The application of nonparametric data augmentation and imputation using classification and regression trees within a large-scale panel study

Dissertation

Presented to the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg in Partial Fulfillment of the Requirements for the Degree of

DOCTOR RERUM POLITICARUM

by

Dipl.-Pol. Solange Diana Goßmann,
born July 23, 1985 in Buenos Aires, Argentina

Date of Submission
March 29, 2016

Principal advisor:	Prof. Dr. Susanne Rässler
	University of Bamberg, Germany
Reviewers:	Prof. Dr. Kai Fischbach
	University of Bamberg, Germany
	Prof. Dr. Guido Heineck
	University of Bamberg, Germany
Date of Submission:	March 29, 2016
Date of defence:	November 07, 2017

URN: urn:nbn:de:bvb:473-opus4-509982
DOI: https://doi.org/10.20378/irbo-50998

Abstract

Generally, multiple imputation is the recommended method for handling item nonresponse in surveys. Usually it is applied as chained equations approach based on parametric models. As Burgette & Reiter (2010) have shown classification and regression trees (CART) are a good alternative replacing the parametric models as conditional models especially when complex models occur, interactions and nonlinear models have to be handled and the amount of variables is very large. In large-scale panel studies many types of data sets with special data situations have to be handled. Based on the study of Burgette & Reiter (2010), this thesis intends to further assess the suitability of CART in combination with multiple imputation and data augmentation on some of these special situations.

Unit nonresponse, panel attrition in particular, is a problem with high impact on survey quality in social sciences. The first application aims at imputing missing values by CART to generate a proper data base for the decision whether weighting has to be considered. This decision was based on auxiliary information about respondents and nonrespondents. Both, auxiliary information and the participation status as response indicator, contained missing values that had to be imputed. The described situation originated in a school survey. The schools were asked to transmit auxiliary information about their students without knowing if they participated in the survey or not. In the end both information, auxiliary information and the participation status, should have been combined by their identification number by the survey research institute. Some data were collected and transmitted correctly, some were not. Due to those errors four data situations were distinguished and handled in different ways. 1) Complete cases, that is no missing values neither for the participation status, nor the auxiliary information. That means that the information whether the student participated were available and the auxiliary information were completely observed and correctly merged. 2) The participation status was missing, but the auxiliary information were complete. That happened when the school transmitted the auxiliary data of a student completely, but the combination with the survey participation information failed. 3) The participation status was available, but there were missings

in the auxiliary information and 4) there were missings in participation status as well as in the auxiliary information.
The procedure to handle the complete data situation 1) was a standard probit analysis. A Probit Forecast Draw was applied in situations 2) and 4) which was based on a Metropolis-Hasting algorithm that used the available information of the maximum number of participants conditional on an auxiliary variable. In practice, the amount of male and female students that participated in the survey was known. This number was used as a maximum when the auxiliary information were combined with a probable participation status. All missings in auxiliary information, that was situations 3) and 4), were augmented by CART. That means that the imputation values were drawn via Bayesian Bootstrap from final nodes of the classification and regression trees. Both, the imputation and the probit model with the response indicator as the dependent variable resulted in a data augmentation approach. All steps were chained to use as much information as possible for the analysis.
The application shows that CART can flexibly be combined with data augmentation resulting in a Markov chain Monte Carlo method or more precisely a Gibbs sampler. The results of the analysis of the (meta-)data showed a selectivity due to nonparticipation which could be explained by the variable sex. Female students tended to participate more likely than male students. The results based on the usage of CART differed clearly from those of the complete cases analysis ignoring the second level random effect as well as from those outcomes of the complete cases analysis including the second level random effect.

Surveys based on flexible filtering offer the opportunity to adjust the questionnaire to the respondents' situation. Hence, data quality can be increased and response burden can be decreased. Therefore, filters are often implemented in large-scale surveys resulting in a complex data structure, that has to be considered when imputing. The second study of this thesis shows how a data set containing many filters and a high filter-depth that limits the admissible range of values for multiple imputation can be handled by using CART. To get more into detail, a very large and complex data set contained variables that were used for the analysis of household net income. The variables were distributed over modules. Modules

are blocks of questions reffering to certain topics which are partially steered by filters. Additionally, within those modules the survey was steered by filter questions. As a consequence the number of respondents on each variable differed. It can be assumed that due to the structure of the survey missing values were mainly produced by filters or caused by the respondent intentionally and only a minor part were missing e.g. by interviewers overseeing them.
The second application shows that the described procedure is able to consider the complex data structure as the draws from CART are flexibly limited due to the changing filter structure which is generated by imputed filter steering values as well. Regarding the amount of 213 chosen variables for the household net income imputation, CART in contrast to other approaches obviously leads to time savings as no model specification is needed for each variable that has to be imputed.

Still, there is a need to get some feedback concerning the suitability of CART-based imputation. Therefore, as third application of this thesis, a simulation study was conducted to show the performance of CART in a combination with multiple imputation by chained equations (MICE) on cross-sectional data. Additionally, it was checked whether a change of settings improves the performance for the given data. There were three different data generating functions of Y. The first was a typical linear model with a normally distributed error term. The second included a chi-squared error term. The third included a non-linear (logarithmic) term. The rate of missing values was set to 60% steered by a missing at random mechanism. Regression parameters, mean, quantiles and correlations were calculated and combined. The quality of the estimation for before deletion, complete cases and the imputed data was measured by coverage, i.e. the proportion of 95%-confidence intervals for the estimated parameters that contain the true value. Additionally, bias and mean squared error were calculated.
Then, the settings were changed for the first type of data set, that was the ordinary linear model. First, the initialization was changed to a tree-based initialization instead of draws from the unconditional empirical distribution. Second, the iterations of the tree-based MI approach were increased from 20 to 50. Third, the number of imputed data sets that were combined for the confidence intervals was doubled from 15 to 30. CART-based MICE showed a good performance

(88.8% to 91.8%) for all three data sets. Additionally, it was not worthwhile changing the settings of CART for the partitioning of the simulated data.
Moreover, the third application shows some insights about the performance and the settings of CART-based MICE. There were many default settings and peculiarities that had to be considered when using CART-based MICE. The results suggest that the default settings and the performance of CART in general lead to sufficient results when conducted on cross-sectional data. Respective the settings, changing the initialization from tree-based draws to draws from the unconditonal empirical distribution is recommendable for typical survey data, that is data with missing values in large parts of the data.

The fourth application gives some insights into the performance of CART-based MICE on panel data. Therefore, the first simulated data set was extended to panel data containing information from two waves. Four data situations were distinguished, that was three random effects models with different combinations of time-variant and time-invariant variables and a fixed effects model. The last was defined by an intercept that is correlated to a regressor, the missingness steering variable X_1. CART-based MICE showed a good performance (89.0% to 91.4%) for all four data sets. CART chose the variables from the correct wave for each of the four data situations and waves. That means that only first wave information was used for the imputation of the first wave variable $Y_{t=1}$, respectively only second wave information was used for the second wave variable $Y_{t=2}$. This is crucial as the data generation for each of both waves was conducted as either independent of the other wave or the variables were time-variant for all four data situations.

This thesis demonstrates that CART can be used as a highly flexible imputation component which can be recommended with constraints for large-scale panel studies. Missing values in cross-sectional data as well as panel data can both be handled with CART-based MICE. Of course, the accuracy depends on the availability of explanatory power and correlations for both, cross-sectional and panel data. The combination of CART with data augmentation and the extension concering the filtering of the data are both feasible and promising.

In addition, further research about the performance of CART is highly recommended, for example by extending the current simulation study by changes of the variables over time based on past values of the same variable, more waves or different data generation processes.

Keywords: missing data, multiple imputation by chained equations, data augmentation, classification and regression trees

Basis for this thesis

This thesis is partly based on the following publications which are joined work with other authors:

- Aßmann et al. (2014a)
- Aßmann et al. (2014b)
- Aßmann et al. (2015)

However, the thesis focuses on my contribution to those publications and only refers to the joint work when it is crucial for a better understanding.

The data in chapter 3 differ from the scientific use file 'Organizational Reform Study in Thuringia (TH)' data from the National Educational Panel Study (NEPS) as it includes sensitive data which are only available for intern staff. The scientific use file is available at http://dx.doi.org/10.5157/NEPS:TH:1.0.0. In chapter 4 data from the NEPS are used as well: Starting Cohort 6 - Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS: SC6:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. More information about the NEPS can be found in Blossfeld et al. (2011).

The software used is *R*, see R Core Team (2014).
The approach of Burgette & Reiter (2010) which gave the impulse to this thesis is provided as *R*-Syntax at http://www.burgette.org/software.html. New is an implementation of CART within the *MICE*-package using *mice.impute.cart* which was not used for this thesis as it was not available when the process started and continued.

Declaration of academic honesty

I hereby confirm that my thesis is the result of my own work. I did not recieve any help or support from commercial consultants. All sources or materials applied are listed and specified in the thesis. I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Solange Goßmann
Bamberg, March 29, 2016

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Susanne Rässler for her support. It was her who planted the idea in my head to be a statistician. I enjoyed working for her as a student and as a PhD student. She always had, has and will have big projects in her head and due to her social network the right persons to execute them. I am thankful that she allowed me to join the NEPS team when I asked her for it. Working in the NEPS was a great experience as it was at the beginning when I started to work there and now it is a renowned Leibniz institute. Working there and handling the challenges that came up was the springboard for my thesis. I want to thank my colleagues and friends who always helped me with great discussions and/or real friendship. I want to thank my family, my older and my newer one, for the great company in all those years. Mom and dad, you were always supporting me in every way. I am conscious about that and I am happy to stand on my own feet because of the way you were there for me. My brothers and sisters, Sonia, Maximiliano and Federico, are the best brothers and sisters someone can have. We argued and we loved each other in all these years, but - most important - we always respected each other. Last but not least I would like to express my deepest love and gratitude to my husband Frank and my kids Fabian and Dana. You made and still make my life blithesome and delightful.

Chapter 1

Introduction

In a large-scale panel study various data types such as survey data or meta data and different challenges related to them occur. As especially survey data are often inflicted by nonresponse, a flexible imputation scheme is needed to avoid invalid statistical inference. The common procedure to correct for item nonresponse is multiple imputation (MI) which was invented and comprehensively shown in Rubin (1977, idea), Rubin (1978, first proposed), Rubin (1987, treatment) and many more. Unit nonresponse is often corrected by weighting, see e.g. Little & Vartivarian (2003). Multiple imputation for example implemented as multiple imputation by chained equations (MICE) is mostly based on parametric models. Those parametric models are hard to implement when for example the amount of variables is high or many nonlinear relations or interaction effects have to be included, compare Burgette & Reiter (2010). In addition, those models have to be specified for each variable with missing values. The same challenges occur when data augmentation is used to get inference about parameters from the data. Data augmentation is an imputation method that alternately combines imputations with an analytic model which results in Markov chains, see for example Tanner & Wong (1987) or K.-H. Li (1988). It can be seen as the stochastic Bayesian version of the famous EM-algorithm (EM: Expectation-Maximization), see Dempster et al. (1977). The aim of using data augmentation is to get independent random draws from the stationary distributions for the imputation.

Burgette & Reiter (2010) suggest a procedure for imputing flexibly using classification and regression trees (CART) in combination with MICE. CART is a nonparametric algorithm for recursive partition respective to a dependent variable that splits the values of this variable into subgroups using the information of other variables. Those subgroups are generated with the goal to include values that are as homogenuous as possible given a defined criterion which is the least squares deviation for continuous variable and usually the Gini impurity, also referred to as Gini index, for categorical variables. The homogenous value groups served as possible donor values for a missing value when CART was combined with multiple imputation as they represent the nonparametric characterization of the full conditional distribution.

The first paper about CART, used in combination with multiple imputation, was published in 2010. Since then, a large amount of articles were published about approaches using CART. Still, there is much more research to do in the field of CART or more general recursive partition algorithms. In the following, a limited literature review on current research is presented. All fields of research mentioned, that is Item-Response-Theory, handling interaction effects, clustering of individuals, e.g. in institutions, by group membership or by time, are closely related to challenges that occur when working with large-scale panel data as this is the focus of this thesis. Following this literature review, the applications that are handled in this thesis are introduced.

In the area of Item-Response-Theory approaches using trees came up in the last years. Research in the field of plausible values was made e.g. by Aßmann et al. (2014c). Mislevy (1991) presented the idea to combine multiple imputation with latent variables that were used to estimate population characteristics when individual values were missing in complex surveys. An example of a latent variable was the "examinees' tendencies to give correct responses to test items", see Mislevy (1991, p.179). Aßmann et al. (2014c) used CART as a component of a Markov chain Monte Carlo procedure to impute missing values in background variables and estimated plausbile values iteratively. In the field of Rasch models, that is a model that divides personal competencies and item difficulty, Strobl et

al. (2013) also used trees to impute missing data.

Doove et al. (2014) showed that CART can outclass standard applications in handling models including interaction effects when the data are multiply imputed. The results were relativized, as the potential of CART "depends on the relevance of a possible interaction effect, the correlation structure of the data, and the type of possible interaction effect present in the data", see Doove et al. (2014, p.92). Stekhoven & Bühlmann (2012) showed similar results of a tree-based approach handling interaction effects. Though, they used an alternative method, that is Random Forest, in using the *R*-package *missForest* instead of *tree* or *rpart*, which are packages for the usage of CART. Another difference is that Stekhoven & Bühlmann (2012) focused more on mixed-type data than Doove et al. (2014).

A typical challenge of large-scale panel studies arises with the clustering of individuals within e.g. institutions, families or states. When cross-sectional data include an additional multilevel structure it has to be correctly considered when the data are imputed. This task becomes even harder when the multilevel structure is included in longitudinal data as the already existing cluster of individual measurements over time is enlarged by another level that has to be considered. The question is whether CART can identify those different levels correctly and automatically. Some research has been done in the field of longitudinal and clustered data by Sela & Simonoff (2012) and Fu & Simonoff (2014). Sela & Simonoff (2012) added the consideration of random effects to trees and call their approach random effects EM tree or short RE-EM tree. Though, as trees are not fitted with maximum likelihood methods, the name is misleading. The reason for the EM within that name is the alternating estimation of regression trees and random effects. See for the exact formal description of this method Sela & Simonoff (2012, p.175). According to Sela & Simonoff (2012, p.205) the approach has the advantage that it is superior to trees ignoring the random effects within the data as it constructs different trees if the trees split on time. Additionally, it showed comparable results to linear models considering the random effects.
Fu & Simonoff (2014) adapted the algorithm of Sela & Simonoff (2012) by using a different tree approach, that is the conditional inference tree proposed by

Hothorn et al. (2006) instead of CART by Breiman et al. (1984). The decision was based on the property of CART to prefer covariates with higher amounts of possible split points.
All of these approaches ignored possible missing values within the data.

In this thesis the approach of Burgette & Reiter (2010) was adapted as basis to handle nonresponse related challenges occuring within a large-scale panel study, that is the National Educational Panel Study (NEPS). Data from the NEPS were used to demonstrate two applications of CART.

First, the approach was used to analyze the unit nonresponse on metadata to decide whether nonrespondents and respondents of a study differ as correction methods should be applicated when they do. Therefore, the approach of Burgette & Reiter (2010) was extended to a data augmentation procedure by conducting it as a component of a Markov chain Monte Carlo approach using a Gibbs Sampler. Here, the participation status as a dichotmous response indicator containing missing values was analyzed by a Bayesian Probit model.
Second, a large amount of variables were multiply imputed considering the high-complex filter structure of the data on the topic of household net income. Therefore, the method of Burgette & Reiter (2010) was used as a multiple imputation approach as originally inteded by them, but extended with a matrix that allows for correct implementation of all filter combinations. Thus, this matrix contained lists of proper donor values for each possible filter combination steering a variable's values. This construct allowed for a correct implementation of the high-complex filter hierarchy within the imputation.
Then, a simulation study was conducted to show the performance of CART-based MICE on cross-sectional data. Additionally, it was checked whether a change of settings improved the performance for the given data. All analysis were based on three different data generating functions of Y.
Finally, in order to assess if CART-based MICE is suitable for imputing panel data another simulation study was conducted. The first data generating model of the previous simulation study was extended to two waves. Here, the performance on three different combinations of time-variant and time-invariant variables for

a random effects model and an additional fixed effects model were tested. For both simulation studies, the quality of the multiple imputation mechanism was measured by coverages comparing the results of the before deletion, complete cases and multiply imputed data. For this purpose, coverage was defined as the proportion of 95%-confidence intervals for the estimated parameters that contained the true value.
The objective of all four applications was to assert if CART can be flexibly combined with other approaches or extended to work for various challenges of data imputation and analysis that occur within large-scale panel studies on a high-level performance.

Summarized, the focus of the first real data based applications was the imputation and analysis of unit nonresponse when auxiliary information, that could contain missings values as well, are available. The target of the second real data based application was the correct implementation of complex filter structures within the data while imputing using CART. The objective of the third application on simulated data was to evaluate the performance of the tree-based imputation on cross-sectional data and to test the influence of changed settings. The fourth application on simulated data focused on the performance of the CART-based imputation on panel data.
The thesis is organized as follows. In chapter 2, the theoretical foundations of MI, CART and the combination of CART with MICE on the one hand and data augmentation on the other hand is described. In chapter 3, an application with metadata from the NEPS on respondents and nonrespondents demonstrates the usage of CART combined with data augmentation to analyze the unit nonresponse process. Chapter 4 deals with survey data from the NEPS that is imputed using CART considering the objective to analyze household income questions considering the filter hierarchy of the data. In chapter 5, a simulation study evaluates the performance of CART-based MICE on cross-sectional data with three variables and three different types of data generating models. Another simulation study checks the performance of CART-based MICE for panel data including different time-variant and time-invariant variable combinations and a fixed effects model in chapter 6. Finally, chapter 7 concludes.

Chapter 2

Theoretical foundations

In social sciences survey data typically is afflicted by missing values. Analyzing this data and ignoring the missing values can lead to invalid statistical inference. Multiple imputation is the preferential treatment for missing values due to item nonresponse at the moment. It was invented and comprehensively shown in Rubin (1977, 1978, 1987) and many more. This chapter is at first an introduction to the multiple imputation theory. Additionally, classification and regression trees are introduced, as they ease some of the difficulties that emerge when standard multiple imputation approaches are used on complex data containing not only continuous variables, as described by Burgette & Reiter (2010). Besides the usage of CART in combination with MI, the usage in combination with data augmentation is presented as an alternative possibility to get statistically valid inference from data with missing values. Both, MI in its common application by chained equations and the iterative CART have special advantages. However, both approaches in general have limitations to address the special high-dimensional complex survey design occuring within the NEPS.

2.1 Missing Data Mechanisms and Ignorability of missing values

As a starting point, it is assumed that values are not only missing, but are missing for a reason. The reasons for the appearance of missings can be manifold. Impor-

tant for the decision whether to impute data is the question how the mechanisms (reasons) for missing values influence the analysis of observed data.
Therefore, Rubin (1976) as well as Little & Rubin (2002, pp.14-17, 89f) defined three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR).
For clarification of these three missing data mechanisms let Y be a $n \times p$ data matrix with $i = 1, \dots, n$ individuals and $j = 1, \dots, p$ variables. These variables are partially observed, that is Y_{obs}, and partially not observed (missing), that is Y_{mis}, so that $Y = [Y_{obs}, Y_{mis}]$. Another matrix R with elements r_{ij} indicates whether an element y_{ij} is missing ($r_{ij} = 0$) or not ($r_{ij} = 1$). The matrix R is called response indicator, see e.g.Rubin (1987, p.30). A simplified illustration to explain the usage of both matrices can be seen in figure 2.1.

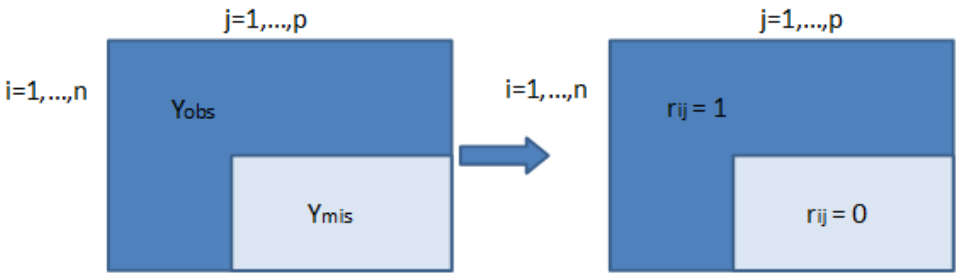

Figure 2.1: Matrices Y and R indicating observed and missing values

On the left side there is the matrix as we usually see it when we look at survey data. There are cells that are observed, illustrated with darker blue. An example of a value in one of these cells is '32' in the first row and first column. This number includes the information that the first person ($i = 1$) was asked about his age ($j = 1$) and answered '32 years'. The lighter blue cells indicate values that are not observed. For example a person $i = 40$ did not answer about the household net income $j = 10$. The fortieth row and tenth column would be empty or have a label for a missing value. The matrix Y is the matrix that is used as the basis for analysis of missingness. The matrix R replaces each concrete (observed) value of Y_{obs} with a one and each missing value, that is Y_{mis}, with a zero. This matrix R can be used to get an idea of the structure of the missings

within the data. Additionally, combining both matrices helps to understand the definition of missing data mechanisms and how R can be explained with Y.

MCAR occurs when Y_{obs} and Y_{mis} can be interpreted as a random subsample of all cases. That means that the missingness does not depend on any other variables or the variable containing missing values itself. The conditional distribution $f(R|Y, \psi)$, where ψ describes unknown parameters, can then be reduced to $f(R|\psi)$. In a more detailed notation the equation can be written as $f(R = 0|Y_{obs}, Y_{mis}, \psi) = f(R = 0|\psi)$.

MAR occurs when the Y_{mis} depend on observed values in the data, that is Y_{obs}, but not on missing values, Y_{mis}. The conditional distribution $f(R|Y, \psi)$ can then be written as $f(R|Y_{obs}, \psi)$. Typically, multiple imputation is conducted when the missing data mechanism is assumed to be MCAR or MAR, where the MCAR assumption is testable, see e.g. Little (1988) and the MAR assumption in general is not, see e.g. Glynn et al. (1993), Graham & Donaldson (1993) and Little & Rubin (2002, chapter 11).

"The observed data are observed at random (OAR) if for each possible value of the missing data and the parameter ψ, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, is the same for all possible values of the observed data.", see Rubin (1976, p.582). When the missing data are MAR and the observed data are OAR, the missing data can be described as MCAR, see Little & Rubin (2002, p.14)

NMAR occurs when the missingness can not be explained by the observed data itself. So the Y_{mis} depend on nonobserved values, the dropout is informative or non-ignorable, see Diggle & Kenward (1994). The conditional distribution $f(R|Y, \psi)$ can not be simplified. Assumptions have to be made by the scientist to handle data under NMAR which can be based on "scientific understanding or related data from other surveys", see Rubin (1987, p.202). Furthermore, Rubin (1987, p.202) stresses that it is important to display the sensitivity under different assumptions for the response mechanism when analyzing data under the

NMAR assumption.
A very catchy description of missing data mechanisms can be found in Koller-Meinfelder (2009). Detailed information about missing data mechanisms in social science data and ways to work with them can be found in Little & Rubin (1989).

In the context of missing data mechanisms two important concepts have to be explained: distinctness and ignorability. Distinctness is formally defined as $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$. "From the perspective of a Bayesian statistician this means that the joint prior distribution can be split into the product of the marginal prior distributions.", see Koller-Meinfelder (2009, p.4f). θ is the unknown (vector) parameter that steers the distribution of Y, in other words, the 'explanatory variable' of the data analyst.
Moreover, according to Rubin (1976) and Little & Rubin (2002, p.90) the requirements for multiple imputation are that the missing data should be missing at random, the observed data should be observed at random and the parameter of the missing data process should be distinct from the parameter θ which steers the distribution of Y. Combining those requirements the missing data mechanism should be ignorable. Thus, θ can be estimated without modelling the missing data mechanism explicitely, that is the model for R, see for example Little & Zanganeh (2013, p.2).
The following describes typical imputation methods with focus on multiple imputation.

2.2 Imputation approaches

When missing values occur there are some ways to handle them. The most common are: Ignore them or impute them. Ignoring them by deletion of the affected cases is the default way of handling missing values in most statistical programms (van Buuren, 2012, p.8). Listwise deletion (each case with at least one missing value is deleted) and pairwise deletion (each case that is needed for the actual analysis with at least one missing value is deleted) are options when the missing data mechanism is MCAR, as it leads to unbiased estimates for the reduced data. Compared to the original data, that include missing values, the standard errors

and significance levels for the subset of the data are often larger, see van Buuren (2012, p.8).
Imputation is a collective term that includes all techniques that replace a missing value with one or several predicted ones. About 30 years ago Sedransk (1985, p.451) ended a conference proceedings with the result that "whenever possible, model the missing data process, do a complete data analysis and avoid imputations". Other researchers, as for example Sande (1982), came to similar results. Time has changed: imputation techniques were modified and became a common method. The requirements on imputation techniques are summarized by Rubin (1987), but already denoted in Rubin (1978): standard complete-data analysis methods, valid inference, display of the sensitivity of inferences. Most of the procedures allow for the mentioned standard complete-data methods, but lack for valid inferences and the display of the sensitivity of inferences.
Multiple imputation is a technique that takes the uncertainity of missing values into account and differs concerning this matter clearly from single imputation. Nevertheless, for a better illustration and as single imputation is sufficient and reasonable in some cases, see e.g. Rao & Shao (1992), it is illustrated as well.

2.2.1 Single Imputation

There are several approaches to replace a missing value with a single value based on different assumptions about the absence of data. Deductive imputation is based on logical values, for example it is easy to understand that 'yes' or 'no' can be imputed to the question whether a women has children or not if she answered the question 'How many children do you have?'. Mean imputation replaces each missing value with the mean of the variable. It undererstimates the variance and biases almost every estimate besides the mean even when the missing data mechanism is MCAR. Hot-deck imputations use information of donor units, chosen for example sequentially, randomly or by nearest-neighbor approaches. For a general discussion of hot-decks, see Andridge & Little (2010). Cold-deck imputation uses information from previous time points (last observation carried forward or baseline observation carried forward), for example from a previous wave in a panel study. Another approach is regression imputation. The observed values are used

within a model and the predicted values of the fitted model serve as imputed values. The last can be unbiased even under MAR when the variables that steer the missingness are included within the regression model. For more information about those procedures see for example Lohr (2009, chapter 8.6), Little & Rubin (2002, p.61) or van Buuren (2012, p.8-13).
In general, the 'best' single imputation method is seen in the stochastic regression imputation, which produces reasonable results, even under the MAR assumption. See for a description and a comparison to other methods Schafer & Graham (2002, p.159-162). The approach is identical to the regression imputation, besides a residual error is added to the predicted values. For a standard linear model that error is normal distributed with mean zero and the variance estimated by residual mean square from the model, see Schafer & Graham (2002, p.159).
Single imputation techniques allow for analysis with standard complete-data techniques, but standard complete-data techniques do not differentiate between observed and imputed values. Inference can be biased and the variability that is caused by missing values is not taken into account. The latter causes bias on estimates which depend on that variability as e.g. correlations or p-values, compare e.g. Rubin (1987, p.12-15) and K.-H. Li et al. (1991). Hence, we focus on multiple imputation, as it "retains the virtues of single imputation and corrects its major flaws", see Rubin (1987, p.15).

2.2.2 Multiple Imputation

The idea of multiple imputation is to replace missing values by a set of plausible values drawn from the posterior predictive distribution of the missing data given the observed. Thus, a propability model is needed on the complete data: $Y_{mis} \sim f(Y_{mis}|Y_{obs})$, compare Schafer & Olsen (1998, p.550). As it is often too complex to draw from $f(Y_{mis}|Y_{obs})$ directly, a two-step procedure can be used: θ is drawn according to its observed data posterior distribution $f(\theta|Y_{obs})$. Then the Y_{mis} are drawn according to their conditional predictive distribution $f(Y_{mis}|Y_{obs}, \theta)$, as $f(Y_{mis}|Y_{obs}) = \int f(Y_{mis}|Y_{obs}, \theta) f(\theta|Y_{obs}) d\theta$.
It might be too complex to derive $f(\theta|Y_{obs})$ as can be seen for example by

the quote from Schafer & Olsen (1998, p.549): "Except in trivial settings, the probability distributions that one must draw from to produce proper MI's tend to be complicated and intractable". A solution then is to draw from $f(\theta|Y_{obs}, Y_{mis}^{(t)})$ with t as time index which leads to a data augmentation procedure, compare chapter 2.3.3.
In contrast to single imputation, multiple imputation imputes M times with $m = 1, \dots, M$ and $M \geq 2$. Each missing value is then replaced not by a single value, but by a vector. After those M imputations there are M complete(d) data sets on which standard complete-data methods can be applied, see e.g. Rubin (1987, p.15), Little & Rubin (2002, pp. 86-87) and Lohr (2009, chapter 8.6.7). The results of these methods can then be combined by Rubin's combining rules which are described in chapter 2.2.4. In general, multiple imputation has important advantages compared to single imputation: 1) MI is more efficient in estimation when imputations are randomly drawn, 2) due to the variation amongst the M imputations MI takes the additional variability, caused by missing values, into account and 3) MI allows for the display of sensitivity, see Rubin (1987, p.16) and Little & Rubin (2002, pp. 85-86).

2.2.3 Imputation with chained equations

The chained equations approach, see e.g. van Buuren & Oudshoorn (1999) and van Buuren & Groothuis Oudshoorn (2011), also known as fully conditional specification (FCS), see van Buuren (2007), or sequential regressions according to Raghunathan et al. (2001), specifies an individual imputation model, that is typically a univariate general regression model, for each variable with missing values, see Azur et al. (2011). These models are iteratively chained as each dependent variable is used in the following model as one of the explanatory variables, following Little (1992) and Little & Raghunathan (1997). At first, the missing values in all variables are initialized and afterwards the algorithm iteratively runs through all specified (conditional) imputation models. The chained equations are repeated several, say M, times. As each iteration consists of one cycle through all variables considered, the algorithm provides M completely imputed data sets, see van Buuren (2007). Before starting the multiple (sometimes called multivariate)

imputation via chained equation algorithms, the data matrix is arranged to ensure that the number of missing values per variable is ascending which is favorable in terms of convergence. The implementation of chained equation imputations are available for example in *R* (packages *mice* and *mi*), *SAS* (package *IVEware*) and *Stata* (package *ice*). Information about these implementations are available e.g. in van Buuren & Groothuis Oudshoorn (2011), Su et al. (2011), Raghunathan et al. (2010), and Royston (2004), Royston (2005a) and Royston (2005b).

2.2.4 Rubin's combining rules and the efficiency of an estimate based on M imputations

When multiple imputation is conducted with for example $M = 5$ imputations there are five complete(d) data sets that can be analyzed. Instead of choosing one of them, the results are all used in a combined form. Rubin (1987, chapter 3) lays out the following rules for multiple imputation confidence intervals. Let $\widehat{\theta}$ be the estimate of interest. Then the multiple imputation estimate $\hat{\theta}_{MI}$ can be calculated as mean of all estimates $\widehat{\theta}_m$ from each of the M data sets which are interpreted as completely observed for the calculation:

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}_m.$$

The total variance of the multiple imputation estimate that is needed for the width of the confindence interval as well as for tests has to be split into two components, which are the within-imputation variance and the between-imputation variance. The within-imputation variance W is calculated analogous to the estimate above as the mean of the estimated variances for the estimate $\widehat{\theta}$:

$$W = \frac{1}{M} \sum_{m=1}^{M} \widehat{var}(\widehat{\theta}_m).$$

The between-imputation variance B then can be described as a variance of the multiple imputation estimate $\widehat{\theta}_{MI}$, calculated as:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\theta}_m - \widehat{\theta}_{MI})^2.$$

The total variance T is calculated by summing both variances up, taking into account, that when the number of imputations M is increased, the simulation error for $\widehat{\theta}_{MI}$ decreases. Hence, a correction factor is added:

$$T = W + \left(1 + \frac{1}{M}\right) B.$$

Using all this information a multiple imputation confidence interval can be calculated by:

$$\widehat{\theta}_{MI} \pm t_{df} \sqrt{T}$$

with the degrees of freedom (df) for the quantile of the Student's t-distribution calculated by:

$$df = (M-1)\left(1 + \frac{M \cdot W}{(M+1)B}\right).$$

For a high number of imputations ($M \to \infty$) the normal distribution can be used instead of the Student's t-distribution.

Typically, about $M = 5$ imputations are conducted before the results of the analysis of each of the imputed data sets are combined by Rubin's rules. This number of imputations seems pretty low. But when calculating the efficiency of an estimate based on M imputations by Rubin (1987) as:

$$\frac{1}{1 + \frac{\gamma}{M}},$$

where γ is the fraction of missing information given by:

$$\gamma = \frac{\frac{r+2}{df+3}}{r+1} \text{ with } r = \frac{1 + M^{-1}B}{W},$$

it can be seen, that for $M = 5$ the efficiency is higher than 90% for fractions of missing information up to 50%. A table for several M- and γ-values can be seen in Schafer & Olsen (1998). Bodner (2008) who was motivated by Royston (2004) showed an alternative table (page 666) and argued for increased numbers of imputations. His basis was a simulation study with a comparison of interpercentile ranges of 5,000 simulated 95% confidence interval half-widths, null hypothesis significance test p-values and fractions of missing information. These interpercentile ranges were interpreted as measure of variability between independent multiple imputation runs. For 95% confidence interval half-widths those interpercentile ranges for $M = 5$ and $\gamma \leq 0.50$ (the exact values for γ were $0.05, 0.1, 0.2, 0.3, 0.5$) lay between 0.02 and 0.30. These values are high when compared to the ranges of $M = 20$ which lay between 0.01 and 0.10 and very high when compared to interpercentile ranges of $M = 100$ which lay between 0.0 and 0.04. Summarized, the confidence intervals are narrower and more accurate with higher numbers of imputation than with lower. This result is not suprising, but the impact of the differences in accuracy have to be considered when deciding about the amount of imputations in practice.

2.2.5 Using multiple imputation does not make you a wizard

Using multiple imputation seems pretty charming and in a lot of cases it is. For instance, van Buuren (2012, p.25) calls multiple imputation the "best general method to deal with incomplete data in many fields". But still there are some 'disadvantages' or better said limitations and requirements that have to be considered when using it.

According to Rubin (1987, p.17f), compared to single imputation there are three (negligible) disadvantages when using multiple imputation: 1) More work and

knowledge about the procedure is needed, 2) multiply-imputed data sets need more storage space and 3) it is more difficult to analyze them (when the goal to get proper inference is ignored). The impact of these disadvantages depends on the number of imputed values. So when M is high, the impact is high. But as already mentioned in chapter 2.2.4 $M = 5$ is already suitable in most cases. Otherwise as already mentioned, Bodner (2008) recommended a much higher number of imputations which depends on the fraction of missing information that has to be managed. Due to the computer power and mass storage possibilities of our time even those increased numbers can be evaluated as unproblematic.

As already mentioned in chapter 2.1, a basic requirement of multiple imputation is that the missing data mechanism is ignorable, see Rubin (1976) and Little & Rubin (2002, p.90).

As described by Rubin (1987), Rubin (1996) and Allison (2000), the quality of the imputation depends on the 'correct' imputation model and the congruency of the imputation model with the analyst model ('uncongeniality'), see Meng (1994).

According to e.g. Glynn et al. (1993), Graham & Donaldson (1993) and Little & Rubin (2002, chapter 11) it can not be tested whether the missingness mechanism is NMAR or MAR.

