NEPS WORKING PAPERS

Christian Aßmann, Ariane Würbach, Solange Goßmann, Ferdinand Geissler, Anika Biedermann

# A NONPARAMETRIC MULTIPLE IMPUTATION APPROACH FOR MULTILEVEL FILTERED QUESTIONNAIRES

NEPS Working Paper No. 36
Bamberg, March 2014

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

**Working Papers of the German National Educational Panel Study (NEPS)**
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at
**https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers**



**Contact**: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

# A nonparametric multiple imputation approach for multilevel filtered questionnaires

*Christian Aßmann, Ariane Würbach, Solange Goßmann,*
*Ferdinand Geissler, Anika Biedermann*
*Leibniz Institute for Educational Trajectories, Germany*
*Humboldt-Universität zu Berlin, Germany*

March 2014

**E-Mail-Adresse des Autors:**
christian.assmann@uni-bamberg.de

**Bibliographische Angaben:**

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F. & Biedermann, A. (2014). *A nonparametric multiple imputation approach for multilevel filtered questionnaires* (NEPS Working Paper No. 36). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

# A nonparametric multiple imputation approach for multilevel filtered questionnaires

**Abstract**

Despite high efforts in field work and questionnaire design, low rates of missing values inevitably occur. The principles of multiple imputation allow for addressing this issue enhancing the analytical potential of the surveyed data. Large scale surveys provide rich data structures characterized by manifold discrete variables in combination with multilevel filtering in questionnaires. This requires multiple imputation techniques to preserve possible nonlinear relationships among the surveyed variables and full conditional distributions incorporating the information from multilevel filtering rules on an individual basis. To meet these requirements, a tree-based sequential regression approach is adapted addressing both the issues of possibly nonlinear relationships between categorical variables and complex multilevel filtering. Handling of filters within imputation is thereby adapted in a way to ensure consistency of the sequence of full conditional distributions. The suggested approach is illustrated in the context of income imputation in the adult cohort of the National Educational Panel Study.

**Keywords**

Nonparametric Imputation; CART; Missing Income Values; Filtering

# 1. Introduction

High data quality is decisive to all empirical analyses based on survey data. As surveys usually fail to provide complete information due to nonresponse, missing data has to be taken into account when analyzing data. In general, unit-nonresponse and item-nonresponse are distinguished. While unit-nonresponse occurs if individuals refuse to participate in the survey or the attempt to contact the respondent fails, item-nonresponse arises, besides technical errors, if respondents are not able or not willing to give a valid answer. This applies to most either sensitive or difficult questions. Especially when missing data can not be characterized as completely at random ($MCAR$), see Rubin (1976), convenient approaches like reducing the data set to complete observations, i.e. listwise deletion, are problematic both in terms of bias and variance of estimators. Rubin (1976) further distinguishes missing values occurring at random ($MAR$) or not at random ($MNAR$). If factors can be identified which influence the occurrence of missing data (in case of $MAR$), analyses should be based on multiply imputed data sets to enhance the analytical potential of the data.

In fact, the use of imputation techniques to deal with item-nonresponse is the common strategy pursued by almost all large panel surveys, where the literature documents a large variety of multiple imputation approaches. For example, with regard to imputation of missing values in household income, the German Socio-Economic Panel Study (SOEP) applies the regression based row-and-column method using information from previous and current waves by Little and Su (1989), see Frick and Grabka (2005). As information from previous waves is required, alternative strategies are necessary, if no information from preceding waves is available. Among these, Frick and Grabka (2005) list logical imputation, median substitution, median share substitution and regression-based substitution. The Swiss Household Panel (SHP) uses a variant of the approach of Little and Su (1989), when longitudinal information is available, see Lipps (2010). Following Watson and Starick (2011), imputation classes are build by age and education to obtain more similar characteristics between source and target of imputation. Moreover, a carryover procedure is adapted which is mainly used for pension variables. In this approach the value from the closest previous wave with a reported value is imputed without modification.[1] If no data from previous waves is available, the imputation process is initialized by a regression-based imputation using cross-sectional data. As Hayes and Watson (2009) report, in the Household, Income and Labour Dynamics in Australia (HILDA) Survey predominantly an extension of the approach of Little and Su (1989) by imputation classes by age is applied. Additionally, predictive mean matching as introduced by Little (1988) is used if no longitudinal information is available. According to Jenkins (2010), the British Household Panel survey (BHP) mainly applies either predictive mean matching for continuous variables or a hotdeck approach for categorical variables. For the imputation of missing data in the Panel Study of Income Dynamics (PSID), Duffy (2011) mentions the predominant application of hotdeck methods, but also prior wave carryovers or median substitution for some income variables. Each imputation method thereby imposes implementation problems of its own when put on real data, e.g., addressing filtering and logical constraints, categorical variables as well as mutual dependencies among the model variables.

The National Educational Panel Study (NEPS) aims to provide data allowing for investigation of various educational research questions. As the survey questionnaires of the NEPS are designed and orientated along central dimensions, e.g. educational decisions or returns to education, the collected data exhibit rich mutual dependencies among the often categorical surveyed variables. Further, as is typical for large-scale survey questionnaires, the surveyed data incorporates a complex filter structure. In order to cope with the dependence structures in combination with multi-

[1] However, Lipps (2010) mentions the consideration of an inflation factor to account for inflation of old age pensions, for example.

level filtering, we adapt the nonparametric tree-based sequential regression approach to impute missing values as proposed by Burgette and Reiter (2010). Other studies using such classification and regression trees for the purpose of multiple imputation are among others Drechsler and Reiter (2011), Burgette and Reiter (2012), and Hapfelmeier et al. (2012). Focusing on imputation of missing values in income variables, we establish a highly flexible initialization scheme allowing for treatment of a wide range of missing patterns and handling the filtering structure within the set of reference variables considered for income imputation. The filter structure thereby restricts the range of admissible values for imputation at the individual level, i.e. for each individual missing value the admissible range for imputed values may differ due to the individual filter structure. The suggested imputation approach is based on nonparametric characterizations of the full conditional distributions for the missing values taking the restrictions arising from filters at the individual level into account. The nonparametric characterizations arise from sequential regression trees documenting the set of valid conditioning factors. Via intersecting the ranges of the nonparametric distributions with the range of admissible values arising from the individual filter structure, the filter information available on an individual level is consistently incorporated in the characterization of the full conditional distributions used to generate imputations. Next to defining restrictions at the individual level on the admissible range the filter structure defines a hierarchy concerning the sequence of variables within the imputation procedures. Sticking to this hierarchy consequently ensures the necessary consistency of the set of full conditional distributions as pointed out by Si and Reiter (2013).

With regard to imputation of missing values in net household income, we find typical variables like bracketed information, occupational status, and age as important predictors. In combination with chained equations as discussed by van Buuren and Groothuis Oudshoorn (2011), we provide multiple imputed data sets thus accounting for the caveat entered by Allison (2002) that single imputation results in underestimation of standard errors and biased test statistics.

The paper is organized as follows. Chapter 2. demonstrates the incidence of missing values on income questions within the NEPS and discusses factors influencing the occurrence of missing values, while chapter 3. describes the adapted imputation method and our strategy to handle filters. In chapter 4. the presented method is applied for imputation of missing values in income variables. Finally, chapter 5. concludes.

