Masters by Coursework and Research Report

Mathematical Statistics

School of Statistics and Actuarial Science

Title: Categorical Data Imputation Using Non-Parametric or Semi-Parametric Imputation Methods

Names:	Floyd Vukosi Khosa
Student number	718043 (Part Time 2013)
Email	vukosikhosa@yahoo.com
Supervisor	Y. Chhana

A research report submitted to the Faculty of Science, University of the Witwatersrand, for the degree of Master of Science by Coursework and Research Report.

Declaration

I declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

_____

(Signature of candidate)

_____

(Date)

Abstract

Researchers and data analysts often encounter a problem when analysing data with missing values. Methods for imputing continuous data are well developed in the literature. However, methods for imputing categorical data are not well established. This research report focuses on categorical data imputation using non-parametric and semi-parametric methods. The aims of the study are to compare different imputation methods for categorical data and to assess the quality of the imputation. Three imputation methods are compared namely; multiple imputation, hot deck imputation and random forest imputation. Missing data are created on a complete data set using the missing completely at random mechanism. The imputed data sets are compared with the original complete data set, and the imputed values which are the same as the values in the original data set are counted. The analysis revealed that the hot deck imputation method is more precise, compared to random forest and multiple imputation methods. Logistic regression is fitted on the imputed data sets and the original data set and the resulting models are compared. The analysis shows that the multiple imputation method affects the model fit of the logistic regression negatively.

Keywords

Imputation. Categorical Data. Semi-Parametric. Multiple Imputation. Hot Deck Imputation. Multiple Hot Deck Imputation. Missing Data.

Acknowledgements

Firstly I express my sincerest gratitude to my supervisor, Mrs Yoko Chhana, who guided me throughout my research report. I thank her for being patient with me, her willingness to help and encouraging me to go the extra mile whenever I needed the push. I am grateful for all her inputs and advice. I would also like to thank Ms Elsabé Smit for her advice at the initial stage of my research report. Secondly I thank my parents, Mr Cecil Khosa and Mrs Asnath Khosa for funding my Masters studies, for their emotional support and encouragement to study further. And last but not least I appreciate the LORD God for giving me the strength to persevere to the end of this research report.

Table of Contents

List of Figures

List of Tables

Notation and Terminology

- BFHD- Back-Forward Hot Deck
- EM- Expectation Maximisation
- GHS- General Household Survey
- HDI- Hot Deck Imputation
- MAR- Missing At Random
- Maxpost- Maximum Posterior Distribution
- MCAR- Missing Completely At Random
- MCMC- Markov Chain Monte Carlo
- MHDI- Multiple Hot Deck Imputation
- MI- Multiple Imputation
- ML- maximum Likelihood
- MNAR- Missing Not At Random
- MS- Mean Square
- OLS- Ordinary Least Squares
- PMMI- Predictive Mean Matching Imputation
- Proppost- Posterior Distribution
- RF- Random Forest
- RFI- Random Forest Imputation
- RHD- Random Hot Deck
- RI- Regression Imputation
- SHD- Sequential Hot Deck
- SS- Sum of Squares
- Stats SA- Statistics South Africa

Chapter 1. Introduction

Researchers and data analysts frequently encounter the issue of non-response or missing data in data sets that they need to analyse. This is problematic as missing data often produces biased estimates and invalid conclusions (Wayman, 2003). Data imputation which is a method of filling in missing values is often performed so that standard statistical techniques can be used to analyse data. Data imputation makes use of observed auxiliary data to impute the cases with non-response thus preserving the precision of the estimates (Enders, 2010).

Methods for imputing numerical data are well established in the literature mainly by Rubin (1987), Schafer and Graham (2002), Allison (2000) and Little and Rubin (2002). Methods for imputing categorical data however are not well established. Most of the available methods for imputing categorical data involve the categorical variable being converted to dummy variables and the dummy variables are then imputed. This process however introduces bias (Allison, 2001; Xiao, Song, Chen and Hall, 2012; Cranmer and Gill, 2012). For example, when imputing a binary gender variable with possible outcomes 1 (male) and 2 (female), the imputation model could impute a value of 1.5 and rounding the value to a 1 or a 2 introduces bias. Frequently used techniques (for example, hot deck imputation and multiple imputation) for imputing missing data tend to fail when a categorical variable with a small number of categories has missing values. Conventional techniques which are dependent on assumptions of continuous distributions can yield unsound imputations, biased results and insignificant standard errors (Graham, 2009).

1.1 Aims of the Study

Most of the data imputation literature that exists concentrates on numerical data imputation and parametric assumptions. Little has been done when it comes to categorical data imputation, particularly with regard to non-parametric methods. This research report focuses on categorical data imputation using non-parametric or semi-parametric methods. The aims of this study are to compare the prediction accuracy of the different categorical imputation methods, outlining their assumptions and assessing the quality of the imputation.

Chapter 2. Literature Review

Data analysis is often problematic when the data contains missing values or non-response. Non-response means that the data required are missing for some respondents in the selected sample. A differentiation is made between item non-response and unit non-response. Unit non-response happens when it is not possible to interview or obtain data from certain sampled members, or when a sampled member does not want to participate in the survey (Durrant, 2005; Enders, 2010). Durrant (2005) states that instances where item non-response occurs include the following; when the interviewer fails to ask a question, does not document the response, the respondent refuses to respond to the question or the respondent does not know (or recall) the answer. Non-response can be classified into two types namely: univariate and multivariate (Durrant, 2005). Univariate implies that data is missing in one variable only (figure 1 (a)). Multivariate implies that data is missing in more than one variable (figure 1 (b)).

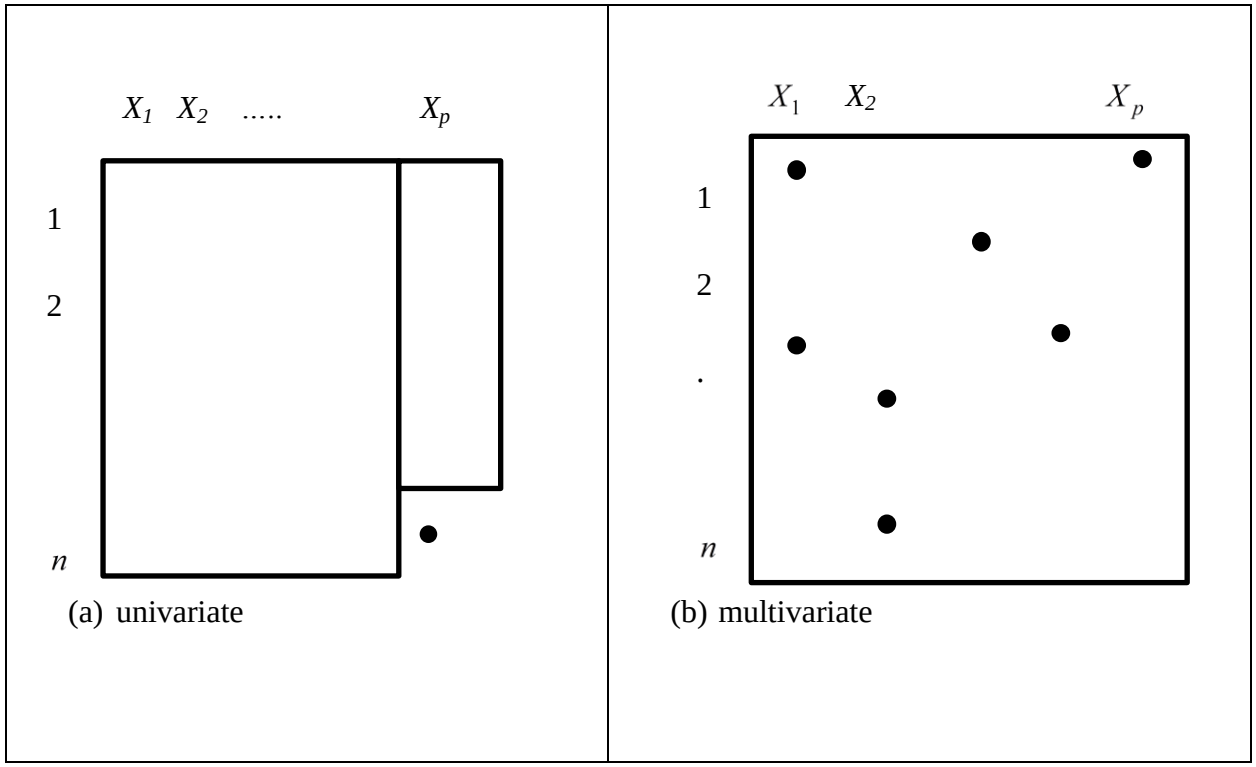

Figure 1. Non-response Patterns

Data may be missing because of various reasons, for example, data collection instruments failed, inclement weather made it difficult or impossible to collect data, people got sick or data were entered incorrectly. Researchers and data analysts often resort to ad-hoc approaches (e.g. mean imputation, which replaces missing data with the average of the variable with missing data, or listwise deletion which involves deleting the entire case with the missing values) to handle missing data. This may ultimately do more harm than good (Wayman, 2003), e.g. by increasing the variance for mean imputation and data loss for listwise deletion. The point of any data analysis is to draw inference about the population of interest. Missing values compromise this objective since the missing values could lead to a biased sample being obtained i.e. it could result in the sample being different to the population from which it has been drawn.

2.1 Missing Data Patterns

Missing data patterns describe the position of the missing values in a data set, however it does not give details for the reasons behind missing data. A multivariate data set has a monotone missing data pattern if variables in the data set can be sorted in an order such that, when an observation is missing for a certain variable then all subsequent observations are missing for that variable (Xiao *et al*, 2012). Mohd Jamil (2012) noted that the monotone missing data pattern is generally linked with longitudinal studies, where a respondent drops out and does not continue with the study. This occurs for example in an experimental study researching a new drug where members in the study do not continue with the study because they have a bad reaction to that drug (Enders, 2010).

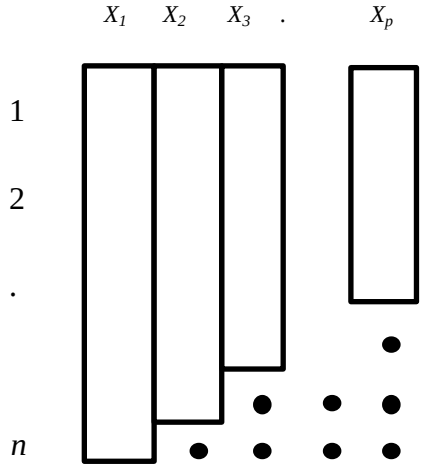

Figure 2. Schematic Representation of Monotone Missing Data

The monotone missing data pattern looks like a staircase (i.e. see figure 2). Methods that handle data with a monotone missing pattern include the propensity score method and discriminant function method. Allison (2001) states that a variable is called latent or unobserved if data are missing on a variable for all cases.

2.2 Mechanisms Responsible for Missing Data

Rubin (1987) formulated a grouping procedure to classify missing data which he named "missing data mechanisms". These mechanisms define the mathematical relationship between the observed variables and the possibility of missing data (Enders, 2010; Cranmer and Gill, 2012; Graham, 2009; Wayman, 2003; Rubin, 1987). Rubin (1987) states that missing data mechanisms are classified into one of the following groups namely; missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR).

2.2.1 Missing Completely At Random

In the MCAR mechanism, the missing data is not related to the collected data or other unobserved values, hence the missingness is not related to the variable itself or any other variable (Støvring, 2013; Allison, 2001; Allison, 2000; Hox, 1999; Rubin, 1987). The missing observations can therefore be thought of as randomly missing, meaning that the missingness is determined by chance alone (Støvring, 2013). The following example explains the MCAR mechanism. Consider a study where a patient in an experimental drug test drops out of the study. Missing data is MCAR if the cause of dropping out is not associated with other variables in the data. Hence, in the case where the patient dropped out of the drug test, it could be that the respondent died in a car accident and had not dropped out for causes related to other variables in the data set.

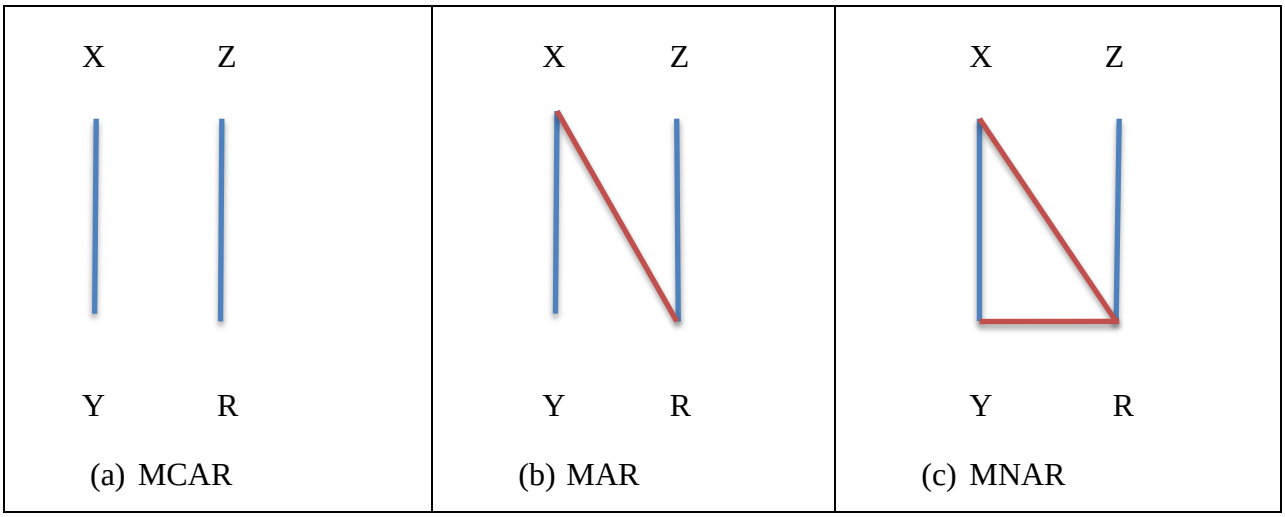

Figure 3. Missing Data Mechanisms

Figure 3 shows a graphical representation of (a) MCAR, (b) MAR, and (c) MNAR, in a univariate data pattern. X represents a complete variable (i.e. no missing values), Y represents a variable which is partially missing (i.e. contains both observed and missing values), Z represents the causes of missing data that is not related to X and Y and the variable R represents the missingness. The above figure is adopted from Schafer and Graham (2002).

2.2.2 Missing At Random

In the MAR mechanism the missingness depends on observed data and it is explained by the variables that are completely observed in the data set (Schafer and Graham, 2002; Rubin, 1976; Wayman, 2003). The missingness R is dependent on the variable X, see figure 3(b). For example, consider a variable X that indicates if the respondent is the primary caregiver or not and the variable Y indicating the amount of baby products used per week, household members that are not involved in looking after the baby may be less likely to respond to questions about baby products, hence the missingness in Y is dependent on X.

2.2.3 Missing Not At Random

In the MNAR mechanism the missing data is dependent on the unobserved data or the available data (i.e. missingness R may be dependent on variables Y or X, figure 3(c)). Generally when

the data is not MCAR or MAR it is considered to be MNAR (Wayman, 2003; Allison, 2000; Rubin, 1987; Rubin, 1976). An example of MNAR data, is income data which could be missing because it was not collected.

2.3 Data Imputation

Data imputation is a technique used to substitute missing values with credible values to create a complete data set. Data imputation is performed so that standard statistical methods can be used to analyse the data and to reduce non-response bias. Non-response bias arises when the distribution of missing data is different from the distribution of the observed data (Durrant, 2005). Imputation uses auxiliary variables that are statistically correlated to the variables in which item non-response occurs to fill in the missing value (Schafer, 1997). Auxiliary variables are variables that can be included in an analysis because they correlate the missingness of the incomplete variable with the complete variable (Enders, 2010; Graham, 2009). As a result the variables that are predictive of the missingness are used to improve the quality of the imputation. The primary purpose of auxiliary variables is to fine tune the missing data analysis by increasing statistical power and reducing non-response bias. For example, if the variable gender has a missing value on a respondent, then the gender variable which is fully observed on another respondent may be used to impute the missing value. Schafer (1997) and Durrant (2005) state that imputation methods are generally classified into two categories namely; deterministic and stochastic imputation.

