International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# ANALYSIS OF MISSING DATA USING MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE) IN R

**O. Mrudula [1], Dr. A. Mary Sowjanya[2]**

[1,2]*Department of Computer Science and Systems Engineering, College of Engineering, Andhra University, Visakhapatnam (India)*

## ABSTRACT

*Data Mining is a process of exploration and analysis of large quantity of data in order to uncover previously unknown patterns. At present data mining is highly used for processing and accessing large volumes of data. In this project, analysis of missing data has been implemented for identifying and replacing missing values using Multivariate Imputation by Chained Equations (MICE). In MICE, the analysis of imputed data is made completely general, whereas the range of models under which pooling works is substantially extended. MICE adds new functionality for imputing multilevel data, automatic predictor selection, data handling, post-processing imputed values, specialized pooling and model selection. MICE V2.0 is freely available from CRAN as R package mice. This paper provides a stepwise approach to using mice for solving incomplete data problems in real data. The Air Quality dataset which is available in R was used to address missing data problem by multiple imputation.*

***Keywords: Data mining, missing value, imputation technique, MICE, multiple imputations, R language***

## I.INTRODUCTION

Data mining has made a great progress in recent year but the problem of missing data or value has remained great challenge for data mining. Missing data or value in a datasets can affect the performance of classifier which leads to difficulty of extracting useful information from datasets. It refers to extracting knowledge from large amounts of data. The data may be spatial data, multimedia data, time series data, text data and web data. It is the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data [1].

### 1.1 MISSING VALUES

Missing values lead to the difficulty of extracting useful information from that data set [2]. Missing data are the absence of data items that hide some information that may be important [1].

**Type of missing data**

There is different type of missing value:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

**MCAR**

The term "Missing Completely at Random" refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset. [3]. MCAR is the prospect of a record possessing a missing value of the attribute but it does not depends on the missing data or the observed data [4].

**MAR**

Sometimes data might not be missing at random but may be termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable. [4]. MAR is to be a condition in which the probability that data are missing depends only on the observed data but on the missing data, after controlling for observed data.

**NAMR**

If the data is not missing at random or informatively missing then it is termed as "Not missing at Random". Such a situation occurs when the missingness mechanism depends on the actual value of missing data. [4]. NMAR is the probability of a record containing missing value of field that depends on the value of attire [2].

## 1.2 Methods of Imputation

The process of replacing attributed values from the available data is known as Imputation. There are some imputation techniques from various experiments are done to choose the best one to handle missing value in a data set. Here are some approaches that have been empirically assessed to find the missing values [12].

- **Treating missing attribute as special value**: - This is totally different approach. Instead of finding some new value, the missing value is considered itself a special value for the instance that containing missing value.
- **Closest fit**:- This method based on replacing a missing attribute value with an existing attribute from the other case of the same attribute that resembles as much as possible the case with missing attribute values.
- **K-nearest neighbor**: - This method finds the K-nearest neighbors, and among all neighbors the most common value is considered for nominal attributes.
- **Weighted imputation with K-nearest neighbor**: - In this the distance of each missing value instances from its neighbors is calculated. Here, the distance used for calculating weight. Missing values are calculated by weighted mean for numerical attributes.
- **K-means clustering imputation**: - This method is similar to "K-Nearest Neighbor" the instances are clustered by using K-means clustering. And the instances in each cluster are considered nearest neighbors of each other.
- **Fuzzy k-means clustering imputation**: - This method is better than K means imputation. Here the data object can be a part of more than one cluster centre. It is best method for overlapping data. In this unreformed attributes for every - uncompleted data are substituted. Missing values are calculated by weighted sum of all entroids on the basis of membership degrees.
- **Support vector machines imputation**: - This method is efficient in memory consumption and large dimensional spaces. This model is used to predict missing attributes from the complete instances which do not have missing values. Here the condition attributes and decision attributes are considered.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

- **Local least squares imputation**: - This fits a least squares model between the instances and known part of the record with missing values.
- **Regression Imputation**: - In this method the missing values are first observed and the predicted values are used for handling the missing values
- **Hot-Deck Imputation**: - This is a traditional method failed to give a complete simulation associated with the missing data. In this type of method each missing value is replaced with an observed response from a similar unit.
- **Missing Data Using Neural Networks-**. A neural network which is an information processing paradigms that is inspired by the biological neural nervous system it consist of four parts- processing units having a certain activation level at any point in time. The activation of one unit leads to the input for another which is resolved by weighted interconnections between different processing units. It works on an activation rule and a learning rule which specified how the weights are adjusted for a given input/output pair. They have the ability to analyze meaning from complicated data and also used to extract patterns.

