A Comparison of Statistical Packages in R Tool to Impute Missing Values

Mrs.D.Suganthi\(^1\), Mr.K.Dheenathayalan\(^2\)

\(^1\)(Department of Computer Science, Sankara college of science and commerce, Coimbatore, India)  
\(^2\)(Department of Computer Science, Kamban college of arts and science, Coimbatore, India)

**Abstract:** Data mining has pushed the realm of information technology beyond predictable limits. Missing value is one of the major factor, which can render the obtain result beyond use attained from specific data set by applying data mining technique. There could be numerous reasons for missing values in a data set such as human error, hardware malfunction etc. It is imperative to tackle the labyrinth of missing values before applying any technique of data mining; otherwise, the information extracted from data set containing missing values will lead to the path of wrong decision making. Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present. Several methods have been, and continue to be, developed to draw inferences from data sets with missing values. In this work, we experimented and results are compared for three methods of imputation of missing values in numerical dataset. We compare MICE (PMM, CART and SAMPLE), HMISC, HOT.DECK, AMELIA, kNN and MISSFOREST, which are likely the most sophisticated imputation methods currently employed for imputing missing values.

**Keywords:** Data mining, distance measure, Imputation, missing values

I. Introduction

Missing data is one of the issues which are to be fathomed for real-time application. Improper imputation produces predisposition result. For example, manual information entrances system, inaccurate estimations, gear blunders, and numerous others. Hence legitimate consideration is expected to credit the missing values. Research data are prone to missing data. Missed observations may occur due to human error. For example, a researcher may forget to take a measurement such as the patient’s pulse. Or, a malfunction which eliminates a measurement [1].

Databases also have missing data. Whenever there is a mismatch of variables between databases, there are missing occurrences. For example, a database analyst is analyzing sales databases consolidated from three regions. If the central region did not record one variable such as the educational background of sales representatives, then this variable would have missing occurrences when the three databases are merged [2]. Missing data can be serious Missing data can be treacherous because it is difficult to identify the problem. Each question or variable may only have a small number of missing responses, but in combination, the missing data could be numerous. Only thorough analysis on missing data can determine whether missing data are problematic [3].

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. The concept of missing values is important to understand in order to successfully manage data. Until now, this analysis has been time consuming and error prone. Missing data can cause serious problems. First, most statistical procedures automatically eliminate cases with missing data. This means that in the end, we may not have enough data to perform the analysis. Second, the analysis might run but the results may not be statistically significant because of the small amount of input data. Third, results may be misleading if the cases you analyze are not a random sample of all cases. Also, it is not always obvious when missing data will cause a problem.

II. Dataset

The example dataset taken is health dataset, which contains Height in meter, Weight in Kg, BMI and Fat percentage of individual person. Totally there are 4 variables with 92 observations. Initially there were no missing values, for testing purpose some values are randomly deleted. After random deletion there are 18 missing values (9 values in BMI and 9 values in Fat percentage). Figure 1 shows the snapshot of the dataset contains missing values.
A Comparison of Statistical Packages in R Tool to Impute Missing Values

Next Generation Computing Technologies
Sankara College Of Science And Commerce

Figure 1. Dataset with missing values (NA)

In the above figure NA represents missing value (Not Available). This work is carried out using R tool. Figure 2 Represents missing values map in dataset.

The figure shows that there are no missing values in Weight.Kg and Height.M. Variable X.Fat contains 9 missing values and variable BMI also contains 9 missing values. The numbers on left side of the figure represents number of observations, totally 92 observations are there.
III. Experimental Setup

3.1 Packages Used For Imputation

a. MICE
Multiple imputation is the method of choice for complex incomplete data problems. Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Methods of MICE are PMM (predictive mean matching), CART, Sample [4].

b. Hmisc
Contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values, advanced table making, variable clustering, character string manipulation, conversion of R objects to LaTeX and html code, and recoding variables.