1. Deterministic imputation is a method that produces the same imputed values each time for a given selected sample for units with the same attributes.
2. Stochastic imputation also known as random imputation is one that may produce different imputed values for a given selected sample.

Data imputation primarily aims to reduce non-response bias. Instead of deleting cases with item non-response, the use of imputation maintains the sample size resulting in higher efficiency. Imputation can have a bad effect if the imputed data is treated as observed data. Every so often distinct adjustment methods are needed to correct for the increase in the variance due to non-

response and imputation (Durrant, 2005). Durrant (2005) advised that the following attributes should be considered when choosing an imputation model:

1. The nature of data analysis that has to be conducted.
2. Whether there is a formula for estimating the variance.
3. The time it takes to compute and implement the procedure.

Durrant (2005) states that the conventional variance estimation methods are not suitable for imputed data because they may underestimate the variance. Different methods for calculating the variance of an estimator under imputation include model assisted approaches and replication methods such as the jack-knife variance estimator. Allison (2001) states that a good imputation method is one that minimises bias, maximises the use of available information and yields good estimates of uncertainty. The method must also yield accurate estimates of the *p*-values, confidence intervals and standard errors. Maximum likelihood and multiple imputation perform well in satisfying the above mentioned criteria and conventional methods often fail in one or more of these attributes.

2.4 Traditional Imputation Methods

There are a number of ways of dealing with non-response. Unit non-response can be addressed through weighting methods. Item non-response can be addressed by data imputation techniques such as; listwise deletion, pairwise deletion and model-based procedures for example maximum likelihood estimation (Durrant, 2005). Listwise deletion, also called complete case analysis, involves deleting the entire case with at least one missing value, analysing only the complete cases with available data on each variable (Cranmer and Gill, 2012; Enders, 2010). The disadvantage of this method is that it reduces the statistical power because the sample size is reduced and the sample might not be representative of the population resulting in the estimates being biased. Another deficiency of listwise deletion is that it needs the data to be MCAR. It can create biased estimates when the hypothesis is not met. Pairwise deletion or available case analysis only analyses the available cases in which the variables of interest are present (Enders, 2010). The advantage of this method is that it keeps as many cases as possible. The disadvantage of this method is that one cannot compare the analyses because the sample differs each time

since the cases used may differ for different variables and thus different samples. The difference between listwise and pairwise deletion is that, listwise deletion does not analyse cases containing missing values. Pairwise deletion also does not include cases with missing values, however, the cases with non-observed values are used when analysing other variables which do not contain missing values.

2.5 Methods for Imputing Numerical Data

Methods for imputing numerical data include mean imputation, regression imputation, mode imputation, nearest neighbour imputation and the propensity score method. Mean imputation fills in the missing data with the overall average for the variable with the missing values (Finch and Margraf, 2008). The regression imputation method uses a regression equation to estimate the missing values based on the available data (Durrant, 2005). An attractive property of regression imputation is its ability to handle both categorical and continuous variables. Two types of regression models exist namely; the discriminant method and ordinary least squares (OLS). The OLS method is utilised for continuous data, and the discriminant method is used for categorical data (Grannell and Murphy, 2011; Durrant, 2005). Mode imputation fills in the missing values with the mode of the available data (i.e. the value with the highest frequency in the variable being imputed). The nearest neighbour imputation method is a donor method, where a donor is chosen by identifying the closest variables that have similar attributes to the variables that have missing values (Grannell and Murphy, 2011). Nearest neighbour imputation uses a distance metric like the Mahalanobis distance function to find nearest neighbours. The propensity score method applies a methodology based on propensity scores and uses the approximate Bayesian bootstrap to impute missing values (Xiao *et al*, 2012). This method is the probable likelihood that a certain component of the data is missing. The missing values are imputed by sampling from instances that have comparable propensity scores to the missing observations.

2.6 Methods for Imputing Categorical Data

A range of methods for imputing categorical data exist in the literature. The following methods are considered in this research report; regression imputation, hot deck imputation, multiple imputation, affinity score imputation, multiple hot deck imputation, predictive mean matching,

Markov Chain Monte Carlo, Bayesian imputation, random forest imputation, and imputation using association rules.

2.6.1 Regression Imputation

Regression imputation (RI) requires the use of auxiliary variables with known values to predict unknown values. An imputation model that relates the dependent variable (Y_i) to the independent variables (X_i) is fitted (Durrant, 2005).

$$Y_i = \beta_0 + \beta_i X_i$$

The estimated values are used to fill in the missing data in Y_i. Multiple imputations are created by means of a regression technique of the imputation variable Y_i that is based on a collection of identified covariates X_i. "The imputations are generated via randomly drawn regression model parameters from the Bayesian posterior distribution based on the cases for which the imputation variable is observed. Each imputed value is the predicted value from these randomly drawn model parameters plus a randomly drawn error-term. The randomly drawn error-term is added to the imputations to prevent over-smoothing of the imputed data" (Grannell and Murphy, 2011, p4). The limitation of regression modelling is that it alters the structure of the distribution and the correlation between the variables being imputed. Another deficiency of this method is that it is sensitive to model specification. Little and Rubin (2002) state that the predictive capacity of the model might be poor if the regression model does not fit the data well.

2.6.2 Hot Deck Imputation

The hot deck imputation method uses information from similar respondents, from the current data set to impute missing values (i.e. missing values are filled in with observed values from the data set). The respondent offering the value is known as the donor and the respondent with the missing value is known as the recipient. The hot deck imputation technique is a non-parametric method as it avoids distributional assumptions. Andridge and Little (2010) state that even though the hot deck imputation method is often used in practice, theoretically the method is not well

developed compared to other imputation procedures. The advantage of hot deck imputation is that actual values from the existing data set are used for imputing missing values. This method requires a large sample to work well (Durrant, 2005).

A major problem with conventional hot decking is that it does not reflect uncertainty in the imputed values. All imputations are treated as real values instead of probabilistic values and there is no direct way to account for uncertainty. Another drawback of hot deck imputation is that it assumes perfect correlation between variables (Munguia and Armando, 2014). There are a range of hot deck imputation methods in the literature, namely; random hot deck (RHD), back-forward hot deck (BFHD), sequential hot deck (SHD), and nearest neighbour hot deck (NNHD). Two hot deck imputation procedures are considered in this research report namely; RHD and NNHD.

2.6.2.1 Random Hot Deck

Coutinho and de Waal (2012) recommend the following steps to implement the RHD imputation procedure. Create an ordered list of possible donors for every entry with a missing value. Potential donors are numbered 1 to n where n is the last donor, thereafter a random number table is used to select donors and put them on the list in an ordered fashion. Thereafter a simple random sample of donors is drawn without replacement, until all the potential donors have been selected and included on the list of the recipients. The missing value for a recipient is imputed with the first value on the donor list for which the corresponding value is observed. The donor value is checked to see whether it is within the range of the variable with a missing value being imputed, and if it is, the value is used to fill in the missing value. If not, then the second donor on the list is checked to see if it falls within the range of the variable with the missing value. This procedure is repeated until a suitable donor is found. In the case where a suitable donor value is not found (i.e. there is no value that falls within the range of the variable), the missing value is imputed with any value within the range of that variable that is the nearest to the first donor on the ordered list. The procedure is done for each variable with missing values. Cranmer and Gill (2012) state that the RHD imputation method was inspired by simple random sampling since it randomly selects donors from a list of potential donors and that it works best for categorical data,

because it is more likely that several respondents will have the same value for categorical variables (e.g. gender).

2.6.2.2 Nearest Neighbour Hot Deck

The NNHD imputation method uses a distance metric to find similar respondents (i.e. donors) in the data set, in which a variable of interest is observed. Some of the distance metrics commonly used in the literature include the Mahalanobis distance and Euclidean distance functions. The Euclidean distance metric is considered in this research report. The Euclidean distance is defined as

$$E(a,b) = \sqrt{\sum_{i \in D} \left(x_{ai} - x_{bi} \right)^2}$$

where "$E(a,b)$ is the distance between two cases a and b, x_{ai} and x_{bi} are the values of attributes i in both cases a and b respectively and D is a set of attributes with non-missing values in both cases" (Jönsson and Wohlin, 2006, p12). The above formula is used to calculate the distances to the various donors and in the process to create a pool of donors.

The donor with the smallest distance is used to fill in the missing value, provided that the donor value falls within the range of the variable being imputed. If the first donor value with the smallest distance is not in the range of the variable being imputed, the second donor with the smallest distance is used. This process is repeated until a suitable donor is found. If a suitable donor is not found, the missing value is imputed with any suitable value which is the closest to the first potential donor (Coutinho and de Waal, 2012). An alternative procedure can be used where the Euclidean distance is confined to columns in which the variable of interest is observed. When the nearest neighbours are found, then the missing value is imputed with the mode of the nearest neighbours (Cranmer and Gill, 2012).

2.6.3 Multiple Imputation

Multiple imputation (MI) is a statistical method for dealing with missing data (Rubin, 1987; Rubin, 1996; Cranmer and Gill, 2012; Durrant, 2005).

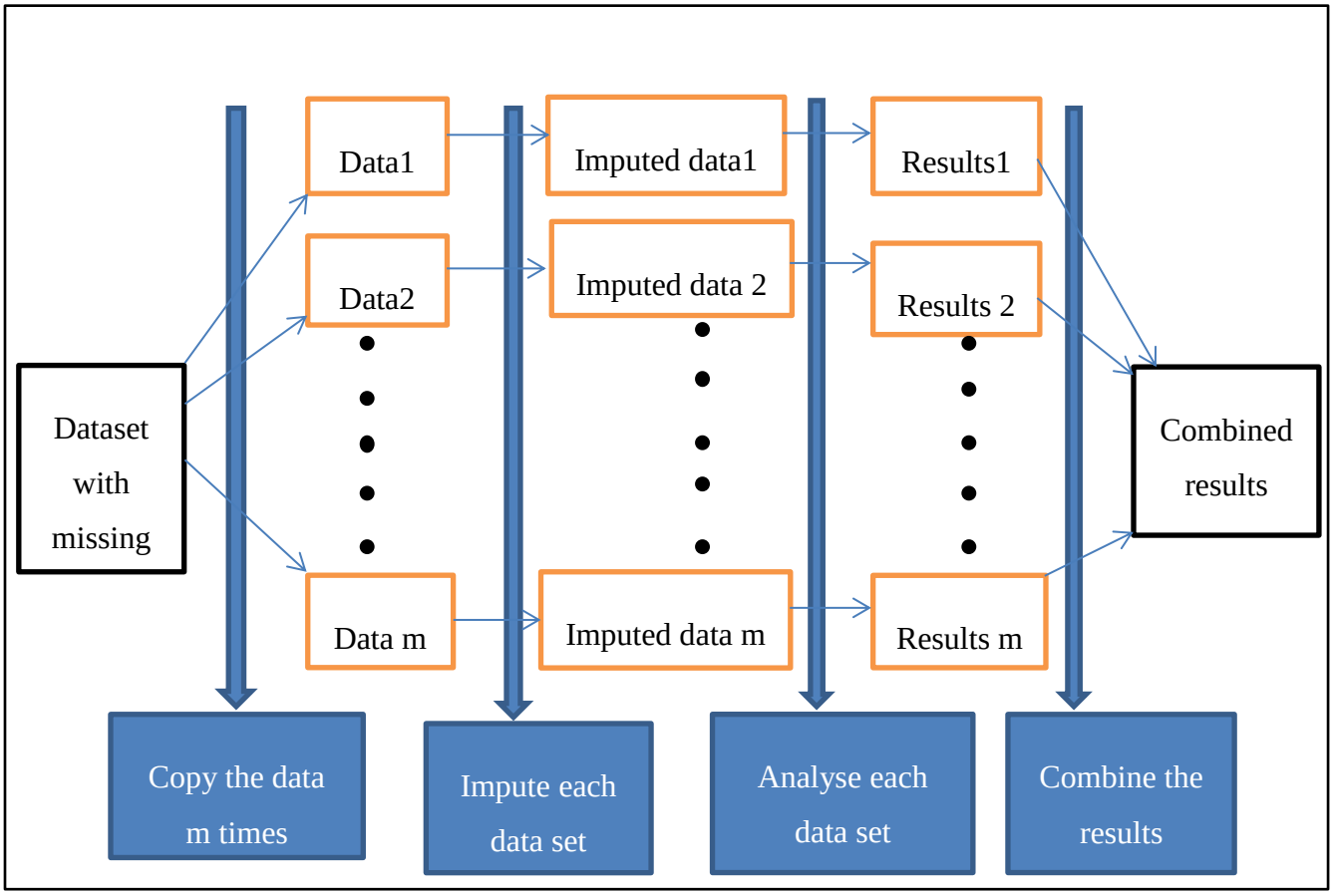

Figure 4. Multiple Imputation Process

Figure 4 above is adopted from Cranmer and Gill (2012). It is the schematic representation of the MI process. Firstly a data set with the missing values is copied several (m) times. Data imputation is then performed on each data set. Each of the data sets are analysed and the estimates of the m analyses are then combined.

When the missingness in the data set is very large (i.e. > 60%) more than 10 imputations may be required (Bodner, 2008). Allison (2001) states that the fraction of missing data refers to how much data are lost about each coefficient because of the unobserved data. This means that missing information is specific to each parameter of interest. Schafer (1997) recommends using a minimum of twenty imputations to decrease the sampling error due to data imputation. MI has two interesting characteristics; the capacity to perform different data analyses using existing

statistical methods on the complete data sets, and the detachment of the imputation step from the analysis step (Schafer 1997; Rubin 1987).

The imputation step imputes missing data and the analysis step provides inference about the multiple results (Shafer, 1997). The multiple imputation procedure makes a hypothesis about the probability model underlying a data set. The multivariate normal probability model is utilised for continuous data and the multinomial distribution model is commonly applied to categorical variables (Finch and Margraf, 2008; Rubin, 1996). When the probability model is selected, parametric estimations are made utilising the Bayesian posterior distribution centred on the following attributes; the observed data, the likelihood function of the recommended model and a prior distribution.

The MI process transforms the categorical data into continuous data. For many applications the imputed numerical values for those categories must be converted back to categorical values. This conversion introduces bias which negatively affects the analysis (Xiao *et al*, 2012; Cranmer and Gill, 2012). The solution to this variable conversion of categorical variables is to round continuous imputations to their nearest discrete value. Wayman (2003) states that MI was designed to maintain the variability in the population and to preserve the relationships amongst the variables. MI is an appealing method as an answer to missing data imputation since it characterises a good sense of balance, between the qualities of the results and because it is easy to implement. The estimates of the M analyses are combined into a single value using the following formulae:

$$\bar{Q} = \frac{\sum_{m=1}^{M} \hat{Q}_m}{M} \qquad (1)$$

where $\bar{Q}$ is the average value of the M data sets, and $\hat{Q}_m$ is the imputed value for the m^{th} data set. The variance of these estimates comprises of two parts namely; the between imputation variance denoted by B and the within imputation variance denoted which is the mean of the estimated variances across imputations by $\bar{U}$. The between imputation variance is given by the following formula

$$B = \frac{\sum_{m=1}^{M} \left(\hat{Q}_m - \bar{Q}\right)^2}{M-1} \tag{2}$$

The within imputation variance is the average of the projected variances through the M imputations. The total variance (T) is estimated using the formula

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B \tag{3}$$

2.6.4 Affinity Scores Imputation

Cranmer and Gill (2012) use affinity scores to find nearest neighbours or donors which are similar to the respondents with missing values, as opposed to the *k*-nearest neighbour approach which uses a distance metric to find nearest neighbours. Affinity is defined in terms of the degree to which each potential donor matches the recipient values, across all variables other than the one being imputed. In order to measure an extent to which a respondent with a missing value is similar to the respondent, a set of affinity scores bounded by zero and one are created. The affinity is denoted by $a_{i,j}$ which measures the degree of similarity that recipient i has to potential donor j.

$$a_{i,j} = \frac{k - q_i - z_{i,j}}{k - q_i} \tag{4}$$

Each respondent has a vector (y_i, x_i) where y_i indicates the outcome variable and x_i is a k-length vector of purely discrete exploratory variables, either of which may contain a missing value. The i^{th} case under consideration has q_i missing values in x_i. A potential vector x_j, $j \neq i$ will have between 0 and $k - q_i$ exact matches with recipient i. The variable $z_{i,j}$ is the number of variables for which a potential donor j and the recipient i differ in values. The formula $k - q_i - z_{i,j}$ then gives the number of variables for which j and i are perfectly matched. As the number of the matches decline, the affinity score moves towards zero. Imputation or estimation cells are a set of donors with the highest affinity scores. They are a subset of a sample to which

respondents are assigned based on characteristics (e.g. gender). If the imputation cell C contains the set of best possible donors (i.e. those with the highest affinity scores) random draws from the observed X_i in the cell will be unbiased and have the least amount of imputation variance. A random sample of these donors is selected from the imputation cell to fill in missing values.

