Knowledge-Based Systems 160 (2018) 104–118

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Handling missing values: A study of popular imputation packages in R

Madan Lal Yadav*, Basav Roychoudhury

Indian Institute of Management Shillong, Shillong, Meghalaya, India

ARTICLE INFO

Keywords:
Missing value handling
VIM
MICE
MissForest
HMISC
Imputation Time
Imputation Accuracy

ABSTRACT

In real world data are often plagued by missing values which adversely affects the final outcome of the analysis based on such data. The missing values can be handled using various techniques like deletion or imputation. Of late, R has become one of the most preferred platform for carrying out data analysis, and its popularity is growing further. R provides various packages for handling missing values through imputation. The presence of multiple packages however, calls for an analysis of their comparative performance and examine their suitability for handling a given set of data. The performance of different R packages may differ for different datasets and may depend on the size of the dataset and richness of the missing values in the datasets. In this paper, the authors perform comparative study of the performance of the common R packages, namely VIM, MICE, MissForest, and HMISC, used for missing value imputation. The authors measured the performances of the said packages in terms of their imputation time, imputation efficiency and the effect on the variance. The imputation efficiency was measured in terms of the difference in predictive performance of a model built using original dataset vis-à-vis a dataset with imputed values. Similarly, the variance of the variables in the original dataset was compared that of corresponding variables in the imputed dataset. A missing value imputation package can be considered to be better if it consumes less imputation time and provides high imputation accuracy. Also in terms of variance, one would like to have the imputation package maintain the original variance of the variables. On analysing the four imputation packages on two datasets over three predictive algorithms–Logistic Regression, Support Vector Machines, and Artificial Neural Networks–it was observed that the performances varies depending on the size of the dataset, and the missing values present in them. The study highlights that certain missing value package used in conjunction with a given predictive algorithm provides better performance, which is again a function of the dataset characteristics.

1. Introduction

In most scientific investigations involving data analysis, presence of missing data is a common issue, and determining the right approach to mitigate this often becomes a major challenge [20]. With growing interest in data dependent tools like machine learning and the like, the quality of data often becomes a major driver for the final outcome–better the data quality, better the final outcome. Missing values in datasets adversely affect the data quality and thereby the final knowledge discovered therefrom [4,47].

One of the preliminary tasks in such data intensive research is the collection of relevant data. In practice, collected datasets are incomplete and are riddled with missing values. The reasons for this may vary: manual error at the time of data entry, values not being available, equipment errors, incorrect measurement, behavioural issues, and so on. The datasets wherein values of one or more attributes are missing across certain number of records are called datasets with missing values [18].

Missing values are categorized into three groups, each requiring different handling of the problem. They could be Missing Completely at Random (MCAR), in which the missed value of an attribute does not depend on the value of the attribute itself nor on the values of other attributes; Missing at Random (MAR), in which missed attribute value does not depend on the value of the concerned attribute itself, but on that of some other attribute; or Missing Not at Random (MNAR), in which missed attribute value depends on value of the concerned attribute itself [1,4,29,39].

To handle missing values, one can use deletion techniques wherein certain records containing variables with missing values are not considered for analysis. In imputation techniques for missing value handling, an appropriate value is used to replace the corresponding missing values [4,39]. The imputed value could be mean, mode, median, or any predefined value of the variable having the missing value, or could be the one obtained through some predictive models [32]. Imputation

* Corresponding author.
E-mail addresses: madan.fpm14@iimshillong.ac.in (M.L. Yadav), brc@iimshillong.ac.in (B. Roychoudhury).

https://doi.org/10.1016/j.knosys.2018.06.012
Received 21 December 2017; Received in revised form 23 April 2018; Accepted 28 June 2018
Available online 06 July 2018

techniques are used to handle missing values if they are of MAR or MCAR type, and when each record or variable in the dataset is very crucial and a single record does not have missing values across many variables [36]. For MCAR type, missing values in a dataset can be handled also by using deletion, preferably listwise deletion or maximum likelihood methods. There are no such general methods to handle MNAR type of missing values. [2,5,41].

Along with the increasing push towards data based decision making, the popularity of R, the open source statistical and data analytics application, is also increasing day by day. Growing number of researchers and practitioners are using this application for their data analysis tasks. To handle the aforesaid issue of missing values in datasets through the techniques of imputation, R provides multiple packages; VIM, MICE, missForest and HMISC being the commonly used ones. In this study, the authors study the imputation time and the imputation accuracies of these four imputation packages, together with the impact of dataset size and the percentage of missing values thereon. The motivation for this study is the immense popularity of the R software among the researchers across various disciplines, and the fact that the issue of missing value is discipline agnostic. The contribution of this study, through comparison of efficiency and accuracy of the considered R packages used with dataset of different sizes and with different percentage of missing values, is to provide guidelines to researchers for picking the right R package for missing value imputation depending on the characteristics of their dataset.

The rest of the paper is organised as follows: Section II presents the review of existing literature dealing with missing value, Section III presents the four R imputation packages used in this study, Section IV details out the methodology followed and discusses the experiments carried out, Section V presents the results and findings, and Section VI concludes the paper.

2. Literature review

With the researchers requiring some means to take care of missing values in their datasets, multiple techniques ranging from deletion to imputation has been suggested in the literature. There are also maximum likelihood techniques that can be used to handle missing values.

Deletion could be 'complete deletion', 'list-wise deletion', or 'Complete Case Analysis', in which all rows having one or more of their attribute values missing are deleted or 'specific deletion' in which only those rows are deleted which have more than a predefined percentage of their attribute values missing ([22,39,25]). There can also be 'variable deletion' or 'pair-wise deletion', where the rows having missing values in the variables involved in the current analysis are deleted, these rows being however used for other analyses which does not involve the concerned variables having the missing values; in the extreme case of each variable having missing values across many records may result in the deletion of the complete dataset [36].

Imputation methods use the information available in the selected dataset to estimate the missing value, wherein an appropriate value is imputed in place of the missed value [4,39]. This value could be mean, mode, median or any predefined value of the variable having missing value [32], or the value can be obtained from case substitution. Imputed value can also be computed using regression models [1], KNN, EM (expectation maximisation) imputation [23], hot deck imputation [39], cold deck imputation [16], etc. In techniques involving prediction models, a model is developed based on existing information, which is then used to predict appropriate values for the missing data. [4].

In mean and mode imputation methods, missing value of an attribute is replaced by mean or mode of the all known values of that attribute [42]. Case substitution is mainly used to handle missing values in sample surveys, where one instance of missed data is substituted by another non sampled instance. In hot deck method, missed values are estimated from current data, while in cold deck, they are estimated using a data source other than current dataset [16]. KNN uses the k nearest cases to impute missing values [22,25,32]. Among KNN and EM imputations, KNN imputation performs better (García-laencina, Abreu, & Abreu, 2015).

Further, imputation could either be single imputation or multiple imputation. In single imputation, a single appropriate value is imputed in place of missed value [16]. Multiple imputation was originally proposed by Rubin [20]. In multiple imputation, 'm' complete datasets are obtained by imputing the missing values 'm' times [36], the final imputed dataset being the weighted average of these 'm' datasets. Multiple imputation has various advantages over other alternative methods, namely single imputation, maximum likelihood techniques [19], and deletion; it however, requires more resources [17] . Single imputation techniques treat all values, including imputed values as true values, and does not account for missing value uncertainty which lead to inflated type I error rates. Maximum likelihood techniques are difficult to implement for complex or nonstandard models [45]. There are also Inverse Probability Weighing (IPW) methods to handle missing data, which use inverse of the observed probability to weight observed records thereby representing the whole data including the missing values. However, imputation method has better performance [11].

Local least squares (LLS), least squares adaptive (LSA) and Bayesian Principal Component Analysis (BPCA) [2] are some good imputation methods with none of these being consistently superior to the other two for every dataset [6]. Some other methods used for multiple imputation are singular value decomposition (SVD), partial least squares (PLS), and ordinary least squares (OLS) [6]. Global based imputation methods like SVD, PLS, and BPCA provide better results on datasets having lower complexity, while neighbour based methods like OLS, KNN, LSA and LLS do better on those with higher complexity. Existing literature includes other models and their evaluations as well [6]. Local Least square_gene (LSI_gene), LSI_adaptive, LSI_array, LSI_combined, Local Least Square imputation (LLSI) and EM_gene methods outperform BPCA [10]. Fully Conditional Specification (FCS) and Multi Variate Normal Imputation (MVNI) approaches are generally less biased and produce similar results despite datasets having binary and ordinal variables. MVNI provide ease of model specification, but some people may have problem with its unrealistic nature of the multivariate normal assumption. FCS needs a separate regression model for each variable whose value is going to be imputed and hence have complex model specification [27]. FINNIM is an efficient nonparametric iterative multiple imputation method which uses KNN to estimate missing value [38]. Kernel extension methods show competitive performance (predictive accuracy) in handling missing values in datasets having binary values, as compared to multiple imputation using Support Vector Machines [5,34].

