity of the two attributes is determined using a distance function. Creation of predictive model for each attribute is not required, but it has got disadvantages also. It is very time consuming in analyzing large dataset. Choice of *k* value is also critical.

Multiple Imputation

In single imputation methods it is assumed that single imputation value is correct one and precision is overstated. However, there can never be absolute certainty about validity of imputed values. Therefore uncertainty around these imputed values has to be incorporated in the missing data methods (Little and Rubin 1989). Rubin (1987) developed a method for averaging the outcome across multiple imputed datasets. Thus, in multiple imputation instead of replacing single value for each missing observation it substitutes multiple plausible values to reflect uncertainty about the right values to impute. Thus, Multiple Imputation method generates “m” different complete datasets with observed and imputed values. All multiple Imputation Method follows three steps: (1) Imputation: Similar to single imputation missing values are imputed; however, imputed values are generated "m" times rather than just once. So there could “m” different complete datasets after imputation. (2) Analysis of each dataset: After imputation and generating “m” different datasets each of “m” datasets is analyzed. (3) Pooling: Finally results obtained from each analyzed datasets are consolidated.

Related Work

There are several studies that report various aspects of data imputation methods. The study conducted by Kyureghian, Capps, and Nayga (2011) compares imputation methods by measuring error of predicting the missing values and parameter estimates from subsequent regression analysis. The result of paper shows that multiple imputation methods have best coverage of both parameter estimates and prediction of the dependent variable. Mishra and Khare (2014) in their study explored efficiency and appropriateness of various imputation methods using small size dataset with varying size of missingness. Brown and John (2003) discussed the impact of missing data on various data mining algorithms including decision trees, *k*-nearest neighbor, association rules, and neural network. A study conducted by Geert et al. (2006) found that single imputation method results in too small estimated standard errors, whereas multiple imputation results in correctly estimated standard errors and confidence interval. The study conducted by Penone et al. (2014) evaluates performance of four approaches, for estimating missing values in traits database, namely kNN, mice, missForest, and Phylopars. Schmitt, Mandel, and Guedj (2015) compares

six methods for data imputation. Comparison was performed on four real datasets of various sizes under missing completely at random assumption. Their result suggests that Bayesian principle component analysis and fuzzy *k*-means imputation methods deserves further consideration in practice.

A comprehensive handling of multiple imputation is discussed by Rubin and Schenker (1986), Rubin (1987), Herzog and Rubin (1983), Rubin and Schenker (1986). Tutz and Ramzan (2015) in their paper proposed improved methods for imputation of missing data by nearest neighbor methods. They found that proposed improved nearest neighbor method outperforms competing nearest neighbor methods. Troyanskaya et al. (2001), (2003) compared *k*-nearest neighbor imputation (KNNimpute) with the mean imputation and singular-value decomposition (SVD) techniques for gene expression data. Their simulation study showed that the KNN impute method performs well compared to mean imputation and SVD approaches. In a comparative study of single imputation methods, Malarvizhi and Thanamani (2012) found that median or standard deviation substitution perform better than mean substitution. The study by Nguyen, Wang, and Carroll (2004) compared KNNimpute with mean, ordinary least squares (OLS), and partial least squares (PLS) imputation methods in microarray data and demonstrated good performance of KNN method.

Poulos and Valle (2016) compared methods for missing categorical data for supervised learning task using two datasets. Gustavo and Monard (2010) analyzed *k*-nearest method as an imputation method. The result of their analysis showed that kNN method outperforms C4.5 and CN2 method to treat missing data. Ghorbani and Desmarais (2017) analyzed effect of missForest(MF), Multiple Imputation based on Expectation Maximization(MIEM), sequential hot-deck, and multiple imputation based on logistic regression (MILR) on prediction accuracy over binary data. The effect is assessed using four different models namely Tree Augmented Naive Bayes, Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). The result shows that MIEM method gives best results for all the classifiers across different percentages of missing data.

MICE is increasingly popular method for doing multiple imputations (Patric and White 2011; Sterne et al. 2009; van Buuren and Groothuis-Oudshoorn 2011; White, Royston, and Wood 2011). Therefore, we have used MICE package to analyze performance of multiple imputation methods which includes: (a) Predictive Mean Matching (PMM): PMM imputes missing values of a continuous variable "z" such that imputed values are sampled only from the observed values of "z" by matching predicted values as closely as possible. (b) Bayesian Linear Regression: Imputes univariate missing data using Bayesian linear regression analysis. (c) Linear Regression (non-Bayesian): This creates imputation using spread around the fitted linear regression line of "y" given "x" as fitted on the observed data by ignoring model error. (d) Sample: This method takes a simple random sample from the observed data, and imputes these into missing cells. The mathematical

details of how these methods works is described by White, Royston, and Wood (2011).

The related work shows that there exists several literature describing different methods of data imputation. From the implementation perspective, it is also very important to understand and evaluate performance of different imputation methods so that appropriate method can be used while performing data mining task. Though there exist some literatures that analyzed performance of different imputation methods, in this paper, we intend to analyze performance of different imputation method for numeric datasets that uses single and multiple imputation methods namely mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob), and random sample.