2.6.5 Multiple Hot Deck Imputation

The Multiple Hot Deck Imputation (MHDI) method is a non-parametric imputation method. It is an alternative of the HDI method combined with repeated imputation and estimation methods (Cranmer and Gill, 2012). "Multiple hot deck avoids assumptions of normality, and it works best where traditional multiple imputation fails: with discrete data. Versions of this method have been shown to be unbiased and efficient" (Cranmer and Gill, 2012, p12). What differentiates this method from other forms of hot decking (e.g. RHDI) is that several values are used for a single missing observation, implying that several donors are used for a single recipient. MHDI takes imputation variance into account unlike the conventional hot decking that fails to account for the variance. According to Cranmer and Gill (2012) MHDI usually produce more accurate imputations than the parametric MI methods. The MHDI procedure is as follows;

1. Duplicate the data set multiple (m) times.
2. Search for missing values in each column of the data set.
    i. Once a missing value is found, a vector of affinity scores are calculated for the missing value (i.e. using equation 4).
    ii. Create the best imputation cell and draw a simple random sample from it to create a vector of imputations.
    iii. One of these values is selected randomly to fill in the missing value.
3. Step 2 is redone until all the missing values are replaced.
4. The desired analysis is implemented on each of the m data sets.
5. The m results are combined, using the MI combination procedure defined in section 2.6.3 (i.e. equations 1 to 3).

The MHDI method is suited to cases where there are relatively few variables but many observations; in this case there will be more potential donors (Cranmer and Gill, 2012). A

potential problem with MHDI is that the best imputation cell could contain only one donor thus rendering the multiple part of MHDI irrelevant.

2.6.6 Predictive Mean Matching Imputation

Durrant (2005) states that the predictive mean matching imputation (PMMI) method is a combination method since it combines the components of HDI, RI and nearest neighbour imputation (Schenker and Taylor 1996). PMMI is a semi-parametric imputation method since it uses an imputation model, however it is not fully dependent on the imputation model. Grannell and Murphy (2011) state that PMMI makes use of OLS regression to fill in missing values. For creating imputations, let the variable Y_i, be the variable with missing values and let X_i be a set of covariates. Let Y_{obs} be the non-missing cases for variable Y_i and Y_{mis} be the missing values in Y_i. X_{obs} is the observed values relating to Y_{obs}. The linear regression model regresses Y_{obs} on X_{obs} to attain the regression equation of this form:

$$\hat{Y}_i = \beta_0 + \beta_i X_i$$

The predicted values are used to identify similarities between respondents and donors, instead of using them for imputation. A pool of potential donors is then created and random donors are drawn to impute missing values (Grannell and Murphy, 2011). Another form of PMMI is the HDI within classes where the imputation classes are described based on a range or intervals of the predicted values from the imputation model. "Randomisation can be introduced by defining a set of values that are closest to the predicted value, and choosing one value out of that set at random for imputation" (Durrant, 2005, p12).

2.6.7 Markov Chain Monte Carlo

Allison (2001) and Xiao *et al* (2012) proposed an approach utilising the Markov Chain Monte Carlo (MCMC) method to deal with categorical data imputation. Each categorical variable is expressed by a group of dummy variables. Graham (2009) states that variables with at least two levels must be dummy coded. If a categorical variable has n levels then $n-1$ dummy variables must be generated to denote the categorical variable. During the imputation process the dummy

variables are treated as numerical variables. After the imputation process, the imputed values for the dummy variables are converted back to categorical values for the original categorical variable. Xiao *et al* (2012) state that the MCMC technique is not biased under appropriate assumptions but that the variable conversion introduces bias. The disadvantage of this imputation method is that it assumes that variables with missing data are normally distributed.

Allison (2001) states that the MCMC algorithm under the multivariate model assumes the following general form. Select a set of initial values to be estimated namely; the covariance matrix and the mean. These initial values are usually attained using the expectation maximisation (EM) algorithm. The EM algorithm is a computational tool for computing maximum likelihood (ML) estimates of the mean and the covariance matrix. For every missing data pattern, current parameter values are used to compute a linear regression of variables with missing data based on variables with complete data. The regression coefficients are used to generate predicted values of the missing data. This procedure is replicated until all the missing values have been filled.

2.6.8 Imputation Using Association Rules

Kaiser (2010) introduced association rules as a method of imputing categorical data. If there are three categorical variables namely A, B and C an association rule takes the general form: "If $A=1$ and $B=1$ then $C=1$ and $P=P(C=1 \mid A=1, B=1)$. The conditional probability P is called the confidence of the rule and $P(A=1, B=1, C=1)$ is called the support of the rule. The support can be used as a constraint of minimum count of the cases supporting the rule. The 'if part' of the rule is often called the antecedent and the 'then part' is called the consequent" (Kaiser, 2010, p111). Kaiser states that procedures for association rules are not able to handle missing data. To get the rules to work he proposed two options, the first is to get a complete data set by reducing the data, or by handling missing values as special values. This is done practically by replacing all the missing values with the value "MISSING". The second option is attractive because it is able to handle data sets with large amounts of missing values. The designed algorithm uses two data sets namely; the training data set and data for missing values imputation. The training data set contains data for generating association rules, data for missing value

imputation, and data with imputed values. The proposed method uses three variants to impute missing values.

2.6.8.1 The First Variant

To get association rules to work, the following procedure is followed; all the missing values on the training data set are filled with the special value ("MISSING"). Using the training data set, generate association rules. Once the list of association rules is created, remove the rules with the support lower than required, and rules with a consequent that has a combination of longer than 1. From the list of the rules, remove those rules with the consequent containing the special value. The confidence of the association rules is sorted in descending order. For each special value (i.e. "MISSING") in the training data set, pass through the list of association until a suitable rule is found, or until the end of the list of the rules is reached. A suitable rule is one that satisfies the following two conditions.

1. The value of the attribute that is being searched is contained in the consequent.
2. The antecedent corresponds to the values of other given case attributes.

If a suitable rule is found, then the missing data is imputed with the values in the consequent of the association rule.

If the association rules in the first variant, are not able to impute the missing value, or if a suitable rule is not found, then association rules are combined with the most common attribute value (for example, the mode of a variable). The first variant of the association rule is used, and if no suitable rule is found, then the most common attribute is used.

2.6.8.2 The Second Variant

The second variant of the algorithm is as follows: all the missing data are filled with a special value (i.e. "MISSING"). The training data set is used to create association rules. Remove association rules with support lower than required from the list of association rules, and those rules with a consequent with a combination longer than 1. Association rules with a consequent that contains the special value are also removed from the list. For every attribute, find the most

common value, excluding the special value "MISSING". The confidence of the rules is sorted in descending order. For each special value (i.e. "MISSING") in the training data set, pass through the list of associations until a suitable rule is found, or until the end of the list of rules is reached. A suitable rule is then one that has the following characteristics.

1. The consequent of the association rule contains the attribute of the value being searched.
2. The antecedent of the association rule corresponds to other given case attributes.

The missing value is filled with the consequent value if a suitable association rule is found. If a suitable rule is not found, then the missing value is filled in with the most common attribute value, provided the most frequent value is not the special value (i.e. "MISSING").

2.6.8.3 The Third Variant

This variant of the algorithm is used for improving the imputation accuracy of the missing values. It combines association rules and the most common attribute technique. The difference between this variant and the other two variants is that only association rules containing confidence higher than the relative frequency of the most common attributes are used. An acceptable rule is the one meeting the following requirements;

1. The consequent of the association rule must contain the value being searched.
2. The antecedent of the association rule must correspond to values in other given cases, and
3. The confidence of the rule must be higher than the relative frequency of the most common value, excluding the special value.

If the association rule is found, then the missing value is filled with the consequent of the rule, if not, the missing value is filled with the most common value. Kaiser. J (personal email communication, 03 July 2014, see Appendix B2) indicated that the major drawback of this approach is that it is not able to handle large data sets.

Wu, Song and Shen (2008) define association rules differently from Kaiser (2010). They define association rules as follows "Let $I = \{i_1, i_2, ...i_m\}$ be a set of attribute values, called items and let D be a set of database transactions where each transaction T is a set of items where $T \subseteq I$.

An association rule is an implication of the form $A \subseteq I, B \subseteq I$ and $A \cap B = \phi$. A is called the antecedent of the association rule, and B is called the consequent of the rule. The rule $A \Rightarrow B$ holds in the transaction set D can be described with the support and confidence" (Wu *et al*, 2008, p1158). That is

$$\text{Support}(A \Rightarrow B) = P(A \cup B)$$

$$\text{Confidence } (A \Rightarrow B) = P(B \mid A)$$

$$= \text{Support } (A \Rightarrow B) / \text{ Support } (A)$$

Support (A) is the percentage of transactions in D that contain A. The support and the confidence of the rule are used to measure the certainty of the association rule. Rules that fulfil the minimum threshold and confidence are described as "strong association rules". In contrast to the Kaiser (2010) approach of association rules, Bashir *et al* (2006) and Wu *et al* (2008) used association rules together with the *k*-nearest neighbour approach. The *k*-nearest neighbour is an algorithm that is used to identify the closest donors to the variable with a missing value. The technique scans the data set, and then creates association rules on a given support and confidence thresholds. Missing values are then imputed using association rules and in case where there is no existing relationship, the missing values are imputed using the k -nearest neighbour approach.

2.6.9 Bayesian Approach

This method proposed by Li (2009) uses two techniques for imputing missing values. The first technique imputes the missing value with the value having the estimated maximum probability, and the second technique uses a value that is selected with probability proportional to the estimated posterior distribution. Xiao *et al* (2012) state that the Bayesian approach is designed for data with categorical variables. The method is non-parametric, thus the prior knowledge about the distribution of the data is not required.

Let $c_1, ..., c_L$ be a portion of the sample space, then for any event x in the sample space

$$P(c_k \mid X) = \frac{P(c_k)P(X \mid c_k)}{\sum_{k=1}^{L} P(c_k)P(X \mid c_k)}, \qquad k = 1, ..., L$$

$P(c_k)$ is termed the prior probability then $P(c_k \mid x)$ is named the posterior probability. It is occasionally assumed that the attributes are conditionally independent of each other given the class value. Under this hypothesis $P(x \mid c_k)$ is given by

$$P(X \mid c_k) = \prod_{j=1}^{M-1} P(x_j \mid c_k)$$

where $M-1$ is the non-class attributes, $X_1, ..., X_{M-1}$. The classifier created in this form is termed a "naïve Bayes classifier" (Li, 2009). The proposed method for predicting missing values draws upon the inspiration of naïve Bayes of predicting class values. The Bayesian classifiers make use of the Bayes theorem which is given by

$$P(c_j \mid d) = \frac{P(d \mid c_j)P(c_j)}{P(d)}$$

where $P(c_j \mid d)$ is the probability of a variable d belonging to class c_j, $P(d \mid c_j)$ is the probability of generating instance d given class c_j. $P(c_j)$ is the frequency of c_j in a data set. $P(d)$ is the probability of the event d occurring. Simple Bayes method can estimate probabilities of multiple missing variables using observed values. The advantage of this algorithm is that it is effective for large data sets, both in space and time. This method is limited to categorical data estimation of non-missing values and by the conditional independence assumption (Li, 2009).

2.6.10 Random Forest Imputation

Breiman (2001, p6) defined a random forest as "a classifier consisting of a collection of tree-structured classifiers $\{h(X,\Theta_k), k=1,.....\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input X". This method creates regression trees from bootstrap, and the set of trees constructed constitute a forest hence the name random forest (RF). Each one of the trees vote and the vote is used to classify each of the instances from the trees, and the mode of the votes is used to impute the missing values (Munguia and Armando, 2014; Hapfelmeir and Ulm, 2014; Svetnik *et al*, 2003).

Stekhoven (2013) and Breimann (2001) state that the procedure uses a complete response variable to train the forest. The missing values are estimated using a RF model trained using the non-missing parts (y_{obs}) of the data set. For any random variable X_n with missing values at records ranging from 1 to m. The data set is split into the following four parts: the non-missing parts of variables X_n, the non-observed parts or missing values (y_{mis}) of variable X_n, the variable except X_n with observations and the variable excluding X_n with missing observations. Initiating the process requires predicting the missing value for the variable X using an imputation method for example using mean imputation or mode imputation. The variables X_n (i.e. n ranges from 1 to p) are ordered with regard to the amount of missing data in ascending order. For every variable X_n the missing values are imputed by fitting a RF with observed parts of the response variable and the observed parts of the predictor variables, then predicting the missing parts of the response variable by applying the trained RF to the missing parts of the predictor variables. The imputation process is replicated until all the missing values are imputed. Some of the advantages of the Random Forest Imputation (RFI) technique include the following: it does not assume normality, linearity and homoscedasticity. Furthermore, it is capable of dealing with mixed data types both continuous and categorical (Stekhoven and Bühlmann, 2012), and lastly it is robust to outliers and noise (Munguia and Armando, 2014).

2.7 Strength, Weakness and Suitability of the Methods

The strength of the RI, MI and RFI models is their ability to handle both categorical and continuous data. Logistic regression and MCMC are suited for categorical data imputation. MI however, can be problematic when applied to categorical data, because it is dependent on a continuous distribution and as a result it produces imputed values that are continuous (Durrant, 2005; Grannell and Murphy, 2011; Little and Rubin, 2002; Breiman, 2001). MI and MCMC require that the variables be normally distributed. The limitation of the RI model however, is that it alters the shape of the distribution. The strength of the HDI and MHDI models is that they preserve the integrity of the data, because actual values are used to fill in the missing values and as a result they do not impute impossible values (Allison, 2001; Schafer, 1997; Cranmer and Gill, 2012; Wayman, 2003; Xiao *et al*, 2012).

The HDI and MHDI are suited for situations where the sample size is large, because more donors will be available for the sample. MI however is able to provide sufficient results regardless of the sample size as opposed to HDI and MHDI (Cranmer and Gill, 2012; Durrant, 2005; Munguia and Armando, 2014). MI is efficient in cases where the amount of missing values is high, as opposed to association rules which are efficient where the variables with missing values are few (Kaiser, 2010). MCMC and MHDI are restricted to situations where the variables are few. MI and MHDI are mostly suited to cases where the data are MAR. MHDI is also suited to cases where the data are MCAR. PMMI is suited for the monotone missing data pattern as opposed to MCMC which is suited for non-monotone missing pattern (Durrant, 2005; Allison, 2001). MCMC and PMMI experience difficulty when applied to multivariate data (Allison, 2001). The strength of the RFI and the Bayesian approach methods is that they can be computed in parallel making them computationally efficient both in space and time (Li, 2009; Stekhoven and Bühlmann, 2012).

