Missing Data Imputation: Predicting Missing Values

Amber van der Meijs (1278896)

Master's Thesis Data Science: Business and Governance

THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE OF DATA SCIENCE
AT THE FACULTY OF HUMANITIES
OF TILBURG UNIVERSITY

Thesis committee:

Prof. Dr. A.G. Ton de Waal

A.T. Drew Hendrickson

Tilburg University
School of Humanities
Tilburg, The Netherlands
July 2018

Preface

The master Data Science: Business and Governance has been an eventful period. I started my time at the Tilburg University after working for 3 years, and was excited to start a new education. What started as a year, became two years. It appeared I had trouble to focus on my school work, and rather kept myself busy with work. I was probably the least structured thesis-student from current year. However, as the end of the two years came closer I was determined to finish this Master. Here I am, writing the preface of this thesis. I came finally this far. This thesis means that it completes my education at the Tilburg University, resulting in my very own Master of Science degree in Data Science.

It was not possible to finish this thesis without the support of my two supervisors, Ton de Waal and Drew Hendrickson. Ton, who introduced me to a whole new and important field of statistics what resulted in me learning a lot about this interesting field and how it had and still evolves through time and Drew, who took the role as 2nd supervisor to a whole new level. Drew mostly helped me with my code in R and made me see certain subjects of machine learning more clearly. Thanks to both my supervisors, I am able to hand in this thesis. A thesis where I am more than content with.

Amber van der Meijs

Tilburg, July 2018

Abstract

This research is motivated by the importance of reliable results in real-world studies. Missing data has been recognized as a major issue to scientist and enterprises and poses a threat to the validity of scientific research. Especially for Statistics Netherlands and other NSIs, for whom it is an important task to provide high quality statistical information. This study aims to investigate to what extent the imputation procedure at Statistics Netherlands benefits from imputing missing values by advanced methods as compared to traditional methods. The techniques considered are traditional methods, such as Mode imputation, Random Hot Deck imputation and Multiple imputation, and the advanced methods *k*-Nearest Neighbors imputation, Decision Tree imputation and Random Forest imputation. These methods have been applied to the data of the Dutch Population Census 2001 (*ipums*), which contains socio-economic information on almost 190,000 persons in the country. Multiple missing data sets are created from the Census data set, the imputation methods are tested on these data sets and the prediction accuracies, execution times, bias percentages and the method's stability are compared. This study shows that the advanced imputation methods do outperform the traditional imputation methods in terms of accuracy and bias. However, the traditional imputation methods compute far more stable outputs and are faster than the advanced imputation methods. From the obtained results, it may be concluded that choosing the best imputation method depends on which evaluation metric weighs more heavily to the researcher. For this study, the encompassing goal of NSIs was taken into account. This goal emphasizes the importance of high quality statistical information on many aspects of society, as up-to-date and accurate as possible. Therefore, it can be said that NSIs will benefit by implementing the advanced imputation methods, which offer improvement over the older techniques.

Keywords missing values · missing data imputation · MCAR · Mode imputation · Multiple imputation · Hot Deck imputation · *k*-Nearest Neighbors imputation · Decision Tree imputation · Random Forest imputation

Contents

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 analyzes of these observations, and therefore, missing data poses a threat to the validity of scientific research (P. McKnight, K. 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.

1.1. Imputation of missing values

One manner 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.

1.2. The onset of missing data

Various kinds of causes occur at times that result in missing values, such as 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 (Brand, 1999; Goswami, Patel & Suthar, 2012). According to Longford (2005), 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 & Kasprzyk, 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). Thus, according to these statements imputation can be applied on two levels: for unit non-response and for item non-response. However, there is another level where imputation can be applied on that was not mentioned by aforementioned researchers: mass imputation. It imputes every variable for which no value was observed for all population units (De Waal, 2015). Considering the aforementioned levels, this thesis will focus on the level of item non-response. Rubin (1987) translated the aforementioned reasons into several 'missing data mechanisms', to clarify how missing data should be handled.

1.3. Data imputation using traditional methods

Prior to the 1970s, missing data were solely 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 first leading/broadly considered traditional imputation methods were developed. After a decade, Little and Rubin (1987) and Rubin (1987), documented the shortcomings of case deletion and single imputations. Consequently, new imputation methods where multiple plausible values replaces each missing value were developed (Graham & Schafer, 2002).

> **Definition 3** (traditional methods): *missing data imputation methods which are based on case deletion, mean and probabilistic models.*

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 advancements in computational techniques developed quickly and other techniques came to light. Because of these advancements in computational techniques, research has been conducted to try reconstitute the most probable values 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 plausible values (Van Buuren, 2012).

1.4. 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). An advantage of machine learning techniques over statistical ones is that the latter require underlying explicit or implicit 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.*

Classical statistical techniques are most often too stringent for the oncoming Big Data era, because those 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 for 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 (Marwala, 2009; Rahman & Davis, 2012; Silva-Ramirez, Pino-Meijas, López-Coello, & Cubiles-de-la-Vega, 2011).

1.5. 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), it is an important task to provide high quality statistical information on many aspects of society, as up-to-date and accurate as possible (De Waal et al., 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; Manrique-Vallier & Reiter, 2016; Norberg, 2009).

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, other kinds of imputation methods have been developed 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.

1.6. Scientific relevance and goals

As stated in Subsection 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 can be said, with some degree of certainty, that advanced methods outperform the traditional imputation methods, and to see which methods yield the best results on a categorical data set from Statistics Netherlands. This coincides with the goal of Statistics Netherlands itself. This also includes finding out if it can be predicted what the missing entries are. 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.

1.7. 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 (on a categorical data set), and thus, which methods could be qualified for Statistics Netherlands and why. Hence, the problem statement of the thesis reads as follows:

> **Problem statement:** *To what extent does the imputation procedure at Statistic Netherlands benefits from imputing missing values with advanced methods as compared to traditional methods?*

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 two aspects: the characteristics of the imputation methods and the results of the experiment. In order to answer the problem statement, it is divided into two research questions, which will be introduced below. Before it can be stated if Statistics Netherlands could benefit from using advanced methods, the performance of both methods on the categorical data set must first be examined:

> **Research question 1 (RQ1):** *To what extent do traditional imputation methods and the advanced imputation perform well on the categorical data set with missing values?*

Furthermore, to check if Chu and Poirier's (2015) and Marwala's (2009) statements, who advocate advanced methods, are in fact true, the second research question states:

> **Research question 2 (RQ2):** *To what extent do the advanced methods outperform the traditional imputation methods in terms of the evaluation metrics, and offer improvement over the older techniques?*

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

1.8. Outline of the thesis

The remainder of this thesis is structured as follows. Section 2 contains a literature review regarding missing data imputation. Section 3 describes the experimental setup and the corresponding procedure will be explained in detail. In Section 4, the experiment results will be presented and in Section 5 the problem statement will be answered. This section includes answering the research questions and describing directions for future research.

2. Literature Review

First, the researchers' approach on how missing data is still commonly managed is discussed in Subsection 2.1. In Subsection 2.2 the theory of the three missing data mechanisms (MCAR, MAR and MNAR) is explained. This is followed by how data imputation has changed throughout the years in Subsection 2.3. To be able to state why the methods behave a certain way, the characteristics of these methods are summarized in Subsection 2.4. In Subsection 2.5 can be read how data imputation is used within the walls of Statistics Netherlands (and other NSIs), and because this research is focused on data imputation on a categorical data set, Subsection 2.6 is devoted on this subject.

2.1. Managing missing data

McKnight et al. (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 it has on conclusions and ultimately on one's knowledge. The impact of missing data on quantitative research can be severe, and subsequently, leading to loss of information, increased standard errors, decreased statistical power, biased estimates of parameters, and damaged 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 known for producing biased and/or inefficient estimates in most situations (Rubin, 1987; Schafer, 1997).

The most frequent reason data is missing in the NSIs' data sets, is because respondents may be unwilling to answer certain questions (item non-response) or refuse to participate in a survey (unit non-response). These reasons concerns the relationship between 'missingness' and the values of variables in the data set (Little & Rubin, 2002). The beliefs based on those reasons for the missing data are translated into mechanisms of missingness. 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).

2.2. 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. MCAR means that causes of the missing data are unrelated to the data. An example of MCAR is when a child in an educational study moves to another neighborhood 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 (Baraldi & Enders, 2010).

If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random. Thus, if the explanation for a variable entry being missing is not related to the missing variables themselves, then the cause may be related to other observed variables (Marwala, 2009). 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, when a sample is been taken from a population where the change to be included depends on some known property. MAR is a much broader class than MCAR. Mostly, modern missing data methods start from the MAR assumption (Baraldi & Enders, 2010; Van Buuren, 2012).

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. For example, when students have to answer questions on their money spending behavior, the people who spend a lot of money at a casino are more likely to skip questions out of fear of getting in trouble. Thus, future observations cannot be predicted without bias by the model. This makes MNAR the most complex case (Baraldi & Enders, 2010; Van Buuren, 2012).

Rubin's distinction is important for understanding why some methods will not work as well as you might expect. This theory shows the conditions under which a missing data method provides valid statistical conclusions (Van Buuren, 2012), resulting in higher effectiveness and prediction accuracy. The present study assumes a MCAR data set. This mechanism may cause loss of statistical power but the advantage of MCAR is that the analysis remains unbiased; the estimated parameters are not biased by the absence of the data (Kang, 2013).

2.2.1. Proportions of missing data

The missing data mechanisms are broadly supported in the academic world. Mainly because this distinction has proved to have an effect on the degree of success of a method (Van Buuren, 2012). The proportion of

missing data, however, has not. The opinions about the acceptable percentage of missing data in a data set differ. Bennet (2001) stated that when the amount of missing data is greater than 10 percent, the values should be imputed and Schafer (1999) claims that 5 percent or less is inconsequential, and consequently, the data set should already be computed when 5 or more percentage is missing. Therefore, researchers might feel like imputing missing values whenever there is a certain proportion of missing data present, even a small one.

2.3. Data imputation: then and now

The overall imputation goal is to carefully substitute missing values, trying to avoid the imputation bias in the data set (Hruschka, Hruschka & Ebecken, 2007). Multiple approaches to resolve the problem of incomplete data exist. Throughout the years, such approaches have been studied, evaluated and implemented, and a sufficient portion of these methods are summarized in this Subsection. In this study, these approaches are divided into two different categories: traditional and advanced methods.