Research Methodology

This section describes procedure followed for analyzing performance of the imputation methods. Objective of this study is to analyze performance of imputation methods that includes: (a) Single Imputation Methods: Mean Imputation, Median Imputation, KNN Imputation and (b) Multiple Imputation Methods: Predictive Mean Matching (pmm), Bayesian Linear Regression (norm), Linear Regression non-Bayesian (norm.nob), and Sample method. All these imputation methods are applied only on numeric datasets. The datasets used in this study are obtained from UCI Machine Learning Repository (Lichman 2013). The description of each dataset is given in Table 1.

We first obtained five different datasets described in Table 1 from UCI machine learning repository. Then we injected varying percentage (10%, 20%, 30%, 40%, and 50%) of missing values in each original dataset. The simulated missing values are then imputed using imputation methods namely mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear regression, Linear regression–non-Bayesian, and sample imputation method. Mean and Median imputation is done by calculating mean and median value of the feature in the dataset. kNN imputation is done by using VIM package in R. The details about VIM package is described by Kowarik and Templ (2016). For multiple imputation methods: predictive mean matching, Bayesian Linear Regression, Linear Regression–non-Bayesian, and Sample imputation we have used mice package in R. The mice package is described in detail by van Buuren and Groothuis-Oudshoorn 2011), Sterne et al. (2009), Patric and White (2011), and White, Royston, and Wood (2011). The next step after imputation is to analyze performance of each imputation method.

There exists different ways to measure performance of imputation method such as accuracy, relative accuracy, MAE (mean absolute error), and RMSE (root mean square error). However, RMSE is one of the most representative and widely used performance indicators in the imputation research (Schmitt,

Table 1. Description of the dataset used in the study.

Sr. No	Dataset	Dataset Description	No. of Instances	No. of Attributes
1	Wine Dataset	These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars	178	13
2	Glass Identification	Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis, In determining whether the glass was a type of "float" glass or not	214	11
3	Concrete Comprehensive Strength	Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, age, and concrete comprehensive strength.	1030	9
4	Indian Liver Patient Dataset	This dataset contains 416 liver patient records and 167 nonliver patient records. The dataset was collected from north east of Andhra Pradesh, India.	583	10
5	Seeds Dataset	Measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment.	210	7

Mandel, and Guedj 2015). We have used Mean of Normalized RMSE (NRMSE) as a performance indicator. The reason for using Normalized RMSE is that scales are different for different features of the dataset. Once NRMSE is calculated for each variable in the dataset then Mean of NRMSE is calculated for the dataset and is used as a measure to assess performance of the imputation methods. The formula for calculation of NRMSE and Mean NRMSE is described in the next section. For the purpose of this study, we have used R and RStudio as a tool for data manipulation, data imputation, and analyzing performance of different imputation methods.

Results and Discussion

This section describes evaluation of performance of seven different imputation methods namely mean, median, kNN, pmm, norm, and norm.nob, and sample. Out of these seven methods mean, median, and kNN methods are single imputation methods and pmm, norm, norm.nob, and sample are multiple imputation methods. Approach of single imputation method is to replace missing value by a single value without taking into consideration uncertainty of the imputation. The pmm, norm, norm.nob, and sample are multiple imputation methods. Approach of multiple imputation method is to take into consideration imputation uncertainty by running single imputation multiple times so that it can provide precise estimate of missing values. Multiple imputation approach imputes incomplete dataset "m" times and analyzes "m" imputed datasets. The "m" results of analysis are then pooled in final result. We have

chosen Multivariate Imputation by Chained Equations (MICE) implemented as mice package in R for multiple imputation methods namely pmm, norm, norm.nob, and sample. The pmm method imputes univariate missing data using predictive mean matching. The norm method imputes univariate missing data using Bayesian linear regression analysis. The norm.nob method imputes missing data using linear regression analysis (non-Bayesian version). The norm.nob method creates imputation using the spread around the fitted linear regression line. The sample method imputes missing value by random sample from the observed data. The details about these multiple imputation methods can be found in van Buuren and Groothuis-Oudshoorn (2011).

To assess performance of the imputation methods we first calculate normalized root mean square error (NRMSE) for each variable in the dataset. The formula for calculation of NRMSE for each variable in the dataset is given as follows:

$$\text{NRMSE} = \sqrt{\frac{mean\left((original\ value - imputed\ value)^2\right)}{\max(\text{original value}) - \min(\text{original value})}}$$

Where original value is actual value of the variable and imputed value is value of the variable after imputation.

After calculating NRMSE for each variable in the dataset Mean NRMSE is calculated as follows:

$$\text{Mean NRMSE} = \frac{\sum_{i=1}^{n} NRMSE}{n}$$

Where n is number of variables in the dataset.

Lower is value of Mean NRMSE; better is estimate of the missing values. The Mean NRMSE for each dataset for different percentage of imputed data using different imputation methods is calculated and given in the Tables 2–6. Each column in the table indicates percentage of imputed data and each row indicates method used for imputation of data. The value in bold indicates lowest Mean NRMSE. It means that bold value indicates the imputation method that gives better imputation result when applied on the given dataset. The plot of Imputation Method and corresponding Mean NRMSE for different percentages of the missing values for all datasets used in the study are shown in Figures 1–5. It is observed that as percentage of missing values increases Mean NRMSE also increases. It is also observed that Mean NRMSE for kNN Impute method is lowest across all datasets and all missing data percentages.