2.8 Summary

Chapter 3 is a review of data imputation literature. The chapter begins with reviewing different mechanisms for missing data, missing data patterns and some of the earlier methods used to impute data in the literature. The chapter briefly discusses methods for imputing numerical data

and discusses in greater detail methods for imputing categorical data since the study is aimed at imputing categorical data. The concluding section of the chapter review, the strength, weaknesses and highlighted the suitability of the methods for imputing categorical data.

Chapter 3. Methodology

3.1 Selected Imputation Methods

3.1.1 Multiple Imputation

The method of MI as outlined by Rubin (1987), Rubin (1996) and Durrant (2005) is as follows:

The data set with missing values is copied several times (e.g. $M = 5$). Thereafter data imputation is performed on each one of the multiple data sets. Each of the multiple data sets (i.e. M) are analysed, and the results of the M analyses are merged into a one value $(\bar{Q})$ using the following equation,

$$\bar{Q} = \frac{\sum_{m=1}^{M} \hat{Q}_m}{M}$$

where $\hat{Q}_m$ denotes the imputed value(s) for the m^{th} data set (i.e. $m = 1, \ldots, M$). The between imputation variance is computed using the equation,

$$B = \frac{\sum_{m=1}^{M} \left(\hat{Q}_m - \bar{Q}\right)^2}{M - 1}$$

And the within imputation variance $\bar{U}$ is the mean of the estimated variances across M imputations. The total variance is estimated using the following equation,

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

Yuan (2010) recommended that MI efficiency (r) be calculated after the m results have been combined. MI efficiency is calculated using the following formula,

$$r = \frac{\left(1 + m^{-1}\right) B}{\bar{U}}$$

and measures the increase in the variance due to missing data. When there are no missing observations the values of r and B are 0 (i.e. zero).

3.1.1.1 Assumptions of Multiple Imputation

MI assumes that data imputation may be performed under any missing data mechanism, and that the resulting estimates will be acceptable under that mechanism (Schafer, 1999).

3.1.2 Random Hot Deck Imputation

Coutinho and de Waal (2012) describe the process for the RHD imputation procedure as follows:

1. First create an ordered list (i.e. from 1 to n) of potential donors for each respondent with a missing value. Use a random number table to randomly select donors and put them on the list.
2. Select a simple random sample of donors without replacement using a random number table, until all the potential donors have been drawn and put on the list for the recipients.
3. Impute the missing value with the first value on the donor list, for which the corresponding value is not missing.
4. Examine the donor to see if it is within the range of the variable with a missing value being imputed (for example, the gender variable ranges between 1 and 2, 1 representing males and 2 representing females).
5. If the donor value is in the range, it is used to impute the missing value.
6. If not, the second donor on the list is checked, to see whether it falls within the range of the variable with the missing value.
7. Steps 4 to 6 are repeated until a suitable donor is found or until the last donor on the list is reached.

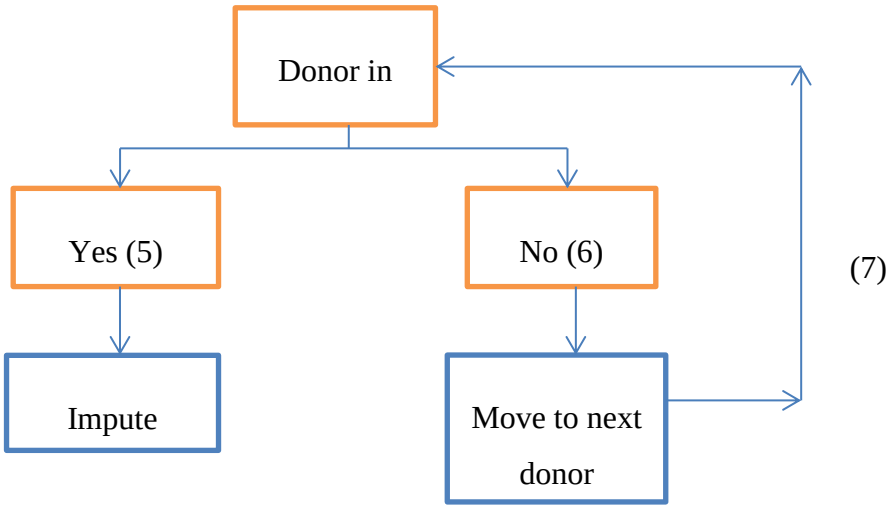

Figure 5. Donor Selection Process RHD

Figure 5 summarises the donor selection process for the RHD particularly steps 4 to 7. Consider the following example to explain the above mentioned process; the gender variable is imputed with either a 1 or 2, meaning that if the first donor value on the list is a 3 it is skipped until a 1 or a 2 is found on the list of donors. In the case where a suitable donor value is not found (i.e. there is no value that falls within the range of the variable), then the missing value is filled in with any value within the range of that variable that is the nearest to the first donor on the list. The procedure is done for each variable with missing values.

3.1.2.1 Assumptions of HDI

Andridge and Little (2010, p2) state that instead of using imputation based on a parametric model "hot deck makes implicit assumptions through the choice of metric to match donors to recipients, and the variables included in this metric, so it is far from assumptions free". They also mention that the hot deck procedure is not dependent on model fitting for the variable to be imputed, meaning that it is less responsive to model misspecification. The HDI procedure assumes that the model predicts values that are within the range of each variable (Cranmer and Gill, 2012).

3.1.2.2 Hot Deck Imputation code

The HDI procedure is executed in the "R" programming software using the "VIM" package. The following code is used to perform the HDI procedure.

$$\text{hot\_deck\_imp} = \text{hotdeck}(\text{data}, \text{impNA} = \text{TRUE}, \text{imp\_var} = \text{FALSE})$$

3.1.3 Random Forest Imputation

Stekhoven and Bühlmann (2012) used an interactive imputation algorithm based on RF. "The random forest algorithm has a built-in routine to handle missing values by weighting the frequency of the observed values in a variable with the random forest proximities after being trained on the initially mean imputed data set. However, this approach requires a complete response for training the forest" (Stekhoven and Bühlmann, 2012, p3). For a random variable X_n with missing values at entries ranging from 1 to m the data set is divided into four parts namely:

1. The observed values of variable X_n, are symbolised by y_{obs};
2. The missing values of the variable X_n, are indicated by y_{mis};
3. The variable other than X_n with observations ranging from 1 to m denoted by X_{obs};
4. The variable other than X_n with missing observations represented by X_{mis}.

To initiate the procedure, predict the missing value in X using an imputation method (e.g. mean imputation, mode imputation). Thereafter order the variables X_n, $n = 1, ..., p$ with regard to the volume of missing data from the smallest to the highest. For every variable X_n the missing values are filled in by initially fitting a RF with response y_{obs} and predictors X_{obs}, and thereafter predicting the missing value y_{mis} by applying the trained random forest to X_{mis}. The imputation process is repeated until all the missing values are imputed.

3.1.3.1 Random Forest Imputation Assumptions

RFI does not depend on distributional assumptions and is able to accommodate non-linear relations and interactions. The technique also assumes conditional normality and constant variance (Shah *et al*, 2014).

3.1.3.2 Random Forest Imputation Code

The random forest imputation technique is implemented using the "missForest" package. The method is implemented as follows; hundred regression trees are constructed, each of the 100 regression trees predicts a missing value in each variable. Each of the trees vote and the mode of the votes is used to fill in the missing values.

```
RF = missForest(data, maxiter = 10, ntree = 100, verbose = FALSE, replace = FALSE,
                classwt = NULL)
```

3.2 Reasons for Methods Selection

The proposed methods do not assume that the missingness is MAR and they work regardless of the mechanism of missing data. The methods are non-parametric and are designed for categorical data imputation. MI takes imputation variance into account, i.e. it has a variance estimation procedure (Durrant, 2005; Rubin, 1987). The reasons for selecting HDI include the following: it does not produce impossible values, i.e. values outside the range of the variables being imputed (Munguia and Armando, 2014). Furthermore, it uses information from similar respondents, and it works best for categorical data (Cranmer and Gill, 2012). The RFI model is selected as it does not assume linearity and normality and it is robust to outliers (Stekhoven and Bühlmann, 2012).

3.3 Binary Logistic Regression

A binary variable is one that contains two outcomes termed "success" and "failure", e.g. the variable gender consist of two outcomes male and female (Gandhi, 2011; Rodriguez, 2007). Logistic regression aims to quantify the relationship between the probability of dependent and independent (or exploratory) variables. The exploratory variables can be either discrete or

continuous or both. The linear regression model uses the following formula to predict an outcome (Gandhi, 2011).

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

where $E(Y)$ is the expected value of Y or the mean of the outcome of the variable Y. The binary variable of interest Y is usually dummy coded $Y = 1$ and $Y = 0$, with the outcome 1 representing success and 0 representing failure. In contrast to the equation of a linear regression model, logistic regression is given by the following equation,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

where π represents the probability of a success and $1-\pi$ represents the probability of failure. Therefore the logistic regression is the model of the probability of success as opposed to the linear regression which is a model of the mean (Scheaffer, 1999). The logistic regression model can be expressed as the log odds which can be calculated using the following two formulas (Gandhi, 2011). Log odds are the beta (β) values of the regression equation and can be written as

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k ,$$

or

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_k X_k}}$$

The logistic regression distribution is S-shaped as compared to the linear regression model which is a straight line.

3.3.1 Assumptions of Binary Logistic Regression

Binary logistic regression assumes that the dependent variable possesses a binomial distribution. Furthermore, the dependent variables need not be normally distributed. The model does not presume a linear correlation between the independent and dependent variables. Logistic regression uses maximum likelihood estimation to estimate parameters which rely on large samples (Gandhi, 2011).

3.3.2 How to Assess Model Fit

The logistic regression model uses deviance to test model fit, as opposed to linear regression which uses the coefficient of determination $\left(r^2\right)$ (Mood, 2010). There are two types of deviance statistics that are used to assess the model fit namely; null deviance and residual deviance. Null deviance measures how well the response variable is predicted by the model with only the intercept (Lillis, 2008). Residual deviance is used to assess the overall fit of the model. The null deviance value and the residual deviance values are chi-square tests with the associated degrees of freedom. The deviance (i.e. residual deviance) can be likened to the residual sum of squares in linear regression (Lillis, 2008). A smaller deviance implies a better fit of the logistic regression model. Some of the statistical tools used to assess model fit in logistic regression include the following; Chi-square goodness of fit tests, classification tables, and ROC curves. Residual deviance is used in this research report to assess the model.

3.4 One-Factor ANOVA

"One-way ANOVA examines equality of the population means for a quantitative outcome and a single categorical explanatory variable within any number of levels" (Seltman, 2007, p1). The one-way ANOVA model is also known as one factor ANOVA. The term one-way implies that there is a single explanatory variable having at least two levels, and only one level of treatment is applied at any time to a given subject. "For example, data collected on, say five instruments have one factor (instrument) at five levels. The ANOVA tests whether instruments have a significant effect on the results" (NIST/SEMATECH, 2012, p1). The one-way ANOVA make use of the F

statistic to test the null hypothesis. The F statistic is defined as the mean squares of the hypothesis divided by the mean squares of the error.

The F statistic is calculated using the following formula

$$F = \frac{MS_h}{MS_e}$$

Mean square (MS) denotes the quantity resulting from dividing the sum of squares by its degrees of freedom. The between group mean squares which is termed the hypothesis mean squares, is calculated using the formula,

$$MS_h = \frac{SS_h}{df_h}$$

where SS represents the sum of squares and df is the degrees of freedom. The within group mean squares, which is called the error mean square is given by the following equation

$$MS_e = \frac{SS_e}{df_e}$$

3.4.1 Assumptions of One-Factor ANOVA

Assumptions of one-way ANOVA include that the population from which samples are taken have to be normally distributed, the samples must be independent, and the variance of the samples must be equal.

3.5 Summary

In this chapter the three selected methods for imputing missing data MI, HDI and RFI were reviewed. The methodology and assumption for Binary Logistic regression and one factor ANOVA were also summarised.

Chapter 4. Data Analysis and Discussion

4.1 Data Set

The General Household Survey (GHS) 2013- (person file) data is used for the analysis. The data are downloadable from the Statistics South Africa (Stats SA) website. The original data set contains 169 variables. Since the study only focuses on categorical variables, all numerical variables are deleted. Four categorical variables are retained for the data analysis, namely: "gender", "age group", "health status" and "geography type". The new data set contains 374996 observations for the four selected variables.

Figure 6 below shows the response rate per category for each of the categorical variables. The "health status" variable contains five categories namely; poor, good, fair, do not know (DNK), and unspecified or no response (99). The variable "gender" contains two categories (or outcomes) namely; male and female. The variable "geography type" contains four categories namely; urban informal, urban formal, tribal areas, and rural informal. Finally, the "age group" variable contains sixteen categories ranging from zero to above seventy five.

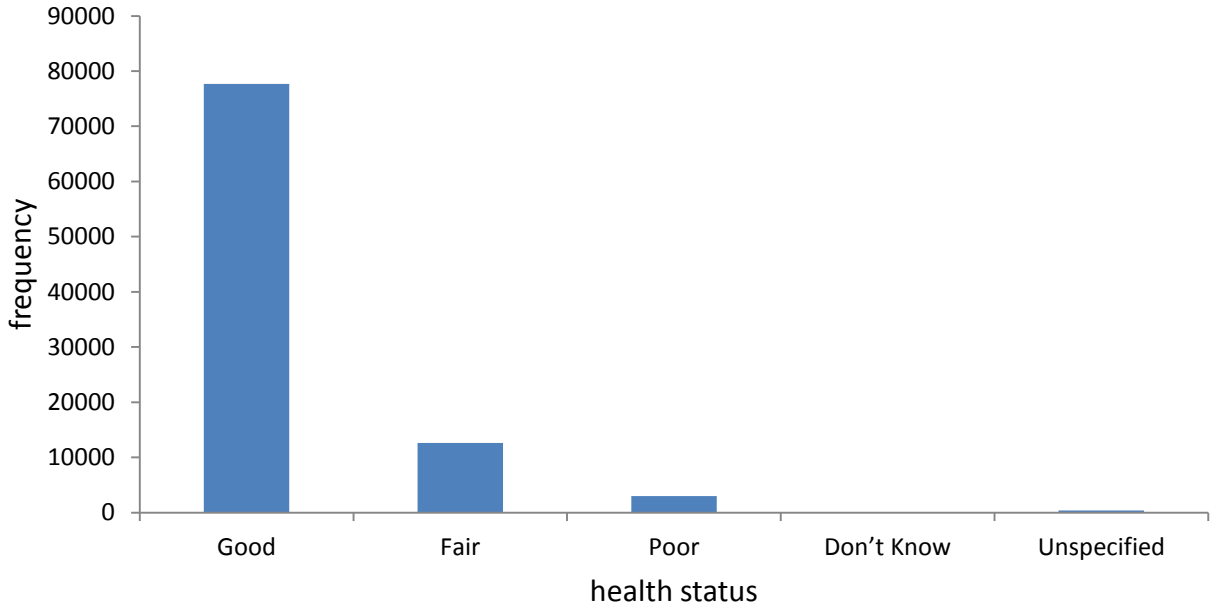

Figure 6. Health Status (Q22GENHEALTH)

Figure 6 shows that the category “good health” has the highest frequency followed by those with “fair” health. The health status variable is positively skewed.

