Austrian Journal of Statistics
April 2014, Volume 43, 5–28.
http://www.ajs.or.at/

Multiple Imputation in the Austrian Household Survey on Housing Wealth

Nicolás Albacete
Oesterreichische Nationalbank

Abstract

This paper presents the multiple imputation model for the imputation of the missing values of the Austrian Household Survey on Housing Wealth 2008. It is based on Bayesian inference and on the fully conditional specification approach. Both theoretical framework and model specification are discussed in detail and, finally, some results about the performance of our imputations are presented.

Keywords: Household wealth survey, imputation methods.

1. Introduction

In 2008 the Oesterreichische Nationalbank carried out a survey called the Household Survey on Housing Wealth (HSHW) which covered, among other, questions concerning housing wealth, housing debt, or intergenerational transfers of Austrian households. Similar surveys, covering the whole household wealth, have been done by the Federal Reserve (Survey of Consumer Finances, SCF), by the Banca d'Italia (Survey on Household Income and Wealth, SHIW), or by the Banco de España (Survey of Household Finances, EFF).

Because of the particular sensitivity and difficulty of wealth questions, a common problem in such surveys is that more households than usual refuse to participate in the interview (unit nonresponse), or do participate in the interview, but refuse to answer specific questions (item nonresponse). When such data is analyzed by just excluding those households that have nonresponse for any of the variables involved in the analysis (complete-case-analysis), then in general estimates are going to be biased, because they measure the portion of the target population that provides responses on all relevant variables in the analysis, rather than the entire target population.[1] A further disadvantage of complete-case-analysis is the estimates' loss of efficiency due to the loss of information.

[1]Most statistical packages do complete-case-analysis by default.

Nonresponse bias is typically a problem in wealth surveys, where nonresponse is usually positively correlated with wealth. In the USA in the SCF which is a conceptual base for the HSHW the unit-nonresponse rate in the general sample is much lower (70 percent) than in the special sample with wealthy households (about 90 percent) (see e.g. Kennickell (1998a)). This illustrates the typical problem of surveys on wealth: wealthy households refuse more often to participate and, therefore, wealth is underestimated in voluntary household surveys. Moreover, as net wealth is a complex variable derived from many different components of gross wealth and debt which have to be collected one by one, such surveys are particularly prone to item-nonresponse issues.

Therefore, it is important to find a more appropriate technique for handling unit and item nonresponse. Little and Rubin (2002) offer an extensive survey of current methodology. For addressing unit nonresponse, a common technique is weighted complete-case-analysis. The idea is to differentially weight households with complete observations to adjust for the nonresponse bias. Although simple, this method still has the disadvantage of loosing efficiency by excluding households with incomplete observations. We will not focus on unit nonresponse in this paper. See Wagner and Zottel (2009) for some more information on unit nonresponse in the HSHW.

Item nonresponse is typically handled by imputation methods. The idea is to fill in the missing values in order that the resultant completed data can be analyzed by standard methods. If we only impute one value for each missing value (single imputation), then standard variance formulas underestimate the variance of estimates, because they do not take into account the uncertainty behind the imputed values. With multiple imputation more than one value for each missing item is imputed allowing for a differing status of the real and the imputed values and therefore, the problem of too low variance is largely corrected.

Of course, the technique that creates the imputed values is also relevant. For example, mean imputation is a simple method that substitutes missing values with means from recorded values. The problem is that it implicitly assumes like a complete-case-analysis that missing values are independent from household characteristics. Furthermore, it biases the correlations between the variables and, therefore, the variance. Another common imputation technique is hot deck imputation, where recorded values in the sample are randomly chosen to substitute missing values, or regression imputation, where missing variables for a household are estimated by predicted values from the regression on the known variables for that household. In general, according to Little and Rubin (2002), imputations should generally be (1) conditional on observed variables, to reduce nonresponse bias, improve precision, and preserve association between missing and observed variables; (2) multivariate, to preserve association between missing variables; (3) draws from the predictive distribution of missing values rather than means, to provide valid estimates of a wide range of estimands.

The imputation methods mentioned in the last paragraph tend to have an ad hoc character, often being solutions worked out by practitioners with limited research into theoretical properties.[2] During the last decades the imputation literature has developed towards sys-

[2]"Ad hoc" in the sense of producing univariate imputations where each variable is imputed independently

tematizing these methods (model-based approach) to provide a basis for future advances. Many of the ad hoc imputation methods mentioned before can be derived as examples (or approximations) of the model-based approach. The idea is to define a model for the observed data, then, based on the likelihood under that model, to estimate the parameters by procedures such as maximum likelihood and, finally, to combine these results to estimate the joint distribution of the observed and unobserved data.[3] The advantages of model-based procedures are (1) flexibility; (2) the avoidance of ad hoc methods, in that model assumptions underlying the resulting methods can be displayed and evaluated; (3) the availability of estimates of variance that take into account incompleteness in the data.