c. HotDeck
Historically, the term “hot deck” comes from the use of computer punch cards for data storage, and refers to the deck of cards for donors available for a non-respondent [5]. The deck was “hot” since it was currently being processed, as opposed to the “cold deck” which refers to using pre-processed data as the donors, i.e. data from a previous data collection or a different data set. Hot deck imputation methods to resolve missing data. Hot deck imputation involves replacing missing values of one or more variables with observed values.

d. Knn Impute
k Nearest Neighbor (kNN) is a data mining technique to impute missing expression data in microarray. For each gene with missing values, it finds the k nearest neighbors using a distance metric, confined to the columns for which that gene is NOT missing. Each candidate neighbor might be missing some of the coordinates used to calculate the distance. In this case we average the distance from the non-missing coordinates. Having found the k nearest neighbors for a gene, it imputes the missing elements by averaging those (non-missing) elements of its neighbors.

e. Amelia
Amelia II performs multiple imputation, a general-purpose approach to data with Missing values. Multiple imputation has been shown to reduce bias and increase efficiency compared to list wise deletion. Multiple imputation involves imputing m values for each missing cell in your data matrix and creating m completed data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data.

f. Missforest
One of the nice “extras” of the random forest algorithm is its use for mixed data type (numeric and categorical) imputation. Unlike other modeling approaches such as generalized linear models which employ a case-wise deletion, random forests will throw an error when it encounters missing values [6]. Out of the box, the random forest algorithm will perform imputation by leveraging a “proximity matrix” which also allowing for random forests to be used for outlier detection and clusters. As each tree in the RF is built, all cases are “run down” the tree and if two cases (i and j) end up in the same terminal node – they are considered to be relatively alike according to the model. A NxN matrix is ultimately produced with element (i,j) incremented each time case i and j end up in the same node. This is normalized for the number of trees in the forest.
IV. Results And Discussion

a. Mice

![Figure 3. MICE-PMM imputation](image)

![Figure 4. MICE-CART imputation](image)

![Figure 5. MICE-Sample imputation](image)
b. Hmisc

Figure 6: HMISC imputation

Figure 6 Shows result of HMISC method after imputing the missing values in the health dataset

c. Hot.Deck

Figure 7: Hot.Deck imputation

Figure 7 Shows result of Hot.Deck method after imputing the missing values in the health dataset

d. Knn Impute

Figure 8: kNN imputation

Figure 8 Shows result of kNN-impute method after imputing the missing values in the health dataset
e. Amelia

Figure 9: Amelia imputation

Figure 9 Shows result of Amelia method after imputing the missing values in the health dataset

f. Missforest

Figure 10: MissForest imputation

Figure 10 Shows result of MissForest method after imputing the missing values in the health dataset

The mentioned methods are used to impute the missing values in the given health dataset. All of the missing values are imputed and replaced in their appropriate position.
Figure 11 shows the missingness map of the health dataset after imputing the missing values. From the figure, it is observed that there is no more missing values after imputation.

V. Discussion

a. Distance measure - Euclidean distance method

The results from all methods are analyzed using distance measure (Euclidean distance method). Here the method Amelia has the minimum distance between the original values and imputed values (Figure 12 and Figure 13). Detailed after the analyses it is come to know that Amelia is the best method to impute missing values in numerical dataset.

![Distance Measures of Imputation Methods for Variable BMI](image1)

**Figure: 12**

Figure 12 Represents bar chart of distance measures of imputation methods for variable BMI in health dataset

![Distance Measures of Imputation Methods for Variable FAT](image2)

**Figure: 13**

Figure 13 Represents bar chart of distance measures of imputation methods for variable FAT in health dataset.

VI. Conclusion And Future Work

Incomplete data comes (missing values) from NA data value when collected. Different considerations between the times when the data was collected and when it is analyzed. Human/hardware/software problems. Missing values can cause major risks, so it is necessary to impute the missing values. There are many statistical methods and tools available for imputation.

Among only few familiar methods are selected and tested in R using health dataset. From the analyses made it shows Amelia is the best method to impute missing values. In future this work can be extended by using other statistical methods and large datasets.

References


A Comparison of Statistical Packages in R Tool to Impute Missing Values

