Master Thesis for Data Science: Business and Governance MSc

Missing Data Imputation

*Predicting Missing Values in a Categorical Data Set for application at Statistics Netherlands*

**Name:** Amber van der Meijs

**ANR:** 1278896

**Supervisor Tilburg University / Statistics Netherlands:** Ton de Waal

**Supervisor Tilburg University:** Drew Hendrickson

**Date:** September 25, 2017

**Faculty:** Tilburg School of Humanities

1. Introduction

Missing data pervade all academic fields and frequently complicate any real-world study. These studies rely on subjects’ cooperation, but expecting a complete cooperation of all subjects is an unattainable ideal; missing data are amongst us, whether one is delighted by it or not. In this thesis, missing data is defined as follows:

**Definition 1** (missing data): *instances wherein no data is present for the variable in question.*

Missing data has been recognized as a major issue to scientists and enterprises, and even the most carefully designed and executed studies produce missing values. Missing data hinders the ability to explain and understand the phenomena that are studied because one seeks to explain and understand these phenomena by collecting observations. Research results depend largely on the analyses of these observations, and therefore, missing data poses a threat to the validity of scientific research (McKnight, McKnight, Figueredo & Sidani, 2007). In one way or another, most scientific, business and economic decisions are made based on or related to the information that research publicize at the time of making important decisions. Therefore, how to manage missing values in a proper manner should become common knowledge.

* 1. **Imputation of missing values**

The most common solution to handle missing data is imputation. The definition of imputation in this thesis is built upon the definition from Rubin and Little (1987, 2014).

**Definition 2** (imputation): *imputation is a general and flexible method for handling missing-data problems. Imputations are means or draws from a predictive distribution of the missing values, and require a method of creating a predictive distribution for the imputation based on the observed data.*

An imputation method predicts a missing value using a function of auxiliary variables, the predictors. There is a vast literature on imputation since it plays an important role, not only in official statistics, but in many other fields in statistics as well. See, e.g., Rubin (1987), Schafer (1997), and Little and Rubin (2002) (De Waal, Pannekoek & Scholtus, 2011).

* 1. **The onset of missing data**

Various kinds of causes occur at times that result in missing values, such as measurements are not made, human or machine error in processing a sample, malfunction of equipment, transcription errors, drop-out in follow-up studies and clinical trials, refusal of respondents to answer a certain question, and joining two not entirely matching data sets (Goswami, Patel & Suthar, 2012; Brand, 1999). According to Longford (2006), the missing data, resulting from those causes, refers to the difference between the data we planned to collect and what we have managed to collect; this difference can also be referred to as ‘non-response’. Non-response can be distinguished at two different levels: ‘unit non-response’ and ‘item non-response’, which explains why and how missing values can occur. Unit non-response arise when none of the survey responses are available for a sampled element because of refusals, inability to participate, not-at-homes, and untraced elements, or that the unit responded to so few questions that their response is deemed useless for analysis or estimation purposes. Item non-response arise when some but not all of the responses are available, because of item refusals, inability to participate, not-at-homes, and untraced elements (Kalton & Kardinsky, 1986; Särndal, Swensson & Wretman, 1992). The respondent may, for instance, refuse to answer the question because he considers the answer to the question as private information (e.g. income or sexual habits) or it takes too much time to complete the questionnaire (de Waal, Pannekoek & Scholtus, 2011).

The reason why the data is missing, is important to resolve, because an appropriate technique for imputation needs to be selected, leading to higher efficiency and prediction accuracy (Marwala & Nelwamondo, 2008). To detect the reason behind the missing values, Rubin (1987) distinguishes several ‘missing data mechanisms’.

* 1. **Data imputation using traditional methods**

Prior to the 1970s, missing data were solved by editing, whereby a missing item could be logically inferred from other data that have been observed. A framework of inference from incomplete data was only developed in 1976. Shortly afterwards, the Expectation maximization (EM) algorithm led to the use of Maximum Likelihood (ML) techniques in managing the missing data problem. After a decade, Little and Rubin (1987) and Rubin (1987), documented the shortcomings of case deletion and single imputations and introduced Multiple Imputations (Graham & Schafer, 2002).

**Definition 3** (traditional methods): *missing data imputation methods which do not include machine learning algorithms.*

From the year 1995 until today, there have been many techniques developed for solving the missing data problem in different applications (Marwala & Nelwamondo, 2008); the methods that are most commonly referred to, concerning item non-response, are the methods documented in Kalton and Kasprzyk (1986), Rubin (1987), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), Longford (2006), Andridge and Little (2010), De Waal, Pannekoek and Scholtus (2011) and Van Buuren (2012). These studies contain different traditional imputation methods, from which hot deck donor imputation, multiple imputation and the Maximum Likelihood method stand-out the most.

Because of the advancements in computational techniques, research has been conducted to try reconstitute the most probable values through processes and to determine new approaches to approximating missing variables, such as with computational intelligence and machine learning methods. It would be extremely valuable to extend data-driven computational techniques to yield a series of plausible values (Van Buuren, 2012).

* 1. **Data imputation using advanced methods**

Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed (Samuel, 1959). These techniques are designed to find models that best fit data, except that these machine learning models are no longer restricted to probabilistic ones. Therefore, an advantage of machine learning techniques over statistical ones is that the latter require underlying explicit probabilistic models.

**Definition 4** (advanced methods): *missing data imputation methods which are based on machine learning techniques. Techniques that gives computers the ability to learn without being explicitly programmed.*

Those classical statistical techniques are most often too stringent for the oncoming Big Data era, because data sources are increasingly complex. Machine learning provides a broader class of more flexible alternative analysis methods better suited to modern sources of data (Chu & Poirier, 2015). Throughout the last decades, multiple machine learning methods have been explored as a method of missing value imputation. Algorithms such as, multilayer perception, self-organizing maps, decision tree and k-nearest neighbors were used as missing value imputation methods in different domains. These machine learning methods have been found to perform better than the traditional statistical methods (Rahman & Davis, 2013; Marwala, 2009; Cubiles-de-la-Vega, López-Coello, Pino-Mejíasm & Silva-Ramírez, 2011).