2.3 CART used in Multiple Imputation and Data Augmentation

2.3.1 Classification and Regression Trees

Classification and regression trees (CART) were originally used in the machine learning area. Machine learning follows the idea that the computer extracts the algorithm automatically, see Alpaydin (2009, p.2). The statistical usage was adapted by Breiman et al. (1984). CART is a nonparametric algorithm for recursive partition respective to Y (dependent variable). A classification tree is used when Y is categorical (nominal or ordinal), a regression tree when Y is continuous. The basic assumption for nonparametric estimation is not a model, as in

parametric estimation, but the idea "that similar inputs have similar outputs", see Alpaydin (2009, p.185). CART uses nonparametric recursive binary splits to partition the data so for continuous variables the values are split in a group with values less than or equal to the splitpoint ($x \leq x_{splitpoint}$) and a group with values greater than this splitpoint ($x > x_{splitpoint}$). For categorical values two groups of values are defined, one equals a defined group of values (e.g. $x = A \cup B \cup C$) and the other one is defined as the remaining values (e.g. $x = D \cup E$).

In figure 2.2 there is an example of what a regression tree can look like. A classification tree would look very similar giving proportions instead of a mean value. The ovals are value groups that still have to be partitioned (pink, blue and orange). The rectangles (green and yellow) are the final groups that fulfill the stop criterion (explained later), i.e. no further partition is conducted. Those final groups are called 'final nodes', 'end nodes' or 'terminal nodes' whereas the others fields (ovals) are just called 'nodes'. In the presented tree the dependent variable Y is e.g. representing the individual net income. Y is continuous and has a mean of 4,000€ with a total number of respondents of $N = 10,000$. Both, the mean and number of respondents are those of Y before any split is done, where split is a synonym for partition. The whole unpartitioned group of values is on top of the figure in pink.
A variable is chosen for the first split by CART (the rules for splitting will be explained later) which is X_1 in this example. X_1 is a categorical variable with answer choices A, B, C, D and E. In this example X_1 has a split point that divides all respondents of variable Y that answered A, B or C to variable X_1 in one group and all respondents that answered D or E to X_1 in another group.
The group on the left side of the tree (blue), that is the group that answered A, B or C to variable X_1, contains 8,000 respondents and has a mean of 2,500€. The group on the right side of the tree (orange), that is the group that answered D or E to variable X_1, contains 2,000 respondents and has a mean of 10,000€. It is important to understand that the label X_1 is only about the chosen split variable and that the values of X_1 only describe the split point. The variable that is getting more homogenous by the split is still Y and the mean and the number of respondents refer to Y as well.

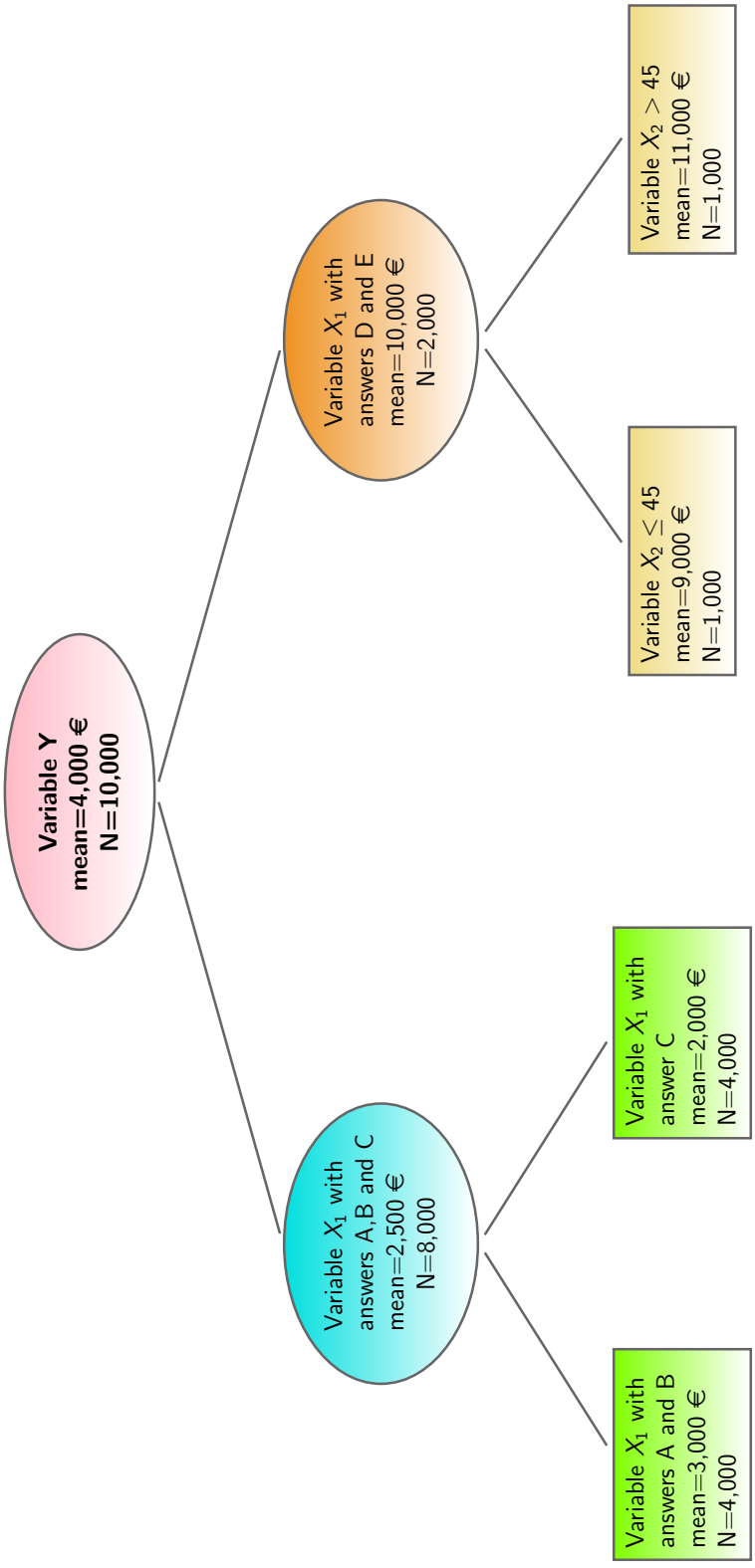

Figure 2.2: Example of a regression tree

The group on the left side (blue) can be divided again using the same variable as another binary split of this variable X_1 is decreasing the heterogenity of the variable Y for the 8,000 respondents of this group better than a split of any other variable. So the group of respondents can be divided by X_1 being answered with A or B in one group and C in the other group. The final nodes of this side are illustrated by the rectangles. Note that the number of splits on each side does not have to be the same. Finally, we have two, according to the stop criterion, most homogenous final nodes on this side with each of them containing 4,000 respondents (the number of respondents does not have to be equal, see second row: 8,000 vs. 2,000 respondents). One group has a mean of 3,000€ whereas the other one has a mean of 2,000€.

On the right side of the tree (orange) we have a group of respondents that answered D or E to variable X_1. The best split to make this group become more homogenous is now a split of a continuous variable X_2 at split point 45. So on the left side (yellow, on the left) we have all respondents of variable Y that answered D or E to variable X_1 and have a maximum of 45 at variable X_2. This group contains 1,000 respondents and has a mean of 9,000€. The second group (yellow, on the right) comprises values of X_2 that are higher than 45. This group is (randomly) as large as the other one and has a mean of 11,000€.
All four rectangle groups (green and yellow) consist of respondents that are as homogenous as they can be within their group and as heterogenous compared to the other groups based on the given split and stop criteria.

The decision whether a binary split is conducted depends on the reduction of heterogenity in the group of values by a possible split. Thus, for all possible split points the least squares deviations for continuous variables or an adequate measure for categorical variables are calculated and the split is conducted at the split point where the reduction of heterogenity is maximized. A threshold defining a minimum reduction of heterogenity serves as stop criterion. So the values within one partition get more homogenous with every split whereas the values across the partitions get more heterogenous compared to each other. According to e.g. Breiman (1996) an adequate measure of heterogenity which

can be used as splitting criterion for categorical variables are the entropy, Gini index or the classification error:

- Entropy: $\sum_{g=1}^{G} -p_g log_2 p_g$
- Gini Index: $1 - \sum_{g=1}^{G} p_g^2$
- Classification error: $1 - max\{p_g\}$

with p_g as relative frequency of an attribute $g = 1, \cdots, G$ within a group.

As an example for the calculation and evaluation, imagine a variable with three attributes, e.g red, blue and green. A possible split leads to the following proportions within one of the two resultings node: 0.2 (red), 0.5 (blue) and 0.3 (green). Another possible split leads to the following proportions within one of the two resulting nodes: 0.1 (red), 0.4 (blue), 0.5 (green). For simplification, only one side of the split is used as basis to calculate the heterogenity. The best split is evaluated by comparing the values of entropy, Gini index or classification error of one possible split with values of these measures of the other possible split(s). The best split is the one with the lowest resulting value, that is the lowest heterogenity.

- Entropy Split 1: $-0.2log_2 0.2 - 0.3log_2 0.3 - 0.5log_2 0.5 = 1.4855$
- Gini Index Split 1: $1 - (0.2^2 + 0.3^2 + 0.5^2) = 0.62$
- Classification error Split 1: $1 - 0.5 = 0.5$
- Entropy Split 2: $-0.1log_2 0.1 - 0.4log_2 0.4 - 0.5log_2 0.5 = 1.3609$
- Gini Index Split 2: $1 - (0.1^2 + 0.4^2 + 0.5^2) = 0.58$
- Classification error Split 2: $1 - 0.5 = 0.5$

As can be seen in this example, the entropy and the Gini index lead to a clear result, that is to prefer the second split as it leads to a lower heterogenity. The classification error does not prefer one or the other.

Alternatively to CART there are other algorithms using different numbers of splits (binary/multiway) and/or other decision rules as splitting criteria. Kim & Loh (2001) outlined some procedures when presenting their algorithm CRUISE. They mentioned CART by Breiman et al. (1984) and QUEST by Loh & Shih (1997) as binary methods. As multiway split methods FACT by Loh & Vanichsetakul (1988), C4.5 by J. R. Quinlan (1992), CHAID by Kass (1980) and FIRM by Hawkins (1997) were mentioned. Note that the algorithm C4.5 is a follower of ID3 which has not separately been mentioned by Kim & Loh (2001), see thereto J. Quinlan (1986). A newer tree-based algorithm, hence not mentioned by Kim & Loh (2001), is CTree by Hothorn et al. (2006). The algorithm DIPOL is a follower of Cal5 by Müller & Wysotzki (1994).
There are many more tree-based algorithms available. Dependent on the structure of the data and the measurement criteria for the performance of those algorithms the results and the consequently preferred algorithm might differ. One of many examples of the results of a performance test can be seen from the creators of DIPOL available at the 'Technische Universität Berlin' website (https://www.ki.tu-berlin.de/menue/team/fritz_wysotzki/cal5_dipol/, date of access: 17.03.2016).

2.3.2 Nonparametric sequential classification and regression trees for multiple imputation

Using MICE, conditional models have to be specified for all variables with missing data, including interactive and nonlinear relations between variables if necessary. However, when knowledge about the conditional distribution is low or appropriate specifications involve high estimation costs, Burgette & Reiter (2010) proposed specifying the full conditional distribution within the MICE algorithm via CART (CART-based MICE). The resulting binary partition of the data along the set of conditioning variables defines the nonparametric characterization of the full conditional distribution. Hence, the final nodes can be used as donor value groups for imputation. All respondents can be assigned to one of these identified donor groups. Each missing value is imputed via a draw from the empirical distribution within this donor group using a Bayesian Bootstrap. Thus, the uncertainty of

the unobserved values is taken into account, see Burgette & Reiter (2010).

As the Bayesian Bootstrap has a very central role it is described in more detail in the following. At first, a tree is built with the non-missing (observed or initialized) data, as for example can be seen in figure 2.2. A missing value in Y with an answer A in X_1 would here be imputed by a value of the green node on the left. A Bayesian Bootstrap gives a random weight or more exact a posterior probability to each of the 4,000 (observed) donor values of this node. Those weights are drawn from a uniform distribution and then are scaled to one.
The steps in more detail and described with an example using four donor values are as follows:

- Draw a value from the uniform distribution (with a minimum of zero and a maximum of one) for the number of donor values minus one, that is for example [0.3, 0.4, 0.24] for a number of four donor values.
- Sort those values from the smallest to the largest, here: [0.24, 0.3, 0.4].
- Create two vectors from these draws: one that is added by one (last position) and one that is added by zero (first position), here: [0.24,0.3,0.4,1] and [0,0.24,0.3,0.4].
- Calculate the first minus the second vector, here: [0.24,0.06,0.1,0.6]. The weights then always sum up to one.
- Use those weights to draw the needed amount of values from the donor values with replacement.

The corresponding *R-Syntax* for CART defined by Burgette & Reiter (2010) can be found in the following. The command is called *bayesianboot* in their provided syntax and is defined as:

```
 a <- sort(runif(length(eligibles) - 1))
values <- sample(eligibles, n, replace = TRUE, c(a, 1) - c(0, a))
```

with *eligibles* as a vector of donor values and *n* as the number of missing values that have to be replaced by donor values from this final node. The theoretical

background of this procedure is explained for example in Rubin (1981). As the Bayesian Bootstrap is based on the Bootstrap which can be interpreted as generalized jackknife further information can be found in Efron (1979) and Miller (1974).

The default of the *tree*-command in the *R-package tree* is a minimum size of 5 units in the final node and a minimum reduction of heterogenity of 0.01 that is needed to conduct a split. Tree growth is limited to a maximum of 31 levels and the amount of levels of a categorical variable to 32 levels. Those settings can all influence the imputation.

2.3.3 Nonparametric sequential classification and regression trees for data augmentation

Data augmentation, as described by Tanner & Wong (1987) or K.-H. Li (1988), is a process to calculate the posterior densities $f(\theta|Y)$ and $f(Y_{mis}|Y_{obs})$ iteratively with θ as the parameter of interest, Y_{obs} as observed data and Y_{mis} as unobserved data with $Y = [Y_{obs}, Y_{mis}]$. Data augmentation can be used for imputation chaining the two steps iteratively. At first, the imputation step imputes values for the missing values by using the information of the estimate θ from the oberserved data. Then θ is 'updated' using the observed and imputed data. Repeating both chained steps, the procedure is a Gibbs Sampler, a Markov chain Monte Carlo (MCMC) method, as described by Geman & Geman (1984), Gelfand & Smith (1990) or Casella & George (1992) among many others.
Formally the procedure can be described as follows, compare for example Schafer (1997, chapter 3.4.2).

1. Starting values either for θ or for Y_{mis}
2. For an arbitrary step t:
   - Imputation-Step: $Y_{mis}^{(t+1)} \sim f(Y_{mis}|Y_{obs}, \theta^{(t)})$
   - Posterior-Step: $\theta^{(t+1)} \sim f(\theta|Y_{obs}, Y_{mis}^{(t+1)})$

with $t+1$ as iterative step following t and anaologous $Y_{mis}^{(t+1)}$ as imputed values of $Y_{mis}^{(t)}$ and $\theta^{(t+1)}$ as imputed values of $\theta^{(t)}$. Let t run until the chain converges to the desired stationary distributions, compare Geweke (1992).

The advantage of MCMC methods is, that no joint multivariate model has to be constructed (but its existence has to be assumed), only the families of conditional models have to be specified, which are already programmed, compare e.g. Liu et al. (2013). Liu et al. (2013) additionally described the conditions under which iterative imputation Markov chain equivalences the posterior distribution of a joint Bayesian model and gave practical implications. Kropko et al. (2014) compared joint and conditional approaches and concluded that one approach does not outperform the other in general and that imputation algorithms should always be chosen appropriate to the characteristics of the data which they are applied to.

The presented data augmentation procedure can be extended using CART. First, the parameter θ is estimated by the observed data. Then, initial values are either drawn unconditionally from the data or drawn from final nodes from an initial (surrogate) tree to fill up the missing values. Afterwards, the Imputation-Step is conducted by a classification or regression tree.
The Imputation-step, as well as the Posterior-step are repeated several (for example, a thousand) times, discarding the burn-in phase. A burn-in phase is therefore defined as an amount of iterations that is ignored for the analysis as it is assumed to be too dependent on the starting values. One implementation is to start a chain with one set of starting values. Only a few iterations from the chain are chosen, the rest is discarded. For this purpose the distance of $L+1$ iterations is defined with L iterations discarded and the iteration of $L+1$ used as imputation draw. Another implementation is to get M different cycles by starting with different starting values. Then, the first L iterations of each cycle are defined as burn-in phase. The remaining iterations of the M cycles can then be analyzed.

2.3.4 A traveling salesman points out some problems

Regarding the partition algorithm, the splitting rule and the evaluation of how well both work, the Traveling Salesman Problem (TSP) has to be mentioned. The TSP is described by Dantzig et al. (1954) with an example of a salesman that has to visit several cities and then returns to the starting point. These cities have in the easiest case different distances between them, and the salesman wants to find out which route is the shortest. So the aim is to find the minimum sum of distances. In a more difficult case, not only the distances, but other variables as for example the traveling costs have to be taken into account. What makes the TSP so outstanding is that it is easy to explain, but hard to solve. According to Dantzig et al. (1954) it is said, that the problem was firstly approached in a seminar talk by Hassler Whitney in 1934. It is not solved until this day. When partitioning the data, one has to start with a splitting criterion. However, the splitting criterion can only be a best splitting criterion for the split that has to be done next. The challenge, that has not been solved yet, is to find the best criterion for all splits, derivated theoretically instead of just using trial and error. Another limitation of CART, as it is usually based on the Gini index, is that variables with many levels are preferred to variables with few, see e.g. Breiman et al. (1984, chapter 4) and Kim & Loh (2001).
In addition, based on conditional models the corresponding joint distribution might not exist, see Si & Reiter (2013). This problem especially manifests in changes in the order of the variables within the tree structure (which can be interpreted as a consequence of TSP) which impacts the imputation, see for example Baccini et al. (2010) and F. Li et al. (2012).

Chapter 3

Analysis of unit nonresponse combining CART and data augmentation

Unit nonresponse, panel attrition in particular, is a problem with a high impact on survey quality (not only) in social sciences, see e.g. Lugtig (2014) or Hillygus & Schnell (2015). Unit nonresponse can be interpreted as a 100% item nonresponse, see e.g. Messingschlager (2012, p.107). For this interpretation auxiliary information which are not surveyed from the target person are not taken into account, see for an illustration figure 3.1.

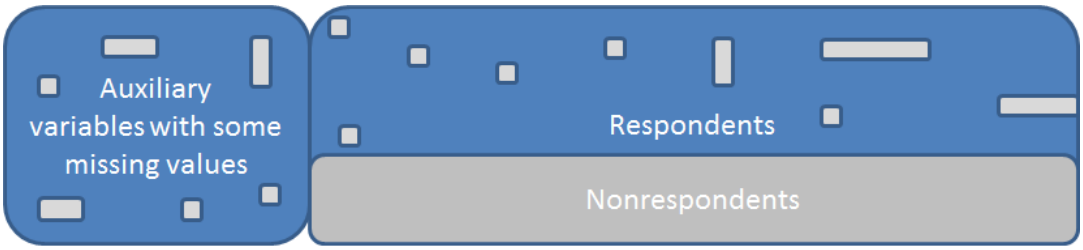

Figure 3.1: Data with auxiliary variables

On the left side there are additional information in the form of auxiliary variables. These can be completely observed (observed values are blue) or limited

by some missing values (missing values are grey). On the right side we have the survey data. They are divided in two parts. One part, that is the upper one, is survey data from the respondents which can have some missings, but are generally observed. The other part, that is the lower one, is the survey data for nonrespondents which are completely missing.
There are many procedures to correct for unit nonresponse, mostly in the field of weighting, see e.g. Little & Vartivarian (2003). Weighting is an approach which is usually applied when the probability to be selected is not the same for all units. Design and base weights (inverse of the inclusion probability) correct for these unequal selection probabilites in the first step, see Kish (1990) and Kish (1992). In a second step, the design weights are adjusted by sample weighting adjustment to correct for unit nonresponse, see Kalton & Kasprzyk (1986). Population weighting adjustment is the third step to correct for potential bias resulting from incomplete coverage, non-coverage or sampling error, see Brick (2013). More detailed information about weighting can be found e.g. in Cochran (1977), Särndal et al. (1992) or Bethlehem (2002).

However, there is an ongoing discussion about whether weighting should be generally applied or whether multiple imputation is an alternative.
Already Rässler & Schnell (2003) compared weighting to multiple imputation when unit nonresponse has to be handled. They concluded with an encouragement to use multiple imputation. Nevertheless, the question whether to weight or to impute is still a topic in the last years' publications. Peytchev (2012) implied that multiple imputation can address (unit) nonresponse and measurement error and so reduce bias and avoid an increasing variance. Messingschlager (2012, p.181) came to the chastening result, that multiple imputation is not always superior compared to weighting, but criticized that weighting leads to unpredictable results. Both, Brick (2013) and Messingschlager (2012, p.182) intended to focus more on the process that leads to unit nonresponse. Little (2013) focused in his discussion on Brick (2013) on the relationship between unit nonresponse and survey outcomes and underlined that the application of weighting procedures is limited. He concluded, that multiple imputation should be used for the correction of item and unit nonresponse and synthetic data sets should be offered.

So the literature is very conflictive. The requirements for both weighting and multiple imputation concerning unit nonresponse are that relevant auxiliary information are available for respondents as well as nonrespondents. Those information can for example be information from the sampling frame, but also from respondents connected to these respondents and nonrespondents (information about third persons). The quality of both, imputation and weighting, depend in their efficiency on the predictive power of the auxiliary variables on the variable of interest, compare respective weighting e.g. Little & Vartivarian (2003, 2005). In the following we distance ourselves a little from the discussion and look at the characteristics of data that are crucial for the decision whether to react for example by weighting, that is whether the participants differ from nonparticipants. Thus, in the remainder of this chapter we use data augmentation combined with CART to get information about the unit nonresponse process. The purpose of this procedure is to decide whether correction methods such as the calculation of nonresponse adjusted weights are necessary for this application. The proceeding is a Markov chain Monte Carlo method, more precisely a Gibbs sampler. The method is shortly described as it is more extensively shown in section 2.3.3. Then, the application on the Thuringia study of the National Educational Panel Study (NEPS) is presented. Section 3.3 concludes this chapter and points out some alternative strategies as well.

Note that the whole procedure, containing an extension on calculating nonresponse adjustment weights and a simulation study about the whole approach is described in detail by Aßmann et al. (2014a). The whole setting is a joint work. The following focuses on my contribution, that is the application to the Thuringia study using CART. All figures and tables about the real data application used for this thesis are identical to those in Aßmann et al. (2014a).
Note as well that the application focuses on the individual level even when the cluster structure, that is schools, is taken into account. The decision whether the requirements to use weights for schools not participating are met, for example based on sampling information, is not part of this illustration. We only focus on the demands of the correction of unit nonresponse on the individual level.

3.1 Nonparametric data augmentation using CART

When auxiliary variables are available, data suffering from unit nonresponse can be imputed for example by using multiple imputation. Furthermore, data augmentation is a possibility, especially when the interest is on the analysis of unit nonresponse, interpreting the response indicator as dependent variable Y and explaining it by the available data. Data augmentation is a MCMC technique, more precisely a two-step Gibbs sampler, see Geman & Geman (1984). Samples from both $f(\theta|Y)$ and $f(Y_{mis}|Y_{obs})$ are drawn iteratively instead of sampling directly from $f(\theta|Y_{obs})$ with $Y = [Y_{obs}, Y_{mis}]$ and θ as unknown parameter steering the distribution of Y. After initializing the data, a Markov chain is performed iteratively drawing $Y_{mis}^{(t+1)}$ from $f(Y_{mis}|Y_{obs}, \theta^{(t)})$ and $\theta^{(t+1)}$ from $f(\theta|Y_{obs}, Y_{mis}^{(t+1)})$, with the values in $t+1$ as the values of the next iteration step with basis t, see for example van Dyk & Meng (2001). Using this data augmentation procedure, the response indicator can be iteratively imputed and an explanatory model can be conducted. Due to the binary nature of the response indicator, a binary logit or probit model is appropriate. As imputation step CART can be applied using the 'updated' information from the data in every chain. An application of this combination of data augmentation and CART can be seen in the following using the Thuringia data from the NEPS.

3.2 (Non)Participants in the Thuringia study of the NEPS

3.2.1 The data

NEPS is a voluntary study in Germany with six starting cohorts (SC1,...,SC6). These six main samples include newborns, Kindergarten children, secondary school children (fifth and ninth grade), first-year undergraduate students and adults. Those starting cohort are accompanied over time, see for a better understanding figure 3.2.

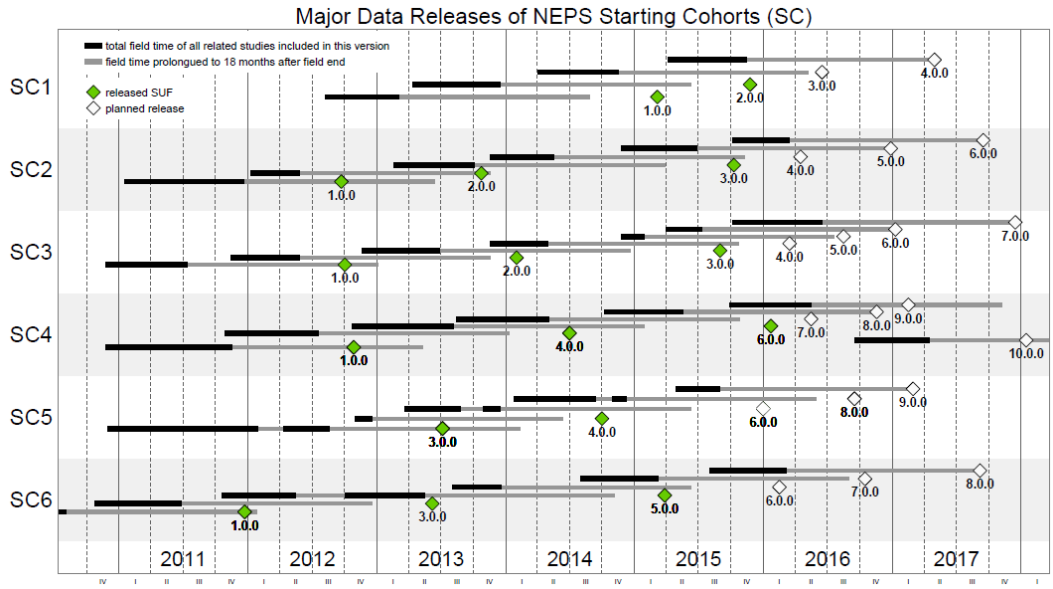

Figure 3.2: NEPS Data Releases, available from https://www.neps-data.de/en-us/datacenter/overviewandassistance/releaseschedule.aspx (Date of download: 22.02.2016)

Beside those starting cohorts there were additional studies to explore the effect of school reforms. One of these additional studies, the organizational reform study in Thuringia, was about the curricular reform of the 'Gymnasiale Oberstufe' (upper Gymnasium level, that is the last years of the upper secondary school) in Thuringia. 32 upper secondary schools in Thuringia took part in 2010 (last year group that was not affected by the reform) and 31 schools in 2011 (first reformed year group). All students of the 12th grade that participated were surveyed and tested once. Achievement tests (Fachleistungstests) in the fields of mathematics, physics, biology and English, questions about the students' social background, a test on cognitive abilities as well as questionnaires were applied. This information can be used to capture possible effects of the reform. Besides, parents and subject teachers were interviewed. In 2010 the number of students within the 12th grade of the 32 upper secondary schools in Thuringia was 1,857. In 2011 the same schools minus the one not participating again contained 1,374 students.

In addition to the information from survey and testing, auxiliary information about the participation status and other related variables were available for all students (participants and nonparticipants). As the focus of this study was on the students, only their participation status was crucial for the definition as non-respondent. Both information, auxiliary infomation and the participation status, should have been combined by the survey research institute. Due to some problems in data collection and transmission in 2010, missing values occured for the participation indicator and the auxiliary information. The missing values that were based on a transmission error can be interpreted as measurement error. Software used in 2011 eliminated most of those problems, so the uncertainty declined. An overview for missing values for both years can be seen in table 3.1.

variable	percent of missings	
	2010	2011
participation status	1.3 %	0.4 %
sex	3.8 %	2.4 %
field of subjects 1	11.9 %	5.0 %
field of subjects 2	12.0 %	5.0 %
field of subjects 3	12.1 %	5.0 %
mean school mark	1.0 %	0.0 %
complete cases	85.0 %	93.6 %

Table 3.1: Overview of missing values

Mosts missings happened for the marks in the fields of subjects (fs). Those fields of subjects were constructs that aggregate the individual marks on related subjects. Students in upper secondary schools in Germany have to chose their subjects within three fields: 1) linguistic-literary-artistic (fs1), 2) social (fs2) and 3) mathematical-natural-scientific-technical (fs3). So each student can choose a preferred combination of subjects, but has to chose a minimum number of subjects within each field. In 2010 about 12% of all information were missing for the three fields of subject. Only 3.8% were missing for sex, 1.3% for participation status and 1.0% for the mean school mark. In 2011, all missings were reduced by the new software resulting in 5.0% missing in the three fields of subject, 2.4% for

sex and 0.4% for the participation status. The mean school mark was completely observed. The missings were not caused by students, but the analysis of the data led to information about the nonresponse process of students.
To evaluate the quality of the data and for further analysis, it was analyzed whether nonrespondents differed compared to respondents. Thereby, the variables available for all students, apart from the information that were missing, could be used to answer this question. A binary probit is appropriate for this requirement. Even though there were just a few missing values the data were imputed since the occured missings might have led to improper inference. The combination of Gibbs sampling and imputation using CART results in a data augmentation approach that will be described in more detail in the following.

3.2.2 Method

Unit nonresponse was handled as item nonresponse manifestated as values of the variable 'participation status' as the binary response indicator. The analysis was performed using a binary probit model. The decision was made for a binary probit model instead of a binary logit, because the distributional assumptions of the binary probit model were more suitable for the analyzed model. The missing values in the data and the coefficients for the probit model were initialized at first. With the completed data CART was applied to replace the starting values with new draws from the donor values (final nodes). The parameters of the binary probit model were renewed as well based on those new values. Both approaches, CART and the binary probit analysis, were alternately conducted combined within a Gibbs-based sampler.
Out of a large list relevant variables have been identified for the analysis. These were the participation status, the information about sex, all marks of all students for the last four semesters (marks from chosen subjects), marks of final exams and the final mean mark.
The individual marks were aggregated as arithmetic mean within the fields of subjects for the analysis as there would have been too many structural missings using the single subject marks. In table C.1 there is an overview of the aggregation for all relevant subjects.

From 2010 to 2011 the rules of subject choice changed, so the data of the mean marks of the students referenced a different calculation basis. Due to the fact that the calculation rule only changed slightly and in order to maintain comparability between the estimations, both mean marks were calculated with the 2010 calculation rule for the analysis.
The students were clustered in schools. So the cluster structure has to be taken into account when CART is used. Only the mean school mark was used as the clustered variable, that is the mean of all marks of all students of the 12th grade for each school. So the mean school mark (as an initialized or updated value) was used in a first (level 1) CART process as additional information. It was the same for all students within a school. Then the aggregated data for all students within a school was used in a second (level 2) CART model and the mean school mark value was updated.

The data situation can be distinguished into four missing value situations that were relevant for the estimation, as can be seen in figure 3.3 with Y as the dependent variable, that is the participation status.

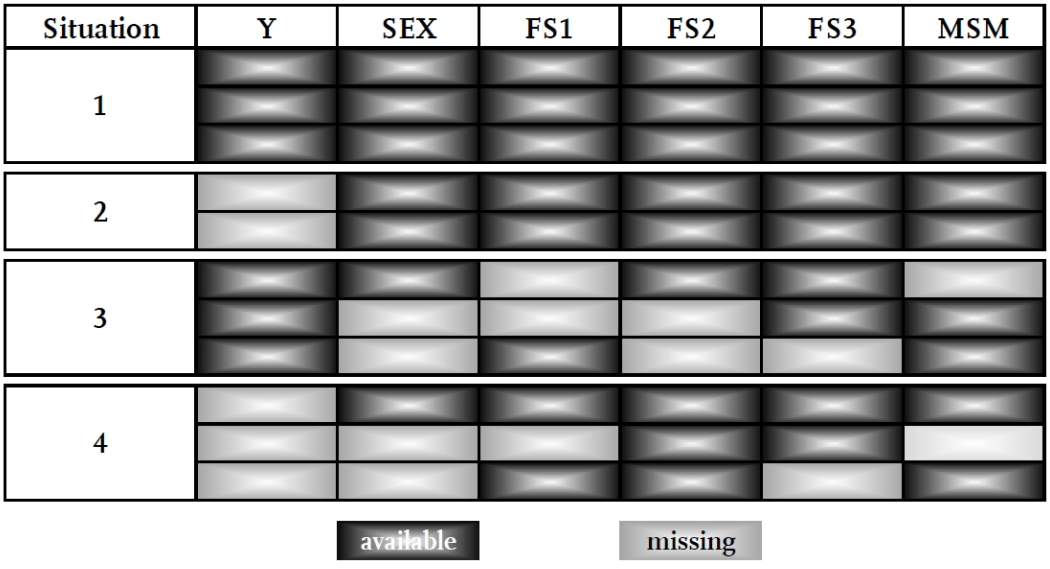

Figure 3.3: Missingness pattern of the Thuringia study data

1) Complete cases: no missing values neither for the participation status, nor in the explanatory variables

2) Complete explanatory variables, but missing values in the participation indicator (measurement error)

3) Complete participation indicator, but missing values in the explanatory variables

4) Missing values in the participation indicator (measurement error) and the explanatory variables

When all variables are complete, as in situation 1), a standard probit regression could be used. For the situations 2) and 4) a so called *Probit Forecast Draw* was used for the participation status which was based on a Metropolis-Hastings algorithm, see Chib & Greenberg (1995). This approach used the information of the maximum number of participants in each class of students conditional on sex. So the participants were drawn from all students within a class, using the information of all explanatory variables (that have to be augmented for situation 4) before) and the information of the maximum number of participants.
For the situations 3) and 4) the missing explanatory variables were augmented by CART.
In the following, the Bayesian Probit model is described. A more detailed description can be found in Aßmann et al. (2014a). y_{ij} were values of a dichotomous dependent variable with $i = 1, \ldots, N_j$ as an index for the students within a school $j = 1, \ldots, J$ with N_j denoting the total number of students of a school and J as number of schools. Whereas the observed variable is binary, a latent variable z_{ij} is assumed which works as link between explaining factors X_{ij} and y_{ij}:

$$y_{ij} = \begin{cases} 1, & \text{if } z_{ij} \geq 0, \\ 0, & \text{if } z_{ij} < 0, \end{cases}$$

where $z_{ij} = X_{ij}\beta + u_j + e_{ij}$ and e_{ij} is an independent identically normal distributed error term with unit variance and u_j a cluster-specific random error term with $\mathcal{N}(0, \sigma_u^2)$.

Pooling hence yields the complete likelihood

$$\mathcal{L}_P(Y|\beta, X, u_j) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} \Phi\left[(2y_{ij} - 1)(X_{ij}\beta + u_j)\right],$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution.

