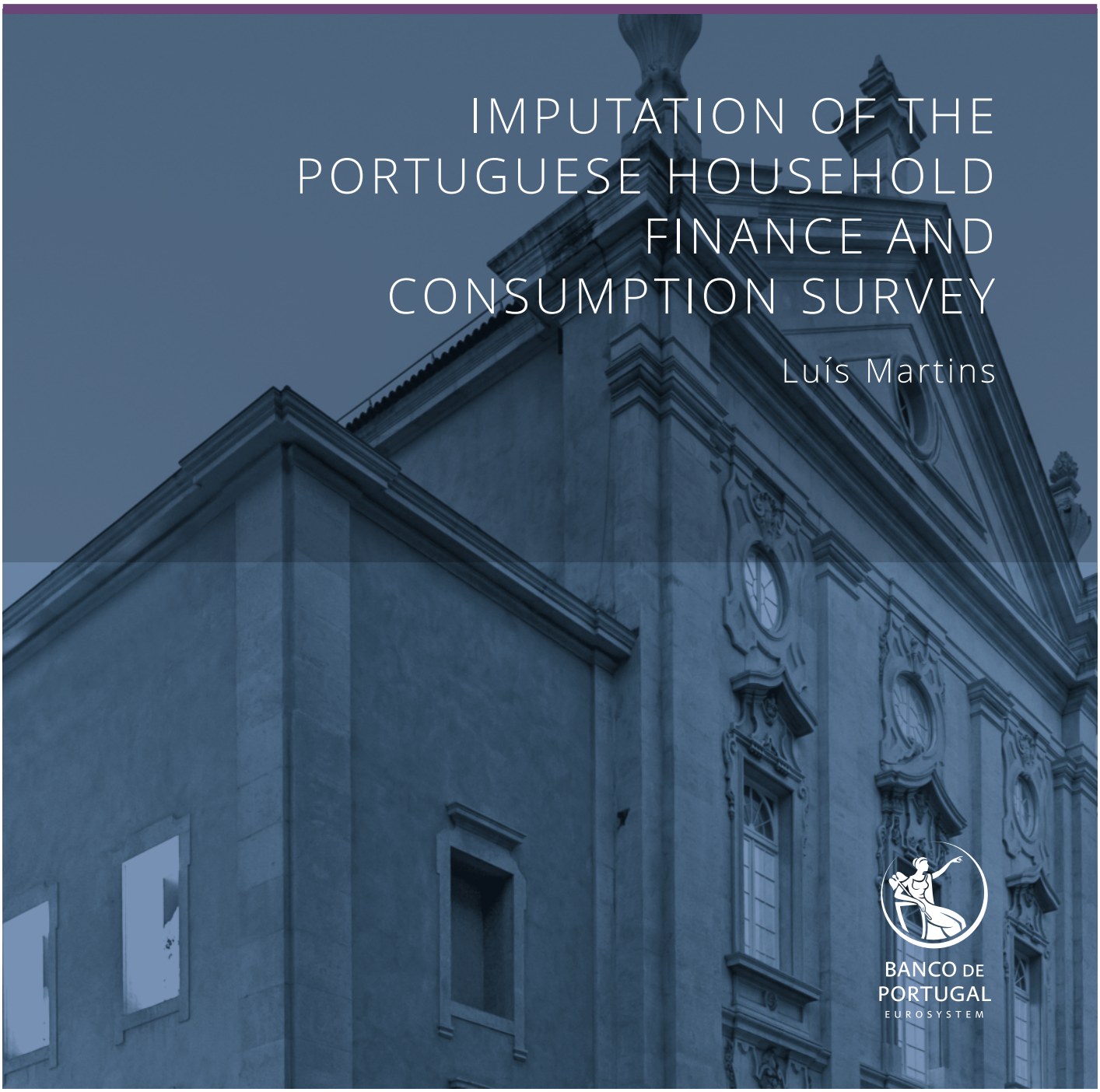
02

OCCASIONAL PAPERS 2020

IMPUTATION OF THE PORTUGUESE HOUSEHOLD FINANCE AND CONSUMPTION SURVEY

Luís Martins



BANCO DE PORTUGAL
EUROSYSTEM

02

OCCASIONAL PAPERS 2020

IMPUTATION OF THE PORTUGUESE HOUSEHOLD FINANCE AND CONSUMPTION SURVEY

Luís Martins

JANUARY 2020

Please address correspondence to
Banco de Portugal, Economics and Research Department
Av. Almirante Reis, 71, 1150-012 Lisboa, Portugal
Tel.: +351 213 130 000, email: estudos@bportugal.pt

BANCO DE PORTUGAL
EUROSYSTEM

Lisboa, 2020 • www.bportugal.pt

Occasional Papers | Lisboa 2020 • Banco de Portugal Av. Almirante Reis, 71 | 1150-012 Lisboa • www.bportugal.pt •
Edition Economics and Research Department • ISBN (online) 978-989-678-710-3 • ISSN (online) 2182-1798

Imputation of the Portuguese Household Finance and Consumption Survey

Luís Martins
Banco de Portugal

January 2020

Abstract
For the most important variables in the Portuguese Household Finance and Consumption Survey (ISFF), missing data is imputed using a stochastic multiple imputation algorithm, as agreed in the Household Finance and Consumption Network of the Eurosystem (HFCN). This paper describes the implementation of this methodology in the ISFF. The objective is to get interested readers, namely data users and other producers of survey data, acquainted with one of the most complex and time-consuming stages of the data preparation.

JEL: C15, C81, D10

Keywords: Multiple imputation, Household Finance and Consumption Survey.

Acknowledgements: I am greatly indebted to Arthur Kennickell for his invaluable advice and help in implementing the imputation of the Wave 3 of ISFF and comments for this paper. I wish to express my sincere gratitude to Sónia Costa and Luísa Farinha for their outstanding support and suggestions in the imputation and comments. I wish to show my appreciation to António Antunes for the revision of the paper. I would like to express my acknowledgement to João Lopes, who worked with the ISFF team of the Banco de Portugal in the imputation project. I would like to express my appreciation to Peter Lindner, Cristina Barceló and Junyi Zhu and all of the remaining participants of the ECB-HFCN workshop held in Frankfurt, in March 19th 2019, for sharing their imputation expertise. I would like to thank Fátima Teodoro, Pedro Próspero, Lucena Vieira, Fernando Graça and António Leite for all of the technical and programming support. All errors are my responsibility. The analyses, opinions and findings of this paper represent my views, which are not necessarily those of the Banco de Portugal or the Eurosystem.

E-mail: lpmartins@bportugal.pt

1. Introduction

This paper describes the imputation process developed for the third wave of the Portuguese Household Finance and Consumption Survey (ISFF, the Portuguese acronym for *Inquérito à Situação Financeira das Famílias*). This survey collects household-level information about real assets and their financing, other liabilities and credit constraints, private businesses, financial assets, intergenerational transfers and gifts, consumption and saving. It also gathers individual-level information on demographics, employment, pension entitlements, and income. The ISFF is part of the Household Finance and Consumption Survey (HFCS), a project developed by a network of the Eurosystem, the Household Finance and Consumption Network (HFCN), which aims at gathering harmonized micro-level information on households' finance across the euro area. The survey is carried out at country level, through national surveys, but countries are required to collect a set of harmonized variables, using common methodological principles, namely about the treatment of missing data.

One of the common difficulties among wealth surveys is non-response. When reporting complex and sensitive issues such as real and financial assets, debt, income and consumption, inevitably some respondents are reluctant, unwilling or unable to provide all of the requested information. This leads to both unit and item non-response. Unit non-response refers to cases where a household refuses to participate in the survey. Item non-response corresponds to situations where the household accepts to participate in the survey, but does not report all of the required data. In either case, non-response patterns are not likely to be random. There are characteristics of the households that affect the likelihood of not answering some questions or even the whole survey. Ignoring the presence of missing data in the analysis of survey data can lead to misleading conclusions.

Unit non-response can be addressed by oversampling households with higher non-response rates, by the replacement of non-responding households or simply by correcting the final sample weights, taking into account non-response patterns. In the ISFF, unit non-response is addressed by oversampling richer households (which usually have higher non-response rates), and correcting the final sample weights.

There is also a variety of approaches to deal with item non-response, which is the object of this paper. Sometimes interviews that have missing values in at least one variable are merely discarded. Inference based on data subject to this procedure will only be unbiased under the strong assumption of Missing Completely at Random (MCAR), i.e., if missing observations are independent of both observable and unobservable characteristics of households.

An alternative solution is to impute the missing values, which consists on assigning a value to each observation that is missing. In the ISFF, as agreed in the HFCN, item-non response is treated using stochastic multiple imputation techniques. These techniques were developed by Rubin (1987) and are applied, for example, in the Survey of Consumer Finances (SCF) of the Federal Reserve Board. This method takes into account observable determinants of non-response and

thus only needs to assume that non-response is independent of the unobservable household characteristics, i.e., it assumes the data is Missing at Random (MAR).

As referred in Rubin (1996), the main objective of this method is to provide the tools for valid statistical inference, as opposed to optimal point prediction. This means that the main objective of stochastic multiple imputation is not to replace missing data by values that best fit the variables of interest, but to preserve the characteristics of their distribution and the relationships between different variables. In fact, as opposed to stochastic imputation methods, the deterministic methods (for instance, methods that replace the missing values with means or medians or with other predictions obtained from a simple regression), do not preserve the characteristics of the joint distributions of variables, undermining the validity of statistical inference. Additionally, multiple imputation (i.e., the existence of several imputed values for each missing observation) allows to take into account the imputation uncertainty in statistical inference. As explained in Barceló (2006) and Barceló (2008), multiple stochastic imputation has the advantage, compared with single stochastic imputation, of taking into account not only the within-imputation variance of the statistics, computed using a single imputed data set, but also the between-imputation variance due to the uncertainty about the imputed values. Thus, it avoids situations of erroneous statistically significant results.