### 1.3 MICE (Multivariate Imputation by Chained Equations) in R

Before a formal data analysis can be done it is important to know the details of the data set like variables included, missing values etc. Data comes in a form that is not easy to analyze it needs to be clean and checked for consistency.

R is a programming language which includes many functions for data analytics. Also R provides graphical facilities for data analysis and reporting. As such the proposed system makes use of R studio at the back end. The various packages and functionalities available in R have been used for predicting missing values, imputation of missing values and visualization of the imputed data set.

**Features of MICE:**

R package MICE 2.9 for multiple imputations is used for generating multiple imputations, analyzing imputed data and for pooling analysis results.

The specific features are as follows:

1) Column wise specification of imputation model.
2) Arbitrary patterns of missing data.
3) Passive imputation.
4) Subset selection of predictors.
5) Suppose of arbitrary complete-data methods.
6) Support pooling various types of statistics.
7) Diagnostics of Imputations.
8) Callable uses-written imputation functions.

Package MICE have a simple architecture, are highly modular and allow easy access to all program code with in R environment.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

## II. RELATED WORK

A number of studies related to data processing of missing value in a dataset have been carried out. The analysis and extraction of new patterns which are error free, complete with correct labels from available real world dataset is difficult. Imputation is the process of estimating or deriving values for fields where data is missing [5].

D.T. Larose and C.D. Larose have examined the methods for imputing missing values for continues and categorical variables [6].

Statistical reconstruction of incomplete datasets using multiple linear regressions to predict missing values was proposed by Shelke, M.B. et. Al [6], using weka tool.

Multiple imputations is the method of choice for complex incomplete data problems. There are two general approaches for imputing multivariate data.

1) Joint modeling (JM) by SCHAFER [7]
2) Fully conditional specification (FCS) also known as multivariate imputation by chained equations (MICE) [9, 10]

## III. METHODOLOGY

Missing data is quite common in real world. When dealing with data set having missing values one way to perform analysis is by discarding observation with any missing values. But it can lead to biast estimate, errors and incorrect results another way to deal with missing data is by imputing all missing values before analysis.

The goal of this paper is to show how to handle missing values using Multivariate Imputation by Chained Equations (MICE) implemented in R. The proposed approach of analysis of missing data and imputation on Air Quality data using MICE is as follows.

Step-1: Load the Air Quality data set into R.

Step-2: Predict missing values.

Step-3: Impute missing values.

Step-4: Now generate the filled data set.

Step-5: Visualize the imputed data set.

### 3.1 Implementation Procedure

Step-1: A simple imputation such as imputing the mean is performed for every missing value in the data set.

Step-2: These mean imputations are set back to missing.

Step-3: The observed values in step 2 are regressed on the other variables in the imputation model.