In order to assess consistency in performance of each imputation method for different datasets and for different percentage of missing values we have ranked each imputation method based on Mean NRMSE value. Ranks are

Table 2. Mean NRMSE for wine dataset.

Method Used for Data Imputation	Percentage of Imputed Data				
	10%	20%	30%	40%	50%
Mean Imputation	0.222146	0.233951	0.456639	0.401967	0.469818
Median Imputation	0.237216	0.231556	0.468412	0.390347	0.486157
KNN Imputation	**0.087871**	**0.077085**	**0.168043**	**0.159184**	**0.183731**
Predictive Mean Matching (pmm)	0.175954	0.138432	0.333917	0.328648	0.396121
Bayesian Linear Regression(norm)	0.199844	0.178192	0.345408	0.375055	0.388085
Linear Regression, non-Bayesian (norm.nob)	0.175866	0.183636	0.338890	0.371935	0.393525
Random sample from observed values(sample)	0.240903	0.262651	0.489474	0.430074	0.508505

Table 3. Mean NRMSE for glass dataset.

Method Used for Data Imputation	Percentage of Imputed Data				
	10%	20%	30%	40%	50%
Mean Imputation	0.027400	0.035618	0.043044	0.051462	0.055981
Median Imputation	0.029164	0.036457	0.045766	0.053873	0.059491
KNN Imputation	0.011371	**0.016015**	**0.022129**	**0.026034**	**0.029958**
Predictive Mean Matching(pmm)	**0.010689**	0.023725	0.032421	0.041404	0.052936
Bayesian Linear Regression(norm)	0.013462	0.027308	0.033820	0.047581	0.054514
Linear Regression, non-Bayesian (norm.nob)	0.013063	0.029579	0.032958	0.046865	0.056759
Random sample from observed values(sample)	0.030792	0.039062	0.047025	0.054459	0.060170

Table 4. Mean NRMSE for concrete compressive strength dataset.

Method Used for Data Imputation	Percentage of Imputed Data				
	10%	20%	30%	40%	50%
Mean Imputation	0.584359	0.814185	0.993566	1.165786	1.313276
Median Imputation	0.639579	0.872116	1.066602	1.261226	1.426217
KNN Imputation	**0.206486**	**0.362306**	**0.491548**	**0.633543**	**0.718348**
Predictive Mean Matching(pmm)	0.359238	0.591370	0.856465	1.021827	1.263514
Bayesian Linear Regression(norm)	0.401028	0.619730	0.869532	1.061656	1.269752
Linear Regression, non-Bayesian (norm.nob)	0.404694	0.618787	0.869513	1.076473	1.261497
Random sample from observed values(sample)	0.645438	0.877418	1.103078	1.284036	1.426888

Table 5. Mean NRMSE for liver patient dataset.

Method Used for Data Imputation	Percentage of Imputed Data				
	10%	20%	30%	40%	50%
Mean Imputation	0.383758	0.458487	0.599017	0.758199	1.114143
Median Imputation	0.395483	0.478442	0.629800	0.792934	1.149098
KNN Imputation	**0.178745**	**0.230518**	**0.312517**	**0.390083**	**0.574943**
Predictive Mean Matching(pmm)	0.272707	0.396284	0.532760	0.765251	1.098783
Bayesian Linear Regression(norm)	0.308782	0.437789	0.628613	0.809780	1.099082
Linear Regression, non-Bayesian (norm.nob)	0.322347	0.439493	0.635758	0.807315	1.092817
Random sample from observed values(sample)	0.436485	0.555811	0.734842	0.864281	1.160755

Table 6. Mean NRMSE for seeds dataset.

Method Used for Data Imputation	Percentage of Imputed Data				
	10%	20%	30%	40%	50%
Mean Imputation	0.068072	0.106639	0.120149	0.142043	0.160996
Median Imputation	0.069302	0.108637	0.123685	0.141440	0.165698
KNN Imputation	**0.013803**	**0.025372**	**0.035587**	**0.047773**	**0.066471**
Predictive Mean Matching(pmm)	0.017844	0.035328	0.044072	0.057366	0.081049
Bayesian Linear Regression(norm)	0.017917	0.033387	0.046552	0.057673	0.077796
Linear Regression, non-Bayesian (norm.nob)	0.017933	0.033459	0.044446	0.055386	0.072130
Random sample from observed values(sample)	0.072683	0.119587	0.131316	0.153484	0.169020