2.3.1. Traditional methods

In Subsection 2.1, it is discussed that one of the standard approaches to missing data is still to delete missing values, e.g. by 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, and eliminates all cases with one or more missing values on the 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. 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) exhibit a neutral opinion. They state that by discarding just a small part of the sample, the problem of missing data can be resolved. Such 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 and inconsistent with each other. Furthermore, there is no basis for estimating standard errors (Graham, 2009). Van Buuren (2012) stated that only when the procedure that follows is designed to take deletion into account, pair-wise deletion could be used.

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 is only used for 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. This kind of imputation may accurately predict missing data but will change the characteristics of the data set, and will introduce bias estimates (Donders, van der Heijden, Stijnen & Moons, 2006; Peng & Lei, 2005; Zhang, 2016). 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).

Mode imputation shares some common features with **Hot Deck imputation** but instead of using the mode of a certain variable, it uses an observed response from a similar unit. In other words, Hot Deck imputation involves replacing missing values 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 has its drawbacks. It especially requires good matches of respondents that reflect available covariate information, which can never be guaranteed and the method finds it hard to find matches if the number of variables is large (Andridge & Little, 2010). Furthermore, this technique is appealing to NSIs but the applicability by individual researchers could be hindered by the huge memory and storage capacity this method requires (Gyimah, 2001).

Unfortunately, some researchers, as the ones mentioned below, believe that the above-mentioned methods are simple imputation solutions that proved to be merely working (Marwala & Nelwamondo, 2008). They lead to inefficient analyzes and commonly produce severely biased estimates of the association(s) investigated (Donders et al., 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 (Allison, 2001; Baraldi & Enders, 2010; Enders, 2006; Schafer & Olsen, 1998). These approaches are believed to be superior to the aforementioned 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 discarded. **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 to identify the parameter values that have the highest likelihood of producing the

sample data (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 randomly fills in the missing values. It creates several copies of the data set with each different imputed values. After performing analyzes on each data set separately, the data sets are 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's 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 (Allison, 2012). `Mice`, multiple imputation by chained equations, is a method that researchers use to perform Multiple imputation with as strength that each variable can be modelled separately. However, the drawbacks researchers should be aware of include model selection and computing limitations (Stuart, Azur, Frangakis & Leaf, 2009).

2.3.2. Advanced methods

Besides the widely recommended Maximum Likelihood and multiple imputation (which are methods where methodologists and statisticians are still content with), newly developed computational intelligence and machine learning techniques have also proven very successful in modeling complex problems (Marwala, 2009). These methods are designed to find models that are the best fit for the data, and are more flexible and less ad hoc than the traditional imputation models (Jerez, 2010). 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 (Batista & Monard, 2003).

There is a wide family of advanced imputation methods from imputation techniques like *k*-Nearest Neighbor, to methods that analyze the relationships between attributes such as Random Forest-based methods. The literature on imputation methods in data mining applies 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 analyze and compare one imputation method against a few others under the same amounts of missing values in the data sets, and impute the missing values with the known artificial mechanisms and probability distributions (García, Herrera & Luengo, 2011). In this Subsection, a selection of machine learning methods, mentioned in literature regarding this subject, are discussed.

***k*-Nearest Neighbor classification (*k*-NN)** 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). The classification is achieved by identifying the nearest neighbors to a problem example and using those neighbors to determine the class of the problem (Cunningham & Delany, 2007). The main drawback of the *k*-Nearest Neighbor is that the algorithm searches through all the data set and becomes time-consuming. This limitation can be critical for Statistics Netherlands (and other NSIs), since NSIs perform, 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. *K*-NN computes the *k*-Nearest neighbors. Then the *k*-nearest neighbor's 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 et al., 2011; Waljee et al., 2013). Other drawbacks that *k*-NN has to be proven to have are the lack of precision in imputing variables and the introduction of spurious associations where they do not exist (Beretta & Santaniello, 2016). Good to know is that *k*-NN is a hot deck method (when *k*-NN with $k = 1$), in which *k* donors are selected from the neighbors (Jönsson & Wohlin, 2004).

Another popular model is the **Decision Tree**. It is basically a classifier that shows all possible outcomes and the paths leading to those outcomes in the form of a tree structure. If a node has no outgoing edges, this node is called a leaf node; otherwise, it is an internal node. Each leaf node is labelled with one class label and each internal node is labelled with one predictor attribute (the splitting attribute), based on these predictor attributes the target or class can be predicted. The predicted value is shown in the lead node. Trees can partition the predictor into distinct groups, so there is 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. CART is the only algorithm that handles categorical variables. CART is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. Furthermore, this algorithm can easily handle both numerical and categorical variables, and will itself identify the most significant variables and eliminate the rest. However, CART can deliver an

unstable tree which can cause changes in complexity and/or location where the tree decides to split (Singh & Gupta, 2014). A complex tree is known to be time consuming, but tend to have a low bias. Furthermore, Burgette and Reiter (2010) also state that there is a relationship between the complexity of the tree and the amount of levels in the categorical data set. Subsequently, as the complexity of the tree increases, the bias becomes smaller (Rokach, 2016). Decision Tree imputation is a method that builds decision trees to determine the missing values of each variable, and then fills the missing values of each variable by using its corresponding tree (Twala, 2009; Quinlan, 1987).

A collection of the above-mentioned classifier is called the **Random Forest** classifier. 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 (Biau, Devroye & Lugosi, 2008; Breiman, 2001). The Random forest algorithm is seen as a valuable alternative, as this algorithm can deal with highly dimensional data, is highly accurate, does not rely on distributional assumptions and the computation can be done in a little amount of time (Penone et al., 2014). 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). Much literature relating to Random Forest as an imputation method, commonly mention the use of the corresponding R package `missForest` to impute missi values with (Penone et al., 2014; Bühlmann & Stekhoven, 2011; Carranza & Laborte, 2015; Tang & Ishwaran, 2017; Waljee et al., 2013). However, 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 (Shah, Barlett, Carpenter, Nicholas & Hemingway, 2014).

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 consistent with the set of attributes 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 (Peng & Lei, 2005).

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

2.4. The similarities and differences between imputation methods

In the previous Subsection, the development from traditional to advanced methods has been walked through, and the characteristics of each method are separately explained. In that Subsection, the distinction of 'old methods' and 'new methods' was emphasized, and consequently seemed as two separate categories. This was mainly done to emphasize the development throughout the years. However, imputation methods are ordinarily classified into two categories: *stochastic* (random) and *deterministic*, depending on whether or not there is some degree of randomness in the imputation process (Kalton & Kasprzyk, 1995). These are two categories that transcends the traditional and advanced distinction and help to clarify the encompassing picture of the similarities and differences between imputation methods.

2.4.1. Stochastic and deterministic

When a method is called stochastic, it is a method which draws imputation values randomly from the observed data or the predicted distribution (Lee, 2001). Because of the random element, the imputation process may be repeated many times and produces a different completed data set each time. Thus, the variability is preserved (Little & Rubin, 2002). Commonly, Random Hot Deck imputation and Multiple imputation are part of the stochastic imputation category, as well as the machine learning imputation methods, as they are based on random draws. Because of the random component, it is possible that a stochastic imputation method behaves differently when it runs multiple times. Therefore, the output will be less constant than the deterministic methods. Thus, deterministic methods produce constant estimates, which helps for large samples (Weisberg, 2009). A deterministic method determines only one possible value for imputing each missing case. This method deduct missing values from available information (Lee, 2001) but are known to distort the shape of the distribution (Kalton, Lepkowski & Lin, 1985). Deterministic methods, are however, very fast (Kalton & Kasprzyk, 1986). Typically, methods as Mean, Median and Mode imputation fall within these category.

Another distinction that can be made is between single and multiple imputation. In single imputation, a single value is imputed for each missing value and in multiple imputation, multiple values for each missing value are imputed.

At last, any deterministic method can be made stochastic by adding a randomly assigned residual (Kalton, 1985; Kovar & Whitridge, 1995; Van Den Boogaard, El Serafy, Weerts & Gerritsen, 2005). It depends if the imputation methods used in this study will have a stochastic or deterministic approach. Some methods in this study can be both due to the content of the corresponding packages (Scholtus, 2014), e.g. it depends how the splitting of the Random Forest is determined, how *k*-NN is choosing his neighbors or which component of the Hot Deck imputation (random or sequential) will be used.

In conclusion, the theory states that stochastic methods produces better estimates but produce results that are less stable, and the deterministic methods are less reliable but have proven to be fast.

2.5. Data imputation at Statistics Netherlands

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 (2005), Andridge and Little (2010), De Waal et al. (2011) and Van Buuren (2012). These studies contain different traditional imputation methods, from which Hot Deck 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, personal communication, December 12, 2016). 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 categories, depending on the kind of data to be imputed: techniques for numerical data and techniques for categorical data (De Waal et al., 2011). Because the data used in this study consists of categorical variables, this thesis will only take the characteristics of a categorical data set into account.

2.6. Data imputation applied on categorical data

The data set 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 et al., 2011).

Categorical data is data which can only take a finite of countable number of values (Andersen, 2012). Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical data can be distinguished into different levels of measurement: nominal, ordinal, interval and ratio. In this study the categorical data set that will be used has nominal levels, which are unordered scales. For nominal variables, the order of listing the categories is thus irrelevant (e.g. types of music: classical, country, folk, jazz, rock) (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 machine learning imputation methods for numerical or continuous data but there is not much research devoted to categorical data

imputation despite the fact that many real life data sets contain categorical attributes (Nishanth & Ravi, 2016). Methods of imputation specifically designed for categorical data are limited in terms of the number of variables they can accommodate (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. However, for categorical variables, donor methods are frequently used because it has been proven that a row in a data set is chosen such that it resembles as much as possible the row with missing values. Nonetheless, the characteristics of the data set and the research goals should always be considered to find out which imputation method is best suited for a particular situation (De Waal et al., 2011). Graham (2009) has another view on the implementation of imputation methods for categorical data. According to Graham (2009) do some researchers believe that missing categorical data requires tailored missing imputation methods but Graham stated that this is not true in general. He believes that important characteristics of the variable are preserved if the right measures are accounted for (e.g. rounding or dummy coding), and therefore, missing imputation methods that are known to be good ones, as Multiple imputation and Maximum Likelihood, should work as good for categorical data than for continuous data.