The aim of this paper is to deliver a detailed description of the procedures entailed to impute the third wave of the ISFF, while attempting to provide approachable examples of the different steps undertaken.

Section 2 discusses the general features of multiple imputation. Section 3 describes characteristics of the ISFF that are relevant for imputation. Section 4 analyses a set of indicators of non-response. Section 5 gets the reader acquainted with the Federal Reserve Imputation Technique Zeta (FRITZ) software package, which is used to implement the multiple imputation techniques in the ISFF. Section 6 illustrates the imputation algorithm used. Section 7 reviews the data preparation procedures. Section 8 describes the data structure and presents some examples to motivate the operational decisions concerning imputation. Section 9 shows the criteria for covariate selection and imputation order. Section 10 describes the evaluation of the imputation models and results. Section 11 concludes.

2. General features of imputation in the ISFF

Imputation consists on assigning a value to an observation that is missing in the dataset. Ideally, all survey variables should be imputed. Nevertheless, given time and computational restrictions which exist at country level, the HFCN agreed on a minimum set of variables that should be imputed by all countries. In the ISFF, the set of imputed variables includes not only the agreed minimum set, but also almost all the remaining monetary variables and other variables that were considered to be important for the imputation process. In the third wave of the ISFF, 317 variables, out of 545 variables with missings in total, were imputed, of which 139 correspond

to monetary variables, 75 to binary variables, 21 to categorical variables and 82 to other type of numerical variables.

In general terms, the stochastic multiple imputation technique used in the ISFF has the following characteristics:

- The process involves the estimation of models for the variables with missing observations, using as much information as possible. This means that the methodology follows the broad conditioning approach, in which ideally one should include all the variables in the dataset as covariates in the estimation models;
- The process is sequential, which means that imputed values for one variable are used in the imputation of the following ones;
- The imputation process is iterative, so that the whole imputation sequence is repeated, based on the values obtained from the previous iteration, as many times as necessary until the process converges;
- The process is stochastic, since there is a randomization process applied to the point estimates obtained from the imputation models;
- The imputation is multiple, which means for every missing observation there will be five imputed values (five implicates) as established in the HFCN[1];

This technique, which was proposed by Rubin (1987), draws its theoretical framework from the Expectation Maximization algorithm and Gibbs sampling.[2]

Proposed by Dempster *et al.* (1977), Expectation Maximization is defined as an iterative method to find maximum likelihood estimates of the model parameters in the presence of missing data. Using observed data, Expectation Maximization computes starting values of the model parameters to simulate the distribution of the missing values. Then, it uses the collected information along with the previously simulated values to adjust parameter estimations. This deterministic process proceeds iteratively until the estimates are close to a fixed point. Gibbs sampling, or stochastic relaxation, described in Geman and Geman (1984), can be used to apply the Expectation Maximization intuition to complex data structures. It is a Markov Chain Monte Carlo algorithm that simulates the distribution of variables, conditional on observed data and simulated distributions of preceding variables within the same iteration. Geman and Geman (1984) show that the process converges under regularity conditions and the simulated distribution of missing data iteratively draws closer to the true latent distribution.

Multiple Imputation, described in Rubin (1987), aims to reflect the uncertainty about the true imputation and non-response models in the statistical inference. This is accomplished through an additional source of variability, in the form of

1. According to Rubin (1976), this number of implicates is a reasonable compromise between potential estimation efficiency gains from having an additional implicate and the corresponding computational burden

2. See Kennickell (1991), Kennickell (1998) and Barceló (2006) for a more detailed review of imputation theory.

multiple stochastically imputed values (implicates) for each missing observation in a dataset. These values are computed by adding a stochastic shock to the point estimates obtained from the imputation model. Multiple imputation provides an honest picture of the limited knowledge about missing data and avoids situations of erroneous statistically significant results. [3]

3. Survey characteristics relevant for the imputation process

This section describes the technical characteristics of the ISFF that have implications for the imputation process.

Variable types

According to their type of information, the survey variables can be classified in four groups:

- Household-level (H) variables: provide information about the household as a whole, containing mostly data related with assets, liabilities, consumption and intergenerational transfers and gifts;
- Person-level (P) variables: yield information for each individual with at least sixteen years old, referring to employment, income and pension rights;
- Demographic-related (R) variables: information for every individual related with demographic aspects (e.g., age, gender, marital status);
- Sample (S) variables: include information about the sample and contacts made by the interviewers (e.g., number of contacts, dwelling rating, dwelling outward appearance);

In the imputation process, not only the collected survey variables are used, but also some derived (D) variables, which are calculated from the collected variables.

Relative to their format, variables can either be:

- Numerical: monetary values, years, interest rates, number of units;
- Binary choice: yes/no;
- Categorical: e.g. education level, employment status, loan purpose;

The existence of all these different types of variables has implications for the imputation process. For example, the existence of variables at the household and person-level implies the need to use a special data structure, as will be described in Section 8. Different data formats imply the need to use different imputation models, as will be described in Section 5, and different variable transformations, as described in Section 7.

3. See Zhu and Eisele (2013) for analysis of Multiple Imputation vs. other imputation methods.

Logical tree: parent-child variable relationship

The design of HFCS has a built-in hierarchical and logical relationship between variables, which determines the questions that are made during the interview. Variables standing in the beginning of each logical tree, called *parent variables* must be replied by all households, regardless of the answers to previous questions. In the branches of the logical trees there are *child variables,* whose response is conditional on the value of one or more *parent variables*. Some *child variables* may also have their own branches in a multi-layered flow. The imputation routine must account for this highly stratified relationship between variables, by defining an imputation sequence were *parent variables* are always imputed before the corresponding *child variables*.

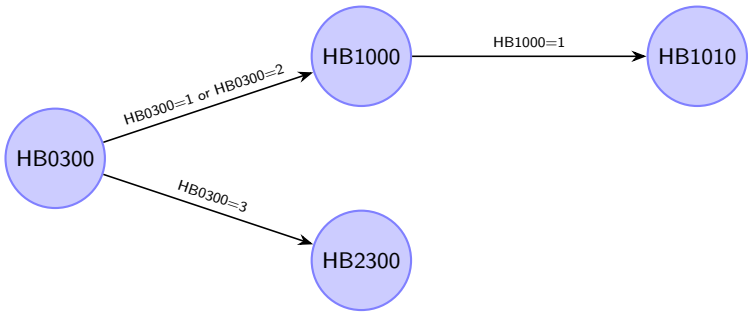


Figure 1: Logical tree example

In the simplified example of Figure 1, the branch starts with HB0300, which is a categorical variable about Household Main Residence (HMR) tenure status. If the household's response equals 1, *"own all"*, or 2, *"own part"* the interview proceeds to HB1000, a binary variable about the existence of loans using HMR as collateral. If the household has any of those loans (HB1000=1), the next question (HB1010) is about how many are there and then the questionnaire proceeds on that branch, collecting further loan information.

However, if the HMR is rented (HB0300=3), the questions of the HB1000 branch will be skipped, because they are not applicable to that household. In this case, the household will be asked about the monthly amount paid as rent (HB2300).

Suppose HB1000 is applicable, but the household does not respond, either because it is unwilling to or simply does not know the answer. If so, HB1000 will be missing and its *child variables* (such as HB1010) will be also missing due to higher order missing.

For the imputation routine, it is very important to distinguish the non-applicable cases from the missing values, since only the latter have to be imputed. Additionally, it is important to distinguish the first order missings from the missings due to higher order missing. One of the roles of the flag variables presented later on in this section is to make this distinction easier.

Variable bounds

The variables in the ISFF might have three different types of bounds, which have to be taken into account simultaneously in the imputation process, in order to ensure the imputed values comply with them. First, the survey variables have upper and lower absolute values bounds for acceptable responses (e.g., age must not be lower than zero, the percentage of business ownership must not be greater than one hundred percent).

Second, monetary variables may also have bounds reported during the interview. When asking questions about monetary items, the respondent often does not know the exact answer or does not want to provide a point value. In those cases there is the possibility of answering with a range of values.

Finally, in ISFF, survey responses must also comply with critical validation rules, which are variable relationships that must hold, in order to ensure their consistency (e.g., the number of years an individual has worked cannot be greater than its age; the household cannot have become the owner of its main residence before the oldest person in the household was born). The enforcing of these rules is done through dynamic bounds, which are bounds that depend on the values of other variables.

Flag variables

One of the most important pieces of the imputation puzzle is the information about the origin of the variables' content, which is provided by flag (F) variables. Each collected survey variable has its corresponding F variable. There are many different flag values in the ISFF.[4] Nevertheless, as far as imputation is concerned, flags can be classified in three categories:

- 0: inapplicable (filtered) cases;
- $1000 \leq$ flag value < 2000: missing values;
- other flag values: non-missing values;

The flags of the missing values identify observations to possibly be imputed and correspond to the following cases:

- 1050: don't know;
- 1051: no answer;
- 1052: higher order missing (missing due to don't know or no answer in a *parent variable*);
- 1053: value collected in range;
- 1054: value deleted (considered incorrect or unreliable);
- 1057: value not collected due to a software or interviewer error or to editing of *parent variable*.

4. More information about flags can be consulted here.

The different flags values for the missing cases have different implications for the imputation process. Flag values 1050, 1051 and 1053 have always to be imputed.

Observations with flag value 1052 are imputed conditional on the imputation outcome of the corresponding *parent variables*. Taking the example in Figure 1, if HB1000 is imputed with a value of 1, HB1010 will be imputed as well. If HB1000 is imputed with value 2, HB1010 will become inapplicable.

Finally, flag values 1054 and 1057 can both refer to first order missing or to higher order missing. This happens because, once one observation is missing and has a flag value 1054 or 1057, all the *child variables* also become missing with the same flag.