## 2. Factors influencing missing values in NEPS income data

Especially with regard to the measurement of income, item-nonresponse is of crucial importance. Questions about earnings are known to be very sensitive for many persons and therefore often lead to relatively high rates of item-nonresponse, see e.g. Riphahn and Serfling (2005). In the SOEP, Frick and Grabka (2007) document missing information on household income in 8 % of the cases. For the collection of labor income, which is composed by up to ten different items, Frick and Grabka (2007) report a share of 14 % of the observations with at least some information missing. Similar numbers are found by Jenkins (2010) in the BHPS with a rate of 15 % missing data on labor income questions and by Hayes and Watson (2009) in HILDA with about 16 % missing values on at least one income question in the first wave.[2]

For illustration, we take a closer look at the household net income as our main variable of interest.[3] In the data of NEPS Starting Cohort 6 – adults (SC6 version 1.0.0, DOI: http://dx.doi.org / 10.5157 / NEPS:SC6:1.0.0) we find missing data on household income questions in 13 % of the

[2] On the household-level, Hayes and Watson (2009) report a rate of at least one missing value on income questions in the first wave of almost 30 %.

[3] While all considered variables are imputed simultaneously, we orientate the maximum set of variables considered within imputation towards meaningful predictors for household income.

cases. However, if respondents refuse to answer the first open question about their household income, they are asked bracketed income questions. Having answered this request, they are asked to assign once more into narrower income brackets. Following this strategy, we get detailed categorized information for 8 % of the whole sample and another 2 % with at least rough information about the household income. Hence, there remain about 4 % of the respondents for whom we have no income information at the household level at all. Note that the information attained via bracketed questions provide valuable extra information and has to be incorporated on an individual level into the conditional distribution used for drawing imputations via restricting the range of admissible values.

Although the share of missing data on household income questions is relatively low, item-nonresponse could bias the estimates if it occurs selectively. To exemplify this missing mechanism, Table 1 presents results from two probit-models estimating the probability of having at least one missing value on household income questions (model $I$) and having missing values on all household income questions (model $II$). According to model $I$, we find an effect of gender with women tending to be more likely to have at least one missing value on income questions at the household level. Furthermore, older respondents tend to be more likely to have missing values. Living in a household with other adults increases the probability for nonresponse, whereas the number of additional children living in the household is not important. Moreover, the occupational status matters with civil servants being less likely to have a missing value on household income questions especially with regard to employees, self-employed and persons who are not working at all. For respondents not born in Germany information on household income tend to lack more often. Finally, the satisfaction with the financial situation influences the response rate as well: if interviewees are less satisfied with their financial situation, they are less likely to report an exact estimate of their household net income. We do not find a significant effect of unemployment, education (measured via CASMIN) or prestige of the occupational position (measured via ISEI). The size of the living area does not influence the probability of item-nonresponse but we find a negative effect for respondents being located within East Germany. We add a missing indicator variable for missing values on the covariates which can be understood as a measurement of the general tendency to refuse to answer questions. Although we are aware of the fact that a missing indicator variable may bias the estimates, see e.g. Jones (1996), we consider it less severe than not controlling for this missing indicator since it is not unconditional of the dependent variable. Consequently, the significant effect of the missing indicator variable gives strong incidence, that persons with missing values on covariates strongly tend to be more reluctant to give income information.

The dependent variable in model $II$ distinguishes between respondents for whom all household income information is missing and those for whom full or at least rough information is available. The results differ considerably from model $I$. Several indicators become insignificant. However, households with more than two adults still remain to have a higher probability to lack all information. Whereas gender does not matter in model $II$, the occupational status and age remain important factors with civil servants being less likely to have missing values on all household income questions compared to employees, self-employed and persons not working. The satisfaction with the financial situation seems to become more important in model $II$, suggesting, that persons who reject to give any information on household income do so, because they are not satisfied with their financial situation. Also, we find a negative effect on item-nonresponse for higher educated respondents (CASMIN group 3). Respondents from greater cities (more than 500.000 inhabitants) tend to be more likely to give no information at all on household income questions. Living in East Germany is no factor in model $II$. Once more, the missing indicator variable has the expected positive effect on item-nonresponse.

Summarizing, clear evidence for factors influencing the likelihood for missing values on household income questions is present. To allow for meaningful analysis with the surveyed income data,

multiple imputation of missing values in income variables seems to be necessary.

# 3. Imputation method and filter handling

When setting up an imputation procedure, two problems need to be addressed in the context of large scale survey data. The first problem is the specification of valid approximations of the full conditional distributions reflecting the uncertainty related to the missing values. The second problem is the integration of all information available at different levels due to the multilevel filter structure corresponding to individual restrictions on the range of full conditional distributions.

## 3.1 Multiple imputation using nonparametric full conditional distributions

In order to reflect the uncertainty within the imputations for missing values, Burgette and Reiter (2010) suggest the usage of nonparametric full conditional distributions obtained via Classification and Regression Trees (CART), see Breiman et al. (1984). CART constitutes a nonparametric recursive partition algorithm. It aims to provide an optimal partition of the data via recursive binary splits. The decision rule determining the partition is illustrated as follows. Let $y$ denote a variable with $n$ observations. Define a binary partition for the variable $y$ via sets of elements $D_1$ and $D_2$ with $|D_1| = n_1$, $|D_2| = n_2$ and $n_1 + n_2 = n$. The partition is preferred against non partitioning, when a defined homogeneity measure $H(y \in D)$ fulfills the condition

$$H(y \in D_1 \cup D_2) > H(y \in D_1) + H(y \in D_2). \tag{1}$$

For metric variables, the variance can serve as an homogeneity index, i.e.

$$H(y \in D) = \sum_{y \in D} (y - \frac{1}{|D|} \sum_{y \in D} y)^2, \tag{2}$$

while for (unordered) categorical variables the entropy

$$H(y \in D) = \sum_{k \in K_y} f_y(k) \ln f_y(k) \tag{3}$$

can be used, where $K_y$ denotes the set of values for discrete variables $y$, and $f_y(k)$ denotes the relative frequency of category $k$. Hence, the objective of CART is to split up the observations into different groups, fulfilling the condition that respondents and thus observations assigned to one group show highest intra group homogeneity with respect to the relevant variable, whereas the inter group homogeneity is intended to be as small as possible. There exist manifold possibilities to define a partition. CART defines binary partitions via a set of $p$ conditioning variables $X = \{X_1, \ldots, X_p\}$ with $n$ observations per variable $X_i$, $i = 1, \ldots, p$. To ensure computational feasibility the CART algorithm does consider univariate splits only, i.e. only binary partitions defined upon a single variable are considered. The total number of considered univariate binary splits, say $S$, depends thereby on the number of metric and categorical variables within the considered set of conditioning variables.[4] The sequential partitioning algorithm proceeds via

[4] Note that for each metric variable there exist $n-1$ potential binary splits, while for each ordinal or nominal variable with $k$ categories, there exist $\sum_{i=1}^{\frac{k-1}{2}} \binom{k}{i}$ for $k$ odd or $\sum_{i=1}^{\frac{k}{2}-1} \binom{k}{i} + \frac{1}{2}\binom{k}{\frac{k}{2}}$ for $k$ even possible binary splits.

consideration of all $S$ partitions defining a binary partition and choose the optimal partition according to

$$\{D_1^*, D_2^*\} = \arg \max_{\substack{D_1^s, D_2^s \\ s=1,\ldots,S}} H(y \in D_1^s \cup D_2^s) - H(y \in D_1^s) - H(y \in D_2^s).$$

Of all possible univariate splits, the split with the maximum reduction in heterogeneity is chosen. The resulting binary partition of the data along the set of conditioning variables provides sets of admissible values defining the nonparametric characterization of the full conditional distribution and serve as donors 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 missing values is taken into account. With regard to the parameters of the CART algorithm, concerning stopping criteria and minimum requirements for the size of donor groups, we follow the suggestions of Burgette and Reiter (2010). Hence, no further split is considered when the resulting reference groups contain less than 50 or the gain in homogeneity is less than 0.01.