* 1. **Data imputation at Statistics Netherlands**

By means of this case study, the thesis will focus on data that are produced and collected by Statistics Netherlands. For Statistics Netherlands, and other National Statistics Institutes (NSIs), is it an important task to provide high quality statistical information on many aspects of society, as up-to-date and accurate as possible (de Waal, Pannekoek & Scholtus, 2011). Knowing that important decisions are based on NSIs’ data, releasing files with erroneous values could cause the public to lose confidence in the validity of the data and in the organization more broadly. (Granquist & Kovar, 1997; Norberg, 2009; Manrique-Vallier & Reiter, 2016).

Currently, Statistics Netherlands refers to imputation methods that have been mentioned in subsection 1.3. Besides imputation methods that have been developed by (mathematical) statisticians, also other kinds of imputation methods have been developed, for instance imputation methods based on computational intelligence and machine learning, which examples are mentioned in subsection 1.4. Such imputation methods are, however, hardly known at NSIs, and the quality of applying these methods on data from NSIs has hardly been studied.

Chu and Poirier (2015) explain why statistical agencies should consider machine learning. They state that machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. They find it crucial for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

According to De Waal, Pannekoek and Scholtus (2011), is that it has been estimated that NSIs spend approximately 40 percent of their resources on editing and imputing data. Thus, any improvement in the efficiency of the editing and imputation process should therefore be highly welcomed by NSIs.

* 1. **Goals**

As stated in section 1.4. Chu and Poirier (2015) emphasized that statistical agencies should commit to machine learning techniques because machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. Considering this statement, and the statement made by Marwala (2009) about the better performance of machine learning methods, the aim of this thesis is to explore if, with some degree of certainty, can be predicted what the missing entries are, and to see which methods yield better results on a categorical data set from Statistics Netherlands. This coincides with the goal of Statistics Netherlands itself. Evidently, this goal includes finding out which algorithm performs the best, and if, indeed, the advanced methods outperform the traditional imputation methods. Currently, there is barely literature of NSIs’ specific research on missing data imputation with machine learning techniques (on categorical data sets) to find, which should make this thesis an interesting addition to this field of research.

Of course, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest, not to estimate, predict or recover missing observations. However, it could be extremely valuable if missing values can be recovered. If so, the results of this thesis will lead to a more efficient pre-process at Statistics Netherlands, and presumably at other NSIs too.

* 1. **Problem statement and research question**

This paper emphasizes the problem of missing data that occur in real-world studies, and how this problem is currently approached by scientists and statisticians, i.e. by case deletion or data imputation. The latter technique is explored by studying traditional and advanced methods. The aim of this study is to result in an advice towards Statistics Netherlands about how certain methods perform, and thus, which methods could be qualified for Statistics Netherlands and why. The main question is, if, with some degree of certainty, can be said what the missing data entries are, and to see which method (traditional or advanced) yield better results on a categorical data set. This question is answered for Statistics Netherlands, a National Statistics Institute, which currently uses traditional imputation methods, as mentioned in subsection 1.2. Therefore, the problem statement of the thesis is formulated as follows:

**Problem statement:** *To what extent are machine learning algorithms needed in predicting missing values in a discrete data set? / To what extent are traditional imputation methods sufficient in comparison with machine learning algorithms?*

This problem statement will be answered by means of a literature review and the results of an experimental study. The comparison between traditional imputation methods and advanced methods will be made on the basis of using the literature review and the experimental study provided by this paper. The comparison will be made along three aspects: the characteristics of the imputation methods, the results of the experiment and which method(s) could be properly implemented at Statistics Netherlands. In order to answer the problem statement, we divide it into four research questions, which will be introduced below.

**Research question 1 (RQ1):** *To what extent do traditional imputation methods and the advanced imputation methods, successfully applied to the dataset without missing values, perform well on the dataset with missing values?*

**Research question 2 (RQ2):** *Which imputation method (traditional and advanced) yield the best results?*

Furthermore, to check if Chu and Poirier’s (2015) and Marwala’s (2009) statements are in fact true,

**Research question 3 (RQ3):** *To what extent do the advanced methods outperform the traditional imputation methods?*

If the statements concerning the outperformance of the advanced methods are true, then research should be done if the methods could be successful in practice at Statistics Netherlands,

**Research question 4 (RQ4):** *Are the advanced methods easy to implement in practice at Statistics Netherlands?* Because

In conclusion, answering the research questions, will result in enhancing the pre-process at Statistics Netherlands.