Table 1 shows the flag value proportions calculated as a percentage of the total number of observations to be imputed. Most of the observations to be imputed are associated with monetary variables reported in ranges (flag value 1053), either for H or P variables. The cases where the respondent does not know or is unwilling to provide an answer (flag values 1050 and 1051, respectively) account for almost one third of the total missing values. The proportion of missings due to higher order missings (flag value 1052) is lower for H variables. The sum of flag values 1054 and 1057 corresponds to less than two percent of the total number of observations to be imputed.

	1050	**1051**	**1052**	**1053**	**1054**	**1057**	**Total**
All variables	27.26	4.16	8.00	58.82	0.53	1.23	100
H	28.72	2.84	4.62	62.05	0.44	1.33	100
P	22.57	8.36	18.78	48.56	0.83	0.90	100

Table 1. Flag values in percentage of the total number of observations to be imputed

4. Item non-response in ISFF

The amount of missing information in each variable of the database has a great influence in the imputation process, since it impinges on the specification of the imputation models and also on the imputation sequence.

One of the difficulties of calculating item non-response rates is related to the treatment of higher order missing cases. Their status is undetermined, meaning that before the imputation of the *parent variables* it is not possible to know whether these cases are not applicable or missing.

Additionally, when calculating non-response rates, it is also important to take into account that missing cases where the value was reported in range differ fundamentally from other types of missings. The narrower the interval reported by the household, the more precise and accurate the imputation will be.

Table 2 illustrates item non-response rates (proportion of missing over applicable cases) for selected aggregates, in order to provide a broad picture of

missing data in ISFF. The bottom of the table includes non-response rates for some variables to illustrate the variety of cases that had to be addressed during the imputation process.[5] Non-response rates for a set of variables are the sum of the number of missings, over the sum of the number of total applicable cases in each variable. See the non-response rate for n H variables, nr_h:

$$nr_h = \frac{\sum_{i=1}^{n} \#miss_{hi}}{\sum_{i=1}^{n} \#cases_{hi}} = \frac{\#miss_h}{\#cases_h} \tag{1}$$

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
All variables							
H	1.6	3.8		1.3	4.1	3.9	
P	2.5	4.4		0.9	3.0	2.6	
R	0.0	0.0		0.0	0.0	0.0	
Binary variables							
H	0.1	0.6		0.1	0.2	0.1	
P	0.4	0.8		0.4	0.6	0.5	
Monetary variables							
H	5.9	21.8	69.2	5.9	21.8	21.4	69.2
P	5.7	22.7	58.5	5.6	22.0	20.1	61.5
Includes non-imputed?	Y	Y	Y	N	N	N	N
PE0100A: Main labour status				0.2	0.2	0.2	na
HC0901: Non-collateralised loan i-rate				46.6	47.2	47.1	na
PG0110: Gross annual employee income				3.5	16.7	16.6	76.9
PG0410: Gross annual income from OPPP				5.7	41.0	18.6	56.3
HB0900: Current price of HMR				7.4	26.6	26.6	71.9
HD0801: Business value				22.3	47.0	47.0	48.1
HB2100: Money owned other HMR loans				0.0	52.9	0.0	0.0

(1), (4): $\frac{\#F[1050]+\#F[1051]}{\#F[\neq 0]}$;

(2), (5): $\frac{\#F[1000,2000[}{\#F[\neq 0]}$;

(3), (7): $\frac{\#F[1053]}{\#F[1000,2000[}$;

(6): $\frac{\#F[1000,2000[}{\#F[\neq 0]}$, excluding the *ex post* non-applicable cases;

Table 2. Item non-response rates (%)

Column (1) shows the non-response rate including as missings only the cases where the respondent does not know or is unwilling to provide an answer, which correspond to flag values of 1050 and 1051, respectively. Column (2) treats every observation with flag value between 1000 and 2000 (including values collected in

5. S variables are not imputed and so they are excluded from the calculation of non-response rates.

ranges) as missing. Thus, (1) and (2) are respectively the lower and upper bounds of the many different non-response rates that may be computed. Column (3) shows the proportion of missing values identified in (2) that are actually values collected in ranges. Columns (4) and (5) display the same indicators as (1) and (2), respectively, excluding all the variables that were not imputed but were eligible for imputation (had some form of missing or range-reported value).[6]

Column (6) was inspired in Zhu and Eisele (2013) and presents non-response rates based on the imputed data. Only after imputation it is possible to know if the higher order missing observations were actually missing, or non-applicable. This will influence the non-response rates, since the more higher order missing cases are deemed as non-applicable, the lower the non-response rate will be, because those will be excluded from the missing counting. Therefore, while (4) and (5) provide the *ex ante* lower and upper bounds for non-response, (6) shows its *ex post* materialization.

Considering all survey variables, (1) and (2) show that H variables have lower missing rates, on average, compared to P variables. R variables have no missing information. Once non-imputed variables are excluded, the previously established relationship between H and P variables is reversed, as seen in (4) and (5). Column (6) shows that *ex post* non-response rates are close to their upper bound, meaning that most of the higher order missing cases were considered to be applicable in the imputation process.

As expected, binary variables have the lowest non-response rates among the survey variable types. Respondents have no trouble and are willing to answer most of the yes/no questions. Accounting for the filtering due to higher order missing in this type of variables does not have a significant impact in the non-response rates, since most logical trees start with a binary variable and those are seldom found in deep branches. The latter also impinges on the decision to impute almost every binary variable, which explains the similar results between columns (1) and (4).

A great number of respondents has trouble in providing monetary amounts, either because they do not know or are unwilling to. Not surprisingly, monetary variables have higher-than-average non-response rates, as reported in columns (1) and (2). However, if values collected in ranges are treated as non-missing, the upper bounds for non-response in column (2) decline to 6.7 and 9.4 for H and P variables respectively.

Most of the monetary variables were imputed, which explains the similar results for monetary variables in columns (1) and (4) and also in (2) and (5).

The bottom of the table shows variables that were imputed, so data is presented in columns (4) to (7).

The first variable, PE0100A (main labour status), illustrates a relatively simple situation, as far as imputation is concerned: it stands on top of the logical tree,

6. Correspondence for column (3) was not presented, because all the monetary variables were imputed, except for a small number of national P variables.

meaning that it is unaffected by the survey filtering structure and only a residual amount of respondents (0.2%) does not provide a valid answer. This should not pose an imputation challenge. On the other hand, some variables such as HC0901 (non-collateralized loan interest rate) have a high non-response rate, as reported in its *ex ante* lower and upper bounds in columns (4) and (5), respectively. Questions about loans usually require the respondent to browse documentation and a fair share of people has trouble answering some loan details.

Turning to monetary variables, PG0110 (gross annual employee income) has a low non-response rate. Notice that the difference between 3.5% in the lower bound and 16.7% in the upper bound is mainly driven by interval values. The upper bound is actually 4%, not accounting for those values. However, some income variables such as PG0410 (gross annual income from occupational and private pension plans) have higher non-response rates and higher order missing cases, on top of a low number of applicable cases.[7]

Despite being one of the main assets of most households, a non-negligible proportion of respondents does not know how (or is unwilling) to price their main residence (HB0900). The minimum of the non-response rates is 7.4%, in column (4), and it has a high proportion of interval values. HD0801 (value of the main business) illustrates monetary variables with high *ex ante* non-response rates, which builds on the uncertainty about the imputation outcome.

The last case to be illustrated is related to variables that stand on the bottom of long logical trees and are applicable to a very low number of households. This is the case of HB2100 (money still owned on additional HMR loans), which applies to households with more than three loans using the main residence as collateral. The complexity of the filtering structure becomes apparent on the wide non-response range (between 0% and 52.9%), which in this case turned out to be the lower bound, given that all higher order missing values became inapplicable after imputation.

5. FRITZ outline

In waves one and two of the ISFF, imputation closely followed the program €mir, which was developed by the ECB team for the HFCN and provides a baseline for countries to impute their national surveys. The core of €mir is based on FRITZ (Federal Reserve Imputation Technique Zeta), the SAS program developed by Arthur Kennickell to impute the Survey of Consumer Finances (SCF) in the US. FRITZ was designed to impute the SCF using multiple imputation techniques and Gibbs sampling methods, as described in Kennickell (1991) and Kennickell (1998).

In wave three of ISFF, €mir was used as a starting point to create a more flexible program. The new program still relies on FRITZ as the core of the multiple

7. PG0110 has close to 6000 applicable cases, while PG0410 has around 100.

imputation routines, but enables the usage of the implemented algorithm and data structure, which will be described in Sections 6 and 8, respectively. Arthur Kennickell was also instrumental for the design of the current ISFF imputation program and the conceptualization of the data structure.

Initially, FRITZ restricts the dataset to cases where the variable to be imputed is applicable and to the variables chosen to be used as covariates. Then, it uses the restricted dataset to compute a Sum of Squares and Cross Products (SSCP) matrix or Conditional Frequency (CF) table, which contains all the information required to estimate the parameters of the imputation models. Next, the program identifies the observations that are going to be imputed, using the logical filter of the variable to be imputed and its flag values. Notice that both the filter and the flags must be used, since the flags alone do not suffice in the identification of missing values in the higher order missing cases. Then, the program proceeds sequentially for each identified observation, estimating the imputed values using the observation-specific parameters computed from the SCCP (or CF) and applying a randomization process.

Finally, the program updates the dataset with the imputed values and moves to the next variable, repeating the whole process until all selected variables have been imputed.

FRITZ provides three imputation models: continuous, binary and frequency, which are suited for different variable types. Binary and frequency models are used to impute binary response and categorical variables, respectively. All of the remaining variables are imputed using the continuous model. Zhu and Eisele (2013) provide a narrative description of FRITZ models, while Barceló (2006) takes a formal approach. The remaining of this section illustrates the different types of models by providing examples of the processes described above.