Based on the nonparametric approximation of the full conditional distribution delivered by CART, multiple imputation via chained equations (MICE) can directly be utilized to conduct imputation. The chained equations approach, see e.g. van Buuren and Groothuis Oudshoorn (2011)[5], specifies an individual imputation 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, thus following Little (1992) and Little and Trivellore (1997). So at first, the missing values in all variables are initialized and afterwards the algorithm iteratively runs through all specified (conditional) imputation models. This procedure is repeated several, say $M$ times, resulting in $M$ imputed data sets. As each iteration consists of one cycle through all variables considered, the algorithm provides a completely imputed dataset, see van Buuren (2007). Before starting the MICE algorithm, the data matrix is usually arranged to ensure that the number of missing values per variable is ascending, which is favorable in terms of convergence. 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 and Reiter (2010) propose to specify the full conditional distribution within the MICE algorithm via CART. Using the CART approach of Burgette and Reiter (2010) reduces the users effort to impute the data, because the conditional models do not have to be specified explicitly.

## 3.2 Handling of filters

The usage of individualized questionnaires navigated by filters is a special feature of large scale surveys. These filters tackle single questions as well as complete question blocks mostly separated in so called modules. Advantages of filters are that they keep the questionnaire at an individual level as short as possible for the surveyed individuals and reduce burden entailing from the survey. Disadvantages of filters relate to the fact that filtering causes lots of missing values when appearing as multilevel filters and by definition make the missing pattern individual specific. Thus, asking for characterizations of full conditional distributions differing between individuals with respect to available conditioning variables and the admissible range for values. These missings by filter have to be differentiated from true missing values caused by nonresponse due to refusal or unawareness.

[5] This approach is also known as fully conditional specification (FCS), see van Buuren (2007), or sequential regressions according to Raghunathan et al. (2001).

Each of the different types of filtering requires special handling. The first type includes item-specific missings, that occur when the question does not fit the respondents specific situation and causes the classification of this respondent into a residual category for this question. For example, all respondents are asked to evaluate their work climate, but some do not have any colleagues. Second, filters establishing a link between two questions within one module need to be regarded. The information from the filter variable is passed to a residual category of another question, and if possible filled up with zero in a further edition step. For example, when a person stated, that no special payment was received. Thus, for the following question about the amount of special payments zero was inserted. The third type relates to questions in different modules. An example is provided by the year of birth – asked in the NEPS within the sociodemographic module – as a filter for questions concerning the part-time work for older employees – part of the employment history module. Fourth, whole modules are filtered because they do not fit to the respondents biography. For example, if a respondent has never been employed, he is filtered over the module concerning his employment history. Fifth, there are missing values due to varying instruments for re-interviewed and first-time respondents or instrument modifications from wave to wave. In addition to programmed filters, which determine the interview sequence, logically derived filters are taken into account. Empirically implausible values, e.g. the reported individual net income exceeds the gross income, are explored and remedied to ensure logical consistency before implementing the extended CART method. Each of these different kinds of filters influence the variables and admissible values at hand for imputation on an individual level.

In order to accomplish proper imputation, two requirements have to be fulfilled to ensure appropriate treatment of the filters. First, the initialization step has to regard the hierarchy of the implemented filters. Second, the imputation has to preserve the sequence of variables implied by the filtering which in turn ensures the consistency of full conditional specifications. Initialization of missing values is performed via draws from the empirical distribution with replacement. In order to account for the filtering, one has to screen the whole list of variables used for the imputation. During the screening process logical constraints and linkups should be captured comprehensively. This screening then provides the admissible range of the empirical distribution fulfilling the restrictions implied by the filter structure. Usually, the admissible ranges can be deduced from filtering tables used for the programming of the questionnaire instrument. The restrictions on the admissible range concerning the empirical distributions are regarded for imputation as well as they are taken into account when setting up the corresponding full conditional distributions. These candidate donors then can be delimited to meet the filters or restrictions given by the data structure. This is done at an individual scale by forming intersects between the values of the donors and the admissible range. For this purpose a matrix is constructed that contains all admissible ranges at each missing position, which ensures a highly flexible response to the variety within imputation. That means, for each imputation cycle the updated data will be employed to extract the range of admissible values with respect to the inherent filter structure. From all of these observed values in the delimited donor group a case is randomly drawn and taken for imputation.

### 3.3 Imputation algorithm and combining rules

The presented imputation approach adapted from Burgette and Reiter (2010) initializes the missing values by draws from the empirical distribution without replacement but incorporating the restrictions and admissible range arising from filtering or bracketed questioning. Compared with MICE, the processing of the variables is arranged by sequences instead of an ascending order according to their number of missing values. The initialized and during the imputation iteration continuously updated data is used as training data to fit the tree models and for prediction of the terminal nodes. Ten prior imputation steps are considered for convergence of the training data. The following $M$ steps are stored and used for analyses. Summarizing the algorithm has

the following structure.

**Step 1:** Initialize the missing values for all variables by drawing from the unconditional distributions taking restrictions and hierarchy of variables arising from filtering and bracketed questioning into account.

**Step 2:** Given the initialized values, the CART algorithm is used sequentially in the hierarchical ordering of variables implied by filtering to obtain a nonparametric approximation of the full conditional distribution. Intersecting the corresponding set of donors with the restrictions arising from filtering or bracketed questioning allows for updating the missing values. The originally missing values are replaced by draws from the predictive distribution conditional on all other variables besides the one for which values are drawn.

**Step 3:** Repeat step 2 $M + L$ times, where $L = 10$ iterations have been found sufficient to mitigate the effect of initialization.

The main advantage of the concept of multiple imputation introduced by Rubin (1976) and discussed in detail by Rubin (1987, 2004) is the possibility to directly assess the uncertainty due to imputation. After multiple imputation the data can be analyzed by standard-complete methods. The difference is, that there are now $M$ data sets. So the estimates have to be calculated with those $M$ data sets, considering the uncertainty caused by the originally missing values. Therefore, the combining rules established by Rubin (1987) have to be used. Following Little and Rubin (2002), inference based on multiply imputed data sets for multidimensional quantities can be performed as follows. Let $\theta$ denote the vector of all model quantities of interest. Further, let $\hat{\theta}$ be an estimator with complete data properties such that

$$\hat{\theta} \overset{\text{asy}}{\sim} \mathcal{N}(\theta, \Sigma).$$

This property can usually be established via use of central limit theorems or results directly from the properties of maximum likelihood estimation. Based on each imputed data set, we obtain an estimator $\hat{\theta}^{(m)}$ with corresponding variance $\Sigma^{(m)}$, $m = 1, \ldots, M$ conditional on the imputed values, i.e.

$$\hat{\theta}^{(m)} = E[\theta|y_{\text{mis}^m}] \quad \text{and} \quad \Sigma^{(m)} = Var[\theta|y_{\text{mis}^m}] = E[(\theta - E[\theta|y_{\text{mis}^m}])(\theta - E[\theta|y_{\text{mis}^m}])'].$$

As the missing values are imputed as draws from the full conditional distributions, the law of iterated expectations can be used to mitigate the effect of conditioning on the imputed values. Hence,

$$\hat{\theta} = E_{f(y_{\text{mis}})}[\hat{\theta}^{(m)}]$$

and

$$\begin{aligned} Var[\hat{\theta}] &= E_{f(y_{\text{mis}})}[\Sigma^{(m)}] + Var_{f(y_{\text{mis}})}[\hat{\theta}^{(m)}] \\ &= E_{f(y_{\text{mis}})}[\Sigma^{(m)}] + \left(E_{f(y_{\text{mis}})}[(\hat{\theta}^{(m)} - E_{f(y_{\text{mis}})}[\hat{\theta}^{(m)}])(\hat{\theta}^{(m)} - E_{f(y_{\text{mis}})}[\hat{\theta}^{(m)}])']\right). \end{aligned}$$

Based on the weak law of large numbers, the involved expectations can be approximated for large $M$ as arithmetic means resulting in

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)} \quad \text{and} \quad Var[\theta] = \frac{1}{M}\sum_{m=1}^{M}\Sigma^{(m)} + \left(\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}^{(m)} - \hat{\theta})(\hat{\theta}^{(m)} - \hat{\theta})'\right).$$

Note that these formulas apply for large $M$, for small $M$ a finite sample correction is routinely applied, see Little and Rubin (2002).

