Multiple imputation in a complex household survey - the German Panel on Household Finances (PHF): challenges and solutions

Martin, Eisele and Zhu, Junyi

Deutsche Bundesbank, Deutsche Bundesbank

28 November 2013

Multiple imputation in a complex household survey - the German Panel on Household Finances (PHF): challenges and solutions[1]

(VERSION 2013.11.28)

Martin Eisele
Deutsche Bundesbank

Junyi Zhu
Deutsche Bundesbank

Abstract

In this paper, we present a case study of the imputation in a complex household survey - the first wave of the German Panel on Household Finances (PHF). A household wealth survey has to be built on a questionnaire with rather complex logical structure mainly because the probes of many wealth items have to be proceeded on both intensive and extensive margins. Hence the number of potential predictors for each imputation model grows and more non-compliance can confront standard modelling due to, e.g., irregular missing patterns, interdependent logical constraints, data anomalies etc. Our model selection procedure borrows the techniques for the out-of-sample prediction to handle the overfitting often associated with the introduction of a large number of predictors. We also take the measures to produce ex ante evaluation for modelling which can be more efficient than the common diagnosis done after imputation in practice. Solutions for the difficulties in the real data and questionnaire structures are also presented. On the other hand, we incorporate the rich flagging information in developing various measures of item-nonresponse to access this complication from logical structure. We find that information loss due to the contagion of item-nonresponse between variables is not serious in our imputed data.

Keywords: Multiple imputation, Model selection, Panel on household finance, item-nonresponse evaluation

JEL-Classification: C15, C52, C42

[1] We are greatly indebted to Arthur B. Kennickell for lending us the FRITZ package he developed for years in the imputation of Survey of Consumer Finance (SCF). Beyond the practical application, this is also a full-fledged model to teach us many specific aspects tailored to a complex household finance survey. In addition, the experience of another FRITZ user Cristina Barceló has been rather constructive for us to build the starting infrastructure, and we learned from her hands-on insights on this package and other imputation techniques. Claudia Biancotti provided us many detailed technical instructions on using this package and valuable auxiliary procedures. Dimitris Christelis shared with us his rich knowledge in multiple imputation for wealth surveys. We often received the advice by discussing with the imputation colleagues from other NCBs and ECB - Nicolás Albacete, Alessandro Porpiglia, Michael Ziegelmeyer, Juha Honkkila, Laurent Van Belle, Antonis Loumiotis to name a few. Last but not least, our development of the imputation infrastructure cannot survive without the frequent support and encouragement from our Bundesbank colleagues – Ulf von Kalckreuth, Tobias Schmidt, Julia Le Blanc and Christoph Weisser. All of them also deserve a special thank from us.

1 Introduction

Missing data poses a tremendous challenge for both the data constructors and users. In one end, the whole data production process for any survey can be deemed as an imputation effort to fill the information into a completely missing data: the interviewer "imputes" with the values without the uncertainty induced by item-nonresponse, the editor "imputes" the implausible values by either replacing it with certain valid values or with missing values with uncertainty. And the imputer imputes bearing the mind that any value imputed contains this uncertainty. The imputation model is a scientific endeavor to translate this uncertainty to the data users.

The multiple imputation (MI) in PHF replaced the missing cells in the data by *m* simulated values. Similar to many other imputation practices, this *m* is five. The simulation draws from a posterior distribution given the observed data. This Bayesian perspective is introduced by Rubin (1978) and well explored in Rubin (1987) and Rubin (2004). Suppose the complete-data are $Y = (Y_{obs}, Y_{mis})$, which contains the observed subsample Y_{obs} and missing subsample Y_{mis}. The estimand is Q. The fundamental motivation is described by

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis},$$

where this integral is calculated via simulation. Rubin (1987) shows the final estimate of Q is an average of multiple complete-data posterior means of Q and the final variance is the sum of the average of within-implicate variance and the between-implicate variance.

In this paper, we present the main structure and important features of the core module FRITZ we used in section 2. Section 3 illustrates data preparations, data format and flag variable construction required for the imputation. We outline the general specification process for imputation models and many data peculiarities in PHF in sections 4 and 5. Section 6 presents an assessment of the item-nonresponse in the first wave of PHF. Finally, the sections 7 and 8 outline the convergence and the evaluation of the imputation model respectively.

2 Imputation package: FRITZ

We only provide an outline of our core imputation routine, FRITZ, its main modules we used, as well as some of its important features and the strengths in this section. For more detailed discussions on FRITZ, MI theories and Gibbs sampling, Kennickell (1991), Kennickell (1998) and Barceló (2006) could be the references. As far as the reader's interest in knowing the comparison of various imputation packages is concerned, the section 3.6 of Drechsler

(2011) should serve as a comprehensive guide for many practical aspects in economic surveys.[2]

2.1 Outline of FRITZ

In FRITZ, the central simulation mentioned in section 1 is achieved by an approximation to an actual Gibbs sampling. Raghunathan et al. (2001) illustrate this process: "*the new imputed values for a variable are conditional on the previously imputed values of other variables, and the newly imputed values of variables that preceded the currently imputed variable"*. These conditional distributions are usually modeled parametrically, e.g., as a regression. This approach avoids the complication of drawing directly the parameters in the formal Gibbs sampling setting which can be practically difficult when the editing constraints, bounds and miscellaneous variable types are present. Raghunathan et al. (2001) also assert that this approach has been proved to be empirically comparable to those based on an explicit Bayesian model in many real applications.

Like many other Gibbs sampling procedure, the program is invoked iteratively. It initially computes the appropriate sum-of-squares-and-cross-products (SSCP) matrix or conditional frequency table by restricting the data using the logical condition leading to the applicable response of the imputed variable. This matrix or table is based on the variable to be imputed and all the covariates specified in each imputation model. Next, it loops over each case where there is the imputation flag value and the logical condition leading to the applicable response is met.[3] Within each loop, the imputation is performed by combining the estimation information of the observation-specific covariates drawn from the SSCP matrix or conditional frequency table and the randomization process following user-specified constraints.

2.2 Modules

We discuss the three main modules for the corresponding data type: continuous, binary and categorical variables. Only main mechanisms used in our imputation are introduced here.

2.2.1 Continuous

The imputation value is an aggregation of the predicted value of a regression of the missing value on the set of explanatory variables available for a given observation and a randomized residual. The SSCP is constructed using all the nonmissing (observed or imputed in the last iteration) target variable and covariates specified (which can be a larger set than that from the estimation model tailored for each imputed case when some of the covariates are missing).[4]

---

[2] Yucel (2011) presents a much broader overview together with many other papers in this special volume.
[3] See the detailed exposition in section 4.3.
[4] Since it uses pairwise deletion, this matrix might not be invertible, esp. during the first iteration when the missing pattern can cause the number of pairwise nonmissing cases varies across each cell in the SSCP. On the other hand, sometimes the observed sample is too small which can cause the SSCP to have less than full rank. Generally, FRITZ uses SWEEP operator in SAS to produce generalized inverse. Alternatively, FRITZ can optionally use single value decomposition to handle this case.

The residual is mostly drawn from a normal distribution with the variance estimated from regression standard error (sigma) of each observation-specific model.

FRTIZ takes a conservative stepwise approach for the randomization: it first draws within +/- 1.29 sigma (80%) range; if it cannot hit within the user-specified bounds after 100 draws, it will draw within a +/-1.96 (95%) sigma range for the other 100 draws; if it still fails, the value is forced to the nearest bound. This can avoid producing the extreme values and/or values from "seeming" truncated/restricted distribution.[5]

2.2.2 Binary

Binary variables are imputed in an analogue way using randomized linear probability models. FRITZ trims the "outlier" case: when the predictive probability is very close to zero or one, the routine will impose that value and ignore the randomization. Thus, the "rare" events can be excluded.

2.2.3 Categorical

The algorithm randomly draws the imputed values from the cell of the conditional frequency table with the margins matching the covariates conditioned. The randomization is based on the frequency distribution of the cell. There are two classifying variables we can specify. But we can always incorporate more covariates by forming the multiplicity variable.

2.3 Options and advantages

Some useful options are used in our imputation to tackle with some quite common imputation issues. Besides, we also summarize a couple of general strengths of FRITZ.

One notorious issue in imputing categorical variable is the perfect prediction (White, et al., 2010): in the FRITZ's setting, when the cell size is too small or even zero, the point estimate $P(Q|Y_{obs})$ can be excessively biased and the between-imputation standard error can be erroneously exaggerated.

FRITZ provides two options to maintain a minimum cell size threshold. One is to only condition on the first covariate and examine the required minimum cell size; if it is still too low, the routine will then condition only on the second covariates; if it fails again, an unconditional frequency distribution is used. Alternatively, the program collapses the second covariate symmetrically around the value for the case to be imputed until the cell size threshold is hit. Additionally, like many other treatments, FRITZ always displays adequate warning messages for each steps involved.

Two kinds of nearest-neighbor matching methods can be invoked in FRITZ: matching using the value or the sample residual of the nearest neighbor. The details will be explored in the section 5.1. There are also options to determine whether the user-specified bounds are respected when they are incompatible with the bounds of the observed sample.

[5] The contribution of this approach will be elaborated in section 6.4.

Weight can be incorporated in calculating the SSCP or conditional frequency table.

FRITZ calculates the SSCP matrix in a unique way: it uses the pairwise deletion such that the correlationship between any two variables is measured with the maximum information content observed. There are two advantages:

1) It simply makes use of as much information as possible. The treatment in other imputation packages is the simple listwise deletion when missingness is presented. Obviously, the information loss can be severe.
2) Only one calculation is required for each imputation model/variable which is computationally more efficient. The estimation of each case to be imputed only has to draw the cells corresponding to the covariates observed out of the shared SSCP in order to build its own SSCPs.

Compared to many other packages, in general, it avoids those nonlinear estimations which require iterative optimization. There are usually some regularity conditions to insure the convergence of them. Unfortunately, in real survey data with complicated missing pattern, many of them cannot be fulfilled. For example, observed distribution is too skewed and/or multimodal, the size of the observed eligible sample is too small, the covariates might have (near) collinearities… These can cause the troubles in practice: the routine cannot converge or collapse. To diagnose each can often become very difficult during imputation, particularly because all the imputation model/variables are interdependent and missing pattern varies across the variables and cases. Plus, many other packages do not supply adequate warnings even when these situations occur. FRITZ instead makes great efforts to enhance the information reported in the log file. In all, FRITZ is superior in this aspect for the practical imputation projects which are often constrained by time and other resources.

3 Imputation infrastructure

We describe the problems before kicking off the specification of imputation model and actual imputation. They are addressed following the sequence in which they are solved.