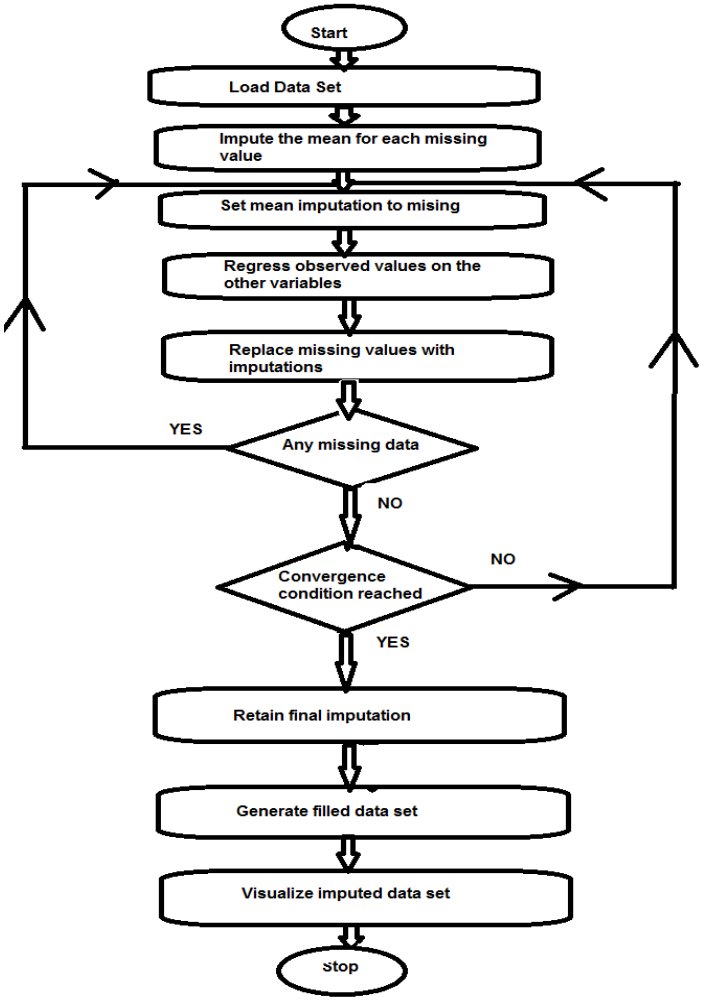
Step-4: The missing values are then replaced with imputations from the regression model.

Step-5: Repeat steps 2, 3, and 4 for each variables that have missing data.

Step-6: Repeat steps 2, 3, 4 for a number of user's specified cycles with the imputation being updated and at each cycle.

Step-7: Retain final imputation which results in one imputed data set.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346




**3.1.1 Flowchart for the Proposed System**

### 3.2. PMM (Predictive Mean Matching Method)

The predictive mean matching method is also an imputation method available for continuous variables. It is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are the closest to predicted value for the missing value from the simulated regression model.

Predictive Mean Matching Method Procedure

Begin

Step-1: Read the dataset

Step-2: Calculate variance

Step-3: Calculate regression coefficients

Step-4: For each missing value, a predicted value $y_{i*}$ to be calculated

Step-5: Compute the covariate values

Step-6: Generate predicted values

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Step-7: Missing values is the replaced by a predicted value

Step-8: Generate reconstructed and completed data

End

### 3.3 Linear Regression Models

The linear regression models are used to find the missing values and impute the missing values in the Air Quality data set. Estimating linear models is easy to predict.

### 3.4 Special box plot

Like the previous methods, the well-known boxplot method is also altered to display information about imputed values. The plot consists of three boxplots: First, a normal boxplot of the variable of interest is produced. Second, two boxplots are drawn, which are grouped by observed and imputed values in the additional variables and the values of them in the variable of interest. Additionally, the frequencies of observed and imputed values will be given for each boxplot.

A lot of information can be retrieved from this graphic. It gives a good overview about the distribution of the variable of interest. Also, outliers can be identified easily. Furthermore, by grouping between observed and imputed values, it can be analyzed.