Based on the outlined imputation algorithm, the next section presents the results concerning imputation of missing income values.

# 4. Imputation of missing income values in the NEPS

## 4.1 Data description

The basis for multiple imputation, which is discussed here in detail, is provided by the NEPS Scientific Usefile of Starting Cohort 6.[6] The total number of respondents is 11,649. Among these 5,154 are first-time respondents in NEPS and 6,495 are re-interviewed.[7] The main scientific usefile release consists of 23 modules with either surveyed or generated variables covering a large diversity of research topics. The total number of variables surveyed or generated is equal to 1,192.

Because our focus lies on income imputation, all variables considered within the imputation process have been selected based on an assumed direct or indirect relationship towards income variables. This results in 151 variables chosen out of the main data file and a generated data file containing panel and cross sectional information and basic socio-demographic information about all respondents. In addition, variables from the employment history module are selected for those respondents who have ever had any employment, thus contributing another 62 income related variables.With regard to the overall 11,649 respondents, all 213 variables are available for 11,516 respondents, whereas available information for 133 persons refers to 151 variables only. A complete list of all 213 selected variables and corresponding descriptive statistics is given in the Appendix A. To match these different sets of information related to the different numbers of available variables, the imputation is done for both, the 11,649 respondents based on 151 variables and the 11,516 respondents based on 213 variables. The final imputed data set then consists for 11,516 respondents on imputations based on 213 variables and for 133 respondents on imputations based on 151 variables.

Further, large parts of the original data are provided in form of episodes, where episode data is originally stored in long format, one row representing one respondent at a given time or episode respectively. One of the prerequisites for the application of the CART algorithm is the availability of individual information in one single row. Data preparation proceeds in three steps. First, information from different waves is harmonized, i.e. information is taken from the preceding wave and, if updated, from the current wave. Second, information from episodes is selected or aggregated. Selection criterion for episodes within the employment history module is to choose the latest episode whether it is continuing or not. The handling of aggregation is threefold, i.e. information is either dichotomized, summarized or indices are built. Finally, all the modules are merged into one single dataset. For the employment history module a homonymous dummy is created, indicating if the module is answered by a respondent or not which would indicate a structural missing. Analogously a panel dummy is applied to wave information.

In the NEPS, gross as well as net income data is available at the individual level. At the household level net income is surveyed. Because of the high sensitivity of financial issues and to obtain at least partial information, respondents who refuse or are not able to provide an exact estimate for their income are asked to assign themselves to certain income brackets, see Figure

[6] SC6 version D-1.0.0, see Blossfeld et al. (2011) for a general discussion of the study design and Leopold et al. (2011) for a general documentation of the scientific usefile.

[7] The NEPS sample integrates the sample of the study "Working and Learning in a Changing World" (ALWA) conducted by the Institute for Employment Research (IAB, Nuremberg).

1. Income data is either given as an exact estimate or in two-staged income brackets depending on the willingness of the respondent to provide concise information. In Table 2, a description of the exact income data estimates is given. Table 3 gives an overview regarding missing values due to refusal or unawareness of the exact estimate and remaining missing values when the first step of the income brackets is considered additionally.

### 4.2 Empirical results

To highlight the results of the imputation algorithm, we analyze for some income variables the characterizations of the full conditional distributions accessible via trees delivered by the nonparametric sequential regressions. Note that 100 iterations are conducted after burn-in and each of these iterations delivers a nonparametric approximation that characterizes the full conditional distribution, which can be graphically shown by a tree. Hence, slight variation resulting from multiple imputation is expected to capture the uncertainty due to imputation. However, only modest variation between the trees over the 100 replications is observed. Thus the following descriptions are made for the last resulting tree.

Figure 2 displays the set of variables selected out of all 151 variables for characterization of the full conditional distribution used for imputation of missing values in the household net income with regard to the main panel file and two generated files. The most important variable to characterize the household income are the household income brackets with detailed information for incomes of more than 3,000 € (split 2c, see Figure 1). The split point of the tree is set between '1' and '2'. That means, respondents in the 'does not apply'-category ('-99') and with an income of 3,000 € up to 4,000 € ('1') are regarded as more homogenous compared to respondents with an income of more than 4,000 € ('2', '3'). For those respondents with less than 4,000 € household net income, a further split point is characterized by the first stage income brackets. Respondents with less than 3,000 € (split 1 '1', '2') are divided into donors with detailed income brackets according to split 2b, whereas respondents grouped in split 1 '3', having an income of more than 3,000 €, denote a final node. Filtered respondents ('-99') and those with 1,500 € up to 2,000 € ('1') are separated from those having 2,000 € up to 3,000 € ('2','3'). Continuing on the right side of the tree, respondents are divided by the both remaining categories of split 2c. Respondents having more than 5,000 € are then partitioned via the occupational status. Workers ('1'), employees ('2'), civil servants ('3'), soldiers ('4'), assisting family members ('6'), freelance workers ('7') and respondents in an employment as preparation for a profession ('8') constitute an end node, and self-employed respondents ('5','-99')

-99

are finally partitioned according to their age. The split point for the age is set to 62.21 years. Respondents younger than the split point have a lower mean in household net income compared to those older than 62.21 years. Note that imputations based on 151 variables were only used for those 133 individuals without employment history.

Figure 3 shows the last tree used for imputation of the household net income for all 11,516 persons with additional information from the employment history module. Here, the bracketed information and the individual net income are the most important explanatory variables. The tree is quite similar to the tree described previously. Just the right side differs slightly. Instead of partitioning respondents with an household income of more than 5,000 € according to their occupational status and age, they are divided by their individual net income, split point at 8,900 €.

Finally, Figure 4 describes the full conditional for the individual net income. The only variables to divide the respondents in binary splits are the first stage income brackets for the individual net income (split 1), which is partitioned several times, the second stage income brackets (split 2a, 2b,

2c) and the current individual gross income. Adding the variables from the employment history module does not change so much in the construction of the trees. Sufficient information is given by the income variables themselves and few variables are chosen from the basic modules. This is supported by a by-product of imputation evaluation. The accuracy of prediction of the household income – regardless of bracketed information – was compared for the imputation considering the basic modules only versus imputation with additional information from the employment history module. Test criterion was if the imputed value lies somewhere within the additionally provided income brackets. The improvement was only modest (53.9 % vs. 54.0 %).

To assess the quality of imputations, we look at distributional similarity before and after imputation, which is according to van Buuren (2012) often more relevant than model fit for the validity of inference. The following tests and plots are performed to check for distributional discrepancy between observed and imputed data across all 100 iteration steps. Categorical and ordinal variables are tested using the Chi-squared goodness of fit test. None of these variables differed significantly after applying the imputation. For continuous metric variables the Kolmogorov-Smirnov goodness of fit test is performed. The individual gross income and the sum of special payments differ significantly between observed and imputed data (in all imputation steps, significance level is set to $\alpha = 0.05$). The Q-Q plots for both variables are given in Figure 5 and reveal discrepancies in the higher quantiles indicating the imputation of higher values. The column charts and Q-Q plots for all variables are given in Appendix B.