3. Experimental Setup

In this section, the data set and experimental procedure are described in detail in order to answer the research questions. In Subsection 3.1, the dataset is described and in Subsection 3.2 a description about the features and their characteristics will be given. A short summary about the software that has been used in R is presented in Subsection 3.3. Subsequently, an explanation about the pre-processing of the data set is written down in Subsection 3.4 and in the last Subsection (3.5) it is discussed which evaluation criteria have been used.

3.1. Description of the data set

Statistical Netherlands provided a categorical data set for this study. This data set is a subset of the Dutch Population Census 2001, which was protected against disclosure of confidential information by means of recoding and other techniques. 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 categorical 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 provided this data set by e-mail, and this data set is allowed to leave the Statistics Netherlands system.

3.2. Feature description

Two columns that are part of this data set could be dismissed, namely the 1st column *nr* and the 13th column *Gewicht*. The *nr* (no.) column stands for an identification number of the rows (or records) in the data set. Statistics Netherlands uses the variable *Gewicht* (Weight) to obtain population estimates. Basically, *Gewicht* indicates the number of persons in the population a record stands for. These two columns were not relevant for this study, and therefore, in deliberation with Statistics Netherlands, deleted from the data set. After the adjustments, the data set contained 12 categorical variables with nominal scales (see Appendix A for a table showing the characteristics of the variables).

3.3. Software

To perform this study, additional packages had to be installed. The R package `readxl` (Wickham et al., 2018) was installed to read the provided Excel files. Furthermore, the `imputeR` (Feng, Moritz, Nowak, Welsh & O'Neil, 2017) R package is installed for introducing the missing values. The missing value

imputation was conducted using the programming language R in RStudio (version 1.1.383). The following R packages for imputation were used: `ForImp` (Barbiero, Ferrari & Manzi, 2015) R package for Mode imputation, `hot.deck` (Cranmer, Gill, Jackson, Murr & Armstrong, 2016) R package for Hot Deck imputation, `MICE` R package (Van Buuren et al., 2017) for multiple imputation and Decision Tree imputation, `missForest` (Stekhoven, 2013) for Random Forest imputation, and `bnstruct` (Sambo & Franzin, 2016) for *k*-Nearest Neighbor imputation. At last, `devtools` (Wickham, Hester & Chang, 2018) is installed to acquire the package `tictoc` (Izrailev, 2014) from Izrailev's GitHub, which itself is used to measure computing time and `plyr` (Wickham, 2016) for counting the frequencies of variables in the data sets.

3.4. Data preparation

Statistics Netherlands supplied the Census data set in two parts. Therefore, the two parts needed to be bound together to form one data set. This action was performed in R with the `rbind` function. Binding the two data sets together, resulted in the data set named *ipums* and was used for this experiment. As mentioned in the previous paragraph, the column *nr* and *Gewicht* were set to `NULL`, and therefore, deleted.

Because the original data set was without missing values, data sets with missing values needed to be created. In deliberation with Statistics Netherlands, the missing values are created 'completely at random' (MCAR). Multiple probabilities of missing data were taken into account, probabilities of 2, 5 and 10 percent. To introduce the missing values randomly, the `SimIm` function from the `imputeR` R package was used. No variables were excluded from this action. While generating these MCAR data sets, three different versions (e.g. 2.1, 2.2. and 2.3) were made for each kind of probability. Hence, all nine data sets have a different missing data pattern. After creating the data sets, they were saved to ensure that every model uses the same randomly created data set. Before imputing the data, the variables were transformed to factors (not ordered) for categorical prediction. The methods Mode imputation and *k*-NN were an exception and did not have to be transformed into factors. These two methods used the data sets as a data frame on a numerical level.

For computational reasons sometime, subsets of the data have been used. The sizes of these data sets were set on 18,792 rows and 9,486 rows, which are 10 and 5 percent of the total number of rows in the current data set. The rows were selected randomly.

3.5. Applying methods on the data

In the theoretical background multiple methods are discussed, traditional and advanced. For every single method mentioned in that section, research was done to find fitting R packages that performed imputation

based on that particular theory. For all methods, a corresponding R package was found that could perform imputation on the data set. Because Maximum Likelihood is not available for non-commercial software, this method was not included in the code.

3.5.1. Parameter estimation

In Table 1 can be read which functions from the aforementioned packages have been used and how the parameters were set.

Table 1

An overview of the used imputation methods with the corresponding R package, function, parameters and to which imputation category the imputation methods belong.

Imputation method	**R package**	**Function**	**Parameters**	**Category**
Mode imputation	ForImp	modeimp	No parameters set	Deterministic
Random Hot Deck imputation	hot.deck	hot.deck	m = 5 method = 'p.draw'	Stochastic
Multiple imputation	MICE	mice	m = 5	Stochastic
Random Forest imputation	missForest	missForest	ntree = 30 maxiter = 5 replace = TRUE	Stochastic
k-Nearest Neighbor imputation	bnstruct	knn.impute	k = 10	Deterministic
Decision Tree imputation	MICE	mice.impute.cart	meth = 'cart' minbucket = 1	Stochastic

In this study, the default values mentioned in the corresponding CRAN reference manual were used. If there were multiple options for the prediction method parameter, the appropriate method for categorical prediction was chosen. Furthermore, in Subsection 2.4.1 was mentioned that it depended on the content of the corresponding package that the method can be stochastic or deterministic. For example for categorical variables, `knn.impute` uses the mode of the neighbours, which makes *k*-NN in this case deterministic.

3.6. Evaluation method

It is crucial to evaluate the performance of the imputation methods to see which method would be the best fit in real world studies. The six models are evaluated by two evaluation methods: accuracy and bias. Because bias is also known as accuracy in statistics, the equations of both evaluation methods are stated to avoid confusion. In this study, accuracy (ML) is the performance measure that is simply the ratio of correctly predicted observations to the total observations:

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives +

False Negatives + True Negatives) (1)

For example, when the score is 0.654, it means that the model is approximately 65% accurate. Moreover, bias of an estimator is defined as the difference between the true value and the mean of measurements:

$$\text{Bias} = E(\hat{\theta}) - \theta \quad (2)$$

By measuring the bias, statements can be made about how far off the imputed values ultimately were. The bias of each value in the variables of all data sets were measured and were each transformed to an absolute percentage. From all these percentages, an average percentage was calculated. As a result, every data set had an overall bias percentage, which gave insights about the average distance between the true and imputed value in the whole data set.

Furthermore, by processing the 2, 5 and 10 percent data sets, and all their versions, multiple accuracies and biases were computed. These accuracies and biases were reported as confidence intervals in order to conclude how stable the imputation method is. Moreover, the execution time from each model was reported. As mentioned in section 3.3, the models are timed by using the `tictoc` (Izrailev, 2014) R package. The execution times were reported in minutes and also presented as confidence intervals.

In conclusion, the variation of this study was characterized by multiple factors: the different missing value probabilities, the variety in imputation methods and the times a data set with the same percentage of missing data was run.

4. Experiments and results

In this Section, the data set and experimental procedure are described in detail. Subsection 4.1 presents the results of the experiment from the methods that have been tested on the Census data set. This Subsection also includes the results of the methods that have been applied on small subsets from the Census data set.

4.1. Results of imputing the complete Census data set

Missing data imputation techniques based on both statistical and machine learning methods were applied to impute missing values in the data from Statistics Netherlands. The objective was to study the performance of both kinds of methods and to see which method, traditional or advanced, yielded better results. Several imputation methods were used to predict the missing values in the different data sets, and now, the performance can be compared. The imputation methods considered are Mode Imputation, Hot Deck Imputation, Multiple Imputation, *k*-Nearest Neighbors imputation, Decision Tree imputation and Random Forest imputation. The accuracy scores are taken as measure to compare between the six methods, and the corresponding execution times, the average bias in the data set and the length of the confidence intervals are also taken into account. All measures are presented as confidence intervals. The methods' accuracy scores, the average bias and the corresponding execution times are shown in Table 2. More information is presented in Appendix B and C, where all the accuracy score's, bias percentages and execution times can be found.

Table 2

Results of imputing the complete Census data sets with the imputation methods from this study.

Imputation method	Accuracy confidence intervals	Execution time in minutes	Overall bias in percentages
1. Mode imputation			
2% data set	[0.9904 – 0.9904]	[0.0045 – 0.0055]	[1.96 – 2.03]
5% data set	[0.9759 – 0.9762]	[0.0042 – 0.0047]	[5.04 – 5.40]
10% data set	[0.9520 – 0.9524]	[0.0043 – 0.0052]	[10.15 – 10.46]
2. Hot Deck imputation			
2% data set			
5% data set	Not retrieved	Not retrieved	Not retrieved
10% data set			

3. Multiple imputation			
2% data set			
5% data set	Not retrieved	Not retrieved	Not retrieved
10% data set			
4. *k*-Nearest Neighbors imputation			
2% data set	[0.5397 – 0.5397]	[38.19 – 39.01]	[60.07 – 60.09]
5% data set	[0.5363 – 0.5364]	[87.14 – 94.16]	[60.57 – 60.60]
10% data set	[0.5305 – 0.5306]	[138.37 – 144.07]	[61.41 – 61.48]
5. Decision Tree imputation			
2% data set	[0.9943 – 0.9943]	[528.42 – 784.19]	[0.23 – 0.28]
5% data set	[0.9851 – 0.9854]	[515.17 – 548.16]	[0.10 – 0.86]
10% data set	[0.9697 – 0.9702]	[491.07 – 507.42]	[0.96 – 1.36]
6. Random Forest imputation			
2% data set	[0.9944 – 0.9945]	[29.45 – 30.23]	[1.85 – 2.16]
5% data set	[0.9858 – 0.9860]	[28.24 – 29.35]	[5.28 – 6.12]
10% data set	[0.9707 – 0.9711]	[27.04 – 27.25]	[10.95 – 11.67]

As can be seen in Table 2, Multiple imputation and Hot Deck imputation were not able to retrieve results from the complete Census data set. In this study, the implemented imputation R packages `mice` and `hot.deck` did not scale up to the large Census data sets.