Multiple imputation using sequential regression trees as conditional model has the ability to capture complex relation and require minimal tuning by the user [7]. In normal imputation (NM) method, posterior normal distribution is used to impute missing values and is based on the regression coefficients [37]. Mean and variance (MV) method uses empirical distribution of standardized residual wherein each missing value is calculated by adding its predictive means to the residuals. In the predictive mean matching (PMM) method, missing value is imputed using an observation randomly drawn from a set of observed cases whose predictive mean is close to predictive mean of missing values. Local residual draw (LRD) method imputes the missing value using the predictive mean as in case of PMM, plus a residual randomly drawn from the residuals of a set of observed cases with predictive means close to that of the missing value [40]. In case of bounded missing value imputation, Proportioned Residual Draw (PRD), Predictive Mean Matching-Proportioned Residual Draw (PMM-PRD) methods perform better than Normal imputation Method (NM), Mean and Variance (MV), PMM and Local Residual Draw (LRD) methods, when the number of boundaries increase [26].

Decision tree and Sampling based missing value Imputation (DSMI) method uses correlation concept to handle missing values through

imputation. It divides dataset in different segments horizontally based on non-missing attributes of the missing records and missing value imputation is done using samples from the distributions induced by the correlation. DSMI has better performance than k Nearest Neighbour Imputation (KNNI) and other popular imputation algorithm [13]. Multiple imputation method obtained from combining Gaussian Mixture Model (GMM) and Extreme Learning Machine (ELM) gives better accuracy than methods based on conditional mean imputation but takes more computational time [43].

Reinforcement programming (RP) method, based on Reinforcement Learning, has better performance as compared to mean per category imputation, zero imputation, genetic algorithm (GA) in terms of sum of square error (SSE) and computational time consumption [35]. MOGAImp (multi-objective genetic algorithm), based on Non-dominated Sorting Genetic algorithm II (NSGA-II), is another multiple imputation algorithm suitable for handling missing values in datasets having mixed attributes (continuous and categorical). MOGAImp performs better than Concept Most Common (CMC) Attribute Value for Symbolic Attributes [31], the Weighted Imputation with KNN (WKNNI), and Global Most Common (MC) Attribute Value for Symbolic Attributes and Global Average Value for Numerical Attributes [30]. Entropy based selection (EBS) and simulation-based self-training selection (STS) schemes help to select optimal imputation algorithm [6].

Fuzzy-Rough Nearest Neighbour Imputation (FRNNI) method is based on implicator/ t-norm based fuzzy-rough sets. Ordered Weighted Average Nearest Neighbour Imputation (OWANNI) method is based on OWA-based fuzzy-rough sets and Vaguely Quantified Nearest Neighbour Imputation (VQNNI) method is based on vaguely quantified rough sets. Among FRNNI, OWANNI and VQNNI fuzzy based imputation methods, FRNNI has best performance and VQNNI has the worst. Performance of OWANNI and FRNNI are similar but FRNNI is slightly better [2]. Nuovo [15] used Fuzzy C Means (FCM) algorithm to obtain effective data imputation. Aydilek and Arslan [3] used a hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm to obtain sensible performance. Tang, Zhang, Wang, Wang, & Liu, [46] proposed a hybrid method which also integrates the FCM imputation method and GA optimization technique and shows better performance than 'Double Exponential Smoothing', 'Historical method' and 'Autoregressive Integrated Moving Average model' .

Problem of missing values in a dataset is worst when dataset size is huge and data is taken from heterogeneous sources. To handle missing values in such type of datasets, Dual rePopulated Bayesian Ant Colony Optimization (DPBACO) techniques can be used as it perform better than other missing value handling algorithms like missForest, and wNN [14].

Another approach is to involve experts in dealing with missing values. Wu & Chiclana, [50] presented a method to handle missing values, where multiplicative consistency of Intuitionistic Reciprocal Preference Relation (IRPRs) is used to find out the missing values in a dataset. Here experts' IRPRs are aggregated to find out final missing values based on the experts' consistency level. This however does not use other factors like knowledge, experience, and reputation of the experts in decision making process. Another problem with this approach is the availability of experts and the time taken in the process as compared to the machine based systems [50]. Capuano, Chiclana, Fujita, Herrera-Viedma, & Loia, [9] proposed a similar way of dealing with missing values as Wu & Chiclana, [50] which involves experts in deciding the incomplete information, but here experts' opinions are influenced by other experts and the degree of influence is based on the level of trust one has on the expert . This approach has similar drawback as that of Wu & Chiclana, [50]. As incomplete information is not same as low quality information, use of experts' opinion can lead to biased results. Urena, Chiclana, Morente-Molinera, & Herrera-Viedma, [48] reviewed various alternative approaches to deal with missing values such as deletion, penalising the experts, and estimating the missing values. GDM-R is an open source framework developed in R to deal with group decision making process. If there is any incomplete information during the process, this framework substitutes an appropriate value. This framework uses same method to handle all type of missing values, while it may be possible that different datasets, based on their size and percentage of missing values, may require different packages to handle them [49].

3. R packages for imputation

In this paper, four popular imputation packages available in R has been analysed in terms of their performance.

VIM[1] package of R has hot-deck imputation, regression imputation, robust model-based imputation and KNN imputation methods to handle missing value imputation [24]. This package can impute missing values of numerical, semi-continuous, categorical or ordered variables.

MICE[2] stands for Multivariate Imputation via Chained Equations [36] and handles missing value of MAR and MNAR type [8]. MICE creates different imputation models for different type of variables: Predictive Mean Matching method for numeric variables, Logistic Regression method for binary variables, Bayesian polytomous regression for Factor variables, and Proportional odds model for ordered variable [8,20].

MissForest[3] package uses non-parametric imputation method [44], wherein a random forest model created for each variable to predict missing values of that variable [21]. The variable having missing value is the response variable and a random forest is developed from the information of other variables using regression tree of resampling based classification for the prediction of missing value [28]. MissForest shows attractive computational efficiency and can work effectively with high-dimensional data [44].

HMISC[4] package performs imputation using additive regression, bootstrapping, and predictive mean matching [12,33]. It has two methods, one for single and other for multiple imputation.

VIM and MissForest deals with missing values through single imputation while MICE and HMISC deal missing values with multiple imputation. But one can performed single imputation by specifying value of 'm' to one in case of MICE or HMISC.

While surveying the literature, the authors did not come across any study comparing the capabilities and performance of the aforesaid popular libraries of R for handling missing values, which thus became the motivation for this study.

4. Dataset and methodology

The study looked into the impact of dataset size and percentage of missing values on imputation time in addition to the accuracy of such imputation. The authors used the two different datasets for this study - 'poker hand' [5] and 'BNG_heart_statlog'.[6]

- 'poker hand' dataset has 11 attributes, namely *s1, c1, s2, c2, s3, c3, s4, c4, s5, c5* and *class*, with *si* representing "Suit of card *#i*", *ci* representing "Rank of card *i*", and *class* representing the 'poker hand'. This dataset comprises of 1,000,000 records with each record representing a hand consisting of five playing cards drawn from a standard deck of 52, and each card described in terms of two attributes – suit and rank.
- 'BNG_heart_statlog' dataset has 14 attributes, namely age, sex, chest, resting_blood_pressure, serum_cholestoral, fasting_blood_sugar,

[1] https://cran.r-project.org/web/packages/VIM/VIM.pdf
[2] https://cran.r-project.org/web/packages/mice/mice.pdf
[3] https://cran.r-project.org/web/packages/missForest/missForest.pdf
[4] https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf
[5] https://archive.ics.uci.edu/ml/datasets/Poker+Hand
[6] https://www.openml.org/d/267

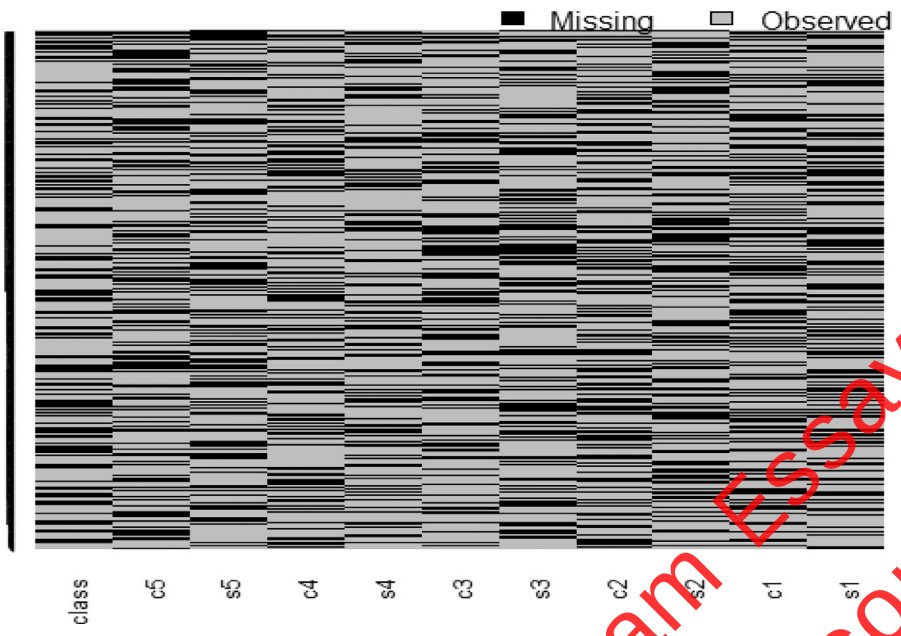

Fig. 1. Distribution of 40% missing values in sub dataset having 20,000 records of the 'poker hand' dataset.

resting_electrocardiographic_results, maximum_heart_rate_achieved, exercise_induced_angina, oldpeak, slope, number_of_major_vessels, thal, and class. This dataset also has 1,000,000 records.