For a deeper look, a binary and an ordinal variable are selected for demonstration. Figure 6 shows the observed and imputed data for the expectations of friends to achieve success on a professional level and the distributions for having a friend who attended courses interesting for the respondent. The confidence intervals are too small to accentuate in the plots. As could be seen, the empirical distribution is very well preserved. The kernel densities are given for the household income and the individual net income as examples and show only minor differences between the distributions before and after imputation, see Figure 7. To search out for differences in more detail, the income variables are compared as classified data. Figure 8 shows slightly higher percentages after imputation in the middle class for household income, and slightly higher percentages in the highest category for individual net income. Nonetheless, these differences are not significant. The modest effect is, among other things, due to the small number of missing values in our data. However, the availability of imputed data sets will nevertheless enhance accessibility of the data.

Building up on the summary of Drechsler (2011) regarding different packages for multiple imputation several advantages of the extended CART can be pointed up. With the extended CART it is easy to handle categorical as well as continuous data. As a result, it can deal with multicollinearity and skewed distributions, is highly flexible to fit interactions and nonlinear relations, and is robust against outliers as well. Because these features are inherent to the modeling approach, and because of the automated model fitting, it is a manageable tool for multiple imputation purposes as already stated by van Buuren (2012). With the additionally implemented features for initialization, sequence control as well as delimiting donor groups in accordance with the filter structure we could make it capable with bracketed information, skip patterns and constraints. As criticized by Si and Reiter (2013) multiple imputation by sequential regression often fail to capture such complex dependencies.

Note that all computations including data preparation, imputation and analyses have beed performed using `R version 3.0.1`, required packages are `Tinn-R`, `foreign`, `tree`, `lattice` and `plyr`. Our code is an adaption of the basic implementation of `treeMI` from Burgette and Reiter (2010), which is available at `http://www.burgette.org/software.html`. Written functions and `R` code are available from the second author upon request. The duration of one single imputation for the three basic modules is about 15 minutes and one single imputation with additional information of the employment history module lasts a little bit more than half an hour, with

slight variation within each imputation iteration (PC-System: Windows 7 Professional, 64bit Clone-Version 3.3.3, Intel(R) Core(TM) i7CPU 860 @2.80GHz, 8GB RAM).

# 5. Concluding remarks

Questions regarding someone's financial situation are highly sensitive and often lead to considerable rates of item-nonresponse. With regard to the used data set, we demonstrated that there are different factors influencing the probability of refusal or unawareness. These include individual characteristics like gender or the reported satisfaction with the financial situation as well as the number of cohabitating adults in the household.

In order to impute missing income information within the NEPS Scientific Usefile of Starting Cohort 6 (adults) version D-1.0.0, we use a nonparametric tree-based sequential regression approach combining the partition algorithm CART (Classification and Regression Trees) and the imputation technique MICE (Multivariate Imputation by Chained Equations). This particular choice was motivated by the ability of this nonlinear approach to deal with rich nonlinear relationships among the surveyed variables. A further advantage is the incorporated flexibility to deal with multilevel filtering rendering the full conditional distributions necessary for imputation individual specific. Operationalizing the full conditional distributions via set of identified donor observations, restrictions arising from multilevel filtering on an individual level can be incorporated via the intersections.

On completion of necessary data preparation and imputation steps we produced imputed datasets containing 213 variables and 11,649 observations. Missing income information was replaced by values, which were drawn from the empirical distribution of the partitions identified by CART. Regarding income data, for the majority of respondents these partitions were established by considering the bracketed income information, occupational status, age as well as individual net and gross income. Furthermore, all missing information in all enlisted variables were imputed simultaneously. The inspection of all variables revealed that distributional similarity is very well preserved. From these, an imputed data set is provided for the scientific community as an extra scientific usefile in long format containing several rows per respondent each representing a single imputation step.

Given the flexibility of the suggested approach, further research may address the topic of heaping thus focusing on preservation of distributional similarity before and after imputation in the presence of rounded values.

**Acknowledgements**

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 '- Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:1.0.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

## References

Allison, P. (2002). *Missing Data. Thousand Oaks.* Sage.

Azur, M. J., Stuart, E. A., Frangakis, C., and 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.

Blossfeld, H.-P., Roßbach, H.-G., and Maurice, J. v., editors (2011). *Education as a Lifelong Process – The German National Educational Panel Study (NEPS).* Number Special Issue 14. VS Verlag.

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

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076.

Burgette, L. F. and Reiter, J. P. (2012). Nonparametric bayesian multiple imputation for missing data due to mid-study switching of measurement methods. *Journal of the American Statistical Association*, 107(498):439–449.

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

Drechsler, J. and Reiter, J. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243.

Duffy, D. (2011). Income and wage imputation methodology. *Technical Series Paper. Survey Research Center - Institute for Social Research.*

Frick, J. and Grabka, M. (2005). Item nonresponse on income questions in panel surveys: Incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv*, 89:49–61.

Frick, J. R. and Grabka, M. M. (2007). Item non-response and imputation of annual labor income in panel surveys from a cross-national perspective.

Hapfelmeier, A., Hothorn, T., and Ulm, K. (2012). Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. *Computational Statistics & Data Analysis*, 56(6):1552–1565.

Hayes, C. and Watson, N. (2009). Hilda imputation methods.

Jenkins, S. P. (2010). The british household panel survey and its income data. *IZA Discussion Paper*, 5242.

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.

Leopold, T., Raab, M., and Skopek, J. (2011). Data Manual: Starting Cohort 6 - Adult Education and Lifelong Learning.

Lipps, O. (2010). Income imputation in the swiss household panel 1999-2007. *FORS Working Paper Series.*

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data.* Wiley, Hoboken.

Little, R. and Trivellore, R. (1997). Should imputation of missing data condition on all observed variables? *Proceedings of the Survey Research Methods Sections; American Statistical Association*, pages 617–622.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

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

Little, R. J. A. and Su, H. L. (1989). *Item nonresponse in panel surveys*. John Wiley.

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

Riphahn, R. T. and Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30(2):521–538.

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

Rubin, D. (2004). The desing of a general an flexible system for handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 48:298–302.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Si, Y. and Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38:499–521.

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*. CHAPMAN & HALL/CRC.

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

Watson, N. and Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics*, 27:693–715.

## Figures

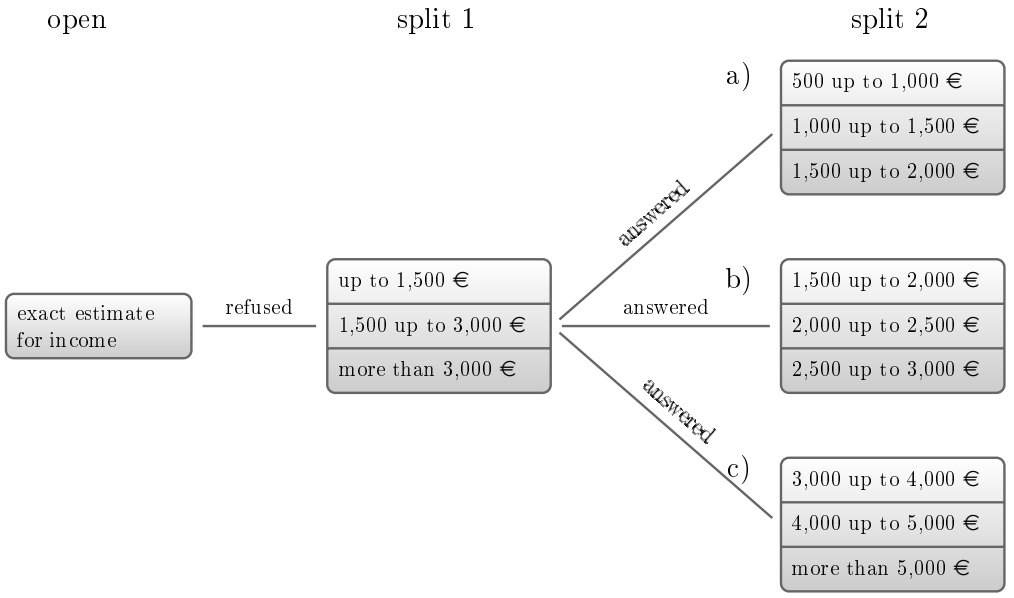


Figure 1: Income questions in the NEPS SUF SC6 – exact estimate and two-stage income brackets.