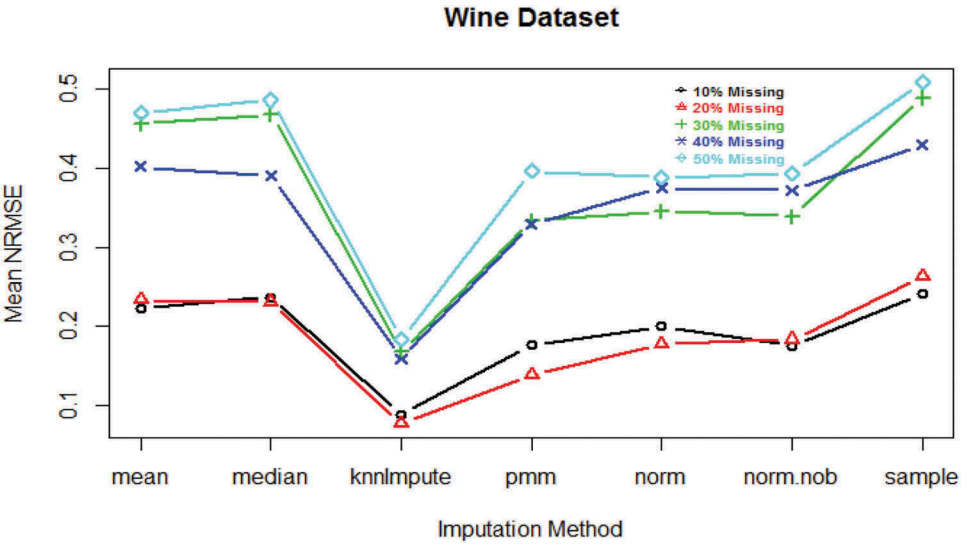

Figure 1. Plot of imputation method versus mean NRMSE for wine dataset.

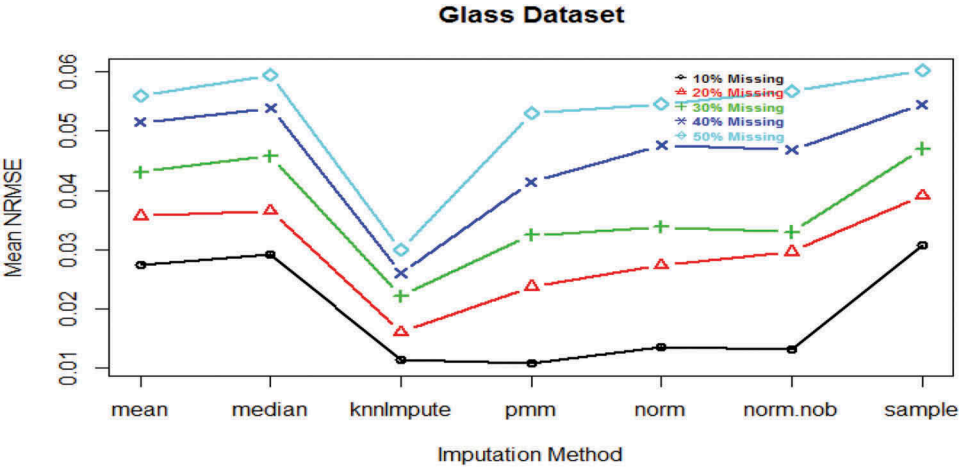

Figure 2. Plot of imputation method versus mean NRMSE for glass dataset.

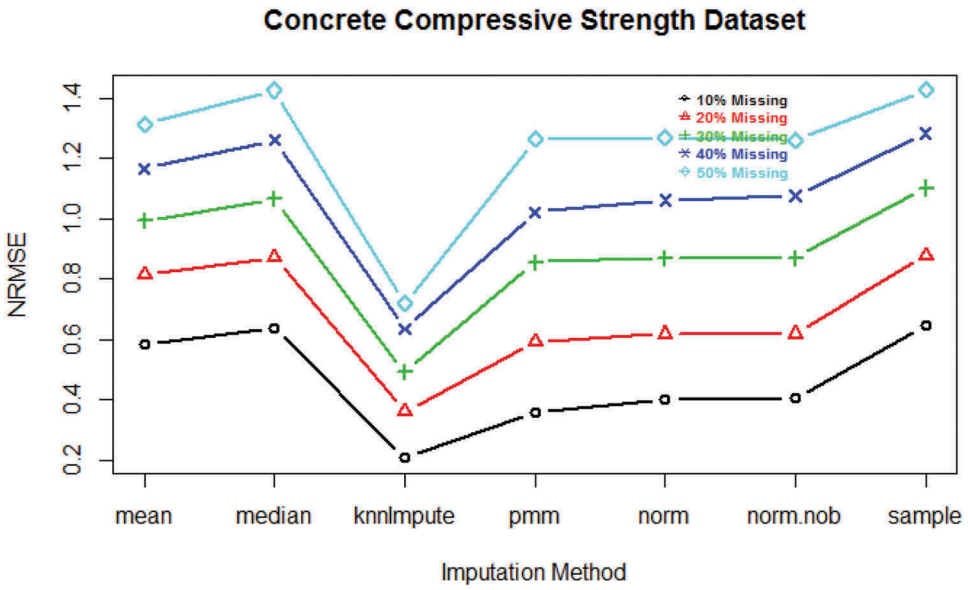

Figure 3. Plot of imputation method versus mean NRMSE for concrete dataset.