3.1 Preliminary data preparations for imputation

The foremost thing to prepare is the list of variables to be imputed. We agreed to generally impute all variables of the questionnaire, with some few exceptions. These exceptions are:

- Variables with verbatim answers (e.g. job description PNE2010)
- Variables that specify the time period of another question (e.g. DHB2010)
- Variables belonging to questions that only had been asked as “back-up” during the interview in the case of item non response of other questions (e.g. HD0851, DHB2600)
- Person IDs (e.g. HD0601A)

- Other Variables where imputation had been considered as not meaningful (e.g. car make DHB82001A)

It is important that every imputed variable contains values of the same dimension. Several data preparation steps are necessary to guarantee this:

a) All net figures are converted into gross. This conversion takes into account the complex German taxation rules, using the information given by the respondent's answers on employment status, marital status etc.

b) All time-related information (in many questions the respondent had the choice to give monthly, quarterly or yearly values) has to be converted into yearly figures.

c) All foreign-currency- or legacy-currency-amounts in Euro value questions have to be calculated into Euro amounts.

To do the net-gross-conversion described above, we need all possible information about the variables which are used in the conversion algorithm, e.g. DPA0100 (marital status) and DHH0900 (church tax). Due to the fact that these variables also have item non response, they are imputed ex ante using hot deck imputation. This preliminary imputation is solely used for the net-gross-conversion. All missing values are imputed again later in the main imputation.

3.2 Data format combining household data and personal data

How do we run the imputation jointly involved with variables of household level and personal level? The complications are attributed by two aspects:

1) Generally, there are two classes of models often required.
   - The covariates are constant within each household. For example, for imputing many household variables, the basic aggregate statistics over household members can be quite informative, e.g. maximum education level, average age, existence of employed household member and so on. Also a household variable (e.g. HI0100) can often serve as a covariate for the imputation of a personal variable.
   - The covariates are not constant within each household. For example, the covariates needed to impute the wife's employment status can include the employment status of the husband or even of the adult offspring (e.g. mother might have to work if her son is still a "lifelong" student).
2) The information in the household level and person level is interdependent. Therefore, we should consider the stability of a joint distribution over all of them as the ultimate imputation goal. This can be achieved only if we allow simultaneous imputation of household data and personal data.

In result, we decided to use one big dataset for imputation that contains all household level variables and all personal level variables.[6] Then, each row contains the P-variables for each person. Additionally, H-variables are attached to each row in different degree according to the status of the person. If he is the FKP (Financially knowledgeable Person), all the H-variables are attached. Otherwise, we only insert the values of those which are used as covariates in the specification of any P-variable imputation. The rest are set as missing. This measure can greatly save the data size.

When we want to impute an H-variable, we can subset the data to keep only one row for one household (i.e. the row for FKP). The estimation is run over them. When we want to impute a P-variable, we actually pool all the household members together in a regression. Then, those "outlier" family members (e.g. the fifth child) are not an issue any more when the sample size of observed cases is concerned: they share the same estimation with all the rest.

We constructed particular macros in SAS to realize two kinds of specification aforementioned: aggregate statistics over household members and the variables of another household member as covariates. The specification can be very flexible – this can also be contingent on the missingness: we can calculate an average employee income across household members and force this to be missing when, for example, some of the household members have this variable to be imputed. We consider this feature in order to avoid the exaggeration of the within household heterogeneity and/or the distortion if missing not at random is present. Another important step in our routine is to always update these particular covariates whenever they (or one member value of them, e.g. any member's employee income when average employee income is considered) are changed during imputation.[7]

3.3 Logical trees

Every variable has a related flag variable. The coding of this flag variable consists of the standard values 1 (answer from respondent), 0 (system missing) and several four-digit special codes. Values greater or equal to 2000 indicate that they will be imputed. Thus during the preparation steps before imputation, all of the flags for the cases which will be imputed potentially, have to be replaced by a value greater or equal to 2000. The values -1 (don't know) and -2 (no answer) usually have flags 1000, 1001 or 1004 (if they were not changed during edit-

[6] We are grateful to Dimitris Christelis for his suggestion on this data format.

[7] The alternative is to use the wide-format. Here every variable from the P-file appears n-times (n=max.(observed HH-size)), representing the values for the 1st, 2nd, … person in the households.
Another possibility to create a wide-format is to assign determined family patterns to the duplicates of the p-variables (e.g. 1st=KT, 2nd=partner, 3rd=child#1…). The second one is more efficient since the first one still needs the efforts to determine the position of each person in the family before most specification can be established.
There are two disadvantages concerning the wide-format: 1) the computation time for the imputation grows dramatically; 2) the data observed for the 5th, 6th, … person is very sparse, so it does not make much sense to build imputation models for them.

ing). So all the flags that correspond to values -1 or -2 are, if necessary, increased by 1000 to conform to the imputation flag definition.

During the imputation procedure no flags are changed.

In many cases a non-response of one variable causes missing values in other related variables. If, for example, the question relating to the ownership of property was answered with "don't know", the entire property section of the questionnaire is left out during the interview. If the initial question is imputed as "yes", then the subsequent property-related questions also have to be imputed.

The logical relationships in the PHF questionnaire are reflected by logical trees of variables. In each tree, one variable is called head variable, and the remaining ones are branch variables which are logically dependent on the head variable. Of course, the branch variable of one tree can also be the head variable of another tree in the lower order of logical sequence.

It is very critical to assign the correct flag value to the branch variables according to the values of the head variable since many candidates for imputation are among the branch variables due to the non-response of the head variables. Consequently, we assigned the flag 2002 (meaning *imputed, originally not collected due to missing answer to a previous question* to every branch-variable) if the head-variable contains the values -1 or -2.

Example: In the raw data, there is a case with:

HD0100=-1 , HD0100FL=1000 , HD0200=-3 and HD0200FL=0.

(HD0100: Investments in businesses; HD0200: Investments in self-employed businesses)

The question for HD0200 had been asked only if HD0100=1. In the above case it was skipped during the interview due to the non-response in the head variable HD0100. After imputation preparation these turn into:

HD0100=-1 , HD0100FL=2000 , HD0200=-3 and HD0200FL=2002.

Branch variables in the logical trees can be imputed only if the value of the head variable imputed or observed satisfies the conditions that allow questioning the branches (in the above example: only if HD0100 is imputed to be 1, a value of HD0200 can be possibly asked and hence imputed). To make sure that imputation causes no inconsistencies within the logical structure of the questionnaire, the imputation code of every single variable contains a WHERE-condition that reflects the logical condition that an observed or imputed answer in this variable can be expected. For example, the imputation code for HD0200 contains the WHERE-condition "HD0100=1".

Since the PHF variables are imputed sequentially, we had to make sure that the head variables had always been imputed before their branch variables. This is also facilitated by the establishment of logical trees. In summary, by building the logical tree, a value is imputed if and only if the flag value is greater or equal to 2000 and the WHERE-condition is fulfilled.

3.4 Further preparations for imputation

If the interviewed persons were not able to provide an exact answer to an Euro value question, they could instead give lower and/or upper bounds for the value. And if this was not possible, they could choose an interval, which fits closest to the possible value, from a list of ranges prepared in the questionnaire.

In both cases, the midpoint of the interval is calculated as starting value for the imputation. We then specify the bounds using the intervals provided into each imputation model.
All continuous variables are transformed into their inverse hyperbolic sine (IHS) before imputation. The whole imputation is based on the distributions of the transformed values.[8] Year values are first transformed into the difference of the provided year to the year of the interview (e.g. 1998 -> 2011-1998=13) and then in a second step transformed to the inverse hyperbolic sine. After imputation all of these values are retransformed.

Categorical variables have been transformed into an ordinal scale, if possible. With this ordinal treatment, the categorical variables can fit the requirement of the randomized hot-deck imputations when we have to collapse the neighboring values of the conditional variables, i.e., these categorical variables in order to maintain enough cell size for random draw. For example, DPA0300 (highest level of education) was reordered to reflect ascending educational level.

Before imputation we constructed several auxiliary variables. They are mainly used as regressors for imputation, e.g. age squared and average personal income within household.
During the editing of the PHF data, miscellaneous consistency checks were used to assure the data plausibility. This was mainly done by imposing the bounds for the imputation of continuous variables or by building the value set feasible for the binary or categorical variables.

4 Model specification

In section 5.1 we present the problems and the corresponding solutions of our model selection process. The practical implementations are revealed together with the diagnostic analysis in sections 4.2 and 4.3.

4.1 The purposes

In terms of the practical consideration for a large household survey, we need to consider a huge number of candidate covariates (main data and many auxiliary data) in imputation model. Besides, accounting for enough interaction effects (they can be really many when there are categorical variables and many candidate covariates) can tame the impact of influential obser-

[8] Even though log transformation is often used in the economic literature, the imputers might choose the other proper transformations: e.g., Christelis (2011) used IHS too and Drechsler (2011) adopted cubic root for various reasons. We will explain the motive for IHS later.

vations. To cope with this large scale of candidates, we resort to the technique of model selection.

There are also many other practical considerations which require a robust model selection routine:

a) We are working with real world data. There are always some anomalies. For example, some variables presumed to be predictive might not have enough variation (e.g. due to notorious issue of self-rounding/bracketing). We do not have resource to detect one by one without automatic algorithm.
b) Congeniality requirements (according to Meng (1994)), or broad conditioning, often suggests the imputation model to include as many predictors as possible so that any potentially important variables in the analyst's model will not be ignored. Otherwise, there can be bias. On the other hand, a "Full Model" with many insignificant predictors can accommodate various missing pattern so that there are always enough "substitute" predictors available if some others are missing. However, the major pitfall is it is likely to overfit the data with a model performing poorly in out-of-sample prediction (i.e. imputation). Additionally, this can lead to a much longer computation time which might not be economic in practice. Thus, to seek a balance, we adopt the stepwise selection to produce an economic model with proper fitness. This should not increase the risk of uncongeniality since those highly correlated variables in a joint setting will always be maintained when enough model diagnosis are applied (see this argument from Drechsler (2011)).
c) Partial F-test in traditional selection method does not follow F distribution. The alternative solution is to use information criterion as selection and stopping rule.
d) Models selected by the traditional criteria often do not perform well in out-of-sample prediction such as the case of imputation. To mimic such an environment, we always randomly select a test sample which has the imputation target variable all observed and is separate from the training sample (used for fitting the model). This test sample is then used for a predictive performance check. Since it is independent from the role of the model selection, this criterion is actually an out-of-sample prediction criterion.
e) Due to missingness, the sample sizes available for training and test data could be often quite small. Hence, there can be no space left for a validation sample used to obtain prediction error for determining the moment to stop the selection procedure and/or inclusion of candidate covariates. We then reply on the cross-validation which allows the sharing of one sample between the test and validation stages.[9] Besides, this technique is also known for enhancing the out-of-sample prediction performance of the selected model.

[9] Basically, the sample is split into k parts. When model fitting is performed on one part, the omitted parts are used to calculate the prediction error. The process is repeated for k times. We then sum up these k sets of prediction errors as one criterion.