In case of small sample sizes, such as the HSHW with 2081 household observations, maximum likelihood estimates could yield to unsatisfactory inferences, as their large sample properties might no longer be valid. One approach to this limitation is to adopt a Bayesian perspective and instead of basing inferences on the likelihood, to base them on the exact posterior distribution for a particular choice of prior.[4] Most model-based methods, especially in the context of multiple imputation, were developed in a Bayesian framework.

When the data has a complex structure, it might be very difficult to explicitly specify a joint distribution for the data that reflects the data well (joint modeling (JM) strategy). Instead one could define it implicitly, by specifying a set of conditional distributions relating each variable to a set of the other variables (fully conditional specification (FCS) strategy). For each of the variables, a draw of parameters is made, the missing data are imputed for that variable, and the procedure cycles through the variables, replacing variables that are being conditioned in any regression by the observed or currently imputed values. In data like the HSHW, with a large number of variables and where many of them have bounds, skip patterns, bracketed responses, interactions, or constraints with other variables, separate regressions for each variable as in the FCS approach often make more sense than postulating a joint model. However, one main drawback of FCS is that the implied joint distribution of the data may not exist theoretically and therefore little is known about the quality of the resulting imputations. Despite this, simulation studies provide evidence that this strategy works quite well in many applications and yields estimates that are unbiased, at least in the cases investigated.[5]

Recent research by Rubin (2003) or Baccini, Cook, Frangakis, Li, Mealli, Rubin, and Zell (2010) proposes to limit this incoherence by combining the JM and the FCS strategy. The idea is to split the data into monotone missingness blocks and use the JM strategy within each block and the FCS strategy across the blocks.

For handling item nonresponse in the HSHW, we choose a Bayesian-based FCS multiple imputation approach for the following main reasons: (1) it seems to be very successful in reducing nonresponse bias according to the above mentioned literature, (2) the approach

from the other variables.

[3]For example, it can be shown that if the data are assumed to be multivariate normally distributed and to have a monotone missing data pattern, maximum likelihood analysis is equivalent to regression imputation (Rubin (1974)).

[4]Even in cases where prior knowledge for the parameters is limited, the Bayesian approach with dispersed priors often yields better inferences than the frequentist approach.

[5]See Van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006) for a more detailed discussion on the fully conditional specification strategy.

preserves the complex HSHW data structure, (3) statistical packages that include this approach are available, and (4) other similar surveys as the SCF or the EFF also successfully employ this approach.

This paper presents the implementation of the HSHW multiple imputation model and is structured as follows. Section 2 introduces the data and provides descriptive statistics about nonresponse. Section 3 discusses the implementation of the HSHW imputation model including a brief explanation of the theoretical framework. In Section 4 some results of the imputed data are presented. Finally, Section 5 concludes.

2. Nonresponse in the HSHW

The HSHW was conducted based on computer assisted personal interviews (CAPI) in the year 2008 and provides the only data source on household housing wealth in Austria. Apart from the focus on housing wealth, the questionnaire also covers housing debt, intergenerational transfers (inheritances, education of parents) and a series of socioeconomic and sociodemographic characteristics of the household. See Fessler, Mooslechner, Schürz, and Wagner (2009) for some general results of the survey.

As already mentioned in the Introduction, due to the sensitivity and complexity of the HSHW questions, the nonresponse rate is relatively high. In addition, the survey has a large number of variables (168 questions), a complex structure due to filtering, and long interview duration (42.3 minutes on average) with a high bandwidth (from 30 minutes to over an hour, depending on the filtering). To reduce unit nonresponse interviewers were instructed to do until five contact attempts to each household. Also five interviewer training sessions and a pretest were organized in different regions of Austria. The resulting unit nonresponse rate was on average 34.9 percent (in Vienna even 50.1 percent), which is in line with other similar surveys. To correct for the typically higher unit nonresponse bias in Vienna, households coming from this region were already oversampled when drawing the sample. See Wagner and Zottel (2009) for more details on the HSHW unit nonresponse.