* 1. **Outline**

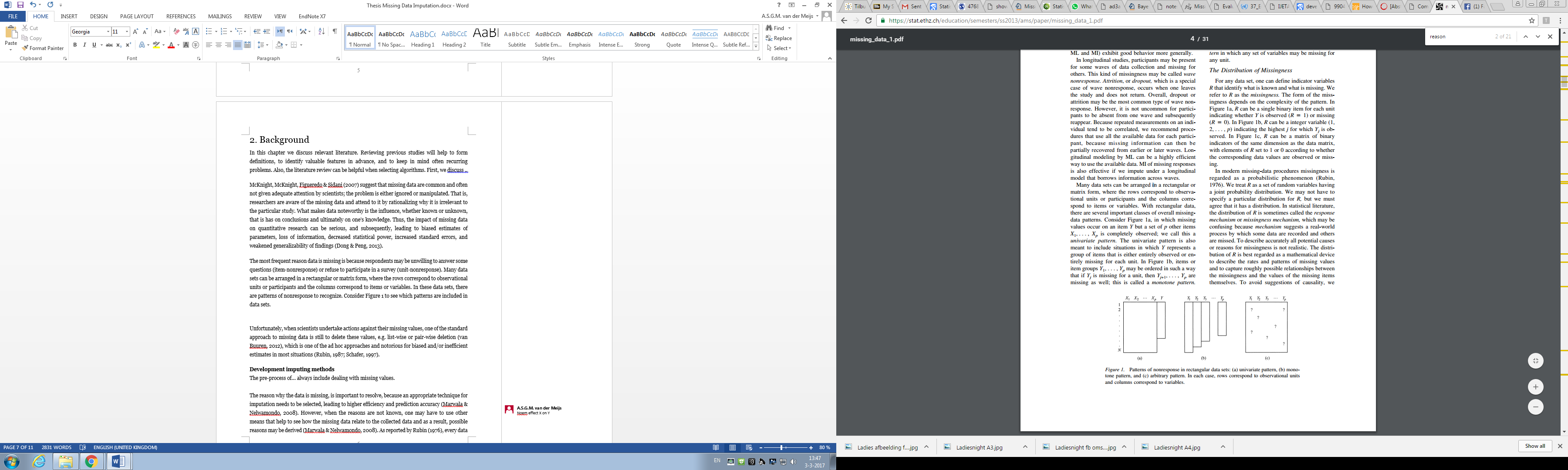
2. Background

In this chapter we discuss relevant literature. Reviewing previous studies will help to form definitions, to identify valuable knowledge in advance, and to keep in mind often recurring problems. Also, the literature review could be helpful for evaluating the methods, which should be selected for the experiment of this thesis. First, the researchers’ approach on how missing data is still commonly managed is discussed in section 2.1. In section 2.2. the theory of the three missing data mechanisms (MCAR, MAR and MNAR) is explained. Followed by how data imputation is has changed throughout the years in section 2.3., and in section 2.4. how data imputation is used within the walls of Statistics Netherlands (and other NSIs). Because this research is focussed on data imputation on a categorical data set, section 2.5. is devoted on this subject. To summarize what this thesis is contributing to, the arguments in section 2.6. will explain how.

* 1. **Managing missing data**

McKnight, McKnight, Figueredo & Sidani (2007) suggest that missing data are common and often not given adequate attention by scientists; the problem is either ignored or manipulated. That is, researchers are aware of the missing data and attend to it by rationalizing why it is irrelevant to the particular study. What makes data noteworthy is the influence, whether known or unknown, that is has on conclusions and ultimately on one’s knowledge. Thus, the impact of missing data on quantitative research can be serious, and subsequently, leading to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings (Dong & Peng, 2013). Unfortunately, when scientists undertake actions against their missing values, one of the standard approach to missing data is still to delete these values, e.g. list-wise or pair-wise deletion (Van Buuren, 2012), which are the ad hoc approaches and notorious for biased and/or inefficient estimates in most situations (Rubin, 1987; Schafer, 1997).

The most frequent reason data is missing, is because respondents may be unwilling to answer certain questions (item-nonresponse) or refuse to participate in a survey (unit-nonresponse). It is useful to distinguish the missing-data pattern because it describes which values are observed in the data and which values are missing. Many data sets can be arranged in a rectangular design, where the rows correspond to observational units or participants and the columns correspond to items or variables. In these data sets, there are patterns of nonresponse (missing values) to recognize. Consider Figure 1 to see which patterns are included in data sets.



*Figure 1.* Patterns of nonresponse: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern.

The aforementioned missing data patterns also distinguish the missing data mechanism(s), which concerns the relationship between ‘missingness’ and the values of variables in the data set (Little & Rubin, 2002). To have a solid understanding of these missing data mechanisms is considerably important; researchers should be aware of the options how missing data should be handled. Applying this knowledge will subsequently lead to higher efficiency and prediction accuracy (Marwala & Nelwamondo, 2008).

* 1. **Theory of MCAR, MAR and MNAR**

As reported by Rubin (1976), every data point has some likelihood of being missing, and therefore, Rubin created mechanisms that govern these probabilities. He distinguished missing data problems into three missing data mechanisms. In other words, missing data can take one of three forms: ‘Missing Completely at Random’ (MCAR), ‘Missing at Random’ (MAR) and ‘Missing Not at Random’ (MNAR).

One can refer to MCAR if the probability of being missing is the same for all cases. This implies that causes of the missing data are unrelated to the data. An example of MCAR is to consider a child in an educational study that moves to another district in the middle of the study. The missing values are MCAR if the reason for the move is unrelated to other variables in the data set. MCAR is convenient, but is often unrealistic for the data at hand.

If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random. In other words, if the explanation for a variable entry being missing is not related to the missing variables themselves. However, the cause may be related to other observed variables (Marwala, 2009). Thus, the word ‘random’ is in fact confusing, because a MAR mechanism is not random and describes systematic missingness where the bias for missing data is correlated with other observed variables in an analysis. For example, a sample is been taken from a population where the probability to be included depends on some known property. MAR is a much broader class than MCAR, and modern missing data methods generally start from the MAR assumption.

Finally, if neither MCAR nor MAR holds, then there can be spoken of missing not at random. MNAR entails that the probability of being missing varies for reasons that are unknown to us. So, it depends on unobserved measurements. The value of the unobserved responses depends on information not available for the analysis. Thus, future observations cannot be predicted without bias by the model. This makes MNAR the most complex case (Van Buuren, 2012; Thompson, 2013).

Rubin’s distinction is important for understanding why some methods will not work successively. His theory lays down the conditions under which a missing data method can provide valid statistical inferences (Van Buuren, 2012), resulting in higher effectiveness and prediction accuracy.