These two datasets were considered as both these were large openly available datasets with nominal as well as numeric attributes and without any missing value across their attributes. The large size allowed for analysis of the effect of dataset size on imputation performance, while the presence of no missing value allowed for random introduction of missing values and to compare the imputation efficiency of considered packages in the presence of both nominal and numeric variables.

To study the impact of dataset size, analysis was performed on randomly sampled sub datasets of 10,000, 15,000, 20,000, 50,000 and 100,000 rows of the each datasets. The authors introduced 10%, 20%, 30% and 40% missing values in each of the aforesaid sub datasets using the R function *produceNA()*. The missing values were introduced across the attributes and the records and were spread out all across; as an illustration, Fig. 1 exemplifies the distribution of missing values in the sub dataset of 20,000 records with 40% missing values of 'poker hand' dataset. The missing values were then imputed using all the four R packages one by one, and their performance recorded.

There is an attribute 'class' in both the considered datasets which contains a classification for the concerned record. In case of 'poker hand' dataset, this attribute contains 10 classifications ranging from 0 to 9. However, the number of records with classifications 2 to 9 are few, and these records were not considered in the sampled datasets, thereby restricting the classification of records to 0 or 1. The 'class' attribute in 'BNG_heart_statlog' contains classification of records as heart disease being 'present' or 'absent'.

To ascertain the accuracy of the imputations performed, three predictive models were developed for each sampled sub-datasets to predict the 'class' attribute. These models were build using Support Vector Machine (SVM), Logistic regression (LR) and Artificial Neural Network (ANN) classifiers. While developing the classifiers using supervised learning methods, two sets of classifiers were developed in each case using two different training sets. The first of training sets were the sampled sub-datasets of varying sizes as mentioned earlier. The second sets of training sets were those derived after introduction of different percentages of missing values in the aforementioned sampled sub-datasets and subsequent imputation using the considered R packages. Two validation sets of 15,000 records for each – one from 'poker hand' and the other from 'BNG_heart_statlog' datasets – were randomly sampled such that they did not contain any record considered in the training sets. These two validation sets were used to evaluate the predictive accuracy of all the corresponding models, i.e., the same validation set derived from 'poker hand' dataset was used for all the classifiers built using data sampled from this dataset, and the similar being true for the other dataset. The difference in the classification accuracies of the corresponding classifier – one developed with data with no missing value and the other developed using the imputed data–as computed based on the classification performance on the validation set was then used as a metric for the imputation accuracy of the concerned imputation methods. It may be worth noting that the variables in the sampled sub-datasets comprised of both nominal and numeric attributes, and the missing values were introduced randomly across all the variables to analyse the imputation performance on both type of variables.

The above methodology was decided upon to mimic the real life situations where it might sometime be impossible to get a dataset devoid of missing values. In such cases, the training and validation sets will comprise of imputed values, and the 'correctness' of such imputation will decide on the correctness of the model so developed. The importance of the imputation package for handling the missing values in the given type of dataset thus becomes paramount.

As part of the experiment on the sub datasets, the following parameters were recorded and compared:

- Time taken for imputation
- Accuracy of imputation based on predictive performance of models developed on imputed sub datasets as compared to those on original sub datasets.
- Variance of imputed sub datasets as compared to that of the original ones–only numeric variables are considered for this purpose.

The Accuracy Variance Percentage (AVP), used as a measure for imputation performance, is the ratio of difference in prediction accuracy of the model built using training data extracted from original

dataset from that of the model built from the training set after inserting missing values and subsequent imputation to that of the model built using the training data extracted from the original dataset, expressed as percentage. The prediction accuracy is computed in either case is based on the classification performance on the same validation set.

$$Accuracy\ Variance\ Percentage(k, x, y, z) = \left| \frac{Accuracy_{Original} - Accuracy_{Imputed}}{Accuracy_{Original}} \right| \times 100\% \quad (1)$$

where

$Accuracy_{Original}$ = predictive classification accuracy of model built using training set of x records extracted from original dataset
$Accuracy_{Imputed}$ = predictive classification accuracy of model built using training set of same x records after introducing y% missing values and subsequent imputation
k = the R package used for imputation and could be VIM, MICE, missForest, or HMISC package
x = the number of records in the training dataset, and could be 10,000, 15,000, 20,000, 50,000, or 100,000
z = the algorithm used in developing the model and could be SVM, LR or ANN

Another measure of a good imputation is that the variance in any attribute of the imputed dataset should be as close to the variance of that attribute in the original dataset. Simple imputations like mean imputation reduces this variance considerably. Thus, the performance of imputation was also examined by Variance Decrease Percentage (VDP), which is defined as

$$Variance\ Decrease\ Percentage(k, x, y, c) = \left| \frac{Variance_{Original} - Variance_{Imputed}}{Variance_{Original}} \right| \times 100\% \quad (2)$$

$Variance_{Original}$ = Variance in attribute c in the training set of x records extracted from original dataset
$Variance_{Imputed}$ = Variance in attribute c in the training set of same x records after introducing y% missing values and subsequent imputation

The above experiments were carried out using a computer with the following configuration:

- Memory: 3 GB
- Processor: Intel core 2 Duo Processor (2 GHs)
- Hard Disk Drive: 250 GB

5. Results and discussion

This section is divided into three subsections. The first subsection

Fig. 2. Time Consumed for imputation for datasets with different percentages of missing values.

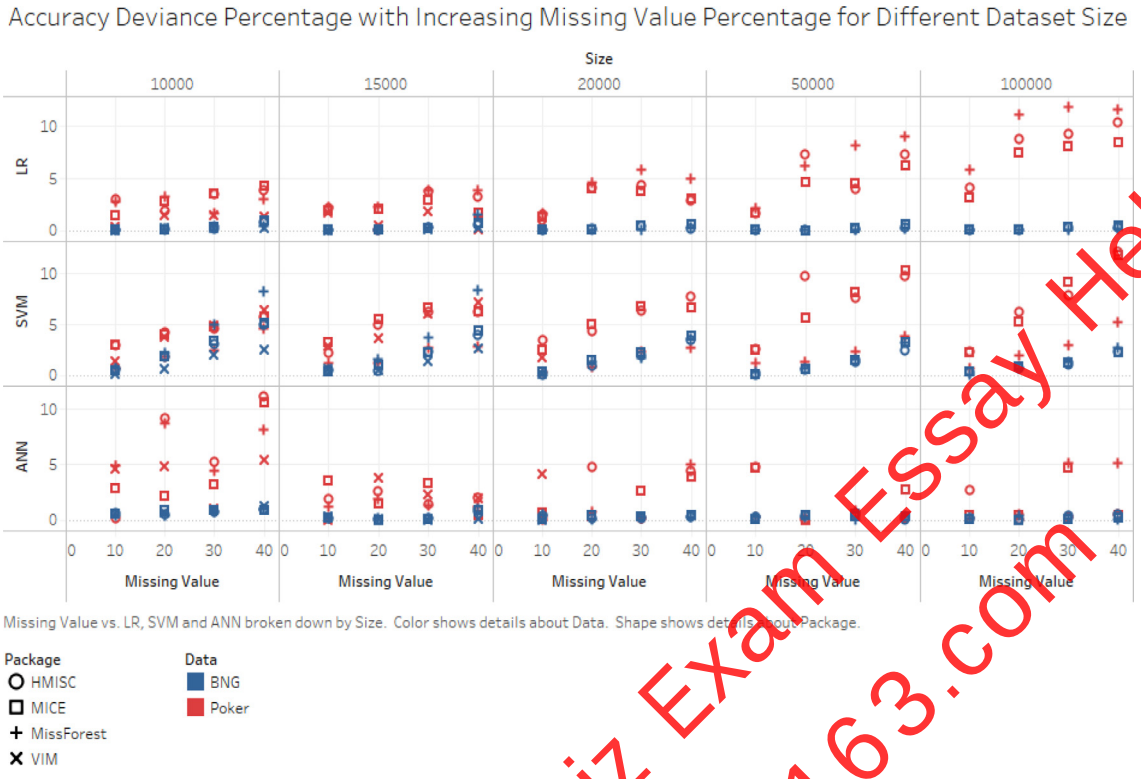

Fig. 3. Accuracy Variance Percentage across dataset sizes and missing value percentage.

discusses the imputation time, the second presents an analysis on imputation accuracy, and the third deals with the comparison of variance in the original and imputed sub-datasets.

5.1. Imputation time analysis

The time taken for imputation using VIM, MICE, MissForest and HMISC packages on sub-datasets of 10,000, 15,000, 20,000, 50,000, and 100,000 randomly sampled rows from the original datasets of 'poker hand' and 'BNG_heart_statlog' with 10, 20, 30, and 40 percentages of missing values respectively are shown in Fig. 2. Appendix A contains the actual data.