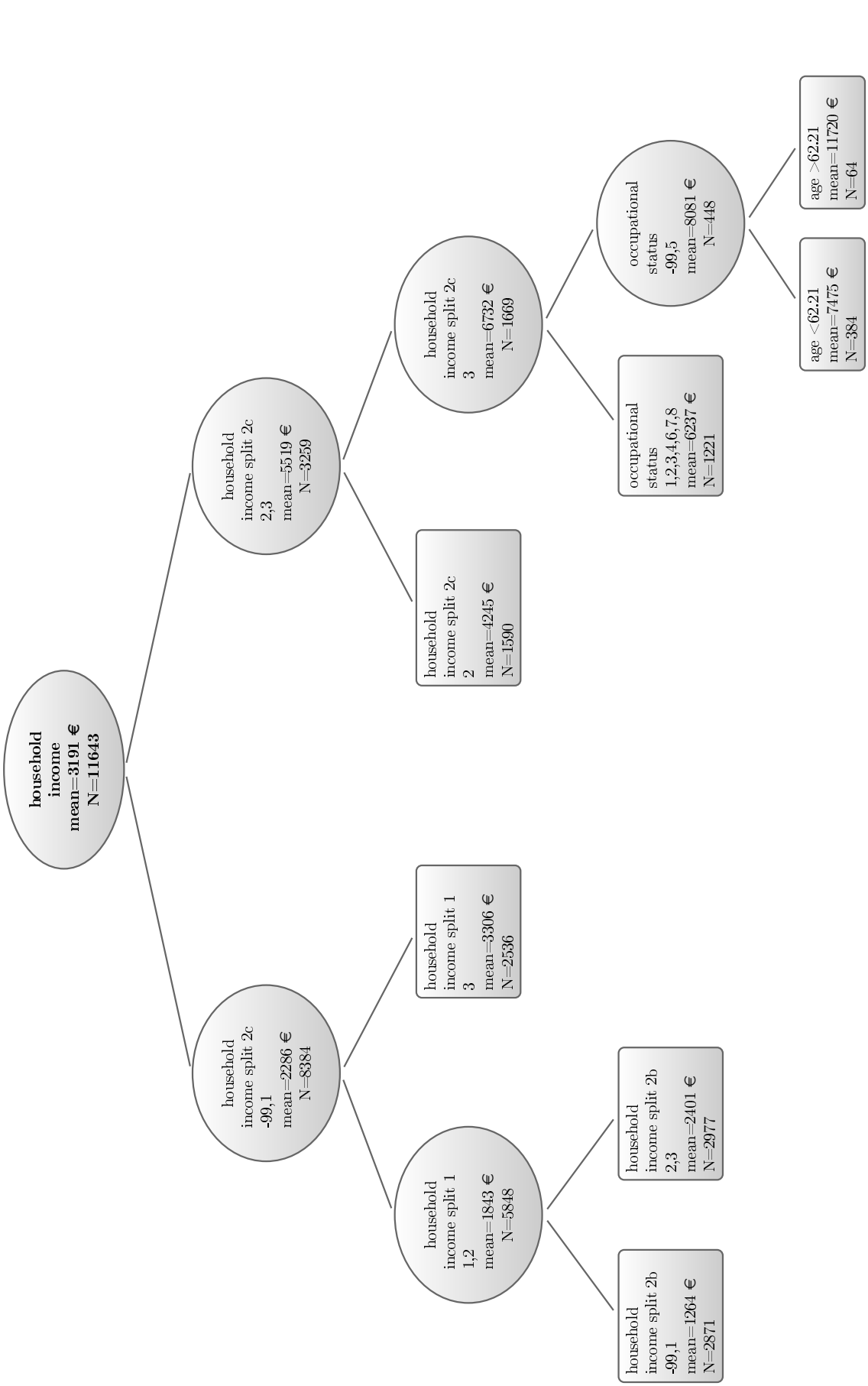


Figure 2: Household income imputed via the main panel file and two generated files. Notes: mean is always with reference to household income, $N$ is the number of respondents in each node.

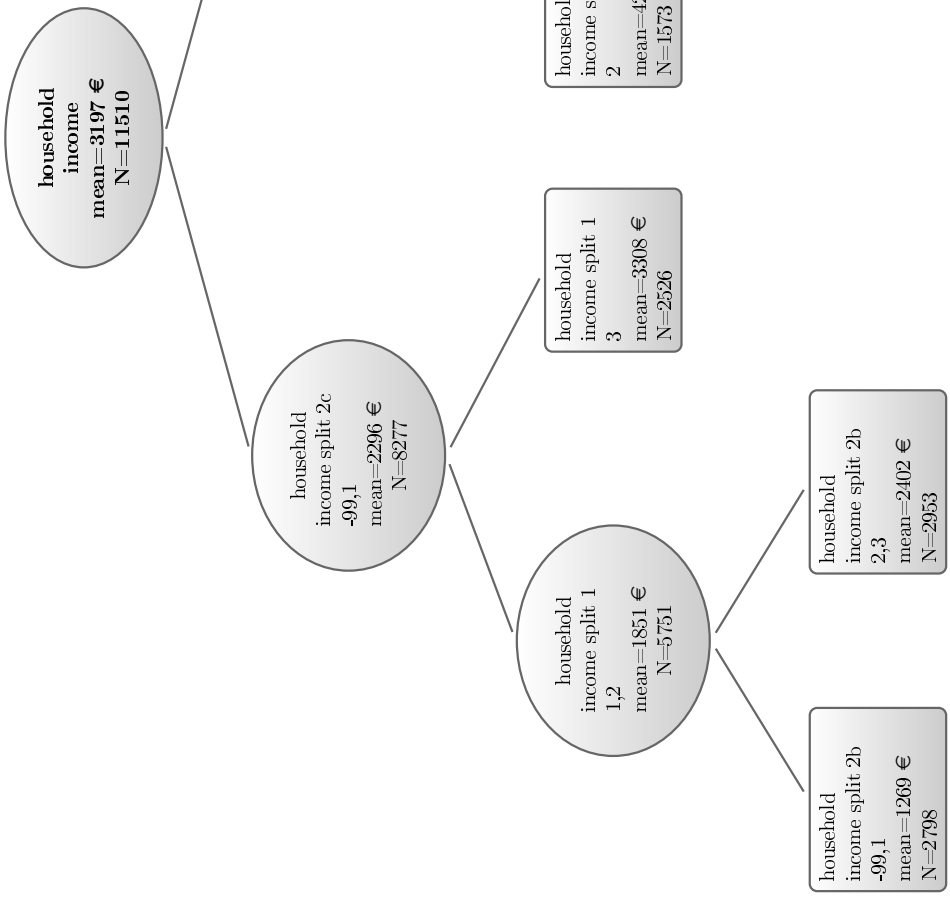


Figure 3: Household income imputed via the main panel file and two generated files. Notes: mean is always with reference to household income, $N$ is the number of respondents in each node.