Abayomi et al. (2008) and Drechsler (2011) recommend both internal and external diagnosis should be evaluated in the model selection procedure. The intention of the former diagnosis is to examine the fitness of the selected model because our imputation is model-based. The latter emphasizes a subjective evaluation on the plausibility of the imputed outcome. Particularly, they investigated if any difference in the distributions of the covariates can explain any possible discrepancy between the distributions of observed and missing samples. Next we will elaborate the implementation of these two aspects in the stage of model specification. Especially, as an extension of external diagnosis, before kicking off the imputation, we also have an ex ante evaluation of potential imputation quality by performing an out-of-sample prediction using the selected model on the missing sample.[10]

4.2 The core procedure[11]

We feed the module with a broad range of candidate covariates: basic demographics, household aggregate wealth and income, the variables from all the auxiliary datasets which are almost nonmissing such as stratification, characteristics explaining non-response behavior and other design features, almost all of the other variables from the same section and the case specific ones (e.g. those economically correlated but in other sections, head variables determining the participation of those predictive covariates included in order to correct potential selection issue, …).[12][13]

The SAS manual for PROC MI (see Imputation Model on P. 3798 of SAS/STAT ® 9.2 User's Guide) summarizes a guide line for model specification which is well shared among many imputation practitioners:

"Generally you should include as many variables as you can in the imputation model (Rubin 1996), at the same time, however, it is important to keep the number of variables in control, as discussed by Barnard et al. (1999). For the imputation of a particular variable, the model should include variables in the complete-data model, variables that are correlated with the imputed variable, and variables that are associated with the missingness of the imputed variable (Schafer, 1997; van Buuren et al., 1999)."

---

[10] Missing or observed samples in this paper refer to the cases where the dependent/targeted variable is to be imputed / missing or observed.

[11] This procedure is only applied to a core set of variables. The diagnostic analysis and comparision between competing models are still heavily involved with human interference and determination no matter what degree of automation we have programmed in the rest of the procedure. A full scale of application will be confronted with resouce constraint. Therefore, we decide to focus on the set of variables which bear these features: economically and/or logically pivotal, with relatively high missingness, significant in determining the household net wealth and potentially MNAR. The list of these variables can be available upon request. The model specification for the rest of the variables follows the principle of broad conditioning, which is an economical version of the "starting point" model as described below. The sensitivity analysis shows some variation of the model specification on them, introducing very minor impact. The relatively low item-nonresponse rates in our survey might be the explanation.

[12] Without including this participation (dummy) variable, the model might neglect the fact that all the non-participants can be systematically different from the participants which can bias the prediction. Additionally, we replace the missing covariate of those non-participants by a constant (e.g. zero). A detailed discussion is extended in the section 6.2.

[13] We always include survey weights in all the final specifications to account for the stratification and oversampling effect (Reiter et al., 2006).

Our choice of candidate covariates mentioned above follows these suggestions, particularly the broad conditioning which intends to cover all the variables any potential data analyst might use. The set of these covariates themselves builds a "starting point" model. Ideally, this "starting point" model should not be single. One model including one particular covariate might exclude quite a large subsample which has missingness on this variable. For example, a model including the value of HMR (household main residence) cannot be applied to those renters. In this case, we should also consider the other "starting point" model without the value of HMR to cover the renters. Since the model selection can be deemed as a refinement process following each "starting point" model, there should also be equal numbers of independent selected models to cover each subpopulation. Theoretically, this can lead to a large number of independent model selection procedures and constructing imputation models for each subset. For practical matters, we rarely do this. Instead, we parameterize these covariates as we elaborate in the section 5.2, which allows us to run one single set of model selection over the whole sample. Actually, after we select some competing models, we also run both the internal and external diagnoses on this "starting point" model. If there is no clear sign of overfitting or that the selected model outperforms the "starting point" model in an ex ante evaluation of the imputation outcome, the "starting point" model is always preferred.

Now we start to discuss the refinement process and the internal diagnosis:

a) Selection by missing pattern. Since the model selection procedure uses the casewise deletion to form the samples for fitting and scoring, there is a tradeoff between larger training sample with enough statistical power and keeping the predictive covariates with high missingness. We trim the variables by these rules:
   1) The very first set to be dropped contains those with high pairwise missingness w.r.t. the dependent variable.
   2) The next group consists of those pivotal variables appearing in many missing patterns.[14]

   However, we prefer the sacrifice of the size of the training sample to the exclusion of any covariate bearing the mild correlationship with the dependent variable. In practice, we observe our model selection procedure can be quite robust even under small sample size.
b) Selection by the model selection technique. After the screening, we begin to run the PROC GLMSELECT and regression diagnosis.[15] Here are the major steps:
   1) We use PROC GLMSELECT to randomly form the training and test subsamples from the observed sample. The cross validation will be performed within the training data to make decisions on the covariate to choose on each step and when to stop. The scoring will be performed on the test subsample. We check the sequence

[14] We apply PROC MI in SAS/STAT to list all the missing patterns.
[15] The core of this selection routine is PROC GLMSELECT, a procedure in SAS/STAT package.

of average square error (ASE) for the test data. If this goes up, this can be a sign of overfitting. Under this circumstance, we could adjust the procedure for another run (e.g. modifying the pool of the candidate covariates, allowing more interaction terms to be tested, changing the selection criterion, …).

2) The selected model is then further examined by regression diagnosis, esp. on the issues of fitness and influential observations: SAS/STAT provides many fit diagnostics, among them, the alignment of the residual and the difference between fitted value and mean is a powerful tool to assess the fitness (i.e. "Proportion Less" plot / quantile - quantile plot). The plot of the residual and leverage is the other tool used to detect the influential observations. The routine will always perform the diagnosis both before and after dropping some outliers. This can help us to determine the impact of the outliers on the model fitness. When they cause the overfitting, we will exclude them in the imputation model.

Figure 0 Internal diagnosis plots for DHI0700 (self-reported household total wealth): before and after the trimming of outliers[16]

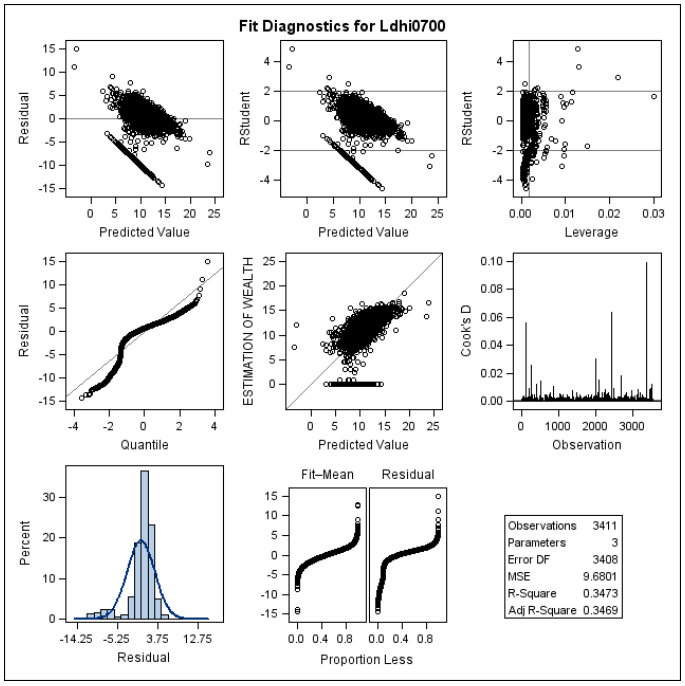

Figure 1-A Before the trimming of outliers

[16] This set of plots is produced by the default diagnosis graphics of PROC REG in SAS/STAT.

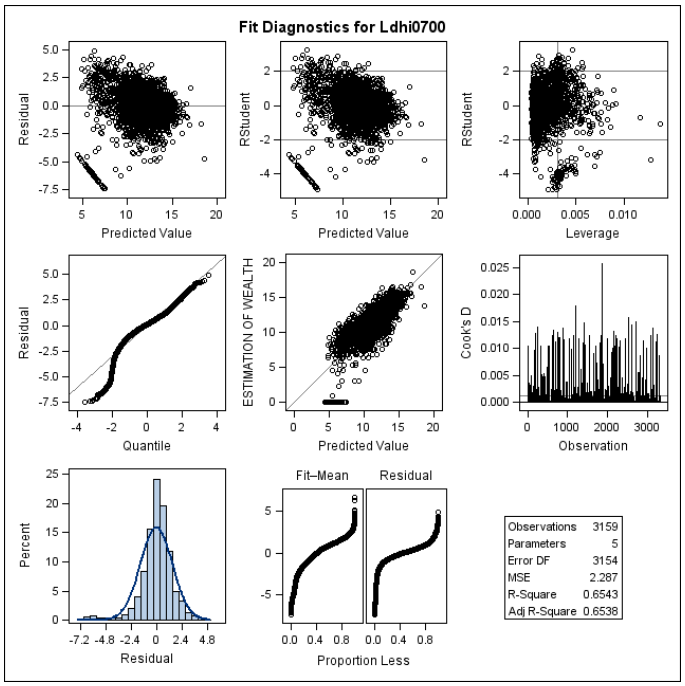

Figure 1-B After the trimming of outliers

Figure 1 presents a case study on how the diagnosis assists us in trimming the outliers. Many respondents answered zeros for DHI0700 (self-reported household total wealth) which might cause the underestimation as suggested by the opposite relationship between predicted value and residual in Figure 1-A (the strip results from these zeros). We then drop all the outliers using a common criterion: |RSTUDENT|>2.[17] Among them are many zeros. Figure 1-B is the diagnosis result after refitting. The opposite relationship between predicted value and residual is not prevalent any more. "Proportion Less" (quantile/quantile plot) shows our model fitness improves and we can indeed reliably predict/impute those in the lower end of the distribution (due to maintaining some valid zeros which are not the outliers).

3) As illustrated in the above example, the regression diagnosis can guide us on whether to accept the selected model for this stage or make further adjustments by repeating from the step 1).

Sometimes, we have to run the sensitivity analysis to produce various competing models due to irregular missing pattern, logical constraint, and specific modeling concern etc.

4.3 Predications of the imputed and observed samples

Although out-of-sample prediction error is assessed by using test and validation samples during the model selection procedure, it is still necessary to perform a prediction on the missing sample using the selected model and compare it with the predictive distribution of the ob-

[17] RSTUDENT is a studentized residual with the current observation deleted.

served sample. The prediction on the missing sample can be considered as an approximate view of the imputed outcome. By our subjective evaluation on this distribution and comparison with the predicted observed sample, we could make a first judgment on whether the imputed distribution and/or single values do make sense economically and/or statistically.

As discussed above, it is critical to capture the variables explaining the missingness. The difference between the predictive distributions on the observed and missing samples does signify the success of including such covariates. However, to decide the sufficient numbers of such variables selected in our model, we need to compare with a baseline benchmark: the difference of the predictive distributions from the observed and missing samples using a model containing those covariates which are almost always nonmissing (basic demographics and those from the paradata).[18] Our decision can be justified by the additional variation of the difference due to the newly selected variables. It is not uncommon that we might obtain several competing models when tuning up our selection procedure. Then this baseline benchmark becomes further indispensible because of the reason discussed below.