The salient observations are listed below:

- As far as VIM package is concerned, imputation time increased with the percentage of missing values up to the size that this package could handle. This package failed to impute datasets of sizes of 20,000 records with 20% missing values and beyond.
- As for MICE, the imputation time increased with increase in missing value from 10% to 30% and then decreased marginally for missing value of 40% in datasets sized of up to 15,000 records. For datasets of size 20,000 and beyond, the imputation time increased with the missing value percentage.
- In case of MissForest, the imputation time decreased with the increase in missing value percentage in case of all datasets up to 50,000 records. Its behaviour was complex for dataset of size 100,000.
- HMISC behaved somewhat similar to MICE, with the imputation time increasing with increase in missing value from 10% to 20% and then decreased marginally for missing value of 30% in datasets sized of up to 15,000 records. For datasets of size 20,000 and beyond, the imputation time increased with the missing value percentage.
- HMISC was the fastest in its performance across all cases, except for the particular case of the sub dataset of size 10,000 with 10% missing values, where VIM was the fastest.
- The authors also performed missing value imputation on sub dataset having 500,000 rows but all four packages in that case failed to impute missing values.

5.2. Predictive accuracy analysis of models on imputed dataset

The predictive accuracy of models built from imputed dataset was compared with those from original dataset using Accuracy Variance Percentage (AVP) as in Eq. (1). The behaviour without considering the imputation package used is shown in Fig. 3, while that based on the imputation packages is shown in Fig. 4. The actual data is placed as Appendix B.

One can observe the following from Fig. 3:

- Across all packages, LR delivers a good performance in terms of low AVP, especially when used with smaller datasets (10,000 to 20,000 records), and deteriorates somewhat at higher missing values. It may thus not be suitable for use in dataset where both the size and the percentage of missing value are large.
- ANN does a better job at larger datasets as well as at higher missing value percentages, but performs poorly at smaller sized datasets.
- For SVM, the results are better for low missing value percentage (at 10%) across dataset sizes with its performance falling at higher missing value percentages.

It may be noted once again that AVP is not a measure of absolute performance of the model, but the percentage variation it shows in accuracy for models developed using imputed datasets as compared to a model developed from the original dataset, the one without any missing value.

Considering the imputations package used (Fig. 4), the following could be observed:

- In general, AVP for VIM package increases, with a few exceptions, with increasing missing value percentage. It is also observed that in

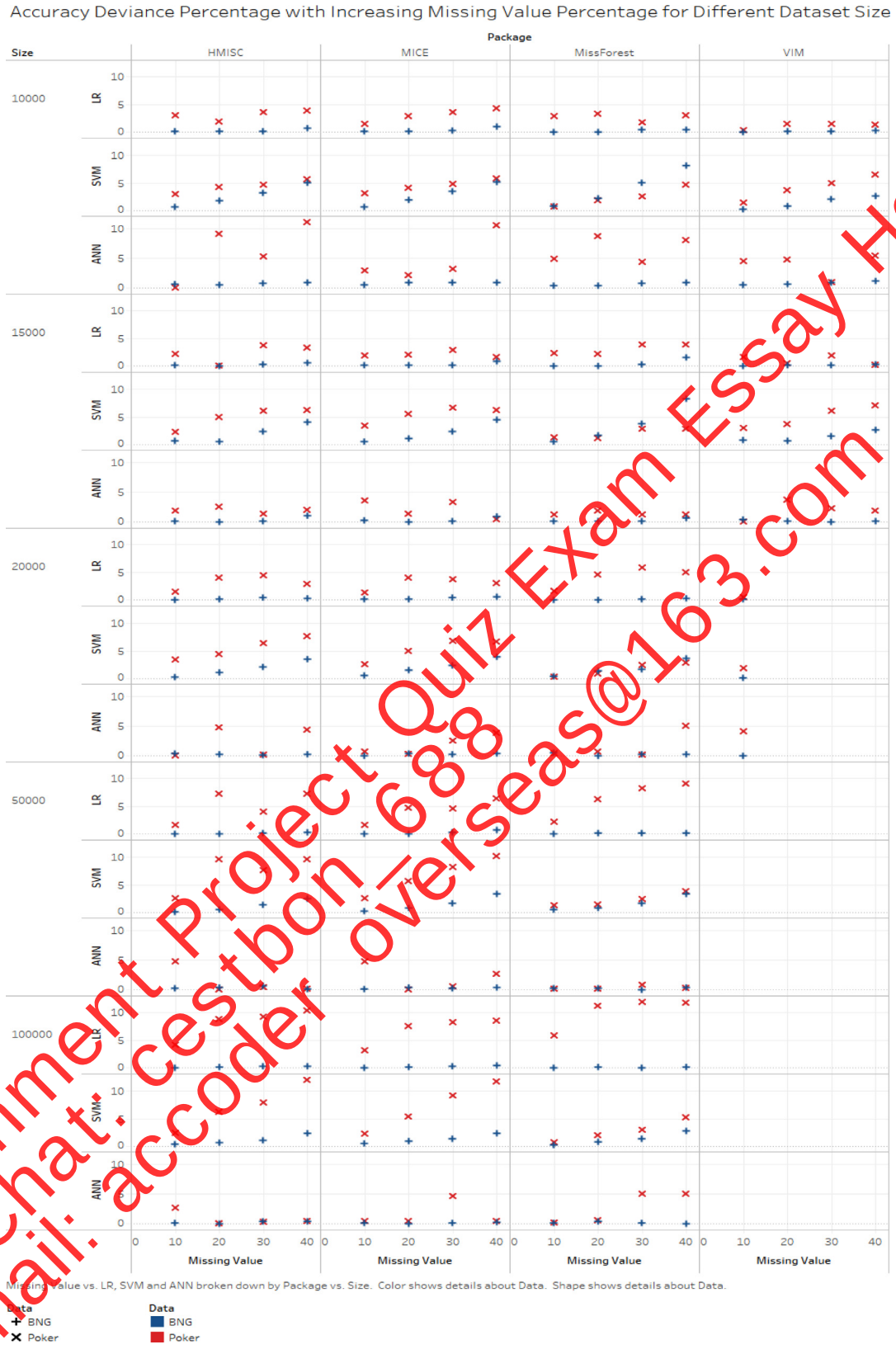

Fig. 4. Accuracy Variance across dataset sizes, missing value percentage, and imputation package.

most of the cases for VIM, LR performance in terms of accuracy is best.

- In general and with minor aberration, AVP for MICE package increases for LR and SVM classifier with the missing value percentage. In case of ANN classifiers, AVP did not show any fixed trend. It is also observed that in most of the cases, AVP is minimum with LR for less percentage of missing values and for high percentage of missing values it is minimum with ANN.
- In most cases, AVP for misForest package showed an increase in case of LR and SVM classifiers with increasing missing value percentage.

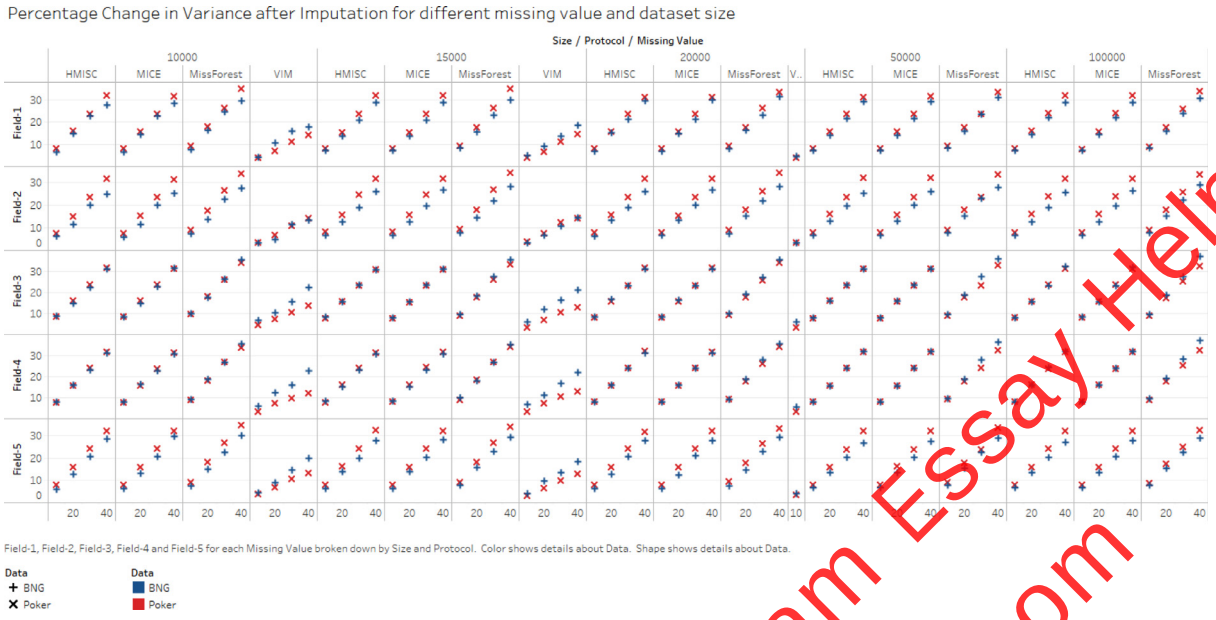

Fig. 5. Variance Decrease Percentage across imputation package, missing value percentage, and dataset sizes.

Fig. 6. Variance Decrease Percentage across missing value percentage, imputation package, and dataset sizes.