The covariance matrix σ_u^2 of the random coefficients is sampled from independent inverse gamma distributions $\mathcal{IG}(\alpha_{\sigma_u^2}, \beta_{\sigma_u^2})$ with parameters

$$\alpha_{\sigma_u^2} = \frac{J}{2} + \alpha^0_{\sigma_u^2}$$

and

$$\beta_{\sigma_u^2} = \frac{1}{2}\sum_{j=1}^{J} u_j^2 + \beta^0_{\sigma_u^2}$$

where the parameters of the conjugate inverse gamma prior distribution $\mathcal{IG}(\alpha^0_{\sigma_u^2}, \beta^0_{\sigma_u^0})$ are $\alpha^0_{\sigma_u^2} = 1$ and $\beta^0_{\sigma_u^2} = 1$.
As mentioned above, there were four data situations which were relevant for the estimations. All four were handled by an initialization step and a Gibbs Sampler step including the presented Bayesian Probit model. The whole estimation routine can then be described by the following with X_{mis} and X_{obs} representing the missing and observed values of the explanatory variables, Y_{mis} and Y_{obs} representing the missing and observed values of the participation status.

Initialization:

1. Unconditionally draw new values for X_{mis} from X_{obs} (with replacement).
2. Use the maximum likelihood estimation results based on complete cases as starting values for the β coefficients (informative prior for β).
3. Generate one run of the Metropolis-Hastings sequence to draw new values for Y_{mis} (measurement error) based on the complete values from the conducted initialization steps.

Gibbs Sampler:

1. Generate new values for X_{mis} for level 1 and level 2 from full conditional distributions provided by CART analysis.

2. Generate one run of the Metropolis-Hastings sequence to draw values for Y_{mis} (measurement error) based on the complete values from step 2 of the initialization step for $m = 1$ and from step 4 of the preceding iteration for $m > 1$.

3. Generate new random effects variance-components σ_u^2 and u_j.

4. Calculate new β coefficients based on conducted steps of the Gibbs Sampler.

5. Repeat the whole Gibbs procedure M times with iterations $m = 1, \dots, L, \dots, M$ with L as the last iteration of the burn-in phase.

The initialization differed from Burgette & Reiter (2010) where the initialization equaled the imputation step with limited variable range as only completely observed variables were used and stepwise imputed variables were added. As there was no completely observed variable in the application data unconditional draws with replacement were sampled from the observed values.

Following the practical advice of Cowles & Carlin (1996) and Raftery & Lewis (1992) multiple long chains of length $M = 20{,}000$ with various starting values were running. The burn-in phase had to be discarded for more correct estimates at iteration L. Then, the values from the remaining iterations after the burn-in phase had to be combined. The Bayes posterior mean vector of unknown parameters $\hat{\Theta}_m = \{\hat{\beta}, \hat{\sigma}_u^2\}$ was then calculated as the mean of the remaining iterations

$$\overline{\hat{\Theta}} = \frac{1}{M - L} \sum_{m=L+1}^{M} \hat{\Theta}_m.$$

3.2.3 Empirical results

The data from the Thuringia study of the NEPS was analyzed in three steps. First, the complete cases were analyzed with a standard binary probit model. Second, a second level random effect was added taking the multilevel structure of the data into account. Third, the data were augmented using CART as imputation step in combination with the binary probit model with the second level random effect.

The results of the complete case analysis led to the suggestion that there was a selection effect caused by nonrespondents. In the upper part of table C.2 on the left there are the results for 2010. The confidence intervals (no null contained) showed that the participation status depends significantly on the marks of fs1, i.e. German, English, arts and music, fs3, i.e. maths, physics, biology and computer sciences and on the mean school mark (msm). In 2011 the participation status depended significantly on sex and the marks of fs3 as well.

Adding a second level random effect changed the results for the school completely. In 2010 there were no significant effects on the participation status and in 2011 only the random effect showed a significant effect. So on the individual level there was no effect, but concerning the homogeneous context within schools the students differed in their participation.

Still, it could not be excluded that due to the uncertainty stemming from unit and item nonresponse the effects change. So the data were initialized and a Gibbs sampler was conducted with a 5,000 inital iteration burn-in phase. Another 15,000 iterations were the basis for the calculation of the relevant estimates. The whole Gibbs sequence showed a good mixing behavior, see figure B.1 and the autocorrelation function (ACF) had only moderate dependencies up to the last 10 iterations for sex and the three mean marks fs1, fs2 and fs3, see figure B.2. As shown in C.3 results were independent of the chosen prior specifications and did not change substantially. With both good mixing behaviour and only moderate dependencies in the autocorrelation function given the results could be interpreted. The marginal effects are presented in table 3.2.

	2010				2011			
(I.1) Gibbs Cart MH P = 0.01					(I.2) Gibbs Cart MH P = 0.01			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	0.3180	0.3462	-0.3587	0.9947	-0.1926	0.5988	-1.3899	0.9858
sex	-0.0426	0.0209	-0.0830	-0.0016	-0.0472	0.0238	-0.0936	-0.0004
fb1	-0.0154	0.0089	-0.0329	0.0021	0.0069	0.0098	-0.0118	0.0264
fb2	0.0051	0.0082	-0.0110	0.0212	-0.0024	0.0089	-0.0201	0.0147
fb3	0.0103	0.0058	-0.0012	0.0214	0.0025	0.0060	-0.0094	0.0141
msm	-0.0425	0.1476	-0.3356	0.2396	0.1304	0.2727	-0.4065	0.6803
(II.1) Gibbs Cart MH P = 0.02					(II.2) Gibbs Cart MH P = 0.02			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	0.3005	0.3399	-0.3721	0.9807	-0.1420	0.5981	-1.3144	0.9957
sex	-0.0427	0.0210	-0.0843	-0.0020	-0.0475	0.0242	-0.0955	-0.0001
fb1	-0.0151	0.0088	-0.0322	0.0024	0.0069	0.0098	-0.0122	0.0266
fb2	0.0047	0.0081	-0.0110	0.0206	-0.0028	0.0089	-0.0201	0.0147
fb3	0.0105	0.0058	-0.0009	0.0217	0.0025	0.0060	-0.0093	0.0141
msm	-0.0351	0.1459	-0.3267	0.2510	0.1078	0.2710	-0.4148	0.6345
(III.1) Gibbs Cart MH P = 0.05					(III.2) Gibbs Cart MH P = 0.05			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	0.3206	0.3513	-0.3640	1.0114	-0.1219	0.5709	-1.2220	1.0239
sex	-0.0421	0.0210	-0.0832	-0.0010	-0.0470	0.0244	-0.0954	0.0007
fb1	-0.0151	0.0089	-0.0324	0.0027	0.0070	0.0098	-0.0120	0.0264
fb2	0.0049	0.0081	-0.0110	0.0210	-0.0026	0.0089	-0.0202	0.0149
fb3	0.0103	0.0057	-0.0009	0.0214	0.0023	0.0060	-0.0097	0.0139
msm	-0.0442	0.1498	-0.3362	0.2473	0.1010	0.2575	-0.4175	0.5945

Table 3.2: Marginal effects of the Bayesian Probit estimation with different prior precision; Note: Initial 5,000 draws were discarded for burn-in, MH: Metropolis-Hastings algorithm

Sex had a very small effect on the participation status for both years with female students more likely to take part than male students. In the model with a prior precision of $P = 0.05$ for 2011 the effect of sex was not significant which was considered as a random variation due the estimation process. This assumption was furthermore supported by the corresponding high density region, which included the null.

Summarizing all these results based on the Gibbs sampling procedure an effect of variable sex was found in all models except for the mentioned model with prior precision $P = 0.05$. The results differed completely from the other two procedures without correcting for unit and item nonresponse. The effect of sex on the participation status was very small, but had to be regarded. When weighting is conducted the nonresponse adjustment weights should correct for the selectivity of the variable sex.

3.3 Conclusion and differences to the original approach

In most surveys unit and item nonresponse occurs. Often weighting is used to correct for unit nonresponse whereas multiple imputation is mostly used for item nonresponse. This chapter was concerned about the analyis of unit nonresponse using CART combined with a Bayesian Probit analysis as the data augmentation procedure, but also marking out the general discussion about weighting and multiple imputation as an alternative. As shown, the results differ a lot comparing the complete cases analysis and the augmented data. The CART algorithm of Burgette & Reiter (2010) was extended and adapted for the current application. It used empirical values as donors for the missing values, searching for 'close to the truth'-values by structuring the data with classification or regression trees and drawing one of the donor values with a Bayesian Bootstrap from corresponding nodes (that had the same predictive distribution), as extensively described in chapter 2.3.2. It was necessary to start this process by defining initial values. The approach of Burgette & Reiter (2010) used the circumstance that the tree structure is based on complete(d) variables, drawing the initial values from complete or stepwise imputed variables. The problem of the implementation to real data is that sometimes there are no completely observed variables available as in the application with the Thuringia study data, so the creation of a tree gets impossible. Consequently, the initial values were drawn unconditionally with replacement from the observed values for each variable. The disadvantage of this initialization is, that it invites an imputer the use of CART for every missing rate. Theoretically, using samples from the observed values without replacement allows to use CART even when there is only one value available for each variable and all others are missing, see for a better understanding figure 3.4. The data on the left side is the empirical data. The darker blue fields are observed values, the light blue fields are missing values. Most of the data are missing for each variable. The available data of each variable, that are the observed values, is used to complete the missing values of the data. The completed data, displayed on the right side of the figure consits of the observed values (dark blue) and their unconditionally drawn replicates (middle-dark blue). Using this data for further

analysis is possible, but not recommendable.

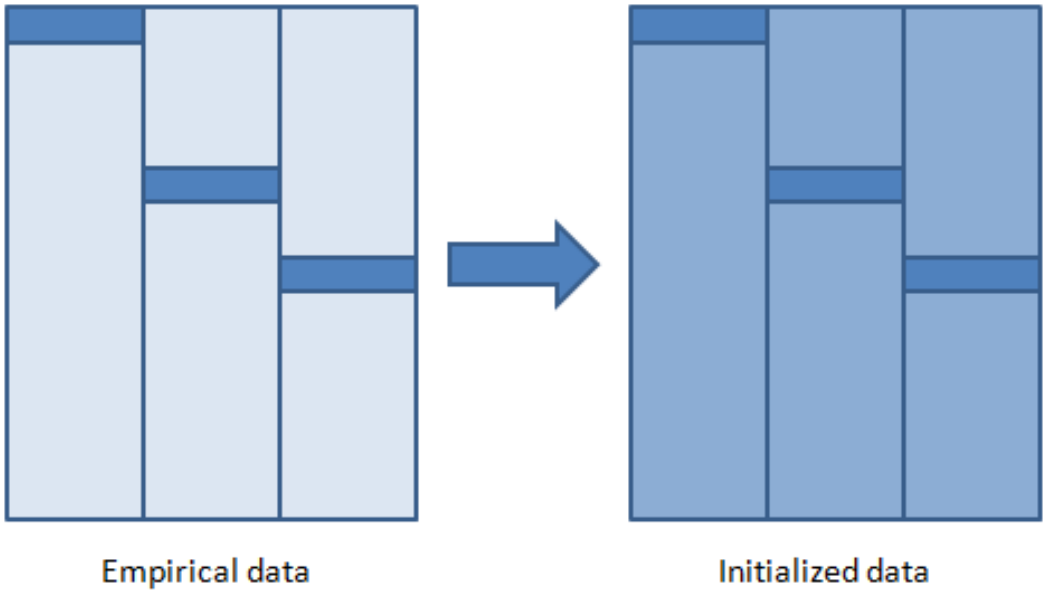

Figure 3.4: Empirical and initialized data

Only with a moderate missing rate it can be assured that the structure of the data can be 'restocked' using CART within the presented Markov chains.
Another slight change from the approach of Burgette & Reiter (2010) is that the cluster structure in the data was taken into account by chaining two CART-steps, one for the individual data and one for the clustered data which was the school context. For both CART models the data from the other level was used but not modified itself. For richer data on the second level (the Thuringia data has only one variable on the school level) alternatives can be considered.
The most important extension and adaption was, that CART was not used as a single approach to impute the data, but within a Gibbs sampler in combination with a Bayesian Probit analysis as a data augmentation procedure. As the usage of CART is very rare in contrast to alternative parametric procedures at the moment, the presented algorithm shows one of many further applications based on combinations of CART with standard algorithms.

The analysis of unit nonresponse was needed in the application of NEPS data to decide whether the necessity for nonresponse adjustment weights was given or not. That is whether the data included a selectivity of nonrespondents compared to nonrespondents. The selectivity was given by the variable sex. For both years, female students tended to participate more often than male. When weighting is conducted the nonresponse adjustment weights should correct for that selectivity. With a higher amount of auxiliary information it can in general be suggested to impute the item nonresponse concerning the uncertainty of unit nonresponse, compare Rässler & Schnell (2003).

As can be seen in table 3.1 the fraction of missing values was very low in 2011 with up to 5.0% missing values. In contrast to that the missing values in 2010 were up to 12.1%. That allows for the suggestion that most of the missings were caused by the transmission error in 2010. As the transmission error effects the data randomly, the missigness mechanism was missing completely at random for those values. Consequently, only about up to 5% of the variable values might be not missing completely at random. On the one hand, it can be doubted if this extensive approach was needed to impute the data and get information about the nonresponse process. On the other hand, it was a good access to the subject of embedding CART within a MCMC approach.

Chapter 4

Nonparametric imputation of high-dimensional data containing filters

The most common application for multiple imputation is for item nonresponse, see e.g. Rässler & Schnell (2003). Item nonresponse is the absence of a variable's value which can for example occur when respondents decide not to answer a question or questions are filtered. In the current literature 'skip patterns' is sometimes used as an alternative term for the patterns of missing values that occur by filters. When filtering is integrated in surveys, the questionnaire can be individualized and consequently increase in quality as no inappropriate questions are asked and as the response burden reduces as the survey becomes shorter and irrelevant questions are avoided, see e.g. Bosley et al. (1999). Besides these advantages of using filters, there are disadvantages as for example different amounts of responses for each variable even when the question was answered by each person to whom it was relevant. Additionally, a coding has to be used that makes it possible for the analyst to see which missing values occured for what reason. The analyst has to decide how to handle these missings, as for example when the data has to be imputed. Below, a short overview is given about some codings that can be used for filter-concerned values. In this overview, not the 'name' of the missing is relevant, that is whether a value is coded as 'NA' or

'-99', but the way how the value's missing by filters is handled due to this coding.

NA: NA is the acronym for 'not available' and hence can be problematic for the analyst if there is no more differentiation from conscious refusal. This can be prevented with an *additional variable* that indicates the status of the original variable's values as answered, not answered or irrelevant. However, an additional variable for each original variable leads to a high data volume.

NR: NR is the acronym for 'not relevant' and is an *additional category* for variables that are filtered. Additional variable categories can be problematic when the data are analyzed, especially when variables are continuous. In contrast to NA-coded values the implemented exclusion, e.g. by the subcommand *na.rm=TRUE* in the statistical program *R*, does not automatically recognize the value NR as missing value when e.g. a mean is calculated.

Zero: Some values can be set to *zero* as the logical consequence of the filter answer. So when a person did not get a Christmas bonus, the height of the Christmas bonus was zero. This can lead to a bias, when for example a mean for Christmas bonus receivers should be calculated as the mean calculation only considers the values of the Christmas bonus not the filtering. Consequently, the filtering has to be regarded when analyzing and imputing the data as the original distribution is changed.

Concerning multiple imputation, filtering forces the data imputer to regard the limited value bounds, i.e. to regard the variables' different value ranges conditional on values of steering variables. The presentation of the implementation of CART-based MICE handling this task on empirical data is the objective of this chapter.

The remainder of this chapter is structured as follows. First, various types of filters are introduced. Then, the imputation method considering the filtering is illustrated. Afterwards, an application on data from the NEPS relevant for the analysis of household net income is described. The chapter concludes with a

short summary and mentioning alternative ways of handling filters when the data are imputed.
Note that the application on NEPS data is described in detail by Aßmann et al. (2014b) and Aßmann et al. (2015). Since the whole setting is a joint work, the chapter focuses on my contribution, that is using CART with data containing filters. Only steps relevant for this thesis are mentioned, thus, the description of the data and the results are shortened. The figures and tables used in section 4.3.1 show identical content to those published by Aßmann et al. (2015).

4.1 Imputation of data with filters

The usage of filters is common, especially in large-scale surveys. Filters can make a long questionnaire shorter and more individual and consequently reduce survey costs and response burdens, see e.g. Bosley et al. (1999). Usually questions in large surveys are arranged by topic, see for illustration figure 4.1.

Figure 4.1: Pathway through a survey with questions arranged by topic

Usually, those topics are arranged by response burden. In the beginning there is a short introduction and the questionnaire starts with harmless questions. Then, the response burden increases over time, compare Schnell et al. (2011, pp. 336-339).

There are filters that tackle single or several questions in each of these topics, but there are also filters that steer whether all questions from a topic, a so called module, are asked or not. Four main types of filters can be distinguished:

Type 1: Figure 4.2 shows the most simple type of filters. No other question within the same or another module, nor a different module as a whole is affected. There is a single question that is asked containing a 'not relevant' (NR) category. This type of coding answers can be interpreted as filtering, because it is the shortform of asking if a question is relevant or not in one question and then asking the question of interest. An example would be to ask only 'What amount of money did you get for your Christmas bonus last year?' and add a 'Did not get a Christmas bonus.' as an answer category. The 'NR' option is problematic when used as category within continuous questions, because the scale of measurement changes. In this example the NR category can be converted to a value, that is zero. In contrast, a refusal of this question can not be interpreted as a zero.

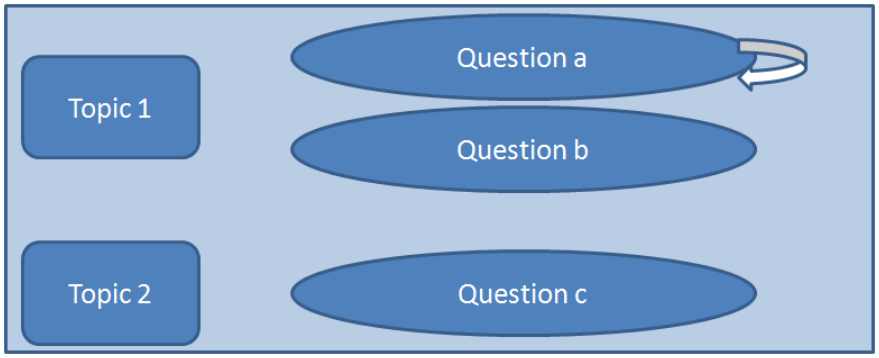

Figure 4.2: Filter type 1, only affecting the same variable

Type 2: As shown in figure 4.3, type 2 is the extension of type 1: At first a question is asked concerning the relevance of the question of interest, e.g. 'Did you get a Christmas bonus last year?' and then a second question 'What amount of money was it?' is asked. This type of filtering can also include connections of different referenced questions, e.g. 'Are you working full-time, part-time or not working at all?' and 'Are you getting unemployment compensation?', where only persons who are unemployed are asked the second question. This connection can happen between one to one or one to many questions.

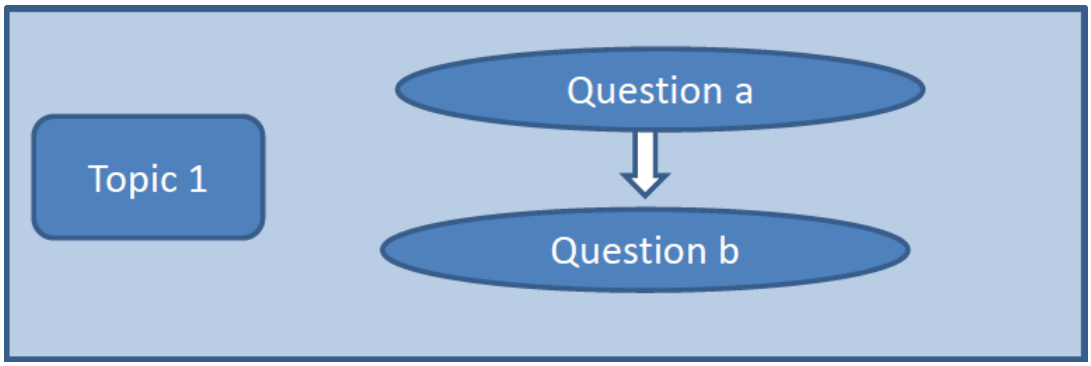

Figure 4.3: Filter type 2, affecting a variable of the same topic

Type 3: Figure 4.4 shows a filter affecting a variable that concerns a different topic. For example topic 1 could be a collection of questions from the children module, where topic 2 could be about income. The question about the amount of child allowance a person gets (topic 2) would only be asked to people living together with at least one child (topic 1).

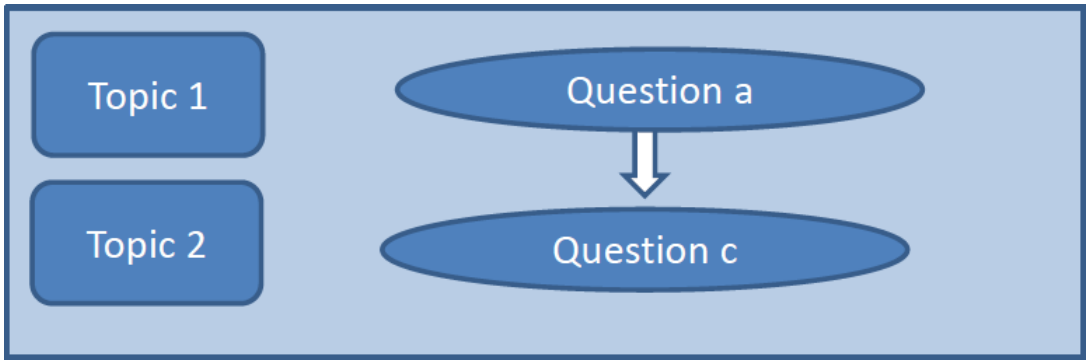

Figure 4.4: Filter type 3, affecting a variable of another topic

Type 4: As shown in figure 4.5, type 4 indicates whether a whole module is asked. Filters of type 4 are an extended version of filters of type 3. An example would be that the marital status is asked as a question of the topic module sociodemography and there is a complete topic module on partners.

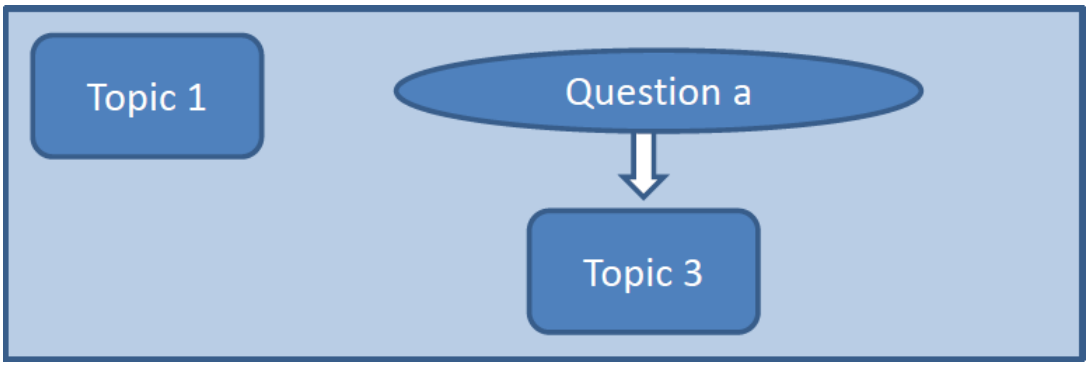

Figure 4.5: Filter type 4, affecting a whole topic module

Additionally to these four main types of filters, two procedures that create missings by design can be mentioned.

Wave-specific: Additionally to programmed filters, different questions can be asked over time when the survey is a longitudinal survey. Those questions can be revised, extended, deleted or skipped for certain waves. Then the interviewed persons are asked different questions or different sets of questions at different time points.

Validity check: A validity check of the answers can be interpreted as a filter. Given answers are checked for their consistency and can be changed to other values or be set to a missing. For example, the net income for a person that is not self-employed should be lower then the gross income. If it is higher, the person perhaps confused net and gross income. The value can then be changed as for example by substituting the net value by the gross and the other way round or both can be set to a missing value.

When filters are considered when multiply imputing missing values, they influence the imputation procedure by changing the admissible range of values that can be imputed. Type 1 ('NR' category) allows only for imputation models that can handle categorical variables, except the case that 'NR' can be set to the value zero. Types 2-4 need to be taken into account when the filtered variable has to be imputed. Type 4 reduces the number of observations for a whole module. Survey questions that change over time or are skipped, for example questions which are

asked only in each third wave, make it difficult to decide whether and how to include additional information in the imputation process from other time points. Whether to conduct a validity check or not is a very general decision, influencing the number of missing values and the consistency of the relations of the variables.

4.2 Nonparametric imputation using CART allowing for a complex filter structure

The basis of the following imputation strategy was the approach of Burgette & Reiter (2010) in which the parametric model of the imputation step of MICE was replaced by a classification or regression tree. At first sight, one might think that CART is able to consider the filter structure by itself. But the criterion for a split is only the homogenity (more precisely the reduction of heterogenity), so 'unlogical' splits can happen when the relation between the variables combined within filter restrictions are overruled by the homogenity criterion. Especially with a *high-complex filter structure* the risk of these unlogical splits exists. Additionally, a variance of values within the final nodes is accepted. The draws via Bayesian Bootstrap might include values that do not fit the filter logic. Therefore, the filters within the data have to be considered to limit the admissible range of donors.
To implement the filter structure in the approach of Burgette & Reiter (2010) the initialization step has to regard the filter hierarchy. Thus, the draws from the observed values with replacement are not unconditional any longer, but unconditional within the limitations of filters. Here, the order of variables has to equal the hierarchical order of the filter structure. The filter hierarchy has to be provided with the data or has to be reproduced by the imputer. More about the use of draws from the unconditional observations instead of tree-based draws can be found in chapter 3.3.

The filter hierarchy can look like in figure 4.6, where there are one or many filters influencing the values of a variable. Those filter variables (questions a,b and

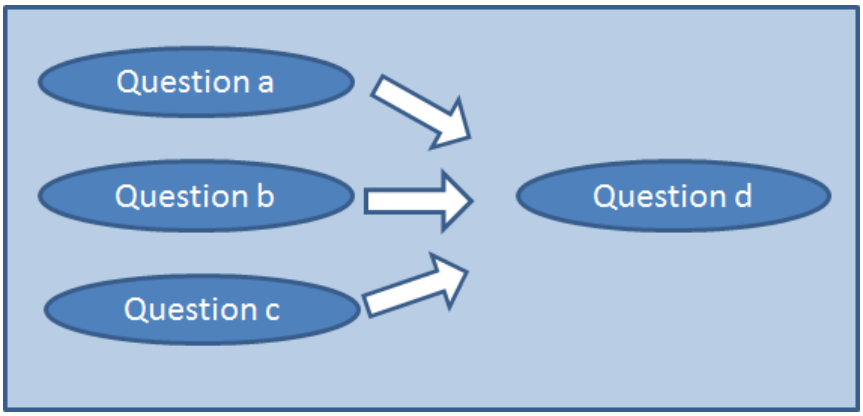

Figure 4.6: Unchained filters influencing one variable

c) have to be imputed first when they have missing values. An example is the question "Do you get an annual bonus?" which is only asked to people who have an employment and are working a minimum number of months.
The filter hierarchy of figure 4.7 is more complex as it shows a higher filter-depth. First, missing values of the higher-order filter variable have to be imputed. In other words, missing values of question a have to be imputed first, as they steer the values of question b. Second, missing values of the next filter variable, question b, have to be imputed conditional on the observed and imputed values of question a. Third and last, the variable which is no filter variable itself, question c, has to be imputed conditional on the observed and imputed values of question b. An example of question c is "Since when does your partner live in Germany?" which is only asked to people who have a partner which would be the content of question a and people who answered "No" to "Was your partner born in Germany?" as question b.

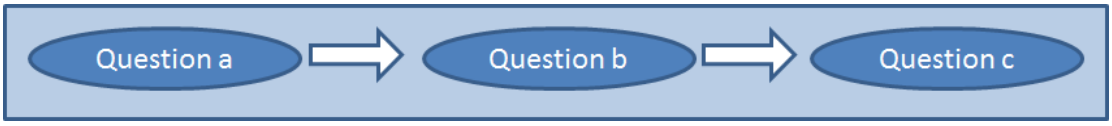

Figure 4.7: Chained filters influencing one variable

After initialization, the CART algorithm imputes the values of the former missing values sequentially, regarding the hierarchical order of the filter structure. When using MICE usually an ascending order corresponding to the amount of missing values is necessary. This order is potentially changed.
The sequential imputation by CART is repeated $L + M$ times with L being the number of iterations needed to mitigate the effect of initialization (burn-in phase).

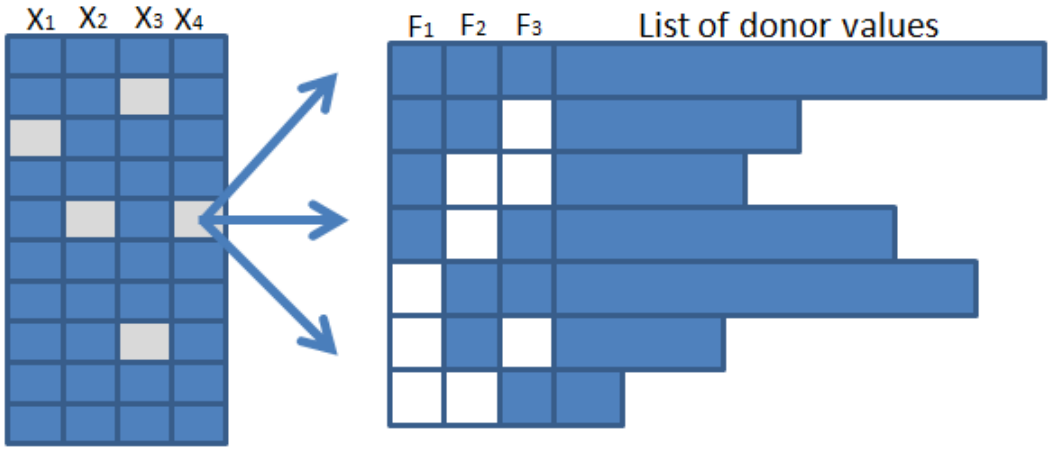

Figure 4.8: Imputation of missing values when there is a filter hierarchy to be regarded

A simplified illustration of how the missing values can be imputed regarding the filter structure can be seen in figure 4.8. There are four variables that have to be imputed: X_1, X_2, X_3 and X_4 as they include not only observed (blue), but also missing (grey) values. The first three variables are not steered by filters, thus, they are initialized by a random draw from the unconditional empirical distribution of the observed values of this variable. Likewise, the draws via Bayesian Bootstrap from a final node for the following iterations are not changed. For the fourth variable, that is X_4, there are three filters (F_1, F_2 and F_3) that steer the admissible values. The simplest filter rule would be a dichotomous one, here illustrated by the two colors blue and white, influencing the values, e.g. having the value 1 (blue) in the filter variable F_1 does not limit the value range of X_4 whereas the value 0 (white) reduces the admissible values. The three filters lead

to seven possible combinations steering the admissible value range. Each combination is linked to a list with donor values. The random draw for initialization and the Bayesian Bootstrap that is usually conducted within the CART-step are both restricted to these values.

As filter variables can be struggled by missing values as well, at first the filter variables have to be imputed. Therefore, imagine that one of the filters, F_1, that is for this purpose defined as the variable X_2, has a missing value in the same row (same respondent) as X_4. First the value of X_2 has to be imputed as 0 or 1 in our example, then, based on the values of the other two filter variables, F_2 and F_3, the dedicated list of donor values is chosen. Finally, one of those donor values is drawn for a missing in X_4.

What struggles the imputer most is to define the admissible values for each combination of relevant filter values. This gets even harder when variables steering the filtering process have missings themselves. The following application uses parts of the data from the NEPS adult cohort study which were selected by relevance for the analysis of household net income. It is an extreme application respective the amount of filters and the maximum filter-depth of five chained filter variables that have to be taken into account for the imputation of an amount of 213 selected variables.

4.3 Nonparametric imputation of income data from the NEPS adult cohort data

4.3.1 The data and methodological consequences for the imputation method

Further NEPS data are used for the application, see for a better understanding of the structure of the NEPS figure 3.2. The Scientific Use File (SUF) of Starting Cohort 6 (SC 6) contained data from adults, so in contrast to the Thuringia study presented in chapter 3 there was no institutional context. The number of respondents in this study was $N = 11{,}649$. The adult cohort is an 'inhereted' study as it was former part of the Working and Learning in a Changing World

study (Arbeiten und Leben im Wandel; ALWA) conducted by the Institute for Employment Research (IAB, Nuremberg). Some of the respondents of the NEPS study were already interviewed within the ALWA study, so the information were used as additional information for NEPS. The number of persons already interviewed in the ALWA study and then interviewed in NEPS was $N_{ALWA} = 6{,}495$ and the number of first time interviewed in NEPS was $N_{NEPS} = 5{,}154$. Note that the data were not interpreted as panel data as there were changes in the questionnaire and the sampling frame, so the data were regarded as combination of two cross-sectional data sets.
The SUF contained 22 files with different topics of questions or generated variables, the earlier mentioned modules. The total number of variables within those 22 modules was 1,125, see figure 4.9.

The aim of the analysis of the data is concerned on household net income. For that reason, the total number of variables was reduced to income variables and all variables that possibly had a direct or indirect effect on household net income. Consequently, out of the 1,125 variables only (with the word 'only' referred to the comparison with the total number) 213 variables were chosen, that is 18.93% of all available variables. Those 213 variables were e.g. sociodemographic variables and variables about the school and employment history, compare figure 4.9. The variables about the employment history were completely steered by the filter question whether the respondents ever had an employment. That filter question was part of the sociodemo-graphy module (filter type 4), so 62 variables related to income were not asked to all participants of the survey. Hence, 151 variables were asked to all 11,649 respondents, whereas the module for the employment history with 62 relevant variables was only asked to 11,516. 133 respondents had a reduced number of questions respective to the chosen variables. The imputation therefore was conducted once for 11,649 respondents based on 151 variables and for 11,516 respondents based on 213 variables. The overlap was intended to maximize the number of information available from the data.

Another feature that had to be considered was that the SUF contained information in longformat. Longformat was used for the episodes of the respondents

within life time. As CART identifies one row as one individual, the data were converted to wideformat which means that each episode of a variable was interpreted as unique variable. The chosen variables from the different modules were combined, the data were harmonized, aggregated and dummies for the study (ALWA/NEPS) and the filtered module employment history were added, see for further information about this Würbach et al. (2014).

File	Module*	Short description	Number of variables
Panel file	pTarget	Socio-demographic information and panel data	415
Generated files	Basics	Basic information	68
	Biography	Integrated life course data	10
	Children	History of (formerly) cohabitating children	9
	Education	Transitions in educational careers	10
	FurtherEduc	Integrated course file	17
	MaritalStates	Marital biography	6
	Methods	Methods	37
Spell files	spSchool	General education history	41
	spVocPrep	Vocational preparation schemes	16
	spVocTrain	Vocational education history	79
	spCourses	Courses and trainings	32
	spEmp	Employment history	117
	spUnemp	Unemployment history	24
	spParLeave	History of parental leaves	17
	spMilitary	Military/ civilian service	16
	spFurtherEdu1	Additional courses	15
	spFurtherEdu2	Selected courses	26
	spFurtherEdu3	Courses in German	10
	spGap	Gap episodes	18
	spPartner	History of partners	72
	spChild	History of children	70

*Gray-colored modules were not considered for our purpose.