Mode imputation

Mode imputation's results imply that this method performed extremely well, on both aspects. The minimum and maximum accuracy scores that were retrieved are 0.9521 and 0.9905, which means that almost 100 percent of all the missing values were predicted correctly. However, the more missing values in the data set, the less the accuracy score became. The accuracy scores were each obtained within a fraction of a second. The duration of the execution time is not in relation with the amount of missing values in the data set, because the duration did not became longer when the missing values in the data set became larger. The outcome of the overall bias is average when compared to the other imputation methods. The bias is approximately as high as the amount of missing values. Thus, the bias got larger when the amount of

missing value got larger. The true values and the imputed value are at least 2 percent apart and at most 11 percent. As can be seen, the bias gets considerably larger when the amount of missing value gets larger, which subsequently caused relatively large confidence intervals. The coverage of the confidence intervals of both the accuracies and the execution times are relatively small.

k-Nearest Neighbors imputation

The accuracy scores from the *k*-NN imputation method imply that this method did not perform well on the Census data set. It yielded a score with a minimum of 0.5306 and a maximum of 0.5397, which means that not much more than 50 percent of the imputed values were correct predicted. The maximum accuracy score is the outcome of the 2 percent missing value data set, and the lowest from the 10 percent missing value data set. This method obtained the highest overall bias percentage. The true values and the imputed values are on average at least 60 percent apart (about 62 percent at most). The accuracy score became smaller whenever the missing value percentages became larger, what also applies for the execution time. The execution varied substantially between the three data sets. The data set with the smallest amount of missing values took around a half-hour and the data set with the most missing values took more than two hours. The coverage of the execution time intervals is getting larger compared to the data sets with a smaller missing percentage; the execution times of the 2 percent missing value data set have a smaller coverage of CI's than the next two CI's. However, the confidence intervals from the accuracy and bias of all the data sets are relatively short.

Decision Tree imputation

The Decision Tree imputation method obtained accuracy scores with a minimum of 0.9697 and a maximum of 0.9943. This means that a minimum of 96 percent of all the missing values are correctly predicted. The accuracy however drops, when the amount of missing values become larger. Obtaining these accuracies took this method around 8 or 9 hours. The execution time is not correlated with the amount of missing values in the data set. Processing the 10 percent missing data set took the shortest amount of execution time. The execution time increased, when the amount of missing values became smaller. The overall bias from the Decision Tree imputation method retrieved the smallest average percentage. The true values and the imputed values differed at most 1.5 percent. The bias is however correlated with the amount of missing values in the data set; when the amount of missing values became larger, so did the overall bias. Furthermore, the coverage of the confidence intervals from the accuracy score and execution time are the longest. The CI's of the overall bias is relatively long but not the longest. This implies that this method behaved relatively unstable.

Random Forest imputation

The Random Forest imputation method is one of the methods that performed well on the data set. The accuracy scores imply this. The accuracy score of 0.9707 for the 10 percent missing data set is the lowest. The data set with the smallest amount of missing values in the data set yielded the highest accuracy score of 0.9946. This means that the Random Forest imputation predicted at least 97 percent of all the values correctly. It took around a half-hour to impute all the values. Also for this imputation methods, the execution time gets shorter whenever the amount of missing vales in the data set become larger. Furthermore, the overall bias is comparable with the percentages of the Mode imputation method. The true values and the imputed values differ on average at most nearly 12 percent (and at least 2 percent). At last, the coverages of the accuracies scores and the execution time are the second longest of the imputation methods. The Random Forest did however obtain the longest confidence intervals of the overall bias.

The results show that the Random Forest imputation method yielded the highest accuracy scores on all missing value amounts. Extremely closely followed by the Decision Tree imputation method. These two advanced imputation methods performed best in accuracy, however the traditional Mode imputation method also performed extremely well. The Mode imputation yielded the accuracy scores the quickest of all methods. Moreover, the Decision Tree retrieved the lowest average bias percentages. The accuracy scores of the Decision Tree method are so significantly close to those from the Random Forest, that the Decision Tree method is more appealing if the overall bias is also taken in to account. However, the substantially long execution times are still a disadvantage.

Although, the *k*-Nearest Neighbors imputation method performed relatively poor on accuracy and bias, it is the most stable method in terms of accuracy and bias. The coverages of execution time and the accuracy scores from the Decision Tree are also the longest. Therefore, the Decision Tree method is the most unstable imputation method, which makes this is a supplementary disadvantage for this method.

Finally, to get an idea about how the R packages `mice` and `hot.deck` behave in comparison to the aforementioned imputation methods, the Subsection below will provide some clarifications.

4.1.1. Results of imputing subsets of the Census data set

Given the chance that some of the imputation methods might not scale well to large data sets, the methods were also evaluated on smaller sets of data. It should be considered that these results are only consulted to compare the behavior of the Multiple imputation and the Hot Deck imputation to the other methods and not to assess their performance. The subsets were set on 10 percent and 5 percent of the Census data, resulting in data sets with the sample sizes of 18,972 and 9,486. The results are presented in Table 3.

Table 3

Results of imputing the two subsets by the imputation methods from this study.

Imputation method	Missing data percentage	Accuracy confidence intervals	Execution time in minutes	Overall bias in percentage
1. Mode imputation				
10% sample size	2% data set	[0.4301 – 0.4302]	[0.0005 – 0.0008]	[89.71 – 89.72]
	5% data set	[0.4332 – 0.4333]	[0.0005 – 0.0007]	[89.87 – 90.35]
	10% data set	[0.4378 – 0.4378]	[0.0006 – 0.0007]	[90.18 – 90.69]
5% sample size	2% data set	[0.4310 – 0.4349]	[0.0005 – 0.0007]	[95.10 – 95.11]
	5% data set	[0.4343 – 0.4338]	[0.0005 – 0.0007]	[95.17 – 95.19]
	10% data set	[0.4384 – 0.4391]	[0.0007 – 0.0007]	[95.32 – 95.32]
2. Multiple imputation				
10% sample size	2% data set	[0.9939 – 0.9945]	[67.05 – 68.22]	[90.34 – 90.40]
	5% data set	[0.9852 – 0.9856]	[65.21 – 67.38]	[90.34 – 90.35]
	10% data set	[0.9703 – 0.9938]	[60.19 – 83.12]	[90.19 – 90.34]
5% sample size	2% data set	[0.4295 – 0.4296]	[32.07 – 33.12]	[94.92 – 95.06]
	5% data set	[0.4294 – 0.4296]	[31.44 – 33.34]	[94.96 – 94.99]
	10% data set	[0.4289 – 0.4293]	[30.03 – 30.50]	[94.84 – 94.91]
3. Hot Deck imputation				
10% sample size	2% data set	[0.4235 – 0.4237]	[2.30 - 2.39]	[109.16 – 110.56]
	5% data set	Not retrieved	Not retrieved	Not retrieved
	10% data set	Not retrieved	Not retrieved	Not retrieved
5% sample size	2% data set	[0.4249 – 0.4253]	[0.49 – 0.50]	[91.66 – 92.03]
	5% data set	[0.4185 – 0.4186]	[1.20 – 1.25]	[91.17 – 92.95]
	10% data set	[0.4062 – 0.4072]	[2.25 – 2.32]	[99.04 – 99.70]
4. *k*-Nearest Neighbors imputation				
10% sample size	2% data set	[0.4286 – 0.4287]	[0.30 – 0.33]	[89.65 – 89.67]
	5% data set	[0.4292 – 0.4294]	[1.06 – 1.09]	[89.73 – 90.21]

	10% data set	[0.4299 – 0.4300]	[1.08 – 1.11]	[89.90 – 90.41]
5% sample size	2% data set	[0.4299 – 0.4301]	[0.08 – 0.11]	[95.07 – 95.08]
	5% data set	[0.4303 – 0.4307]	[0.17 – 0.21]	[95.10 – 95.12]
	10% data set	[0.4313 – 0.4315]	[0.27 – 0.30]	[95.17 – 95.19]
5. Decision Tree imputation				
10% sample size	2% data set	[0.4282 – 0.4284]	[32.04 – 32.27]	[89.60 – 89.62]
	5% data set	[0.4282 – 0.4283]	[32.01 – 32.28]	[89.57 – 90.03]
	10% data set	[0.4279 – 0.4282]	[30.17 – 31.08]	[89.44 – 90.12]
5% sample size	2% data set	[0.4294 – 0.4297]	[14.23 – 14.55]	[95.04 – 95.06]
	5% data set	[0.4295 – 0.4298]	[13.41 – 14.10]	[94.99 – 95.04]
	10% data set	[0.4289 – 0.4294]	[13.04 – 13.28]	[94.97 – 95.00]
6. Random Forest imputation				
10% sample size	2% data set	[0.4278 – 0.4279]	[5.18 – 5.31]	[89.52 – 89.54]
	5% data set	[0.4270 – 0.4272]	[4.40 – 5.21]	[89.38 – 89.85]
	10% data set	[0.4256 – 0.4260]	[4.14 – 4.31]	[89.20 – 89.69]
5% sample size	2% data set	[0.4292 – 0.4294]	[1.59 – 2.16]	[95.03 – 95.04]
	5% data set	[0.4217 – 0.4288]	[3.02 – 3.04]	[94.96 – 94.99]
	10% data set	[0.4277 – 0.4278]	[2.02 – 2.49]	[94.89 – 94.95]

Multiple imputation

The Multiple imputation method yielded very different accuracy scores in both sample sizes; in the largest subset the method obtained a maximum accuracy score of 0.9945 and in the smallest subset a score of 0.4260. This is a large difference in performance. The performance in the largest subset does get lower when the amount of missing values get higher, but this is not the case with the smallest subset. It took the Multiple imputation method approximately an hour to retrieve the highest accuracy score, and a half-hour to obtain the lowest score. The execution time did not increase when the amount of missing values in the data set became larger, except for one 10 percent missing value data set. The coverage of the corresponding confidence interval of the execution time is therefore long. The remaining CI's are relatively short. The results of the average bias percentages in all the data sets imply that the bias is high but the coverages of

the intervals show that this method behaved stable in this case. However, the Multiple imputation method did behave unstable in terms of accuracy scores and corresponding execution times.

Hot Deck Imputation

The results imply that this method did not perform well on both subsets. The used Hot Deck imputation method managed to impute the 2 percent missing data set of the 10 percent subset of the Census data but did not succeed on the 5 percent and 10 percent missing value data sets. The maximum performance score was 0.4238, which is below average. The execution time however took no longer than three minutes. The Hot Deck method did impute the 5 percent subset fully, and the corresponding minimum accuracy score is 0.4062 and the maximum score is 0.4253. Again, these scores are below average. The method did manage to impute the values within three minutes. The accuracy scores and the execution times become larger when the amount of missing values in a data set become larger. However, the confidence intervals of both the accuracy scores and the execution time are short, and therefore, it can be said that this method behaved stable.