In case of ANN classifier, AVP first increases then decreases for each dataset with increasing missing value percentage.

- AVP for HMISC package when used with SVM increases with increasing missing value percentage. For LR and ANN, the AVP falls in most cases at moderate values of missing value percentage. Additionally, for large datasets, ANN used with HMISC provides better AVP.
- For VIM package LR performance in terms of accuracy is best for smaller missing values while ANN takes the place at the higher end.
- For all packages, LR perform better in most cases than other two classifiers when data size is small, and for large dataset ANN performance is better.
- For small size datasets, it is better to use VIM package for imputation.

5.3. Variance analysis

The variance of the imputed datasets vis-à-vis that of the original dataset is compared for both the datasets. As 'poker hand' dataset only has five numeric attributes, the authors considered first five numeric attributes from second dataset 'BNG_heart_statlog' for the ease of comparison. Field1 represents c1 attribute of dataset 'poker hand' and 'age' attribute of 'BNG_heart_statlog'. Field2 represents c2 attribute of dataset 'poker hand' and 'chest' attribute of 'BNG_heart_statlog'. Field3 represents c3 attribute of dataset 'poker hand' and 'resting_blood_pressure' attribute of 'BNG_heart_statlog'. Field4 represents c4 attribute of dataset 'poker hand' and 'serum_cholestoral' attribute of 'BNG_heart_statlog'. Field5 represents c5 attribute of dataset 'poker hand' and 'maximum_heart_rate_achieved' attribute of 'BNG_heart_statlog'. The variance, as expected, decreased in all cases of

imputed subdatasets of both the datasets 'pker hand' and 'BNG_heart_statlog', but by different percentage. The behaviour on the basis of imputation packages is seen in Fig. 5, while Fig. 6 shows the same based on missing value percentages. The actual data for the aforesaid five attributes is placed as Appendix C.

The variance decrease percentage, as computed using Eq. (2), shows an increase with increasing missing value percentage, exhibiting a linear trend. Moreover, there is not much difference among the imputation packages, except for VIM. While VIM did not work with larger datasets, it showed better performance till the point it worked. Also, other than VIM, MICE and HMISC showed comparable performance while missForest was the worst in most cases.

6. Conclusion and future work

This paper investigates the effect of missing value imputation, using four popular missing value imputation packages available in R, on the predictive performance of models developed from dataset with missing values post imputation. The decrease in variance in the cases of imputed datasets was also examined.

In terms of time consumed during the imputation process, HMISC exhibited the fastest performance in almost all the scenarios considered.

While it is observed that the variance decreased in the imputed datasets, as to be expected, the decrease was less in case of one imputation package as compared to the other three; the three also differing in their capacity to maintain the variance . Thus, while VIM exhibited the least decrease in variance for smaller datasets, one may prefer to use HMISC or MICE for larger datasets where VIM fails to execute.

An interesting observation is that the performance of predictive models depended on the imputation package and the modelling algorithm used for developing the predictive model. While certain algorithms did better for lower percentage of missing values, others performed better with higher percentage. The same also was the case for imputation package, some performing better at certain spectrum of dataset sizes and missing value percentages. For smaller datasets, it is found that VIM package is better for imputation used along with LR for predictive modelling. For lager datasets, it is the HMISC package along with ANN that performed the best.

While this paper conducted an experimental study on the performance of four missing value imputation packages of R, it did not include in its scope the analysis of the reasons behind the reported performances. An investigation to understand the reasons for the aforesaid behaviour in terms of imputation package and modelling algorithms can be a way forward for this work.

Appendix A. Imputation time

Size of Dataset	Data	Percentage of missing value (%)	VIM	MICE	MissForest	HMISC
10,000	Poker	10	00:01:44	00:16:06	00:10:55	00:01:45
10,000	Poker	20	00:03:47	00:17:12	00:10:17	00:02:11
10,000	Poker	30	00:04:10	00:21:15	00:07:53	00:02:01
10,000	Poker	40	00:05:03	00:16:45	00:03:55	00:01:56
15,000	Poker	10	00:06:47	00:23:51	00:41:30	00:03:01
15,000	Poker	20	00:10:57	00:25:38	00:19:14	00:03:04
15,000	Poker	30	00:22:09	00:27:35	00:15:18	00:03:02
15,000	Poker	40	00:22:36	00:27:07	00:10:30	00:02:58
20,000	Poker	10	00:07:21	00:34:34	02:05:51	00:04:03
20,000	Poker	20	Fail	00:38:07	00:53:47	00:04:32
20,000	Poker	30	Fail	00:41:18	00:31:19	00:04:43
20,000	Poker	40	Fail	00:45:28	00:19:37	00:05:07
50,000	Poker	10	Fail	00:43:37	16:20:54	00:14:00
50,000	Poker	20	Fail	02:42:17	12:36:03	00:18:42
50,000	Poker	30	Fail	03:41:30	10:59:52	00:22:52
50,000	Poker	40	Fail	03:47:50	06:51:46	00:24:32
100,000	Poker	10	Fail	06:37:40	64:58:52	00:51:51
100,000	Poker	20	Fail	09:10:57	50:35:29	01:18:03
100,000	Poker	30	Fail	13:00:00	45:49:29	01:32:06
100,000	Poker	40	Fail	14:42:57	32:04:51	01:34:50
10,000	BNG	10	00:02:10	00:11:10	00:09:41	00:02:16
10,000	BNG	20	00:04:07	00:11:17	00:07:50	00:02:46
10,000	BNG	30	00:05:49	00:11:55	00:06:10	00:02:39
10,000	BNG	40	00:06:29	00:11:41	00:03:48	00:02:27
15,000	BNG	10	00:05:38	00:18:35	00:45:11	00:04:46
15,000	BNG	20	00:10:26	00:21:11	00:31:05	00:04:59
15,000	BNG	30	00:13:37	00:22:23	00:18:03	00:04:51
15,000	BNG	40	00:14:38	00:22:21	00:10:01	00:04:24
20,000	BNG	10	00:09:18	00:26:11	01:40:55	00:06:26
20,000	BNG	20	Fail	00:32:19	01:14:34	00:07:14
20,000	BNG	30	Fail	00:44:43	00:42:44	00:08:46
20,000	BNG	40	Fail	00:45:28	00:27:35	00:08:59
50,000	BNG	10	Fail	01:51:23	13:41:03	00:22:43
50,000	BNG	20	Fail	02:38:41	10:32:39	00:33:07

50,000	BNG	30	Fail	03:34:28	08:26:21	00:32:21
50,000	BNG	40	Fail	03:14:35	05:37:02	00:32:57
100,000	BNG	10	Fail	06:28:35	52:08:32	01:06:01
100,000	BNG	20	Fail	09:26:27	41:30:19	01:32:23
100,000	BNG	30	Fail	11:36:35	33:12:06	01:49:11
100,000	BNG	40	Fail	12:21:04	22:13:29	02:17:22

Appendix B. Accuracy variance percentage

Size	Data	% Missing Value	Package	SVM	LR	ANN
10,000	Poker	10	VIM	1.43	0.38	4.54
10,000	Poker	20	VIM	3.73	1.41	4.81
10,000	Poker	30	VIM	4.9	1.43	1.04
10,000	Poker	40	VIM	6.45	1.35	5.38
15,000	Poker	10	VIM	2.91	1.66	0.04
15,000	Poker	20	VIM	3.64	0.47	3.74
15,000	Poker	30	VIM	6.04	1.82	2.24
15,000	Poker	40	VIM	7.13	0.14	1.91
20,000	Poker	10	VIM	1.78	0.47	4.06
10,000	Poker	10	MICE	3.08	1.45	2.89
10,000	Poker	20	MICE	4.04	2.84	2.11
10,000	Poker	30	MICE	4.81	3.59	3.24
10,000	Poker	40	MICE	5.75	4.31	10.53
15,000	Poker	10	MICE	3.32	1.93	3.53
15,000	Poker	20	MICE	5.51	2.01	1.43
15,000	Poker	30	MICE	6.58	2.89	3.32
15,000	Poker	40	MICE	6.23	1.64	0.48
20,000	Poker	10	MICE	2.55	1.38	0.66
20,000	Poker	20	MICE	4.98	3.97	0.39
20,000	Poker	30	MICE	6.75	3.73	2.57
20,000	Poker	40	MICE	6.59	3.04	3.89
50,000	Poker	10	MICE	2.59	1.66	4.71
50,000	Poker	20	MICE	5.67	4.68	0.02
50,000	Poker	30	MICE	8.16	4.52	0.57
50,000	Poker	40	MICE	10.21	6.29	2.7
100,000	Poker	10	MICE	2.29	3.15	0.44
100,000	Poker	20	MICE	5.32	7.45	0.44
100,000	Poker	30	MICE	9.14	8.11	4.68
100,000	Poker	40	MICE	11.69	8.4	0.45
10,000	Poker	10	MissForest	0.77	2.8	4.85
10,000	Poker	20	MissForest	1.85	3.25	8.63
10,000	Poker	30	MissForest	2.53	1.67	4.4
10,000	Poker	40	MissForest	4.6	3.02	8.07
15,000	Poker	10	MissForest	1.25	2.23	1.2
15,000	Poker	20	MissForest	1.18	2.19	1.87
15,000	Poker	30	MissForest	2.77	3.83	1.29
15,000	Poker	40	MissForest	2.89	3.79	1.24
20,000	Poker	10	MissForest	0.31	1.64	0.43
20,000	Poker	20	MissForest	0.84	4.56	0.78
20,000	Poker	30	MissForest	2.36	5.76	0.18
20,000	Poker	40	MissForest	2.76	4.91	4.96
50,000	Poker	10	MissForest	1.27	2.14	0.24
50,000	Poker	20	MissForest	1.39	6.23	0.17
50,000	Poker	30	MissForest	2.39	8.14	0.88
50,000	Poker	40	MissForest	3.83	8.94	0.29
100,000	Poker	10	MissForest	0.72	5.79	0.15
100,000	Poker	20	MissForest	2.04	11.13	0.54
100,000	Poker	30	MissForest	3.03	11.86	5.07
100,000	Poker	40	MissForest	5.18	11.63	5.05
10,000	Poker	10	HMISC	3	2.96	0.12
10,000	Poker	20	HMISC	4.22	1.92	9.06
10,000	Poker	30	HMISC	4.65	3.49	5.22
10,000	Poker	40	HMISC	5.7	3.85	11.08