## IV. RESULTS

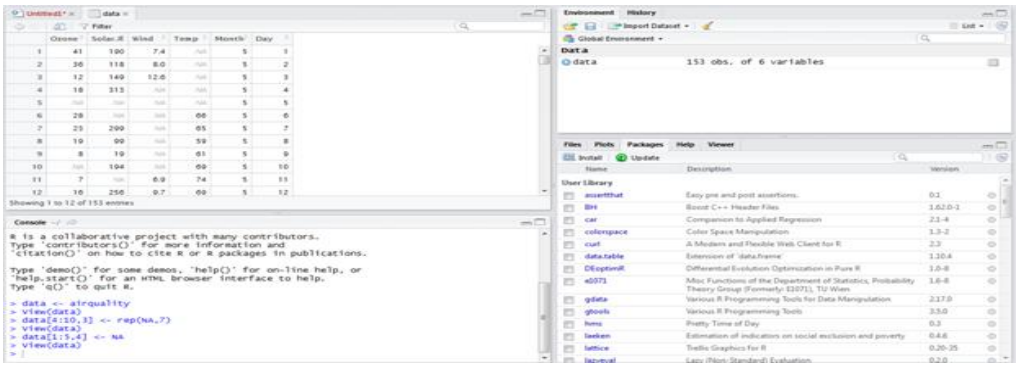


**Fig4.1. Handling missing data**

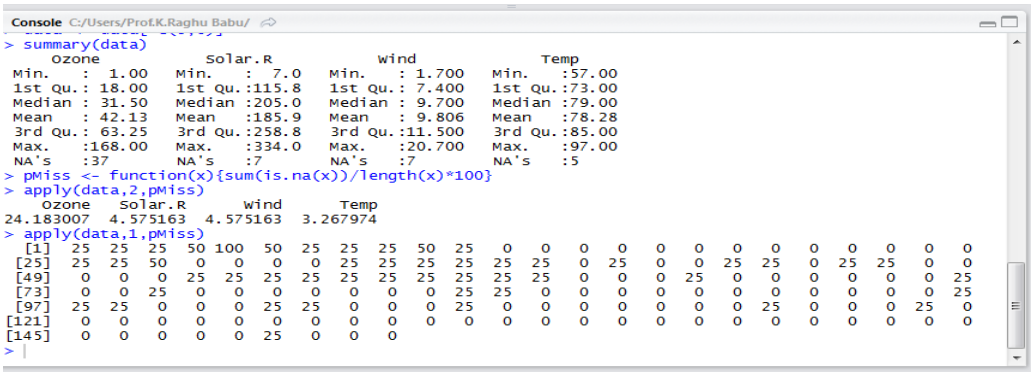



**Fig4. 2: Summary of the data**

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

**#Ozone is missing almost 25% of the data points**

```
> install.packages("mice")
Installing package into 'C:/Users/Prof.K.Raghu Babu/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.3/mice_2.30.zip'
Content type 'application/zip' length 1446459 bytes (1.4 MB)
downloaded 1.4 MB

package 'mice' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Prof.K.Raghu Babu\AppData\Local\Temp\RtmpKKmyYX\downloaded_packages
> library(mice)
Warning message:
package 'mice' was built under R version 3.3.3
> md.pattern(data)
    Temp Solar.R Wind Ozone
104    1       1    1     1  0
 34    1       1    1     0  1
  4    1       0    1     1  1
  3    1       1    0     1  1
  3    0       1    1     1  1
  1    1       0    1     0  2
  1    1       1    0     0  2
  1    1       0    0     1  2
  1    0       1    0     1  2
  1    0       0    0     0  4
       5       7    7    37 56
>
```

Fig4.3. Installing MICE

- 104 samples are complete,
- 34 samples miss only the Ozone measurement,
- 4 samples miss only the Solar.R value.
- 3 samples miss only the Wind value.
- 3 samples miss only the Temp value and so
- on