Concerning item nonresponse, the mean number of missing values per household in the HSHW is only 3. The median is 2 and the 90th percentile is 8. In other words, 50 percent of the households deny the answer to not more than two questions and 90 percent of the households deny to not more than eight questions. The total number of missing values over all households is $6,322$ which is equivalent to 2.6 percent of all the questions asked over all households. Although these statistics are surprisingly low, they are not necessarily a good measure of the degree of information missing due to nonresponse, as noted by Kennickell (1991). This is because all missing values are added up equally, but not all variables with missing values are of equal importance for the objectives of the survey. For example, the estimated total value of real estate is more important than the current value of the second additional house or condominium. The more concentrated missing values are on the important questions of the survey the higher the degree of information missing should be. In general, it is difficult to find a good measure of the information missing due to nonresponse and it will become easier once we have imputed (see Section 4).

An alternative illustration of item nonresponse is presented in Table 1, where response rates are shown per item instead of per household. For example, for the question of the

outstanding loan amount for acquiring the primary residence, we see that 17.3 percent of the households arrive to this item (column 1) and 39.2 percent of those who arrive give a number (column 2). Thus, the nonresponse rate here is 60.8 percent. For the annual total of rental or leasing income the nonresponse rate is 11.5 percent. Most of the other nonresponse rates lie between these two numbers. The estimated total value of real estate, for example, has a nonresponse rate of 35.4 percent. These nonresponse rates will play an important role for imputations because they determine the order of imputations during the whole imputation process (see Section 3.1).

Table 1: Unweighted Item-nonresponse in selected items of the HSHW

		Value reported by respondent, for those who responded having the item			
Item	**Have item**	**Number**	**Range***	**DK**	**NA****
HOUSING WEALTH					
Current value of the primary residence	52.1	73.5	14.8	7.3	4.4
Current value of the first additional house/condominium	11.0	64.6	x	3.9	31.4
Current value of the second additional house/condominium	1.8	64.9	x	5.4	29.7
Current value of plots of land/building lot	7.1	72.1	x	5.4	22.4
Current value of agricultural or forestry real estate, fields, forest etc.	5.7	78.0	x	6.8	15.3
Current value of office, business premises, company site	0.7	60.0	x	6.7	33.3
Current value of other real estate	0.9	77.8	x	0.0	22.2
Estimated total value of real estate	22.2	64.6	20.0	8.2	7.2
HOUSING DEBT					
Outstanding amount of loans for acquiring the primary residence	17.3	39.2	34.4	17.8	8.6
Outstanding amount of loans for acquiring additional houses	3.0	30.2	x	14.3	55.6
Outstanding amount of loans for acquiring plots of land/building lots	1.3	14.3	x	0.0	85.7
Annual repayment for acquiring the primary residence	17.3	49.4	33.3	8.6	8.6
INTERGENERATIONAL TRANSFERS					
Value of the properties given away as a gift	3.6	68.0	x	12.0	20.0
Value of the properties inherited	20.1	69.6	x	6.0	24.4
OTHER CHARACTERISTICS OF THE HOUSEHOLD					
Total monthly net household income	100.0	67.3	22.6	0.4	9.7
Household head's monthly net income from paid employment	100.0	72.8	14.1	0.0	13.2
Homeowners' imputed rent	52.1	75.0	11.6	8.4	5.0
Tenants' monthly rent paid	43.6	77.3	13.9	5.8	3.0
Tenants' deposit to the housing association (Genossenschaftsbeitrag)	15.7	67.9	27.8	1.8	2.4
Annual total of rental or leasing income	5.4	88.5	x	8.8	2.7
Time since the household head's father passed away	51.4	88.0	x	0.0	12.0
Time since the household head's mother passed away	37.7	88.8	x	0.0	11.2

* x means that the item has no range question

** Includes some editing cases

Although some of these nonresponse rates are rather high, column 3 of the table[6] shows that the use of range questions after euro variables is extremely helpful in significantly reducing pure non-response rates of these variables: many households who do not give an amount for a certain item are at least willing to select a range from a given list in which

[6]The x values in this column mean that, unfortunately, no range question was posed for these items.

this missing amount is lying. For example, in the question of the outstanding loan amount for acquiring the primary residence, 34.4 percent of households who arrive to this item do not provide an amount but do at least provide a range for this amount. Thus, after range responses just 17.8 percent still completely "don't know" the answer (column 4), and 8.6 percent are households that intentionally do not want to provide any response (column 5). In case of the current value of the primary residence, 73.5 percent of homeowners report a number and 14.8 percent a range, such that just 11.7 percent do not provide any information.