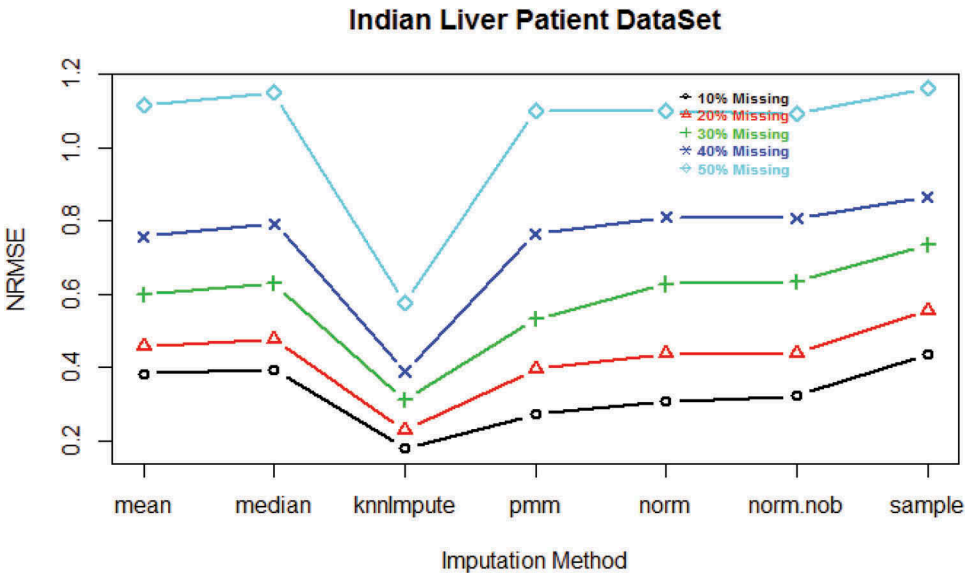

Figure 4. Plot of imputation method versus mean NRMSE for liver patient dataset.

given in ascending order of Mean NRMSE. It means that lowest Mean NRMSE value get first rank.

Tables 7–11 provide rank of imputation method for varying percentage of missing data for five different datasets. Each table indicates performance of imputation method on different datasets for given percentage of imputed data. The reason for doing this is to assess consistency in performance of the imputation methods on five different datasets when percentage of imputed

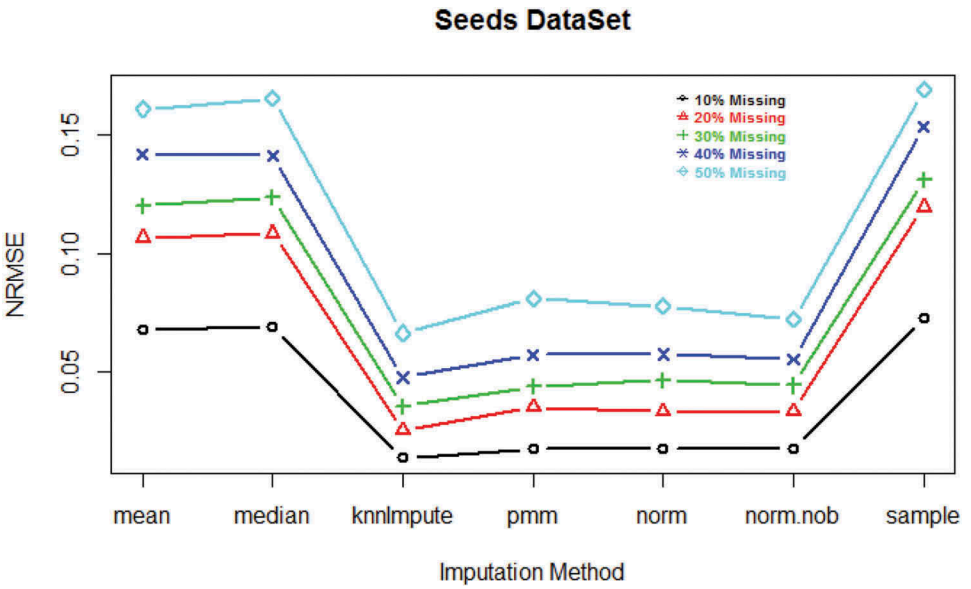

Figure 5. Plot of imputation method versus mean NRMSE for seed dataset.

Table 7. Rank of imputation method for 10% missing data for each dataset.

	Rank of Imputation Method when 10% Missing Values Are Imputed in Each Dataset						
Imputation Method	Wine	Glass	Concrete	Liver	Seed	Rank by Mean	Rank by Mode
Missing Percentage of Data: 10%							
Mean Imputation	5	5	5	5	5	5	5
Median Imputation	6	6	6	6	6	6	6
KNN Imputation	1	2	1	1	1	1.2	1
Predictive Mean Matching (pmm)	3	1	2	2	2	2	2
Bayesian Linear Regression (norm)	4	4	3	3	3	3.4	3
Linear Regression, non-Bayesian (norm. nob)	2	3	4	4	4	3.4	4
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.949, Chi-sq = 28.5, *p* value = 7.70E-05						

data is same. The last two columns in each table indicate the average rank and rank obtained using mode method. The last row in each table provides Kendall's test statistics which is used for testing agreement among the rankings of imputation methods when datasets are different but percentages of imputed data is same.

Tables 12–16 provide rank of each imputation method for given dataset for different percentage of missing values. Each table indicates performance of different imputation method for different percentage of missing data for a given dataset. The reason for doing this is to assess consistency in performance of each imputation method for different percentage of missing data for

Table 8. Rank of imputation method for 20% missing data for each dataset.