Figure 4.9: Modules in the NEPS SUF SC6

The filter structure of the data came to a maximum filter-depth of five variables steering the values of another variable with different combinations of the steering variables possible.

Additional to the filter structure of the data, income variables were asked as bracketed questions when the exact amount had not been answered for gross or net income on individual or household level. As can be seen in figure B.3 those brackets were disjoint. Bracketed questions reduce the number of available values for the imputation as well and increase the accurateness of the imputation. Drechsler (2011) described the special efforts that have to be made when imputation bounds defined by bracketed questions have to be included in parametric models and generally indicated on CART-based imputation models.
A validity check showed up implausible or inconsistent values which were set to 'NA'. The 'NR'-category of missing values was recoded to '-99' as '-99' does not change the measurement of scale for continuous data.
As a consequence of all these data features, the imputation step by CART has a limited range of values. So for every cell in the data set, that is the value for each person at each variable, a list of all possible combinations of filters and their corresponding values was generated. Combined within a matrix these lists were flexibly reduced by the observed and imputed filter variable values. So within the CART process donor values from the final remaining values are drawn via Bayesian Bootstrap. For variables without filters steering them the ordinary CART procedure was done.

4.3.2 Empirical results

At first, the descriptive statistics of the variable of interest are presented and it is illustrated how it is affected by missing values. Table C.4 shows a description of the income data estimates (quartiles, mean) obtained by complete cases analysis for household net income, individual net income and individual gross income. Table C.5 displays how the bracketed questions reduced the amount of missing values of the given income questions and that they were replaced with an interval of values. The proportion of missings reduced considerably when the exact income questions (any income information missing) are enriched by bracketed values. For the household net income the proportion reduced from 13.4% to 3.8%, for the individual net income from 8.0% to 2.1% and for the individual gross income from 10.7% to 3.5%.

Second, it was checked whether the MCAR assumption held or if there was a selectivity in responses. Model *I* of table 4.1 displays the results of a probit model with 'Any income information missing' as dependent variable. Additionally, a second probit model with a 'All income information missing' dummy was estimated as model *II*. Both dummy variables refered to the household net income. The results illustrated that there were significant effects in both models that implied that the MCAR assumptions did not hold.

According to model *I*, older respondents (61 years and older) refused to answer any income question more often than young respondents (up to 30 years). Women and people living with other adults (two or more than two adults in the household) tended to be more likely to have missing values as well. The occupational status had an effect on the tendency to have at least one income information missing with workers, employed and self-employed persons having a higher effect (more likely not to answer) compared to civil-servants (reference category). Additionally, people with a higher ISEI-score (International Socio-Economic Index of Occupational Status: prestige of the occupational position) tended to respond less often. Respondents who were unemployed (in contrast to not employed) tended to refuse the household net income question less likely. The satsifaction with the financial situation realized as an u-shaped-effect: people with very low and very high satisfaction with their finacial situation were less likely to report an exact estimate of their household net income.

Note that a generated variable 'Number of missings on covariates' was part of both models. It can be interpreted as a tendency to refuse answers in general. According to Jones (1996) missing indicators may bias the estimates, but nonetheless it was included in the model as it had a central explanatory role. The significant effects gave incidence that respondent who refused other variables were more likely to refuse the answer to the household net income as well.

	I Any household income information missing	*II* All household income information missing
31 to 40 years	−0.0393	0.1220
41 to 50 years	0.0089	0.2353**
51 to 60 years	0.0971	0.1983*
61 years or older	0.1346*	0.4207***
Gender: Female	0.1932***	0.0383
Adults in the household: Two	0.2442***	0.0919
Adults in the household: More than two	0.8909***	0.5115***
One child in the household	−0.0348	−0.0885
More than one child in the household	0.0237	−0.0579
Occupational status: Worker	0.2425**	0.1826
Occupational status: Employed	0.2753***	0.2582*
Occupational status: Self-employed	0.3734***	0.2600*
Occupational status: Other	−0.0229	−0.1785
Occupational status: Not working	0.1708	−0.0966
Unemployed	−0.1854**	−0.0405
Satisfaction with fin. situation	−0.1001*	−0.0842
Satisfaction with fin. situation (squared)	0.0086***	0.0083*
CASMIN: Group 2	−0.0364	−0.0764
CASMIN: Group 3	−0.0941	−0.1796*
ISEI-Score	0.0026*	0.0010
Born in Germany	0.0850	−0.1113
Living area: 20,000 up to 100,000 inh.	0.0256	0.1921*
Living area: 100,000 up to 500,000 inh.	0.0561	0.1395
Living area: More than 500,000 inh.	0.0981	0.2233**
Number of missings on covariates	0.4068***	0.4017***
Constant	−1.8546***	−2.2862***
Observations	11649	11649
Log-Likelihhod	−4371.6744	−1821.6165
Log-Likelihood, constant only	−4579.5030	−1882.8145

Reference Categories: 18 to 30 years; Male; One adult in the household; No child in the household; Occupational status: civil-servant; Working; Born abroad; CASMIN: group 1; Not unemployed; Living area: up to 20,000 inhabitants;
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4.1: Estimating the probability for item-nonresponse on household income questions - Results from probit models

In model *II* the respondents did not only refuse to answer the exact household net income, but all (rough) bracketed questions as well. The results indicated that refusing only the exact household net income or all household net income

questions was based on different motivations. Again, older respondents (but now 41 years and older) in comparison to the youngest (up to 30 years) as well as people living with other adults (households with more than two adults) tended to be more likely to have missing values. Employed and self-employed persons refused the answer more likely than civil-servants (reference category). The u-shaped effect of the satisfatction with the financial situation remained, but lost significance. Respondents that were part of CASMIN (Comparative Analysis of Social Mobility in Industrial Nations) group 3 (high education) refused less often compared to respondents of CASMIN group 1 (low education) whereas people from small (20,000 to 100,000 inhabitants) or very large cities (more than 500,000 inhabitants) refused more often than people from living areas up to 20,000 inhabintants (reference category). The tendency to refuse answers in general again had a positive effect on having a missing value.
Giving a résumé, the results showed that there was a selectivity in the data and that a complete cases analysis was not recommendable. Although MAR can not be tested, multiple imputation seemed reasonable.

Finally, the imputation was conducted. The focus on the imputation of the adult cohort data from the NEPS was on the household net income. After a burn-in phase of 10 iterations, 100 iterations of CART-based MICE were conducted and analyzed. Each iteration served as imputation, resulting in $M = 100$ imputed data sets. For each iteration a tree was built up with CART for each variable. Those trees can vary slightly between iterations because of the uncertainty within the imputation process. For the 100 iterations only modest variations within the trees were found. Figures B.4 to B.6 show the results for the household and the individual net income. Many interaction effects were captured by the tree structure. Each tree level divides the value range in two value groups for categorical variables and in ranges lowerequal and greater a cut-off value for continuous variables. The most important variable explaining the income variables were the bracketed questions. Note that the labeling of the bracketed questions in the trees refers to figure B.3 with e.g. 'income split 2c - 2' referring to split 2c, answer 2, that is '4,000 up to 5,000€'. The selection of the bracketed questions was not very suprising, but, the variables additionally chosen were of special in-

terest.

For the household net income which is shown in figure B.4 (all respondents, without regarding the employment history module), occupational status and age were additional explanatory variables to the bracketed questions, but only for the highest income group (income split 2c, answer 3, that is 'more than 5,000€', see figure B.3). When the employment history module was added (the number of respondents decreases by 133) for the highest income group the individual net income replaced occupational status and age, see figure B.5.

For the individual net income which is shown in figure B.6 the bracketed questions were the most important explanatory variables as well. The exact individual gross income was the only variable with additional explanatory impact.

Based on the advises of van Buuren (2012), the distributional similarity of all variables were compared before and after imputation to asses the quality of the imputation. Categorical variables were checked via Chi-square goodness of fit test and continuous variables via Kolmogorov-Smirnov goodness of fit test comparing the distributions before and after imputation. No significant changes were found for categorical variables, whereas for continuous variables the individual gross income and the sum of special payments differed significantly between observed and imputed data (with $\alpha = 0.05$). In figure B.7 the Q-Q plots for both variables show discrepancies in the higher quantiles indicating the imputation of higher values.

For most variables the imputation only changed the absolute frequencies without changing relative frequencies of categories or the distribution, as demonstrated on 'Expectations of friends: achieve success on a professional level', an ordinal variable and 'Social circle: further education', a binary variable, both displayed in figure B.8. For the variable of interest, the household net income and as additionally provided for the individual net income, the kernel densities showed only minor differences compared for before and after imputation, see figure B.9. As this form of illustration makes it hard to see differences we decided to recode the income information as classified data. In figure B.10 it can be seen that the imputation added especially values in the middle category (1,500 up to 3,000€) for the household net income and in the highest category (more than 3,000€) for the individual net income. As mentioned, those differences were not signifi-

cant. One of the reasons for the insignificant differences might be the very small amount of missing values, compare table C.5.

4.4 Conclusion

Filters are a useful way to individualize surveys and decrease the response burden by for example lowering the length of the surveying process, see e.g. Bosley et al. (1999). Though, when the data have to be imputed, there is a marked increase of the imputer's effort. Each combination of filters has to be taken into account, adding the case that the steering variable of the filter might have missings too. In addition, as shown in chapter 4.1 there are many types of filters that have to be taken into account. A nearby suggestion is that the imputer should avoid high-dimensional data and focus on a small set of possibly relevant variables only. As shown in the application on NEPS data, only a few variables were chosen for the imputation of the variable of interest by CART. The problem of limiting the number of variables is that the explantory variables have to be properly imputed as well. A pyramide scheme still has to be avoided, that is dropping the variables explaining the explanatory variables and keeping only the explanatory variables for the variable of interest is to repudiate. So foregoing checks about the variance of the explanatory variables could minimize the imputers effort to offer an imputation process considering the filter structure of the data.

When filters are still part of the data, the proposed approach is to define the possible value ranges for each variable depending on the filter structure. If available, filter schemes provided by the survey programmers can be used to facilitate this task. The whole list of value ranges depending on filter variables is then added for each cell of the data within a matrix. These lists are then flexibly reduced within the imputation procedure by observed or imputed filter steering variables. For this it is necessary that the order of the variables that are imputed follow the hierarchy of the filter structure, whereas usually an ascending order corresponding to the amount of missing values is necessary as a prerequisite for the consistency of the set of full conditional distributions, compare Si & Reiter (2013).

When the number of donor values gets too low for the tree to partition the data

the remaining process is henceforth not always CART-based for the variables steered by filters, but constrained to the Bayesian Bootstrap on the remaining values.
An alternative strategy to the described procedure using a matrix with a list for all cells of the data frame is to add the filter structure in the imputation process by defining value ranges by *'if'*-conditions. This can, dependent on the complexity of filters, extend the run-time of the approach considerably, but might be an option for simple filter structures. Another option would be an *accept-reject*-procedure within the imputation process enforcing new draws when the value drawn via Bayesian Bootstrap from the corresponding node does not fit to the filter structure. A disadvantage of these alternative strategies is that the produced tree structure might not end in nodes that contain reasonable values. Another strategy, that seems to be the easiest way of handling filter structures is to simply ignore them while multiply imputing. If existing, the discrepancies in values could then be corrected manually which can be interpreted as manual version of an accept-reject procedure. With many values possible for a filtered variable one of these could be chosen by different mechanisms as for example by sampling one of the values unconditionally. On the one hand this strategy would ease the task of imputing, but on the other hand it would lead to improper predictions for filtered variables as the conditional distribution of the values is not regarded. In other words, the available information about the filtered variable's values would not be used.
Above all, the filter structure of the data can be interpreted as additional information. The original CART might not regard the filter structure automatically, especially when the filter-depth is high. Thus, it might not lead to correct tree structures. Hence, using the filter structure as extension of the knowledge about the data increases the quality of the whole CART-based imputation approach.

As an incidental conclusion, it can be mentioned that CART as a nonparametric method, leads to time-savings, especially if conducted on large application data. These time-savings are based on the fact that no specification of models or model families as by parametric methods is needed.

Chapter 5

Some insights into the performance of CART

To get some insights into the performance of a new method, simulation studies have to be performed. The literature still lacks studies with coverage statistics and exploration of the potential of CART-based multiple imputations to create proper imputations, see van Buuren (2012, p.84). Nevertheless, there are a few works with promising results. The first simulation study, conducted by Burgette & Reiter (2010), was based on nine regressors and an additional variable that was not included in the data generation model of Y. All ten variables were multivariate normally distributed. Data were deleted in Y and eight of those regressor variables, that were X_1 to X_8, based on a MAR mechanism. The fraction of missing data was on average 17% resulting in fewer than 25% complete cases. The results of CART-based MICE were compared to those of the standard MICE. Burgette & Reiter (2010) conclude that CART-based MICE is superior to standard MICE respective the simulated data. Doove et al. (2014) extended that simulation study by creating three different data generating models for Y. The fraction of missing data of Y was approximately 50%. They showed that for variables whose full conditional distribution include interaction effects of other variables the performance of CART-based MICE results in more reliable inferences compared to the parametric MICE. Though, the results are relativized, as the potential of CART "depends on the relevance of a possible interaction effect,

the correlation structure of the data, and the type of possible interaction effect present in the data", see Doove et al. (2014, p.92). Shah et al. (2014) evaluated the performance of Random Forest-based MICE on survival data. The basis of one of the two conducted simulation studies was real data. The fraction of missing data ranged from 1.5% to 56.7% based on a MAR mechanism. The other simulation study comprised three variables with one variable having 20% missing values. Random Forest is an alternative machine learning algorithm based on decision trees and is listed here due to the absence of manifold research publication in the decision tree area. Shah et al. (2014) show that Random Forest-based MICE is superior to parametric MICE concerning the bias of the estimates of (log) hazard ratios. Additionally, the parameter estimation is more efficient and narrower confidence intervals are produced. Conducted on nonlinear dependent data, Random Forest-based MICE leads to less biased parameter estimates and higher coverages respective parameter estimates' confidence intervals. Stekhoven & Bühlmann (2012) evaluated the performance of Random Forest-based imputation based on ten different data sets with 10%, 20% or 30% data that were randomly removed, resulting in a MCAR mechanism. The results were differed by continuous variables only, categorical variables only or mixed-type data. They conclude that the full potential of their algorithm *missForest* is reached "when the data include complex interactions or non-linear relations between variables of unequal scales and different type", see Stekhoven & Bühlmann (2012, p. 171). Valdiviezo & Aelst (2015) evaluated the performance of tree-based methods on predicitions comparing imputation and surrogate decision methods. They used five data sets, four based on real data (with 80% of the original data used as training set and 20% as test set) and one simulated data. The amount of predictors ranged from three to thirteen. The fraction of missing data was set to 10%, 20%, 30% and 40% for MCAR, MAR and NMAR. They conclude that the performance clearly depends on the fraction of missing values within the data. Ensemble methods combined with surrogates and single imputation lead to sufficient results for small proportions. Conditional inference trees combined with multiple imputation are the best choice for moderate and large fractions of missing data. Conditional bagging using surrogates can be considered as alternative, especially for high-dimensional prediction problems. Overall, Valdiviezo & Aelst

(2015) show that multiple imputation ensembles are superior for all of their applications. Akande et al. (2015) evaluated the performance of CART-based MICE compared to MI-GLM (chained equations using generalized linear models) and MI-DPM (a fully Bayesian joint distribution based on Dirichlet Process mixture models) on categorical data. The simulation study was based on random samples from real data. 30% respective 45% of the data were deleted randomly, resulting in a MCAR mechanism. The results show that if only main effects appear in the data MI-GLM is superior. Adding more complex structure MI-CART and MI-DPM have to be preferred, thus, there is no clear winner.
The following simulation study is guided by the simulation study of Koller-Meinfelder (2009, chapter 5.3), especially respective the data generation process. This simulation study aims at evaluating the performance of CART-based MICE concerning three variables with different data generating functions and a high fraction of missing values. More precisely, the performance of CART-based MICE is assessed respective the imputation of a non-metric variable that is created violating linear model assumptions in two of the three data situations. The fraction of missing data is 60% steered by a MAR mechanism. The remainder is structured as follows. First, the setup of the data, that is especially the data generating functions and the MAR mechanism, are presented. Then, peculiarities of CART are illustrated. This is followed by the description of the analysis and the presentation of the results. A conclusion ends this chapter.

5.1 Setup of the data

All simulated data sets consisted of three variables, i.e. a variable with missing values Y and two variables, X_1 and X_2, that were used for imputation. X_1 and X_2 were completely observed. Three different types of data sets were generated based on three different data generating functions of Y whereas the missing data mechanism was missing at random (MAR). The sample size was $n = 2,000$.

The variables X_1 and X_2 were generated identically throughout all data sets with $X_1 \sim U(0, 3)$; and $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.

There were three different data generating functions used for Y:

1. $y_1 = [1.75 + x_1 - 0.5x_2 + u_1]$, with $u_1 \sim N(0, \frac{11}{48})$.
2. $y_2 = \left[1.75 + x_1 - 0.5x_2 + (u_2 - \frac{107}{96})\right]$, with $u_2 \sim \chi^2_{\frac{107}{96}}$.
3. $y_3 = \left[4 + 1.5(x_1 - 1.5)^3 - 0.25log(abs(x_2 + 9)) + u_3\right]$, with $u_3 \sim N(0, 0.2)$.

All values of Y were rounded resulting in **integer** values which added further disturbance to the relation of the variables.

The rate of missing values was fixed to 60% for Y with the MAR mechanism related to X_1 which was defined by:

$$y_i = \begin{cases} missing, & \text{if } F_z(z_i) > 0.4 \\ y_i & \text{if } F_z(z_i) \leq 0.4 \end{cases} \forall i = 1, \dots, n$$

where $F_z(z)$ is the empirical distribution function of Z, and

$$z = \frac{1}{1 + exp(0.2x_1\phi + \epsilon)} \text{ with } \phi \sim N(0, 16) \text{ and } \epsilon \sim N(0, 36).$$

Bias, mean squared error (MSE) and coverage were used to check the performance of CART-based MICE. The coverage was therefore defined as the proportion of 95%-confidence intervals for the estimated parameters that contain the true value.

The chosen parameters were

mean: $E(Y)$

proportions: $P(Y < 3)$, $P(Y < 4)$, $P(Y < 6)$

correlations: $\rho(X_1, Y)$, $\rho(X_2, Y)$

linear model estimates: α, β_1, β_2.

	DS1	DS2	DS3
$E(Y)$	4.0001	4.0001	4.4907
$P(Y < 3)$	0.2033	0.2564	0.1335
$P(Y < 4)$	0.3943	0.4357	0.2187
$P(Y < 6)$	0.7967	0.7821	0.7837
$\rho(X_1, Y)$	0.7500	0.5810	0.8784
$\rho(X_2, Y)$	-0.8278	-0.6413	-0.3136
α	1.7501	1.7498	3.9976
β_1	1.0000	1.0001	1.5004
β_2	-0.5000	-0.5000	0.2512

Table 5.1: Overview of the mean estimates of 20,000 data sets

All 'true values' for those estimands were calculated as mean from $20,000$ generated data sets with $10,000$ values for each variable. The results can be seen in table 5.1 for all three data sets (DS1, DS2 and DS3). If at all, the values differ only by sampling errors from the values presented by Koller-Meinfelder (2009, chapter 5.3.3).

The number of imputations was set to $M = 15$ and 500 runs were conducted to get proper results especially for the coverage. The results were compared to those of the complete data (before deletion) and the complete cases.

5.2 Peculiarities of CART

CART needs the dichotom information *factorvar* for each variable, that is the information whether a variable is an (ordered) factor (*factorvar=1*) or not (*factorvar=0*). Based on this assignation, a classification or a regression tree is chosen. To define the values of that variable is an easy task when the number of variables is as small as in this simultation study, but gets stressfull for large data.

Special in this task was the rounding of Y. It seems obvious to define Y as factor (*factorvar=1*) due to the integer values. However, as the creation of Y

was based on a linear model with non-integer outcomes, a regression tree is the better choice which needs the assignation as non-factor (*factorvar=0*).
CART detects proper values for the imputation based on the structure given within the data. The results will show whether having only three variables leads to sufficient results, especially as the missing values of Y are initialized unconditionally from the empirical distribution of the observed values which affects the structure within the data.
Concerning the CART-based MICE *R*-command which was provided by Burgette & Reiter (2010), multiple variables with missing values are needed as the programming is made for vectors. Instead of retyping the syntax an easy solution is to set one missing value additionally in an originally completely observed variable. This solution comes along with a (small) loss of information. Note that when this part of the thesis was conducted the implementation of CART within the *R*-package *MICE* as *mice.impute.cart* did not exist.
The message *incrementing minCut by one* might pop up in *R* when multiply imputing using CART. The message informs about a change in settings. If the message appears repeatedly it can be interpreted as a warning that there is an error within the procedure.

5.3 Analysis

Rubin's combining rules which are explained in chapter 2.2.4 are based on the assumption that the estimates which are combined are (approximately) normally distributed. Since mean, proportions and regression parameters fullfil this requirement Rubin's combining rules are used to calculate confidence intervals of those estimates. Correlations are neither normally nor approximately normally distributed. However, correlation values can be transformed with the Fisher ('ρ to z') transformation and then be combined. The confidence intervals were calculated with the z-transformed values and were then transformed back to get the ρ-values. The *R*-package *psych* assists perfectly for this task and was used for this simulation study.

The standard method to calculate confidence intervals for proportions is the 'normal approximation interval', known as Wald interval or asymptotic interval, firstly described by Laplace (1812). Alternatively, the 'exact'-method, also known as 'Clopper-Pearson'-method by Clopper & Pearson (1934) could be applied. The name refers to the exact binomial distribution, not to exact confidence intervals, which are usually too conservative for this method, compare Tuyl (2001). Agresti & Coull (1998), Brown et al. (2001) and Brown et al. (2002) have shown that the 'Agresti-Coull'-method to calculate confidence intervals is to be favored for large n, that is the total amount of values of a variable, compared to the 'exact'-method and the Wald interval. The 'Agresti-Coull'-method adjusts p, that is the proportion of values of interest. For this purpose X, that is the amount of values of interest within n and n itself, that is the mentioned total amount of values of a variable, are adjusted by an additional term, that is $\frac{\lambda_{1-\frac{\alpha}{2}}^2}{2}$, respectively $\lambda_{1-\frac{\alpha}{2}}$. Then, the resulting values are used for the calculation of the adjusted p. The 'Agresti-Coull' confidence intervals can thus be calculated by:

$$\tilde{p} \pm \lambda_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \text{ with } \tilde{X} = X + \frac{\lambda_{1-\frac{\alpha}{2}}^2}{2},\ \tilde{n} = n + \lambda_{1-\frac{\alpha}{2}} \text{ and } \tilde{p} = \frac{\tilde{X}}{\tilde{n}}.$$

There are some more alternatives to calculate the confidence intervals that can for example be calculated easily **for completely observed** data with the *binom*-package in *R*. For this simulation study the 'Agresti-Coull' method was chosen.

To evaluate the impact of the settings of CART on the performance of CART-based MICE, results using default settings were compared with outcomes from alternative settings for DS1. First, the initialization was changed back to the default tree-based initialization instead of draws from the unconditional empirical distribution. Second, the iterations of the tree-based MI approach were increased from 20 to 50. Third, the number of imputed data sets that were combined for the confidence intervals was doubled from 15 to 30. All changes should improve the performance of CART-based MICE.
As already mentioned in chapter 2.3.2, there are more settings that can be varied. The default of the *tree*-command in the *R-package tree* is a minimum size of 5 units in the final node and a minimum reduction of heterogenity of 0.01 that is

needed to conduct a split. Both can be changed. On the one hand, especially in the context of imputation the depth of the trees can be increased by the heterogenity criterion as the trees are not used to explain the structure of the data. On the other hand, random combinations could be interpreted as 'structure' by mistake. Consequently, the amount of donor values would be decreased by an additional split and the variance of the imputed values would be reduced by mistake. An analogous argumentation can be made for the minimum size of the final nodes. Decreasing the amount of units lowers the variance of donor values and might lead to improper structure identifications. Increasing the amount might prevent CART from detecting relevant structure. Both, heterogenity and minimum leaf size criterion are influencing each other. Finally, no changes of those settings were conducted.
Furthermore, the setting *factorvar* could be changed by interpreting originally numerical values as (ordered) factors or the other way round. However, it is not recommendable to do, especially as the tree growth is limited to a maximum of 31 levels and the amount of levels of a categorical variable to 32. Hence, no changes in that setting were conducted.

5.4 Results

Three different data generating functions were defined for Y. In the simulation study of Koller-Meinfelder (2009, p.51) parametric and semi-parametric procedures worked with misspecified models for the second and third data set. The violation of the normality assumption in the second data set and the violation of the linearity assumption affect imputation procedures which are based on linear models. As CART is a nonparametric procedure it should not be affected by those assumption violations. Correspondingly, the overall coverage of the first data set, shown in table 5.2, and of the third data set, shown in table 5.4, were not that different. Respective CART-based MICE, the first data set had an overall coverage of 91.8% whereas it was 90.7% for the third. The results for the second data set differed more with an overall coverage of 88.8%. The reason for this difference and more detailed results are given in the following.

The results of DS1 in table 5.2 show that the CART-based imputation worked satisfactorily with a minimum coverage of 88.8% for the constant of the linear model, that is α. Actually, the coverages for all estimands of the linear model, that is α (88.8%) as the constant and the two slope estimands β_1 (90.2%) and β_2 (89.8%), were the lowest compared to mean (94.0%), proportion (91.2% to 94.8%) and correlation estimands (93.2% and 90.8%). The relative bias, shown in table C.6, confirmed the relatively bad performance with an amount of 0.4% to 0.9% (absolute values). All other estimands had a relative bias located between 0.01% and 0.4% (absolute values).

The average coverage for all nine estimands was 91.8%. For comparison, the average coverage is 97.2% for the before deletion data estimands and 62.4% for the complete cases estimands. The complete cases analysis lead to proper results for the correlations and linear model estimands, but failed for the mean and the proportions. Whereas the coverage for the estimand of the mean was only 73.8% for the complete cases estimand, the coverage of the CART-based MICE estimand was 94.0%, thus, very close to the before deletion coverage (96.8%).

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	96.8	73.8	94.0
$P(Y < 3)$	0.2033	96.8	5.0	93.4
$P(Y < 4)$	0.3943	96.4	1.8	94.8
$P(Y < 6)$	0.7967	95.2	2.8	91.2
$\rho(X_1, Y)$	0.7500	97.8	96.4	93.2
$\rho(X_2, Y)$	-0.8278	95.4	95.4	90.8
α	1.7501	93.8	94.8	88.8
β_1	1.0000	94.0	95.0	90.2
β_2	-0.5000	97.2	96.2	89.8
Average	-	95.9	62.4	91.8

Table 5.2: Coverages: DS1

In spite of the violation of the normality assumption of the linear model, the performance of CART-based MICE was still good for the second data set, that is DS2, which can be seen in table 5.3. The average coverage for the estimands

of the imputed values was lower with 88.8% compared to 91.8% in DS1. Still 88.8% were high compared to 93.8% for the before deletion estimands and 61.1% for the complete cases estimands of DS2.

Eye-catching was the low coverage of the correlation estimands, especially of $\rho(X_2, Y)$. It was the lowest coverage for before deletion with 86.6% and for CART-based MICE with 83.8%. In contrast, the relative bias, shown in table C.8, was striking for $\rho(X_1, Y)$ with 0.6% and unremarkable for $\rho(X_2, Y)$ with 0.1% (absolute values).

The findings indicate that the low coverages were caused by the violation of central assumptions of the pearson correlation coefficient, that is the assumption of linearity of the correlation and the need for two normally distributed random variables. Above all, the non-linearity of Y in DS2 due to the chi-square distributed error term disturbed the correct calculation leading to too low coverages, especially for $\rho(X_2, Y)$. The highest coverages for CART-based MICE were reached for the proportion estimands (90.6% to 92.4%).

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	95.6	76.2	87.0
$P(Y < 3)$	0.2564	95.4	2.6	92.4
$P(Y < 4)$	0.4357	95.0	3.2	91.2
$P(Y < 6)$	0.7821	95.8	5.6	90.6
$\rho(X_1, Y)$	0.5810	90.6	90.2	87.0
$\rho(X_2, Y)$	-0.6413	86.6	85.4	83.8
α	1.7498	95.0	96.0	89.8
β_1	1.0001	93.4	95.2	88.0
β_2	-0.5000	96.6	95.6	89.6
Average	-	93.8	61.1	88.8

Table 5.3: Coverages: DS2

On the whole, the results of the performance of CART-based MICE did not change for DS3 which can be seen in table 5.4. The average coverage for the imputed values was 90.7% compared to 94.8% for the before deletion estimands and 61.6% for the complete cases estimands. With 82.4%, the lowest coverage

for CART-MICE was reached for the slope β_2 of the 'linear model'. The slope β_2 steered the logarithmic term. As in DS1, the coverages for the estimands of the 'linear model' were the lowest (82.4% to 88.2%) compared to mean, proportion and correlation estimands. The relative bias, shown in table C.10, underlined this low coverages with relative biases of 1.8% for α, 0.2% for β_1 and 14.6% for β_2 (absolute values).
Comparing table 5.2 and table 5.4, that are the two tables showing the coverages for DS1 and DS3, it can be seen that the coverage of the before deletion correlation estimand $\rho(X_1, Y)$ was too high. As Y was rounded to an integer and the assumptions of the pearson coefficient were violated due to the missing linearity and the normality of the variables, this is not remarkable.

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.4907	95.6	69.0	92.8
$P(Y < 3)$	0.1335	94.2	4.2	93.6
$P(Y < 4)$	0.2187	94.0	3.2	93.8
$P(Y < 6)$	0.7837	95.2	2.2	93.2
$\rho(X_1, Y)$	0.8784	98.6	97.8	96.2
$\rho(X_2, Y)$	-0.3136	93.6	94.6	93.8
α	3.9976	94.4	94.4	82.6
β_1	1.5004	93.8	95.0	88.2
β_2	0.2512	93.8	94.2	82.4
Average	-	94.8	61.6	90.7

Table 5.4: Coverages: DS3

All tables containing the relative bias and the mean squared error (MSE) information for all three data sets, that is table C.6 to C.11, can be found in the appendix.

An aggregated version of the results of (Koller-Meinfelder, 2009, p.55) can be seen in figure 5.1. The abbreviations stand for 'ROV': Rounding to the nearest Observed Value, 'PPMM': Posterior Predictive Mean Matching, 'BBPMM': Bayesian Bootstrap Predictive Mean Matching and 'RPMM': Rounded Predictive Mean Matching.

Table 5.3: Average coverage (in %) by various factors

coverage	CC	ROV	PPMM	BBPMM	RPMM
overall	82.02	84.09	90.80	92.37	83.75
big	75.80	77.71	91.82	93.19	80.77
small	88.23	90.47	89.79	91.55	86.73
MCAR	*94.81*	83.92	91.23	92.75	84.09
MAR	69.23	84.26	90.37	91.99	83.41
DS(1)	82.82	95.58	91.37	92.75	86.24
DS(2)	82.53	90.71	88.96	90.48	86.46
DS(3)	80.70	**65.99**	92.08	93.88	78.54

Figure 5.1: Results from the analysis of Koller-Meinfelder (2009)

The analysis was more extensive and with a focus on different aspects. The overall coverages include the coverage of the variance estimator. The missingness was set to both, missing completely at random (MCAR) and missing at random (MAR) and the sample sizes were $n = 200$ (small) and $n = 2,000$ (big). Nethertheless, broadly spoken, it can be seen that at least the coverages were in a comparable range to the results of PPMM and BBPMM and much higher compared to the results of ROV and RPMM. Basis for the comparison is the row MAR of table 5.1. The values represent the average coverage of the three data sets with big and small sample size, including the variance estimation coverage. The average coverage of CART-based MICE of this simulation study was 90.45% for all three data sets with the big sample size, not including a variance estimation.
As the results for each data situation are available in the appendix in Koller-Meinfelder (2009), in table 5.5 the adjusted average coverages are presented for comparison, leaving out the variance estimation coverage for the calculation of the average coverage of all methods.
The results can only be compared with constraints as the methods were not conducted on the same values, only the same data generation processes. Moreover, the results depend on the design of the simulation study, e.g. no interaction effects were included and the complete cases data were not sufficient for a tree-

DS	ROV	PPMM	BBPMM	RPMM	CART-MICE
DS1	95.73	92.09	93.09	81.78	91.80
DS2	89.33	90.76	91.90	87.70	88.82
DS3	45.00	92.51	94.90	70.24	90.73
Average	76.69	91.79	93.30	79.91	90.45

Table 5.5: Coverages in percent for all three data sets (BIG, MAR)

based initialization of CART (more about this circumstance is described below). Still, it can be seen that for the ordinary linear model without any assumption violations, CART-based MICE is only superior to RPMM, close to the values of PPMM and BBPMM, but clearly inferior to ROV. For DS2, that is the model including a chi-squared error term, CART-based MICE is again superior to RPMM, though close to ROV and PPMM and inferior to BBPMM. When the logarithmic term is added, CART remains superior to RPMM, is additionally superior to ROV, close to PPMM and inferior to BBPMM. Averaging all the results, CART-based MICE is superior to ROV and RPMM and inferior to PPMM and BBPMM as already shown respective the comparison with the results from table 5.1.

The results of CART-based MICE were compared to those of the before deletion and complete cases analysis using three different data generating models of Y. Additionally, the performance using different settings on the first data set (DS1) were checked changing the initialization, increasing the number of iterations and the amount of imputed data that had to be combined.
Changing the initialization back from an unconditional draw from the empirical distribution to a tree-based draw like originally intended by Burgette & Reiter (2010) seems to be a good decision at first view. Trying to implement this, the following error appeared:

Error in runif(length(eligibles) - 1) : invalid arguments
In addition: Warning message:
In node.match(nodes, node, treeframevar == "<leaf>") :
supplied nodes 0 are not in this tree

Here, a shortfall of CART and other methods based on empirical data as donor values instead of predicted values becomes apparent. Due to the missingness the bandwith of value combinations is limited. For example, all possible donors of Y combined with $X_1 \leq 0.3$ and $X_2 \geq 5.0$ are missing. Using only complete cases received by listwise deletion the learning data are insufficient to initialize the application data. Drawing unconditionally from the empirical values might weaken distinct combinations, but might recreate some lost structure. Additionally, CART-based MICE can be applied afterwards in contrast to the the tree-based initialization which ended with an error message.

The learning data used were the complete cases data generated via the *R*-command *na.omit*). The command conducts a listwise deletion, that is that each row from the data set was deleted when any missing value existed in a column. With a deletion of 60 percent of the data set the structure of the data got partially lost. Thus, CART was not able to get enough information from the learning data to partition the data containing missing values.

As can be seen regarding the results from DS1 to DS3, the performance of CART was sufficient even with an initialization that was not only non-informative, but in addition might decrease the correlation of the variables. CART was able to mitigate those effects with 20 iterations.

Suprisingly, increasing the number of iterations from the initialization to the imputed data set (burn-in phase) from 20 iterations up to 50 did not increase the quality of imputation. As can be seen in tables C.12, C.13 and C.14 neither coverage and relative bias (in percent) nor the mean squared error differed much. Consequently, it seems not worth worrying about increasing the number of iterations.