* 1. **Data imputation: then and now**

Multiple approaches to the problem of incomplete data exist. Throughout the years, such approaches have been studied, evaluated and implemented, and a sufficient portion are summarized in this section. In this thesis, these approaches are divided into two different categories: traditional and advanced methods.

**2.3.1. Traditional methods**

Prior to the 1970’s, missing data were solved by editing, whereby a missing item could be logically inferred from other data that have been observed. In section 2.1., we discussed that one of the standard approach to missing data is still to delete these values, e.g. list-wise or pair-wise deletion. The analysis of data with missing observations has been dominated by these two approaches (Roth, 1994). ***List-wise deletion*** is the default way of handling incomplete data in many statistical packages, and eliminates all cases with one or more missing values on the analysis variables. This approach can be useful even today, especially if values are MCAR. However, when that is not the case, the concerns are that it may yield biased parameter estimates and that there will always be some loss of power because of the unused partial data (Graham, 2009). The opinions on the value of list-wise deletion vary. The leading authors in the field, Little and Rubin (2002), argue that it is difficult to formulate the best rule to follow since the consequences of using list-wise deletion depend on more than the missing data rate alone. Schafer and Graham (2002, p.156) exhibit a neutral opinion:

If a missing data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective.

***Pair-wise deletion*** attempts to remedy the data loss problem of list-wise deletion. The idea behind pair-wise deletion is to use all available information, which is a good idea. The method calculates the means and (co)variances on all observed data. Nevertheless, when taken together these estimates have major shortcomings, because correlations and variance estimates are based on different subsets and will therefore be biased. Furthermore, there is no basis for estimating standard errors (Graham, 2009). Pair-wise deletion should only be used if the procedure that follows it is specifically designed to take deletion into account (Van Buuren, 2012).

Another simple approach is ***mean imputation***, which is perhaps the easiest way to impute by replacing each missing value with the mean of the observed values for that variable. Mean imputation may accurately predict missing data but deforms relationships between variables by “pulling” estimates of the correlation toward zero, which leads to complications with summary measures including underestimates of the standard deviation (Gelman & Hill, 2006). Thus, averaging the available items is difficult to justify theoretically either from a sampling or likelihood perspective (Schafer & Graham, 2002). A disadvantage of any single imputation method is that standard errors are underestimated, confidence intervals are too narrow and p-values are too low, suggesting a higher precision and more evidence than in fact can be concluded from the observed data (Brand, 1999). Mean imputation works on numerical and continuous data, and is not sufficient for categorical data. Researchers then use ***mode imputation*** to get the most frequent value of a variable to impute.

Mean imputation shares some common features with ***hot deck imputation***; instead of using the mean of a certain variable, it uses an observed response from a similar unit. In other words, hot deck imputation involves replacing missing values (of one or more variables for a non-respondent) with observed values from a respondent that is similar to the non-respondent with respect to characteristics observed by both cases. Despite that hot deck imputation imputes realistic values and is being used extensively in practice, this method is not as well developed as that of other imputation methods. It especially requires good matches of respondents that reflect available covariate information, which can never be guaranteed (Andridge & Little, 2010).

Unfortunately, the above-mentioned methods are simple solutions that proved to be merely working (Marwala & Nelwamondo, 2008). They lead to inefficient analyses and commonly produce severely biased estimates of the association(s) investigated (Donders, Moons, Stijnen & Van Der Heijden, 2006). Therefore, the interesting question that remains is how missing data, ideally, should be managed.

Because of the advancements in computational resources, more sophisticated imputation techniques were developed to handle missing data that, fortunately, give much better results. For example, the imputation methods maximum likelihood and multiple imputation are widely recommended in the methodological literature (Schafer & Olsen, 1998; Allison, 2002; Enders, 2006; Baraldi & Enders, 2010). These approaches are believed to be superior to the aforementioned traditional missing data methods because they produce unbiased estimates. Furthermore, maximum likelihood and multiple imputation tend to be more powerful than the traditional methods because no data are “thrown out”. ***Maximum likelihood (ML)*** treats the missing data’s random variables by removing them from the likelihood function as if they were never sampled. It uses all of the available data, complete and incomplete, to identify the parameter values that have the highest likelihood of producing the sample data. This estimation process uses a mathematical function called a log likelihood to measure the standardized distance between the observed data points and the parameters of interest (e.g. the mean), and the goal is to identify parameter estimates that minimize these distances (Baraldi & Enders, 2010). However, there are also downsides of using maximum likelihood; the good properties of maximum likelihood estimates are all ‘large sample’ approximations, and those approximations may be poor in small samples. Additionally, there is no commercial software for maximum likelihood available (Allison, 2012).

Despite of the numerous similarities between maximum likelihood and multiple imputation, the mechanics of ***multiple imputation*** are quite different. Rather than using all the available data, multiple imputation fills in the missing values. Multiple imputation creates several copies of the data set, each containing different imputed values. Analyses are then carried out on each data set using the same procedures that would have been used had the data been complete. Analysing each data set separately yield multiple sets of parameter estimates and standard errors, and these multiple sets of results are subsequently combined into a single set of results. Most of the traditional imputation methods underestimate standard errors. Multiple imputation solves this problem by incorporating the between-imputation variance in the standard errors. In this way, multiple imputation standard errors account for the fact that the imputed values are faulty guesses about the true data values (Baraldi & Enders, 2010). Compared with maximum likelihood, multiple imputation has one big advantage: it can be applied to virtually any kind of data or model. However, multiple imputation produces different results every time you use it because the imputed values are random draws. Furthermore, there are many different ways to do multiple imputation which possibly leads to uncertainty and confusion (Allison, 2012).