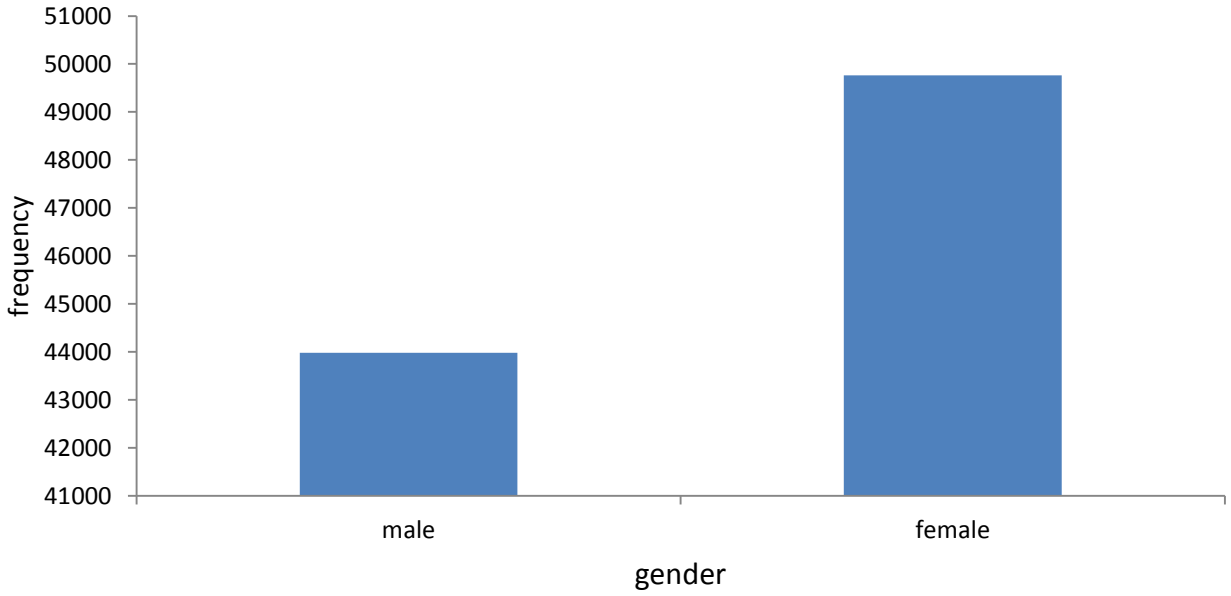

Figure 7. Gender

Figure 7 highlights that in the binary variable gender, category “female” has the highest frequency. The gender variable is negatively skewed.

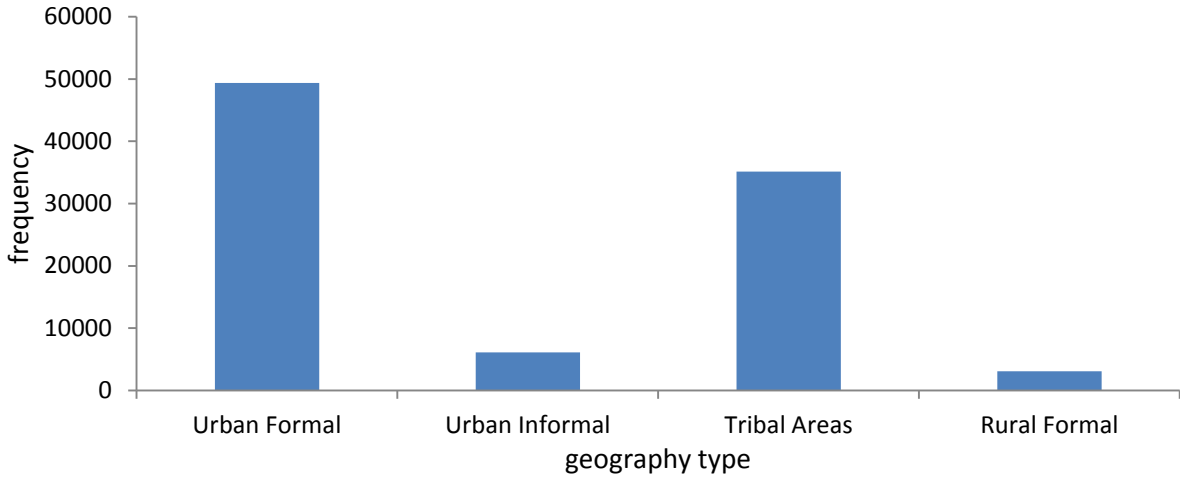

Figure 8. Geography Type (geotype)

Figure 8 illustrates that in the variable "geotype" the category "urban formal" has the highest frequency, followed by "tribal" areas, "urban formal" and lastly "rural formal".

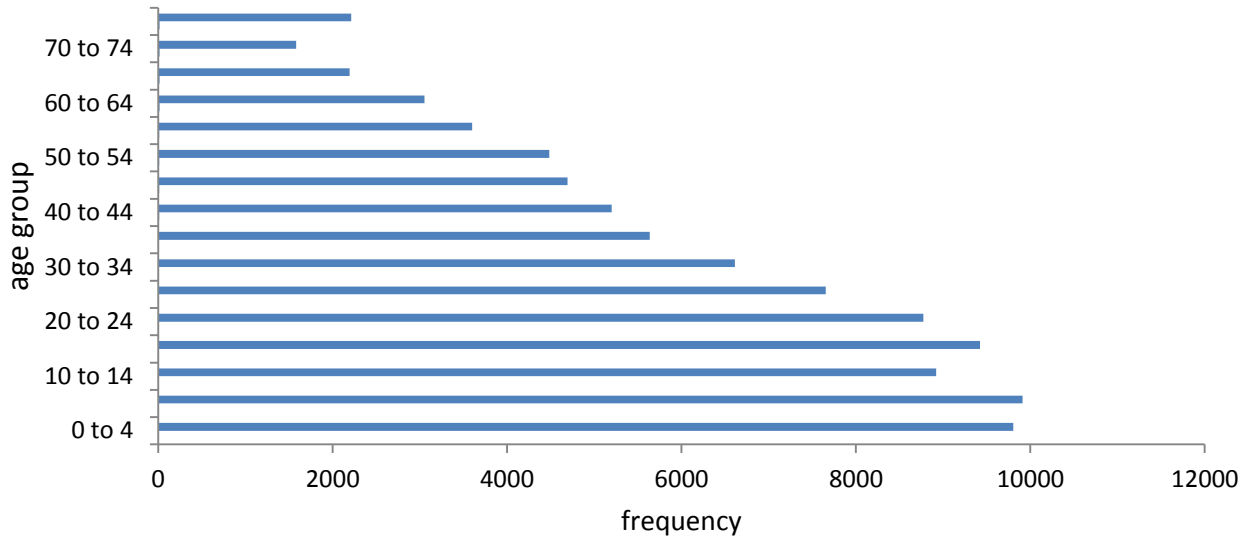

Figure 9. Age Group (Age_grp)

Figure 9 demonstrates that the age category "5 to 9" has the highest frequency followed by the category "0 to 4". The category "15" has the lowest frequency.

4.2 Software

The analysis is performed using the R version 3.1.1 statistical computing software, together with the Microsoft Office Excel (2010 version) software. The Amelia II version 1.7.2 software is used to implement the multiple imputation procedure.

4.3 Analysis procedure

1. Incomplete data sets with the required amounts of missing observations are generated using the MCAR mechanism from a complete data set. The following percentages of missingness are generated from the data set: 5, 15, 20, 30 and 40%. The percentages of missingness were selected to assess imputation with less than half of the data missing. The MCAR mechanism is implemented in R (see Appendix A1).

2. The proposed imputation methods are then used to impute the missing values that were previously removed (i.e. step 1 above).
3. The imputed data sets are compared with the original data set, and the percentages of values correctly imputed are counted (i.e. summed), divided by the number of missing values to get the predictive measure (see Appendix A2). This procedure was also used in the studies conducted by Kaiser (2010), Waljee *et al* (2013), and Li (2009) termed classification or misclassification error.
4. Classification error is also computed for each variable by comparing the imputed variable with the original variable (see Appendix A3). This is done to assess imputation at variable level as opposed to point "c" above that assesses imputation on the overall data set.
5. Binary logistic regression models are fitted to the original data set as well as the imputed data sets.
6. The binary logistic regression models are compared using one-factor ANOVA. The goodness of fit of the models is also assessed using deviance statistics.

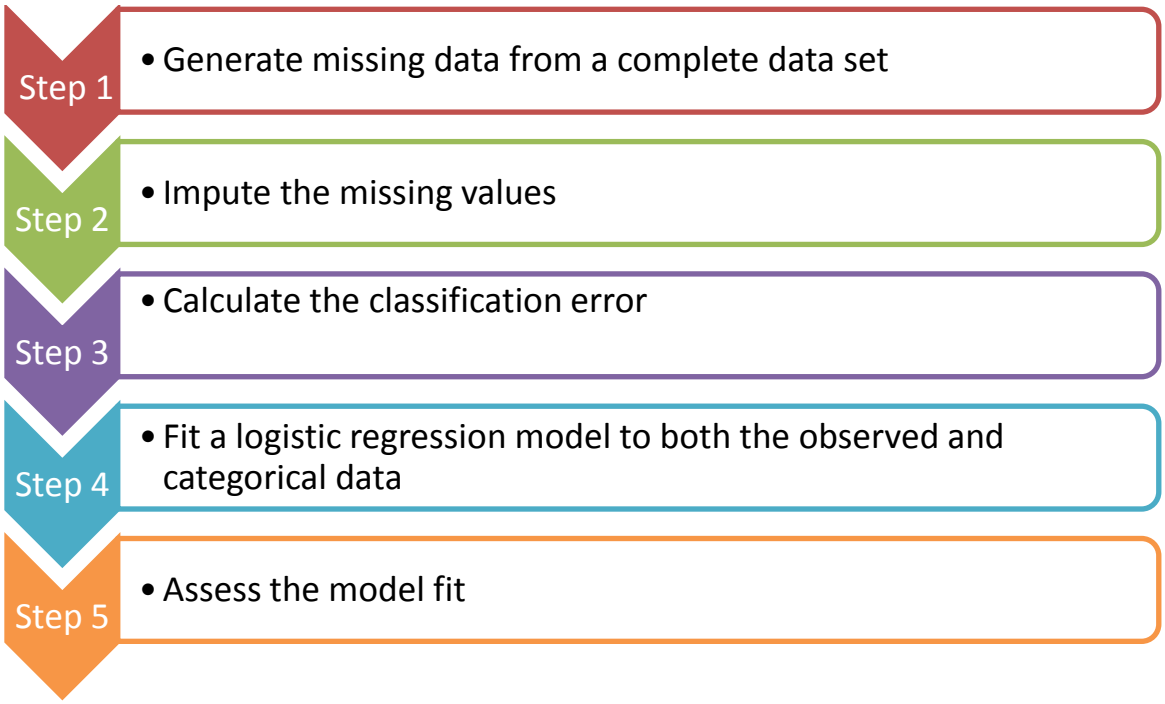

Figure 10. Analysis Procedure

Figure 10 summarises the step-by-step analysis procedure as described in section 4.3.

4.4 Imputation Analysis

4.4.1 Multiple Imputation

The Amelia II version 1.7.2 software is used to implement the multiple imputation procedure. The data set with missing values is copied five times (i.e. $m = 5$). The literature suggest that 5 to 10 imputations are adequate. Thereafter data imputation is performed on each of the five data sets using the Amelia II version 1.7.2 software that produces a graphical user interface in the R programming software. The imputation procedure resulted in five imputed data sets for each percentage of missingness (i.e. 5, 15, 20, 30 and 40%). A code is then created in "R" to compare the original data set with each of the imputed data sets. The model is used to calculate the number of imputed values that are the same as the values that were previously removed. This is done by superimposing the imputed data on the original data set, meaning that the imputed data set is literally placed on top of the original data set and the number of values imputed correctly is counted (i.e. if the imputed value is the same as the observed value).

Table 1 below illustrates the performance of MI on the overall data set under the different percentages of missingness.

Table 1: Multiple Imputation Results

% of missing values	**% of correctly imputed missing values**					
	Data 1	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Average**
5%	72.4	72.3	72.4	72.4	72.3	72.4
15%	45.1	44.3	45.4	45.0	45.2	45.0
20%	39.2	39.6	39.5	39.8	39.0	39.4
30%	34.0	33.7	32.9	33.1	33.1	33.4
40%	31.6	31.1	32.4	31.2	31.7	31.6

Table 1 shows the percentage of values correctly imputed on the overall data set using the MI method. Table 1 illustrates that MI performs best when the percentage of missing values is low (i.e. 5%), and it performs at its worst when the percentage of missing values is high (i.e. 40%). Table 1 show that an average of 72.4% of values is correctly imputed when 5% of the data is missing, whilst an average of 31.6% of values are correctly imputed when 40% of the data is missing. This means that 68.4% of the values are incorrectly imputed when 40% of the data is missing. When 5% of the data is missing 27.6% of observations are incorrectly imputed.

The above findings are in agreement with the study conducted by Kaiser (2010), who also found that the imputation accuracy (i.e. classification error) increases with the decrease in the percentage of missing values. Thus when the percentage of missing values is low the imputation model produces more imputed values that are the same as the observed values, than when the percentage of missingness is high. In a study conducted by Munguia and Armando (2014) it was found that when 5% of the values were missing, the imputation methods used produced more imputed values that were the same as the observed values. They also concluded that at 15% missingness there was no statistical difference between the imputed values and the observed values. The finding of Munguia and Armando (2014) and Waljee *et al* (2013) are also in agreement with the results presented in table 1.

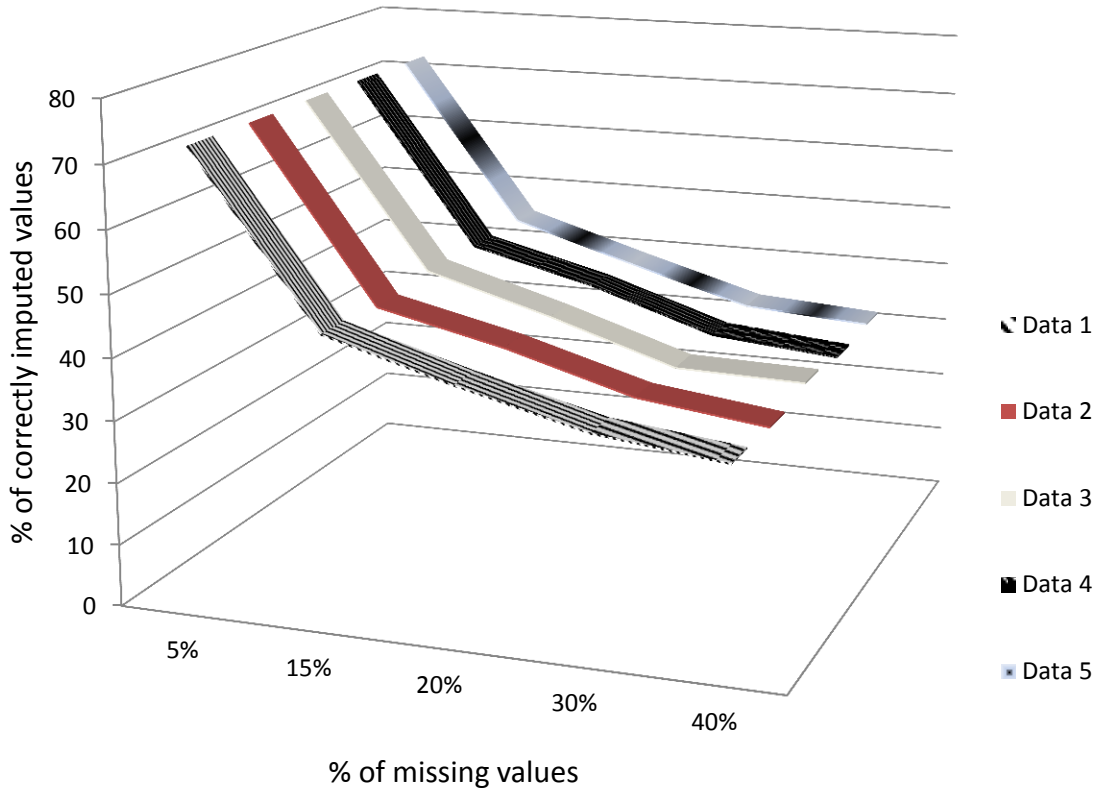

Figure 11. Distribution of the Multiple Imputed Data Sets

Figure 11 compares the distribution of the five imputed data sets (i.e. using MI), with regard to the percentage of values correctly imputed, and percentage of missing values. The figure shows that the distributions of the five data sets are similar. When 5% of the data is missing, in all five imputed data sets, 70% of the values are correctly imputed. The distribution for Data 3 is slightly different from the distribution of the other four data sets when the percentage of missingness increases from 30% to 40% (i.e. slightly elevated compared to the other four data sets which appear similar).