Continuous

For continuous variables, FRITZ relies on the linear regression model to impute the missing observations. Suppose an imputation model is defined for a continuous variable y using x_1 and x_2 as covariates. The linear regression model for y would be described by equation (2):

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{2}$$

The program computes the 3x3 normalized SSCP matrix, using the subset of observations where y is applicable:

$$SSCP = \begin{bmatrix} VAR(y) + \bar{y}^2 & COV(y, x_1) + \bar{y}\bar{x}_1 & COV(y, x_2) + \bar{y}\bar{x}_2 \\ COV(x_1, y) + \bar{x}_1\bar{y} & VAR(x_1) + \bar{x}_1^2 & COV(x_1, x_2) + \bar{x}_1\bar{x}_2 \\ COV(x_2, y) + \bar{x}_2\bar{y} & COV(x_2, x_1) + \bar{x}_2\bar{x}_1 & VAR(x_2) + \bar{x}_2^2 \end{bmatrix} \tag{3}$$

Variances and covariances in SSCP cells are computed using SAS PROC CORR default options, which proceeds with pairwise deletion when observations contain missing values. PROC CORR includes all non-missing pairs of values for each pair of variables in the statistical computations. Therefore, variances and covariances reported may be based on different numbers of observations.

On the one hand, this procedure retains as much data as possible for estimation of the linear regression parameters. Listwise deletion would severely reduce the sample size, particularly in the first iteration, where covariates have the most missing values. On the other hand, pairwise-calculated SSCP is not guaranteed to be positive definite, which may result in pathological situations, such as a negative variance of the estimation residuals $\hat{\sigma}^2$.

The linear regression parameters are computed using subsets of SSCP, according to the non-missing covariates of each case to be imputed. This implies possible different parameter estimations for each missing observation (i.e. for different households).

Suppose a household i where x_1 and x_2 are non-missing. The vector of parameter estimates $\hat{\beta}_i$ will be given by:

$$\hat{\beta}_i = \Sigma(XX)_i^{-1}\Sigma(XY)_i \tag{4}$$

$\Sigma(XX)_i$ is a 2x2 matrix:

$$\Sigma(XX)_i = \begin{bmatrix} VAR(x_1) + \bar{x}_1^2 & COV(x_1, x_2) + \bar{x}_1\bar{x}_2 \\ COV(x_2, x_1) + \bar{x}_2\bar{x}_1 & VAR(x_2) + \bar{x}_2^2 \end{bmatrix} \tag{5}$$

$\Sigma(XY)_i$ is a 2x1 vector:

$$\Sigma(XY)_i = \begin{bmatrix} COV(x_1, y) + \bar{x}_1\bar{y} \\ COV(x_2, y) + \bar{x}_2\bar{y} \end{bmatrix} \tag{6}$$

This results in a $\hat{\beta}_i$ 2x1 vector:

$$\hat{\beta}_i = \begin{bmatrix} \hat{\beta}_{1i} \\ \hat{\beta}_{2i} \end{bmatrix} \tag{7}$$

The linear projection of y for household i is obtained using $\hat{\beta}_i$ and the values of x_1 and x_2:

$$\hat{y}_i = X_i\hat{\beta}_i = \hat{\beta}_{1i}x_{1i} + \hat{\beta}_{2i}x_{2i} \tag{8}$$

The stochastic process consists of adding a random noise term to the point estimate $\hat{y}_i$. First, FRITZ computes the observation-specific variance of the linear regression errors, $\hat{\sigma_i}^2$:

$$\begin{aligned} \hat{\sigma_i}^2 &= \Sigma(Y)_i'\Sigma(Y)_i - \Sigma(XY)_i'\Sigma(XX)_i^{-1}\Sigma(XY)_i, \\ \Sigma(Y)_i &= \begin{bmatrix} VAR(y) + \bar{y}^2 \\ COV(x_1, y) + \bar{x_1}\bar{y} \\ COV(x_2, y) + \bar{x_2}\bar{y} \end{bmatrix} \end{aligned} \tag{9}$$

The imputed value $\hat{imp}_i$ is given by:

$$\hat{imp}_i = \hat{y}_i + draw_i, \quad draw \sim N(0, \hat{\sigma}_i) \tag{10}$$

The random value $draw_i$ is picked from a normal distribution with mean zero and $\hat{\sigma}_i$ standard deviation, in order to account for the uncertainty associated with the estimation of imputed values. Lower regression standard errors contribute to a narrower range of values between the different implicates. First, the program searches the $draw$ distribution, within a $-1.29\hat{\sigma}_i < draw < 1.29\hat{\sigma}_i$ (80%) range, for a value that is compatible with all of the bound restrictions (absolute, reported and dynamic bounds). If it is unable to find such a value after 100 draws, the program widens the search to the $-1.96\hat{\sigma}_i < draw < 1.96\hat{\sigma}_i$ (90%) range. The value is forced to the nearest bound if the program cannot hit an admissible value after 100 more draws.

Notice that if x_2 was missing for another household k, different subsets of SSCP would have been used to compute the model parameters (a single beta, in this case):

$$\begin{aligned}
\Sigma(XX)_k &= \left[VAR(x_1) + \bar{x}_1^2\right], \\
\Sigma(XY)_k &= \left[COV(x_1, y) + \bar{x}_1\bar{y}\right], \\
\Sigma(Y)_k &= \begin{bmatrix} VAR(y) + \bar{y}^2 \\ COV(x_1, y) + \bar{x}_1\bar{y} \end{bmatrix}, \\
\hat{\beta}_k &= \left[\hat{\beta}_{1k}\right] \\
\hat{y}_k &= \hat{\beta}_{1k} x_{1k}
\end{aligned} \tag{11}$$

Finally, it is important to notice that the process described from equations (2) to (10) is performed independently for each implicate. This means that the program always computes a linear projection before adding the random noise term, which differs from estimating $\hat{y}_i$ beforehand and then obtaining the between implicate variability by adding different random draws. In the current setup, $\hat{y}_i$ can be different across implicates, if the covariates have been previously imputed. This ensures the information coherence for each variable in each implicate.

In the higher order missing cases, there can be situations where the *child variable* is deemed as applicable in a given implicate and non-applicable for another implicate. Thus, performing the imputation process independently for each implicate ensures the consistency between the different variables. This makes the process more burdensome, also.

Binary

The imputation of binary variables relies on a linear probability model, which is relatively similar to the continuous model. The imputed value is obtained by comparing draws from a uniform distribution with the predicted probability $\hat{y}_i$. Adapting the example from the continuous case and taking the upper and lower

predicted probability limits as 1 and 0, respectively, the imputed value $\hat{imp}_i$ will be given by:

$$\hat{imp}_i = \begin{cases} 1, & \text{if } draw_i < \hat{y}_i \\ 0, & \text{if } draw_i \geq \hat{y}_i \end{cases} \quad , \quad draw \sim U(0,1) \tag{12}$$

If the predicted probability is very close to extreme values, (lower than 5% or greater than 95%) FRITZ imposes that value and ignores the randomization, which prevents uncanny random draws.

Frequency

The process for imputing categorical variables is similar to a hot-deck procedure. It can be described as randomized imputation from a conditional frequency table. For each categorical variable to be imputed, the program restricts the data to a frequency table based on the dependent variable filter and the covariates chosen. Only two covariates can be specified in the frequency imputation model.

The program computes a conditional frequency table of the variable to be imputed based on the values of the specified covariates. Next, it extracts from that table the observations with the same covariate values as the observation being imputed. The imputed value is obtained from a random process applied to the cumulative frequency distribution of those observations.

If the number of cases that match the covariate values of the observation being imputed is smaller than a user-specified threshold, the program will proceed in one of two ways, depending on the user specification.

If the user chooses not to collapse values, the program will consider only the first covariate to look for a sufficient number of cases. If there are still not enough observations, FRITZ will account only for the second variable. If it still does not find enough cases, it will compute the unconditional frequency distribution of the missing variable, meaning that the imputation will be independent of the covariates specified.

If the user chooses to collapse values, the program will try to collapse adjacent cells of the second classifying variable and then proceed as in the no-collapsing case. Therefore, the specification of the covariate order is relevant in the collapsing case. The second covariate should have an ordinal meaning (e.g., age satisfies this criterion, while the household's main residence tenure status does not).

In the randomization process, the program computes the cumulative frequency distribution of observations with the same (or collapsed) covariate values as the observation being imputed and then draws a value from a uniform distribution to determine the position of the chosen value in the frequency distribution.

Suppose, for example, an imputation model for the level of education (PA0200), using the main labour status (PE0100A) and age (RA0300) as the first and second covariates, respectively. Consider that the program is imputing the implicate i of PA0200 for a retired individual (PE0100A=5) with 56 years of age (RA0300=56).

Table 3 shows the relevant part of the conditional frequency table computed for PA0200:

PA0200	PE0100A	RA0300	Frequency
3	7	48	40
5	3	48	30
6	5	48	55
6	5	48	30
1	5	55	30
3	5	55	55
4	5	55	20
2	5	56	25
2	5	68	5
3	5	68	35
1	5	85	5

Table 3. Conditional frequency table for PA0200

First, the program scopes the conditional frequency table for the peers of this particular individual. There are 25 observations (row highlighted in green) that match exactly its characteristics. If the minimum cell size required to compute the conditional frequency distribution of PA0200 was lower than 25, the program would proceed to the computation of the cumulative frequency distribution. Alternatively, let us suppose the minimum cell size is 30 and the collapsing option is activated.

Since the cell size is below the required threshold, the program will collapse the first adjacent "categories" of RA0300, which corresponds to all rows where RA0300 equals 55 or 68 (highlighted in blue). In this case, the cell will have 170 observations.[8]

The program proceeds by computing the cumulative frequency distribution of PA0200, based on the cell observations, as illustrated in Table 4.

PA0200	Frequency	Relative Frequency	Cumulative Rel. Frequency
1	30	0.18	0.18
2	30	0.18	0.35
3	90	0.53	0.88
4	20	0.12	1.00

Table 4. Cumulative frequency distribution of PA0200

For the randomization process, FRITZ draws a random number from a uniform $U(0,1)$ distribution and compares it with the cumulative relative frequency of

8. Notice that if the collapsing option was not activated, FRITZ would condition only on the first variable, which would be equivalent to collapse all values of RA0300 where PE0100A equals 5, leading to a cell size of 260 observations.