In conclusion, the Multiple imputation method (R package) managed to impute both subsets while the Hot Deck imputation method (R package) did not manage to perform the same action. The performance scores of the Multiple imputation method and the Hot Deck imputation method were below average, except the accuracy scores from the Multiple imputation on the 10 percent subset. However, all performance scores were around 0.40, including the other imputation methods (Mode, *k*-NN, Decision Tree and Random Forest). Additionally, the overall bias percentage of every data set is in all cases close to the 100 percent. (The Hot Deck imputation method even retrieved average percentages above the 100 percent due to relatively large outliers). The same results in all cases implies that the methods behave approximately the same being applied on smaller data sets in terms of accuracy and bias. The corresponding confidence intervals of both these measurements are relatively short, and therefore it implies that the methods behaved stable retrieving these results. Except the CI's from the Multiple imputation 10 percent subset, which are relatively long.

The matter where the methods do differ in, is execution time. A clearly distinction between the behavior of methods can be observed. Where Mode imputation obtained the results in a split second, as it did with the large data sets, Multiple imputation took at least a half-hour to obtain the results on the 5 percent subset and at least an hour on the 10 percent subset. The other methods needed a couple of seconds or a couple of minutes to retrieve their results. The corresponding CI's are relatively short, except the CI's from the Multiple imputation method 10 percent subset which is also in this case relatively long comparing to the

other confidence intervals. Therefore, it can be said that the Multiple imputation method applied on the 10 percent subset behaved relatively unstable in terms of accuracy, execution time and bias.

5. General discussion and conclusions

In this section, a general discussion on predicting missing values with imputation methods is provided. Furthermore, the conclusions of this study, recommendations for Statistics Netherlands and suggestions for future research are presented. First, the research questions that were formulated in Subsection 1.7 will be answered. After that, the answer for the problem statement is provided. Subsequently, the recommendations for Statistics Netherlands are given. This section is concluded with the suggestions for future research.

5.1. Answers to the research questions

In this thesis, the following problem statement was addressed: *To what extent does the imputation procedure at Statistics Netherlands benefit from imputing missing values with advanced methods as compared to traditional methods?* In order to find an answer to this problem statement, two research questions were formulated. In the remainder of this Subsection, a short conclusion and discussion upon each of the research questions is provided.

RQ 1: To what extent do traditional imputation methods and the advanced imputation methods perform well on the categorical data set with missing values?

In the present study, multiple traditional and advanced imputation methods were presented from which the traditional methods are frequently used at Statistics Netherlands. To investigate to what extent these methods perform well, the methods were applied on the categorical data set as provided by Statistics Netherlands. The results show that the Random Forest imputation method is the best performing method in terms of accuracy scores. However, the Decision Tree imputation method is more appealing because the accuracy scores are extremely close to those of the Random Forest and the Decision Tree method retrieved the lowest average bias percentages. Nevertheless, the significantly long execution times need to be taken into consideration. At last, Mode imputation obtained high accuracy scores and was the fastest method to impute all the values, which makes that this method should not be ignored.

As discussed in Subsection 2.3, the Random Forest imputation method can handle any type of data and makes few as possible assumptions about structural aspects of the data. Therefore, no statement can be made about the influence of the categorical data set on the performance. Furthermore, the performance of this method is in line with Penone et al.'s (2014) statement, who emphasized that the Random Forest method performs highly accurate and require little computation time. However, the part about *little computation time* is subjective, it is hard to say if the Random Forest method has met this expectation but it can be said, with certainty, that this method has the second best computation time of the present study. The Random

Forest imputation method behaved unstable and retrieved bias percentages that were not relatively high or low but average. That this method showed some bias, was expected since Shah et al. (2014) stated that bias occurs when using this method, because it strives to predict missing values as accurately as possible rather than take random draws from a distribution.

The Decision Tree imputation method performed well on terms of accuracy and imputed values that showed the lowest bias. The CART algorithm was used while applying the Decision Tree imputation method on the missing data. CART can easily handle categorical variables and identifies the most significant variables, as was addressed by Singh and Gupta (2014). Hence, it ended up being one of the methods that retrieved a relative high accuracy score on the categorical missing data set. However, the Decision Tree imputation method was overall the most unstable method in this study as suggested by Singh and Gupta (2014). An unstable tree caused changes in the complexity of the tree, and therefore showed low bias. This is in line with the statement of Burgette and Reiter (2010) who also stated that a complex tree tends also to be time consuming. From this statement, it can be concluded that the long execution time and low bias can be caused by complex tree created by the Decision Tree imputation method.

The last imputation method that obtained satisfactory results (above average), is the Mode imputation method. Mode imputation was the only traditional imputation method that obtained any results, which also resulted in being in the third highest accuracy score of the experiment. Furthermore, all the imputation was done in fraction of seconds and the bias was comparable with the bias of the Random Forest imputation method. It is a well-known issue that mode imputation introduce bias into the data set, as stated by multiple researchers (Donders et al., 2006; Peng & Lei, 2005; Zhang, 2016).

At last, the performance of these three imputation methods are in line with their characteristics. The Decision Tree and Random Forest imputation methods are stochastic and produce relatively low bias but showed to be the two most relatively unstable methods, which are possible effects due to the random element of the imputation process. Moreover, the Mode imputation method is deterministic and is known to perform the imputation quickly and stable (mentioned in Subsection 2.4.1), as this method did in this study. Additionally, deterministic methods are believed to produce biased estimates which the Mode imputation method also did but scored relatively low in this study.

The answer to the first research question therefore reads: the majority of the advanced methods perform well on the categorical data set with missing values. Mode imputation performed as only traditional method well on the categorical data set with missing values. However, the conclusion is that the Decision Tree imputation method came as best method out of the experiment, closely followed by the Random Forest imputation method.

RQ2: To what extent do the advanced imputation methods outperform the traditional imputation methods in terms of the evaluation metrics, and offer improvement over the older techniques?

In the answer of the first research question is concluded that the two advanced imputation methods Decision Tree and Random Forest obtained the two highest accuracy scores of the experiment. Followed by the Mode imputation, which retrieved the best results of all the traditional methods. The performance of the remaining methods, traditional and advanced, were below average and disappointing. Due to the memory error, it can be said that the used R packages `mice` and `hot.deck` did not scale well to the large Census data sets. Moreover, the results of this study show that these R packages have trouble to impute large data sets when using standard equipment, which is mentioned in earlier studies by Stuart et al. (2009) and Gyimah (2001) in Subsection 2.3. At last, the *k*-Nearest Neighbor imputation method was the only advanced method that performed below average.

The answer to the second research question therefore reads: in terms of accuracy score the advanced methods are outperforming the traditional imputation scores, and do offer improvements over the older techniques. In terms of time, the advanced methods do not outperform the traditional method Mode imputation and, at last, when the results of the overall bias of very data set are taken into account, an advanced method does outperform the traditional methods. At last, the advanced imputation methods do not outperform the Mode imputation in terms of stability; the Random Forest and the Decision Tree imputation methods were the two most unstable methods in this study. In conclusion, the advanced imputation methods did outperform the traditional imputation methods on two of the evaluation metrics, and do offer improvement on bias and accuracy over the older techniques. This conclusion is partly confirming the statements from Chu and Poirier (2015) and Marwala (2009).

5.2. Answer to the problem statement

In this Subsection, the problem statement will be answered. The answer of the second research question was clear: 'the advanced imputation methods did outperform the traditional imputation methods on two of the evaluation metrics, and do offer improvement on bias and accuracy over the older techniques'. Therefore, it cannot entirely be stated that the Statistics Netherlands benefits from imputing missing values by advanced imputation methods. The consideration of which evaluation metric weighs more heavily in the concept of 'performance' should be made by the NSIs and which of these metrics could meet their future needs better. An example of the consideration is that results showed that the Mode imputation scored a relatively high accuracy score, was significantly faster than any other imputation method and performed comparable to the Random Forest imputation method in terms of bias.

It is the task of NSIs to provide, high quality information as accurately as possible (De Waal et al., 2011). Emphasizing 'as accurately as possible', the advanced imputation methods were better in predicting the original values. Therefore, this study recognizes the task of NSIs as the encompassing goal of imputing data. At last, it is known that the traditional imputation methods have already been implemented at Statistics Netherlands, while the advanced imputation methods have not. Before a conclusion can be made, it should be stated that the advanced imputation methods can be easily implemented. Because the knowledge about advanced imputation methods could be scarce at Statistics Netherlands, it is helpful to know that it is easy to explore and apply R packages especially made for machine learning imputation. No difficult code structures have to be written and a greatly amount of literature is available.

In conclusion, there can be said to what extent the imputation procedure at Statistics Netherlands benefits from imputing missing values by advanced imputation methods as compared to the traditional imputation methods; the performance of the advanced imputation methods are benefiting Statistics Netherlands, as they are able to impute more accurate values and create new data sets that show less bias. Additionally, the encompassing goal of NSIs is considered. Using advanced imputation methods means producing more accurate statistics compared to the traditional imputation methods.

5.3. Directions for future research

Pointing out the major limitations of this study can help to suggest directions for feature research. First, this study is limited by the equipment that was used. As an individual researcher, a standard laptop was used with an 8 GB temporary memory. However, this amount of memory is already higher than the average laptop, it was not enough to run the `hot.deck` and `mice` R packages. Where this study left out the results of the Hot Deck imputation and Multiple imputation method, further research can replicate this study by obtaining these results with other R packages. In such way, a sufficient comparison between the advanced imputation methods and the traditional imputation methods can be made.

The second limitation relates to the generalizability of the results from this study. The answer of the problem statement cannot be generalized because there is no universal imputation method that performed best in accuracy, execution time, bias and stability in the different versions of all data sets. Therefore, the ultimate use of the concerning data set and/or the situation in which the imputation is performed needs to be taken into account. By adding these considerations into future research, a clear statement about which explicit method is the best fit for NSIs and therefore which method should be used, can be made.