Another interesting aspect of nonresponse is what determines it. Table C.1 in the Appendix presents the results of a logit regression of a nonresponse dummy for the value of primary residence on some household and interviewer variables.[7]

The estimation shows that the probability of not responding to the question about the value of the primary residence increases significantly when the respondent[8] is female, when she is a farmer, when the household's municipality size is large, or when the interviewer is female. For example, a particularly high item nonresponse for farmers on the value of the primary residence is not very surprising because for farmers it is particularly difficult to sepparate farming wealth (business wealth) from the value of the main residence. On the other hand, the probability of nonresponse decreases significantly, when the education level of the respondent is high, when the standard of living of the household is rather basic or poor, or when the number of persons who provided information during the interview is high. Age has a U-shaped effect on nonresponse: for younger age groups nonresponse is decreasing with age, but for higher age groups it is increasing. The low item nonresponse for households with a rather poor standard of living supports the general problem of wealth surveys that nonresponse is positively correlated with wealth (see Introduction).

The results of Table C.1 also tell us something else. They support our presumption stated in the Introduction that nonresponse in the HSHW does not happen completely at random. The fact that many coefficients in the above regression are significant implies that if we would do complete-case-analysis of this variable our inferences would have a nonresponse bias. Thus, imputations are necessary.

3. Imputation Method of the HSHW

3.1. Theoretical Framework [9]

Let $Y = (y_{ij})$ denote an $(n \times K)$ rectangular data set that would occur in the absence of missing values, with ith row $y_i = (y_{i1}, \ldots, y_{iK})$ where y_{ij} is the value of variable Y_j for household i, for $i = 1, \ldots, n$ and $j = 1, \ldots, K$. With missing data, define the missing-data indicator matrix $M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. The matrix M then defines the pattern of missing data. We write $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} denote the observed components or entries of Y, and Y_{mis} the missing compo-

[7] Household income is not included as a regressor because it has several missing values. By excluding those households with missing values from the regression sample we would introduce a selection bias in the estimation, as nonresponse of income is probably not random.

[8] The respondent of all survey questions is always the owner of the household's main residence.

[9] See Little and Rubin (2002) for more details.

nents.[10]

Furthermore, the probability or density of the joint distribution of Y_{obs}, Y_{mis} and M is denoted by $f(Y_{obs}, Y_{mis}, M \mid \theta, \psi)$ which is indexed by the unknown parameters θ (for Y) and ψ (for M). The likelihood and the prior distribution of these parameters are denoted by $L(\theta, \psi \mid Y_{obs}, M)$ and $p(\theta, \psi)$, respectively, and Bayes inference is obtained by their joint posterior distribution: $p(\theta, \psi \mid Y_{obs}, M) \propto p(\theta, \psi) \times L(\theta, \psi \mid Y_{obs}, M)$. Under the assumption that the missing-data mechanism is ignorable (see next paragraph) it can be shown that Bayes inference about θ simplifies by just dropping M and ψ from the last expression: $p(\theta \mid Y_{obs}) \propto p(\theta) \times L(\theta \mid Y_{obs})$.

Intuitively, the ignorability assumption of the missing-data mechanism means that nonresponse probabilities do not depend on any unobserved information. In the case of the wealth variable, this means that we assume that nonresponse of the amount of wealth does only depend on observed values and not on unobserved ones like the missing amount of wealth itself. Even if we suspect, as we do, that wealthy households may be less likely to report their wealth, this is still compatible with the ignorability assumption as long as we condition on many other observed variables, especially those that are highly correlated with wealth. In this way, we may be able to reduce or eliminate the dependence of missingness on wealth and make the ignorability assumption much more reasonable. See the Appendix for the technical definition of ignorability.

Our aim is to impute by drawing the missing values as $Y_{mis} \sim p(Y_{mis} \mid Y_{obs})$, that is, from their joint posterior predictive distribution. As already mentioned in the Introduction, there are two approaches to draw from this distribution: the JM approach and the FCS approach. The JM approach explicitly assumes a density function $f(Y \mid \theta)$ for the joint distribution of $Y = (Y_{obs}, Y_{mis})$ and uses Markov Chain Monte Carlo methods such as *data augmentation* or the *Gibbs' sampler* to obtain draws from the joint posterior predictive distribution of Y_{mis} consistent with the assumed density function.