PA0200. The imputed value corresponds to the first row where the cumulative relative frequency is higher than the random draw. For example, if the random draw was 0.05, 0.76, 0.14 or 0.29, the imputed value for PA0200 in implicate i would be 1, 3, 1 or 2, respectively. Notice that the whole frequency imputation process must be executed for each implicate, in order to ensure variable consistency, as explained for continuous models.

6. Imputation algorithm

The examples illustrated in the previous section overlooked the fact that multiple imputation is an iterative and sequential procedure. Each iteration is a whole new imputation round, that builds from the results obtained from the previous iteration, in order to impute the five implicates of each missing observation all over again. This process should be repeated as many times as necessary until convergence is achieved.[9]

Figure 2 illustrates the imputation process for the first two iterations. Interpretation of further iterations is straightforward, as the process is similar from the second iteration henceforth.

Output_0 is the ISFF pre-imputation dataset.[10] It is used as input (Input_1) in the first iteration. The first iteration differs systematically from the remaining, since its purpose is to obtain initial values for the missing data. In the beginning of the first iteration all values to be imputed will be empty, except for the range-reported values which contain the midpoints of the respective intervals.

Input_1 is the source for the calculation of the SCCP matrices (or CF tables, for the frequency models). It is also being sequentially updated with the results of the imputation process. Recalling equation (2) this means that, in the first iteration, the estimations of β_1 and β_2 and the covariates x_1 and x_2 are based on the same dataset, which is being updated throughout the imputation procedure. In this iteration, both the estimation of the model parameters and the values of the covariates are based on pre-imputation values as well as on the imputed values obtained in this first iteration for the covariates that have already been imputed (i.e. that have a lower order in the imputation list).

The sequencing of the variable imputation is particularly important in the first iteration, since it impinges both on the parameter estimation, as well as on available observations for the different covariates. Imputed observations are treated as if they were non-missing for imputing the remaining variables.

After the imputation of all the variables in the imputation sequence, the results of the first iteration are available in the Output_1 dataset.

9. Section 10 provides an assessment of convergence.

10. Sections 7 and 8 describe the content and the structure of the pre-imputation dataset.

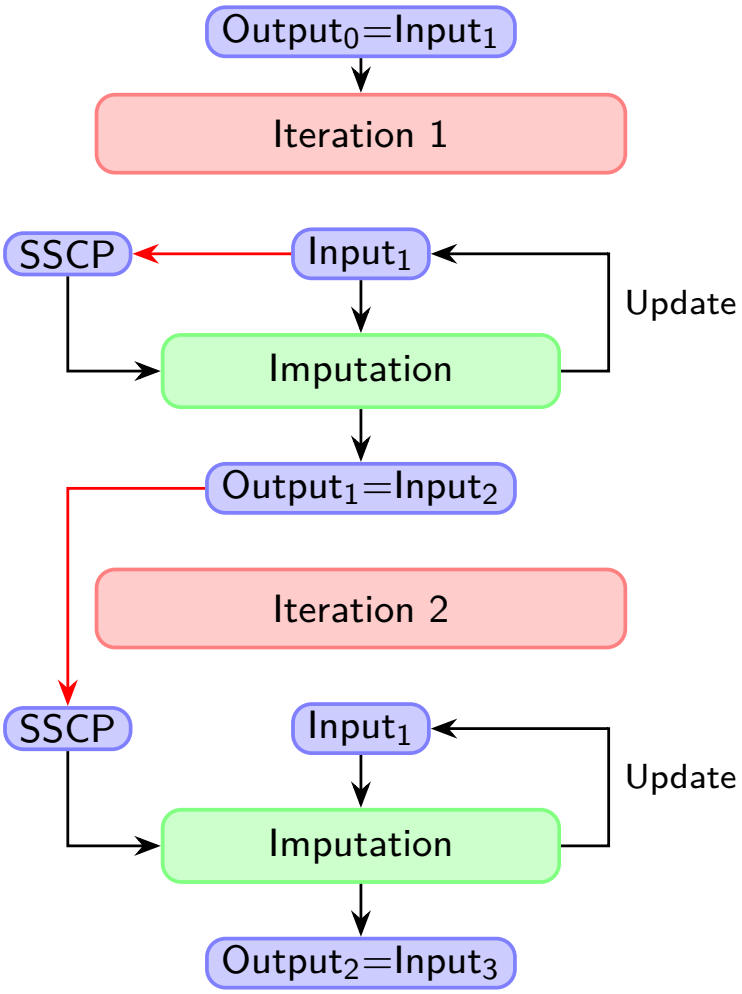


Figure 2: Imputation algorithm

The main difference from the first iteration to the second (and onwards) is that the dataset used in the previous iteration will be used as input (Input_2) of the SSCP matrices. Input_2 will be used only to compute the imputation model parameters and will not be updated with the results obtained during this iteration. Those parameters will be used along with Input_1 in order to calculate the imputed values. Notice that at the beginning of each iteration Input_1 contains only pre-imputation data and will be sequentially updated with the imputation results of the current iteration.

Recalling equation (2), this means that from the second iteration until the end of the imputation process, β_1 and β_2 will be computed using pre-imputation data and the imputed data from the previous iteration. By contrast, the covariates x_1 and x_2 will always refer to the pre-imputation data, plus the imputed data during the

current iteration. It immediately follows that after the first iteration, the estimation parameters will no longer be directly affected by the imputation order, since their computation will be based on the complete dataset obtained at the end of the previous iteration. In other words, after the first iteration, the parameters of the imputation models do not change during the imputation sequence and are only updated at the end of each iteration.

One alternative would be to use the dataset currently being imputed also as the source of the parameter estimations, in line with what happens during the first iteration. Nevertheless, that leads to stability and convergence issues, which are related with the fact that both the parameter estimates and the covariate information is being modified during the imputation sequence of each iteration.

7. Data transformation

In order to properly implement the imputation algorithm, several transformations have to be done to the data. This section motivates and describes those changes.

Categorical variables

Categorical variables have different types of transformations depending on their role as covariates, or as variables to impute. For their use as covariates, it is necessary to compute dummies for the different classes. Additionally, in some cases the categories were aggregated (e.g., foreign nationalities into EU and extra-EU aggregates). This saves some degrees of freedom, while preserving the relevant information. When categorical variables are imputed using the frequency model, in most cases their original format is kept. As referred in Section 5, one caveat of this model is that it does not allow the use of more than two covariates. Such limitation can be overridden by using as covariates computed variables that are the combination of two or more variables (e.g. age$\times$education). In the ISFF, the alternative strategy to overcome this limitation was to split some important categorical variables and to impute them in two steps. As an example, the categorical variable that corresponds to the main labour status (PE0100A), was decomposed in a binary variable that disentangles the cases where individuals are working from the remaining cases and in a categorical variable that includes the labour status other than working. With this artifact, the first variable was imputed with the binary model and only in cases where it was imputed that the individual is not working it was necessary to use the frequency model in a second step.

Monetary variables

Monetary variables were converted to logs when only positive values were allowed. When zeros were an admissible value, they were replaced by 0.1 and

then a log transformation to that variable was applied. Concerning negative values, variables were split into non-negative and negative parts and imputed separately.

Derived variables

Derived variables are calculated from the collected variables and typically refer to household level information (e.g. total income, total debt, age of the oldest person in the household). These variables were computed in order to be used as covariates in the imputation models, because aggregates such as total income may have a different meaning than each income component used separately. Moreover, using aggregates as covariates also saves some degrees of freedom.

Dealing with inapplicables

FRITZ estimates the imputation models using pairwise deletion as described in Section 5. With this procedure, if inapplicable cases are left empty, variances and covariances are potentially calculated based on very different numbers of observations, which can lead to the negative variance issues explained in Section 5, and ultimately restrict the choice of covariates used in each model. In order to minimize this problem, in the pre-imputation dataset inapplicable cases were filled with values that reflect their inapplicable nature. In the case of monetary variables, inapplicable observations were filled with zeros. The same approach was applied to variables related with the number of assets and liabilities (e.g. number of vehicles, number of HMR collateralized loans). Additionally, most of inapplicable binary variable cases were filled with its "no" equivalent, which is typically "2". For example, if households do not own their main residence, the variable referring to the possession of loans that use HMR as collateral will have a value of "2", instead of empty. Similar procedures were applied to other variables on a case-by-case basis.

This approach of using variables with treated inapplicable cases as covariates requires that the corresponding *parent variables* have to be included in the imputation models, in order to ensure the distinction between inapplicables and pre-imputation values.

Splitting person-level variables

Individuals within a given household have different characteristics and roles. Thus, the imputation process is likely to benefit from distinguishing and specifying different models for different types of individuals, provided that there are enough observations to do so.

In the imputation process, person-level variables were split according to the different types of individuals. Whenever there was enough observations, individuals were split into three categories: "Representative Person" (RP), "Married to Representative Person" (MR) and "Others" (O). Thus, for each person-level variable (either P or R variables), three new variables were computed (e.g. PG0110

will lead to PG0110RP, PG0110MR and PG0110O). The Representative Person is the individual with more than 15 years of age, that the household considers as such. Therefore, RP and MR will typically identify the main household couple. In the ISFF dataset, the identification of the type of person is stored in the variable RA0100. RP, MR and O correspond to the observations in the dataset where RA0100 equals one, two and more than two, respectively.

Further fragmentation of the individuals by type would not be feasible, due to problems associated with lack of observations. On the other hand, treating everyone besides the main household couple in the same group may pool widely heterogeneous individuals. In order to minimize this issue, RA0100 dummies were included as covariates when imputing the individuals other than the main household couple.