Imputation Method	Rank of Imputation Method when 20% Missing Values Are Imputed in Each Dataset					Rank by Mean	Rank by Mode
	Wine	Glass	Concrete	Liver	Seed		
Missing Percentage of Data: 20%							
Mean Imputation	6	5	5	5	5	5.2	5
Median Imputation	5	6	6	6	6	5.8	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	2	2	2	4	2.4	2
Bayesian Linear Regression (norm)	3	3	4	3	2	3	3
Linear Regression, non-Bayesian (norm. nob)	4	4	3	4	3	3.6	4
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.943, Chi-sq = 28.3, p value = 8.30E-05						

Table 9. Rank of imputation method for 30% missing data for each dataset.

Imputation Method	Rank of Imputation Method when 30% Missing Values Are Imputed in Each Dataset					Rank by Mean	Rank by Mode
	Wine	Glass	Concrete	Liver	Seed		
Missing Percentage of Data: 30%							
Mean Imputation	5	5	5	3	5	4.6	5
Median Imputation	6	6	6	5	6	5.8	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	2	2	2	2	2	2
Bayesian Linear Regression (norm)	4	4	4	4	4	4	4
Linear Regression, non-Bayesian (norm. nob)	3	3	3	6	3	3.6	3
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.92, Chi-sq = 27.6, p value = .000112						

Table 10. Rank of imputation method for 40% missing data for each dataset.

Imputation Method	Rank of Imputation Method when 40% Missing Values Are Imputed in Each Dataset					Rank by Mean	Rank by Mode
	Wine	Glass	Concrete	Liver	Seed		
Missing Percentage of Data: 40%							
Mean Imputation	6	5	5	2	6	4.8	6
Median Imputation	5	6	6	4	5	5.2	5
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	2	2	3	3	2.4	2
Bayesian Linear Regression (norm)	4	4	3	6	4	4.2	4
Linear Regression, non-Bayesian (norm. nob)	3	3	4	5	2	3.4	3
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.823, Chi-sq = 24.7, p value = .00039						

Table 11. Rank of imputation method for 50% missing data for each dataset.

Imputation Method	Rank of Imputation Method when 50% Missing Values Are Imputed in Each Dataset					Rank by Mean	Rank by Mode
	Wine	Glass	Concrete	Liver	Seed		
Missing Percentage of Data: 50%							
Mean Imputation	5	4	5	5	5	4.8	5
Median Imputation	6	6	6	6	6	6	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	4	2	3	3	4	3.2	4
Bayesian Linear Regression (norm)	2	3	4	4	3	3.2	3
Linear Regression, non-Bayesian (norm.nob)	3	5	2	2	2	2.8	2
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.906, Chi-sq = 27.2, p value = .000134						

Table 12. Rank of imputation method for wine dataset for different percentage of imputed data.

Imputation Method	Rank of Imputation Method for Different Percentage of Imputed Data					Rank by Mean	Rank by Mode
	10%	20%	30%	40%	50%		
Dataset Name: Wine							
Mean Imputation	5	6	5	6	5	5.4	5
Median Imputation	6	5	6	5	6	5.6	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching(pmm)	3	2	2	2	4	2.6	2
Bayesian Linear Regression(norm)	4	3	4	4	2	3.4	4
Linear Regression, non-Bayesian (norm.nob)	2	4	3	3	3	3	3
Random sample from observed values(sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.923, Chi-sq = 27.7, p value = .000108						

Table 13. Rank of imputation method for glass dataset for different percentage of imputed data.

Imputation Method	Rank of Imputation Method for Different Percentage of Imputed Data					Rank by Mean	Rank by Mode
	10%	20%	30%	40%	50%		
Dataset Name: Glass							
Mean Imputation	5	5	5	5	4	4.8	5
Median Imputation	6	6	6	6	6	6	6
KNN Imputation	2	1	1	1	1	1.2	1
Predictive Mean Matching (pmm)	1	2	2	2	2	1.8	2
Bayesian Linear Regression (norm)	4	3	4	4	3	3.6	4
Linear Regression, non-Bayesian (norm.nob)	3	4	3	3	5	3.6	3
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.951, Chi-sq = 28.5, p value = 7.42E-05						

Table 14. Rank of imputation method for concrete dataset for different percentage of imputed data.

	Rank of Imputation Method for Different Percentage of Imputed Data						
Imputation Method	10%	20%	30%	40%	50%	Rank by Mean	Rank by Mode
Dataset Name: Concrete							
Mean Imputation	5	5	5	5	5	5	5
Median Imputation	6	6	6	6	6	6	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	2	2	2	3	2.2	2
Bayesian Linear Regression (norm)	3	4	4	3	4	3.6	4
Linear Regression, non-Bayesian (norm.nob)	4	3	3	4	2	3.2	4
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.966, Chi-sq = 29, p value = 6.16E-05						

Table 15. Rank of imputation method for liver dataset for different percentage of imputed data.