15,000	Poker	10	HMISC	2.25	2.15	1.87
15,000	Poker	20	HMISC	4.99	0.06	2.56
15,000	Poker	30	HMISC	6.15	3.67	1.38
15,000	Poker	40	HMISC	6.2	3.22	1.99
20,000	Poker	10	HMISC	3.46	1.5	0.07
20,000	Poker	20	HMISC	4.33	4.04	4.78
20,000	Poker	30	HMISC	6.34	4.38	0.2
20,000	Poker	40	HMISC	7.69	2.92	4.42
50,000	Poker	10	HMISC	2.61	1.64	4.73
50,000	Poker	20	HMISC	9.68	7.25	0.1
50,000	Poker	30	HMISC	7.62	3.99	0.44
50,000	Poker	40	HMISC	9.68	7.25	0.15
100,000	Poker	10	HMISC	2.4	4.14	2.73
100,000	Poker	20	HMISC	6.18	8.71	0.1
100,000	Poker	30	HMISC	7.88	9.21	0.33
100,000	Poker	40	HMISC	12.09	10.3	0.5
10,000	BNG	10	VIM	0.19	0.03	0.57
10,000	BNG	20	VIM	0.73	0.18	0.7
10,000	BNG	30	VIM	2.1	0.17	0.94
10,000	BNG	40	VIM	2.59	0.2	1.2
15,000	BNG	10	VIM	0.7	0.02	0.34
15,000	BNG	20	VIM	0.6	0.07	0.07
15,000	BNG	30	VIM	1.48	0.16	0.01
15,000	BNG	40	VIM	2.63	0.24	0.16
20,000	BNG	10	VIM	0.13	0.08	0.02
10,000	BNG	10	MICE	0.56	0.14	0.52
10,000	BNG	20	MICE	1.9	0.16	0.95
10,000	BNG	30	MICE	3.45	0.29	0.86
10,000	BNG	40	MICE	5.1	1.01	0.87
15,000	BNG	10	MICE	0.48	0.07	0.25
15,000	BNG	20	MICE	1.07	0.08	0.03
15,000	BNG	30	MICE	2.35	0.18	0.08
15,000	BNG	40	MICE	4.45	0.76	0.93
20,000	BNG	10	MICE	0.48	0.06	0.06
20,000	BNG	20	MICE	1.5	0.15	0.45
20,000	BNG	30	MICE	2.31	0.42	0.29
20,000	BNG	40	MICE	3.85	0.61	0.45
50,000	BNG	10	MICE	0.23	0.05	0.07
50,000	BNG	20	MICE	0.7	0.02	0.42
50,000	BNG	30	MICE	1.55	0.23	0.29
50,000	BNG	40	MICE	3.32	0.63	0.39
100,000	BNG	10	MICE	0.42	0.05	0.15
100,000	BNG	20	MICE	0.94	0.09	0.03
100,000	BNG	30	MICE	1.27	0.29	0.13
100,000	BNG	40	MICE	2.26	0.44	0.21
10,000	BNG	10	MissForest	0.74	0.02	0.36
10,000	BNG	20	MissForest	2.21	0.02	0.4
10,000	BNG	30	MissForest	4.96	0.38	0.74
10,000	BNG	40	MissForest	8.15	0.46	0.91
15,000	BNG	10	MissForest	0.45	0.02	0.17
15,000	BNG	20	MissForest	1.59	0.01	0.19
15,000	BNG	30	MissForest	3.74	0.23	0.19
15,000	BNG	40	MissForest	8.31	1.48	0.69
20,000	BNG	10	MissForest	0.39	0	0.55
20,000	BNG	20	MissForest	1.31	0.02	0.03
20,000	BNG	30	MissForest	1.68	0.07	0.27
20,000	BNG	40	MissForest	3.57	0.32	0.27
50,000	BNG	10	MissForest	0.42	0	0.27
50,000	BNG	20	MissForest	0.73	0.08	0.23
50,000	BNG	30	MissForest	1.66	0.1	0.02
50,000	BNG	40	MissForest	3.28	0.14	0.44
100,000	BNG	10	MissForest	0.15	0.02	0.12
100,000	BNG	20	MissForest	0.71	0.09	0.35
100,000	BNG	30	MissForest	1.38	0	0.16
100,000	BNG	40	MissForest	2.7	0.11	0.03
10,000	BNG	10	HMISC	0.6	0.09	0.61

10,000	BNG	20	HMISC	1.81	0.18	0.56
10,000	BNG	30	HMISC	3.15	0.16	0.74
10,000	BNG	40	HMISC	4.95	0.75	0.92
15,000	BNG	10	HMISC	0.57	0.09	0.17
15,000	BNG	20	HMISC	0.52	0.05	0.01
15,000	BNG	30	HMISC	2.33	0.24	0.18
15,000	BNG	40	HMISC	3.97	0.59	1
20,000	BNG	10	HMISC	0.14	0.05	0.35
20,000	BNG	20	HMISC	1.03	0.19	0.2
20,000	BNG	30	HMISC	2.02	0.39	0.12
20,000	BNG	40	HMISC	3.51	0.22	0.27
50,000	BNG	10	HMISC	0.07	0.01	0.26
50,000	BNG	20	HMISC	0.55	0.01	0.43
50,000	BNG	30	HMISC	1.39	0.16	0.66
50,000	BNG	40	HMISC	2.45	0.23	0.07
100,000	BNG	10	HMISC	0.32	0.05	0.17
100,000	BNG	20	HMISC	0.68	0.09	0.03
100,000	BNG	30	HMISC	1.09	0.24	0.44
100,000	BNG	40	HMISC	2.3	0.29	0.45

Appendix C. Variance decrease percentage

Size	Data	Protocol	Missing Value (%)	Field-1	Field-2	Field-3	Field-4	Field-5
10,000	BNG	HMISC	10	6.26	6.23	8.55	7.89	5.96
10,000	BNG	HMISC	20	14.73	11.23	14.65	15.80	12.56
10,000	BNG	HMISC	30	22.52	19.99	22.49	22.88	20.24
10,000	BNG	HMISC	40	27.58	24.58	31.24	30.84	28.32
10,000	BNG	MissForest	10	7.59	7.34	9.73	9.15	7.29
10,000	BNG	MissForest	20	16.07	13.59	17.44	18.69	14.89
10,000	BNG	MissForest	30	24.50	22.39	26.52	26.42	22.28
10,000	BNG	MissForest	40	29.28	27.23	35.57	35.44	29.89
10,000	BNG	MICE	10	6.42	5.84	8.34	7.91	6.00
10,000	BNG	MICE	20	14.82	11.35	14.72	16.24	12.81
10,000	BNG	MICE	30	22.67	20.04	22.55	22.64	20.43
10,000	BNG	MICE	40	28.27	24.97	31.45	30.60	29.32
10,000	BNG	VIM	10	3.99	3.21	6.56	5.64	4.19
10,000	BNG	VIM	20	10.26	4.82	10.43	12.03	8.73
10,000	BNG	VIM	30	15.77	11.20	15.60	15.92	14.25
10,000	BNG	VIM	40	17.48	13.05	22.37	22.71	19.52
15,000	BNG	HMISC	10	7.05	6.39	8.17	8.25	6.26
15,000	BNG	HMISC	20	13.49	12.39	15.48	14.94	13.56
15,000	BNG	HMISC	30	20.77	18.91	23.60	23.01	19.54
15,000	BNG	HMISC	40	28.57	25.96	30.94	30.46	27.37
15,000	BNG	MissForest	10	8.11	7.55	9.44	9.70	7.55
15,000	BNG	MissForest	20	15.24	14.40	18.41	17.78	15.69
15,000	BNG	MissForest	30	22.87	21.61	27.52	26.54	22.50
15,000	BNG	MissForest	40	29.96	28.15	35.45	34.83	28.99
15,000	BNG	MICE	10	6.95	6.44	8.00	8.30	6.25
15,000	BNG	MICE	20	13.65	12.36	15.64	15.07	13.77
15,000	BNG	MICE	30	20.75	19.62	23.52	22.82	19.99
15,000	BNG	MICE	40	28.69	26.63	31.00	30.43	27.73
15,000	BNG	VIM	10	4.66	3.33	5.75	6.42	4.10
15,000	BNG	VIM	20	8.87	6.34	11.67	10.88	9.35
15,000	BNG	VIM	30	13.47	10.44	16.08	16.45	13.19
15,000	BNG	VIM	40	18.26	14.41	21.17	21.87	18.02
20,000	BNG	HMISC	10	6.60	6.16	8.18	7.82	6.33
20,000	BNG	HMISC	20	14.83	13.04	16.53	15.70	12.35
20,000	BNG	HMISC	30	21.23	18.87	23.14	23.83	20.53
20,000	BNG	HMISC	40	29.36	25.65	31.29	30.79	27.39
20,000	BNG	MissForest	10	7.83	7.33	9.66	9.16	7.43
20,000	BNG	MissForest	20	16.25	14.98	19.08	18.39	14.26
20,000	BNG	MissForest	30	22.87	21.56	27.16	27.58	22.56
20,000	BNG	MissForest	40	31.35	27.89	35.57	35.41	29.02