RA0100 dummies were also used as covariates in the cases where the low number of observations prevented any split of the person-level variables. This situation occurred mainly in pensions-related variables. [11]

Dealing with mass points in the variables' distributions

Interest rates and some monetary variables present very high frequencies of zeros, while in percentage variables there are many reported values of 100 percent. Additionally, some categorical variables such as labour status and business legal form have a mass point in some categories.

The binary model was used in a preliminary stage to deal with these cases and impute them separately. For this purpose, variables are split according to the mass points of the original variables. For example, the percentage of the main real estate property, other than the HMR, belonging to the household (HB2701) had more than 85 percent of the applicable cases with a value of 100 percent. Thus, HB2701 was split in two variables in the pre-imputation dataset:

$$
\begin{aligned}
HB2701\_1 &= \begin{cases} 1, & \text{if } HB2701 = 100 \\ 0, & \text{if } 0 < HB2701 < 100 \\ \text{missing}, & \text{if HB2701 is missing} \end{cases} \\
HB2701\_2 &= \begin{cases} HB2701, & \text{if } 0 < HB2701 < 100 \\ \text{inapplicable}, & \text{if } HB2701 = 100 \text{ or } HB2701\_1 = 1 \\ \text{missing}, & \text{if HB2701 is missing} \end{cases}
\end{aligned}
\tag{13}
$$

The binary variable HB2701_1 is imputed firstly, using the binary model. If the imputed value for HB2701_1 is equal to 1, then the imputed value for HB2701 is going to be equal to 100. Otherwise the imputation routine proceeds with the imputation of HB2701_2 relying on the continuous model, whose outcome is going to be a value greater than zero and lower than one hundred.

11. Section 8 discusses the integration of person-level variable splits in the dataset.

All of the variable transformations described in this section were performed not only in the pre-imputation dataset, but had also to be continuously updated during the imputation procedure. For instance, after imputing a categorical variable in its original format, all the dummy variables corresponding to that categorical variable had to be updated, in order to be used as covariates in the imputation of the following variables.

8. Data structure

In ISFF there are both household and person-level variables and ideally one would consider all variables as covariates in the imputation of the remainder, regardless of their type. However, such is not a realistic expectation due to data and computational constraints. Nevertheless, it is possible to find a reasonable compromise in order to achieve a properly imputed dataset. This section will describe the dataset organization used in the imputation process in order to:

- Provide differentiated treatment of individuals within the same household;
- Use person-level information when imputing household-level variables;
- Use household-level variables when imputing person-level variables;
- Use other than self-person-level information when imputing person-level variables.

Firstly the household and person-level variables were merged into a single dataset in a longitudinal format (one row per person). In this format, the household-level variables are repeated in each row for the individuals belonging to that household, so that H variables can be used as covariates when imputing the data of every individual. Then, person-level variables were split into RP, MR and O, as described in the previous section, in order to have different covariates and differentiated imputation models for each type of individual. Since there is only one RP and a maximum of one MR per household in the third wave of the ISFF dataset, the corresponding person-level variables can be treated as household-level variables and thus made available for every row. Regarding O variables, the information for these individuals is available only in their corresponding rows.[12]

Table 5 illustrates the ISFF dataset with the data structure described above.

SA0010 and RA0010 refer to the household and individual identifiers, respectively. RA0100 provides the relationship of each person with the RP. In this example, the household is composed by the RP (RA0100=1), the MR (RA0100=2) and two children of the RP (RA0100=3). The household-level variable HB0800 corresponds to the HMR value at the time of purchase. The person-level variable PG0110 corresponds to the annual gross employee income. The information of the other individuals, PG0110O, will still only be available on the corresponding row.

12. In Zhu and Eisele (2013), H variables are only available in the rows where they are used as covariates, which is a more efficient approach but may increase operational risk.

SA0010	RA0010	RA0100	HB0800	PG0110	PG0110RP	PG0110MR	PG0110O
1	1	1	200,000	20,000	20,000	15,000	.
1	2	2	200,000	15,000	20,000	15,000	.
1	3	3	200,000	12,000	20,000	15,000	12,000
1	4	3	200,000	8,000	20,000	15,000	8,000

Table 5. Illustration of ISFF dataset

However it is possible to compute a household-level indicator from that information (e.g., average) and then treat it as a household variable that can be used when imputing the data for the other individuals.[13]

In order to illustrate how the dataset is filled during the imputation process, the remaining of this section presents a very simplified example for the household of Table 5.

RA0100	HB0800	HB0900	PG0110RP	PG0110MR	PG0110O	PG0210RP	PG0210MR	PG0210O
1	200,000	.	20,000	15,000	.	18,000	.	.
2	200,000	.	20,000	15,000	.	18,000	.	.
3	200,000	.	20,000	15,000	12,000	18,000	.	2,000
3	200,000	.	20,000	15,000	8,000	18,000	.	.

Table 6. Example of a household with missing information in ISFF

In Table 6, the blue-coloured cells highlight the places in the dataset where there is missing information to be imputed. This household has three missing values: the current value of HMR (HB0900) and the self-employment income (PG0210) for the MR and the second child. Missing data for PG0210 is scattered across PG0210MR and PG0210O, which implies defining different models. The imputation of all of the H, RP and MR variables will take place in the first row. Then, after all of those variables have been imputed, the imputed data will be copied for every row, making it available for every other individual. The data of the remaining individuals (O) will be imputed in the end of this process, using a single imputation model.

Despite being operationally possible to intertwine the imputation of the O variables with H, RP and MR, that requires significantly more computations and processing power. It does not seem a good trade-off, since in general RP and MR would still be imputed before imputing data for the remaining individuals within a household. In fact, the information about the main household couple plays, in

13. An alternative would be imputing the data using a wide-format dataset. In this format, the person-level variables are in a single row per household. This allows for complete differentiation of every individual within a household. However, after some experiments this turned out to be computationally heavier and operationally harder to specify, without noticeable gains. Additionally, the order in which people appear in the household listing does not follow any particular order, after the RP and MR are accounted for. Thus, treating all of the 4th household members as somehow equivalent, as the wide-format approach implies, seems dubious.

general, a more important role in explaining the data of the other individuals than the reverse.

Suppose HB0900 is imputed firstly using the following (very simplified) specification:

$$HB0900 = \alpha + \beta_1 HB0800 + \beta_2 PG0110RP + \beta_3 PG0110MR + \beta_4 PG0210RP + \varepsilon \quad (14)$$

The imputation model of HB0900 uses as covariates HB0800, the RP and MR information about PG0110 (PG0110RP and PG0110MR, respectively) and also PG0210RP. A synthetic indicator of PG0110 for the other individuals could also be included.

Table 7 shows the dataset after the imputation of HB0900.

RA0100	HB0800	HB0900	PG0110RP	PG0110MR	PG0110O	PG0210RP	PG0210MR	PG0210O
1	200,000	350,000	20,000	15,000	.	18,000	.	.
2	200,000	.	20,000	15,000	.	18,000	.	.
3	200,000	.	20,000	15,000	12,000	18,000	.	2,000
3	200,000	.	20,000	15,000	8,000	18,000	.	.

Table 7. ISFF dataset after the imputation of HB0900

As mentioned earlier, the imputation of all household level and also RP and MR variables takes places in the first row, which corresponds to the RP row of each household in the dataset. The cells highlighted in green show the information used in the imputation of HB0900. The variable split, along with the replicated information for every individual, enables the usage of more information about PG0110, namely about MR. Otherwise, only the self-reported information would be accounted for, since person-level data would only be available for individuals in their corresponding rows.

In order to impute PG0210, different model specifications must be defined for PG0210MR and PG0210O. Starting with PG0210MR:

$$PG0210MR = \alpha + \beta_1 HB0800 + \beta_2 PG0110RP + \beta_3 PG0110MR + \beta_4 HB0900 + \beta_5 PG0210RP + \varepsilon \quad (15)$$

Since imputation occurs sequentially, as explained in Section 5, HB0900 can be used as a covariate, because it was already imputed, according to the specification in equation (14). The data structure makes it possible to account for the RP information (PG0110RP and PG0210RP) when imputing missing observations of the MR. Table 8 illustrates the ISFF dataset after the imputation of PG0210MR.

Finally, in the imputation of PG0210O the pre-imputation data of HB0800, PG0110RP, PG0110MR, PG0110O, PG0210RP and the imputed data for HB0900

RA0100	HB0800	HB0900	PG0110RP	PG0110MR	PG0110O	PG0210RP	PG0210MR	PG0210O
1	200,000	350,000	20,000	15,000	.	18,000	17,500	.
2	200,000	.	20,000	15,000	.	18,000	.	.
3	200,000	.	20,000	15,000	12,000	18,000	.	2,000
3	200,000	.	20,000	15,000	8,000	18,000	.	.

Table 8. ISFF dataset after the imputation of PG0210MR

and PG0210MR will be used:

$$\begin{aligned} PG0210O = \alpha &+ \beta_1 HB0800 + \beta_2 PG0110RP + \beta_3 PG0110MR \\ &+ \beta_4 PG0110O + \beta_5 HB0900 + \beta_6 PG0210RP \\ &+ \beta_7 PG0210MR + \varepsilon \end{aligned} \tag{16}$$

Table 9 shows the dataset after the imputation of PG0210O. Before imputing PG0210O, all of the H, RP and MR observations available in the first row had to be made available for the remaining rows, otherwise those covariates in (16) would be missing.

RA0100	HB0800	HB0900	PG0110RP	PG0110MR	PG0110O	PG0210RP	PG0210MR	PG0210O
1	200000	350000	20000	15000	.	18000	17500	.
2	200000	350000	20000	15000	.	18000	17500	
3	200000	350000	20000	15000	12000	18000	17500	2000
3	200000	350000	20000	15000	8000	18000	17500	4000

Table 9. Modified dataset after the imputation of PG0210O

This section illustrated a simplified imputation example, in order to provide the reader with the intuition behind the imputation process. It overlooked some aspects, namely the implicates and iterations, that were explained in Sections 2 to 6. Additionally, while in these examples only a few covariates where used, in practice more than 80 covariates were used per model on average (excluding the frequency models, which only allow for two covariates), in order to comply with the broad conditioning principles. The covariates used in the implemented imputation models were chosen according to several criteria, which will be discussed in the next section.