4.4.2 Hot Deck Imputation

The HDI procedure is executed in the "R" programming software using the "VIM" package. The following code is used to perform the HDI procedure.

```
hot_deck_imp = hotdeck(data,impNA = TRUE,imp_var = FALSE)
```

The HDI procedure results in the following table:

Table 2: Hot Deck Imputation Results

% of missing values	% of correctly imputed values
5%	77.4
15%	55.5
20%	50.5
30%	45.7
40%	44.2

Table 2 shows the percentage of observations that are correctly imputed for the different percentages of missing values. The table illustrates that 77.4% of observations are correctly imputed when 5% of the values are missing. The imputation accuracy (i.e. classification error) decreases to 55.5% when 15% of the data is missing, and eventually reaches 44.2% when 40% of the data is missing. The general trend of missing values versus percentage of correctly imputed values from table 2 is that the hot deck imputation procedure imputes more values correctly when the percentage of missingness is low.

4.4.3 Random Forest Imputation

The random forest imputation technique is performed in the R programming software using the "missForest" package. The method is implemented as follows; hundred regression trees are constructed, each of the 100 regression trees predicts a missing value in each variable. Each of the trees vote and the mode of the votes is used to fill in the missing values.

The code to implement the above mentioned procedure is as follows:

```
RF = missForest(data, maxiter = 10, ntree = 100, verbose = FALSE, replace = FALSE,
                classwt = NULL)
```

Table 3 shows the overall imputation accuracy when using the RFI method. The table shows that the RFI model correctly imputed 75% of missing values when 5% of the data is missing. When the missingness increases to 15%; the prediction accuracy decreases to 51.8%. The table also shows that when 40% of the observations are missing; 37.5% of the observations are correctly imputed. It is observed that 62.5% of the values are incorrectly imputed when 40% of the data is missing.

Table 3: Random Forest Imputation Results

% of missing values	**% of correctly imputed values**
5%	75.0
15%	51.8
20%	54.9
30%	41.0
40%	37.5

From the above results the general trend is that when the percentage of missingness increases, the imputation accuracy decreases. The RFI method predicted values outside the range of the variables (i.e. values lower than the minimum, and larger than the maximum for each variable), as a result the outcomes were rounded to the nearest whole number (e.g. 1.3 was rounded to 1, and 1.5 was rounded to 2). The RFI method is an example of a stochastic imputation method.

4.5 Comparison of the Imputation Methods

Table 4: Imputation Comparison

% of missingness	MI	HDI	RFI
5	72.4	77.4	75.0
15	45.0	55.5	51.8
20	39.4	50.5	54.9
30	33.4	45.7	41.0
40	31.6	44.2	37.5

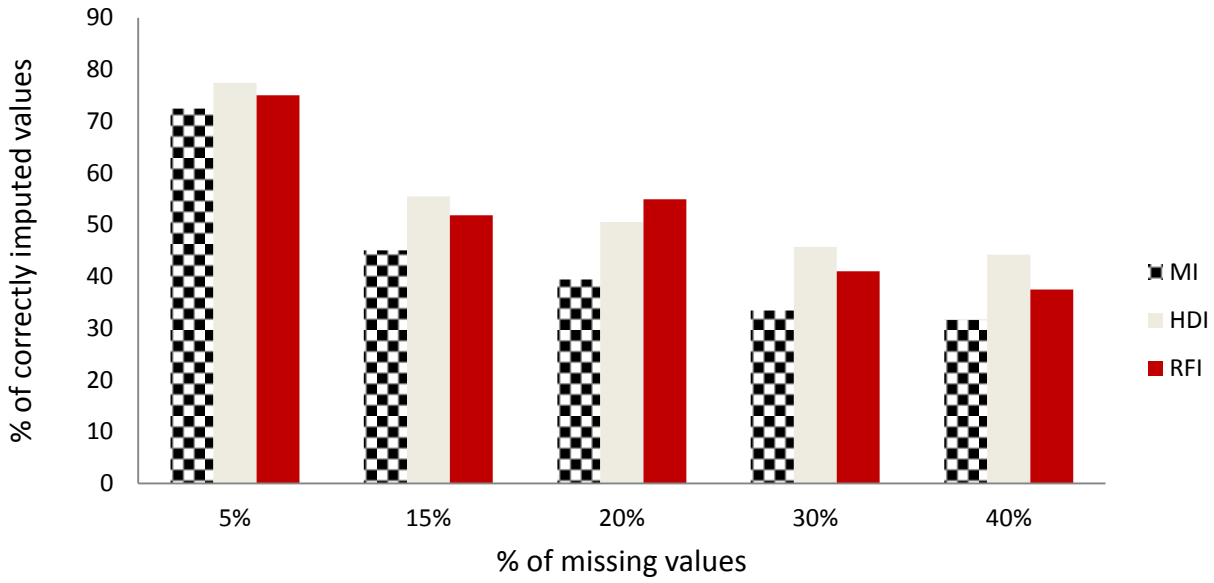

Figure 12. Imputation Comparison Overall Data

Table 4 and figure 12 describes the imputation accuracy of three imputation methods, MI, HDI and RFI respectively. From the table, it is evident that using the HDI method results in a higher percentage of imputed values which are closer to the original data (77.4%), followed by RFI (75%) and MI (72.4%) when 5% missingness is introduced in the data. Furthermore, it is noted

that the imputation accuracy is high when the percentage of missingness is low. The imputation accuracy decreases when the percentage of missingness increases. Figure 10 shows that the HDI method imputed a larger percentage of values correctly, compared to the MI and RFI methods under the various scenarios for the percentages of missingness except for the case of 20% missingness. At 20% missingness the RFI method imputed more values correctly (i.e. 54.9%) compared to HDI (50.5%) and MI (33.4%). HDI also predicts more accurate values (44.2%) when 40% of the data is missing compared to RFI (37.5%) and MI (31.6%).

4.6 Imputation Assessment at Variable Level

Tables 5 to 8 below show the performance of the imputation methods at variable level. A code is created in R which compares the imputed variable with the original variable. The number of values in the imputed variable that are the same as those in the original variable are counted (See Appendix B3).

Table 5: Similarity Gender Variable

% missingness	**Similarity percentage (%)**		
	MI	**HDI**	**RFI**
5	79.8	80.0	79.2
15	59.2	61.2	63.5
20	54.9	57.0	57.3
30	50.2	52.7	55.4
40	49.8	50.9	47.9

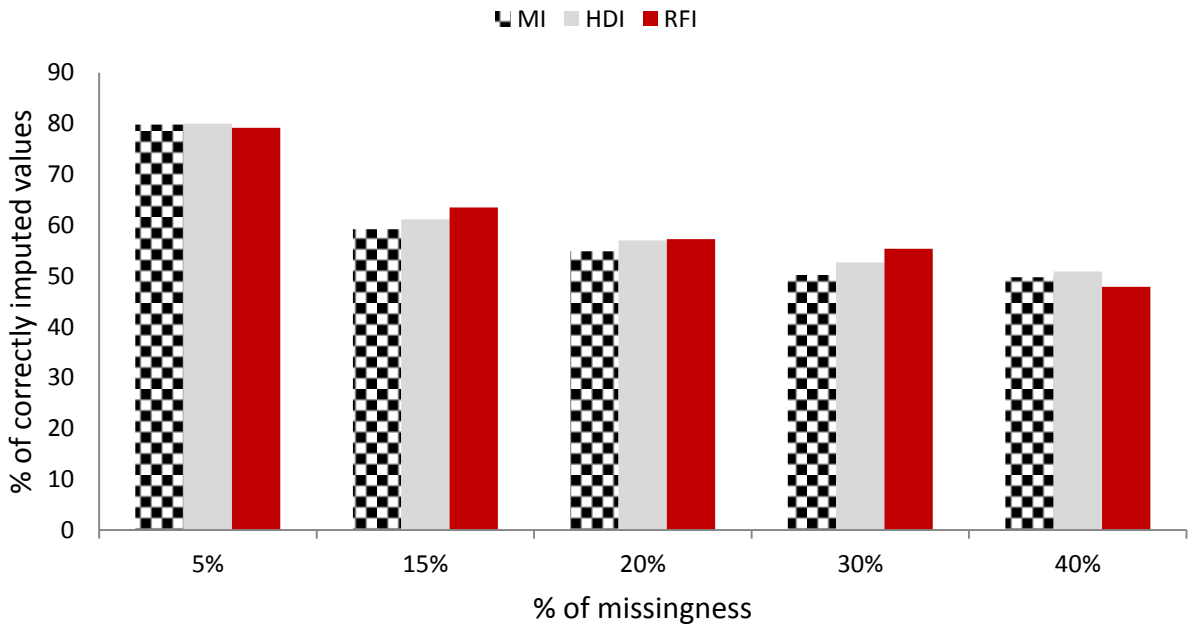

Figure 13. Gender Variable Comparison

Table 5 and figure 13 gives the similarity percentage (i.e. percentage of values correctly imputed) for the gender variable at the different percentages of missingness for the three imputation methods. The table shows that for the data containing 5% missingness, 80% of the missing data was correctly imputed when the HDI approach was used, followed by MI (79.8%) and RFI (79.8%). At 15% missingness, the RFI imputed more values correctly (i.e. 63.5%) followed by HDI (61.2%), and MI (59.2%).

The same trend follows when 20 and 30% of the data is missing (i.e. RFI, HDI, and MI). When 40% of the data is missing, the HDI imputes more values correctly (i.e. 50.9%), followed by MI (49.8%) and lastly RFI (47.9%). From the above results it is evident that when the percentage of missingness is low (i.e. 5%) the imputation method imputes more values that are the same as the original values. As the percentage of missingness increases, the percentage of similarity decreases for all three imputation methods.

Table 6: Similarity Age_grp Variable

% missingness	Similarity percentage (%)		
	MI	**HDI**	**RFI**
5	63.5	63.6	64.3
15	27.4	28.3	29.6
20	19.3	20.2	21.2
30	11.4	12.4	13.8
40	9.0	9.39	11.0

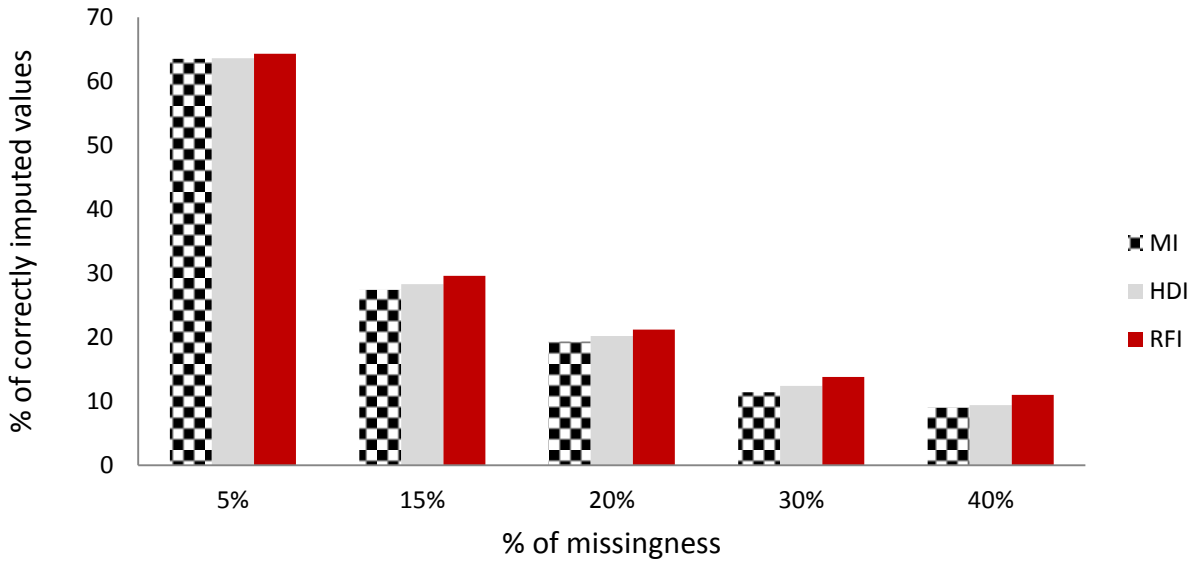

Figure 14. Age_grp Variable Comparison

Table 6 shows that an increase in the percentage of missingness results in a decrease in the percentage similarity for the Age_grp variable. When 5% of the data is missing the percentage of similarity for all three imputation methods is above 63%. Figure 14 shows that for each of the percentages of missingness, the RFI approach imputed more values correctly followed by HDI, and lastly by MI.

The figure shows that the three imputation methods worked well when imputing the Age_grp variable when 5% of the data is missing, but performed badly when the percentage of missingness increases from 15% to 40%. When the percentage of missingness is at its highest (i.e. 40%); the percentage of similarity is at its lowest (i.e. less than 12% for all three methods). From the above findings it is noted that the three imputation methods do not do well when imputing the Age_grp variable when the percentage of missingness is high (i.e. 40%).

The imputation models might have performed badly when imputing the Age_grp variable from 15% to 40% missingness because the variable Age_grp is skewed. The models might have performed better for 5% missingness because there is less variability in the data as compared to 40% missingness with a lot more imputed data value.

Table 7: Similarity Geotype Variable

% missingness	**Similarity percentage (%)**		
	MI	**HDI**	**RFI**
5	70.1	77.2	63.1
15	36.1	55.1	27.4
20	28.7	49.8	56.4
30	21.9	45.6	11.2
40	38.1	43.8	8.2

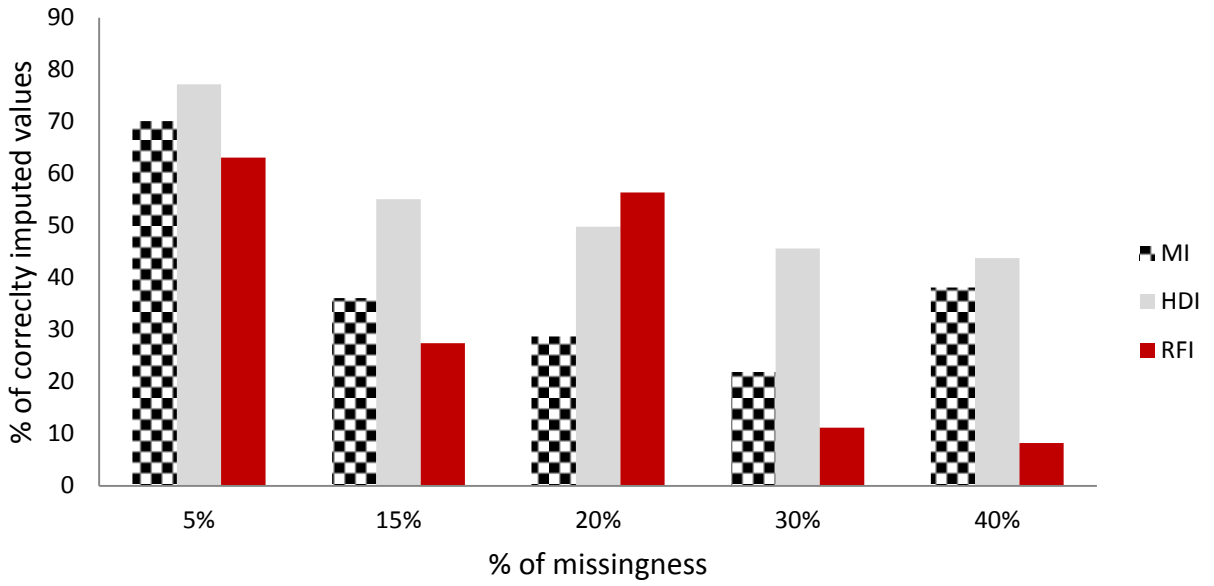

Figure 15. Geotype Variable Comparison

Table 7 shows that when imputing the geotype variable using the HDI method, the missingness decreases as the percentage of missing data increases. But when we look at the MI column we see that when the percentage of missingness increases from 5% to 30%; the similarity percentage decreased. When the percentage of missingness increases from 30% to 40% the similarity percentage increases from 21.9% to 38.1%. This might be as a result that the distribution of the correctly imputed values when the percentage of missingness increase follow the distribution of the original geotype variable. When looking at the RFI column we see that when the percentage of missingness increases, the percentage of similarity fluctuates between 8.2% and 63.1%.