**2.3.2. Advanced methods**

Besides the widely recommended maximum likelihood and multiple imputation, methods where methodologists and statisticians are still content with, newly developed computational intelligence and machine learning techniques have also to be proven very successful in modeling complex problems (Marwala, 2009). Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed (Samuel, 1959). These methods are designed to find models that are the best fit for the data, and these models are no longer restricted to probabilistic ones. Furthermore, these prediction models are sophisticated procedures for handling missing data because the attribute with missing data is used as class-attribute, and the remaining attributes are used as input for the predictive model. An advantage of imputation with advanced methods, is that the missing data treatment is independent of the learning algorithm. This allows the user to select the most suitable imputation method for each situation (Batista & Monard, 2003). The overall imputation goal is to carefully substitute missing values, trying to avoid the imputation bias in the data set (Hruschka, Hruschka & Ebecken, 2007).

There is a wide family of advanced imputation methods from simple imputation techniques like *k*-nearest neighbor, to methods that analyse the relationships between attributes such as support vector machines-based and cluster-based methods, and logistic regression. The literature on imputation methods in data mining employs well-known machine learning methods for their studies, in which the authors show the convenience of imputing the missing values for the mentioned algorithms, particularly for classification. These studies usually analyse and compare one imputation method against a few others under controlled amounts of missing values, and induce them artificially with known mechanisms and probability distributions (García, Herrera & Luengo, 2011). In this section, a selection of machine learning methods, mentioned in literature regarding this subject, are discussed. Methods which are also well considered in the machine learning community.

***K-nearest neighbor classification (kNN)***  is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data (Peterson, 2009). This algorithm is a well-known decision rule that is widely used in pattern classification. The classification is achieved by identifying the nearest neighbors to a problem example and using those neighbors to determine the class of the problem. The kNN classification has two stages: 1) the determination of the nearest neighbors and 2) the determination of the class using those neighbors. KNN can be easily adopted to work with any attribute as class, by just modifying the attributes to be considered in the distance metric (Cunningham & Delany, 2007). The main drawback of the *k*-nearest neighbour, whenever the *k*-nearest neighbour looks for the most similar instances, the algorithm searches through all the data set. This limitation can be very critical for Statistics Netherlands (and other NSIs), since this research area does, as one of its main objectives, the analysis of large databases (Batista & Monard, 2003).

Imputation with *k*-nearest neighbor is used every time a missing value is found in a current instance. KNN computes the *k*-nearest neigbors. Then the *k*-nearest neigbor observations, that have non-missing values for that particular variable, are used to impute a missing value through a weighted mean of the neighbouring values. Therefore, a distance measure between instances is needed for it to be defined, e.g. the Euclidean distance. However, for continuous and categorical variables, the Gower distance is also considered (García, Herrera & Luengo, 2011; Balis, Higgins, Marrero, Mukherjee, Singal, Waljee, Warren, Zhang & Zhu, 2013).

A ***decision tree*** is basically a classifier that shows all possible outcomes and the paths leading to those outcomes in the form of a tree structure. The root of the tree does not have any incoming edges. Every other node had exactly one incoming edge and zero or more outgoing edges. If a node *n* has no outgoing edges, we call *n* a leaf node; otherwise, we call *n* an internal node. Each leaf node is labelled with one class label; each internal node is labelled with one predictor attribute called the splitting attribute. Decision trees can be used to predict the values of the target or class attribute based on the predictor attributes. The value of the target attribute shown in the leaf node is the predicted value. Trees can partition the predictor into distinct groups, so no need to re-encode the data (Marwala & Ssali, 2007; Twala, 2009).

Various algorithms used in the decision trees are CART, ID3, C4.5, OC1 and J48 with comparison of complexity or performance. The algorithms are briefly discussed below:

* ***CART*** is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. An important feature is its ability to generate regression trees. Their leaves predict a real number and not a class. The prediction in each leaf is based on the weighted mean for node. Furthermore, this algorithm can easily handle both numerical and categorical variables, and will itself identify the most significant variables and eliminates the rest. However, CART can deliver an unstable tree which can cause changes in complexity and/or location where the tree decides to split.
* ***ID3*** is considered a very simple decision tree algorithm. It determines the classification of object by testing the values of the properties. It builds a decision tree for the data in a top-down manner. At each node of the tree, one property is tested based on maximizing information gain and minimizing entropy, and the results are used to split the object set. This becomes a leaf node of the decision tree. ID3 is a very efficient algorithm which searches the whole dataset before it builds short, and the fastest, trees. However, it does not handle numeric attributes and missing values.
* ***C4.5*** is an evolution of the ID3. The decision tree grows using depth-first strategy. It considers all the possible tests that can split the data and selects a test that is given the best information gain. Additionally, C4.5 can handle both continuous and discrete attributes but does construct empty branches and is susceptible to noise (Giri & Singh, 2014).
* ***OC1***is an algorithm for generating multivariate decision trees, which classifies examples by testing linear combinations of the features at each non-leaf node of the tree. The trees then partition the space of examples with both multivariate and axis-parallel hyperplanes. OC1 is a decision tree type that is intended for applications where the problem has numeric continuous variables (Beigel, Kasif, Murthy & Salzberg, 1993).
* ***J48*** uses a pruning method to build a tree. A technique that reduces size of tree by removing overfitting data, which leads to poor accuracy in predications. The J48 algorithm classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. (Apte & Dangare, 2012).

Decision tree imputation is a method that builds decision trees to determine the missing values of each attribute, and then fills the missing values of each attribute by using its corresponding tree. The original class is treated as another attribute, while the value of the attribute becomes the “class” to be determined. The formed trees are then used to determine the unknown values of that particular attribute. In the classification phase, where the class attribute is not present, the tree uses, instead of the class attribute, all the attributes in the test set in an alternating fashion. In other words, each of the attributes would become a “class” variable at one point in time. Lobo and Numao (1999, 2000), follow-up the aforementioned (unordered) approach but by first ordering the attributes using mutual information before growing the trees. It makes sense to use the data with filled values in order to construct a decision tree for filling the missing values of other attributes (Quinlan, 1987; Twala, 2009).