9. Covariates selection and imputation order

This section describes the criteria used to choose the covariates for each imputation model and to define the order of the variables to impute. Ultimately, this is an extensive and iterative learning-by-doing process that makes for most of the time allocated to the imputation of ISFF.

Selection of covariates

As mentioned in the beginning of this paper, the main goal of imputation is not to obtain the best prediction of the missing values, but to preserve the characteristics of the variable joint distributions. For this purpose, the covariate selection should not be limited to the variables that are most correlated with the variable to impute.

According to Rubin (1996), the imputation model should include as many predictors as possible, in order to accommodate any potentially important variable. This kind of broad conditioning would also allow to capture the different missingness patterns, because it accounts for enough substitute covariates if others are missing, as referred by Zhu and Eisele (2013). This approach is also in line with the congeniality requirements explained in Takahashi (2017), which imply that the imputation model can be larger than the corresponding substantive analysis model, but it must not be smaller. Moreover, imputation models should also account for the different hypotheses about competing economic theories (e.g. permanent income hypothesis versus precautionary saving motive).

Nevertheless, the degrees of freedom impose a limit on the number of covariates that can be used in the imputation models. To address this issue in the ISFF, a threshold for the number of covariates equal to 20 percent of the number of observations was imposed. Keeping this setup in mind, the definition of continuous and binary models accounted for different sets of covariates:

- A common set of household-level variables, related with geographical location, type of family (e.g. number of household members), income, consumption, assets, liabilities, dwelling characteristics reported by the interviewers and the sample design weights;
- A common set of person-level variables, especially RP and MR features such as education, labour status, types of income received, age, gender and marital status;
- A specific set of covariates, containing almost all variables within the same questionnaire section as the imputed variable, as well as correlated or potentially relevant variables from other sections.

The common sets of variables are used conditionally on the degrees of freedom.

The sets of covariates also contain *parent variables* of covariates whose inapplicable cases were treated, as mentioned in Section 7. For example, when using some monetary variables as covariates, the corresponding yes/no *parent variables* are included in order to disentangle "real" zero values from the zeros due to being inapplicable. This approach minimizes the selection bias, since these zeros may correspond to fundamentally different subsets of the population, while enabling the usage of otherwise jointly exclusive covariates. The alternative would be specifying

different models for every subgroup of the population, whenever such covariates were deemed important.[14]

The pool of covariates includes not only variables that are on the list of variables to impute, but also other relevant survey variables. Variables with high non-response rates were not used as covariates, unless there was an imperative reason, since that introduces an additional source of uncertainty. Moreover, those would be imputed later on in the imputation sequence and therefore would have many missing values before being imputed, which could build on the negative variance issue described in Section 5.

Special case: imputation of variables with very low number of observations

One important obstacle for the definition of a well-specified imputation model is the low number of observations. In the ISFF this issue is particularly important in the case of variables that belong to a high iteration in the questionnaire loops.[15] These variables have, in most cases, a low number of applicable situations and also suffer frequently from high non-response rates. The lack of observations to estimate a model was addressed by applying the coefficients estimated for the first item in the loop to the remaining within the same loop. For instance, in the case of the monthly payments on the HMR mortgage, firstly a model is defined for the monthly payment on the first loan, HB2001. The covariates' parameters obtained for HB2001 are applied to the imputation models of HB2002 and HB2003, in order to impute the values of the monthly payments of second and third loans, respectively.

This strategy assumes that the variable relationships do not differ significantly across the loop, which could arguably be a strong hypothesis. Alternative strategies that might be explored in future waves consist of estimating the coefficients with pooled data from the three iterations, or in the case of monetary variables imputing a sum of the items on the three iterations and then use the total to calculate the missing parts.

Imputation order

Defining the sequence of variables to impute is a puzzle which has to take into account, not just the survey logical tree, but also the number of missing values in the covariates. Since competing standards are often found, the ordering is defined according to the following criteria hierarchy:

1. *Parent variables* are imputed before the corresponding *child variables*;

14. See Zhu and Eisele (2013) for further depiction of this argument.

15. In ISFF, the assets and liabilities of the same type are collected in a loop. For example, after asking about the number of loans that use HMR as collateral, there is a set of questions about loan characteristics that are asked for each one of the three most important loans.

2. Household-level, RP and MR variables are imputed before O and non-split person-level variables;
3. Imputation of RP occurs before the corresponding MR variables;
4. Loop variables are imputed according to the loop order;
5. Variables with low non-response rates (e.g. binary, education, employment status) are imputed first;
6. Variables that are for good predictors of the remaining variables to be imputed are imputed first.

10. Model evaluation and assessment of imputation results

The specification of each imputation model results from an iterative process that involves *ex ante* and *ex post* analysis.

Before running the whole imputation sequence, initial candidate covariates, selected according to the broad conditioning purpose described in the previous section, are included in a regression. This first step allows for some trimming of the covariates by dropping the ones with high pairwise missingness, which could lead to extraneous SSCP outcomes in the first iterations, as explained in Section 5. After this preliminary step, the regressions are ran in the first iteration of the imputation process to address similar issues that were not identified earlier. The models obtained at the end of this process are then used on a full imputation sequence. Finally, the statistical and economic plausibility of the imputed results is evaluated, which may lead to the exclusion from the estimates of outliers or highly influential observations and can also result in new model specifications. This model evaluation is a dynamic process which is entangled with the assessment of imputation results, which will be described in the remainder of this section.

Simulated distributions should converge across the iterative process. During the imputation procedure, as the program executes the different iterations, imputed observations for each implicate are expected to move closer to each other, despite the stochastic disturbance added to the point estimates.

Gelman and Rubin (1992) proposed an indicator to assess the convergence of the imputation process. For a given dataset with $m = 1, ..., M$ implicates and $t = 1, ..., T$ iterations, the Gelman and Rubin (GR) convergence diagnostic examines the ratio between the variation of the estimates (e.g., mean of HB0900) across the implicates in each iteration (BV) and within each implicate across iterations (WV):

$$GR = \sqrt{\frac{T-1}{T} + \frac{BV}{WV}} \tag{17}$$

In order to calculate the between-implicate variability (BV), firstly, for each implicate, the mean of the imputed values in each iteration must be computed:

$$\overline{X}_m = \frac{1}{T} \times \sum_{t=1}^{T} X_m^t \tag{18}$$

The between-implicate variability, BV, corresponds to:

$$BV = \frac{T}{M-1} \times \sum_{m=1}^{M} (\overline{X}_m - \overline{X})^2 \tag{19}$$

where $\overline{X}$ denotes the mean of the M (five in the case of ISFF) implicate means:

$$\overline{X} = \frac{1}{M} \times \sum_{m=1}^{M} \overline{X}_m \tag{20}$$

The within-implicate variability (WV) is the average of the M variances of estimates within each implicate obtained across iterations:

$$\begin{aligned} s_m^2 &= \frac{1}{T-1} \times \sum_{t=1}^{T} (X_m^t - \overline{X}_m)^2, \\ WV &= \frac{1}{M} \times \sum_{m=1}^{M} {s^2}_m \end{aligned} \tag{21}$$

A GR value lower than 1.1 denotes convergence, according to Gelman *et al.* (1996). A high dispersion of implicates, leading to higher between implicate-variance, contributes negatively to the convergence of the estimates, since this signals high uncertainty and instability of the imputed values. On the other hand, movements of the estimates of variables within each implicate along the iterative process indicate that imputation can cover well the domain of the joint distribution, as mentioned by Zhu and Eisele (2013).

In the ISFF, the convergence was tested with both the original GR defined in equation (17), as well as with an alternative, more restrictive, GR, which was calculated using 1 instead of $\frac{T-1}{T}$ in equation (17). The tests were calculated for the mean and several percentiles (10, 25, 50, 75 and 90). Nevertheless, those indicators can only be used on continuous variables with a minimum threshold number of imputed observations.[16]

In the ISFF, the imputation routine included fifteen iterations and according to the GR tests convergence was achieved after six iterations, using a burn-in period

16. These indicators were calculated for all the imputed variables, excluding the categorical and some continuous variables referring to very specific items, such as the third mortgage on the third reported property other than HMR.

of one iteration in order to reduce the dependence of the results on the initial values.

The remainder of this section illustrates the kind of analysis done to evaluate convergence and the plausibility of imputation results, using the example of HB0900.

Table 10 shows the GR and its alternative definition for the different estimates of HB0900. Both indicators signal convergence for the imputed observations of the

	Mean	**Median**	**P10**	**P25**	**P75**	**P90**
GR	0.9788	1.0018	0.9823	0.9688	0.9897	0.9854
Alt GR	1.0146	1.0368	1.0180	1.0050	1.0251	1.0210
BV/WV	0.0294	0.0750	0.0363	0.0100	0.0509	0.0424
BV	40,036	95,727	30,196	6,230	468,833	619,396
WV	1,362,465	1,275,633	831,974	621,747	9,216,885	14,616,227

Table 10. Convergence indicators for imputed values of HB0900

current value of HMR, regardless of the statistics evaluated, since all measures are below the critical value of 1.1.

The highest variation, both between and within implicates, comes from estimates based on the P_{90}. Nevertheless, measures relying on the P_{50} present the highest proportion of between-to-within implicate variability, which explains why both GR indicators are higher in this case.

Figure 3 illustrates the convergence process of the (unweighted) mean for imputed values of HB0900.

The sharp adjustment of the mean of imputed values right after the first iteration is fairly distinguishable. Such is expected since the latter is meant only to provide the necessary values to initialize the imputation process. After that, the implicate mean (solid red line) continues to decrease until it stabilizes from the fifth/sixth iteration onwards. Notice that the means of each implicate in each iteration were not necessarily computed based on the same households, as the imputation of the *parent variable* related with the tenure status may lead to different applicable cases.