In addition to the limitations, the machine learning techniques SVM, naive Bayes and Neural Networks are not included in this study. These three methods have also arisen in the area of missing data treatment and

have stimulated the missing data research to a new stage (Marwala, 2009; Rey Del Castillo, 2012; Yang, Janssens, Ruan, Cools, Bellemans & Wets, 2001). These methods were however not included because they were not developed into pre-fabricated R imputation packages. Future research should strive to include these methods in their study, under the same conditions as in this study to make a sufficient comparison with the other methods.

The fourth limitation concerns the statement of Graham (2009). His statement about his believe that missing categorical data does not requires tailored missing imputation methods, could not be tested. According to Graham, important characteristics of the variable are preserved if the right measures are accounted for (e.g. rounding or dummy coding), and therefore, missing imputation methods that are known to be good ones, as Multiple imputation and Maximum Likelihood, should work as good for categorical data than for continuous data. Because Multiple imputation (`mice`) did not yield any results in this study, his statement cannot be taken into consideration to be (partly) true or not true. The answer on his hypothesis should be an interesting topic for this field of research because it will reduce time spend on debating what the best method is in any kind of situation, because only the right measures have to be taken into account.

Also, apart from the limitation of this study, future research can focus on retrieving the agreement between measures, and therefore, lay more focus on the relationships between the methods and their characteristics. It is a valuable addition the field of research to explain why the methods behave as they do in the examined situation.

References

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. In: D. Banks and F.R. Morris (Eds.), *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation* (pp. 639-647).

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Allison, P.D. (2001). *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Allison, P.D. (2012). Missing Data. In R.E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 72-89). doi: 10.1435/9780857020994.n4

Andersen, E.B. (2012). *The Statistical Analysis of Categorical Data*. Retrieved from books.google.nl/books?isbn=364297225X

Andridge, R.R., & Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Non-Response. *International Statistical Review, 78*, 40-64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Baraldi, A.N., & Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37. doi: 10.1016/j.sp.2009.10.001

Barbiero, A., Ferrari, P.A., & Manzi, G. (2015). ForImp: Imputation of Missing Values Through a Forward Imputation Algorithm. R package version 1.0.3. https://cran.r-project.org/web/packages/ForImp/index.html

Batista, G.E.A.P.A., & Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. *Journal of Applied Artificial Intelligence, 17*, 519-533. https://doi.org/10.1080/713827181

Bennet, D. (2001). How can I deal with missing data? *Australian and New Zealand Journal of Public Health, 25,* 464-469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making, 16*. http://doi.org/10.1186/s12911-016-0318-z

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research, 9*, 2015-2033.

Brand, J.P.L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets* (Doctoral dissertation, Erasmus University, Rotterdam, The Netherlands). Retrieved from https://repub.eur.nl/pub/19790/990408_BRAND,%20Jacob%20Pieter%20Laurens.pdf

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Burgette, L.F., & Reiter, J.P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology, 172*, 1070-1076. https://doi.org/10.1093.aje/kwq260

Carranza, E.J.M., & Laborte, A.G. (2015). Random forest predictive modelling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). *Computer & Geosciences, 74*, 60-70. https://doi.org/10.1016/j.cageo.2014.10.004

Chu, K., & Poirier, C. (2015). *Machine Learning Documentation Initiative* (UNECE Working paper). Retrieved from United Nations Economic Commission for Europe Conference of European Statisticians website: http://www.unece.org.net4all.ch/fileadmin/DAM/stats/documents/ece/ces/ge.50/2015/Topic3_Canada_paper.pdf

Cranmer, S., Gill, J., Jackson, N., Murr, A., & Armstrong, D. (2016). hot.deck: Multiple Hot-Deck Imputation. R package version 1.1. https://cran.r-project.org/web/packages/hot.deck/index.html

Cunningham, P., & Delany, S.J. (2007). *k*-Nearest neighbour classifiers. *Multiple Classifier Systems, 35*, 1-17.

Dangare, C.S., & Apte, S.S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications, 47*, 44-48. doi: 10.5120/7228-0076

De Waal, T. (2015). *General Approaches for Consistent Estimation based on Administrative Data and Surveys* (Statistics Netherlands Discussion Paper No. 2015-11). Retrieved from Statistics Netherlands website: https://www.cbs.nl/nl-nl/achtergrond/2015/37/general-approaches-for-consistent-estimation-based-on-administrative-data-and-surveys

De Waal, T., Pannekoek, J. & Scholtus, S. (2011). *The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values* (Statistics Netherlands

Discussion Paper No. 2013-32). Retrieved from Statistics Netherlands website: https://www.cbs.nl/-/media/imported/documents/2012/03/2011-x10-32.pdf

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation.* Retrieved from books.google.com/books?isbn=0470904836

Donders, R., Moons, K.G.M., Stijnen, T., & Van Der Heijden, G.J.M.G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59,* 1087-1091. doi: 10.1016/j.clinepi.2006.01.014

Dong, Y., & Peng, C.Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*.

Feng, L., Moritz, S., Nowak, G., Welsh, A.H., & O'Neil, T.J. (2017). imputeR: A General Imputation Framework in R. R package version 2.0. https://cran.r-project.org/web/packages/imputeR/index.html

Finch, W.H. (2010). Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches. *Journal of Data Science, 8,* 361-378.

Gelman, A., & Hill, J. (2006). *From data collection to model understanding to model checking.* doi: 10.1017/CBO9780511790942.031

Graham, J.W., & Schafer, J.L. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7,* 147-177. doi: 10.1037///1082.989X.7.2.147

Graham, J.W., (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60,* 549-576. doi: 10.1146/annurev.psych.58.110405.085530

Granquist, L., & Kovar, J.G. (1997). Editing of Survey Data: How Much Is Enough?. In L. Lyberg and P. Biemer (Eds.), *Survey Measurement and Process Quality.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Gyimah, S.O.M. (2001). Missing Data in Quantitative Social Research. *PSC Discussion Papers Series, 15*.

Hruschka Jr., E.R., Hruschka, E.R., & Ebecken, N.F.F. (2007). Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems, 29*, 231-252. https://doi.org/10.1007/s10844-006-0016-x

Izrailev, S. (2014). tictoc: Functions for timing R scripts, as well as implementations of Stack and List structures. R package version 1.0. https://cran.r-project.org/web/packages/tictoc/index.html

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martin, M., & Franco, L. (2010). Missing Data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50,* 105-115. https://doi.org//10.1016/j.artmed.2010.05.002

Jönsson, P., & Wohlin, C. (2004, September) An evaluation of *k*-nearest neighbor imputation using Likert data. Paper presented at the *10th International Symposium on Software Metrics.* Retrieved from https://ieeexplore.ieee.org/abstract/document/1357895/

Kalton, G., & Kasprzyk, D. (1982). Imputing for Missing Survey Responses. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 22-31.

Kalton, G., & Kasprzyk, D. (1986). Treatment of missing survey data. *Survey Methodology, 12,* 1-16.

Kalton, G., Lepkowski, J., & Lin, T.K. (1985). Compensating for Wave Nonresponse in 1979 ISDP Research Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 327-77.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64,* 402-406. doi: 10.4097/kjae.2013.64.5.402.

Kovar, J.G., & Whitridge, P.J. (1995). Imputation of Business Survey Data. In B.G. Cox and D.A. Binder (Eds.), *Business Survey Methods.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Lee, R. (2001). *A Study of Imputation Algorithms* (NCES Working Paper No. 2001-17). Retrieved from National Center for Education Statistics website: https://nces.ed.gov/pubs2001/200117.pdf

Little, R.J.A., & Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Lobo, O.O., & Numao, M. (1999). Ordered Estimation of Missing Values. In: N. Zhong & L. Zhou (Eds.) *Methodologies for Knowledge Discovery and Data Mining, 1574,* 499-503. https://doi.org/10.1007/3-540-48912-6_67

Lobo, O.O., & Numao, M. (2000). On The Applicability of a Machine Learning Method for Estimating Missing Values. Paper presented at the *International Machine Learning Conference.* Retrieved from https://semanticsscholar.org/

Longford, N.T. (2005). *Missing data and small-area estimation.* New York, NY: Springer.

Luengo, L., García. F. & Herrera, F. (2011). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems, 32,* 77-108. doi: 10.1007/s10115-011-0424-2

Manrique-Vallier, D., & Reiter, J.P. (2016). Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data. *Journal of the American Statistical Association, 112,* 1708-1719. https://doi.org/10.1080/01621459.2016.1231612

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques.* Hershey, PA: IGI Global.

McKnight, P.E., McKnight, K.M., Figueredo, Sidani, S., & Figueredo, A.J. (2007). *Missing data. A gentle introduction.* New York, NY: Springer.

Nelwamondo, F.V. & Marwala, T. (2008). *Key Issues on Computational Intelligence Techniques for Missing Data Imputation – A Review*

Nelwamondo, F.V. (2008). *Computational Intelligence Techniques for Missing Data Imputation* (Doctoral dissertation). Retrieved from http://wiredspace.wits.ac.za/bitstream/handle/10539/5345/nelwamondo_PhD.pdf

Nishanth, K.J., & Ravi, V. (2016). Probabilistic neural network based categorical data imputation. *Neurocomputing, 218,* 17-25. https://doi.org/10.1016/j.neucom.2016.08.044

Norberg, A. (2009). Editing at Statistics Sweden – Yesterday, today and tomorrow. *Proceedings of Modernisation of Statistics Production 2009.*

Peng, L., & Lei, L. (2005). A review of missing data treatment methods. *Sixth International Conference on Intelligent Systems Design and Applications,* 633-638. doi: 10.1109/ISDA.2006.194

Penone, C., Davidson, A.D., Schoemaker, K.T., Di Marco, M., Rondinini, C., Brooks, … Costa, C.C. (2014). Imputation of missing data in life-history trait datasets: which approach performs the best? *Methods in Ecology and Evolution, 5,* 961-970. doi: 10.1111/2041-210X.12232

Peterson, L.E. (2009). K-nearest neighbor. *Scholarpedia, 4.* doi: 10.4249/scholarpedia.1883

Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies – Special Issue: Knowledge Acquisition for Knowledge-based Systems. Part 5, 27,* 221-234. doi: 10.1016/S0020-7373(87)80053-6

Rahman, M., & Davis, D.N. (2012). Machine Learning Based Missing Value Imputation Method for Clinical Datasets. *Lecture Notes in Electrical Engineering,* 244-272.