Many models selected can pose the restriction and/or affect the impact of influential observations in various degrees. One typical concern is that the eligible training and missing samples might be a particular subsample. For example, when the value of HMR is included as a covariate, both samples should be no more than the home owner subsample. However, conditional on being homeowner, the distributions of both predictive observed and missing samples might bear exceptional statistical features (e.g. a much more skewed distribution of self-reported wealth). Given the comparison with the baseline benchmark, we are able to quickly identify and justify the source of increasing skewness. The similar comparison could also be informative if the explaining power of some other covariates on the missing pattern varies when conditional on being home owner.

The selected model can introduce more sensitivity on some influential observations relative to the baseline benchmark. The parallel comparisons of both observed and missing samples with the baseline benchmark can provide the justification on the severity of outliers: whether there is overestimation/underestimation on the missing sample due to them. Figure 2-A displays a baseline benchmark for DHI0700 (self-reported wealth): the distributions of prediction produced by the baseline model on both the observed and missing samples.[19] Figure 2-B illustrates the predicted distributions when an imputation model is selected.[20] As presented previously, quite a few of zeros are maintained in the training sample for DHI0700. Compared with the baseline benchmark, they affect the predictive distribution on the selected model by creat-

[18] This construction can serve as a baseline model because all the models selected can be considered as almost always the extension (e.g. containing more covariates). On the other hand, we use only those covariates rarely missing which can cover the most cases in the observed and missing samples.

[19] The value is transformed by HIS.

[20] The selected model forms the slightly different subsamples for both observed and missing samples relative to those in the baseline benchmark: the jointly nonmissing cases are not exactly same due to introduction of additional covariates. We have examined that this difference is not correlated with the following interpretation.

ing a dualism (i.e. adding the other mode in the lower end) and shift the overall distribution. We observe the degrees of these two impacts are equivalent in the observed and missing samples: the relative location and scale of the lower mode is almost same and the means of both distributions decrease by 2%. The caution that the zeros might result in extraneous imputed outcome seem to be unsupportive.[21]

Figure 1 Predictive distributions of observed and missing samples for DHI0700 (self-reported household total wealth): baseline benchmark vs. selected model (histogram, kenal density (red dashed line) and normal fitted curve (blue line) are displayed)[22]

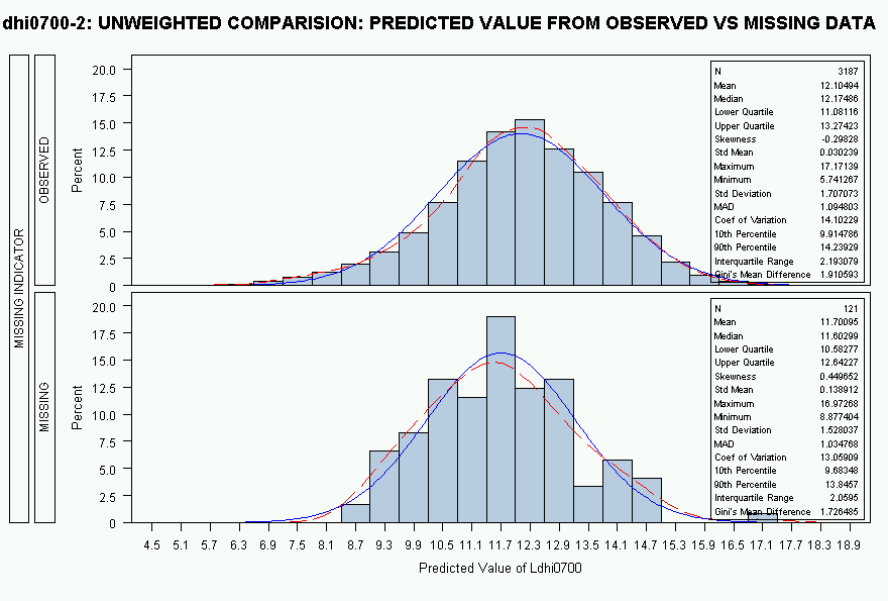

Figure 2-A baseline benchmark

[21] We acknowledge that all the diagnosis in the model selection procedure might not be conclusive in terms of the final imputed outcome due to mainly two reasons: 1) the covariates and observed dependent variables can change when imputation iterations evolve; 2) we cannot observe the prediction on some cases with missing covariates within the selected set due to casewise deletion, which might be quite exceptional. However, in our experience, they are not detrimental. One major reason is our relatively low item-nonresponse.

[22] DHI0700-2 means the second round of model selection (there are a number of competing models). LDHI0700 denotes the DHI0700 transformed (by HIS).

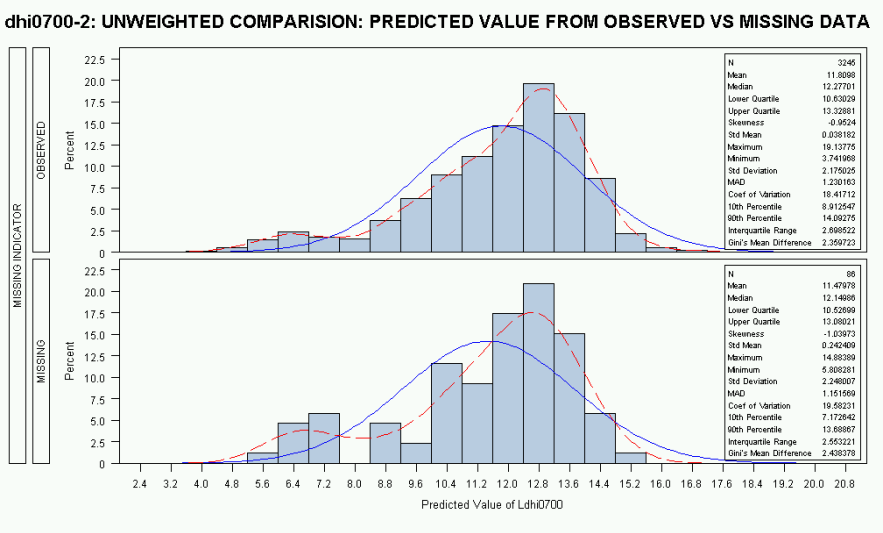

Figure 2-B selected model

5 Specific modelling technique for real-data problems

The other side of challenges on imputation specification comes from the irregular data structure and the logical constraints imposed by a real-time survey. Based on the imputation of German Institute for Employment Research (IAB) Establishment Panel, Drechsler (2011) has a comprehensive discussion on many of real-data issues as well as the possible bias or inconsistency when standard routines are applied. Next, a number of critical problems and solutions in our data are visited.

5.1 Semi-continuity and heteroscadacity

It is often observed that many survey variables designed to be continuous are not fully "continuously answered", e.g., there are spikes and truncations. Many observed values are self-reported which are mostly the multiples of 100 (e.g. 155,000, 3,700 etc.) and concentrate on a small number of points (this is the notorious self-rounding issue). Respondents often do not have enough market information and/or are quite conservative when asked about the value of some assets, e.g., the value of the cars. Then they tend to answer a value of zero for those old and undermaintained cars. We then observe a mass point on zero. There are some institutional restrictions to constrain the variables. For example, almost all the hedge funds have the requirement of least amount invested. The truncation is then created.

To well maintain these data patterns and avoid imputing impossible values in the imputed sample, we widely adopt the predictive mean matching (PMM). This means that the initially imputed value (i.e. the predictive mean plus a random draw from the estimated residual distribution) is matched to (or imputed by) one observed value with the smallest distance between this initially imputed value and the observed value.[23] This mechanism is pioneered by Little

[23] This is a variation of the inital PMM proposed by Little (1988) which defines the distance measure as the difference between the predicted means of both donor and beggar. Another kind of PMM is to add the empirical

(1988). One specific advantage is that it is more robust to model misspecification than simply drawing random residual, which can be sensitive to assumptions about the variance structure (David et al., 1986).

Another advantage of PMM is its superiority in handling heteroscedasticity. For instance, rich respondents are generally more financially sophisticated and better informed about their financial situation than the poor one. Thus there is supposed to be less reporting error for the first group in our survey.

Like many other surveys, the spike on zero is often present in the distributions of many value variables. Many of such variables are headed by a filtering variable asking the participation. And we always impute these filtering variables first and construct the imputation only over whoever participates. This is simply the two-step or hurdle model labeled in the literature (Cragg, 1971; Kennickell, 1998; Raghunathan et al., 2001; Barceló, 2006; Schunk, 2008).

Given the fact there are still many zeros answered even using such a participation filtering, it is plausible that many of such respondents are on the margin of participation.[24] On the other hand, it is possible some of them misreport the values in the sense that their residuals can be rather high in a correctly specified model. To account for these two kinds of zero reporters, three techniques are combined: outlier detection, inverse hyperbolic sine (IHS) transformation and PMM. We first screen out those zero cases when constructing imputation SSCP if they are assessed to be extraneous in the regression diagnosis. We adopt IHS transformation for the imputation model of most continuous variables. Generally, this can improve the approximation to the joint normal assumption which is required by many literatures on multiple imputation and MCMC methods. In the current application of handling zeros in the observed sample, the special consideration is that it is a continuous transformation crossing zero and the steepness on zero is much lower than the other commonly used transformation logarithm. This feature permits the prediction of negative values for those on the margin of participation. Given the PMM, we could impute zeros for them as the nearest neighbor. Instead, using logarithm under PMM, quite a few of them would be matched to some small positive values because they are among the observed values. We would then possibly produce a mass on these values other than zeros in the imputed sample. However, these small positive values are rarely observed to be the spikes in the observed sample. For many users who are less sophisticated in imputation, this outcome can be misleading particularly when they are interested in modeling the zero mass point.[25]

---

residual from the matching donor to the predictive value of the imputed beggar (Kalton et al., 1985). This method can avoid the clustering which can be valuable when missing rate is high. Our version can be deemed as a combination of these two: our distance measure is actually the intial one plus a difference between esti-

[24] mated and empirical residuals. This additional term can introduce a penalty on the match with the outliers.

[25] This is evident from many observed correlated variables.

For instance, those reporting zeros can be a specific subpopulation who might be unconfident, pessimistic, impatient or lack of financial sophistication and/or market information.