	Rank of Imputation Method for Different Percentage of Imputed Data						
Imputation Method	10%	20%	30%	40%	50%	Rank by Mean	Rank by Mode
Dataset Name: Liver							
Mean Imputation	5	5	3	2	5	4	5
Median Imputation	6	6	5	4	6	5.4	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	2	2	3	3	2.4	2
Bayesian Linear Regression (norm)	3	3	4	6	4	4	3
Linear Regression, non-Bayesian (norm.nob)	4	4	6	5	2	4.2	4
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.806, Chi-sq = 24.2, p value = .000486						

Table 16. Rank of imputation method for seed dataset for different percentage of imputed data.

	Rank of Imputation Method for Different Percentage of Imputed Data						
Imputation Method	10%	20%	30%	40%	50%	Rank by Mean	Rank by Mode
Dataset Name: Seed							
Mean Imputation	5	5	5	6	5	5.2	5
Median Imputation	6	6	6	5	6	5.8	6
KNN Imputation	1	1	1	1	1	1	1
Predictive Mean Matching (pmm)	2	4	2	3	4	3	2
Bayesian Linear Regression (norm)	3	2	4	4	3	3.2	3
Linear Regression, non-Bayesian (norm.nob)	4	3	3	2	2	2.8	3
Random sample from observed values (sample)	7	7	7	7	7	7	7
Kendall's Statistics	W = 0.92, Chi-sq = 27.6, p value = .000112						

a given dataset. The last two columns in each table indicate the average rank and rank obtained using mode method. The last row in each table provides Kendall's test statistics which is used for testing agreement among the ranks of imputation methods for a given dataset for different percentage of missing data.

In order to assess consistency in performance of each imputation method we have formulated null and alternative hypothesis as follows:

Null hypothesis (H0): Statistically there is no agreement among rankings of different imputation methods (w = 0).

Alternative Hypothesis: Statistically there is an agreement among rankings of different imputation methods (w = 1).

To test this hypothesis we have used Kendall's W test statistics. The Kendall's coefficient of concordance W ranges from 0 to 1. The value Zero (0) means no agreement on ranking and one (1) means complete agreement on ranking. The statistical significance of Kendall's W can be evaluated using chi-square test with n-1 degrees of freedom.

From the test statistics (the last row in each table from Tables 7–16), it can be observed that W statistics is close to 1 and *p* value is also significant for 5% level of significance, so null hypothesis be rejected in all the cases.

Therefore, we can conclude that there is an agreement among rankings of different imputation methods and the rank of imputation method is independent of dataset and percentage of missing data in the dataset. In other words, we can conclude that the ranking or performance of the imputation method is consistent across five different numeric datasets used in the study and with different percentages of missing data. It means ranking or performance of the imputation method neither changes with percentage of missing data nor with the different datasets. We also found that Mean NRMSE is lowest for kNN imputation method and hence we can conclude that kNN imputation method outperforms the other methods. But these results are applicable only to numeric datasets and one must always consider that there is no universal method always performing best in every situation.

Conclusion

Quality of the data is main concern of the data scientists. Although quality of data depends on several factors, one of the main factors is data incompleteness. Therefore, issues concerning missing data must be dealt with rigor by data scientists before analyzing data and viable decisions are taken by end users of the data mining projects. Data imputation is one of the techniques of handling missing values to make data complete and ready for analysis by replacing missing values with most plausible values. In this paper, we have discussed the concept of data imputation, data missingness mechanisms, handling missing values, Single and Multiple Imputation Methods, and then analysis of performance of different imputation methods namely

mean imputation, median imputation, kNN imputation, predictive mean matching (PMM), Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob), and Random Sample methods.

For the purpose of analyzing performance of different imputation methods we have used only numeric datasets obtained from UCI machine learning repository. The Normalized RMSE (NRMSE) method is used to compare performance of different imputation methods. The mean NRMSE for each dataset for different percentage of missing data is calculated using different imputation methods. Less is value of Mean NRMSE, better is performance of the imputation method. The result of analysis shows that kNN imputation method outperforms the other methods. We have also analyzed whether performance of each imputation method is consistent over five different numeric datasets for different percentage of missing values using Kendall's W test statistics. The results of analysis shows that value of Kendall's coefficient of concordance (W) is close to 1 (one). It means that there is complete agreement on ranking of imputation method over five different datasets for different percentages of missing values. Therefore, performance of the imputation method is independent of dataset and percentage of missing values. Limitation of this study is that it employs imputation method only on numeric datasets therefore outcome of the study is applicable only to numeric dataset.

References

Brown, M. L., and John. F. K. 2003. Data mining and the impact of missing data. *Industrial Management & Data Systems* 103 (8):611–21. doi:10.1108/02635570310497657.

Geert, J. M. G., A. van der Heijdena, R. T. Donders, T. Stijnen, and K. G. M. Moons. 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology* 59:1102–09. doi:10.1016/j.jclinepi.2006.01.015.

Ghorbani, S., and M. C. Desmarais. 2017. Performance comparison of recent imputation methods for classification tasks over binary data. *Applied Artificial Intelligence* 31 (1):1–22.

Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60:549–76. doi:10.1146/annurev.psych.58.110405.085530.

Graham, J. W., S. M. Hofer, and A. M. Piccinin. 1994. Analysis with missing data in drug prevention research. *NIDA Research Monograph* 142:13–13.