Rey Del Castillo, P. (2012, September). Use of Machine Learning Methods for Categorical Data Imputation. Paper presented at the *Conference: UNECE Work Session on Statistical Data Editing.*

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion, 27,* 111-125. https://doi.org/10.1016./j.inffus.2015.06.005

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys.* New York, NY: Wiley.

Sambo, F., & Franzin, A. (2016). bnstruct: Bayesian Network Structure Learning from Data with Missing Values. R package version 1.0.2. https://cran.r-project.org/web/packages/bnstruct/index.html

Samuel, A.L. (1959). Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development, 3,* 210-229.

Särndal, C.E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York, NY: Springer.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545−571.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London, England: Chapman & Hall.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8,* 3-15. doi: 10.1191/096228099671525676

Scholtus, S. (2017). Donor Imputation. *Handbook on Methodology of Modern Business Statistics* (pp. 3-10). Retrieved from https://ec.europa.eu/

Shah, A.D., Barlett, J.W., Carpenter, J., Nicolas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epodemiology, 179,* 764-774. https://doi.org/10.1093/aje/kwt312

Silva-Ramirez, E.L., Pino-Meijas, R., López-Coello, M., & Cubiles-de-la-Vega, M.D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Network, 24,* 121-129. https://doi.org/10.1016/j.neunet.2010.09.008

Simonoff, J.S. (2013). *Analyzing Categorical Data.* Retrieved from books.google.nl/books?isbn=0387217274

Singh, S., & Gupta, P. (2014). Comparative Study ID3, Cart and C4.5 Decision Tree Algorithm: A Survey. *International Journal of Advanced Information Science and Technology, 27,* 97-103.

Ssali, G., & Marwala, T. (2007). *Estimation of Missing Data Using Computational Intelligence and Decision Trees*

Stekhoven, D.J. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4. https://cran.r-project.org/web/packages/missForest/index.html

Stekhoven, D.J., & Bühlmann, P. (2012). MissForest – Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28,* 112-118. https://doi.org/10.1093/bioinformatics/btr597

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, *169*, 1133–1139. http://doi.org/10.1093/aje/kwp026

Suthar, B., Patel, H., & Goswami, A. (2012). A Survey: Classification of Imputation Methods in Data Mining. *International Journal of Emerging Technology and Advanced Engineering, 2,* 309-312.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining, 10.* doi; 10.1002/sam/11348

Twala, B. (2009). An Empirical Comparison of Techniques for Handling Incomplete Data Using Decision Trees. *Journal of Applied Artificial Intelligence, 23,* 373-405. https://doi.org/10.1080/08839510902872223

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Retrieved from books.google.nl/ books?isbn=1439868247

Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., … Gray, B. (2018). mice: Multivariate Imputation by Chained Equations. R package version 3.1.0. https://cran.r-project.org/web/packages/mice/index.html

Van Den Boogaard, H.F.P., El Serafy, G.Y., Weerts, A.H., & Gerritsen, H. (2005). *Conversion of Deterministic Models into Stochastic Models* (Report No. RIKZ-X0327).

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J, Balis, U., … Higgins, P.D.R. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open, 3,* 1-7. doi: 10.1136/bmjopen-2013-002847

Weisberg, H.F. (2009). *The Total Survey Error Approach: A Guido to the New Science of Survey Research.* Chicago, IL: The University of Chicago.

Wickham, H. (2016). plyr: Tools for Splitting, Applying and Combining Data. R package version 1.8.4. https://cran.r-project.org/web/packages/plyr/index.html

Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., … Miller, E. (2018). readxl: Read Excel Files. R package version 1.1.0. https://cran.r-project.org/web/packages/readxl/index.html

Wickham, H., Hester, J., & Chang, W. (2018). devtools: Tools to Make Developing R Packages Easier. R package version 1.13.6. https://cran.r-project.org/web/packages/devtools/index.html

Yang, B., Janssens, D., Ruan, D., Cools, M., Bellemans, T., & Wets, G. (2011). A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models. In: Y. Wang & T. Li (Eds.). *Foundations of Intelligent Systems. Advances in Intelligent and Soft Computing, 122*, 249-257. https://doi.org/10.1007/978-3-642-25664-6_29

Appendices

Appendix A: Variables in the Census data set

Variable	Type	Levels	Meaning categories
Gender	Categorical	1	Male
(Geslacht)		2	Female
		3	Unknown
Age (Leeftijd)	Categorical	1	0 – 4 years
		2	5 – 9 years
		3	10 – 14 years
		4	15 – 19 years
		5	20 – 24 years
		6	25 – 29 years
		7	30 – 34 years
		8	35 – 39 years
		9	40 – 44 years
		10	45 – 49 years
		11	50 – 54 years
		12	55 – 59 years
		13	60 – 64 years
		14	65 – 69 years
		15	70 – 74 years
		16	75 – 79 years
		17	80 years and older
		98	Unknown
Position in	Categorical	1110	Child
household		1121	Married without children
(HH_Pos)		1122	Married with children
		1131	Living together without children
		1132	Living together with children
		1140	Single parent
		1210	Single
		1220	Other in particular household
		9998	Unknown

Size of the household (HH_Grootte)	Categorical	111	1 person
		112	2 persons
		113	3 persons
		114	4 persons
		125	5 persons
		126	6 persons or more
		998	Unknown
Living area previous year (Woonregio vorig jaar)	Categorical	1	Same ‘COROP-area 3’
		2	Other ‘COROP-area or outside the Netherlands
		9	Does not apply, 0 years old
		998	Unknown
Nationality (Nationaliteit)	Categorical	1	Netherlands
		2	Other (Europe)
		3	Other
		98	Unknown
Country of birth (Geboorteland)	Categorical	1	Netherlands
		2	Other (Europe)
		3	Other
		98	Unknown
Education level (Onderwijsniveau)	Categorical	0	Pre-primary
		1	Primary
		2	Lower secondary
		3	Upper secondary
		4	Post-secondary
		5	Tertiary
		9	No education at all
		98	Unknown
Economic status (Economische status)	Categorical	111	Employee, other
		112	Following education with job on the side
		120	Independent employer
		210	Unemployed
		221	Following education
		222	Retired
		223	Houseman/wife

		224	Other inactive
		998	Unknown
Professional occupation (Beroep)	Categorical	1	ISCO 1; legislators, senior officials and managers
		2	ISCO 2; professionals
		3	ISCO 3; technicians and assistant professionals
		4	ISCO 4; clerks
		5	ISCO 5; service, shop, market sales workers
		6	Other
		7	ISCO 7; craft and relative workers
		8	ISCO 8; plant and machine operators and assistants
		9	ISCO 9; elementary occupations
		998	Unknown
		999	Not working
Branch of industry (NACE/ Bedrijfstak)	Categorical	111	NACE A+B; agriculture, hunting, forestry and fishing
		122	NACE C+D+E; mining, manufacturing and electricity
		124	NACE F; construction
		131	NACE G; wholesale, retail trade, repair
		132	NACE H; hotels and restaurants
		133	NACE I; transport, storage and communication
		134	NACE J; financial intermediation
		135	NACE K; real estate, renting and business activities
		136	NACE L; public administration
		137	NACE M; education
		138	NACE N; health, social work
		139	NACE O; other community, social personal service activities
		200	Not working
		998	Unknown
Marital status (Burgerlijke staat)	Categorical	1	Unmarried
		2	Married
		3	Widowed
		4	Divorced
		8	Unknown

Appendix B: The amount of times a value occurs (frequencies) in the Census data set

Variable	Levels	Frequencies
Gender (Geslacht)	1	93,474
	2	96,251
	3	NA
Age (Leeftijd)	1	9,970
	2	9,892
	3	9,725
	4	12,347
	5	10,703
	6	13,223
	7	17,552
	8	18,416
	9	17,350
	10	16,272
	11	16,017
	12	11,921
	13	9,569
	14	5,780
	15	4,475
	16	3,384
	17	2,800
	98	329
Position in household	1110	49,634
(HH_Pos)	1121	36,488
	1122	58,931
	1131	11,361
	1132	4,889
	1140	5,583
	1210	20,700
	1220	2,047
	9998	92

Size of the household	111	20,700
(HH_Grootte)	112	53,335
	113	34,960
	114	50,130,
	125	21,688
	126	8,912
	998	NA
Living area previous year	1	183,614
(Woonregio vorig jaar)	2	4,052
	9	2,059
	998	NA
Nationality (Nationaliteit)	1	184,042
	2	3,034
	3	2,602
	98	47
Country of birth	1	173,727
(Geboorteland)	2	5,620
	3	10,378
	98	NA
Education level	0	14,157
(Onderwijsniveau)	1	30,095
	2	40,005
	3	58,555
	4	5,907
	5	28,527
	9	12,414
	98	NA
Economic status	111	85,736
(Economische status)	112	5,214
	120	7,000
	210	2,656
	221	29,835
	222	19,430
	223	17,664

	224	22,188
	998	2
Professional occupation	1	12,698
(Beroep)	2	16,065
	3	17,799
	4	12,077
	5	2,106
	6	2,106
	7	10,671
	8	6,911
	9	8,102
	998	44
	999	91,774
Branch of industry (NACE/	111	2,877
Bedrijfstak)	122	14,535
	124	6,420
	131	15,990
	132	2,815
	133	5,730
	134	3,618
	135	15,019
	136	6,786
	137	6,454
	138	13,950
	139	3,742
	200	91,774
	998	15
Marital status (Burgerlijke	1	75,919
staat)	2	97,017
	3	6,862
	4	9,918
	8	9

Appendix C: The average bias in each version of the Census data set

Imputation method	Data set version	Overall bias in percentage
1. Mode imputation		
2%	2.1	1.960334
	2.2	2.028322
	2.3	1.964347
5%	5.1	5.290881
	5.2	5.400201
	5.3	5.037321
10%	10.1	10.149033
	10.2	10.333326
	10.3	10.455864
2. Multiple imputation		
2%		
5%		
		Not retrieved
10%		
3. Hot Deck imputation		
2%		
5%		
		Not retrieved
10%		
4. *k*-Nearest Neighbor imputation		

2%	2.1	60.074179
	2.2	60.073031
	2.3	60.092840
5%	5.1	60.599256
	5.2	60.585240
	5.3	60.573939
10%	10.1	61.444424
	10.2	61.408540
	10.3	61.476813
5. Decision Tree imputation		
2%	2.1	0.279252
	2.2	0.231558
	2.3	0.235408
5%	5.1	0.857067
	5.2	0.101683
	5.3	0.387938
10%	10.1	1.357688
	10.2	0.963057
	10.3	1.211382
6. Random Forest imputation		
2%	2.1	1.983295
	2.2	1.852704
	2.3	2.160937
5%	5.1	5.799638
	5.2	6.118846
	5.3	5.277349
10%	10.1	11.670075
	10.2	11.484175
	10.3	10.949237