A collection of the above-mentioned classifier is called the ***random forest*** classifier, and are identically distributed random vectors. The randomizing variable is used to determine how the successive cuts are performed when building the tree, such as selection of the node and the coordinate to split, as well as the position of the split. Random forests grow many decision trees and output the clustering that appears most often in the individual trees. In other words, they take a majority vote among the random tree classifiers (Breiman, 2001; Biau, Devroye & Lugosi, 2008). Random forest algorithms can be a valuable alternative, as they have been shown to be highly accurate and require little computational time. Also, this algorithm can deal with highly dimensional data, does not rely on distributional assumptions and are particularly appropriate for modelling complex interactions and non-linear relationships among variables (Brooks, Costa, Davidson, Di Marco, Graham, Penone, Rondinini, Shoemaker & Young, 2014).

Random forest is used as an imputation method because it works as a regression method. This method of imputation can handle any type of input data and makes a few as possible assumptions about structural aspects of the data (Bühlmann & Stekhoven, 2011). Many literature relating to random forest as an imputation method, commonly mention the use of corresponding R-package ‘missForest’ to impute missing values with (Bühlmann & Stekhoven, 2011; Balis, Higgins, Marrero, Mukherjee, Singal, Waljee, Warren, Zhang & Zhu, 2013; Brooks, Costa, Davidson, Di Marco, Graham, Penone, Rondinini, Shoemaker & Young, 2014; Carranza & Laborte, 2015; Tang & Ishwaran, 2017). This algorithm aims to predict individual missing values accurately rather than take random draws from a distribution, so the imputed values may lead to biased parameters estimates in statistical models (Barlett, Carpenter, Hemingway, Nicholas & Shah, 2014).

Unlike the most classifiers, the ***Support Vector Machine (SVM)*** classifier is explicitly told to find the best separating hyperplane. It maps the data into a higher dimensional input space, in which it constructs an optimal separating hyperplane (Cortes & Vapnik, 1995; Suykens & Vandewalle, 1999). The algorithm seeks to find the optimal separating hyperplane between classes by focusing on the training cases that are placed at the edge of the class descriptors. These training cases are called support vectors. Training cases other than support vectors are discarded. In other words, it only uses ‘difficult points’ close to the decision boundary, called super vectors, and maximizes the margin between super vectors. By maximizing the margins, it guarantees that the best hyperplane is found. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyperplanes, only a few qualify as the solution. It offers the best generalization ability, and it allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data (Tzotsos, 2008; Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand & Steinberg, 2007; Lu & Mizushima, 2013).

Imputation using the Support Vector Machine classifier is using a regression-based algorithm to fill in missing values, such as set the class attributes as the input attributes and the input attributes as the class attributes. Thus, SVM regression can be used to predict the missing input attributes values. In order to do that, first it selects the examples in which there are no missing attribute values. In the next step, the method sets one of the input attributes, some of those values that are missing, as the class attribute, and the class attribute as the input attributes by contraries. Finally, a SVM regression is used to predict the decision attribute values (García, Herrera & Luengo, 2011).

At last, probably the most commonly used generative classifier: ***naïve Bayes***. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Thus, it assumes that the classes are independent from each other. An advantage of the naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means of variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix (Parveen & Pattekari, 2012).

Naïve Bayes works with discrete data and requires only one pass through the training dataset, which makes it computationally efficient. Imputation based on naïve Bayes consists of two simple steps. Each attribute is treated as the class attribute, and the data are divided into two parts: 1) a training database that includes all records for which class attribute is complete and 2) a testing database for which the records are missing. First, prior probability of each non-class attribute value and frequency of each non-class attribute value in combination with each class attribute value are computed on the basis of the training database. The computed probabilities are then used to perform prediction of class attribute for the testing database, which constitute the imputed values (Farhangfar, Kurgan & Pedrycz, 2007; Dy, Farhangfar & Kurgan, 2008).

However, not every researcher is a supporter of a predictive model as an imputation approach. According to Acuña and Rodriguez (2004), are the disadvantages of this approach that the model estimated values are usually more well-behaved than the true values would be, and if there are no relationships among attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values.

In conclusion, a broad view about multiple imputation methods, traditional and advanced, has been given in this section (2.3.), including certain methods from which Statistics Netherlands are already familiar with.

* 1. **Data imputation at Statistics Netherlands**

It is the task of National Statistics Institutes (NSIs) and other official statistical institutes to provide high-quality statistical information on many aspects of society, as up-to-date and as accurately as possible. Because it is a prerequisite for NSIs to publish accurate statistics, data imputation comes in.

Currently, Statistics Netherlands mostly refers to the imputation methods from Kalton and Kasprzyk (1986), Rubin (1987), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), Longford (2006), Andridge and Little (2010), De Waal, Pannekoek and Scholtus (2011) and Van Buuren (2012). These studies contain different traditional imputation methods, from which hot deck donor imputation, multiple imputation and the Maximum Likelihood method stand-out the most. Other methods are rarely used. In conclusion, Statistics Netherlands does not apply imputation methods based on machine learning techniques. Such imputation methods are hardly known at NSIs, and the quality of applying these methods on data from NSIs has hardly been studied (De Waal, 2017). However, the amount of data can be immense, stressing the need for automatic methods.

Imputation techniques, which *are* used at NSIs, can be divided into two main classes, depending on the kind of data to be imputed: techniques for numerical data and techniques for categorical data (De Waal, Pannekoek & Scholtus, 2011). Because the data used in this study consists of categorical variables, this thesis will only take characteristics of a categorical data set into account.