Unexpectedly, the findings indicate in the same way that increasing the amount of imputed data sets from 15 to 30 was not worthwhile in case of the simulation data. As it can be seen in table C.15 to C.17 the results for the coverages, relative bias (in percent) and the mean squared error did not improve due to the doubling.

Summarized, changes of the default settings or the initialization did not increase the performance of CART-based MICE. 20 iterations for each imputed data set and 15 imputations to calculate the MI estimates were sufficient for the simula-

tion data. Furthermore, the 'easy task' of using tree-based initialization values can lead to problems in structure identification in the application data. This happens when the learning data does not contain the needed value combinations due to listwise deletion. In this case, the tree resulting from the learning data fails to represent the correctly related final nodes that are needed to serve donor values for the imputation available.

5.5 Conclusion

The conducted simulation study had the objective to give some insights about the performance and settings of CART in combination with multiple imputation. The data generation was based on the simulation study of Koller-Meinfelder (2009, chapter 5.3). Additionally, a tree-based initialization and changed settings, that is the number of iterations and imputations, were checked. The performance was assessed by coverage, bias and mean squared error.

The performance results of CART-based MICE show that the method can be used as adequate imputation method. The overall coverage was at a minimum of 88.8% for DS2, the highest value was reached for DS1 with 91.8%. Indendent of a direct comparison to the methods presented by Koller-Meinfelder (2009), it can be seen that the CART-based MICE approach performed satisfactorily, that is that the CART-based MICE results were close to the before deletion results respective coverage, bias and mean squared error.

Some changes of the settings were conducted on DS1 to check their influence of the performance of CART-based MICE. First, the initialization was changed back from unconditional draws from the empirical distribution to tree-based. The tree-based initialization was not feasible as the recreation of the structure of the simulated data structure by the learning data (with 60% missing values steered by a MAR mechanism) was not complete. Second, the amount of iterations for each data set from the initialization to the chosen imputed value was set to 20. Increasing that number to 50 did not improve the performance of the procedure. Third, the same result shows for doubling the amount of imputed

data sets. The originally chosen 15 imputations were sufficient. Surprisingly, it can be summarized that none of the setting manipulations had a positive effect on the performance of CART-based MICE.

As CART is a nonparametric method, it was assumed that the violence of the linear model assumptions has no negative effect on the results of CART-based MICE. The first data set, that is DS1, included a typical linear model as the data generation of Y. No violations were included. The results of DS1 show by far the highest coverage (91.8%) of all three data sets. Second, the non-linear term in DS3 leads to a coverage of 90.7%, that is not much lower and could have happened by chance. Certainly, the chi-squared error term in the data generation function of Y in DS2 lead to the lowest result with 88.8%. As the pearson correlation coefficient coverages were affected by the violences too, the two values could be ignored for the calculation of the average coverage. Then, the following values resulted: 91.7% (DS1), 89.8% (DS2) and 89.5% (DS3). There was still a clear difference between DS1 and the other two data sets respective the average coverage. Consequently, it can be assumed that the violences of the linear model assumptions have a negative effect of the recreation of a variable's values by CART.

In chapter 5.2 it was speculated if three variables are enough to recreate the structure of a variable with missing values by CART. The results of this simulation study indicate that they are. However, it can be assumed, that a tree-based initialization and data enabling the creation of surrogate splits to compensate the missingness, increase the performance of CART.

Chapter 6

Nonparametric imputation of panel data

Panel data can be described in contrast to cross-sectional data by including different points in time and in contrast to longitudinal data by including the same units for those points in time. For panel data, especially in the context of survey data, those points in time are also called waves. In contrast to this description, panel data can include units that have information available only for one wave. Consequently, not each wave contains the same amount of units. Alltogether, a definition for panel data is pretty hard to find. A very rough one summarizes that all have in common, that the *objective* is to collect "multiple observations on each individual in the sample", see Hsiao (2003, p.2). This definition points out that panel data are defined by the sample, that can be different to the realization. Furthermore, the original sample can change as refreshments and the loss of units over time is taken into account as well for some panel data. In addition, data can be available for example as survey data, observations or metadata. An example for metdadata are changing statistics such as the unemployment rate in different countries over time. Due to these multifarious kinds of data, there is no embracing literature about panel data, but only on special kinds of panel data.

At least, all panel data have a multilevel structure in common, that is that different time points are clustered in units. With this additional level the amount

of types of missing values that can occur increases. In the following those types are summarized using the example of collecting survey data of individuals over time:

- Individuals that are in the sample can be missing in the data completely as a survey was never realized with them.
- Individuals do not take part in one or several waves, but took part at least once.
- Individuals leave the panel by not taking part any longer after a certain point in time.
- Individuals refuse to answer some questions at some points in time.

As there is no embracing literature about panel data, consequently, there is no embracing literature about the correction of missing values in panel data either. In the following, there is a short literature review with attention to challenges that occur to imputers when handling surveyed panel data.

When we look at the literature, there are some ways of correcting for nonresponse with several definitions of panel data as basis. Shao et al. (2012) define their panel population by the observed entirety of the first wave of their survey and not by a previously defined sample in contrast to the definition of Hsiao (2003). So in the baseline, that is $t = 1$ with the time index defined as $t = 1, \dots, T$ in general, there is no unit nonresponse by definition. Based on the past-value-dependent response propensity assumption by Little (1995) and Little & Rubin (2002) and the work of Vansteelandt et al. (2007) they offer an approach to handle monotone and nonmonotone nonresponse. The imputation regressions are always based on past observations. This procedure is not transferable for survey panel data when survey data are defined by a sample and not by the realizations of this sample. Additionally, the usage of only past observations for imputations limits the information available for imputations very restrictively.
Kleinke et al. (2011) compared different methods to handle missing values in panel data. They summarize that techniques such as case deletion (which is only

suitable when the missing mechanism is MCAR) and single imputation techniques such as mean imputation or regression imputation do not cope with the requirements of panel imputation, as shown e.g. by Schafer & Graham (2002). The missing indicator method, described by Allison (2001), yields biased estimates. Furthermore, as shown by Carpenter et al. (2004) and Cook et al. (2004), the 'last observation carried forward'-method also suffers in producing biased estimates. Kleinke et al. (2011) recommends multiple imputation techniques and show their application in the statistical software *R*. Multiple imputation as used in the *R*-packages *norm:* Analysis of multivariate normal datasets with missing values, *cat:* Analysis of categorical-variable with missing values, *mix:* Estimation/multiple imputation programs for mixed categorical and continuous data and *pan:* Multiple imputation for multivariate panel or clustered data are all recommended. Not shown in their application but mentioned with the hope for further research was predictive mean matching (PMM). Meanwhile, there are many PMM approaches in *R*, as for example *BBPMM* in the *BaBooN*-package, *mice.impute.pmm* in the *mice*-package or *mi.pmm* in the *mi*-package. Kleinke et al. (2011) critisize that the extension to panel data is still missing, especially the expansion to the multilevel structure of the data. At least for some kinds of regressions those expansions exist, as for example an add-on to the *mice*-package for binary data by Zinn (2013).

There is no general recommendation in the literature to deal with nonresponse within panel data. The literature review on simulation studies based on decision trees in chapter 5 includes no data set considering panel structures. The aim of this study is to provide a method to handle this issue using CART. First, the method is presented. Then, some peculiarities of CART relevant for this study are outlined. The next section describes how the performance of the imputation method was assessed. Afterwards, results concerning the suitability of CART-based MICE for panel data are presented followed by a conclusion.

6.1 Proposed procedure for handling nonresponse within a panel study using CART

When looking at panel data from the imputer's perspective, whole units can be missing and item nonresponse can occur. When additional data are available, a sensitivity analysis is recommended to check whether respondents and nonrespondents differ. If they do, a method to correct the data should be applied, for example weighting.
If the additional data about the units besides their surveying are rich enough, unit nonresponse should be handled as item nonresponse, see Rässler & Schnell (2003). When $t > 1$ unit nonresponse can be handled as item nonresponse if people who refused or were not able to take part in some waves have survey data available for at least one wave. Especially sociodemographic variables are often time-invariant. Additionally, other information can be used for imputation of some missing variables as for example a) information from earlier and later waves, b) additional information available for all (non)respondents as for example from a sampling frame or c) information for at least some (non)respondents as for example by third persons. If the additional data are not rich enough to handle unit nonresponse as item nonresponse, for example weighting, a correction based on response prospensity, see Peress (2010), or other approaches can be used. A very commendable reference about unit nonresponse adjustment techniques is the discussion from Little on Brick (2013), see Little (2013).

The following is focused on item nonresponse, respectively handling unit nonresponse (of survey data) as item nonresponse within a panel study. For the imputation of the available panel data at first it has to be checked whether logical imputations are possible. As mentioned above, most of the sociodemographic variables are time-invariant or at least stable for some time span. So the additional variables can complement the earlier wave information.
Then the missing values are initialized by draws from the empirical distribution without replacement. The initialized and during the iterations of the imputation continuously 'updated' data are used as training data to fit the tree models and for prediction of the terminal nodes. The imputation range of a value is not

limited to past data based on the recommendations in Little (1992).
All of the data are imputed together and not separated by time point. Sociodemographic variables that are stable over time, as for example sex, are reduced to one variable for all waves.
It is not necessary to model the response propensity as it is included in the variables chosen by CART when available in the data. The recommended algorithm has the following form.

Step 1: Initialize the missing values with draws from the unconditional empirical distribution.

Step 2: Given the initialized values, the CART algorithm is used sequentially to obtain a nonparametric approximation of the full conditional distribution. The original missing values are replaced by draws from the predictive distribution conditional on all other variables.

Step 3: Repeat step 2 $L + M$ times, where L should be high enough to mitigate the effect of initialization.

6.2 Setup of the data

As in chapter 5, the setup of the data was based on the simulation study of Koller-Meinfelder (2009, chapter5.3).
The simulated data set consisted of three variables, that is a variable with missing values Y and two variables, X_1 and X_2, that were used for imputation. X_1 and X_2 were completely observed. In contrast to the procedure in chapter 5, the data generating functions differed for all three variables, Y, X_1 and X_2, dependent on the given data situation as described below. In addition, two waves were generated. Y was affected by missing values whereas the missing data mechanism was MAR and based on X_1. The sample size was $n = 2,000$.

The basis model for the analysis was a linear panel model, that is
$[y_{it} = x_{it}\beta + a_i + u_{it}]$ with β as a $Kx1$ vector and K as the number of variables within the model, compare e.g. Greene (2012, chapter 11.2.1). There is a broad

range of different panel models that can result from this form, as presented e.g. by Greene (2012, chapter 12.2.2) or Cameron & Trivedi (2005, chapter 21.2.1), that is pooled regression, fixed effects model, random effects model and a random parameters model.

A **pooled regression** is defined by a_i as constant term. Then, that is the one case the ordinary least squares method (incorporated as *lm*-function in *R*, the default linear model method) leads to consistent and efficient estimates. In all other cases, the ordinary least squared method leads to biased and inconsistent results for β. The **fixed effects model** is defined by a_i as unobserved individual heterogenity (time-invariant), that is correlated with x_{it}. The **random effects model** is defined by a_i as unobserved individual heterogenity (time-invariant), that is uncorrelated with x_{it}. The **random parameters model** is defined by an additional random constant term.

The chosen models for the further analysis were the fixed effects model and the random effects model. Four data situations ($DS1, \dots, DS4$) were created:

DS1: A random intercept a_i was defined as $a_i \sim N(0, 0.25)$, X_1 and X_2 were time-invariant with $X_1 \sim U(0, 3)$; and $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.

DS2: A random intercept a_i was defined as $a_i \sim N(0, 0.25)$, X_1 was time-invariant with $X_1 \sim U(0, 3)$; X_2 changed over time with $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$ for $t = 1$ and $\epsilon \sim N(0, 6)$ for $t = 2$.

DS3: A random intercept a_i was defined as $a_i \sim N(0, 0.25)$, X_1 changes over time with $x_{1,t=1} \sim U(0, 3)$ and $x_{1,t=2} \sim U(0, 4)$; $X2$ was affected by X_1 and became time-variant too, as it was defined as $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.

DS4: An intercept a_i was defined that was dependent on X_1, that is $a_i \sim N(0, 0.1x_{1,t=1})$ (fixed effect), X_1 changed over time with $x_{1,t=1} \sim U(0, 3)$ and $x_{1,t=2} \sim U(0, 4)$; $X2$ was affected by X_1 and became time-variant too, as it was defined as $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.

For all four data situations Y was rounded to an integer value. The model estimands, that is the intercept α, and the two slopes β_1 and β_2 were the same for all four (three for α as it was not estimated for the fixed effects model) data situations with

$$\alpha = 1.75,\ \beta_1 = 1.0 \text{ and } \beta_2 = -0.5.$$

The rate of missing values was fixed to 60% for Y based on X_1 for the MAR mechanism which was defined by:

$$y_i = \begin{cases} missing, & \text{if } F_z(z_i) > 0.4 \\ y_i & \text{if } F_z(z_i) \leq 0.4 \end{cases} \forall i = 1, \dots, n$$

where $F_z(z)$ is the empirical distribution function of Z, and

$$z = \frac{1}{1 + exp(0.2 x_1 \phi + \epsilon)} \text{ with } \phi \sim N(0, 16) \text{ and } \epsilon \sim N(0, 36).$$

The missingness mechanism was affected by the changes in data situation three and four, as the creation of X_1 was changed, but the missingness model itself did not change between both waves.

6.3 Peculiarities of CART

In addition to the peculiarities described in chapter 5.2, it has to be mentioned that CART needs the information arranged in wideformat when handling panel data as already mentioned in chapter 4.3.1. Both ways of displaying data contain the same information. As can be seen in figure 6.1 the arrangement in longformat, as shown on the left side, includes the wave information within the variables (columns), identifyable by an index (here: time index t with $t = 1, 2$). An example to explain this way of illustration is the income, that is X_1. In year 2010, that is wave one ($t = 1$), a person with identifying (ID) number $i = 10$ has an income of $1,000$ Euro. In 2011, that is wave two ($t = 2$), the same person ($i = 10$) has an income of $1,105$ Euro. Both income information are saved in

the same column and are identifyable with information from two other columns from the same row (identifier i, time index t).

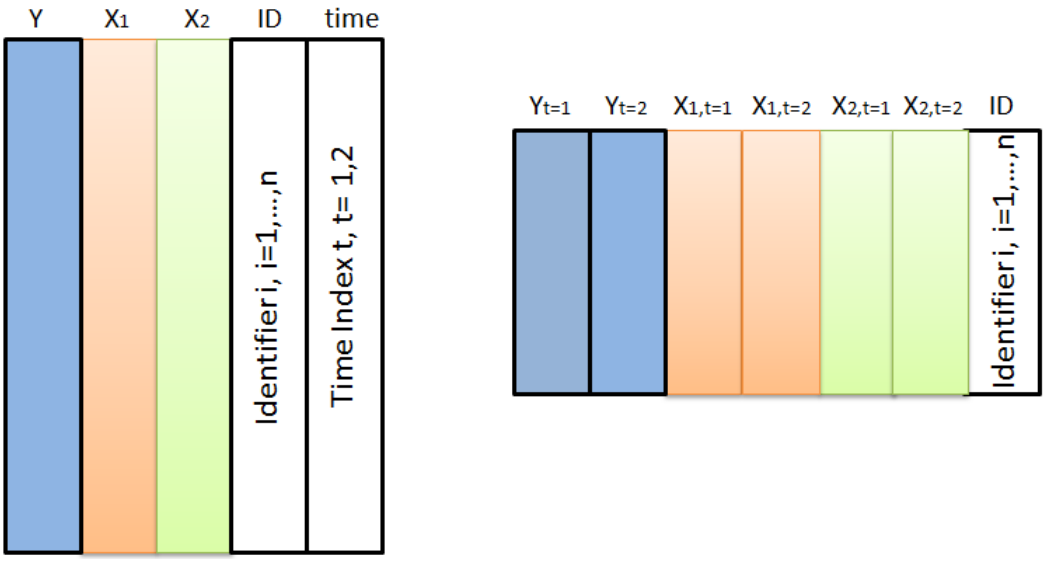

Figure 6.1: Data in longformat and wideformat

The same information can be displayed in different columns arranged as wideformat as shown on the right side of figure 6.1. For example, the income information from 2010 is then available from a variable income, X_1, with the time index as additional information, that is $X_{1,t=1}$. An additional variable $X_{1,t=2}$ contains the information from 2011. The time index is not needed anymore as an additional variable as it is included within the index. The length of the person identifying column is halfed, containing only unique IDs.

6.4 Analysis

To check the performance of CART-based MICE bias, mean squared error (MSE) and coverage were used. The coverage is defined as the proportion of 95%-confidence intervals for the estimated parameters that contain the true value.

The chosen parameters were

mean: $E(Y)$

proportions: $P(Y < 3)$, $P(Y < 4)$, $P(Y < 6)$

correlations: $\rho(X_1, Y)$, $\rho(X_2, Y)$

linear model (panel data) estimates: α, β_1, β_2

with mean, proportions and correlations calculated for each of both waves and the linear panel model estimates calculated for the whole data as random effects or fixed effects.

All 'true values' for those estimands were calculated as mean from $2,000$ generated data sets with $10,000$ values for each variable analogous to the procedure in chapter 5.1. The results can be seen in table 6.1 for all four data sets (DS1 to DS4). a_i was not estimated as the mean of a_i was set to zero with $a_i \sim N(0, 0.25)$ for all four data situations.
Compared to the estimates from DS1 of table 5.1 the correlations in table 6.1, that is $\rho(X_1, Y)$ and $\rho(X_2, Y)$, were lower for DS1. This can be explained by the disturbing effect of the random intercept due to the given variance.

The estimates of DS1 and DS2 were very close to each other, different only due to the data generating variance of X_2 which was time-invariant in one case (DS1) and time-variant in the other case (DS2). The estimates of DS3 and DS4 were likewise similar, different only due to the additional correlation of a_i to X_1 in DS4 in comparison to DS3. The two groups, that is DS1 and DS2 compared to DS3 and DS4, differed a lot due to the change in X_1 from time-invariant to time-variant with different maxima. As X_1 steered the missingness mechanism it had the highest influence on the performance of the imputation.

The number of imputations was set to $M = 15$ and 500 runs were conducted to get proper results especially for the coverage. The results were compared to those of the complete data (before deletion) and the complete cases.

	DS1	DS2	DS3	DS4
$E(Y_{t1})$	3.9996	4.0001	4.0001	4.0005
$P(Y_{t1} < 3)$	0.2110	0.2110	0.2110	0.2088
$P(Y_{t1} < 4)$	0.3970	0.3968	0.3969	0.4005
$P(Y_{t1} < 6)$	0.7891	0.7890	0.7891	0.7927
$\rho(X_{1,t1}, Y_{t1})$	0.7208	0.7205	0.7205	0.7319
$\rho(X_{2,t1}, Y_{t1})$	-0.7954	-0.7953	-0.7952	-0.8079
$E(Y_{t2})$	4.0000	3.9998	4.7506	4.7499
$P(Y_{t2} < 3)$	0.2109	0.2250	0.1581	0.1560
$P(Y_{t2} < 4)$	0.3968	0.4020	0.2983	0.2976
$P(Y_{t2} < 6)$	0.7891	0.7750	0.6225	0.6228
$\rho(X_{1,t2}, Y_{t2})$	0.7207	0.6709	0.8108	0.8200
$\rho(X_{2,t2}, Y_{t2})$	-0.7953	-0.8198	-0.8109	-0.8199
α	1.7497	1.7500	1.7504	—
β_1	1.0003	1.0001	0.9998	1.0001
β_2	-0.4998	-0.5000	-0.5001	-0.5001

Table 6.1: Overview of the mean estimates of 2,000 data sets

The model that was used for the analysis was a linear model for panel data regarding the unobserved individual heterogenity a_i. As a_i was not correlated with the other variables for DS1 to DS3 the model used was a random effects model. For DS4 a_i was correlated with X_1, consequently a fixed effects model was applied. In *R* there are several packages and commands that can be used for panel models. One command applicable for a linear model for panel data is *plm* (from the *R*-package *plm*: Linear Models for Panel Data) which has the option *model* which has to be set to *model="random"* for a random intercept or a random effects model. For a fixed effects model the option has to be set to *method="within"*. The default of the method used to calculate the variance of the unobserved individual heterogenity *random.method="swar"* is the method suggested by Swamy & Arora (1972). The alternatives are the methods of Wallace & Hussain (1969), Amemiya (1971) and Nerlove (1971).
One example of alternative commands that can be used is *lme* (from the *R*-

package *nlme*: Linear and Nonlinear Mixed Effects Models). As the *plm*-package with the default setting *random.method="swar"* leads to appropriate results, it was used to estimate the linear panel model estimands.

As already mentioned in chapter 5, Rubin's combining rules, which are explained in chapter 2.2.4, are based on the assumption that the estimates to combine are (approximately) normally distributed. Regression parameters, proportions and mean fulfill this requirement. Correlations are neither normally distributed nor approximately normally distributed.
Correlation values can be transformed with the Fisher ('ρ to z') transformation and then be combined. The confidence intervals were calculated with the z-transformed values and were then transformed back to get the ρ-values. The *R*-package *psych* assists perfectly for this task and was used for this simulation study.
Likewise to the procedure described in chapter 5.3 the 'Agresti-Coull'-method was used to calculate the confidence intervals for proportions.

Summary of the simulation study settings

The settings for the simulation study can be summarized as follows: The number of rows (individuals) in the (wideformat) data was $n = 2{,}000$. There were two waves, $t = 1, 2$. One dependent variable, that is Y, was generated via a linear function based on X_1 and X_2, which were partially time-invariant (DS1) or time-variant (DS2 to DS4). The simulation study was repeated 500 times with $M = 15$ imputed data sets which were combined by Rubin's combining rules. The 95%-confidence intervals for before deletion (BD), complete cases (CC) and imputed data (CART-MICE) estimands were checked if they contain the true value. The amount of the confidence intervals for each method that contain the true value was used to compute a ratio of appropriate confidence intervals, that is the coverage.
The settings for CART were the default settings as described in chapter 5 since the examined setting manipulations did not improve the performance of the imputation procedure. CART is iterated 20 times for one imputation value.

6.5 Results

Four data sets with different data generation models were created. First, the results of DS1 are shown in table 6.2. DS1 included a random intercept a_i, defined as $a_i \sim N(0, 0.25)$, furthermore, X_1 and X_2, both time-invariant, with $X_1 \sim U(0, 3)$; and $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	3.9996	95.2	75.0	91.2
$P(Y_{t1} < 3)$	0.2110	95.6	3.4	94.8
$P(Y_{t1} < 4)$	0.3970	92.4	4.0	91.0
$P(Y_{t1} < 6)$	0.7891	95.8	3.4	91.0
$\rho(X_{1,t1}, Y_{t1})$	0.7208	95.8	96.8	91.2
$\rho(X_{2,t1}, Y_{t1})$	-0.7954	92.6	95.2	84.2
$E(Y_{t2})$	4.0000	95.0	82.6	89.6
$P(Y_{t2} < 3)$	0.2109	94.0	35.0	89.8
$P(Y_{t2} < 4)$	0.3968	95.2	33.8	92.4
$P(Y_{t2} < 6)$	0.7891	96.0	32.2	91.6
$\rho(X_{1,t2}, Y_{t2})$	0.7207	97.2	97.0	91.6
$\rho(X_{2,t2}, Y_{t2})$	-0.7953	95.2	95.4	87.4
α	1.7497	93.8	94.6	82.4
β_1	1.0003	95.4	95.4	84.0
β_2	-0.4998	94.8	92.8	83.2
Average	-	94.9	62.4	89.0

Table 6.2: Coverages: DS1, Panel

Regarding CART-based MICE, the coverages for the mean and the first proportion were higher for the first wave, whereas the coverages for the other estimands were higher for the second. The lowest coverage, that is 82.4%, was reached for the constant of the linear panel model, that is α. The highest coverage, that is 94.8%, was reached for $P(Y_{t1} < 3)$. Actually, the coverages for all estimands of the linear model, that is α (82.4%) as the constant and the two slope estimands

β_1 (84.0%) and β_2 (83.2%), were the lowest compared to mean, proportion and correlation estimands. This property is similar to the cross-sectional data, shown in table 6.2. The relative bias, shown in table C.18, confirmed the relative 'poor' performance with an amount of 0.7% to 1.1% (absolute values). All other estimands had a relative bias located between 0.02% and 0.6% (absolute values).
The average coverage of all fifteen estimands was 89.0%. For comparison, the average coverage was 94.9% for the before deletion data estimands and 62.4% for the complete cases estimands. The complete cases analysis led to proper results for the correlations and linear panel model estimands, but failed for the mean and the proportions. Whereas the coverage for the estimand of the mean was only 75.0% respectively 82.6% for the complete cases estimand, the coverage of the CART-based MICE estimand was 91.2% respectively 89.6%, thus, close to the before deletion coverage (95.2% respectively 95.0%). The average coverage of the estimands of the cross-sectional data was 91.8%, see table 5.2.

The results of DS2 are shown in table 6.3. DS2 included a random intercept a_i, defined as $a_i \sim N(0, 0.25)$, X_1 which was time-invariant with $X_1 \sim U(0, 3)$ and X_2 which changed over time with $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$ for $t = 1$ and $\epsilon \sim N(0, 6)$ for $t = 2$.
Considering the imputation, the constant of the linear panel model, that is α, and the slope estimand β_2 showed the lowest coverage (83.0%). The highest coverage, that is 93.2%, was gained for $P(Y_{t2} < 6)$. Again, the coverages of all estimands of the linear model were the lowest in contrast to mean, proportion and correlation estimands. Correspondingly, the relative bias, shown in table C.19, emphasized the relative 'poor' performance with an amount of 0.8% to 1.0% (absolute values). The remaining estimands had a relative bias located between 0.02% and 0.7% (absolute values).
The average coverages amounted to 89.5% for CART-based MICE, 95.0% for the before deletion data estimands and 62.8% for the complete cases estimands. In more detail, the complete cases analysis led to good results for the correlations and linear panel model estimands, but failed for the mean and the proportions. Comparing the coverage of the mean, the complete cases reached 76.8% for wave 1 and 80.2% for wave 2 while the CART-based MICE estimation reached

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	94.2	76.8	90.4
$P(Y_{t1} < 3)$	0.2110	93.0	3.2	91.6
$P(Y_{t1} < 4)$	0.3968	95.0	3.2	92.2
$P(Y_{t1} < 6)$	0.7890	95.2	2.8	92.8
$\rho(X_{1,t1}, Y_{t1})$	0.7205	96.0	95.4	89.8
$\rho(X_{2,t1}, Y_{t1})$	-0.7953	95.6	94.8	89.0
$E(Y_{t2})$	3.9998	95.2	80.2	89.6
$P(Y_{t2} < 3)$	0.2250	94.8	52.0	91.8
$P(Y_{t2} < 4)$	0.4020	94.0	42.0	91.2
$P(Y_{t2} < 6)$	0.7750	97.0	13.4	93.2
$\rho(X_{1,t2}, Y_{t2})$	0.6709	97.8	96.4	91.6
$\rho(X_{2,t2}, Y_{t2})$	-0.8198	93.6	95.0	89.0
α	1.7500	93.8	94.8	83.0
β_1	1.0001	94.4	96.4	84.0
β_2	-0.5000	95.0	95.4	83.0
Average	-	95.0	62.8	89.5

Table 6.3: Coverages: DS2, Panel

90.4% respectively 89.6%, thus, close to the before deletion coverage (94.2% respectively 95.2%).
Adding variation over time for X_2 did not decline the coverages compared to the time-invariant variable X_2 in DS1. Both, MSE and relative bias worsened only minimal, see table C.18 and table C.19, respectively C.22 and C.23.

In table 6.4 the results of DS3 are shown. DS3 included a random intercept a_i, defined as $a_i \sim N(0, 0.25)$, X_1 which changed over time with $x_{1,t=1} \sim U(0, 3)$ and $x_{1,t=2} \sim U(0, 4)$ and $X2$ which is affected by X_1 and became time-variant too, as it was defined as $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0, 4)$.
DS3 was the first data set which had a different missingness mechanism for each of both waves caused by the change within the data generation of X_1.

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	95.4	74.4	89.6
$P(Y_{t1} < 3)$	0.2110	96.2	1.6	93.6
$P(Y_{t1} < 4)$	0.3969	95.4	3.4	92.8
$P(Y_{t1} < 6)$	0.7891	94.8	4.6	90.2
$\rho(X_{1,t1}, Y_{t1})$	0.7205	96.2	97.2	93.0
$\rho(X_{2,t1}, Y_{t1})$	-0.7952	94.2	93.8	86.4
$E(Y_{t2})$	4.7506	95.4	80.2	90.4
$P(Y_{t2} < 3)$	0.1581	93.6	0.0	93.2
$P(Y_{t2} < 4)$	0.2983	94.2	0.0	91.8
$P(Y_{t2} < 6)$	0.6225	96.0	0.0	92.6
$\rho(X_{1,t2}, Y_{t2})$	0.8108	97.4	97.4	92.2
$\rho(X_{2,t2}, Y_{t2})$	-0.8109	93.8	95.0	88.4
α	1.7504	94.2	93.6	82.0
β_1	0.9998	96.2	95.2	85.0
β_2	-0.5001	95.2	95.0	84.4
Average	-	95.2	55.4	89.7

Table 6.4: Coverages: DS3, Panel

The lowest coverage based on the imputed data set was 82.0% for the constant of the linear panel model α. $P(Y_{t1} < 3)$ had the highest coverage (93.6%). Once more, the coverages of the mean, the proportion and the correlation estimands exceeded the estimands of the linear model. As shown in table C.20, the relative bias, indicated the relative 'poor' performance only for α with an amount of 0.7% and β_2 with an amount of 1.1% (absolute values). β_1 was the exception with a relative bias of 0.2% (absolute value). The relative bias of the other estimands varied between 0.02% and 0.7% (absolute values).

Respective the results of CART-based MICE, the average coverage of the fifteen estimands was 89.7%. Before deletion data estimands resulted in an average coverage of 95.2%, complete cases estimands yielded an average coverage of 55.4%. Correlations and linear panel model parameters were estimated properly based on complete cases analysis in contrast to the mean and the proportions.

While the coverage of the mean was only 74.4% (wave 1) respectively 80.2% (wave 2) for the complete cases analysis, the coverage of the CART-based MICE estimand was 89.6% (wave 1) respectively 90.4% (wave 2) and approximated to the before deletion coverage of 95.4% in both waves.

The results of DS4 are illustrated in table 6.5. DS4 included an intercept a_i which was defined as $a_i \sim N(0, 0.1x_{1,t=1})$, i.e. a_i was dependent on the first wave values of X_1 which led to a fixed effect. Furthermore, X_1 which changed over time with $x_{1,t=1} \sim U(0,3)$ and $x_{1,t=2} \sim U(0,4)$ and $X2$ which was affected by X_1 and became time-variant too, as it was defined as $x_2 = -x_1 + \epsilon$, with $\epsilon \sim N(0,4)$.
The linear panel model was estimated as fixed effects model due to the dependency of a_i on X_1.

With 85.0% the lowest coverage of CART-based MICE referred to the correlation of Y and X_2 in the second wave, that is $\rho(X_{2,t2}, Y_{t2})$. The highest coverage amounted to 94.4%, for $P(Y_{t2} < 4)$. Contrary to the results of the previous data sets the coverages of the two slope estimands of the linear model were not the lowest compared to the other estimands. The relative bias, shown in table C.21, indicated a relative bad performance on the correlation of Y and X2 with an amount of 0.7% (absolute value) for both waves. All other estimands, except β_2 had a relative bias ranging from 0.01% and 0.7% (absolute values). Despite the good coverage, β_2 had a relative bias of 1.0% (absolute value).
The average coverage of CART-based MICE was 91.4%, before deletion data yielded an average coverage 95.5% and complete cases analysis 52.9%. The complete cases analysis led to acceptable outcomes for the correlations and linear panel model estimands, but not for the mean and the proportions. The coverage for the estimand of the mean was 76.6% respectively 79.2% for the complete cases estimand, while the coverage of the CART-based MICE estimand was 91.2% respectively 92.0% and just a little lower than the before deletion coverage (96.8% respectively 94.6%).

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0005	96.8	76.6	91.2
$P(Y_{t1} < 3)$	0.2088	96.2	4.8	92.0
$P(Y_{t1} < 4)$	0.4005	94.8	3.6	93.8
$P(Y_{t1} < 6)$	0.7927	94.0	3.6	91.6
$\rho(X_{1,t1}, Y_{t1})$	0.7319	96.8	97.4	93.6
$\rho(X_{2,t1}, Y_{t1})$	-0.8079	94.8	94.2	86.2
$E(Y_{t2})$	4.7499	94.6	79.2	92.0
$P(Y_{t2} < 3)$	0.1560	95.8	0.0	91.6
$P(Y_{t2} < 4)$	0.2976	95.4	0.0	94.4
$P(Y_{t2} < 6)$	0.6228	94.4	0.0	92.0
$\rho(X_{1,t2}, Y_{t2})$	0.8200	96.8	97.2	91.4
$\rho(X_{2,t2}, Y_{t2})$	-0.8199	95.2	94.2	85.0
α	—	—	—	—
β_1	1.0001	95.8	95.2	94.2
β_2	-0.5001	96.2	94.2	90.2
Average	-	95.5	52.9	91.4

Table 6.5: Coverages: DS4, Panel

The results show that CART-based MICE leads to sufficient results for panel data. The lowest coverage was reached for the first data set with 89.0% and the highest for the fourth (fixed effects model) with 91.4%. The coverages were rising from the first to the fourth data set. As the data were simulated as panel data the constructed trees that were the basis to impute the missing values are of special interest. Especially whether the data were imputed using information from the same wave or changing between waves. For demonstration the resulting trees for $Y_{t=1}$ and $Y_{t=2}$ are displayed. The trees for DS1 are displayed exemplary in the following, that is figure 6.2 for $Y_{t=1}$ and figure 6.3 for $Y_{t=2}$, whereas the figures of the other data sets (DS2, DS3 and DS4) are located in the appendix, see B.11 to B.16. Note that the created trees might (slightly) change for each imputation cycle. The displayed trees are the ones from the last iteration of one imputation cycle.

The tree-based selection of explanatory variables for $Y_{t=1}$, shown in figure 6.2, were all part of wave one, that is $t = 1$. As the data generation of the variables $X_{1,t=1}$ and $X_{2,t=1}$ was independent of the data generation of all variables from wave two, $t = 2$, and $Y_{t=1}$ and $Y_{t=2}$ only had the random effect in their data generation in common, this was appropriate. The tree had a depth of five levels reaching a minimum end node size of 168, that is a proportion of 8.4% of the whole data set with $n = 2,000$. The highest end node size was 219, that is a proportion of 10,95%. The explaining variables $X_{1,t=1}$ and $X_{2,t=1}$ alternate for the mapping of $Y_{t=1}$.

Analogously, the tree-based selection of explanatory variables for $Y_{t=2}$, shown in figure 6.3, were all part of wave two, that is $t = 2$. Again, the tree had a depth of five levels reaching an even lower minimum end node size of 88, that is a proportion of 4.4% of the whole data set. The highest end node size was 218, that is a proportion of 10,9%. In contrast to the first wave's tree, the explaining variables $X_{1,t=2}$ and $X_{2,t=2}$ were not alternating for the mapping of $Y_{t=2}$. That means that the partition of $Y_{t=1}$ and $Y_{t=2}$ was basically different and not only changing by decimals.