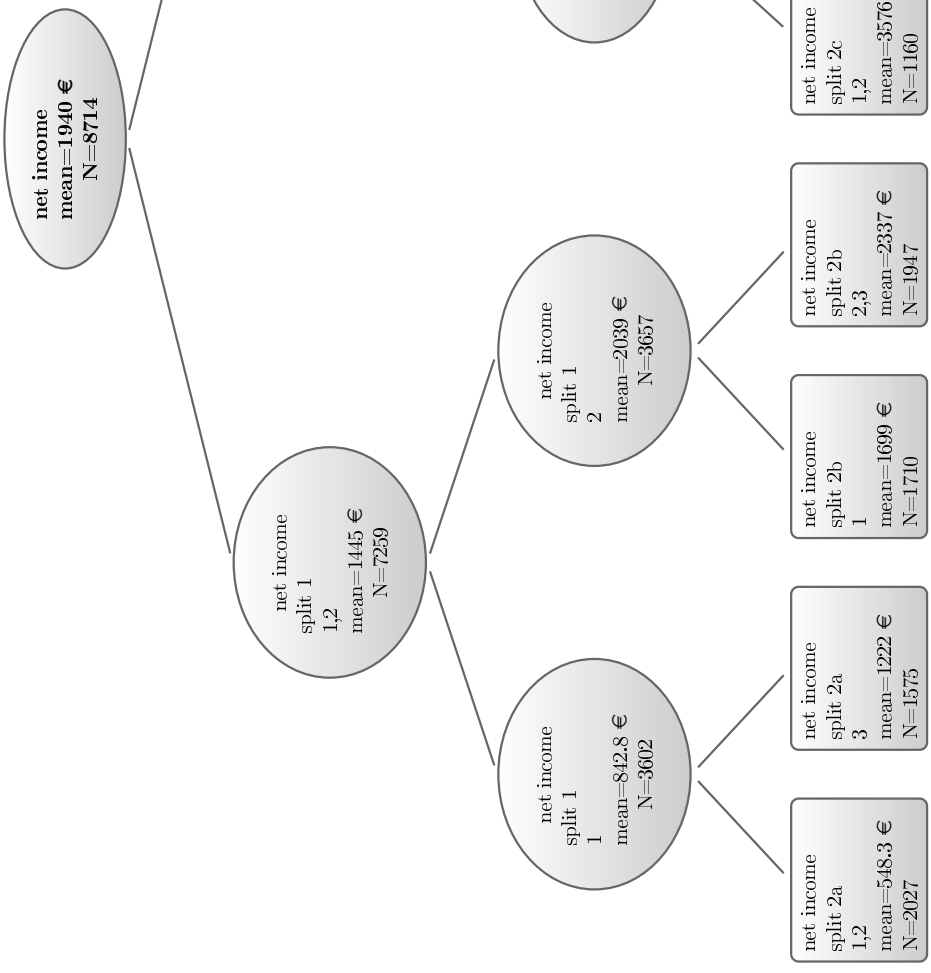


Figure 4: Individual net income imputed via the main panel file, two generated files, and the module for employment history. Notes: mean is always with reference to individual net income, $N$ is the number of respondents in each node.

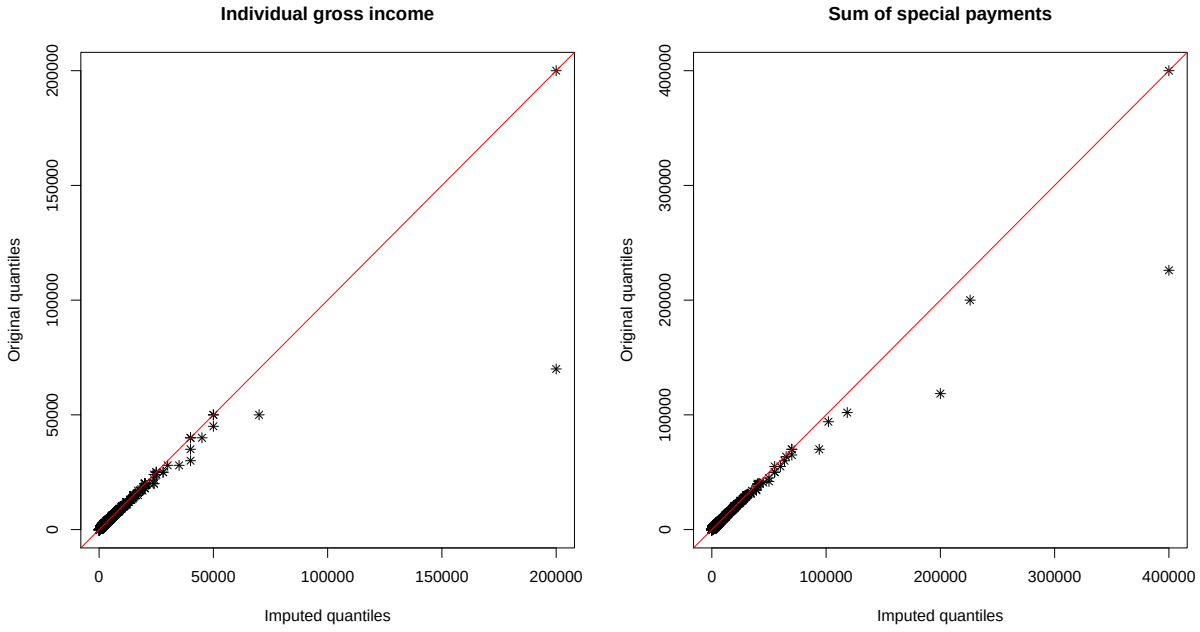


Figure 5: Q-Q plots for the individual gross income and sum of special payments, variables with significant differences between observed and imputed data according to Kolmogorov-Smirnov goodness of fit test (level of significance: $\alpha = 0.05$).

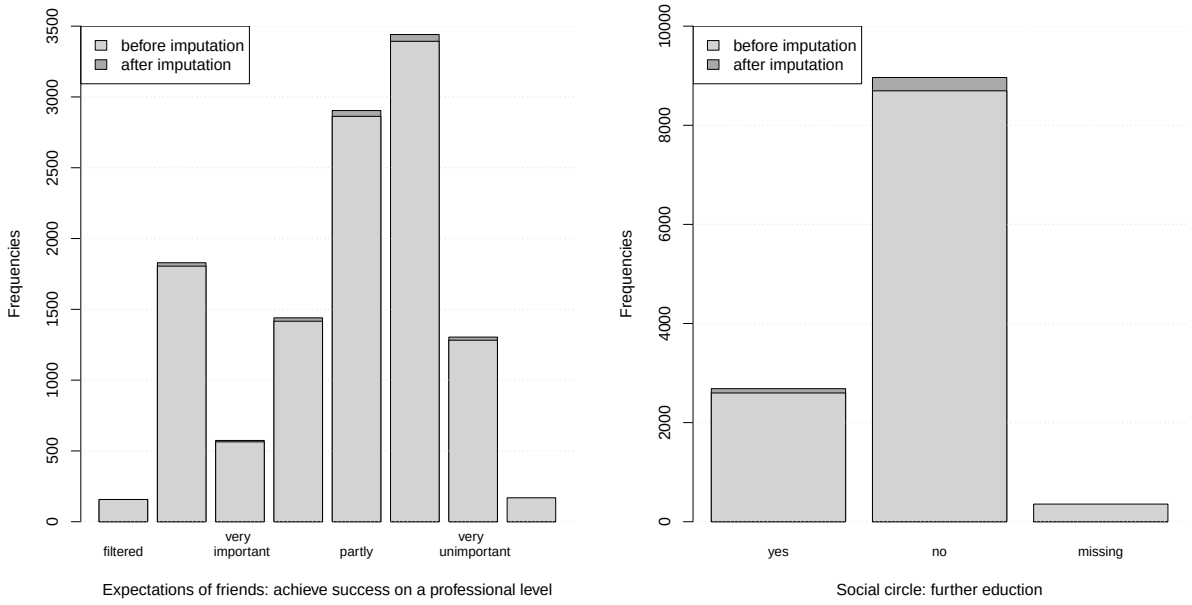


Figure 6: Column charts for one ordinal variable on the left side and one binary variable on the right side. Observed values are indicated with light gray and imputed values with dark gray. Confidence intervals are too small to be plotted.