Figure 15 shows that of the three imputation methods the HDI method imputed more values correctly for all the percentages of missingness, except at 20% missingness. RFI imputed the smallest percentage of values correctly when 40% of the data is missing, and imputed a largest proportion of the values correctly when the percentage of missingness is 5%. The MI method imputed more values correctly compared to RFI for 15, 30 and 40% missingness.

Table 8: Similarity Q22GENHEALTH Variable

% missingness	Similarity percentage (%)		
	MI	HDI	RFI
5	83.9	88.5	93.3
15	58.8	77.2	86.6
20	54.9	74.8	84.9
30	47.9	71.8	83.7
40	61.8	72.4	83.1

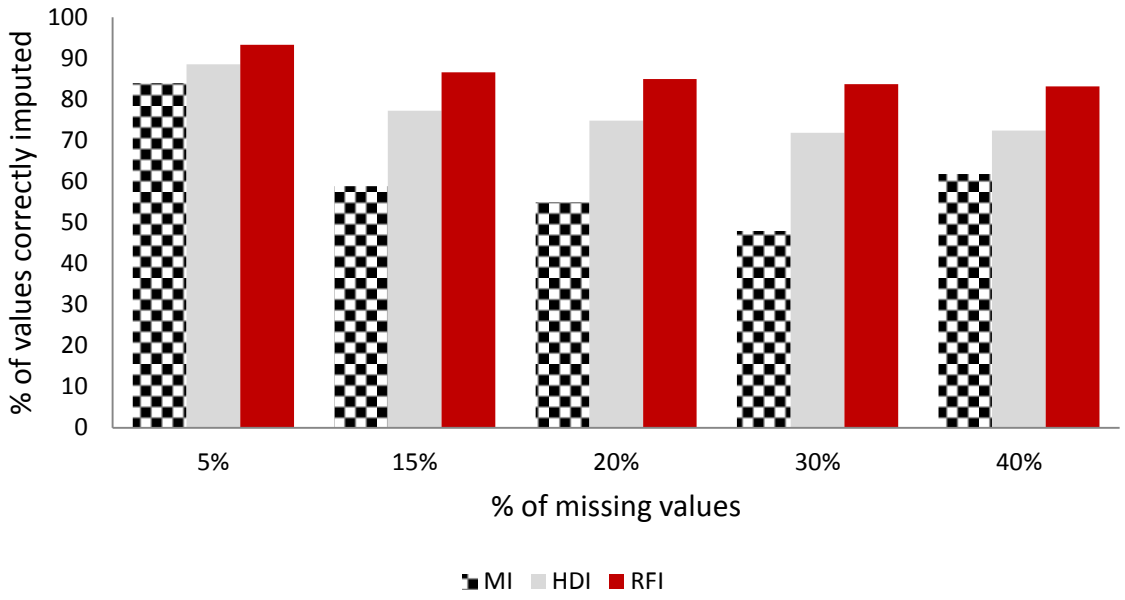

Figure 16. Q22GENHEALTH Variable Comparison

Table 8 shows that for the RFI method, with the increase in the percentage of missingness there is a decrease in the percentage of similarity. The MI and HDI columns show that when the percentage of missingness increases from 5% to 30% the percentage of similarity decreases. When the percentage of missingness increases from 30% to 40%; the percentage of similarity increases from 47.9% to 61.8% for MI. For the HDI when the percentage of missingness

increases from 30% to 40% the percentage of similarity increases from 71.8% to 72.4%. Figure 16 shows that the RFI has the highest percentage of similarity followed by HDI and lastly MI. The RFI method imputed over 80% of the values correctly for all the percentages of missingness. The smallest percentage of values imputed correctly occurs when MI is applied to the data with 30% of missingness.

4.7 Binary Logistic Regression of the Original Data Set

The following logistic regression model is fitted to both the original data set and the imputed data sets. The variable "Q22GENHEALTH" is used as a dependent variable. The following variables are used as independent variables; geotype, Age_grp and gender. Binary logistic regression requires that the variable to be predicted be a binary variable. Thus the variable "Q22GENHEALTH" is recoded into a binary categorical variable with 0 representing good health and 1 representing poor health. Category "good" is not changed and categories "fair", "poor", "don't know" and "unspecified" are combined to indicate poor health. The gender variable is already a binary variable with 1 representing males and 2 representing females, hence conversion is not necessary. The variable Age_grp is coded into a binary variable with the outcome 1 representing young people (i.e. 0 to 34 years) and outcome 2 representing old (i.e. 35 years and above) people. The geotype variable is re-coded to a binary variable with outcome 1 representing urban areas and 2 representing rural areas. See Appendix B1 for the variable conversion code, and see table 9 for the new codes.

Table 9. Binary Logistic Regression Codes

Q22GENHEALTH	$X_p = \begin{cases} 0, \text{good health} \\ 1, \text{poor health} \end{cases}$
Gender variable	$X_1 = \begin{cases} 1, \text{males} \\ 2, \text{females} \end{cases}$
Age_grp variable	$X_2 = \begin{cases} 1, \text{young} \\ 2, \text{old} \end{cases}$
Geotype variable	$X_3 = \begin{cases} 1, \text{urban areas} \\ 2, \text{rural areas} \end{cases}$

Table 10: Binary Logistic Regression Results Original Data

Coefficients	Estimate	Standard Error	Z value	$\Pr(\rangle\|Z\|)$
Intercept	-4.69422	0.05109	-91.887	<2e-16
Gender	0.24355	0.01866	13.055	<2e-16
Age_grp	1.75814	0.01900	92.556	<2e-16
geotype	0.09403	0.01887	4.982	6.29e-07

$$\text{logit}(\pi_i) = -4.69422 + 0.24355X_1 + 1.75814X_2 + 0.09403X_3$$

Table 10 and the logistic regression equation shows that for one unit change in gender the log odds of being healthy increases by 0.243. For a one unit increase in Age_grp the log odds of being healthy increases by 1.758. Lastly for one unit change in geotype the log odds of being healthy increases by 0.094. Furthermore, using the output from table 9 the following regression equation is fitted to predict "Q22GENHEALTH" based on the independent variables.

4.7.1 Assessment of Model Fit

Null deviance and residual deviance are used to evaluate the performance of the model. When we assess the model fit with only the intercept we have a chi-square (null deviance) value of 85850 with 93748 degrees of freedom. The null deviance value is small compared to the degrees of freedom indicating that the model fits the data well with only the intercept. When adding the variables; gender, Age_grp and geotype to the model, the deviance decreased by 9794 with 3 degrees of freedom. The residual deviance is then 76056 with 93745 degrees of freedom. The residual deviance value is small compared to the degrees of freedom indicating that the model fits the data well.

4.8 Binary Logistic Regression of the Imputed Data Sets

4.8.1 Multiple Imputed Data Set

Table 1 shows that MI performs best when 5% of the data is missing and worst when 40% of the data is missing. Hence the two extremes (i.e. 5% and 40% of missingness) are analysed for the comparison of the logistic regression models computed using the MI data.

Table 11: Binary Logistic Regression Results Multiple Imputation 5% Missing Data

Data set	intercept	gender	Age_grp	Geotype
1	-2.89677	0.16803	0.89943	0.12914
2	-2.842177	0.185384	0.835673	0.140894
3	-2.786976	0.184162	0.883792	0.156253
4	-2.785096	0.180601	0.774900	0.157514
5	-2.876363	0.177501	0.873815	0.138769

Table 11 summarises the logistic regression results computed from the five imputed data sets that were constructed when 5% of the observations are missing. Multiple imputation requires that the resulting estimates from the five imputed data sets be combined to form one estimate. The logistic regression estimates are combined using the following formula:

$$\bar{Q} = \frac{\sum_{m=1}^{M} \hat{Q}_m}{M}$$

$$\bar{Q}_{intercept} = \frac{-2.89677 + (-2.842177) + (-2.786976) + (-2.785096) + (-2.876363)}{5}$$

$$\bar{Q}_{intercept} = -2.83748$$

Similarly the averages of gender, age_grp and geotype yield the following estimates; 0.179136, 0.853522 and 0.144514 respectively. Therefore the logistic regression equation of the combined estimates is given by the following equation:

$$\text{logit}(\pi_i) = -2.83748 + 0.179136X_1 + 0.853522X_2 + 0.144514X_3$$

The between imputation variance (B) is calculated using the formula

$$B = \frac{\sum_{m=1}^{M}\left(\hat{Q}_m - \bar{Q}\right)^2}{M-1} = \frac{\left(-4.69422 + 2.89677\right)^2 + ... + \left(0.09403 + 0.144514\right)^2}{5-1}$$

$$B = 1.06813$$

The within imputation variance $(\bar{U})$, is the average of the estimated variances across M imputations which was calculated to be 1.06918. Finally the total variance (T) is given by the formula

$$T = \bar{U} + \left(1 + \frac{1}{M}\right)B$$

$$T = 1.06918 + \left(1 + \frac{1}{5}\right) \times 1.068132$$

$$T = 2.35093$$

and lastly we calculate the MI efficiency using the following formula

$$r = \frac{\left(1 + m^{-1}\right)B}{\bar{U}}$$

$$r = \frac{\left(1 + 5^{-1}\right) \times 1.06813}{1.60918}$$

$r = 0.79652$ this implies that the 5% missing values increased the variance by a factor of 0.79.

Table 12: Binary Logistic Regression Results Multiple Imputation 40% Missing Data

Data set	intercept	gender	age_grp	geotype
1	-1.112810	0.144466	0.129246	-0.002116
2	-1.114425	0.114425	0.119302	0.019367
3	-10.925392	-0.016241	0.139540	9.892087
4	-0.945860	0.030038	0.135835	0.021079
5	-1.030410	0.060366	0.128963	-0.003250

Table 12 describes the logistic regression results computed using the five imputed data sets when 40% observations were missing.

The averages of intercept, gender, age_grp and geotype yielded the following estimates, 3.02578, 0.066611, 0.130577 and 1.985433 respectively. Therefore the logistic regression equation is given by:

$$\text{logit}(\pi_i) = -3.02578 + 0.066611X_1 + 0.130577X_2 + 1.985433X_3$$

The between imputation variance is calculated to be 2.26034, the within imputation variance is calculated to be 10.06962 and the total variance is calculated to be 12.78202.

The value of r is computed to be 0.26936, which implies that imputation increased the variance by a factor of 0.26 when 40% of the data was missing which is much higher than the data with 5% missingness.

Table 13: Comparison of Variances

Variance type	5% missing data	40% missing data
Within imputation $(\overline{U})$	1.06818	10.06962
Between imputation (B)	1.06918	2.26034
Total variance (T)	2.35093	12.78202

Table 13 illustrates that the within imputation variance is low when 5% of the data is missing and high when 40% of the data is missing. The between imputation variance and the total variance follow the same trend.

Table 14: Combined Regression Coefficients of Multiple Imputation

% of missingness	estimate	gender	age_grp	geotype
5	-2.837	0.1791	0.8535	0.1445
15	-4.641	0.0998	0.2366	3.3918
20	-1.271	0.1134	0.2117	0.0205
30	-1.088	0.0718	0.2017	0.0097
40	-3.025	0.0666	0.1306	1.9854

Table 14 shows the combined logistic regression estimates under the different percentages of missingness after they were imputed using the MI method. The gender column shows a decrease in the estimates when the percentage of missingness increases except at 20% missingness. The Age_grp variable estimates decrease when the percentage of missingness increase. The estimates of the geotype variable fluctuate between 0 and 1.9 when the percentage of missingness increases from 5% to 40%. This might have been cause by the Age_grp variable is skewed and the increased variability in the data set caused by imputation.

4.8.2 Hot Deck Imputed Data Set

Table 15: Logistic Regression Estimates Hot Deck Imputation

% of missingness	intercept	gender	age_grp	geotype
5	-3.334982	0.165515	1.048680	0.004792
15	-2.243362	0.071501	0.398499	0.003721
20	-1.980241	0.035790	0.215255	0.009215
30	-1.635485	-0.006904	0.085334	-0.023397
40	-1.756340	0.012856	0.023965	0.004527

Table 15 gives the logistic regression estimates of the imputed data sets using the hot deck imputation method. When looking at the Age_grp column, the estimates decrease when the percentage of missingness increase. For the gender and Geotype column when the percentage of missingness increase from 5% to 30% the estimates decreases; however the estimates increases when the percentage of missingness increase to 40%.

4.8.3 Random Forest Imputed Data Set

Table 16: Logistic Regression Estimates Random Forest Imputation

% of missingness	intercept	gender	age_grp	geotype
5	-6.84182	0.6368	2.3072	0.4083
15	-7.69103	-0.5370	3.2233	1.0202
20	-7.91877	0.0598	3.5023	0.0678
30	-9.38676	-1.1546	4.2920	1.3968
40	-14.6925	1.8403	4.2839	1.6090

Table 16 gives the logistic regression estimates of the imputed data sets using RFI. For the Age_grp column, when the percentage of missingness increases from 5% to 30% the estimates of the variable increases. When the percentage of missingness increased to 40% the estimate decreases from 4.29 to 4.28. The estimates of the gender variable fluctuate between -1.15 and 1.84 when the percentage of missingness increases from 5% to 40%. Estimates of the gender variable fluctuate between 0.06 and 1.60 when the percentage of missingness increases from 5% to 40%. The negative signs in the gender variable for 15% and 30% might have been caused by the skewness of the imputed data set.

4.9 Comparison of the Imputed Models

The logistic regression models resulting from the imputed data sets using the three imputation methods are compared using one-way ANOVA under different percentages of missingness. The following code is used to compare the models.

anova(model1,model2, model3)

Model1 is the logistic regression model created using the MI data, model2 is the model resulting from the HDI data set, and lastly model3 is the model resulting from the RFI data set. One-way ANOVA is used to test whether there is a significant difference between the coeffients of the binary logistic regression models computed using the imputed data sets. The hypothesis testing procedure is as follows:

4.9.1 Hypothesis Statements

H_0 : There is no significant difference between the three logistic regression models.

H_1 : There is a significant difference between the three logistic regression models.

4.9.2 Test Statistic

The F statistical test is used to test for significant difference between the three models.

4.9.3 Decision Rule

Reject H_0 when the value of F is greater than the value of F_{crit}, or if the p-value is less than 0.05 (i.e. 5%) level of significance. The critical value is obtained from the F distribution table with 5% level of significance with 3 and 8 degrees of freedom (see Appendix B3).

Table 17: Analysis of Variance 5%

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	58.1623	3	19.38743	13.95019	0.001524	4.066181
Within Groups	11.11809	8	1.389762			
Total	69.2804	11				

Table 18: Analysis of Variance 15%

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	57.30726	3	19.10242	6.577037	0.014942	4.066181
Within Groups	23.23529	8	2.904412			
Total	80.54255	11				

Table 19: Analysis of Variance 20%

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	42.72115	3	14.24038	3.36288	0.075573	4.066181
Within Groups	33.87664	8	4.23458			
Total	76.59779	11				

Table 20: Analysis of Variance 30%

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	52.93786	3	17.64595	2.497696	0.133701	4.066181
Within Groups	56.51914	8	7.064893			
Total	109.457	11				

Table 21: Analysis of Variance 40%

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	103.9579	3	34.65264	2.084106	0.180743	4.066181
Within Groups	133.0168	8	16.6271			
Total	236.9747	11				

4.9.4 Conclusion

From table 17 and table 18 the value of F is greater than the F_{crit} value hence we reject H_0 at 5% level of significance and conclude that there is a significant difference between the three logistic regression models for the 5% and 15% missingness scenarios respectively. Furthermore, the p-values in table 17 and 18 are 0.001 and 0.014 respectively, hence there is strong evidence to favour H_1 (i.e. there is a significant difference between the three logistic regression models).