5.2 Jointly exclusive covariates

Some important covariates never coexist for any case because of the logical tree in the questionnaire. For instance, the housing value of the main residence and the rental payment are asked separately for different subsamples depending on their residence status. Ideally, we should subset our imputation models: for example, one for the subpopulation of house owners and one for the subpopulation of renters. However, this means an overwhelming effort of data manipulation and model selection if these are two important covariates for many models. We take a compromising approach: the housing value is set as zero if the respondent is a renter and vice versa, plus we include the binary variable on residence status (namely a transformation of the head variable indicating the participation of housing market). This variable will reduce the selection issue when house owners and renters are fundamentally different group of population (see Albacete (2012) for the same argument).[26]

5.3 Imputation of three particular multiple choice/categorical variables

We have many categorical variables to be imputed. Questionnaire imposes some explicit or implicit constraints on them. We separate them into three classes:

1) **Ordered** (e.g. DHI0300A-M, reason of saving). We use sequential hotdeck imputation by respecting the order while controlling the fact that the answers should be exclusive with each other: it is impossible to answer the choice which has been given previously in the sequence. The latter is done by building an exclusion list containing those choice values imputed previously in the sequence.
2) **Unordered, multiple choices allowed and only joint missingness can occur** (e.g for DPA0200A-E (legal status marriage), CAPI presents an answering box where interviewer can type in any choices the respondent picked up from a list or -1/-2. The former scenario will lead to the answers of "NAMED/NOT NAMED" in DPA0200A-E and the latter one will produce -1/-2 to all of these five variables). A sequential imputation might yield an outcome such that each variable becomes "NOT NAMED" which is logically wrong. Therefore, we have to control this implicit constraint. Here is an approach we adopt: taking DPA0200A-E as an example, we map the combinations of these five variables answered to a set of index (they are actually formed as a binary number with each digit representing the outcome of one variable).[27] All the missing cases will be first imputed by the hotdeck method drawing from these observed indexes. Afterwards, a transformation will map the imputed index back to the response in each variable.

[26] If the selection issue is really serious, the other approach is to turn on the other kind of PMM in FRITZ: use sample residual of the near neighbour. It will allow that the estimation of SSCP is tailored to the subset of covariates observed for each case to be imputed. However, this can dramatically extend the computation time since it can be very likely to require a large number of separate calculations of SSCP and its inverse.

[27] This approach is initially proposed in SCF byArthur Kennickell.

3) **Unordered, multiple variables in a list answered and single missingness can occur** (e.g for dhb1200a-h (number of other vehicles), CAPI allows that a non-negative integer or -1/-2 is answered for all items). The implicit constraint is that it is impossible that all items are zero. The property for this kind of variables is very similar to that in 2). But the only exception of single missingness prevents us from easily adopting the same approach as in 2): for example, imagining dhb1200b is two, dhb1200c is -2 and all the rest are zeros, we have to restrict our indexes to be hot decked in order to meet this particular combination. This is not so straightforward to code. Besides, it is possible this particular combination does not exist in the observed cases.

   We then resort to a compromising approach: impute the aggregate value (setting a lower bound of one to make sure it will be positive always) and use it as covariate to impute each component variables. If each component variable (observed or imputed) is zero, we force the lower bound of the last variable "others" to be one while imputing this variable. The extra preparatory work involved is to edit this "other" variable to be -1/-2 whenever there is -1/-2 among other component variables and this "other" variable is zero (this is to assume there is possibility the respondent might have difficulty to make sure of the right category when he reports -1/-2 for one category). This is not an ideal solution because there is no clear rationale why this "other" variable becomes "exit/residual" variable, particularly this might mean some insignificant/very idiosyncratic component in the mind of the respondent. However, it appears this is not the story in our data: there is a quite high frequency of positive numbers answered for this "others" in such kind of questions, which implies this does represent some specific categories ignored in our list. In this situation, it is relatively convincing to assign a positive number to this "other" variable while the respondent enters into these questions and all the rest of variables are imputed to be zeros.

5.4 Bracket imputation and editing constraint

There are two main bounds imposed during imputation: the intervals provided by respondents and many of the editing constraint, e.g., the total years employed should be smaller than the age minus 15.

There are mainly two approaches to handle bounds when imputing: drawing from a truncated distribution using the bounding information or repeated drawing from an untruncated distribution until the value meets or is forced to hit the bounds. The first is econometrically more accurate and computationally more efficient. However, this requires a subsample satisfying the truncation/bounding totally in order to estimate the parameters of the truncated distribution. In practice, this is too restrictive since this kind of subsample can often have a very small size or even does not exist in the observed data given the complex logical structure and missing pattern. In addition, many such observed distributions are highly skewed. These factors can result in a collapse or difficulty in convergence during estimation. FRITZ uses the second approach which is more common in practice. Since it uses the information from all the observed

cases, we will always feasibly estimate and impute the model.[28] But the pitfall is that the model might be misspecified. In another word, only when we do not impute many values hitting the bounds, or alternately, the probability of drawing implausible values is very low, this approach is safe (Drechsler, 2011).

The PMM approach discussed previously can effectively prevent drawing implausible values. To make sure the bounds provided are compatible with those from the observed sample values, we resort to a feature in FRITZ to align both. In addition, it is also involved with tremendous efforts in ex-post validation: we read carefully the log file to monitor the events when the random draws collapse on the bounds and adjust the editing constraint and/or our specification if this count is too high.

5.5 Multiple assets answered in order.

The PHF questionnaire contains several loops asking for details of e.g. the three most important mortgages or the three most important loans. In case there was too much non response in the key variables of the loop (e.g. more than 50% don't know or no answer in hb170$x) the interview continues with an additional question about the total value of all mortgages/loans (e.g. dhb2600). These additional variables have not been imputed; instead they serve as upper bounds for the imputation of the key variables in the loops. Mostly, the answering order means something different and a separate specification for each asset is more proper.

The deviations between the sum of the single values and the estimated sum value are possible and generally not corrected. Schenker (2006) summarized the reasons not to enforce this kind of consistency: the size of inequality is not ignorable in the observed data, the effort to impose such equality constraint tends to distort the marginal distribution and this does not have much impact on the core research question. Jaenichen (2012) also discussed many institutional reasons on the similar issue for the income variables in Panel Arbeitsmarkt und Soziale Sicherung (PASS).[29] Besides these two studies, Barceló (2006) also left alone such inconsistency in Spanish Survey of Household Finances (EEF). Likewise, we take no action for this case as well as for some other implicit constraints with similar characteristics.

[28] We would like to give credit to Cristina Barceló for her comment on this point in the mailing list of imputation subgroup of HFCN.

[29] We thank Tobias Schimdt for drawing our attention to this study.

6 Item-nonresponse in PHF[30]

In order to obtain a comprehensive picture on our item-nonresponse, we selected a number of variables from each section and calculated their item-nonresponse (missing) rate in **Table 1** according to a couple of definitions. Here are the selection criteria which can be jointly satisfied in some cases: they are economically critical (i.e. many of them are pivotal covariates used in many imputation specifications), they are representative for the distribution of the missingness in each section (i.e. those with high, middle and low missing rates are picked or otherwise single variable is presented when the distribution is flat), they well spread across each unit of the section and/or they are the exceptional cases. In doing so, we believe this presents a representative subsample where our analysis of the item-nonresponse can be reasonably well extended to the whole sample.

In order to address the impact from the filtering structure of the questionnaire, availability of additional bounding information and the imputation of filtering variables, we calculate five missing rates:

a) Missingness by DK/NA (% of *ex ante* applicable cases).[31] This *ex ante* applicable cases exclude the case which is filtered by the head variable, and thus, which the response flow never arrives. This is the item-nonresponse rate many surveys report and use as a benchmark for comparison (e.g. Christelis (2011)). However, this ignores two important features:

   i. The filtering can also happen due to the fact that some of the head variables in the higher order of the logical tree are not answered (e.g. some cases of HB0900 (value of household main residence/HMR) are filtered because the respondents answered DK/NA in the head variable DHB0200A-D (ownership share of HMR)). These are actually the cases which will be potentially imputed as long as these head variables are imputed such that the response flow can arrive this variable (this is also called "participation"). Therefore, they should also be counted as the cases to be imputed.

   ii. As illustrated previously, all Euro value variables allow the user to answer a bound information if they respond DK/NA. Whenever this information is given, the imputation should be much more accurate. An effective item-nonresponse rate should exclude the cases when a bound is provided.

[30] Above all, our questionnaire team should be given the credit because we observe some of the careful design helps in reducing item-nonresponse (though a thorough comparative study is required to formally evaluate these observations):
   i. Respondents have various flexibilities to answer many value variables which each might know in terms of different reference period (yearly, quarterly, monthly and some months of the year if the cash flow does not last for the whole year).
   ii. It is creative and effective to allow the respondent to answer either gross or net income. It is quite plausible that many respondents have only the information on one kind of income (e.g. many low-paid employees do not have full access to the payrolls and/or they do not often intend to maintain them well).

[31] The number of DK/NA counts the total number of cases answering "don't know" or "no answer".

b) Missingness by DK/NA in a variable itself or any higher order head variables (% of total number of *ex ante* applicable cases and cases with DK/NA in any higher order head variables): this is to account for the potential imputations discussed in a) i.

c) Missingness by DK/NA in a variable itself or any higher order head variables excluding the cases with bounds (% of total number of *ex ante* applicable cases and cases with DK/NA in any higher order head variables): this is to account for the feature in a) ii.

d) Missingness by DK/NA in a variable itself or any higher order head variables excluding the cases with non-participation imputed *ex post* (% of total number of *ex ante* applicable cases and DK/NA in any higher order head variables): this is to count only those cases with variable participation, or say to condition on the imputed participation.

e) Missingness by DK/NA (% of ex post applicable cases). This corresponds to definition a), but uses the number of realized imputed cases (imputed participation) for enumerator and denominator.

Notice these rates are calculated based on different datasets. a) uses the raw one. As discussed before, there is an imputation preparation step to assign the imputation flag value according to the missingness of the head variables. Both b) and c) are calculated after the raw data went through this step. Finally, d) and e) have to rely on the imputed data. The calculations are as follows:

a) # Missings / # Cases (ex ante)

b) # Missings + # Missings in higher order head variables / # Cases (ex ante) + # Missings in higher order head variables

c) # Missings + # Missings in higher order head variables **- #** Interval values / # Cases (ex ante) + # Missings in higher order head variables

d) # Imputes cases (ex post) / # Cases (ex ante) + # Missings in higher order head variables

e) # Imputes cases (ex post) / # Cases (ex post)

Due to logical constraints, the rate in a) is smaller equal than the rate b), and c) smaller equal b), d) smaller equal b), a) smaller equal e).

We focus on the comparison across sections and variables in our survey to shed light on the potential behavioral interpretation of individual item-non response and its impact on the information content from imputation. There is a baseline friction for the general degree of item-nonresponse. This can be derived from the levels of cooperation, trust, attentativeness as well

as the recall efforts required for household financial questions. However, this paper is not going to address these factors.[32]

Generally, according to the benchmark rate a) and more effective rate c), the item-nonresponse is not severe in our data ((a) will be cited in the following discussion). Notice we sort the Table 1 according to the section average rate (a). Particularly, many economically critical variables were answered quite well in terms of item-nonresponse relative to the other variables in our survey: e.g., most variables about HMR in the housing section (section 3) have quite low item-nonresponse rates (between 0.5% on size and 8.9% on value). The information on employment status and time/history (section 7) is even better (between 0.1% on current employment status and 1.3% on total time in employment). Following them is the consumption section (section 2; between 0.25% on food consumption and 11.1% on estimation of wealth). Our respondents seem to not have much reluctance in answering income questions (particularly 6.8% on income from employment and 4.4% on social transfer income) (section 9 and 9.2), which surprisingly contrasts with the suffering in many other surveys. The exceptions are the private pension incomes (23.9%) and income from financial investment (21.9%).[33] Respondents do not have difficulty in recalling the values of donations and gift/inheritances but church tax seems to be challenging (section 6; between 1.4 and 7.5% for the former two and 25.5% for the latter).