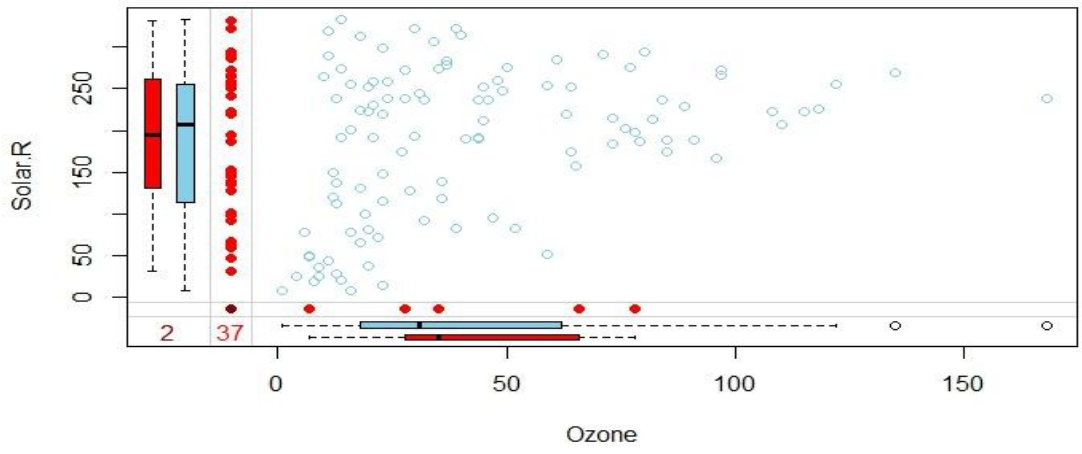



**Fig 7.8. Box Plot**

Almost 70% of the samples are not missing any information, 22% are missing the Ozone value, and the remaining ones show other missing patterns.

# The red box plot on the left shows the distribution of Solar. R with Ozone missing

#while the blue box plot shows the distribution of the remaining data points.

#Likewise for the Ozone box plots at the bottom of the graph.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