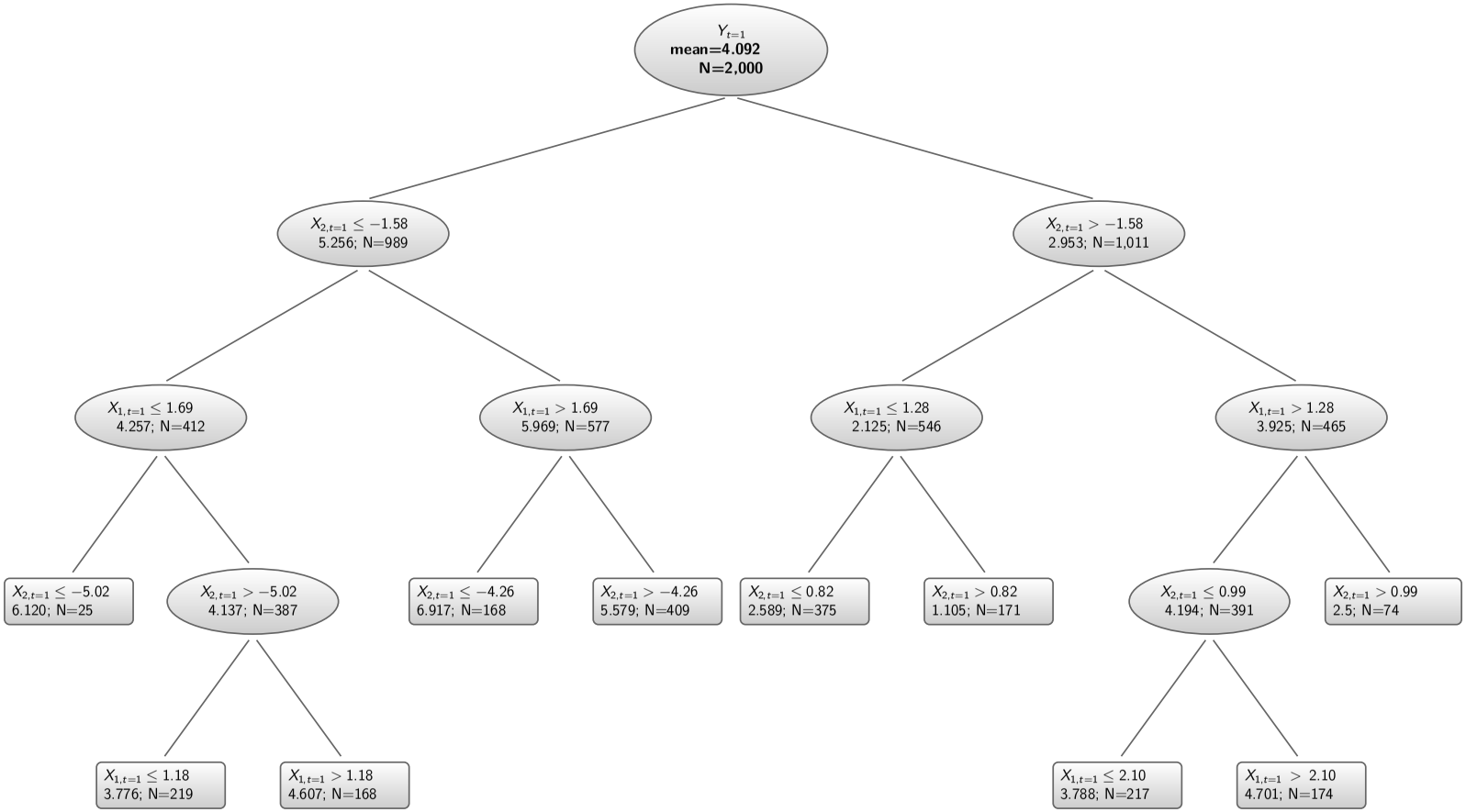

Figure 6.2: Resulting tree of one imputation cycle for the imputation of $Y_{t=1}$ of DS1. Notes: the mean is always with reference to $Y_{t=1}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

Figure 6.3: Resulting tree of one imputation cycle for the imputation of $Y_{t=2}$ of DS1. Notes: the mean is always with reference to $Y_{t=2}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

6.6 Conclusion

The conducted simulation study had the objective to give some insights about the performance of CART in combination with multiple imputation on panel data. The data generation was based on the first data set of the simulation study of Koller-Meinfelder (2009,chapter 5.3) in extension to chapter 5.

A linear panel model was created and four different unobserved individual heterogenity terms were added. In three of those four cases that term was defined as independent from the regressors, consequently the model was a random effects model. The last was defined as correlated to $X_{1,t=1}$ leading to a fixed effects model. X_1 in general steered the missingness mechanism of both waves. The first three data sets included different combinations of X_1 and X_2 as time-variant or time-invariant.

The analysis shows that CART-based MICE leads to sufficient results as the lowest coverage was 89.0% for the first and the highest coverage was 91.4% for the fourth data set. The coverages were rising from the first to the fourth data set. The data generation for each of both waves was conducted either independent of the other wave (besides of the unobserved individual heterogenity) or the variables were time-variant for all four data situations. Thus, it is pleasing that CART detected only information from the same wave for the partitioning of the data regarding $Y_{t=1}$, respectively $Y_{t=2}$.

The results lead to the restricted conclusion that CART-based MICE is able to handle panel data. However, the representation of panel data as a manifold term was very limited. The simulation study included only information from two waves, random effects models and a fixed effects model. The two waves were only connected due to the unobserved individual heterogenity a_i and partially time-invariant regressors, not by, for example, changes of the variables over time based on past values of the same variable. The missingness mechanism was generated identically for both waves. There were only two explanatory variables and no filtering was included. Consequently, there are many extension possible for this

simulation study and it is recommendable to research more on this topic before giving general recommendations. Anyway, the results from the given simulation study are very promising.

Chapter 7

Concluding Remarks

The thesis focused on the application of CART-based MICE and CART in combination with data augmentation on special data situations within a large-scale panel study. As presented, there were multiple applications within a large-scale panel study to use the approach of Burgette & Reiter (2010) in modified or extended ways.

As a first application, the usage of CART combined with a Bayesian Probit model for the analysis of data from the National Educational Panel Study which was affected by unit nonresponse was presented. The application was needed to find possible selection effects within the data for nonrespondents compared to respondents. The implementation of the CART step was very close to the suggested approach of Burgette & Reiter (2010). Differences were conducted for the initialization and the hierarchy of the data. The data were initialized not with CART itself, but with unconditional draws from the empirical distribution. As the application data contained information of students within schools, clusters had to be taken into account. CART was divided into two steps. At first, data from the individual level were augmented using the information of the aggregated level without imputing it. Then the procedure was conducted for the second level analogously. A novelty was the usage of CART within a data augmentation procedure resulting in a Markov chain Monte Carlo approach, more precisely a Gibbs Sampler.

The results of the analysis showed the necessity for nonresponse adjustment weights. The data included a selectivity of nonrespondents compared to nonrespondents given by the variable sex. For both years, female students tended to participate more often than the male ones. The results of the analysis differed clearly when using CART considering the second level random effect compared to the complete cases analysis ignoring as well as including the second level. The basic complete cases analysis suggested an effect of the marks of fs1, fs3 and the mean school mark in 2010 and an effect of sex and the marks of fs3 in 2011. Including a second level random effect there were no significant effects in 2010 and only an effect of the random effect in 2011 for the complete cases analysis.

Second, CART was used in combination with multiple imputation by chained equations considering a high amount of filters and a high-depth filter structure of the data. Several types of filtering were presented as they might influence the selection of valid values for imputation differently. The CART approach of Burgette & Reiter (2010) was extended with a matrix containing several lists of values for each person and filter combination within the data. These lists defined the admissible value range dependent on the filter value or the combination of filter values. All filters had to be regarded for the initialization and the imputation. With a maximum of five chained filter questions, the imputation was very complex. In contrast, the resulting trees showed a very manageable breakdown of relevant variables. Unsuprisingly, the most important variables to capture the structure of the exact household net income and the individual net income variable were the bracketed question variables. Additionally, occupational status and age were relevant as explanatory variables, but only for the highest household income group. Adding the employment history module, the same group was best captured by the individual net income. For the individual net income the bracketed questions and the exact individual gross income were most relevant explanatory variables. Pleasing was the display of interaction effects by CART as the different income groups were influenced dissimilar.

The two applications on real data show that CART can be flexibly combined with other algortihms and used for many challenges that occur within a large-

scale panel study. But diagnostics on real data are very constricted. Hence, the suitability of CART-based MICE can not be assessed in the context of real data applications. Thus, analysis on simulated data was needed to evaluate the performance of CART-based MICE. A simulation study was conducted based on the work of Koller-Meinfelder (2009, chapter 5.3). Three data sets were created. Two variables were drawn from the same distribution for all three data sets, that is X_1 and X_2, whereas the data generation of Y varied. For the first data set (DS1) Y was generated via a linear regression based on X_1 and X_2, the other two data sets included a chi-squared distributed error term (DS2) and a non-linear term or more precisely a logarithmic term (DS3). The results showed that the performance of CART in combination with MICE is sufficient. The lowest coverage, with coverage defined as the proportion of 95%-confidence intervals for the estimated parameters that contain the true value, was 88.8% for the second data set (DS2). As the pearson correlation is, among others, based on the assumption of the linearity of the correlation and the normality of both variables, the correlation estimates were biased even for before deletion results (BD: 93.8% overall coverage, 90.6% for $\rho(X_1, Y)$ and 86.6% for $\rho(X_2, Y)$; CART-based MICE: 88.8% overall coverage, 87.0% for $\rho(X_1, Y)$ and 83.8% for $\rho(X_2, Y)$). Considering this bias, the overall coverage of CART-based MICE could be evaluated as 'high' and definitely sufficient as it was close to the before deletion value. The overall coverages for the other two data sets were high as well with 91.8% (DS1) and 90.7% (DS3) for CART-based MICE.

In a rough comparison to the results of Koller-Meinfelder (2009) for ROV, PPMM, BBPMM and RPMM, the average coverages of CART-based MICE were (a little) lower than the coverages of PPMM and BBPMM and clearly higher than the coverages of ROV and RPMM. Unfortunately, a general recommendation which method is to prefer can not be derived from the current comparison. The methods did not perform on the same values and the results depended on the design of the simulation study.

Additional analysis was performed on the first data set (DS1) changing the settings of CART. It was tested whether changing the initialization from drawing unconditionally from the empirical distribution back to a tree-based (informative) initialization leads to better results. Unfortunately, no results to be compared

were available, as the tree-based initialization was not able to recreate the structure of the application data based on the test data. The test data, that is the complete cases data, were achieved by listwise deletion. Furthermore, it was tested if increasing the amount of iterations of the tree-based MI approach from 20 to 50 or doubling the number of imputed data sets that are combined for the confidence intervals improves the performance. Considering the coverages, relative bias and mean squared error, both changes did not improve the performance of CART-based MICE on the simulated data.

In a second simultation study, the first data set (DS1) from the previous simulation study was extended to panel data. Four data situations were distinguished defined by different data generation procedures of Y. In addition, different combinations of time-variant and time-invariant variables were created for the first three data situations. Those three data situations included an unobserved individual heterogenity which was independent of the regressors. Consequently, the regression models of Y were defined as random effect models. The fourth data situation included an unobserved individual heterogenity which was dependent on $X_{1,t=1}$ with X_1 steering the MAR mechanism. Consequently, the regression model of Y was defined as fixed effects model. The results were sufficient with a minimum coverage of 89.0% reached for the first and a maximum coverage of 91.4% reached for the fourth data set. The coverages are rising from the first to the fourth. The illustration as trees show that CART detected the wave-specific creation of Y correctly for each of both waves of all four data situations.

The additional benefit of this thesis can be found in an example of how manifold the applications with CART can be, the evaluation of the performance of CART-based MICE on cross-sectional and panel data and some insights about the settings of CART. Both content areas, the empirical applications and the performance check based on simulation studies, can be enlarged. In sum, the results of the simulation studies imply that the usage of CART within a large-scale panel study is recommendable with constraints. Reassuring is that the results of the two simulation studies indicate that the default settings are adequate as they were conducted on the empirical data.

Limitations of this work and self-criticism

In all four cases, CART was a highly flexible approach to get imputation values without defining a model for any variable. The modifications and extensions that had been implemented for the applications were needed as there was no syntax available to handle the properties of the data automatically. Meanwhile, at least the basic syntax of Burgette & Reiter (2010) is implemented in the *MICE*-package as *mice.impute.cart*. The modifications of the syntax offered by Burgette & Reiter (2010) that were conducted for the application in chapter 3, that was the change from a tree-based initialization to draws from the unconditional empirical distribution and some changes of the syntax for the adjustment to newer *R*-versions, were minor. However, the combination with the Bayesian Probit analysis as Markov chain was innovative and costly. The extension needed for chapter 4, that was especially the matrix containing lists of the admissible values dependent on filter values or filter value combinations was considerable. It was adjusted for the application on NEPS data which has a high amount of filtered variables and a high filter-depth of up to five filter levels. Hence, it is very application-specific.

As mentioned above, data characteristics which indicate the usage of CART-based MICE instead of other imputation procedures should be precised. Further research is needed to evaluate which challenges can be handled by CART and whether and under which circumstances it performs better compared to other algorithms.
Two real data applications with data from the NEPS were presented in this thesis. Those two applications can not be evaluated by their performance, especially not in contrast to alternative procedures. Consequently, they can not be generalized. It can only be shown that CART is one (of many) approaches that can handle those tasks that come with the imputation and analysis of the presented data. Additionally, the presented applications are only a selection of many possible applications within a large-scale panel study.

Furthermore, it can always be doubted if the approach used for an application was

the best choice. This is especially relevant for the application on the Thuringia study with CART conducted in combination with data augmentation. It can be said that it was a first attempt to learn more about the functionality of the syntax and a first test for changes and possible combinations with other methods. The fraction of missing values was really low, especially for the second data (from 2011), with 5.0% as maximum percentage of missing. The complete cases reached 93.6% of the whole data information. It can be doubted if the afford of implementing CART combined with a two-step Gibbs sampler was appropriate. Still, it was a very constructive examination.

The second NEPS data (from SC6) was very voluminous. The decision to impute all variables that might be connected to income, that were 213 variables, resulted in a very high effort to implement CART-based MICE. Especially the high filter-depth was a big challenge. In the end, the acceptance of this imputed data was very high by our data users and the application was published in an peer-reviewed article, that is Aßmann et al. (2015). Hence, the scientific interest on this new approach can also be interpreted as high. A special characteristic of CART can be illustrated by this application respective the high amount of variables. The time saving of using CART is very high when compared to approaches which need a definition of a model or model family for each variable used.

Alternative approaches to CART were presented only very roughly. CART was used as workhorse, but other nonparametric approaches as for example Random Forest as conducted by Shah et al. (2014) could have been used as well. However, parametric models are always an alternative and are usually preferred when available and suitable. The usage of CART combined with MI is relatively new, as the article of Burgette and Reiter was published in 2010. The evaluation of this approach is only at the beginning as carved out in chapter 5.
Basically, concerning the illustration of more applications the current literature still lacks a lot. However, the present work contributes to this field of research. Overall, there is considerable need for research. As mentioned above, data characteristics which indicate the usage of CART-based MICE instead of other imputation procedures should be precised. Further research is needed to evaluate

which challenges can be handled by CART and whether and under which circumstances it performs better compared to other algorithms.

Simulation studies, as presented in this thesis lead to results that are dependent on the given settings. To give a general recommendation for the usage of an approach the settings have to be diversified in many possible ways. More precisely, the utilization for cross-sectional, longitudinal and panel data has to be evaluated in more detail with varying settings. The aspect of the best splitting criterion, especially respective the TSP, seems to be worth a second look as it steers the performance of CART. Moreover, all the other settings such as stop criterions should be evaluated as well, even with the risk that it is not possible to give general guidelines for the usage of CART. Additionally, when a simulation study is conducted as a test for real data the simulated data should be adjusted to the properties of the real data as far as possible. The given data situations in both simulation studies are very simplified compared to empirical data from a large-scale panel study. Here, the necessity of further research is obvious. However, the results are very promising as a first insight.

References

Agresti, A., & Coull, B. A. (1998). Apprixomate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119-126.

Akande, O., Li, F., & Reiter, J. (2015). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *ArXiv e-prints*. Retrieved from http://adsabs.harvard.edu/abs/2015arXiv150805918A

Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods Research*, *28*(3), 301-309.

Allison, P. (2001). *Missing data. Thousand Oaks*. Sage.

Alpaydin, E. (2009). *Introduction to Machine Learning (adaptive computation and machine learning series), Second Edition*. The MIT Press.

Amemiya, T. (1971). The estimation of the variance in a variance-components model. *International Economic Review*, *12*(1), 1-13.

Andridge, R., & Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*, 40-64.

Aßmann, C., Carstensen, C. H., Gaasch, C., & Pohl, S. (2014c). Estimation of plausible values using background variables with missing values: A data augmented MCMC approach. *NEPS Working Paper Series. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study*(38). Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXVIII.pdf

Aßmann, C., Goßmann, S., & Schönberger, B. (2014a). Bayesian analysis of binary panel probit models: The case of measurement error and missing values in explaining factors. *NEPS Working Paper Series. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study*(35). Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXV.pdf

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., & Biedermann, A. (2014b). A nonparametric multiple imputation approach for multilevel filtered questionnaires. *NEPS Working Paper Series. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study*(36). Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXVI.pdf

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., & Biedermann, A. (2015). Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the national educational panel study. *Sociological Methods and Research*. Retrieved from http://smr.sagepub.com/content/early/2015/11/02/0049124115610346.abstract doi: 10.1177/0049124115610346

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40-49.

Baccini, M., Cook, S., Frangakis, C., Li, F., Mealli, F., Rubin, D. B., & Zell, E. (2010). Multiple imputation in the anthrax vaccine research program. *CHANCE*, *23*, 16-23.

Bethlehem, J. (2002). Weighting adjustments for ignorable nonresponse. In R. Groves, D. Dillman, J. Elinge, & R. Little (Eds.), *Survey nonresponse* (p. 275-287). New York: John Wiley & Sons.

Blossfeld, H.-P., Roßbach, H.-G., & Maurice, J. v. (Eds.). (2011). *Education as a Lifelong Process – The German National Educational Panel Study (NEPS)* (No. Special Issue 14). VS Verlag.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651-675.

Bosley, J., Dashen, M., & Fox, J. E. (1999). When should we ask follow-up questions about items in lists? *Proceedings of the Section on Survey Research Methods: American Statistical Association*, 749-754. Alexandria, VA: American Statistical Association.

Breiman, L. (1996). Technical note: Some properties of splitting criteria. *Machine Learning*, *24*, 41-47.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. New York: Chapman and Hall.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329-353.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101-133.

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, *30*(1), 160-201.

Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, *172*(9), 1070-1076.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Carpenter, J., Kenward, M., Evans, S., & White, I. (2004). Last observation carry-forward and last observation analysis. *Statistics in Medicine*, *23*, 3241-3244.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, *46*(3), 167-174.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327-335.

Clopper, C. J., & Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413.

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Cook, R., Zeng, L., & Yi, G. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation. *Biometrics*, *60*, 820-828.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.

Dantzig, G., Fulkerson, R., & Johnson, S. (1954). Solutions of a large-scale traveling-salesman problem. *Operations Research*, *2*(4), 393-410.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1-38.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C*, *43*(1), 49-93.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, *72*, 92-104.

Drechsler, J. (2011). Multiple imputation in practice - a case study using a complex german establishment survey. *Advances in Statistical Analysis*, *95*, 1-26.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1-26.

Fu, W., & Simonoff, J. S. (2014). Unbiased regression trees for longitudinal data. Retrieved from http://ssrn.com/abstract=2399976

Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398-409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions and the baysian restoration of images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *6*, 721-741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics* (pp. 169–193). Oxford University Press.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with followups. *Journal of American Statistical Association*, *88*, 984-993.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of followup data. *Journal of Applied Psychology*, *78*, 119?128.

Greene, W. H. (2012). *Econometric analysis (7th edition)*. New York: Pearson Education.

Hawkins, D. M. (1997). Firm: Formal inference-based recursive modeling, pc version, release 2.1. *Technical Report 546, University of Minnesota, School of Statistics*.

Hillygus, D. S., & Schnell, S. (2015). Longitudinal surveys: Issues and opportunities. *The Oxford Handbook of Polling and Polling Methods*. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190213299.001.0001/oxfordhb-9780190213299-e-7 doi: 10.1093/oxfordhb/9780190213299.013.7

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Hsiao, C. (2003). *Analysis of panel data*. Cambridge University Press.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222-230.

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

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*, 119-127.

Kim, H., & Loh, W.-Y. (2001). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, *96*(454), 589-604.

Kish, L. (1990). Weighting: Why, when, and how? *American Statistical Association: Proceedings of the Survey Research Methods Section*, 121-130. Retrieved from https://www.amstat.org/sections/SRMS/Proceedings/papers/1990_018.pdf

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, *8*(2), 183-200.

Kleinke, K., Stemmler, M., Reinecke, J., & Lösel, F. (2011). Efficient ways to impute panel data. *Advances in Statistical Analysis*, *95*(4), 351-372.

Koller-Meinfelder, F. (2009). *Analysis of incomplete survey data - multiple imputation via bayesian bootstrap predictive mean matching*. Otto-Friedrich Universität Bamberg.

Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint and conditional approaches. *Political Analysis*. Retrieved from http://pan.oxfordjournals.org/content/early/2014/04/23/pan.mpu007.full.pdf+html

Laplace, P.-S. (1812). *Théorie analytique des probabilités* (V. Courcier, Ed.). Paris.

Li, F., Yu, Y., & Rubin, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. Retrieved from ftp://stat.duke.edu/pub/WorkingPapers/11-24.pdf

Li, K.-H. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, *30*(1), 57-79.

Li, K.-H., Meng, X.-L., Raghunathan, T., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*(1), 65-92.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198-1202.

Little, R. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, *87*(420), 1227-1237.

Little, R. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association*, *90*(431), 1112-1121.

Little, R. (2013). Discussion: Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 363-366.

Little, R., & Raghunathan, T. E. (1997). Should imputation of missing data condition on all observed variables? *Proceedings of the Survey Research Methods Sections; American Statistical Association*, 617-622.

Little, R., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*, 292-326.

Little, R., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition*. Hoboken, New York: John Wiley & Sons.

Little, R., & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, *22*(9), 1589-1599. doi: doi:10.1002/sim.1513

Little, R., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, *31*(2), 161-168.

Little, R., & Zanganeh, S. (2013). Missing at random and ignorability for inferences about subsets of parameters with missing data. *The University of Michigan Department of Biostatistics Working Paper Series*, *Working Paper 98*. Retrieved from http://biostats.bepress.com/umichbiostat/paper98

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika*, *100*(4), 1-19.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815-840.

Loh, W.-Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with comments). *Journal of the Anmerican Statistical Association*, *83*, 715-728.

Lohr, S. (2009). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lugtig, P. (2014). Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods and Research*. Retrieved from http://smr.sagepub.com/content/early/2014/02/06/0049124113520305.abstract doi: 10.1177/0049124113520305

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538-558.

Messingschlager, M. (2012). *Fehlende Werte in den Sozialwissenschaften - Analyse und Korrektur mit Beispielen aus dem ALLBUS*. Otto-Friedrich Universität Bamberg.

Miller, R. G. (1974). The Jackknife – A Review. *Biometrika*, *61*(1), 1-15.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177-196.

Müller, W., & Wysotzki, F. (1994). Automatic construction of decision trees for classification. *Annals of Operations Research*, *52*(4), 231-247.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica*, *39*(2), 359–382.

Peress, M. (2010). Correcting for survey nonresponse using variable response prospensity. *Journal of the American Statistical Association*, *105*(492), 1418-1430.

Peytchev, A. (2012). Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quaterly*, *76*(2), 214-237.

Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81-106.

Quinlan, J. R. (1992). *C4.5: Programs for machine learning*. Morgan Kaufmann.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler? *Bayesian statistics*, *4*(1), 763-773.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85-95.

Raghunathan, T. E., Solenberger, P., & van Hoewyk, J. (2010). *IVEware*. University of Michigan.

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*(4), 811-822.

Rässler, S., & Schnell, R. (2003). Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study, Zentrum für quantitative Methoden und Surveyforschung, Universität Konstanz. Retrieved from http://www.uni-due.de/%7Ehq0215/documents/2003/2003_MultipleImputationUnitNonresponse.pdf

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227-241.

Royston, P. (2005a). Multiple imputation of missing values: update. *Stata Journal*, *5*(2), 188-201.

Royston, P. (2005b). Multiple imputation of missing values: Update of ice. *Stata Journal*, *5*(4), 527-536.

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

Rubin, D. B. (1977). Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, *72*, 538-543.

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20-28.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130-134.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, *91*(434), 473-489.

Sande, I. G. (1982). Imputation in surveys: Coping with reality. *The American Statistician*, *36*(3a), 145-152.

Särndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton and London and New York and Washington D.C.: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Schnell, R., Hill, P. B., & Esser, E. (2011). *Methoden der empirischen Sozialforschung, 9th Edition*. München: Oldenburg Verlag.

Sedransk, J. (1985). The objective and practice of imputation. *Proceedings of the First Annual Research Conference*, 445-452. Washington, D.C.: Bureau of the Census. Retrieved from http://babel.hathitrust.org/cgi/pt?id=mdp.39015079389774

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, *86*, 169-207.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*. doi: 10.1093/aje/kwt312

Shao, Klein, & Xu. (2012). Imputation for nonresponse in the survey of industrial research and development. *Survey Methodology*, *38*(2), 143-155.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of educational and behavioral statistics*, *38*(5), 499-521.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Strobl, C., Kopf, J., & Zeileis, A. (2013). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, 1-28. Retrieved from http://dx.doi.org/10.1007/s11336-013-9388-3

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, *45*(2), 1-31.

Swamy, P., & Arora, S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica*, *40*(2), 261–275.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Tuyl, F. (2001). Were Clopper & Pearson (1934) too careful? *Proceedings of the Fourth Annual ASEARC Conference*, 17-18 February 2011, University of Western Sydney, Paramatta, Australia. Retrieved from http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1018&context=asearc

Valdiviezo, H. C., & Aelst, S. V. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, *311*, 163 - 181. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025515001838 doi: http://dx.doi.org/10.1016/j.ins.2015.03.018

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully condititional specification. *Statistical Methods in Medical Research*, *16*, 219-242.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman & Hall/CRC.

van Buuren, S., & Groothuis Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

van Buuren, S., & Oudshoorn, K. C. (1999). *Flexible multivariate imputation by mice*. Leiden: TNO Prevention and Health.

van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

Vansteelandt, S., Rotnitzky, A., & Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonginorable nonmonotone nonresponse. *Biometrika*, *94*(4), 841-860.

Wallace, T., & Hussain, A. (1969). The use of error components models in combining cross section with time series data. *Econometrica*, *37*(1), 55–72.

Würbach, A., Hammon, A., Geissler, F., & Goßmann, S. (2014). Data Documentation. Imputed Data File of Starting Cohort 6. *Bamberg: Leibniz Institute for Educational Trajectories (LIfBi), National Educational Panel Study (NEPS)*.

Zinn, S. (2013). An imputation model for multilevel binary data. *NEPS Working Paper Series. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study*(31). Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXI.pdf

Appendix A

List of Abbreviations

ACF	Autocorrelation function
ALWA	Working and learning in a Changing World (Arbeiten und Lernen im Wandel)
BBPMM	Bayesian Bootstrap Predictive Mean Matching
BD	Before Deletion
CART	Classification and Regression Tree
CASMIN	Comparative Analysis of Social Mobility in Industrial Nations
CC	Complete Cases
DPM	Dirichlet Process Mixture model
EM	Expectation-Maximization
FCS	Fully Conditional Specification
fs	Field of subjects
GLM	Generalized Linear Model
IAB	Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung)
ISEI	International Socio-Economic Index of Occupational Status
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings algorithm
MI	Multiple Imputation
MICE	Mulitple Imputation by Chained Equations, sometimes also Multivariate Imputation by Chained Equations
MSE	Mean Squared Error
NA	Not Available
NEPS	National Educational Panel Study
NMAR	Not Missing At Random
NR	Not Relevant
OAR	Observed at random
PMM	Predicitve Mean Matching
PPMM	Posterior Predictive Mean Matching
ROV	Rounding to the nearest Observed Value
RPMM	Rounded Predictive Mean Matching
SUF	Scientific Use File
SC	Starting Cohort
TSP	Traveling Salesman Problem

Appendix B

Figures

B.1 Analysis of unit nonresponse combining CART and data augmentation

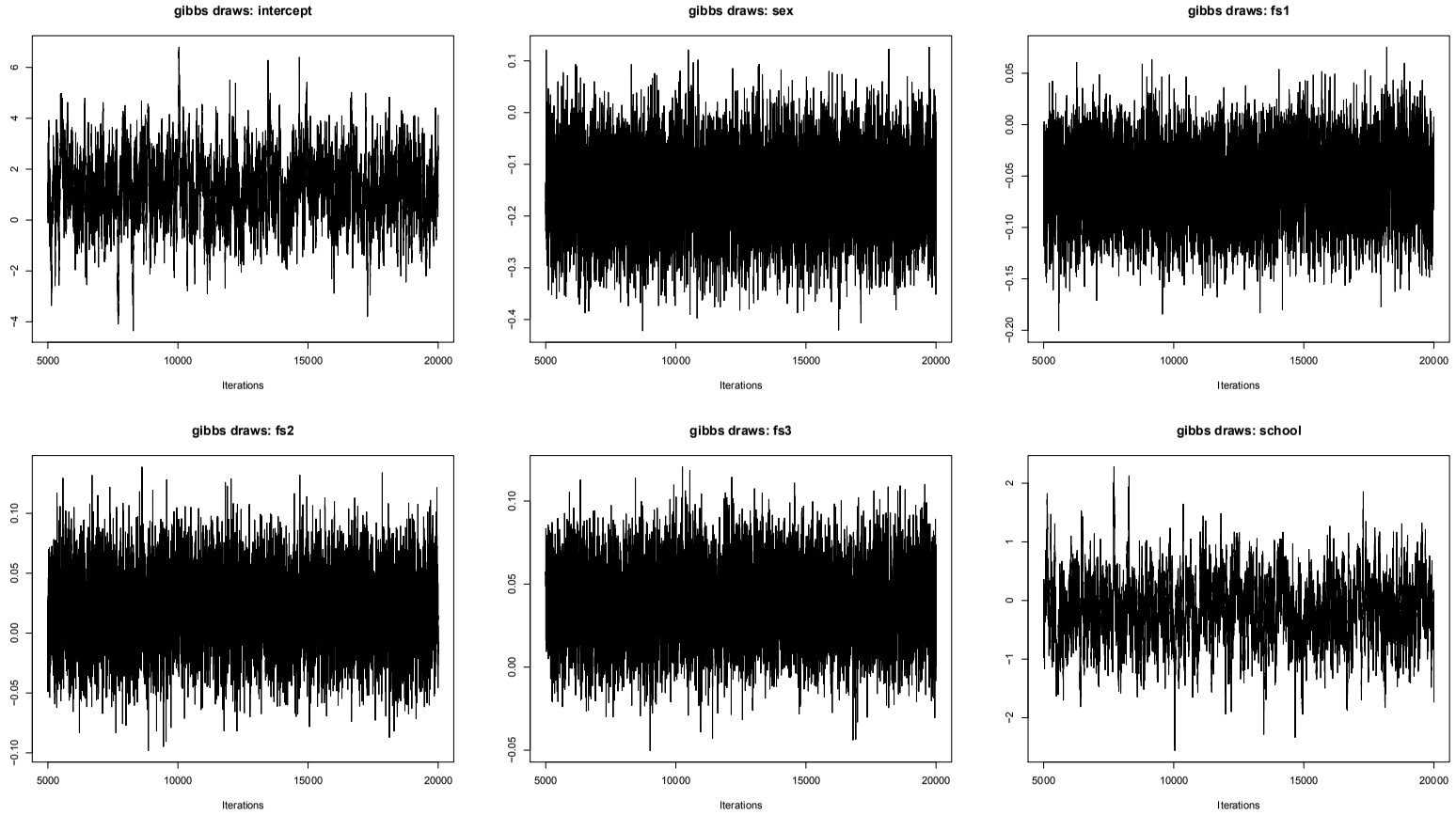

Figure B.1: Draws from the Gibbs sampler

Note: Estimations are based on 20,000 Gibbs iterations, where initial 5,000 draws were discarded for burn-in.

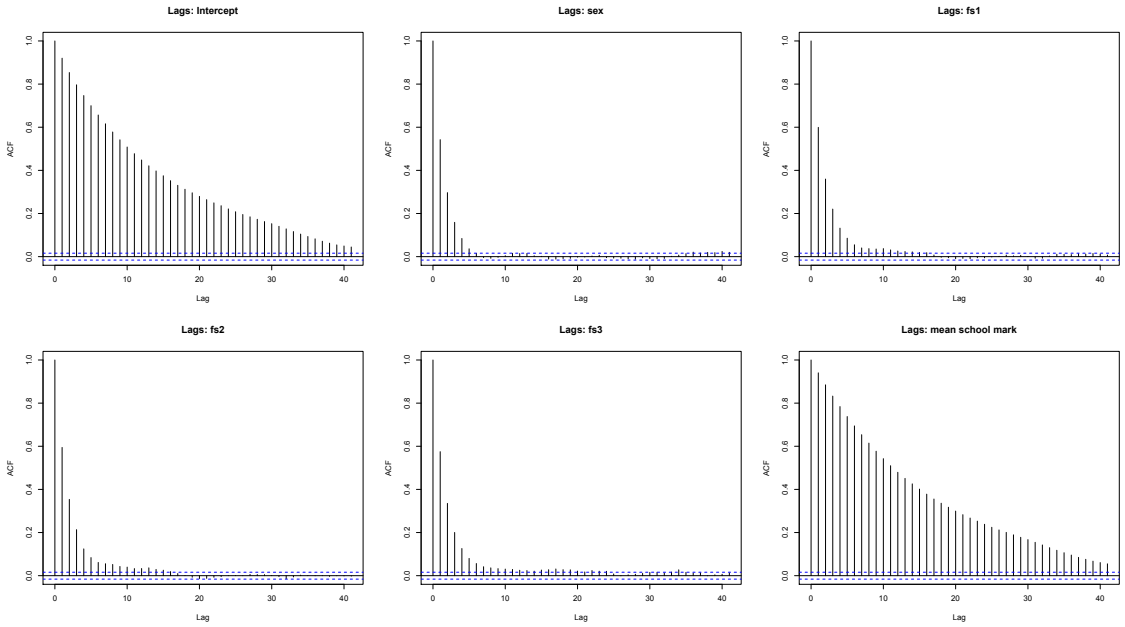

Figure B.2: Plots of the autocorrelation functions (ACF)

Note: Estimations are based on 20,000 Gibbs iterations, where initial 5,000 draws were discarded for burn-in.