* 1. **Data imputation applied on categorical data**

The dataset relevant for this study, and provided by Statistical Netherlands, consists of categorical data. At NSIs, and other statistical institutes, categorical data occur mainly in social surveys, for instance, surveys on persons or households (De Waal, Pannekoek & Scholtus, 2011).

Categorical data is data which can only take a finite of countable number of values (Andersen, 2012). Differently explained, categorical data is qualitative data where the possible values, which a variable can take on, form a set of categories, also known as a measurement scale. Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Furthermore, categorical variables have two main types of measurement scales. Many categorical scales have a natural ordering (e.g. excellent, good, fair, poor). The variables having these kind of ordered scales are called ordinal variables. Categorical variables having unordered scales are called nominal variables. For nominal variables, the order of listing the categories is irrelevant (e.g. types of music: classical, country, folk, jazz, rock). Categorical variables are often referred to as qualitative, to distinguish them from numerical-valued or quantitative variables. Quantitative discrete data are also considered categorical provided that the set of possible responses is small, e.g. a small set of 0, 1, 2, 3, 4, 5 (Agresti, 2007; Simonoff, 2013).

When the data in question are categorical, it is mostly not clear what the appropriate methodology for imputing missing data should be. Numerous studies are reported regarding the imputation of the numerical or continuous data but there is not much research devoted to categorical data imputation with machine learning despite the fact that many real life datasets contain categorical attributes (Nishanth & Ravi, 2016). Methods of imputation specifically designed for categorical data are either limited in terms of the number of variables they can accommodate, or have not been fully compared with the continuous data approaches used with categorical data (Finch, 2010). In addition, significant features regarding the imputation of categorical data are not always taken into account (Rey Del Castillo, 2012).

Current statistical methods for imputing missing categorical data have limited use in practice because of the concern about robustness and/or difficulty in implementation when the number of categorical variables are large. The most common types of imputation models are variants of regression models with parameters estimated from the observed correct data. However, for categorical variables, donor methods are also frequently used. It depends on the characteristics of the data set and the research goals, which imputation methods is best suited for a particular situation (De Waal, Pannekoek & Scholtus, 2011).

However, according to Graham (2009) do some researchers believe that missing categorical data requires special missing data procedures but Graham stated that this is not true in general. In conclusion, onderzoek naar categorical data imputation methods zijn schaars en niet geheel goed onderbouwd, of steeds andere methodes en niet één geluid.

* + 1. **How data imputation methods work in a categorical dataset**

***Multiple imputation***, which can based on log-linear modeling, provides an elegant and sound solution for many missing data problems concerning categorical variables. Which yields unbiased statistical inference, and it is robust against departures from the assumed imputation model. The main limitation of MI under the log-linear model is, however, that it can be applied only when the number of variables used in the imputation model is small. ***Hot deck imputation*** can be used for data sets containing large numbers of categorical variables. This non-parametric imputation method involves a search for complete cases that have (almost) the same values on the observed variables as the case with missing values, and then imputes the missing value of the latter by drawing from the empirical distribution defined by the former (Sijtsma, Van Der Ark, Van Ginkel & Vermunt, 2003). Because data are categorical, we need to replace missing data with the mode rather than the mean. Mean imputation is typically for continuous data (Ästerbro & Chen, 2003). In ***mode imputation (mean),*** the mode of an attribute is used to impute the missing value of the corresponding attribute. The disadvantage of mode imputation is that it leads to underestimation of the population variance (Nishanth & Ravi, 2016).

***maximum likelihood, K-nn, deciscion tree, random forest, svm, naïve bayes***

* 1. **Contributions of this thesis (not finished)**

Throughout this thesis, the relevance of this research has been mentioned. Nevertheless, the contribution of this thesis will be summarized.

Missing values in a data set are a recurring problem in the research-field. Therefore, knowing how to predict these values as accurately as possible, would be of great value.

This thesis aims to enhance the pre-process for every researcher but more especially at NSIs, and especially Statistics Netherlands.

1. Experimental set up

In this section, we describe the dataset and experimental procedure in detail.

* 1. **Description of the data set**

Statistical Netherlands provides a categorical dataset for this study. This dataset is a subset of the Dutch Population Census 2001, which was protected against disclosure of confidential information by means of recoding. This subset contains information on almost 190,000 persons on the variables: gender, age, position in the household, size of the household, living area in the previous year, nationality, mother country, marital status, education level, economic status, occupation and branch of industry. These are mainly categorical (discrete) data. Reasons for using this data set are that these data are actually used by Statistics Netherlands for producing important statistical information about the Netherlands. Statistics Netherlands is providing this dataset by e-mail, and this dataset is allowed to leave the Statistics Netherlands system.

The data is, e.g. by recoding, secured against possible disclosure of confidential information. Below, an overview of the variables and their categories in the data set.