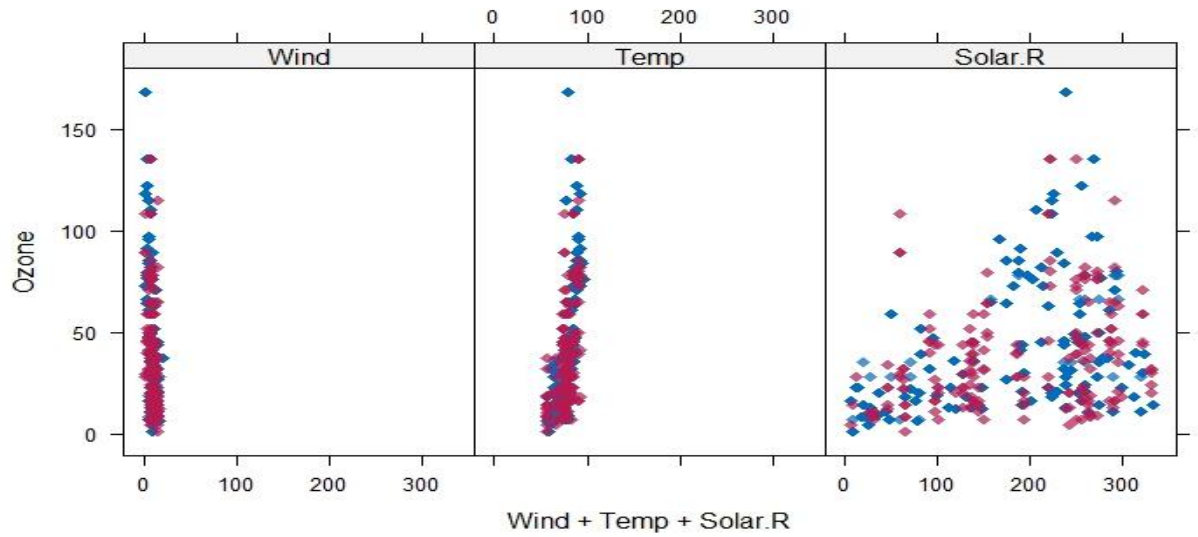



Fig 7.14(a): Scatter plot

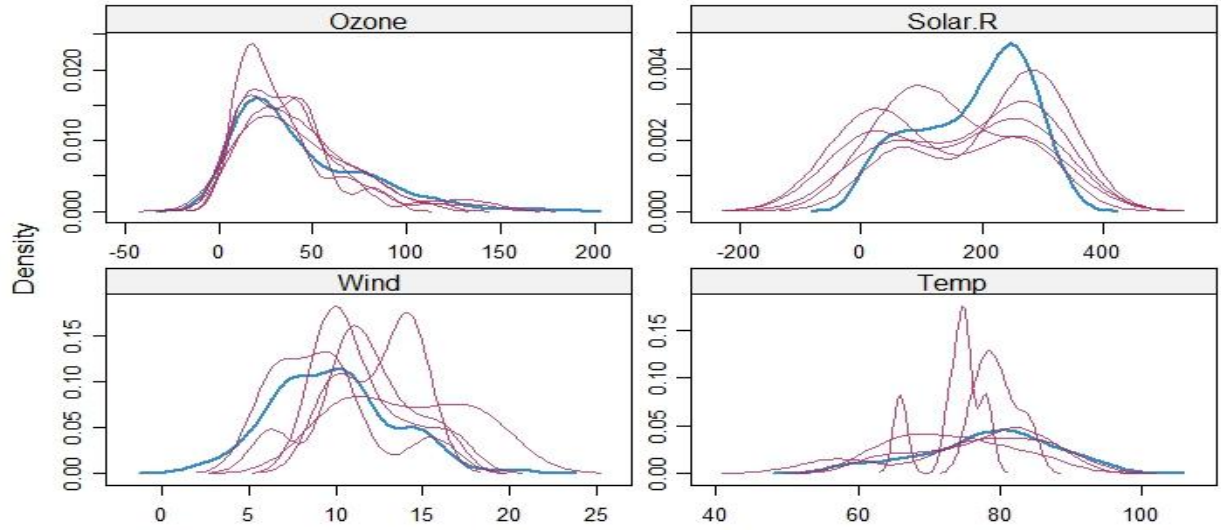



**Fig 7.15(a) Density plot**

## V. CONCLUSION

In this paper, Analysis of missing data using MICE (Multivariate Imputation by Chained Equations) in R has been implemented and also compared with linear regression model. This method helps in reconstructing incomplete data set and produces a complete data set. Although the MICE approach is a good method for addressing missing data it has some limitations while applying data mining techniques with it. For example, clustering is not always automatically incorporated by MICE in R.

In Future, MICE (Multivariate Imputation by Chained Equations) can be applied to impute data with data mining techniques like clustering to better understand and address the limitations of MICE approach in R environment for data analysis.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

## REFERENCES

[1] Dinesh J. Prajapati ,Jagruti H. Prajapat, "Handling Missing Values: Application to University Data Set" .Issue 1, Vol.1 (August-2011), ISSN 2249-6149.

[2] Luai Al Shalabi, Mohannad Najjar and Ahmad Al Kayed, A framework to Deal with Missing Data in Data Sets . Journal of Computer Science 2 (9): 740-745, 2006 ISSN 1549-363.

[3] Bhavik Doshi, Handling Missing Values in Data Mining. Data Cleaning and Preparation Term Paper.

[4] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" Third edition .

[5] De Jonge, Edwin, and Mark van der Loo. An introduction to data cleaning with R Technical Report 201313, Statistics Netherlands, 2013. URL http://www.cbs.nl/nl L/menu/methoden/onderzoekmethoden/discussionpapers/archief/2013/default.htm, 2013.

[6] D.T. Larose and C.D. Larose, "Imputation of missing Data", wiely Online Library, 2014. DOI:10.1002/9781118874059.ch13.

[7] Shelke, M.B., & Badade, K. B "Processing of incomplete Data Set" IJCER 2, no. 5 (2013): 658660.

[8] Schafer JL (1997), Analysis of Incomplete Multivariate data. Chapman & Hall, London.

[9] Van Buuren S, Oudshoorn (k) 2000 "Multivariate Imputation by Chained Equations: MICE VI.0 usris manual, Volume pg/vgz/00.038.

[10] Van Buuren S, Groothuis-Oudshoorn K (2011) "MIltivariate Imputation by Chained Equations" R Package version 2.9.

[11] Purwar, Archana, and Sandeep Kumar Singh. "Hybrid prediction model with missing value imputation for medical data." *Expert Systems with Applications* 42.13 (2015): 5621-5631.

[12] Horton, Nicholas J., and Ken P. Kleinman. "Statistical Computing and Graphics-Statistical Computing Software Reviews-Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *American Statistician* 61.1 (2007): 79.