20,000	BNG	MICE	10	6.47	6.42	8.17	7.92	6.31
20,000	BNG	MICE	20	14.56	13.16	16.40	15.62	12.21
20,000	BNG	MICE	30	21.16	19.70	23.27	23.75	20.75
20,000	BNG	MICE	40	29.64	26.69	31.11	30.90	27.44
20,000	BNG	VIM	10	4.35	3.15	5.69	5.37	4.11
50,000	BNG	HMISC	10	7.07	6.37	7.96	7.74	6.67
50,000	BNG	HMISC	20	13.91	12.78	15.88	15.49	13.30
50,000	BNG	HMISC	30	21.50	19.62	23.60	23.75	19.89
50,000	BNG	HMISC	40	29.15	25.16	31.15	31.70	26.51
50,000	BNG	MissForest	10	8.28	7.61	9.42	9.22	7.71
50,000	BNG	MissForest	20	15.68	14.91	18.76	18.38	15.14
50,000	BNG	MissForest	30	23.52	22.76	27.60	27.78	22.33
50,000	BNG	MissForest	40	30.97	27.62	35.98	36.14	28.68
50,000	BNG	MICE	10	7.14	6.33	8.03	7.70	6.50
50,000	BNG	MICE	20	13.73	12.87	16.06	15.39	13.35
50,000	BNG	MICE	30	21.61	20.03	23.32	23.73	19.83
50,000	BNG	MICE	40	29.23	25.99	31.21	31.63	27.24
100,000	BNG	HMISC	10	7.07	6.41	7.96	7.82	6.63
100,000	BNG	HMISC	20	14.21	12.54	15.57	15.81	13.17
100,000	BNG	HMISC	30	21.80	18.86	23.15	24.01	20.07
100,000	BNG	HMISC	40	28.81	25.36	32.17	31.62	26.74
100,000	BNG	MissForest	10	8.20	7.61	9.51	9.27	7.84
100,000	BNG	MissForest	20	15.93	14.92	18.52	18.81	15.24
100,000	BNG	MissForest	30	23.60	22.27	27.34	28.23	22.35
100,000	BNG	MissForest	40	30.40	28.61	37.00	36.76	28.64
100,000	BNG	MICE	10	7.02	6.35	8.05	7.77	6.65
100,000	BNG	MICE	20	14.15	12.56	15.41	15.90	13.20
100,000	BNG	MICE	30	21.71	19.41	23.11	23.93	20.25
100,000	BNG	MICE	40	28.86	26.27	31.98	31.70	27.32
10,000	Poker	HMISC	10	8.10	7.61	8.34	7.59	7.94
10,000	Poker	HMISC	20	15.83	15.00	16.09	15.71	15.85
10,000	Poker	HMISC	30	23.65	23.33	23.68	23.86	23.85
10,000	Poker	HMISC	40	31.89	31.46	31.62	31.46	31.73
10,000	Poker	MissForest	10	9.23	8.77	9.56	8.95	9.12
10,000	Poker	MissForest	20	17.99	17.35	18.27	17.95	17.95
10,000	Poker	MissForest	30	26.20	26.49	26.12	26.71	26.44
10,000	Poker	MissForest	40	34.95	33.75	34.22	33.68	34.41
10,000	Poker	MICE	10	8.05	7.58	8.28	7.69	7.89
10,000	Poker	MICE	20	15.75	15.12	16.01	15.63	15.78
10,000	Poker	MICE	30	23.55	23.55	23.64	23.43	23.94
10,000	Poker	MICE	40	31.58	31.28	31.40	31.16	31.81
10,000	Poker	VIM	10	3.86	3.40	4.45	3.35	3.57
10,000	Poker	VIM	20	7.02	6.58	7.40	7.07	6.63
10,000	Poker	VIM	30	11.17	10.75	10.40	9.64	10.43
10,000	Poker	VIM	40	14.14	14.14	13.72	11.95	12.91
15,000	Poker	HMISC	10	7.99	8.04	7.83	7.69	7.80
15,000	Poker	HMISC	20	15.22	15.63	15.53	15.99	15.93
15,000	Poker	HMISC	30	23.46	24.34	23.29	24.11	23.93
15,000	Poker	HMISC	40	32.06	31.65	31.08	31.32	32.04
15,000	Poker	MissForest	10	9.22	9.18	8.93	8.88	9.11
15,000	Poker	MissForest	20	17.31	17.96	17.66	18.25	17.97
15,000	Poker	MissForest	30	26.12	26.88	26.16	26.78	26.36
15,000	Poker	MissForest	40	34.97	34.07	33.50	33.83	33.69
15,000	Poker	MICE	10	7.99	8.05	7.73	7.81	7.78
15,000	Poker	MICE	20	15.18	15.75	15.38	16.12	15.65
15,000	Poker	MICE	30	23.40	24.53	23.40	24.25	23.85
15,000	Poker	MICE	40	32.08	31.75	31.13	31.41	31.77
15,000	Poker	VIM	10	3.72	3.90	3.35	3.27	3.07
15,000	Poker	VIM	20	6.58	7.34	6.66	7.19	6.45
15,000	Poker	VIM	30	10.87	12.13	10.53	10.36	9.67
15,000	Poker	VIM	40	14.32	13.95	12.86	12.85	12.70
20,000	Poker	HMISC	10	7.87	7.91	7.95	7.92	7.99
20,000	Poker	HMISC	20	15.51	15.51	15.77	15.64	15.73
20,000	Poker	HMISC	30	23.45	23.26	23.51	23.82	24.01
20,000	Poker	HMISC	40	31.05	31.64	31.65	31.78	31.49
20,000	Poker	MissForest	10	9.02	9.04	9.13	9.08	9.17

20,000	Poker	MissForest	20	17.38	17.72	17.83	17.54	17.75
20,000	Poker	MissForest	30	26.12	25.86	25.81	25.90	26.30
20,000	Poker	MissForest	40	33.38	34.19	34.26	33.78	32.85
20,000	Poker	MICE	10	7.86	7.75	7.91	7.83	7.98
20,000	Poker	MICE	20	15.49	15.33	15.80	15.68	15.79
20,000	Poker	MICE	30	23.67	23.55	23.29	23.85	24.08
20,000	Poker	MICE	40	30.99	31.67	31.60	31.63	31.64
20,000	Poker	VIM	10	3.63	3.28	3.40	3.17	3.25
50,000	Poker	HMISC	10	7.93	7.92	7.59	7.90	7.71
50,000	Poker	HMISC	20	15.71	15.90	15.90	15.70	15.81
50,000	Poker	HMISC	30	23.50	23.43	23.24	23.85	23.58
50,000	Poker	HMISC	40	31.20	31.83	31.27	31.52	31.87
50,000	Poker	MissForest	10	9.06	9.01	8.67	8.99	8.85
50,000	Poker	MissForest	20	17.54	17.77	17.72	17.47	17.65
50,000	Poker	MissForest	30	23.50	23.43	23.24	23.85	23.58
50,000	Poker	MissForest	40	33.59	33.50	33.03	32.41	33.18
50,000	Poker	MICE	10	7.92	7.89	7.46	7.83	7.81
50,000	Poker	MICE	20	15.72	15.76	15.86	15.69	15.90
50,000	Poker	MICE	30	23.44	23.49	23.38	23.85	23.52
50,000	Poker	MICE	40	31.34	32.00	31.40	31.60	31.83
100,000	Poker	HMISC	10	7.81	7.82	7.95	7.88	7.74
100,000	Poker	HMISC	20	15.78	15.84	15.81	15.84	15.76
100,000	Poker	HMISC	30	23.83	23.64	23.88	23.70	23.84
100,000	Poker	HMISC	40	31.85	31.68	31.39	31.43	31.73
100,000	Poker	MissForest	10	8.80	8.89	8.95	8.88	8.76
100,000	Poker	MissForest	20	17.45	17.67	17.29	17.45	17.33
100,000	Poker	MissForest	30	25.84	25.61	25.46	25.23	24.77
100,000	Poker	MissForest	40	33.81	33.25	32.49	32.47	32.23
100,000	Poker	MICE	10	7.77	7.82	7.90	7.81	7.80
100,000	Poker	MICE	20	15.65	15.89	15.75	15.88	15.80
100,000	Poker	MICE	30	23.77	23.62	23.90	23.66	23.90
100,000	Poker	MICE	40	31.87	31.73	31.42	31.40	31.75

Note: in each table *Poker* represents 'poker hand' dataset and *BNG* represents 'BNG heart_statlog'.