Appendix D: The average bias in each version of the subsets of the Census data set

Imputation method	Missing data percentage	Data set version	Overall bias in percentages
1. Mode imputation			
10% sample size	2% data set	2.1	89.720263
		2.2	89.713384
		2.3	89.708835
	5% data set	5.1	90.352377
		5.2	89.874144
		5.3	89.903093
	10% data set	10.1	90.662393
		10.2	90.179886
		10.3	90.686720
5% sample size	2% data set	2.1	95.096878
		2.2	95.096150
		2.3	95.111436
	5% data set	5.1	95.187127
		5.2	95.174159
		5.3	95.173262
	10 % data set	10.1	95.315206
		10.2	95.323586
		10.3	95.311870
2. Multiple imputation			
10% sample size	2% data set	2.1	90.338451
		2.2	90.398315
		2.3	90.338451
	5% data set	5.1	90.350811
		5.2	90.343996
		5.3	90.353842
	10% data set	10.1	90.325185
		10.2	90.185565
		10.3	90.339155
5% sample size	2% data set	2.1	95.025689

		2.2	94.922657
		2.3	95.059773
	5% data set	5.1	94.984513
		5.2	94.993391
		5.3	94.960829
	10% data set	10.1	94.844220
		10.2	94.875045
		10.3	94.906995
3. Hot Deck imputation			
10% sample size	2% data set	2.1	110.564189
		2.2	109.555202
		2.3	109.162250
	5% data set	5.1	
		5.2	Not retrieved
		5.3	
	10% data set	10.1	
		10.2	Not retrieved
		10.3	
5% sample size	2% data set	2.1	92.028093
		2.2	91.655285
		2.3	91.810373
	5% data set	5.1	92.951364
		5.2	92.480671
		5.3	91.173874
	10% data set	10.1	99.697959
		10.2	99.432641
		10.3	99.040126
4. *k*-Nearest Neighbor's imputation			
10% sample size	2% data set	2.1	89.668386
		2.2	89.649807
		2.3	89.651236
	5% data set	5.1	90.213554
		5.2	89.725004

		5.3	89.758612
	10% data set	10.1	90.384794
		10.2	89.897186
		10.3	90.407806
5% sample size	2% data set	2.1	95.070628
		2.2	95.068340
		2.3	95.084877
	5% data set	5.1	95.120200
		5.2	95.106016
		5.3	95.101153
	10% data set	10.1	95.178221
		10.2	95.186974
		10.3	95.169430
5. Decision Tree imputation			
10% sample size	2% data set	2.1	89.617424
		2.2	89.610039
		2.3	89.604531
	5% data set	5.1	90.026095
		5.2	89.565130
		5.3	89.580044
	10% data set	10.1	90.122444
		10.2	89.443285
		10.3	90.090346
5% sample size	2% data set	2.1	95.048462
		2.2	95.042784
		2.3	95.056604
	5% data set	5.1	95.035774
		5.2	94.991279
		5.3	94.994775
	10% data set	10.1	94.974639
		10.2	94.964961
		10.3	94.998852
6. Random Forest imputation			
10% sample size	2% data set	2.1	89.523620

		2.2	89.542809
		2.3	89.523930
	5% data set	5.1	89.850125
		5.2	89.376368
		5.3	89.437566
	10% data set	10.1	89.681896
		10.2	89.196877
		10.3	89.686704
5% sample size	2% data set	2.1	95.036987
		2.2	95.026565
		2.3	95.029439
	5% data set	5.1	94.989468
		5.2	94.970915
		5.3	94.964362
	10% data set	10.1	94.891326
		10.2	94.891840
		10.3	94.952008

Appendix E: An overview of the accuracy scores and execution times of each version of the Census data set

Imputation method	Data set version	Accuracy scores	Execution time (minutes)
1. Mode imputation			
2%	2.1	0.990496	0.0055
	2.2	0.990458	0.0044
	2.3	0.990440	0.0052
5%	5.1	0.976260	0.0043
	5.2	0.976252	0.0047
	5.3	0.975994	0.0042
10%	10.1	0.952498	0.0052
	10.2	0.952450	0.0045
	10.3	0.952075	0.0043
2. Hot Deck Imputation			
2%			
5%		Not retrieved	
10%			
3. Multiple Imputation			
2%			
5%		Not retrieved	
10%			

4. *k*-Nearest Neighbor imputation			
2%	2.1	0.539790	38.19
	2.2	0.539702	38.22
	2.3	0.539725	39.01
5%	5.1	0.536388	87.19
	5.2	0.536438	87.14
	5.3	0.536391	94.16
10%	10.1	0.530560	144.07
	10.2	0.530621	138.37
	10.3	0.530568	140.24
5. Decision Tree imputation			
2%	2.1	0.994314	491.07
	2.2	0.994333	507.42
	2.3	0.994321	494.07
5%	5.1	0.985499	528.42
	5.2	0.985121	784.19
	5.3	0.985361	587.31
10%	10.1	0.970275	548.16
	10.2	0.970020	525.13
	10.3	0.969728	515.17
6. Random Forest imputation			
2%	2.1	0.994564	30.23
	2.2	0.994505	30.18
	2.3	0.994433	29.45
5%	5.1	0.986026	29.35
	5.2	0.985976	28.49
	5.3	0.985889	24.24
10%	10.1	0.971161	27.04
	10.2	0.970741	27.25
	10.3	0.970731	27.24

Appendix F: An overview of the accuracy scores and execution times of each version of the subsets of the Census data set

Imputation method	Missing data percentage	Data set version	Accuracy	Execution time in minutes
1. Mode imputation				
10% sample size	2% data set	2.1	0.4302	0.0005
		2.2	0.4301	0.0005
		2.3	0.4302	0.0008
	5% data set	5.1	0.4332	0.0007
		5.2	0.4332	0.0007
		5.3	0.4333	0.0005
	10% data set	10.1	0.4378	0.0007
		10.2	0.4378	0.0007
		10.3	0.4378	0.0006
5% sample size	2% data set	2.1	0.4310	0.0005
		2.2	0.4314	0.0006
		2.3	0.4349	0.0007
	5% data set	5.1	0.4348	0.0006
		5.2	0.4343	0.0005
		5.3	0.4343	0.0007
	10 % data set	10.1	0.4384	0.0007
		10.2	0.4391	0.0007
		10.3	0.4387	0.0007
2. Multiple imputation				
10% sample size	2% data set	2.1	0.994518	67.05
		2.2	0.994298	67.52
		2.3	0.993982	68.22
	5% data set	5.1	0.985390	67.38
		5.2	0.985605	65.21
		5.3	0.985206	65.03
	10% data set	10.1	0.970342	60.19
		10.2	0.970614	83.12
		10.3	0.993863	64.38

5% sample size	2% data set	2.1	0.429545	33.12
		2.2	0.429668	32.47
		2.3	0.429650	32.07
	5% data set	5.1	0.429633	33.34
		5.2	0.429422	32.07
		5.3	0.429545	31.44
		10.1	0.428903	30.37
		10.2	0.429308	30.03
		10.3	0.429167	30.50
3. Hot Deck imputation				
10% sample size	2% data set	2.1	0.423775	2.30
		2.2	0.423770	2.39
		2.3	0.423537	2.36
	5% data set	5.1		
		5.2	Not retrieved	Not retrieved
		5.3		
	10% data set	10.1		
		10.2	Not retrieved	Not retrieved
		10.3		
5% sample size	2% data set	2.1	0.424934	0.50
		2.2	0.425319	0.49
		2.3	0.425068	0.49
	5% data set	5.1	0.418922	1.20
		5.2	0.418523	1.25
		5.3	0.418666	1.23
	10% data set	10.1	0.406247	2.32
		10.2	0.406978	2.29
		10.3	0.407254	2.25
4. *k*-Nearest Neighbor's imputation				
10% sample size	2% data set	2.1	0.4286	0.30
		2.2	0.4287	0.32
		2.3	0.4286	0.33
	5% data set	5.1	0.4292	1.06
		5.2	0.4294	1.07

		5.3	0.4294	1.09
	10% data set	10.1	0.4300	1.08
		10.2	0.4299	1.08
		10.3	0.4299	1.11
5% sample size	2% data set	2.1	0.4299	0.11
		2.2	0.4301	0.08
		2.3	0.4301	0.08
	5% data set	5.1	0.4303	0.17
		5.2	0.4304	0.20
		5.3	0.4307	0.21
	10% data set	10.1	0.4315	0.30
		10.2	0.4314	0.28
		10.3	0.4313	0.27
5. Decision Tree imputation				
10% sample size	2% data set	2.1	0.4283	32.28
		2.2	0.4284	32.01
		2.3	0.4282	32.04
	5% data set	5.1	0.4282	32.27
		5.2	0.4283	32.09
		5.3	0.4283	32.04
	10% data set	10.1	0.4282	30.17
		10.2	0.4279	30.37
		10.3	0.4282	31.08
5% sample size	2% data set	2.1	0.4294	14.49
		2.2	0.4297	14.23
		2.3	0.4295	14.55
	5% data set	5.1	0.4295	14.10
		5.2	0.4296	14.07
		5.3	0.4298	13.41
	10% data set	10.1	0.4289	13.04
		10.2	0.4293	13.27
		10.3	0.4294	13.28
6. Random Forest imputation				
10% sample size	2% data set	2.1	0.4278	5.19

		2.2	0.4279	5.31
		2.3	0.4278	5.18
	5% data set	5.1	0.4272	4.40
		5.2	0.4270	4.15
		5.3	0.4271	5.21
	10% data set	10.1	0.4257	4.25
		10.2	0.4256	4.31
		10.3	0.4260	4.14
5% sample size	2% data set	2.1	0.4294	1.59
		2.2	0.4294	2.16
		2.3	0.4292	2.01
	5% data set	5.1	0.4288	3.02
		5.2	0.4217	3.04
		5.3	0.4217	3.02
	10% data set	10.1	0.4277	2.02
		10.2	0.4278	2.31
		10.3	0.4277	2.49