From table 19 to table 21 the value of F is less than the value of F_{crit} therefore we do not reject H_0 at the 5% level of significance and conclude that there is no significant difference between the three logistic regression models at 20, 30 and 40% missingness. Furthermore, this is corroborated by the p-values which are 0.075, 0.133 and 0.180 respectively, which means that there is weak or no evidence in favour of H_1.

4.10 Goodness of Fit Assessment for the Imputed Models

Table 22: Model Fit Imputed Data

% of missingness	Residual deviance			Degrees of freedom
	MI	HDI	RFI	
5%	99822	82346	49410	93745
15%	117320	85809	20574	93745
20%	116171	83974	19554	93745
30%	116612	86412	5285.7	93745
40%	118445	80805	2121.8	93745

Table 22 shows the residual deviance with corresponding degrees of freedom for the logistic regression models computed using the imputed data sets, under the different percentages of missingness. Residual deviance is used to assess the model fit for the overall models. Examining the values of residual deviance for MI at different percentages of missingness we note that the deviance values are more than the degrees of freedom.

This indicates that the models does not fit the data well for all for the different percentages of missingness. For HDI and RFI the residual deviance values are smaller than the degrees of freedom which indicate that the models fit the data well for the different percentages of missingness that were introduced in the data.

4.11 Example: Model Prediction

The following example was implemented to investigate if the logistic regression models created using the imputed data sets predicted the same outcomes with the model created using the original data set.

4.11.1 Scenario

Suppose that you want to predict the health status of an individual based on age, gender and geographical area. Assume that you want to predict the health status of an old male person living in an urban area. The prediction model computed using the original data set will be written in this form

$$\text{logit}\left(\text{Q22GENHEALTH}\right) = -4.69422 + 0.24355\left(\text{gender}\right) + 1.75814\left(\text{Age\_grp}\right) + 0.09403\left(\text{geotype}\right)$$

$$\text{logit}(\pi_i) = -4.69422 + 0.24355(1) + 1.75814(2) + 0.09403(1)$$

$$\pi = 0.453$$

If you round the value 0.453 to the nearest whole number the model predicts the outcome of "0" which implies "good health" for the above state person. Similarly when using the model computed using the MI method at 5% missingness the model can be written as follows:

$$\text{logit}\left(\pi_i\right) = -2.83748 + 0.179136X_1 + 0.853522X_2 + 0.144514X_3$$

$$\text{logit}\left(\pi\right) = -2.83748 + 0.179136(1) + 0.853522(2) + 0.144514(1)$$

$$\pi = 0.445$$

which is closer to 0 than 1. The two estimates differ by a factor of 0.008 which is not a big difference between the estimates. Similarly the models of HDI and RFI at 5% missingness predicts 0.355 and 0.306 respectively. All the models computed using the imputed data sets (i.e. HDI, MI and MI) are in agreement in that they all predict good health for the above state scenario.

4.12 Summary

Chapter 4 introduced the data set and software used. The imputation and analysis of the data under varying levels of missingness were explored. The chapter is concluded with an example that compares the predictions of the imputed data sets using the binary regression logistic models.

Chapter 5. Conclusion

The main focus of this research report was to review different non parametric and semi-parametric imputation methods for handling missingness in categorical data and to compare the performance of these methods. Three imputation methods were applied to the data and compared namely; MI, HDI and RFI. The three methods shared the following attributes; they did not assume normality and linearity. For all three methods, it is observed that when the percentage of missingness increased, the prediction accuracy decreased (i.e. percentage of missingness is inversely proportional to the prediction accuracy). When calculating the percentage of values correctly imputed; for the data with 5% missingness, all three methods correctly imputed more than 70% of the missing values correctly; with MI correctly imputing 72.4% of the values, HDI correctly imputing 77.4% of the data and RFI correctly imputing 75% of the values. When 40% of the data were missing MI imputed 36.1% of the values correctly, HDI imputed 44.2% of the values correctly and lastly RFI correctly imputed 37.5% of the observations correctly. It is observed that out of the three methods, HDI is the best imputation method for imputing categorical data, followed by RFI and MI respectively.

An in depth analysis which looked at imputation at the variable level revealed that although the RFI method predicted more values correctly than MI for the overall data set, MI predicted more values accurately at the variable level than RFI. At the variable level, HDI is still the best imputation method followed by MI and lastly RFI. It was also identified that for a variable with a small number of categories (e.g. gender variable which has two outcomes, male and female) the similarity between the observed and the imputed variable was high. For a variable with more than two categories the similarity between the imputed and the observed variables was low especially when the missingness was high (i.e. more than 15%). This might be as a result that HDI imputes values that are within the range of the variables being imputed, and in the case of MI and RFI they could produce values that are outside the range of the variables.

The observations that were outside the range of the variables needed to be rounded to the nearest whole number, and this might have affected the results for both the imputation accuracy and the estimates of the logistic regression. From the above findings it is noted that there is no perfect

imputation method, since they all have certain limitations. When assessing the model fit for the logistic regression models that were computed using the imputed data sets, the following was observed; the logistic regression model computed using the MI data did not fit the data for all the percentages of missingness. The models computed using the HDI and RFI data fitted the data well for all the percentages of missingness. From these findings it is noted that out of the three imputation methods, only the MI imputation method affected the model fit negatively when compared to the model fitted using the original data set.

5.1 Limitations and Recommendations

The findings of this research report are limited to the GHS 2013 (person file) data set and as a result, the imputation models might behave differently when applied to different data sets. Other methods like k-nearest neighbour imputation, MHDI and imputation using association rules were not applied to the data because the data set used was very large and due to the memory limitations of the R programming software these analyses could not be carried out. Future research can focus on applying the methods under different mechanisms of missingness on multiple data sets, and finding solutions to resolve the memory issue in R when using large data sets.

Reference List

Allison, P.D., (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research* 28 (3), 301-309.

Allison, P.D., (2001). *Missing Data*. Sage University Paper Series on Quantitative Application in the Social Sciences. Thousand Oaks, CA: Sage publications. Series no 07-136.

Andridge, R.R., and Little, R.J.A., (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review* 78 (1), 40-64.

Bashir, S., Razzaq, S., Maqbool, U., Tahir, S., and Baig, A.R., (2006). Using Association Rules for Better Treatment of Missing Values. The proceedings of the 10th WSEAS international conference on computers. Vouliagmeni, Athens, Greece, 1080-1085.

Bodner, T.E., (2008). What Improves with Increased Missing data Imputations ?. *Structural Equation Modeling* 15, 651-675.

Breiman, L., (2001). Random Forests. *Machine Learning* 45 (1), 3-32.

Coutinho, W., and de Waal, T., (2012). Hot Deck Imputation of Numerical Data under Edit Restrictions. Statistics Netherlands, Discussion paper, 1-26.

Cranmer, S.J., and Gill, J., (2012). We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data. *British Journal of Political Science* 43 (2), 425-449.

Durrant, G.B., (2005). Imputation Methods for Handling Item- Non Response in Social Sciences: A Methodological Review. National Centre for Research Methods Working Paper Series, 1-42.

Enders, C.K., (2010). *Applied Missing Data Analysis*. New York: The Guilford Press. 401 pages.

Finch, H., and Margraf, M., (2008). Imputation of Categorical Missing Data: A Comparison of Multivariate Normal and Multinomial Methods. *Journal of Data Science* 8, 361-378.

Gandhi, N., (2011). *Chapter 24 Logistic Regression*. [Online], available: http://www.uk.sagepub.com/burns/website%20material/Chapter%2024%20-%20Logistic%20regression.pdf. [12 November 2014].

Graham, J.W., (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* 60, 549-576.

Grannell, A., and Murphy, H., (2011). Using Multiple Imputation to Adjust for Survey Non-Response. Proceedings of the sixth ASC conference, University of Bristol, UK, 123-139.

Hapfelmeier, A., and Ulm, K., (2014). Variable Selection by Random Forests Using Data with Missing Values. *Computational and Statistical Data Analysis* 80, 129-139.

Hox, J.J., (1999). A Review of Current Software for Handling Missing Data. *Kwantitatieve Methoden* 62, 123-138.

Jönsson, P., and Wohlin, C., (2006). Benchmarking K-Nearest Neighbour Imputation with Homogeneous Likert Data. *Empirical Software Engineering: An International Journal* 11(3), 463-489.

Kaiser, J., (2010). Algorithm of Missing Values Imputation in Categorical Data with Use of Association Rules. *ACEEE International Journal on Recent Trend in Engineering and Technology* 6, 111-114.

Li, X-B., (2009). A Bayesian Approach for Estimating and Replacing Missing Categorical Data. *ACM Journal of Data and Information Quality* 1(3), 1-11.

Lillis, D., (2008). *Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output*. [Online], available: http://www.theanalysisfactor.com/r-glm-model-fit/. [12 January 2015].

Little, R.J.A., and Rubin, D.B., (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons. 2nd edition, 408 pages.

Mohd Jamil, J.B., (2012). Partial Least Squares Structural Equation Modelling with Incomplete Data: An Investigation of the Impact of Imputation Methods. PhD Thesis, University of Bradford, 1-258.

Mood, C., (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26 (1), 67-82.

Munguia, T., and Armando, J., (2014). Comparison of Imputation Methods for Handling Missing Categorical Data with Univariate Pattern. *Revista de Métodos Cuantitativos para la Economía y la Empresa* 17 (1), 101-120.

NIST/SEMATECH., (2012). *Engineering Statistics Handbook. 1.3.5.4. One-Factor ANOVA.* [Online], available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda354.htm. [06 January 2015].

Rodriguez, G., (2007). *Chapter 3 Logit Model for Binary Data*. [Online], available: http://data.princeton.edu/wws509/notes/c3.pdf. [12 November 2014].

Rubin, D.B., (1976). Inference and Missing data. *Biometrika* 63, 581-592.

Rubin, D.B., (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York 258 pages.

Rubin, D.B., (1996). Multiple Imputations After 18+ Years. *Journal of the American Statistical Association* 91 (434), 473-489.

Schafer, J.L., (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London. 430 pages.

Schafer, J.L., (1999). *The Multiple Imputation FAQ Page*. [Online], Available: http://sites.stat.psu.edu/~jls/mifaq.html. [06 January 2015].

Schafer, J.L., and Graham, J.W., (2002). Missing Data: Our View of State of the Art. *Psychological Methods* 7 (2), 147-177.

Scheaffer, R.L., (1999). Categorical Data Analysis. NCSSM Statistics Leadership Institute. http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf. 1-20.

Schenker, N., and Taylor, J.M.G., (1996). Partially Parametric Techniques for Multiple Imputation. *Computational Statistics and Data Analysis* 22, 245-446.

Seltman, H., (2007). *Chapter 7 One-way ANOVA*. [Online], available: http://www.stat.cmu.edu/~hseltman/309/Book/chapter7.pdf. [06 January 2015], 171-190.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., and Hemingway, H., (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology* 179 (6), 764-774.

Stekhoven, D.J., (2013). Nonparametric Missing Value Imputation Using Random Forest. Package "missForest", version 1.4, 1-10.

Stekhoven, D.J., and Bühlmann, P., (2012). Missforest-Nonparametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics* 28 (1), 112-118.

Støvring, H., (2013). Missing Data Mechanisms-Taxonomy, Patterns, and Implications. Department of Biostatistics, Aarhus University, 1-28.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P., (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modelling. *Journal of Chemical Information and Computer Sciences* 43 (6), 1947-1958.

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P.D.R., (2013). Comparison of Imputation Methods for Missing Laboratory Data in Medicine. *BMJ Open* 3, 1-7.

Wayman, J.C., (2003). Multiple Imputation For Missing Data: What Is It And How Can I Use It?. Annual Meeting of the American Educational Research Association. Chicago, IL, 1-16.

Wu, J., Song, Q., and Shen, J., (2008). Missing Nominal Data Imputation Using Association Rule Based On Weighted Voting Method. International Joint Conference on Neural Networks, Hong Kong, China 1157-1162.

Xiao, Y., Song, R., Chen, M., and Hall, H.I., (2012). Direct and Unbiased Multiple Imputation for Missing Values of Categorical Variables. *Journal of Data Science* 10, 465-481.

Yuan, Y.C., (2010). *Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0)*. SAS institute Inc, 1-13.

Appendix: A

A1: MCAR Code

```
MCAR=function(data,miss=0.5){ # Miss= % of missingness (i.e. 10=100%)

n_row=nrow(data)

num_miss=round(miss*n_row,0)

variable = 1:4

i=c(0)

i=sample(1:n_row,num_miss,replace=T)

for(x in 1:num_miss){

data[i[x], variable]=NA

}

return(data)

}

MCAR(data)
```

A2: Classification Error

```
original=read.csv("cleaned data-GHS 2013.csv")

impute1=read.csv("data_MCAR_15%-imp1.csv")

mean(impute1==original)*100
```

A3: Variable Similarity Calculator

```
original=read.csv("cleaned data-GHS 2013.csv") # original data

impute1=read.csv("RF_30%_cld.csv") # imputed data

mean(original$Age_grp==impute1$Age_grp)*100
```

Appendix: B

B1: Logistic Regression

```
## data preparation for logistic regression

#Health

data$Q22GENHEALTH[data$Q22GENHEALTH==1]=0

data$Q22GENHEALTH[data$Q22GENHEALTH==2]=1

data$Q22GENHEALTH[data$Q22GENHEALTH==3]=1

data$Q22GENHEALTH[data$Q22GENHEALTH==4]=1

data$Q22GENHEALTH[data$Q22GENHEALTH==9]=1

#Geotype

data$geotype[data$geotype==2]=1

data$geotype[data$geotype==4]=2

data$geotype[data$geotype==5]=2

#Age group

data$Age_grp[data$Age_grp==2]=1
```

```
data$Age_grp[data$Age_grp==3]=1

data$Age_grp[data$Age_grp==4]=1

data$Age_grp[data$Age_grp==5]=1

data$Age_grp[data$Age_grp==6]=1

data$Age_grp[data$Age_grp==7]=1

data$Age_grp[data$Age_grp==8]=2

data$Age_grp[data$Age_grp==9]=2

data$Age_grp[data$Age_grp==10]=2

data$Age_grp[data$Age_grp==11]=2

data$Age_grp[data$Age_grp==12]=2

data$Age_grp[data$Age_grp==13]=2

data$Age_grp[data$Age_grp==14]=2

data$Age_grp[data$Age_grp==15]=2

data$Age_grp[data$Age_grp==16]=2

mylogit=glm(Q22GENHEALTH~Gender+Age_grp+geotype,family="binomial", data)

summary(mylogit)
```

B2: Kaiser Email

Dear Mr. Khosa,

thank you for your interest in the method. I am sorry for my late response.

I wrote and used my own scripts for data preparation and for application of association rules for missing values imputation. These scripts were written only for testing purposes and are practically unable to handle large data sets. Some of these scripts were written in order to prepare data set for extraction of association rules using Weka software. Weka software is able to extract association rules from data sets in csv (comma separated values) files. So I prepared these files and extract association rules from them using Weka software. Then I used another script, that I made for parsing of the Weka output, then another one to impute missing values and finally I used a script that compare imputed data set with original data set. All these scripts were written by myself.

I hope this information helps you a bit.

Regards

Jiri Kaiser

B3: Fcritical value

df2 \df1	1	2	3	4	5	6	7	8	9	10	11
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41