B.2 Nonparametric imputation of high-dimensional data containing filters

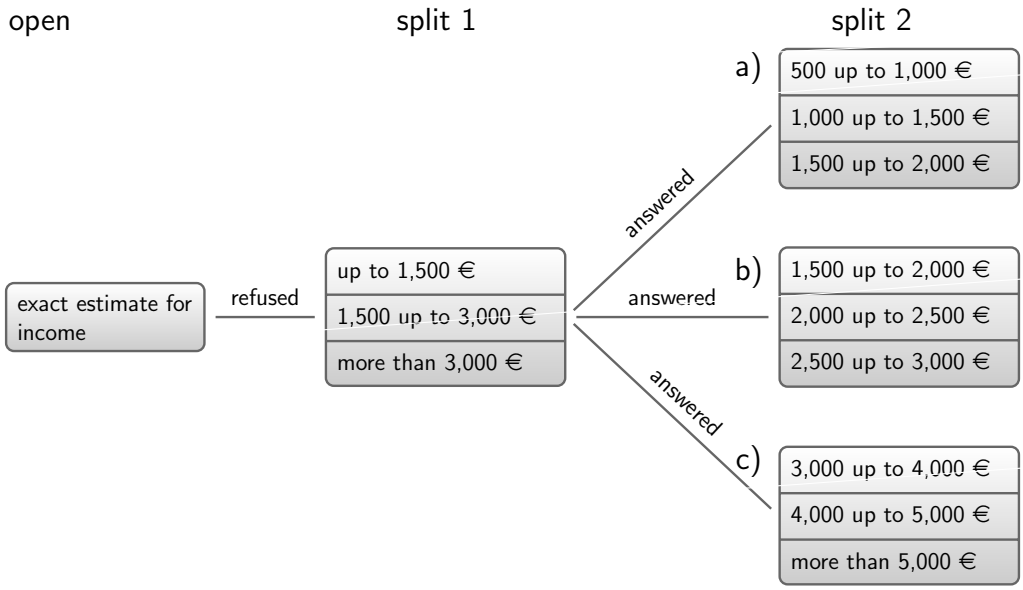

Figure B.3: Income questions in the NEPS SUF SC6 – exact estimate and two-stage income brackets.

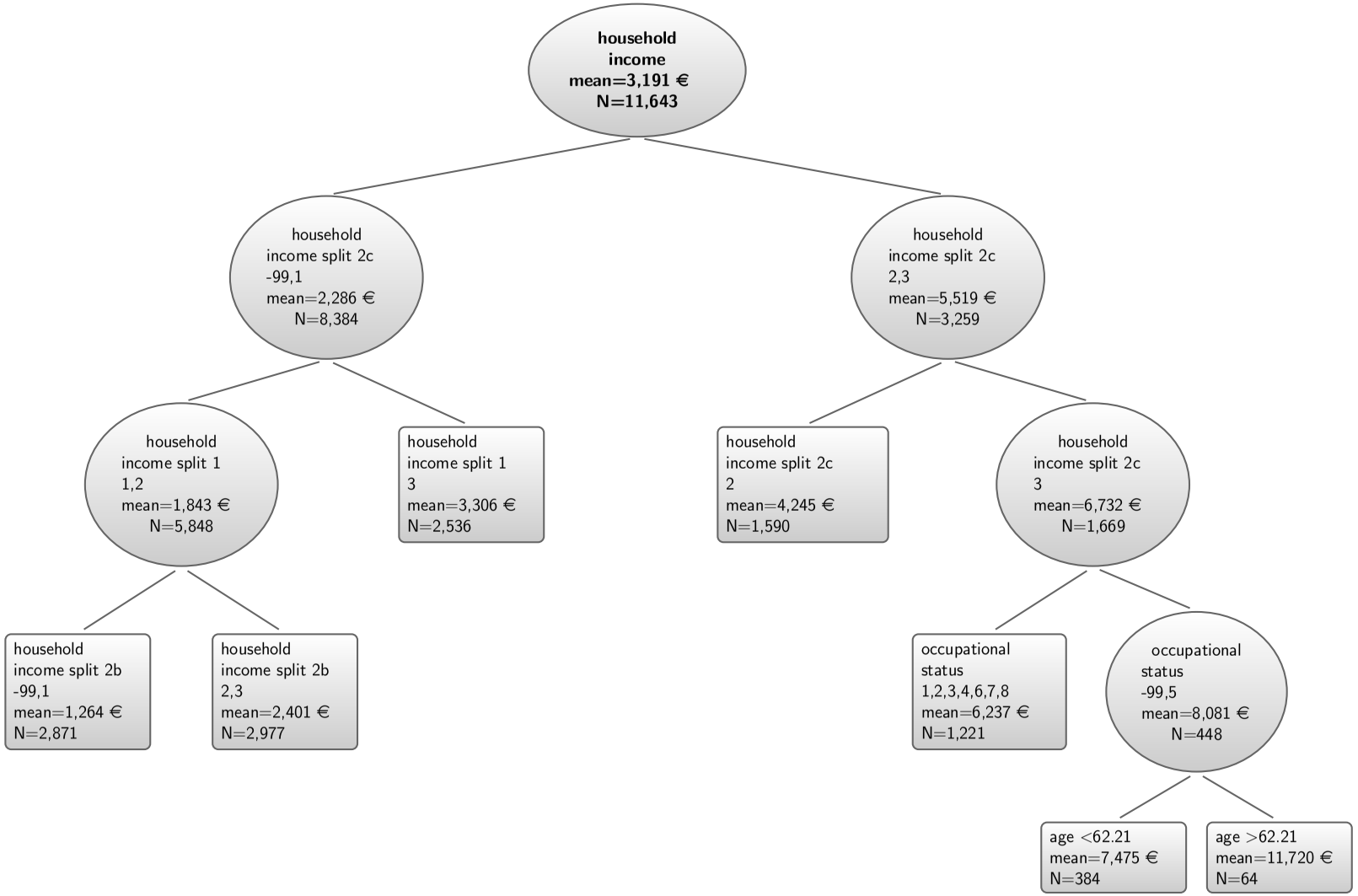

Figure B.4: Household net income imputed via the main panel file and two generated files. Notes: mean is always with reference to household income, *N* is the number of respondents in each node.

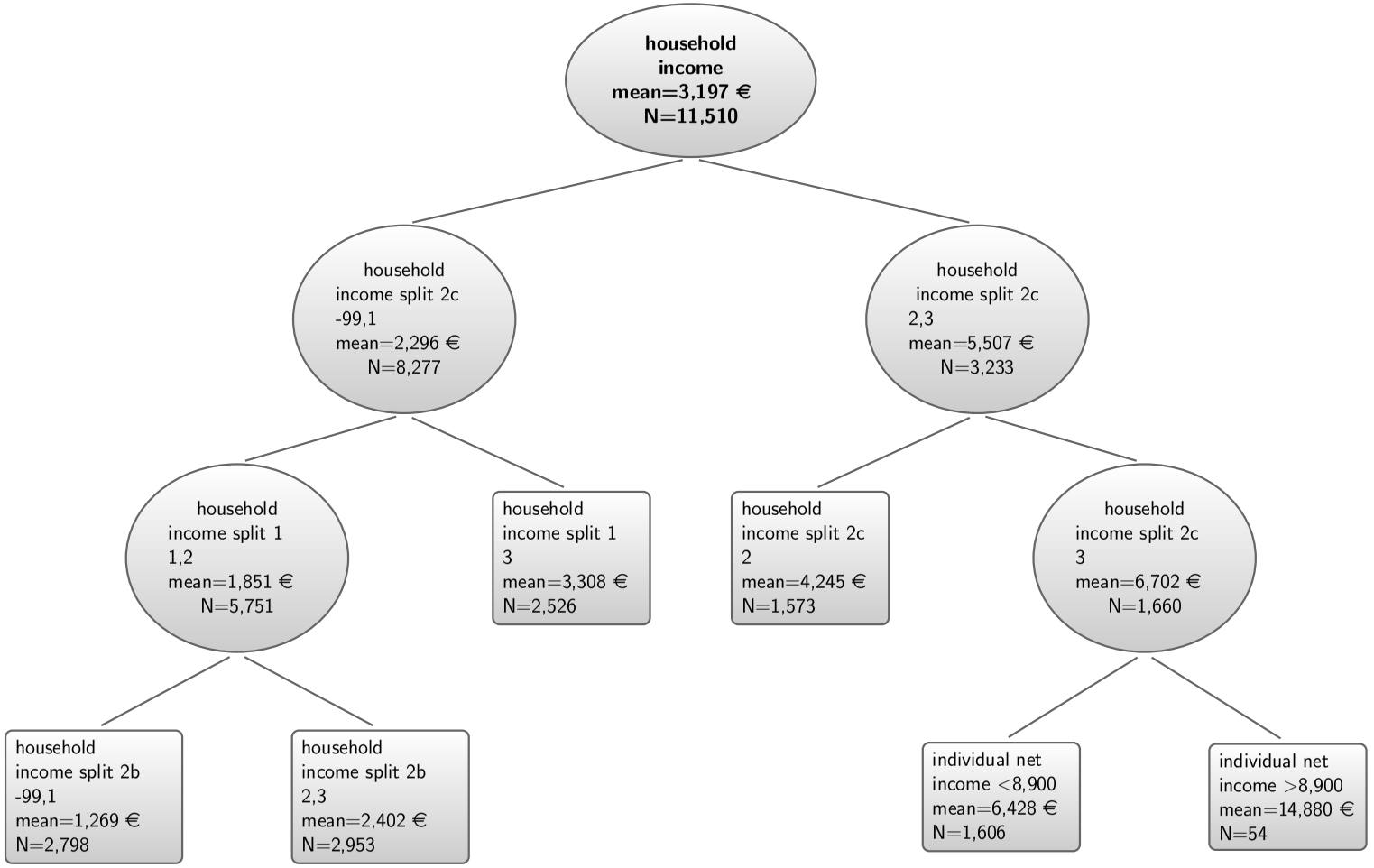

Figure B.5: Household net income imputed via the main panel file and two generated files. Notes: mean is always with reference to household income, *N* is the number of respondents in each node.

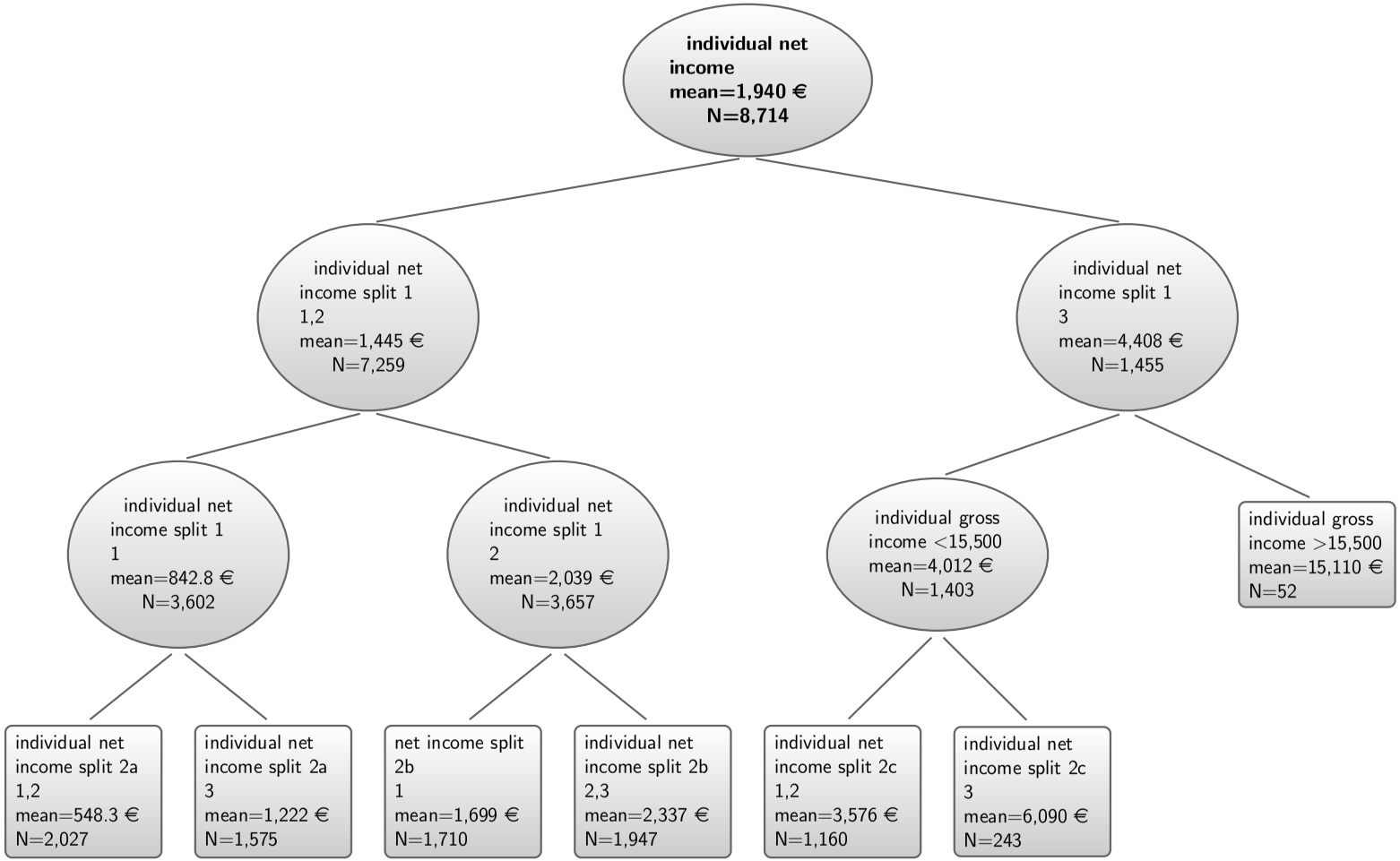

Figure B.6: Individual net income imputed via the main panel file, two generated files, and the module for employment history. Notes: mean is always with reference to individual net income, *N* is the number of respondents in each node.

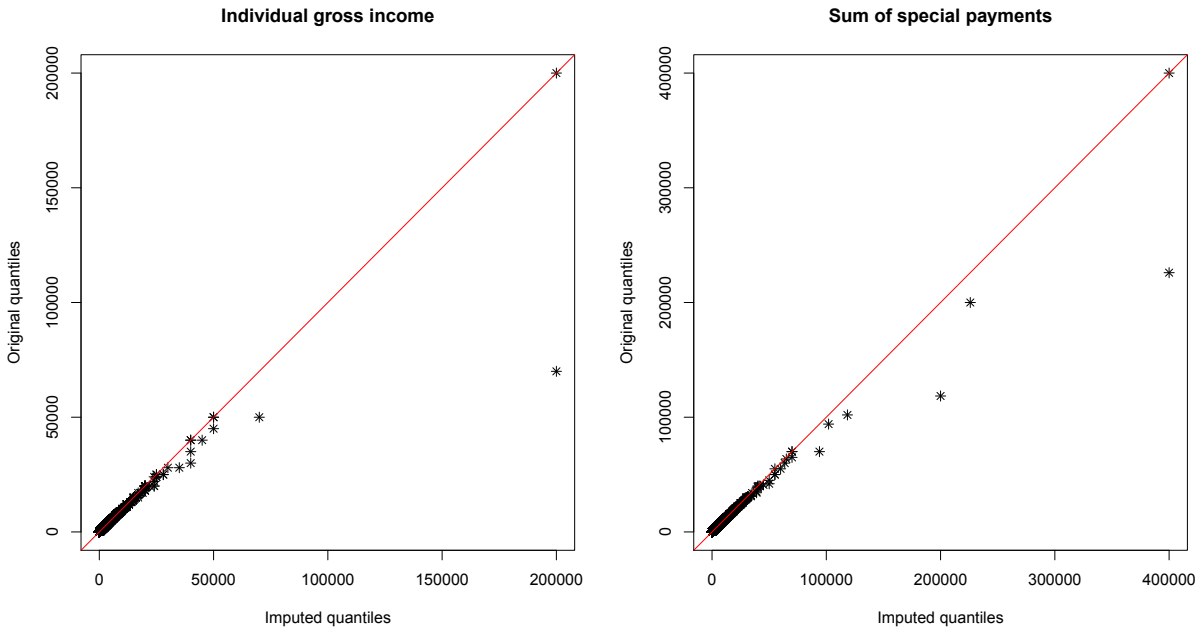

Figure B.7: Q-Q plots for the individual gross income and sum of special payments, variables with significant differences between observed and imputed data according to Kolmogorov-Smirnov goodness of fit test (level of significance: $\alpha = 0.05$).

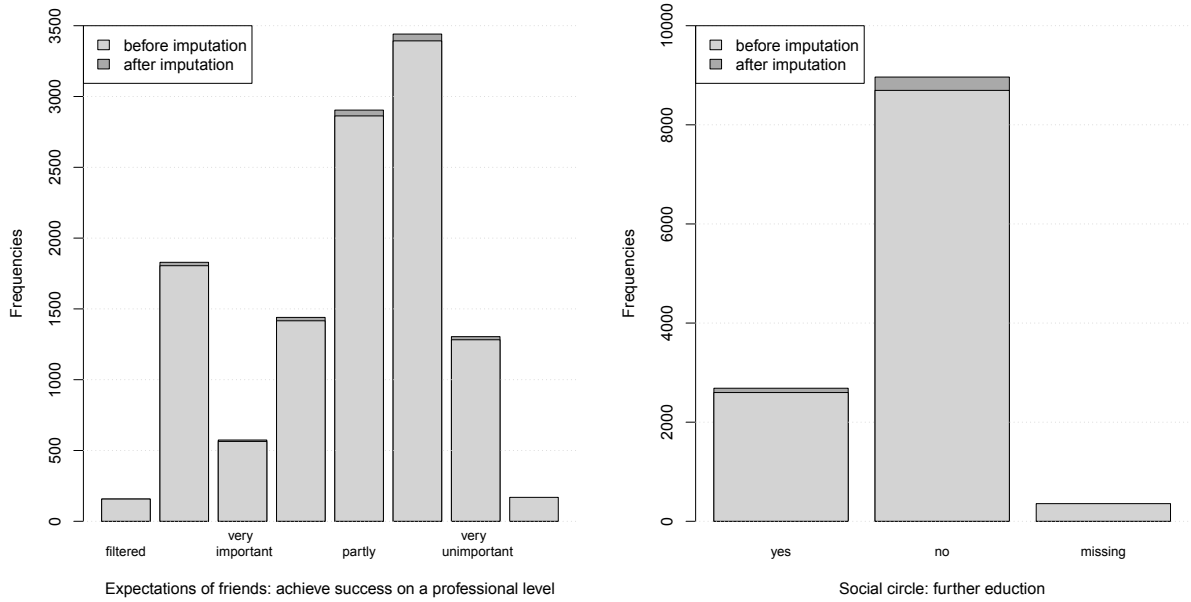

Figure B.8: Column charts for one ordinal variable on the left side and one binary variable on the right side. Observed values are indicated with light gray and imputed values with dark gray. Confidence intervals are too small to be plotted.

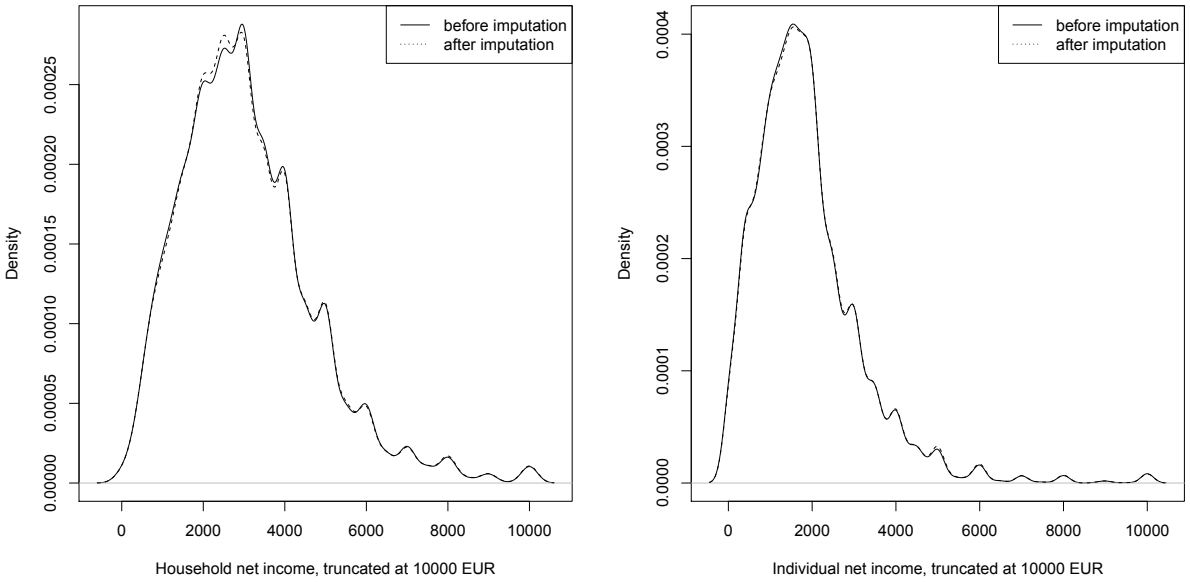

Figure B.9: Kernel densities for household income and individual net income. Solid lines indicate observed data and dashed lines imputed data (bandwidths are: 200 for household income and 150 for individual net income).

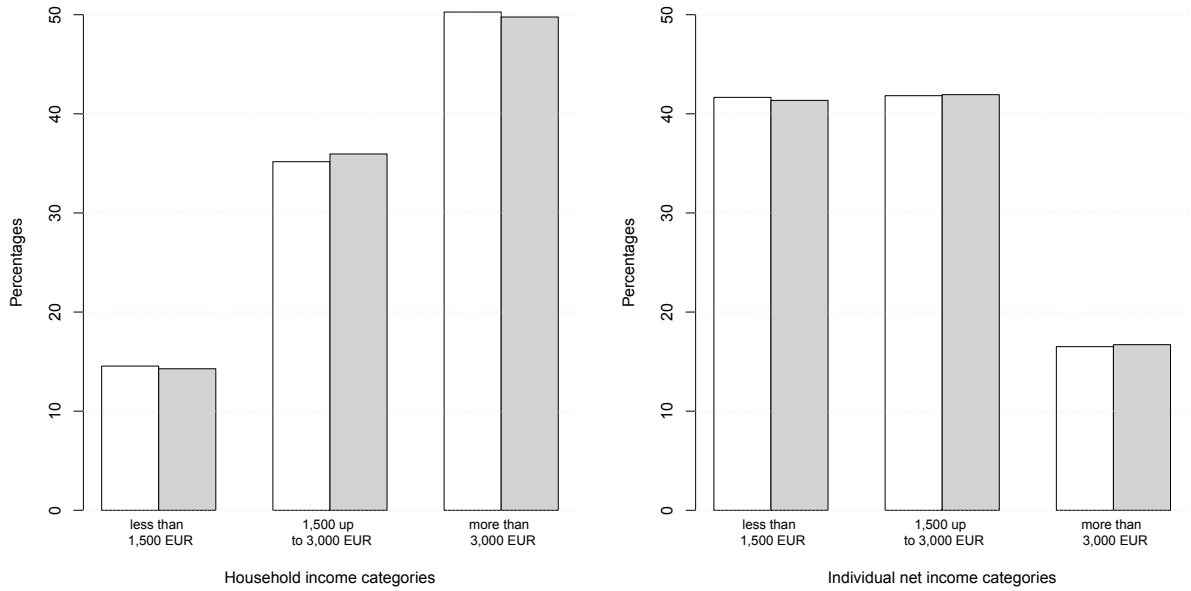

Figure B.10: Classified income information for household income and individual net income. Respondents for which these questions do not apply where excluded. Imputed data are indicated with light gray and observed data white.

B.3 Nonparametric imputation of panel data

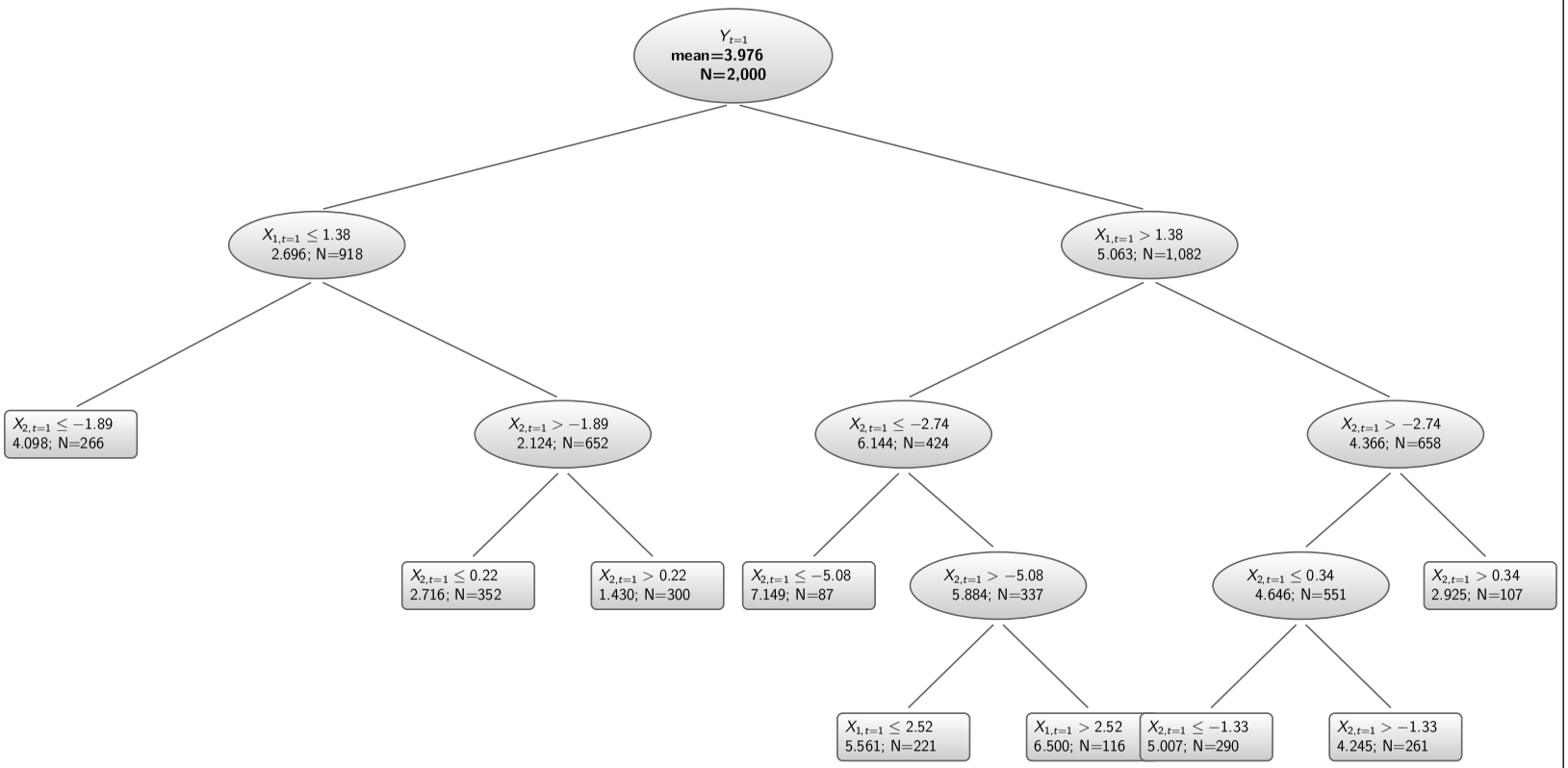

Figure B.11: Resulting tree of one imputation cycle for the imputation of $Y_{t=1}$ of DS2. Notes: the mean is always with reference to $Y_{t=1}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

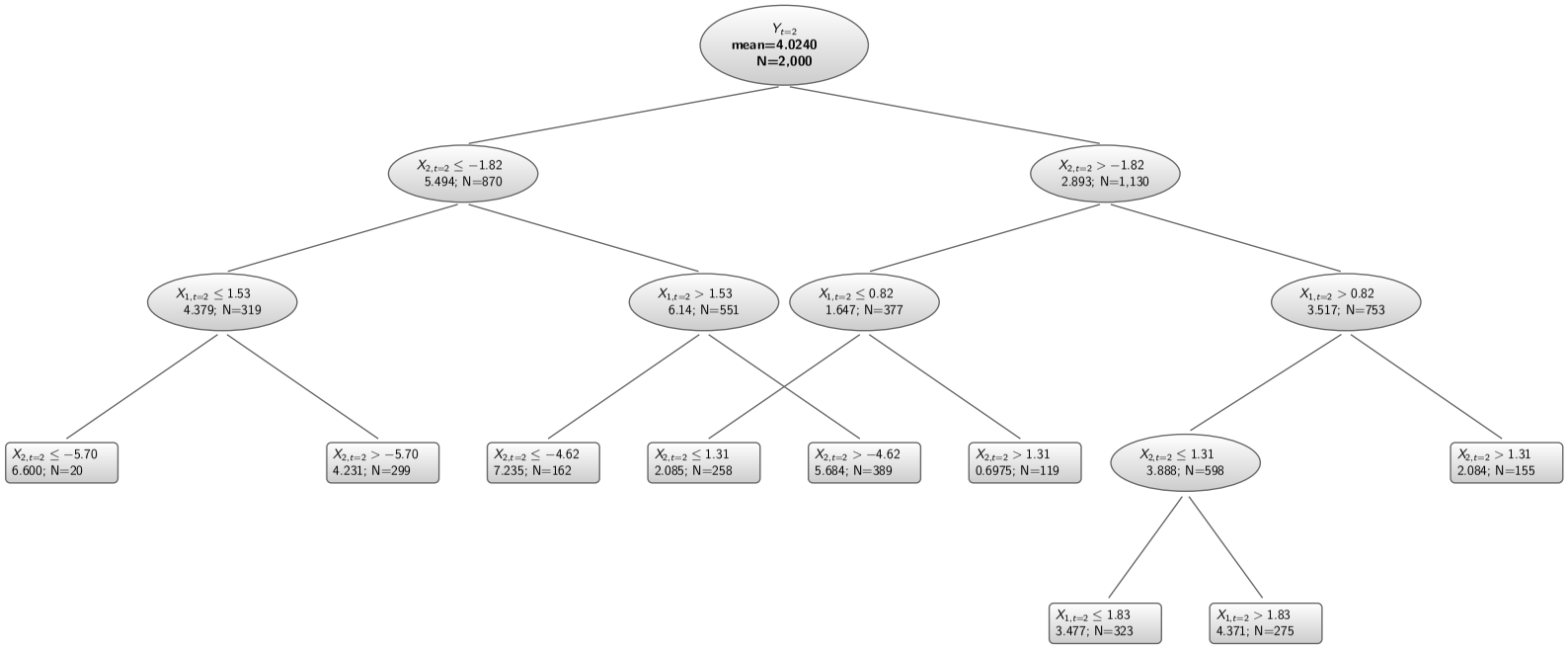

Figure B.12: Resulting tree of one imputation cycle for the imputation of $Y_{t=2}$ of DS2. Notes: the mean is always with reference to $Y_{t=2}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

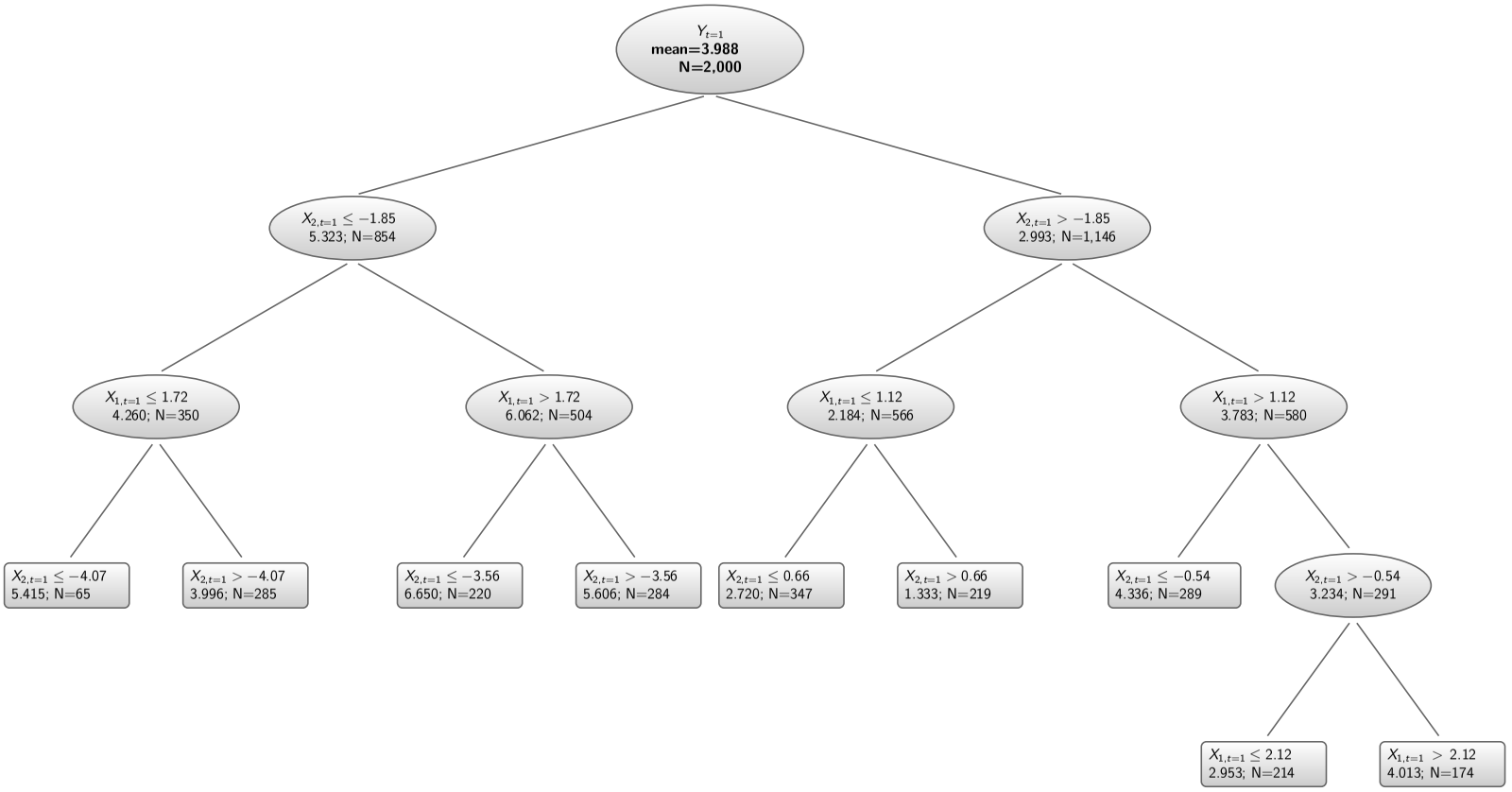

Figure B.13: Resulting tree of one imputation cycle for the imputation of $Y_{t=1}$ of DS3. Notes: the mean is always with reference to $Y_{t=1}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