Graham, J. W., S. M. Hofer, and D. P. MacKinnon. 1996. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research* 31 (2):197–218. doi:10.1207/s15327906mbr3102_3.

Graham, J. W., S. M. Hofer, S. I. Donaldson, D. P. MacKinnon, and J. L. Schafer. 1997. Analysis with missing data in prevention research. *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research* 1, 325–66.

Gustavo, E. A. P. A. B., and M. C. Monard. 2010. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence* 17:5–6,519–533.

Han, J., and Kamber. 2012. *Data mining: Concepts and techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers.

Herzog, T. N., and D. B. Rubin. 1983. Using multiple imputations to handle nonresponse in sample surveys. *Incomplete Data in Sample Surveys* 2:209-245.

Horton, N. J., and K. P. Kleinman. 2007. Much Ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 61 (1):79–90. doi:10.1198/000313007X172556.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95 (1):49–69.

Kowarik, A., and M. Templ. 2016. Imputation with the R Package VIM. *Journal of Statistical Software* 74 (7):1–16. doi:10.18637/jss.v074.i07.

Kyureghian, G., O. Capps, and R. M. Nayga. 2011. A missing variable imputation methodology with an empirical application, in David M. Drukker (ed.) Missing data methods: Cross-sectional methods and applications. *Advances in Econometrics* 27 (Part 1):313–37.

Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Little, R. J., and D. B. Rubin. 1989. The analysis of social science data with missing values. *Sociological Methods & Research* 18 (2–3):292–326. doi:10.1177/0049124189018002004.

Little, R. J., and D. B. Rubin. 1987. *Statistical analysis with missing data*. Second ed. Hoboken, NJ: John Wiley & Sons.

Malarvizhi, M., and A. Thanamani. 2012. K-NN classifier performs better than K-means clustering in missing value imputation. *Iosr Journal Of Computer Engineering (IOSRJCE)* 6:12-15. doi:10.9790/0661.

Mishra, S., and D. Khare. 2014. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: A simulation study. *Journal of Medical Statistics and Informatics* 2. Article:9. doi:10.7243/2053-7662-2-9.

Nguyen, D. V., N. Wang, and R. J. Carroll. 2004. Evaluation of missing value estimation for microarray data. *Journal of Data Science* 2 (4):347–70.

Patric, R., and I. R. White. 2011. Multiple Imputation by Chained Equations (MICE), Implementation in Stata. *Journal of Statistical Software* 45 (4):1-20.

Penone, C., A. D. Davidson, K. T. Shoemaker, M. DiMarco, C. Rondinini, T. M. Brooks, B. E. Young, C. H. Graham, and G. C. Costa. 2014. Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution* 5:961–70. doi:10.1111/2041-210X.12232.

Poulos, J., and R. Valle. 2016. Missing data imputation for supervised learning. *Applied Artificial Intelligence* 32 (2):186–96. doi:10.1080/08839514.2018.1448143.

Raymond, M. R. 1986. Missing data in evaluation research. *Evaluation & the Health Professions* 9 (4):395–420. doi:10.1177/016327878600900401.

Rubin, D., and N. Schenker. 1986. Multiple Imputation for interval estimation from simple random sample with ignorable nonresponse. *Journal of the American Statistcal Association* 91:366–74. doi:10.1080/01621459.1986.10478280.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63:581–92. doi:10.1093/biomet/63.3.581.

Rubin, D. B. 1987. *Multiple imputation for nonresponse in survey*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. 1999. Multiple Imputation: A Primer. *Statistical Methods in Medical Research* 8:3–15. doi:10.1177/096228029900800102.

Schafer, J. L., and J. W. Graham. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7:147–77. doi:10.1037/1082-989X.7.2.147.

Schmitt, P., J. Mandel, and M. Guedj. 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics* 6 (1):1-6.

Somasundaram, R. S., and R. Nedunchezhian. 2011. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. *International Journal of Computer Applications* 21 (10). doi:10.5120/2619-3544.

Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal* 338: doi: 10.1136/bmj.b902.

Troyanskaya, O., D. Botstein, and R. Altman. 2003. Missing value estimation. In *A practical approach to microarray data analysis*, 1-46, ed. D. Berrar, W. Dubitzky, and M. Granzow. US: Springer.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. 2001. Missing value estimation methods for DNA micro arrays. *Bioinformatics* 17 (6):520–25. doi:10.1093/bioinformatics/17.6.520.

Tsikriktsis, N. 2005. A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* 24 (1):53–62. doi:10.1016/j.jom.2005.03.001.

Tutz, G., and S. Ramzan. 2015. Improved methods for the imputation of missing data by Nearest neighbor methods. *Computational Statistics and Data Analysis* 90:84–99. doi:10.1016/j.csda.2015.04.009.

van Buuren, S., and K. Groothuis-Oudshoorn. 2011. Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45 (3):1-67.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple Imputation Using Chained Equations: Issues and guidance for practice. *Statistics in Medicine* 30:377–99. doi:10.1002/sim.4067.

Zhang, J., and H. Aytug. 2016. Comparison of imputation methods for discriminant analysis with strategically hidden data. *European Journal of Operational Research* 255:522–30. doi:10.1016/j.ejor.2016.05.052.