The following variables/sections are relatively worse. Respondents seem to have difficulty in recalling/reporting the debt on credit card (section 4).[34] It is not surprising that owners might not precisely know the value of their business (24.7%) and both the flow (e.g. 42.9% on saving in certificates) and stock (e.g. 35.1% on bond) values of saving in financial vehicles (they are not answered well except the questions on saving account (0.4%; section 5). Many questions in the section of pension and insurance have quite high missingness (section 8; with many well above 30%). In general, we can postulate that

- the respondents make good effort to trace the information as long as the reference documents can be reachable and straightforward (e.g housing transaction statements and bank statements about saving accounts and payrolls),

- they have difficulty in recollecting the information when the question might involve with reference documents which are associated with multiple sources, uneasy to be

[32] ECB will soon publish a metadata report, where a cross country view is provided on missing rate. This might offer some benchmarks for assessing these underlying frictions.

[33] Neverthless, the questions on the ownership of these incomes (e.g. financial investment income) are not subject to high missingness.

[34] One important feature in German credit card industry might explain this: many debitors will allow the banks to pay off the whole or part of the debt by an automatical transfer from their linked current accounts in the end of each billing cycle. They often just ignore the monthly statement or there is almost zero balance duto automatic transfer.The distributions of DHC0610 and HC0320 do mass over some positive neighbor of zero.

identified or even unavailable (e.g. questions on private pension, financial investment, church tax, business value, particular saving contracts),

- they intend to reveal what they are proud of and meaningful in their lives (e.g. donation and gift/inheritance),

- and it seems questions requiring incidental instititutional knowledge are challenging for them (e.g. private pension and particular saving contracts).

Furthermore, we define three types of change: I. between (a) and (b); II. between (b) and (d); and III. between (b) and (c). The first and second comparisions can reveal the *ex ante* and *ex post* impacts of logical trees in the questionnaire (i.e., additional missingness caused by the variables in the higher order of logical trees). The third one can shed light on the quality of the possiblity to anwer a value bound.

The significant increase in type I change of item-nonresponse naturally concentrates on those with multiple head variables and/or in the downstream of long logical trees (e.g. HB3701, DHC0610 and many in section 8). The complication of logical structure in questionnaires is obviously positively related with item-nonresponse.

The type I and II changes are very close for most variables. This is equivalent to state that missing rate (a) and (d) are quite close (most exceptions concentrate on the sections 5, 7 and 8).[35] Relative to (a), both the denominator and numerator increase by the same amount: the additionally imputed cases induced by the missingness in the head variables to be finally imputed as participation. Given this evidence, we can infer that many respondents answering DK/NA in the head variables are really those on the margin of owning the content of these variables (e.g. HMR). Therefore, they are then mostly imputed as not owning them (i.e. non-participation). The potential concern of inflation of missing rate depicted as type I change can be mitigated by this fact. Adding simple logical structure in the questionnaire (e.g. asking the ownership before the values of many asset questions) does not *de facto* aggravate the information loss due to item-nonresponse.

Many noticeable drops in type III change of item-nonresponse occur among the economically critical variables and/or many suffering high missingness in (b) definition: e.g. DHI0700, DHI0600, HB0800, HB0900, DHG0800T1, DHH0905T1 and HD0801. This is quite a positive evidence for the effectiveness of offering a possibility to answer the value bounds as well as information loss due to item-nonresponse because the variables with bound information provided will be much more accurately imputed.

[35] These three sections have quite long multi-level logical trees, or say, there are multiple head variables determining the participation condition parallely. As long as one participation condition is finally imputed, the missing rate (d) and (a) can be different.

7 Convergence

We adopt two classes of convergence measurement:

The first one assesses the distributional stability across iterations. Euclidean distances of the statistics (mean, median and interquantile range) aggregating over almost each continuous variable (about 220 variables in our data) are calculated during the end of each iteration from the second on. A detailed discussion on this measure can be found in section 6.2 of Barceló (2006).

The other one is the Gelman-Rubin convergence criterion which is quite popular among Markov chain Monte Carlo (MCMC) literature. It penalizes a high variance between implicates since this means the imputed values of each implicate are not close enough. It rewards a high variance within implicates which implies imputation can cover well the domain of the joint distribution. This is a necessary condition for the convergence in the context of joint distribution as emphasized by theory for multiple imputation (Gelman et al., 1992; Christelis, 2011). Gelman (1992) suggests that 1.1 is a signal that convergence has reached. We calculated this measure over the mean of each continuous variable. For all the iterations, all the variables have this figure smaller than 1.0.

8 Evaluation

First, we should define a couple of concepts. If the response and sampling mechanism are both ignorable, the missingness in the data obtains the property missing at random (MAR (Rubin, 1987)). Otherwise, the data is missing not at random (MNAR). The sampling mechanism is ignorable if the sampling probability only depends on the observed data. This is true in most scientific surveys including PHF. As long as the sampling mechanism is ignorable, the response mechanism is also ignorable if the response probability depends on the observed data too. A special MAR is missing completely at random (MCAR) which means the response mechanism does not depend on either observed or unobserved data.

If the distributions of the missing sample and observed sample are close, we can at least assert that there is no evidence to turn down MCAR. This is a strong form of MAR. The latter is a necessary condition to allow any imputation based on the observed information to achieve valid inference. When there is discrepancy between these two distributions and it can be attributed to the observed information (i.e. any covariates in the imputation model), the MAR is still not violated.

The external diagnosis proposed in Abayomi (2008) follows this reasoning.[36] We also perform this ex post comparision of observed and imputed samples. For almost all the variables,

[36] See the end of section 5.1 for the introduction of internal and external diagnoses.

the messages we learn from does not deviate from what comes out of the ex ante evaluation of these two distributions as elaborated in the section 4.3.

Figure 3 contains this comparision for HB0900 (value of HMR) where Figure 3-A is a pdf version and Figure 3-B is a cdf version. Since we do not spot the evidence that two distributions differ systematically, no further examination is pursued. However, the imputed sample is obviously richer than the observed one for HD0801 (value of the business) as illustrated by Figure 4 which has the same setup as Figure 3 (the median of the imputed sample is 58,000 euro which contrasts with 25,000 euro in the observed sample). Our investigation reveals respondents with more children, lower stock in saving account and less value of gift/inheritance tend not to disclose the value of business but they also tend to hold higher value of business.[37]

Figure 2 The distributional comparisons of imputed and missing samples for HB0900 (current value of HMR): pdf and cdf versions (histogram, kenal density (red dashed line) and normal fitted curve (blue line) are displayed in the pdf version)

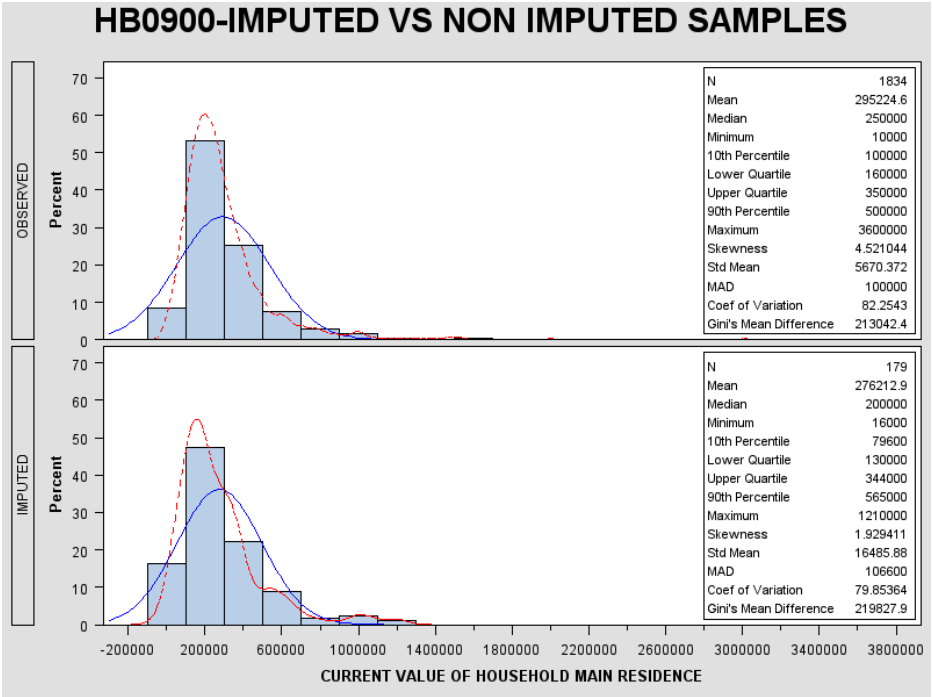

Figure 3-A PDF comparison

[37] These variables appear to be quite significant in the logit regression of the missing indicator of HD0801 but not in a regression of HD0801 itself. However, some descriptive statistics, as we explore, can also conclude this association between missing patterns and the difference in two distributions.