|  |  |  |  |
| --- | --- | --- | --- |
| *Variable* | *Type* | *Categories* | *Meaning categories* |
| Gender (Geslacht) | Categorical | 1  2  3 | Male  Female  Unknown |
| Age (Leeftijd) | Categorical | 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  98 | 0 – 4 years  5 – 9 years  10 – 14 years  15 – 19 years  20 – 24 years  25 – 29 years  30 – 34 years  35 – 39 years  40 – 44 years  45 – 49 years  50 – 54 years  55 – 59 years  60 – 64 years  65 – 69 years  70 – 74 years  75 – 79 years  80 years and older  Unknown |
| Position in household (HH\_Pos) | Categorical | 1110  1121  1122  1131  1132  1140  1210  1220  9998 | Child  Married without children  Married with children  Living together without children  Living together with children  Single parent  Single  Other in particular household  Unknown |
| Size of the household (HH\_Grootte) | Categorical | 111  112  113  114  125  126  998 | 1 person  2 persons  3 persons  4 persons  5 persons  6 persons or more  Unknown |
| Living area previous year  (Woonregio vorig jaar) | Categorical | 1  2  3  98 | Same ‘COROP-area 3’  Other ‘COROP-area or outside the Netherlands  Does not apply, 0 years old  Unknown |
| Nationality (Nationaliteit) | Categorical | 1  2  3  98 | Netherlands  Other (Europe)  Other  Unknown |
| Country of birth (Geboorteland) | Categorical | 1  2  3  98 | Netherlands  Other (Europe)  Other  Unknown |
| Education level (Onderwijsniveau) | Categorical | 0  1  2  3  4  5  9  98 | Pre-primary  Primary  Lower secondary  Upper secondary  Post-secondary  Tertiary  No education at all  Unknown |
| Economic status (Economische status) | Categorical | 111  112  120  210  221  222  223  224  998 | Employee, other  Following education with job on the side  Independent employer  Unemployed  Following education  Retired  Houseman/wife  Other inactive  Unknown |
| Professional occupation  (Beroep) | Categorical | 1  2  3  4  5  6  7  8  9  998  999 | ISCO 1; legislators, senior officials and managers  ISCO 2; professionals  ISCO 3; technicians and assistant professionals  ISCO 4; clerks  ISCO 5; service, shop, market sales workers  Other  ISCO 7; craft and relative workers  ISCO 8; plant and machine operators and assistants  ISCO 9; elementary occupations  Unknown  Not working |
| Branch of industry (NACE/ Bedrijfstak) | Categorical | 111  122  124  131  132  133  134  135  136  137  138  139  200  998 | NACE A+B; agriculture, hunting, forestry and fishing  NACE C+D+E; mining, manufacturing and electricity  NACE F; construction  NACE G; wholesale, retail trade, repair  NACE H; hotels and restaurants  NACE I; transport, storage and communication  NACE J; financial intermediation  NACE K; real estate, renting and business activities  NACE L; public administration  NACE M; education  NACE N; health, social work  NACE O; other community, social personal service activities  Not working  Unknown |
| Marital status (Burgerlijke staat) | Categorical | 1  2  3  4  8 | Unmarried  Married  Widowed  Divorced  Unknown |
| Weight | Continuous |  |  |

* 1. **Data preparation: Creating missing values in data set**

Two datasets supplied. In R with rbind() two parts combined as one dataset.

* + 1. **Missing items of each variable / missing data pattern**
  1. **Dimensionality reduction?**
  2. **Constructing a training and a test dataset**

For most methods constructing a training and test dataset was not necessary. In most methods it was already implemented, except in the naiveBayes, SVM and Decision Trees because those algorithms were not included in an imputation packages and thus had to be written manually. How this is worked out?

* 1. **Applying traditional methods** 
     1. **Mode imputation**
     2. **Multiple imputation**
     3. **…**
  2. **Applying advanced methods**
     1. **Random forest**
     2. **..**
     3. **…**
  3. **Parameters (parameter estimation)**
  4. **Technical summary**

[**http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.7809&rep=rep1&type=pdf**](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.7809&rep=rep1&type=pdf) **page 18**

* 1. **Evaluation criteria**

RMSE, UCE, SCE and computation time (/execution time)

**Figure 1** shows the general principle of the analysis. From the original datasets (without missing values), we introduced in the data a varying percentage of missing values (from 5% to 45%) generated under an MCAR assumption. These simulated missing values were imputed using the 6 methods and the 4 [**evaluation**](https://www.omicsonline.org/searchresult.php?keyword=evaluation) criteria (RMSE, UCE, SCE and execution time) were measured. Difference between the replaced values and the original true values was evaluated by RMSE criterion, the influence of the imputed values on the quality of clustering by UCE and SCE criteria (expressed in %), and finally the execution time in minutes. For the strength of this work, we performed 1000 simulations for each original dataset and for percentage of missing values i.e. 20,000 simulations. The results were averaged over the 1000 simulations. <https://www.omicsonline.org/open-access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.php?aid=54590>

Bias? Variance?

Ideally, an imputation procedure should be capable of effectively reproducing the key outputs from a “complete data” statistical analysis of the data set of interest. However, this is usually impossible, and therefore, imputation methods are evaluated to conclude which method should produce the best results. The global aim of national statistical institutes is to produce aggregated estimates from a data set, which is kept in mind in this section. (Chambers, 2006).

Because not all methods are being tested in this thesis, the following criteria will be taken into account to select methods for the experimental stage.

Graham (2009) judged various methods by three means and thus formulated criteria where a method has to be met. According to Graham:

* The method should yield unbiased parameters. That is, the parameter estimate should be close to the population value for that parameter. Mean imputation may yield a mean for a particular variable that is close to the true parameter value, but other parameters using this method can be seriously biased.
* There should be a method for assessing the degree of uncertainty about parameter estimates. It should be able to obtain reasonable estimates of the standard error or confidence interval. Thus, the parameter estimates should have standard errors correctly reflecting the uncertainty due to missing data.
* At last, the method should have statistical power. Thus, it should be stable in order to avoid unnecessary loss of power in the statistical analysis.

Ideally, these criteria should be met for data sets with both small and large number of variables, sample sizes and percentages of incomplete data, and for both simple and complex associations in the data (Graham, 2009; Van Der Palm, Van Der Ark & Vermunt, 2012).

Generally speaking, prediction methods are used for imputation of continuous variables and classification methods for discrete or categorical. Because predicting values in a categorical data set is also referred as classification, this thesis will handle this as an additional criteria while evaluating advanced methods.