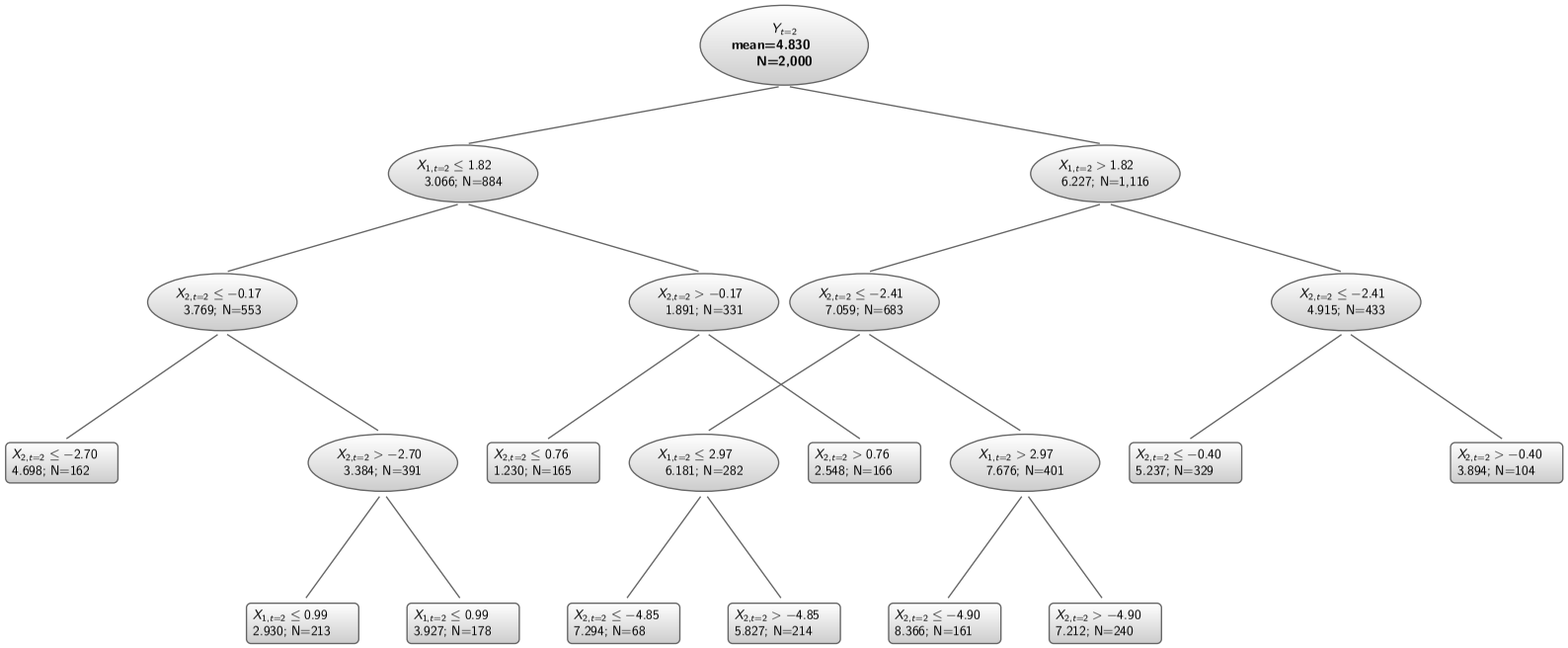

Figure B.14: Resulting tree of one imputation cycle for the imputation of $Y_{t=2}$ of DS3. Notes: the mean is always with reference to $Y_{t=2}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

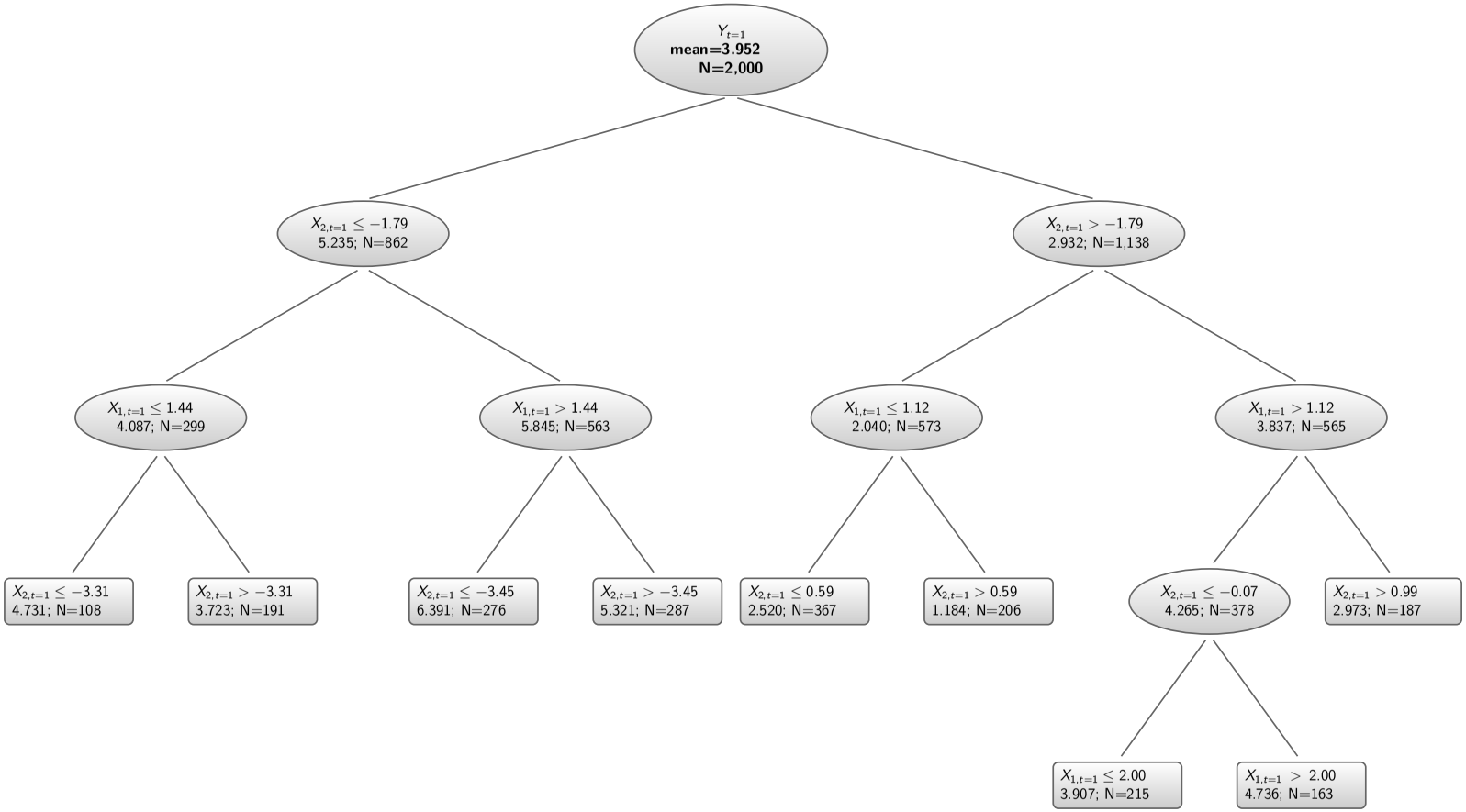

Figure B.15: Resulting tree of one imputation cycle for the imputation of $Y_{t=1}$ of DS4. Notes: the mean is always with reference to $Y_{t=1}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

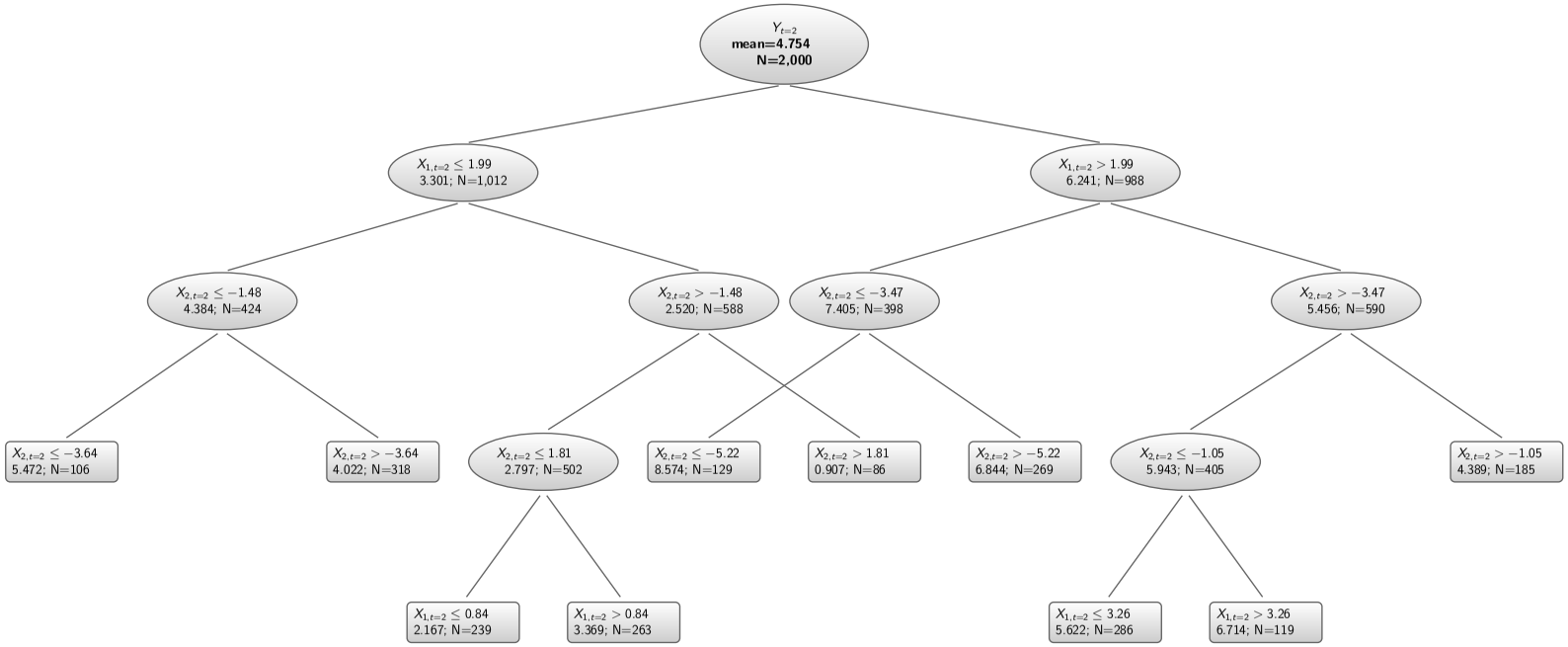

Figure B.16: Resulting tree of one imputation cycle for the imputation of $Y_{t=2}$ of DS4. Notes: the mean is always with reference to $Y_{t=2}$, N is the number of respondents in each node, the split points are rounded to two decimals for better display.

Appendix C

Tables

C.1 Analysis of unit nonresponse combining CART and data augmentation

Field of subject	Subjects contained
(1) linguistic-literary-artistic	German, English, arts, music
(2) social	geography, history, religion
(3) mathematical-natural-scientific-technical	maths, physics, biology, computer sciences

Table C.1: Fields of subjects

	2010				2011			
Standard Probit regression (complete cases: N=1578)					Standard Probit regression (complete cases: N=1304)			
	Estimate	Std. Error	95% CI		Estimate	Std. Error	95% CI	
Intercept	2.2685	0.5812	1.1482	3.4018	0.7200	0.6150	-0.4892	1.9320
sex	-0.1349	0.0763	-0.2842	0.0143	-0.2325	0.0832	-0.3963	-0.0689
fs1	-0.0703	0.0318	-0.1324	-0.0084	-0.0127	0.0340	-0.0792	0.0538
fs2	0.0217	0.0285	-0.0340	0.0773	-0.0141	0.0302	-0.0729	0.0447
fs3	0.0496	0.0211	0.0083	0.0909	0.0468	0.0217	0.0057	0.0879
msm	-0.6324	0.2181	-1.0550	-0.2166	-0.1672	0.2336	-0.6293	0.2934
Bayes Probit incl random effects (complete cases)					Bayes Probit incl random effects (complete cases)			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	1.9000	1.7080	-1.4816	5.2895	-0.8209	2.8212	-6.4948	4.6262
sex	-0.1321	0.0856	-0.2997	0.0366	-0.1885	0.1001	-0.3860	0.0070
fs1	-0.0671	0.0372	-0.1393	0.0065	0.0336	0.0428	-0.0507	0.1165
fs2	0.0280	0.0332	-0.0370	0.0922	-0.0109	0.0392	-0.0864	0.0654
fs3	0.0390	0.0234	-0.0073	0.0843	0.0022	0.0257	-0.0484	0.0526
msm	-0.3698	0.7374	-1.8310	1.1034	0.6009	1.2724	-1.8854	3.0999
σ_u^2	0.6037	0.2326	0.2920	1.1836	1.2770	0.4139	0.6874	2.2772

Table C.2: Comparison of a standard probit model with and without random effects

	2010				2011			
(I.1) Gibbs Cart MH P = 0.01					(I.2) Gibbs Cart MH P = 0.01			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	1.2011	1.3131	-1.3491	3.7941	-0.8278	2.5411	-5.9724	4.1328
sex	-0.1528	0.0749	-0.2975	-0.0058	-0.1912	0.0962	-0.3791	-0.0017
fs1	-0.0582	0.0339	-0.1247	0.0079	0.0286	0.0410	-0.0504	0.1087
fs2	0.0193	0.0309	-0.0417	0.0801	-0.0098	0.0374	-0.0840	0.0622
fs3	0.0387	0.0218	-0.0046	0.0811	0.0105	0.0253	-0.0393	0.0598
msm	-0.1603	0.5586	-1.2717	0.9051	0.5588	1.1597	-1.7002	2.9225
σ_u^2	0.3867	0.1156	0.2195	0.6673	1.1512	0.3549	0.6373	2.0165
(II.1) Gibbs Cart MH P = 0.02					(II.2) Gibbs Cart MH P = 0.02			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	1.1325	1.2841	-1.3911	3.7243	-0.6130	2.5406	-5.6655	4.1640
sex	-0.1530	0.0754	-0.3025	-0.0070	-0.1925	0.0980	-0.3863	-0.0002
fs1	-0.0568	0.0336	-0.1226	0.0090	0.0287	0.0412	-0.0525	0.1098
fs2	0.0178	0.0303	-0.0411	0.0774	-0.0115	0.0376	-0.0845	0.0624
fs3	0.0394	0.0218	-0.0033	0.0820	0.0105	0.0252	-0.0395	0.0596
msm	-0.1318	0.5503	-1.2348	0.9443	0.4624	1.1531	-1.7344	2.7384
σ_u^2	0.3890	0.4201	0.2153	0.6604	1.1531	0.3555	0.6338	2.0121
(III.1) Gibbs Cart MH P = 0.05					(III.2) Gibbs Cart MH P = 0.05			
	Estimate	Std. Error	95% HDR		Estimate	Std. Error	95% HDR	
Intercept	1.2086	1.3278	-1.3657	3.8252	-0.5243	2.4254	-5.2583	4.3361
sex	-0.1509	0.0755	-0.2990	-0.0035	-0.1901	0.0988	-0.3845	0.0028
fs1	-0.0570	0.0338	-0.1233	0.0098	0.0290	0.0409	-0.0514	0.1083
fs2	0.0183	0.0306	-0.0412	0.0785	-0.0107	0.0374	-0.0844	0.0631
fs3	0.0389	0.0216	-0.0034	0.0807	0.0098	0.0252	-0.0401	0.0587
msm	-0.1665	0.5649	-1.2735	0.9383	0.4320	1.0963	-1.7518	2.5616
σ_u^2	0.3867	0.1373	0.2167	0.6627	1.1523	0.3572	0.6388	2.0180

Table C.3: Bayesian Probit estimation with different prior precision
Note: Initial 5,000 draws were discarded for burn-in

C.2 Nonparametric imputation of high-dimensional data containing filters

	n	1st Quartil	Median	3rd Quartil	Mean
Household net income	11,643†	2,000	3,000	4,000	3,192
Individual net income	8,581*	1,000	1,680	2,400	1,929
Individual gross income	8,581*	1,540	2,500	3,800	3,036

† Number of respondents n=11,649, n=6 dropouts at household net income.

* Respondents without an actual employment episode (n=2,975), only a sideline job or an activity with training character (n=93) were excluded from calculation, 11,516 reported in the employment history module.

Table C.4: Descriptives of the NEPS income data

	n	Any income information missing	All income information missing
Household net income	11,643[†]	13.4% (1,556)	3.8% (443)
Individual net income	8,581*	8.0% (695)	2.1% (186)
Individual gross income	8,581*	10.7% (934)	3.5% (309)

[†] Number of respondents n=11,649, n=6 dropouts at household net income.

* Respondents without an actual employment episode (n=2,975), only a sideline job or an activity with training character (n=93) were excluded from calculation, 11,516 reported in the employment history module.

Table C.5: Frequencies of nonresponse in the NEPS income data

C.3 Some insights into the performance of CART

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	-0.0074	0.7213	0.0149
$P(Y < 3)$	0.2033	0.2895	-1.5328	-0.0100
$P(Y < 4)$	0.3943	0.2047	-1.3054	0.1920
$P(Y < 6)$	0.7967	0.0175	-0.5632	0.1092
$\rho(X_1, Y)$	0.7500	0.0517	0.1188	-0.1538
$\rho(X_2, Y)$	-0.8278	0.0473	0.0432	-0.3936
α	1.7501	0.0493	0.0087	0.7671
β_1	1.0000	-0.0570	-0.0041	-0.3909
β_2	-0.5000	0.0061	0.0423	-0.8817
Average	-	0.0669	-0.2746	-0.0830

Table C.6: Relative bias: DS1

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0013	0.0045	0.0016
$P(Y < 3)$	0.2033	0.0001	0.0002	0.0001
$P(Y < 4)$	0.3943	0.0001	0.0003	0.0002
$P(Y < 6)$	0.7967	0.0001	0.0002	0.0001
$\rho(X_1, Y)$	0.7500	0.0001	0.0002	0.0001
$\rho(X_2, Y)$	-0.8278	0.0000	0.0001	0.0001
α	1.7501	0.0006	0.0015	0.0022
β_1	1.0000	0.0003	0.0006	0.0008
β_2	-0.5000	0.0000	0.0001	0.0001
Average	-	0.0003	0.0009	0.0006

Table C.7: Mean squared error: DS1

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	-0.0040	0.6708	0.0352
$P(Y < 3)$	0.2564	0.3293	-1.2043	0.3044
$P(Y < 4)$	0.4357	0.1416	-1.0147	0.1909
$P(Y < 6)$	0.7821	-0.0118	-0.3686	0.0083
$\rho(X_1, Y)$	0.5810	0.1798	0.3324	0.6321
$\rho(X_2, Y)$	-0.6413	0.1206	-0.0564	-0.1216
α	1.7498	-0.2161	-0.6710	-0.5944
β_1	1.0001	0.2086	0.7711	1.1116
β_2	-0.5000	0.0290	-0.1822	-0.6984
Average	-	0.0863	-0.1914	0.0965

Table C.8: Relative bias: DS2

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0024	0.0072	0.0045
$P(Y < 3)$	0.2564	0.0001	0.0003	0.0002
$P(Y < 4)$	0.4357	0.0001	0.0003	0.0002
$P(Y < 6)$	0.7821	0.0001	0.0002	0.0002
$\rho(X_1, Y)$	0.5810	0.0003	0.0007	0.0007
$\rho(X_2, Y)$	-0.6413	0.0003	0.0008	0.0007
α	1.7498	0.0046	0.0104	0.0131
β_1	1.0001	0.0020	0.0046	0.0060
β_2	-0.5000	0.0003	0.0007	0.0009
Average	-	0.0011	0.0028	0.0029

Table C.9: Mean squared error: DS2

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.4907	-0.0310	0.9190	-0.0233
$P(Y < 3)$	0.1335	0.7323	-1.6383	0.7767
$P(Y < 4)$	0.2187	0.4879	-2.0868	0.5288
$P(Y < 6)$	0.7837	0.0403	-0.8790	0.0092
$\rho(X_1, Y)$	0.8784	0.0156	0.0395	0.0440
$\rho(X_2, Y)$	-0.3136	-0.6276	0.0249	0.6719
α	3.9976	-0.0227	0.0800	1.8028
β_1	1.5004	0.0189	-0.0005	-0.2316
β_2	0.2512	0.1037	-0.7078	-14.6252
Average	-	0.0797	-0.4721	-1.2274

Table C.10: Relative bias: DS3

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y)$	4.4907	0.0019	0.0067	0.0022
$P(Y < 3)$	0.1335	0.0001	0.0002	0.0001
$P(Y < 4)$	0.2187	0.0001	0.0002	0.0001
$P(Y < 6)$	0.7837	0.0001	0.0003	0.0001
$\rho(X_1, Y)$	0.8784	0.0000	0.0000	0.0000
$\rho(X_2, Y)$	-0.3136	0.0004	0.0010	0.0005
α	3.9976	0.0059	0.0137	0.0204
β_1	1.5004	0.0001	0.0002	0.0003
β_2	0.2512	0.0015	0.0035	0.0052
Average	-	0.0011	0.0029	0.0032

Table C.11: Mean squared error: DS3

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	95.8	71.0	93.6
$P(Y < 3)$	0.2033	95.0	3.0	93.4
$P(Y < 4)$	0.3943	94.6	3.4	93.6
$P(Y < 6)$	0.7967	95.2	1.6	92.0
$\rho(X_1, Y)$	0.7500	96.6	96.6	91.8
$\rho(X_2, Y)$	-0.8278	96.0	96.4	93.2
α	1.7501	94.2	96.4	90.0
β_1	1.0000	95.0	95.6	88.8
β_2	-0.5000	94.2	93.6	86.0
Average	-	95.2	62.0	91.4

Table C.12: Coverages: DS1, 50 iterations

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0680	0.9195	0.0655
$P(Y < 3)$	0.2033	0.2261	-1.9194	0.0209
$P(Y < 4)$	0.3943	0.0496	-1.9205	-0.0955
$P(Y < 6)$	0.7967	-0.0545	-0.7361	0.0504
$\rho(X_1, Y)$	0.7500	-0.0360	-0.1135	-0.3554
$\rho(X_2, Y)$	-0.8278	0.0238	0.0685	-0.3857
α	1.7501	0.0033	0.0118	0.8732
β_1	1.0000	-0.0148	-0.0201	-0.6231
β_2	-0.5000	0.0163	-0.0169	-0.8069
Average	-	0.0313	-0.4293	-0.1396

Table C.13: Relative bias: DS1, 50 iterations

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0015	0.0049	0.0019
$P(Y < 3)$	0.2033	0.0001	0.0002	0.0001
$P(Y < 4)$	0.3943	0.0001	0.0003	0.0002
$P(Y < 6)$	0.7967	0.0001	0.0002	0.0001
$\rho(X_1, Y)$	0.7500	0.0001	0.0002	0.0002
$\rho(X_2, Y)$	-0.8278	0.0000	0.0001	0.0001
α	1.7501	0.0006	0.0014	0.0020
β_1	1.0000	0.0002	0.0006	0.0008
β_2	-0.5000	0.0000	0.0001	0.0002
Average	-	0.0003	0.0009	0.0006

Table C.14: Mean squared error: DS1, 50 iterations

	True parameters	Coverages in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	96.2	71.4	92.2
$P(Y < 3)$	0.2033	94.8	4.0	92.0
$P(Y < 4)$	0.3943	97.4	3.2	92.6
$P(Y < 6)$	0.7967	95.4	4.4	93.2
$\rho(X_1, Y)$	0.7500	98.0	97.2	93.6
$\rho(X_2, Y)$	-0.8278	97.0	96.6	90.4
α	1.7501	96.6	93.8	87.8
β_1	1.0000	96.0	95.0	90.4
β_2	-0.5000	96.2	95.8	88.8
Average	-	96.4	62.4	91.2

Table C.15: Coverages: DS1, 30 imputations

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0607	0.7741	0.0550
$P(Y < 3)$	0.2033	0.2885	-1.4389	-0.0244
$P(Y < 4)$	0.3943	0.0861	-1.3778	0.0951
$P(Y < 6)$	0.7967	-0.0795	-0.6682	0.0067
$\rho(X_1, Y)$	0.7500	-0.0143	-0.0873	-0.3129
$\rho(X_2, Y)$	-0.8278	-0.0451	-0.0344	-0.4355
α	1.7501	-0.0703	-0.1083	0.6536
β_1	1.0000	0.0915	0.0719	-0.4085
β_2	-0.5000	0.0037	0.0650	-0.7145
Average	-	0.0357	-0.3115	-0.1206

Table C.16: Relative bias: DS1, 30 imputations

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y)$	4.0001	0.0014	0.0049	0.0018
$P(Y < 3)$	0.2033	0.0001	0.0002	0.0001
$P(Y < 4)$	0.3943	0.0001	0.0003	0.0002
$P(Y < 6)$	0.7967	0.0001	0.0002	0.0001
$\rho(X_1, Y)$	0.7500	0.0001	0.0002	0.0001
$\rho(X_2, Y)$	-0.8278	0.0000	0.0001	0.0001
α	1.7501	0.0006	0.0017	0.0023
β_1	1.0000	0.0002	0.0006	0.0008
β_2	-0.5000	0.0000	0.0001	0.0001
Average	-	0.0003	0.0009	0.0006

Table C.17: Mean squared error: DS1, 30 imputations

C.4 Nonparametric imputation of panel data

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	3.9996	-0.0732	0.6875	-0.0493
$P(Y_{t1} < 3)$	0.2110	0.4154	-1.1997	0.2833
$P(Y_{t1} < 4)$	0.3970	0.2663	-1.3102	0.1741
$P(Y_{t1} < 6)$	0.7891	0.0978	-0.5497	0.1310
$\rho(X_{1,t1}, Y_{t1})$	0.7208	0.0230	-0.0114	-0.4178
$\rho(X_{2,t1}, Y_{t1})$	-0.7954	0.0327	0.0134	-0.6290
$E(Y_{t2})$	4.0000	-0.0316	-0.0583	-0.0222
$P(Y_{t2} < 3)$	0.2109	0.3158	1.0073	0.2465
$P(Y_{t2} < 4)$	0.3968	0.2026	0.4733	0.0984
$P(Y_{t2} < 6)$	0.7891	0.0782	0.1110	0.1018
$\rho(X_{1,t2}, Y_{t2})$	0.7207	-0.0225	0.0018	-0.4893
$\rho(X_{2,t2}, Y_{t2})$	-0.7953	0.0562	-0.0144	-0.5618
α	1.7497	0.1167	0.0778	1.0696
β_1	1.0003	-0.0996	-0.0140	-0.6690
β_2	-0.4998	0.0034	-0.0190	-0.9958
Average	-	0.0921	-0.0536	-0.1153

Table C.18: Relative bias: DS1, Panel

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	-0.0364	0.6528	-0.0754
$P(Y_{t1} < 3)$	0.2110	0.4333	-1.1235	0.2892
$P(Y_{t1} < 4)$	0.3968	0.1323	-1.2301	0.2529
$P(Y_{t1} < 6)$	0.7890	0.0916	-0.4777	0.2009
$\rho(X_{1,t1}, Y_{t1})$	0.7205	0.0440	0.1575	-0.3135
$\rho(X_{2,t1}, Y_{t1})$	-0.7953	0.0071	-0.0212	-0.6227
$E(Y_{t2})$	3.9998	0.0208	-0.0052	0.0158
$P(Y_{t2} < 3)$	0.2250	0.3158	1.0072	0.2502
$P(Y_{t2} < 4)$	0.4020	0.0988	0.2594	0.0441
$P(Y_{t2} < 6)$	0.7750	-0.0037	0.0075	-0.0247
$\rho(X_{1,t2}, Y_{t2})$	0.6709	0.0681	0.0854	-0.6806
$\rho(X_{2,t2}, Y_{t2})$	-0.8198	0.0034	0.0241	-0.5120
α	1.7500	0.0350	-0.1287	0.9708
β_1	1.0001	-0.0196	0.0831	-0.7752
β_2	-0.5000	0.0114	0.0506	-0.7788
Average	-	0.0801	-0.0439	-0.1173

Table C.19: Relative bias: DS2, Panel

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	-0.0116	0.7360	-0.0209
$P(Y_{t1} < 3)$	0.2110	0.3088	-1.5025	0.1507
$P(Y_{t1} < 4)$	0.3969	0.1317	-1.4935	0.0275
$P(Y_{t1} < 6)$	0.7891	0.0426	-0.5032	0.1014
$\rho(X_{1,t1}, Y_{t1})$	0.7205	-0.0474	-0.0226	-0.4718
$\rho(X_{2,t1}, Y_{t1})$	-0.7952	-0.0501	-0.0818	-0.6389
$E(Y_{t2})$	4.7506	-0.0220	-0.0353	-0.0283
$P(Y_{t2} < 3)$	0.1581	0.6017	1.5524	0.5614
$P(Y_{t2} < 4)$	0.2983	0.4662	0.7945	0.2350
$P(Y_{t2} < 6)$	0.6225	0.1123	0.3006	0.2191
$\rho(X_{1,t2}, Y_{t2})$	0.8108	0.0215	0.0780	-0.2731
$\rho(X_{2,t2}, Y_{t2})$	-0.8109	0.0176	0.0313	-0.6760
α	1.7504	-0.0219	-0.1095	0.7065
β_1	0.9998	0.0260	0.1103	-0.1882
β_2	-0.5001	0.0090	0.0509	-1.0604
Average	-	0.1056	-0.0063	-0.0904

Table C.20: Relative bias: DS3, Panel

	True parameters	Relative Bias in %		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0005	0.0096	0.7502	-0.0066
$P(Y_{t1} < 3)$	0.2088	0.4211	-1.0521	0.4045
$P(Y_{t1} < 4)$	0.4005	0.1445	-1.3605	0.1268
$P(Y_{t1} < 6)$	0.7927	0.0049	-0.6098	0.0636
$\rho(X_{1,t1}, Y_{t1})$	0.7319	0.0034	0.0180	-0.4707
$\rho(X_{2,t1}, Y_{t1})$	-0.8079	-0.0859	-0.0033	-0.6735
$E(Y_{t2})$	4.7499	0.0118	-0.0556	-0.0185
$P(Y_{t2} < 3)$	0.1560	0.3313	1.3349	0.3098
$P(Y_{t2} < 4)$	0.2976	0.2348	0.7109	0.2862
$P(Y_{t2} < 6)$	0.6228	0.0443	0.1887	0.1204
$\rho(X_{1,t2}, Y_{t2})$	0.8200	-0.0437	-0.1097	-0.3210
$\rho(X_{2,t2}, Y_{t2})$	-0.8199	0.0310	-0.0628	-0.7056
α	—	—	—	—
β_1	1.0001	-0.0422	0.0142	-0.3765
β_2	-0.5001	0.1261	0.0635	-1.0416
Average	-	0.0009	-0.0001	-0.0016

Table C.21: Relative bias: DS4, Panel

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	3.9996	0.0016	0.0050	0.0022
$P(Y_{t1} < 3)$	0.2110	0.0001	0.0002	0.0001
$P(Y_{t1} < 4)$	0.3970	0.0001	0.0003	0.0002
$P(Y_{t1} < 6)$	0.7891	0.0001	0.0002	0.0001
$\rho(X_{1,t1}, Y_{t1})$	0.7208	0.0001	0.0002	0.0002
$\rho(X_{2,t1}, Y_{t1})$	-0.7954	0.0001	0.0002	0.0002
$E(Y_{t2})$	4.0000	0.0016	0.0037	0.0022
$P(Y_{t2} < 3)$	0.2109	0.0001	0.0002	0.0002
$P(Y_{t2} < 4)$	0.3968	0.0001	0.0003	0.0002
$P(Y_{t2} < 6)$	0.7891	0.0001	0.0002	0.0002
$\rho(X_{1,t2}, Y_{t2})$	0.7207	0.0001	0.0002	0.0002
$\rho(X_{2,t2}, Y_{t2})$	-0.7953	0.0001	0.0002	0.0002
α	1.7497	0.0006	0.0015	0.0023
β_1	1.0003	0.0002	0.0006	0.0008
β_2	-0.4998	0.0000	0.0001	0.0002
Average	-	0.0003	0.0009	0.0006

Table C.22: Mean squared error: DS1, Panel

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	0.0017	0.0046	0.0022
$P(Y_{t1} < 3)$	0.2110	0.0001	0.0002	0.0002
$P(Y_{t1} < 4)$	0.3968	0.0001	0.0003	0.0002
$P(Y_{t1} < 6)$	0.7890	0.0001	0.0002	0.0001
$\rho(X_{1,t1}, Y_{t1})$	0.7205	0.0001	0.0003	0.0002
$\rho(X_{2,t1}, Y_{t1})$	-0.7953	0.0001	0.0002	0.0002
$E(Y_{t2})$	3.9998	0.0018	0.0044	0.0027
$P(Y_{t2} < 3)$	0.2250	0.0001	0.0002	0.0002
$P(Y_{t2} < 4)$	0.4020	0.0001	0.0003	0.0002
$P(Y_{t2} < 6)$	0.7750	0.0001	0.0002	0.0001
$\rho(X_{1,t2}, Y_{t2})$	0.6709	0.0001	0.0003	0.0002
$\rho(X_{2,t2}, Y_{t2})$	-0.8198	0.0001	0.0001	0.0001
α	1.7500	0.0006	0.0015	0.0025
β_1	1.0001	0.0002	0.0005	0.0008
β_2	-0.5000	0.0000	0.0001	0.0001
Average	-	0.0004	0.0009	0.0007

Table C.23: Mean squared error: DS2, Panel

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0001	0.0015	0.0048	0.0022
$P(Y_{t1} < 3)$	0.2110	0.0001	0.0002	0.0001
$P(Y_{t1} < 4)$	0.3969	0.0001	0.0003	0.0002
$P(Y_{t1} < 6)$	0.7891	0.0001	0.0002	0.0002
$\rho(X_{1,t1}, Y_{t1})$	0.7205	0.0001	0.0002	0.0002
$\rho(X_{2,t1}, Y_{t1})$	-0.7952	0.0001	0.0002	0.0002
$E(Y_{t2})$	4.7506	0.0023	0.0058	0.0031
$P(Y_{t2} < 3)$	0.1581	0.0001	0.0002	0.0001
$P(Y_{t2} < 4)$	0.2983	0.0001	0.0003	0.0002
$P(Y_{t2} < 6)$	0.6225	0.0001	0.0003	0.0002
$\rho(X_{1,t2}, Y_{t2})$	0.8108	0.0000	0.0001	0.0001
$\rho(X_{2,t2}, Y_{t2})$	-0.8109	0.0001	0.0001	0.0001
α	1.7504	0.0006	0.0016	0.0023
β_1	0.9998	0.0002	0.0004	0.0006
β_2	-0.5001	0.0000	0.0001	0.0002
Average	-	0.0004	0.0010	0.0007

Table C.24: Mean squared error: DS3, Panel

	True parameters	Mean squared error		
		BD	(CC)	CART-MICE
$E(Y_{t1})$	4.0005	0.0015	0.0044	0.0020
$P(Y_{t1} < 3)$	0.2088	0.0001	0.0002	0.0001
$P(Y_{t1} < 4)$	0.4005	0.0001	0.0003	0.0002
$P(Y_{t1} < 6)$	0.7927	0.0001	0.0002	0.0001
$\rho(X_{1,t1}, Y_{t1})$	0.7319	0.0001	0.0002	0.0002
$\rho(X_{2,t1}, Y_{t1})$	-0.8079	0.0001	0.0002	0.0002
$E(Y_{t2})$	4.7499	0.0023	0.0053	0.0028
$P(Y_{t2} < 3)$	0.1560	0.0001	0.0002	0.0001
$P(Y_{t2} < 4)$	0.2976	0.0001	0.0002	0.0001
$P(Y_{t2} < 6)$	0.6228	0.0001	0.0003	0.0002
$\rho(X_{1,t2}, Y_{t2})$	0.8200	0.0000	0.0001	0.0001
$\rho(X_{2,t2}, Y_{t2})$	-0.8199	0.0001	0.0001	0.0001
α	—	—	—	—
β_1	1.0001	0.0003	0.0007	0.0006
β_2	-0.5001	0.0001	0.0002	0.0001
Average	-	0.0003	0.0009	0.0005

Table C.25: Mean squared error: DS4, Panel