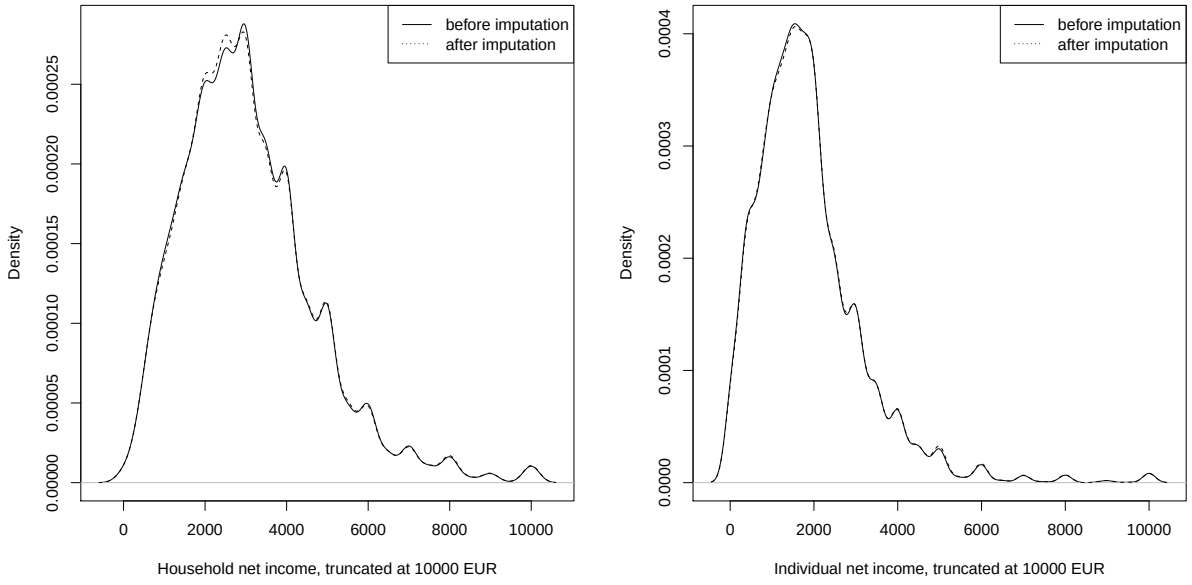


Figure 7: Kernel densities for household income and individual net income. Solid lines indicate observed data and dashed lines imputed data (bandwidths are: 200 for household income and 150 for individual net income).

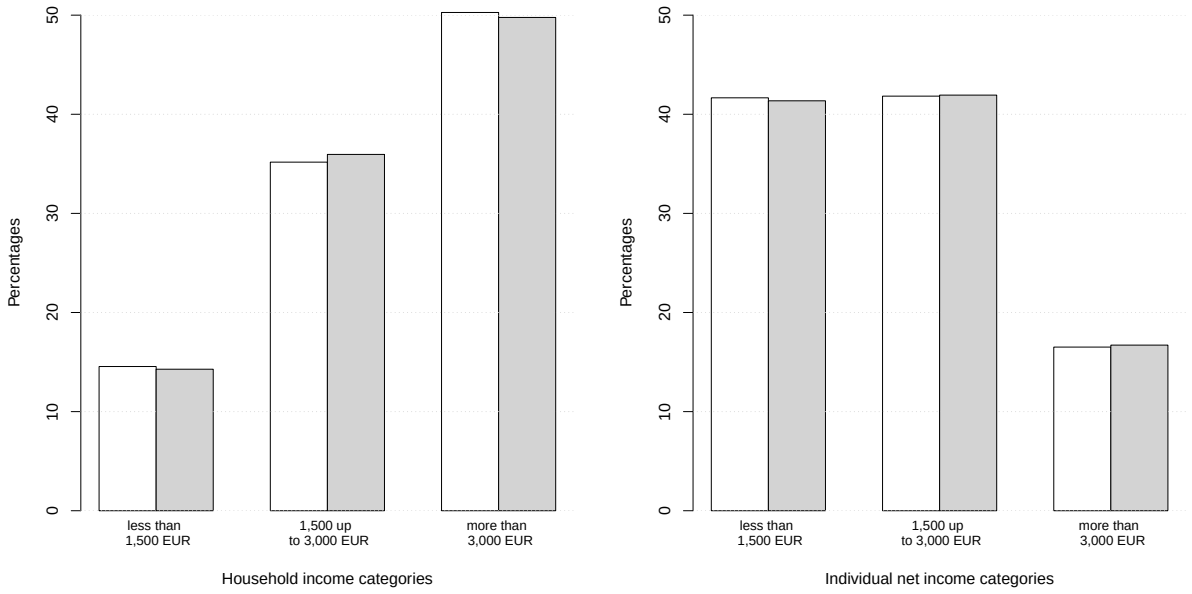


Figure 8: Classified income information for household income and individual net income. Respondents for which these questions do not apply where excluded. Imputed data is indicated with light gray and observed data white.

## Tables

Table 1: Estimating the probability for item-nonresponse on household income questions - Results from probit models

| | Model *I* Any income-information missing | Model *II* All income-information missing |
|---|---|---|
| 31 to 40 years | −0.0251 | 0.1264 |
| 41 to 50 years | 0.0231 | 0.2423** |
| 51 to 60 years | 0.1127* | 0.2043* |
| 61 years or older | 0.1536* | 0.4314*** |
| Gender: female | 0.1948*** | 0.0423 |
| Two adults in the household | 0.2434*** | 0.0900 |
| More than two adults in the household | 0.8867*** | 0.5113*** |
| One child in the household | −0.0439 | −0.0936 |
| More than one child in the household | 0.0100 | −0.0581 |
| Occupational status: worker | 0.2143* | 0.1631 |
| Occupational status: employed | 0.2802*** | 0.2568* |
| Occupational status: self-employed | 0.3898*** | 0.2726* |
| Other Occupational Status | −0.1147 | −0.1826 |
| Not working | 0.4619*** | 0.2608* |
| Satisfaction with fin. situation | 0.0273** | 0.0409*** |
| Born in Germany | 0.1119* | −0.1100 |
| CASMIN: group 2 | −0.0122 | −0.0815 |
| CASMIN: group 3 | −0.0353 | −0.1716* |
| Unemployed | −0.1528* | −0.0259 |
| Living in East Germany | −0.1072** | 0.0246 |
| Living area: 20,000 up to 100,000 inhabitants | 0.0282 | 0.1958* |
| Living area: 100,000 up to 500,000 inhabitants | 0.0405 | 0.1454 |
| Living area: more than 500,000 inhabitants | 0.0912 | 0.2265** |
| Number of missings on covariates | 0.5165*** | 0.4277*** |
| Constant | −2.1949*** | −2.6826*** |
| Observations | 11649 | 11649 |
| Log-Likelihhod | −4373.0973 | −1824.8164 |
| 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 in West-Germany; Living area: up to 20,000 inhabitants; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 2: Descriptives of the NEPS income data

| | 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 3: Frequencies of nonresponse in 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% (301) |

† 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.