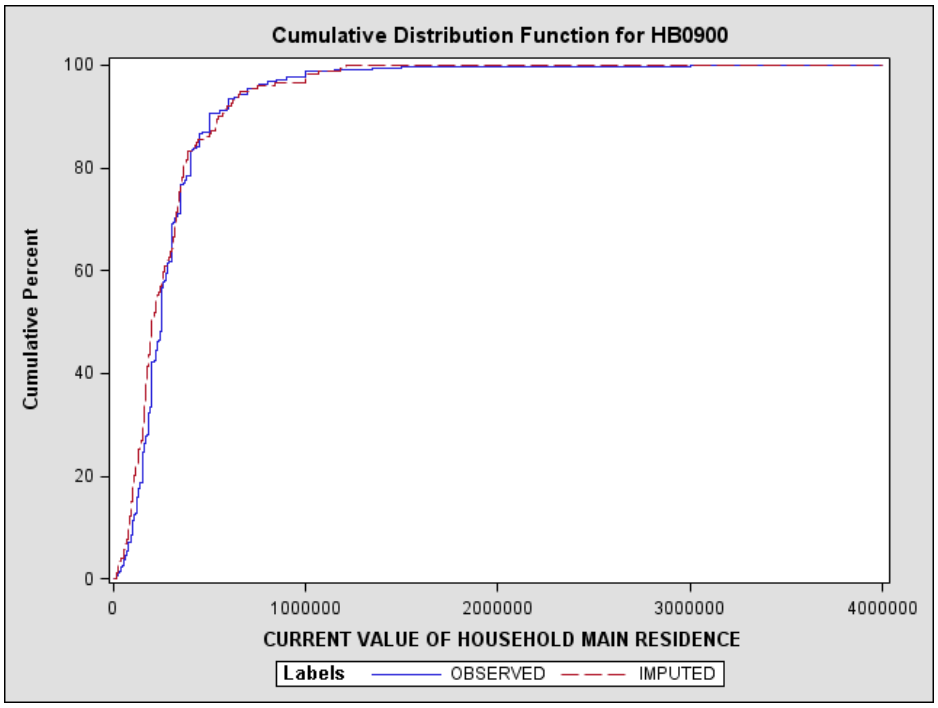

Figure 3-B CDF comparison

Figure 3 The distributional comparisons of imputed and missing samples for HD0801 (value of the business): pdf and cdf versions (histogram, kenal density (red dashed line) and normal fitted curve (blue line) are displayed in the pdf version)

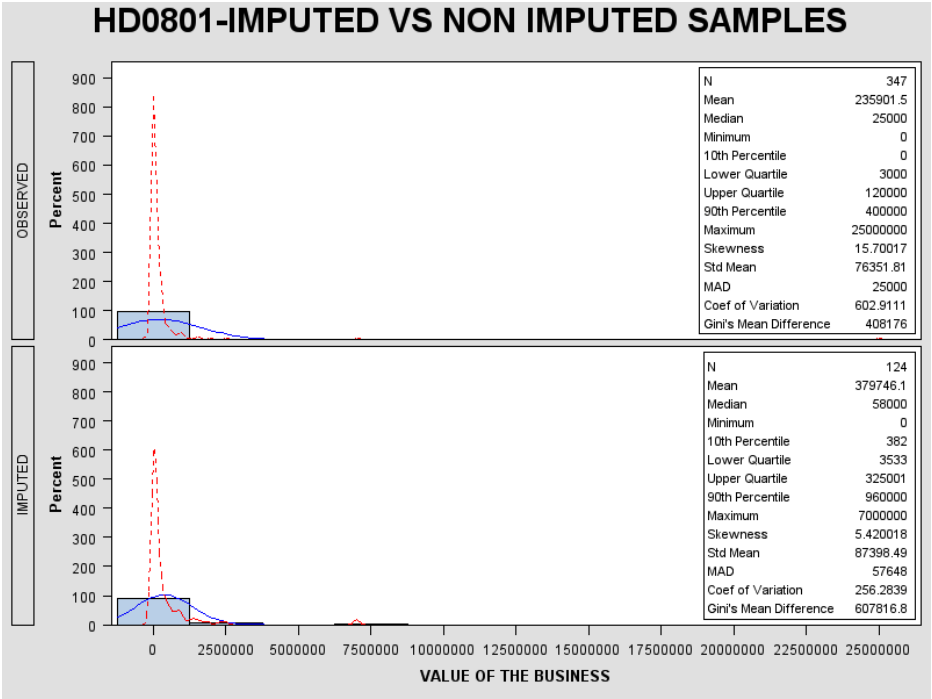

Figure 4-A PDF comparison

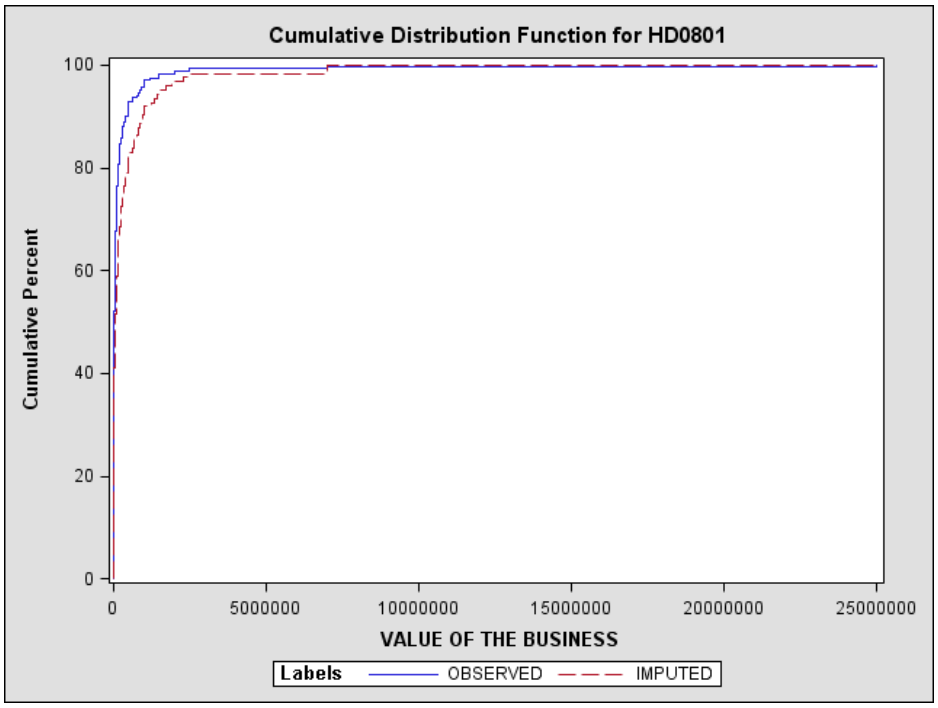

Figure 4-B CDF comparison

9 Conclusions

To disentangle the item-nonresponse in the first wave of PHF, we apply the multiple imputation proposed by Rubin (1987). FRITZ developed by Arthur Kennickell from US Federal Reserve Board is used as a robust imputation implementation. We made huge efforts to build the infrastructure ranging from a net-gross-conversion for income, adapting the data format for imputing jointly household and personal level variables, cleaning the flags etc.

On the other hand, we adopt the measures for improving the out-of-sample prediction performance to the environment of imputation: using validation and test samples in the model selection stage. Beyond the common diagnosis to compare the imputed and observed samples, a comparison of predictive distributions on the imputed and missing samples is performed and evaluated before the stage of imputation. This indeed can provide the early stage signal on the implausible values, singularity in the subsample constrained by the selected model and outliers which allows adjusting the model before the time consuming formal imputation stage.

It is prevalent in our survey that the respondents have to confront with multi-level filtering structures before they can participate to answer the value questions of many asset and liability items. This complication can increase the item-nonresponse and degree of imputation dependency between the variables with logical connection. Using the flagging information, we develop a set of measures of item-nonresponse to access these issues. We find that many respondents turn out to be on the margin of participation which they are hesitant to reveal. In-

formation loss due to the contagion of item-nonresponse between variables is not serious in our imputed data.

Moreover, the treatments for several practical problems are also discussed: semi-continuity (esp. spike in zero), jointly exclusive covariates, imputation of multiple choice variables and so on. Measuring the convergence and imputation evaluation are briefly exposed which can be enriched in the future development and research.

Table 1: Item-nonresponse rate in the first wave of PHF: representative variables (rate in terms of percentage)

Missing rate definitions:

a) Missingness by DK/NA (% of *ex ante* applicable cases)
b) Missingness by DK/NA in variable itself or any higher order head variables (% of total number of *ex ante* applicable cases and cases with DK/NA in any higher order head variables)
c) Missingness by DK/NA in variable itself or any higher order head variables excluding the cases with bounds (% of total number of *ex ante* applicable cases and cases with DK/NA in any higher order head variables)
d) Missingness by DK/NA in a variable itself or any higher order head variables excluding the cases with non-participation imputed *ex post* (% of total number of *ex ante* applicable cases and DK/NA in any higher order head variables)
e) Missingness by DK/NA (% of ex post applicable cases)

Missingness by DK/NA in variable itself or any higher order head variables excluding the cases with non-participation imputed *ex post* (% of total number of *ex ante* applicable cases and DK/NA in any higher order head variables *ex post*)

Section	**Variable**	**Label**	**Number of DK/NA**	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**
1	DPA0500	EMPLOYED	2	0	0	0	0	0
1	DPA0300	HIGHEST LEVEL OF EDUCATION COMPLETED	10	0.2	0.2	0.2	0.2	0.2
7	DPE1275	NUMBER OF CHILDREN	0	0	0.5	0	0.5	0.1
7	DPE0100A	CURRENT EMPLOYMENT STATUS - MAIN STATUS	4	0.1	0.1	0.1	0.1	1.1
7	DPE0500A	TYPE OF EMPLOYMENT RELATIONSHIP - LAST JOB	8	0.3	1.4	1.4	1.1	0.6
7	DPE0200A	TYPE OF EMPLOYMENT RELATIONSHIP - CURRENT	16	0.5	0.6	0.6	0.6	0.5
7	PE0700	TIME IN MAIN JOB	36	1.1	1.2	1.2	1.1	1.1
7	PNE2100	TIME IN LAST JOB	33	1.3	2.4	2.4	2.1	1.6
7	PE1000	TOTAL TIME IN EMPLOYMENT	77	1.3	1.8	1.8	1.6	2.1

7	PE0600	WORKING HOURS PER WEEK - MAIN JOB	66	1.9	2	2	2	2
2	HI0100	AMOUNT SPENT ON FOOD AT HOME	8	0.2	1.7	0.2	1.7	0.2
2	DHI0200	SAVING BEHAVIOUR	7	0.2	0.2	0.2	0.2	1.7
2	DHI0600	ESTIMATION OF MONTHLY HOUSE-HOLD INCOME	182	5.1	5.1	1.7	5.1	5.1
2	DHI0700	ESTIMATION OF WEALTH	394	11.1	11.1	3.6	11.1	11.1
3	HB0100	SIZE OF HOUSEHOLD MAIN RESIDENCE	18	0.5	0.5	0.5	0.5	0.5
3	DHB0300	AMOUNT OF RENT PAID FOR HOUSE-HOLD MAIN RESI-DENCE (EXCLUDING BILLS)	20	1.5	1.6	1.1	1.5	1.5
3	HB1401	INITIAL AMOUNT BORROWED	26	3.3	7.8	6.4	4.3	4.5
3	DHB0810	VALUE OF ALL CARS OWNED BY HOUSE-HOLD	119	4.2	4.6	2.2	4.6	4.6
3	HB0800	PROPERTY VALUE AT THE TIME OF ITS ACQUISITION	158	7.8	7.9	3.9	7.8	7.8
3	HB3701	AMOUNT STILL OWED	24	8.1	20.6	16.2	11.7	12.9
3	HB0900	CURRENT VALUE OF HOUSEHOLD MAIN RESIDENCE	179	8.9	8.9	3.6	8.9	8.9
9.2	HG0400	INCOME FROM FI-NANCIAL INVEST-MENT	46	1.3	1.3	1.3	1.3	1.3
9.2	DHG0200T1	TOTAL INCOME FROM REGULAR SO-CIAL TRANSFERS (YEARLY)	57	4.4	5.2	2.4	4.8	4.9
9.2	DHG0600T1	TOTAL RENTAL IN-	52	7.5	8.7	7	7.6	7.8