References

[1] P.D. Allison, S. Horizons, Handling missing data by maximum likelihood, SAS Global Forum 2012 (2012) 1–21.

[2] M. Amiri, R. Jensen, Missing data imputation using fuzzy-rough methods, Neurocomputing 205 (2016) 152–164.

[3] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Inf. Sci. 233 (2013) 25–35.

[4] G.E.A.P.A. Batista, M.C. Monard, A study of K-nearest neighbour as an imputation method, Soft Comput. Syst. Des. Manag. Appl. 87 (2002) 251–260.

[5] L.A. Belanche, V. Kobayashi, J. Aluja, Handling missing values in kernel methods with application to microbiology data, Neurocomputing 141 (2014) 110–116.

[6] G.N. Brock, J.R. Shaffer, R.E. Blakesley, M.J. Lotz, G.C. Tseng, Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, BMC Bioinform. 12 (2008) 1–12.

[7] L.F. Burgette, J.P. Reiter, Multiple imputation for missing data via sequential regression trees, Am. J. Epidemiol. 172 (9) (2010) 1070–1076.

[8] S. van Buuren, K. Groothuis-oudshoorn, mice : multivariate imputation by chained, J.Stat. Softw. 45 (3) (2011).

[9] V. Capuano, F. Chiclana, H. Fujita, E. Herrera-Viedma, V. Loia, Fuzzy group decision making with incomplete information guided by social influence, IEEE Trans. Fuzzy Syst. 26 (3) (2017) 1–15.

[10] G. Chang, T. GE, Comparison of missing data imputation methods for traffic flow, Proceedings of the International Conference on Transportation, Mechanical, and Electrical Engineering, 2011, pp. 639–642.

[11] Q. Chen, M. Chu, M. Kim, C. Wang, Using link-preserving imputation for logistic partially linear models with missing covariates, Comput. Stat. Data Anal. 101 (2016) 174–185.

[12] X. Cheng, D. Cook, H. Heike, Visually exploring missing values in multivariable, J. Stat. Softw. 68 (6) (2015).

[13] R. Deb, A.W. Liew, Missing value imputation for the analysis of incomplete traffic accident data, Inf. Sci. 339 (2016) 274–289.

[14] R. Devi Priya, R. Sivaraj, N. Sasi Priyaa, Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases, Knowl.-Based Syst. 133 (2017) 107–121.

[15] Di Nuovo, A. G., Missing data analysis with fuzzy C-means: a study of its application in a psychological scenario, Expert Syst. Appl. 38 (6) (2011) 6793–6797.

[16] A. Farhangfar, L.A. Kurgan, W. Pedrycz, A novel framework for imputation of missing values in databases, IEEE Trans. Syst. Man Cybern.—Part A: Syst. Hum. 37 (2007) 692–709.

[17] M. Fichman, J.M. Cummings, Multiple Imputation for missing data : making the most of what you know, Organ. Res. Methods 6 (3) (2003) 282–308.

[18] P.J. Garcia-Laencina, J.-L. Sancho-Gomez, R.A. Figueiras-Vidal, Pattern classification with missing data : a review, Neural Comput Appl. 19 (2010) 263–282.

[19] J.W. Graham, Missing data analysis : making it work in the real world, Ann. Rev. Psychol. 60 (2009) 549–576.

[20] N.J. Horton, S.R. Lipsitz, N.J.H. Orton, S.R.L. Ipsitz, Multiple imputation in practice : comparison of software packages for regression models with missing variables, Am. Stat. 55 (3) (2001) 244–254.

[21] E. John, M. Carranza, A.G. Laborte, Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra, Comput. Geosci. 74 (2015) 60–70.

[22] P. Jonsson, W. Claes, An evaluation of k -nearest neighbour imputation using Likert data per Jönsson and Claes Wohlin, Proceedings of the 10th International Symposium on Software Metrics, IEEE, 2004, pp. 108–118.

[23] A. Karmaker, S. Kwek, Incorporating an EM-approach for handling missing attribute-values in decision tree induction, Proceedings of Fifth International Conference on Hybrid Intelligent Systems, 2005. HIS'05, IEEE, 2005, p. 6.

[24] A. Kowarik, M. Templ, Imputation with the R Package VIM, J. Stat. Softw. 74 (7) (2016) 1–16.

[25] V. Kumutha, S. Palaniammal, An Enhanced approach on handling missing values using bagging k-NN imputation, Proceedings of International Conference on Computer Communication and Informatics, 2013, pp. 1–8.

[26] T.Y. Kwon, Y. Park, A new multiple imputation method for bounded missing values, Stat.Probab. Lett. 107 (2015) 204–209.

[27] K.J. Lee, J.B. Carlin, Multiple imputation for missing data : fully conditional specification versus multivariate normal imputation, Am. J. Epidemiol. 171 (5) (2010) 624–632.

[28] S.G. Liao, Y. Lin, D.D. Kang, D. Chandra, J. Bon, N. Kaminski, G.C. Tseng, Missing value imputation in high-dimensional phenomic data : imputable or not , and how , BMC Bioinform. 15 (2014) 1–12.

[29] R.J.A. Little, D.B. Rubin, Statistical analysis with missing data, J. Educ. Stat. 16 (2) (1987) 150–155.

[30] F. Lobato, C. Sales, I. Araujo, V. Tadaiesky, L. Dias, L. Ramos, A. Santana, Multi-objective genetic algorithm for missing data imputation, Pattern Recognit. Lett. 68 (2015) 126–131.

[31] J. Luengo, S. García, F. Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods, Knowl. Inf. Syst. 32 (2011) 77–108.
[32] M.R. Malarvizhi, A.S. Thanamani, K-nearest neighbor in missing data imputation, Int. J. Eng. Res. Dev. 5 (1) (2012) 5–7.
[33] M. Misztal, Imputation of missing data using R package, Acta Universitatis Lodziensis. Folia Oeconomica 269 (2012) 131–144.
[34] Z. Qiu, X. Chen, Y. Zhou, A kernel-assisted imputation estimating method for the additive hazards model with missing censoring indicator, Stat. Probab. Lett. 98 (2015) 89–97.
[35] I.E.W. Rachmawan, A.R. Barakbah, Optimization of missing value imputation using reinforcement programming, Proceedings of the International Electronics Symposium (IES), 2015, pp. 128–133.
[36] P. Royston, Multiple imputation of missing values, Stata J. 4 (3) (2004) 227–241.
[37] D. Rubin, Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse, Proceedings of the Section on Survey Research Methods, The Association, 1978, pp. 20–34.
[38] Z. Sahri, R. Yusof, J. Watada, FINNIM : iterative imputation of missing values in dissolved gas analysis dataset, IEEE Trans. Ind. Inform. 10 (4) (2014) 2093–2102.
[39] J. Scheffer, Dealing with missing data, R.L.I.M.S. 3 (2002) 153–160.
[40] S. Nathaniel, J.M.G. Taylor, Partially parametric techniques for multiple imputation, computational statistics & data analysis 22 (4) (1996) 425–446.
[41] S.-B. Marina, Dealing with missing data: Key assumptions and methods for applied analysis, Boston University 4 (2013) 1–19.
[42] F. Soltanveis, Using parametric regression and KNN Algorithm with missing handling for software effort prediction, Proceedings of Artificial Intelligence and Robotics (IRANOPEN), 2016, pp. 77–84.
[43] D. Sovilj, E. Eirola, Y. Miche, K. Björk, R. Nian, A. Akusok, A. Lendasse, Extreme learning machine for missing data using multiple imputations, Neurocomputing 174 (2016) 220–231.
[44] D.J. Stekhoven, P. Bühlmann, MissForest — non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.
[45] E.A. Stuart, M. Azur, C. Frangakis, P. Leaf, Multiple imputation with large data sets : a case study of the children ' s mental health initiative, Am. J. Epidemiol. 21205 (2009) 1–7.
[46] J. Tang, G. Zhang, Y. Wang, H. Wang, F. Liu, A hybrid approach to integrate fuzzy C- means based imputation method with genetic algorithm for missing traffic volume data estimation, Transp. Res. Part C 51 (2015) 29–40.
[47] C. Tsai, F. Chang, Combining instance selection for better missing value imputation, J. Syst. Softw. 122 (2016) 63–71.
[48] R. Urena, F. Chiclana, J.A. Morente-Molinera, E. Herrera-Viedma, Managing incomplete preference relations in decision making: a review and future trends, Inf. Sci. 302 (2015) 14–32.
[49] R. Urena, F. Javier Cabrerizo, J. Antonio Morente-Molinera, E. Herrera-Viedma, GDM-R: a new framework in R to support fuzzy group decision making processes, Inf. Sci. 357 (2016) 161–181.
[50] J. Wu, F. Chiclana, Multiplicative consistency of intuitionistic reciprocal preference relations and its application to missing values estimation and consensus building, Knowl.-Based Syst. 71 (2014) 187–200.