Convergence alone is not a guarantee of the quality of the imputation. Therefore, on top of the previous convergence assessment, the plausibility of the imputed data from an economic point of view is also evaluated.

Figure 4 shows, for HB0900, the non-imputed, imputed and non-imputed + imputed distributions and its corresponding normal density curves. Table 11 shows the statistical indicators for the corresponding distributions. This analysis is based on unweighted data, since it aims to assess the impact of imputation on the variable distributions and not to draw conclusions about the population.

The distribution of the non-imputed values of HB0900 is positively skewed with a very long tail, due to the presence of a small number of households with very high main residence values. The distribution of the imputed values shows a similar shape.

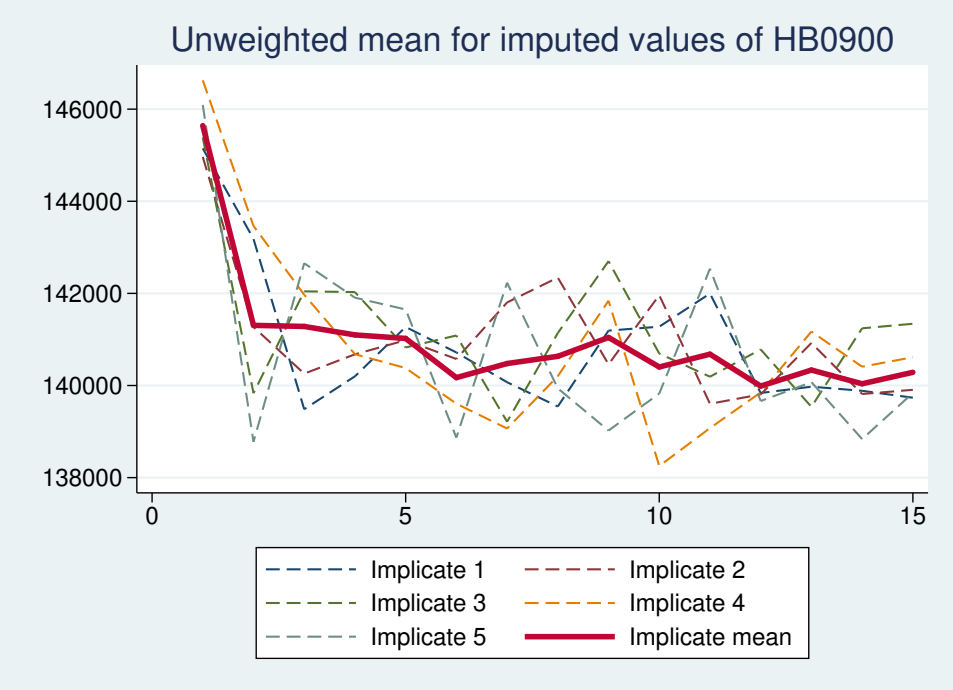


Figure 3: Graphical convergence assessment of the unweighted mean of HB0900

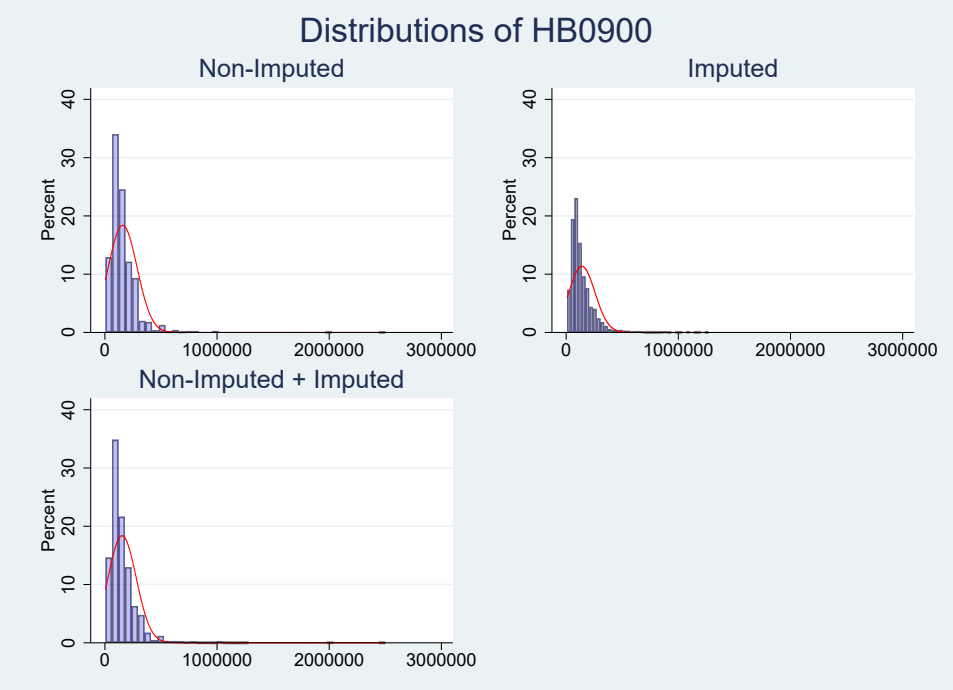


Figure 4: Assessment of the imputed value distribution of HB0900

	Non-Imp	Imp	Non-Imp+Imp
Mean	158,416	140,286	153,601
Median	130,000	107,418	124,466
Max	2,500,000	1,269,059	2,500,000
Min	5,000	7,499	5,000
P10	60,000	46,242	50,000
P25	80,000	71,151	80,000
P75	200,000	174,110	200,000
P90	300,000	262,092	280,000
Std Dev	128,737	116,559	125,871
Coef of Var	0.81	0.83	0.82
Skewness	5.15	3.24	4.75
Kurtosis	63.07	20.52	55.04
N	17,875	6,464	24,339

Table 11. Unweighted statistics for the distributions of HB0900

Nevertheless, the distribution is less skewed and has lower percentile and mean values. The fact that the imputed values were globally lower and less dispersed, compared to the non-imputed ones, does not necessarily signal any problem with the imputation process. As explained in the beginning of this paper, a correlation between the non-response pattern and the values of the variable to be analysed may exist. For example, in this particular case the households with imputed values of HB0900 have lower mean and median values for the main residence size and for the income, compared to the households where HB0900 was collected.

Ultimately, in order to analyse the impact of the imputation on the variables' joint distribution, several relationships between the variables of the same household had to be compared in the pre-imputation and post-imputation datasets. Evaluating the imputation results is a complex endeavour that requires a comprehensive analysis of different indicators besides the convergence assessment.

11. Concluding remarks

Non-response is a major theme in wealth surveys, since households may not be willing or able to disclose sensitive information, such as the value of assets, debt and income. Using only complete interviews would lead towards biased estimates of statistics, as the non-response patterns may be correlated with household features. Multiple imputation provides the means to deal with item non-response, while preserving the characteristics of the joint distributions and ultimately enabling valid statistical inference. It accounts for within and between-imputation variance, thus providing a clear picture concerning the limits of the knowledge about the missing data.

Multiple imputation of survey data is a challenging and complex procedure, due to constraints related with complex logical trees of the questionnaire, critical

relationships, bounds, data structure and missing patterns. This paper provides an overview of the different stages of the imputation process, from data preparation, to evaluation of imputation results.

This paper delivers a comprehensive and approachable description of the imputation process in the ISFF, in order to share the imputation experience with the interested readers, namely other data producers and the data users. Ultimately, it will be used as a starting point for the imputation of future waves of the ISFF.

References

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

Barceló, Cristina (2008). *The Impact of Alternative Imputation Methods on the Measurement of Income and Wealth: Evidence from the Spanish Survey of Household Finances*. Banco de España Working Papers 0829.

Dempster, Arthur, Nan Laird, and Donald Rubin (1977). "Maximum Likelihood from Incomplete Data Via EM Algorithm." *Jornal Royal Statistical Society, Series B*, 39, 1 – 38.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin (1996). *Bayesian data analysis. 2nd ed*, vol. 52.

Gelman, Andrew and Donald Rubin (1992). "Inference From Iterative Simulation Using Multiple Sequences." *Statistical Science*, 7.

Geman, Stuart and Donald Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

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

Kennickell, Arthur (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.

Rubin, Donald (1976). "Inference and Missing Data." *Biometrika*, 63, 581–592.

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

Rubin, Donald (1996). "Multiple Imputation After 18+ Years." *JASA. Journal of the American Statistical Association*, 91.

Takahashi, Masayoshi (2017). "Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations." *Data Science Journal*, 16, 37.

Zhu, Junyi and Martin Eisele (2013). *Multiple imputation in a complex household survey - the German Panel on Household Finances (PHF): challenges and solutions*. EconStor Preprints 100007, ZBW - Leibniz Information Centre for Economics.

Occasional Papers

2016

1|16 Public debt sustainability: methodologies and debates in European institutions
João Amador | Cláudia Braz | Maria M. Campos | Sharmin Sazedj | Lara Wemans

2019

1|19 The Deepening of the Economic and Monetary Union
João Amador | João Valle e Azevedo | Cláudia Braz

2|19 A tentative exploration of the effects of Brexit on foreign direct investment *vis-à-vis* the United Kingdom
Ana de Almeida | Duncan Van Limbergen Marco Hoeberichts | Teresa Sastre

3|19 Sovereign exposures in the Portuguese banking system: evidence from an original dataset
Maria Manuel Campos | Ana Rita Mateus Álvaro Pina

4|19 Economic consequences of high public debt and challenges ahead for the euro area
Cristina Checherita-Westphal | Pascal Jacquinot | Pablo Burriel | Maria Manuel Campos | Francesco Caprioli | Pietro Rizza

2020

1|20 Banco de Portugal TARGET balance: evolution and main drivers
Rita Soares | Joana Sousa-Leite | João Filipe | Nuno Nóbrega

2|20 Imputation of the Portuguese Household Finance and Consumption Survey
Luís Martins