		COME FROM REAL ESTATE PROPERTY (YEARLY)						
9.2	DHG0800T1	AMOUNT OF INCOME FROM FINANCIAL INVESTMENT (YEARLY GROSS)	381	21.9	23.9	10.7	22.8	23.1
6	DHH0805	VALUE OF DONATIONS - AMOUNT	30	1.4	1.9	1.2	1.7	5.6
6	HH0401	VALUE OF GIFT/INHERITANCE	93	7.5	8.4	5.1	7.8	7
6	DHH0905T1	VALUE OF CHURCH TAX - AMOUNT (YEARLY)	508	25.5	26.4	12.9	25.7	24.1
9	DPG0210T1	EMPLOYEE INCOME - AMOUNT OF BONUS PAYMENTS (YEARLY GROSS)	153	5.6	6	4.3	5.4	1.7
9	DPG0200T1	AMOUNT OF EMPLOYEE INCOME (YEARLY GROSS)	218	6.8	7.2	4.7	7	7.8
9	DPG0800T1	TOTAL GROSS INCOME FROM PRIVATE PENSIONS (YEARLY GROSS)	115	23.9	22	18	16.5	25.9
4	DHC0610	AMOUNT OF POSITIVE BALANCE ON CREDIT CARD ACCOUNT	32	14.9	24.4	16.5	14.5	16.1
5	DHD0500	SAVING - SAVINGS ACCOUNT	10	0.4	0.8	0.8	0.7	0.7
5	DHD3200	VALUE OF SIGHT DEPOSITS	223	6.9	6.4	3.5	6.3	6.3
5	DHD0620T1	SAVINGS AMOUNT - HOME PURCHASE SAVINGS - AMOUNT (YEARLY)	153	11.9	12.4	10.2	12	12
5	DHD0610	POSITIVE BALANCE ON SAVINGS AND	167	13	13.5	7.9	13.1	14.5

		LOAN CONTRACT						
5	DHD2610	VALUE OF LISTED SHARES	147	23.1	17.8	13.2	13.9	13.1
5	HD0801	VALUE OF THE BUSINESS	114	24.7	26.8	16	25.7	19.5
5	DHD2520	MARKET VALUE OF GOVERNMENT BONDS	112	35.1	26.1	21.8	17.9	26
5	DHD1010T1	SAVINGS AMOUNT - CERTIFICATES - AMOUNT (YEARLY)	3	42.9	91.1	91.1	7.1	44.4
8	DPF1800ST1	CURRENT OWN CONTRIBUTIONS - RIESTER OR RÜRUP BANK SAV-INGS/LOAN CON-TRACTS - AMNT	8	19.5	97	97	1.4	32.1
8	DPF1300H	CURRENT BALANCE PENSION ACCOUNT - NON-STATE-SUBSIDISED LIFE IN-SURANCE POLICIES	452	24.6	25.9	17	24.1	24.9
8	DPF1910GT1	EMPLOYER CON-TRIBUTION - DIRECT INSURANCE - RIESTER/RÜRUP PLANS - AMOUNT (YEARLY	20	33.3	96.5	96.2	3.3	48.5

Table 2: Changes of item-nonresponse rate in the first wave of PHF: representative variables (rate in terms of percentage; (a)-(d) follows the definitions in Table 1)

Types of changes: I. (b)-(a), II. (b)-(d) and III. (b)-(c)

Section	Variable	Label	I	II	III
1	DPA0500	EMPLOYED	0.0	0.0	0.0
1	DPA0300	HIGHEST LEVEL OF EDUCATION COMPLETED	0.0	0.0	0.0
7	DPE1275	NUMBER OF CHILDREN	0.5	0.0	0.5
7	DPE0100A	CURRENT EMPLOYMENT STATUS -	0.0	0.0	0.0

		MAIN STATUS			
7	DPE0500A	TYPE OF EMPLOYMENT RELATION-SHIP - LAST JOB	1.1	0.3	0.0
7	DPE0200A	TYPE OF EMPLOYMENT RELATION-SHIP - CURRENT	0.1	0.0	0.0
7	PE0700	TIME IN MAIN JOB	0.1	0.1	0.0
7	PNE2100	TIME IN LAST JOB	1.1	0.3	0.0
7	PE1000	TOTAL TIME IN EMPLOYMENT	0.5	0.2	0.0
7	PE0600	WORKING HOURS PER WEEK - MAIN JOB	0.1	0.0	0.0
2	HI0100	AMOUNT SPENT ON FOOD AT HOME	0.0	0.0	1.5
2	DHI0200	SAVING BEHAVIOUR	0.0	0.0	0.0
2	DHI0600	ESTIMATION OF MONTHLY HOUSE-HOLD INCOME	0.0	0.0	3.4
2	DHI0700	ESTIMATION OF WEALTH	0.0	0.0	7.5
3	HB0100	SIZE OF HOUSEHOLD MAIN RESI-DENCE	0.0	0.0	0.0
3	DHB0300	AMOUNT OF RENT PAID FOR HOUSE-HOLD MAIN RESIDENCE (EXCLUDING BILLS)	0.1	0.1	0.5
3	HB1401	INITIAL AMOUNT BORROWED	4.5	3.5	1.4
3	DHB0810	VALUE OF ALL CARS OWNED BY HOUSEHOLD	0.4	0.0	2.4
3	HB0800	PROPERTY VALUE AT THE TIME OF ITS ACQUISITION	0.1	0.1	4.0
3	HB3701	AMOUNT STILL OWED	12.5	8.9	4.4
3	HB0900	CURRENT VALUE OF HOUSEHOLD MAIN RESIDENCE	0.0	0.0	5.3
9.2	HG0400	INCOME FROM FINANCIAL INVEST-MENT	0.0	0.0	0.0
9.2	DHG0200T1	TOTAL INCOME FROM REGULAR SO-CIAL TRANSFERS (YEARLY)	0.8	0.4	2.8
9.2	DHG0600T1	TOTAL RENTAL INCOME FROM REAL ESTATE PROPERTY (YEARLY)	1.2	1.1	1.7
9.2	DHG0800T1	AMOUNT OF INCOME FROM FINAN-CIAL INVESTMENT (YEARLY GROSS)	2.0	1.1	13.2
6	DHH0805	VALUE OF DONATIONS - AMOUNT	0.5	0.2	0.7
6	HH0401	VALUE OF GIFT/INHERITANCE	0.9	0.6	3.3

6	DHH0905T1	VALUE OF CHURCH TAX - AMOUNT (YEARLY)	0.9	0.7	13.5
9	DPG0210T1	EMPLOYEE INCOME - AMOUNT OF BONUS PAYMENTS (YEARLY GROSS)	0.4	0.6	1.7
9	DPG0200T1	AMOUNT OF EMPLOYEE INCOME (YEARLY GROSS)	0.4	0.2	2.5
9	DPG0800T1	TOTAL GROSS INCOME FROM PRIVATE PENSIONS (YEARLY GROSS)	-1.9	5.5	4.0
4	DHC0610	AMOUNT OF POSITIVE BALANCE ON CREDIT CARD ACCOUNT	9.5	9.9	7.9
5	DHD0500	SAVING - SAVINGS ACCOUNT	0.4	0.1	0.0
5	DHD3200	VALUE OF SIGHT DEPOSITS	-0.5	0.1	2.9
5	DHD0620T1	SAVINGS AMOUNT - HOME PURCHASE SAVINGS - AMOUNT (YEARLY)	0.5	0.4	2.2
5	DHD0610	POSITIVE BALANCE ON SAVINGS AND LOAN CONTRACT	0.5	0.4	5.6
5	DHD2610	VALUE OF LISTED SHARES	-5.3	3.9	4.6
5	HD0801	VALUE OF THE BUSINESS	2.1	1.1	10.8
5	DHD2520	MARKET VALUE OF GOVERNMENT BONDS	-9.0	8.2	4.3
5	DHD1010T1	SAVINGS AMOUNT - CERTIFICATES - AMOUNT (YEARLY)	48.2	84.0	0.0
8	DPF1800ST1	CURRENT OWN CONTRIBUTIONS - RIESTER OR RÜRUP BANK SAVINGS/LOAN CONTRACTS - AMNT	77.5	95.6	0.0
8	DPF1300H	CURRENT BALANCE PENSION ACCOUNT - NON-STATE-SUBSIDISED LIFE INSURANCE POLICIES	1.3	1.8	8.9
8	DPF1910GT1	EMPLOYER CONTRIBUTION - DIRECT INSURANCE - RIESTER/RÜRUP PLANS - AMOUNT (YEARLY	63.2	93.2	0.3

Reference

Abayomi, K., A. Gelman, et al. (2008). Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 3: 273--291.

Albacete, N. (2012). Multiple Imputation in the Austrian Household Survey on Housing Wealth. Oesterreichische Nationalbank (Austrian Central Bank) Working Paper.

Barceló, C. (2006). Imputation of the 2002 wave of the Spanish survey of household. Banco de España Occasional Papers 0603.

Barnard, J. and X. L. Meng (1999). Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. Statistical Methods in Medical Research. 8: 17-36.

Christelis, D. (2011). Imputation of Missing Data in Waves 1 and 2 of SHARE. CSEF Working Papers.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica: Journal of the Econometric Society: 829--844.

David, M., R. J. A. Little, et al. (1986). Alternative methods for CPS income imputation. Journal of the American Statistical Association: 29-41.

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

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. Statistical science. 7: 457-472.

Jaenichen, U. and J. W. Sakshaug (2012). Multiple imputation of household income in the first wave of PASS. Institute for Employment Research (IAB).

Kalton, G. and D. Kasprzyk (1985). The Treatment of Missing Survey Data. Survey Methodogogy. 12: 1-16.

Kennickell, A. B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. the Annual Meetings of the American Statistical Association.

Kennickell, A. B. (1998). Multiple imputation in the Survey of Consumer Finances. Proceedings of the Section on Business and Economic Statistics, 1998 Annual Meetings of the American Statistical Association.

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

Little, R. J. A. and D. B. Rubin (2002). Statistical analysis with missing data. New York, Wiley.

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

Raghunathan, T. E., J. M. Lepkowski, et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 27: 85-96.
Reiter, J. P., T. E. Raghunathan, et al. (2006). The importance of modeling the sampling design in multiple imputation for missing data. Survey Methodology. 32: 143.
Rubin, D. B. (1978). Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association: 20-34.
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, Wiley.
Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association. 91: 473--489.
Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. The American Statistician. 58.
Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. New York, Chapman and Hall.
Schenker, N. a. R. T. E., P. L. Chiu, et al. (2006). Multiple imputation of missing income data in the National Health Interview Survey. Journal of the American Statistical Association. 101: 924-933.
Schunk, D. (2008). A Markov Chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Advances in Statistical Analysis. 92: 101-114.
van Buuren, S., H. C. Boshuizen, et al. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. Statistics in Medicine. 18: 681-694.
White, I. R., R. Daniel, et al. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational Statistics \& Data Analysis. 54: 2267--2275.
Yucel, R. M. (2011). State of the Multiple Imputation Software. Journal of statistical software. 45.