For example, according to Raghunathan, Lepkowski, Hoewyk, and Solenberger (2001), one can develop a Gibbs sampling algorithm, that partitions the missing data Y_{mis} and the parameters θ into a sequence of p conditional distributions of the form

$$p(\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p, Y_{mis,1}, \ldots, Y_{mis,p}),$$

$$p(Y_{mis,j} \mid Y_{mis,1}, \ldots, Y_{mis,j-1}, Y_{mis,j+1}, \ldots, Y_{mis,p}, \theta_1, \ldots, \theta_p), \tag{1}$$

for $j = 1, \ldots, p$, where p is the number of variables with missing values and θ_j is a vector of parameters in the joint distribution $f(Y \mid \theta_1, \theta_2, \ldots, \theta_p,)$ (e.g. regression coefficients and dispersion parameters). Each conditional distribution is computed based on this joint distribution and values from the conditional distributions are drawn sequentially and iteratively. It can be shown that the sequence converges to a draw from the posterior predictive

[10] For example, if there are two households and three variables:

$$Y = (Y_{obs}, Y_{mis}) = \left(\begin{pmatrix} 2 & 1 & \cdot \\ 4 & \cdot & 3 \end{pmatrix}, \begin{pmatrix} \cdot & \cdot & 2 \\ \cdot & 3 & \cdot \end{pmatrix} \right) \text{ with } M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

distribution of Y_{mis} and a draw from the posterior distribution of θ.[11]

Although this approach is theoretically preferable when the underlying model, $f(Y \mid \theta)$, is well justified, in situations with multivariate data like the HSHW involving nonlinear relationships such as bounds, skip patterns, bracketed responses, interactions, or constraints with other variables, it might be difficult and time-consuming to find a coherent model, program the draws of the conditional distributions, and assess convergence.

The FCS approach is a simpler method that *approximates* draws from the posterior predictive distribution of Y_{mis}. FCS is also known under several other names like stochastic relaxation (Kennickell (1991)), regression switching (van Buuren, Boshuizen, and Knook (1999)), chained equations (van Buuren and Oudshoorn (2000)), or incompatible MCMC (Rubin (2003)).[12] Although it is less formally rigorous than JM, it is easier to implement and yields approximately valid inferences. It may be even more effective, if the assumed model (in JM) is not a good reflection of the data.

Instead of assuming *explicitly* a density function $f(Y \mid \theta)$ for Y, the FCS approach assumes it *implicitly* by explicitly assuming a model (e.g. linear regression or logit regression) for each one of the P conditional distributions of variables with missing values, that relates each variable to a set of other variables. These models are reasonable when taken one at a time, but incoherent in the sense that they might not be derivable from a single joint distribution $f(Y \mid \theta)$ for Y (although *implicitly* assumed). For each one of the modelled variables $Y_{mis,j}$, a draw of parameters (regression coefficients and residual variance) and subsequently of missing data (predictions) is made, the missing data are imputed for that variable, and the procedure cycles through the variables, replacing variables that are being conditioned in any regression by the observed or currently imputed values.

The algorithm is as follows (see also van Buuren and Groothuis-Oudshoorn (2010)). Start with an initial draw $Y_{mis}^{(d,0)}$. These starting values are obtained by randomly drawing from the marginal distribution of Y_{obs}; that is, by filling the incomplete entries of each variable with random draws from its observed values. Given a value $Y_{mis}^{(d,t)}$ of Y_{mis} drawn at iteration t:

$$1.\ \begin{cases} \theta_1^{(d,t+1)} & \sim p\left(\theta_1 \mid Y_{obs}, Y_{mis,2}^{(d,t)}, \cdots, Y_{mis,p}^{(d,t)}\right) \\ \theta_2^{(d,t+1)} & \sim p\left(\theta_2 \mid Y_{obs}, Y_{mis,1}^{(d,t+1)}, Y_{mis,3}^{(d,t)}, \cdots, Y_{mis,p}^{(d,t)}\right) \\ & \vdots \\ \theta_p^{(d,t+1)} & \sim p\left(\theta_p \mid Y_{obs}, Y_{mis,1}^{(d,t+1)}, Y_{mis,2}^{(d,t+1)}, \cdots, Y_{mis,p-1}^{(d,t+1)}\right) \end{cases}$$

$$2.\ \begin{cases} Y_{mis,1}^{(d,t+1)} & \sim p\left(Y_{mis,1} \mid Y_{obs}, Y_{mis,2}^{(d,t)}, \cdots, Y_{mis,p}^{(d,t)}, \theta_1^{(d,t+1)}\right) \\ Y_{mis,2}^{(d,t+1)} & \sim p\left(Y_{mis,2} \mid Y_{obs}, Y_{mis,1}^{(d,t+1)}, Y_{mis,3}^{(d,t)}, \cdots, Y_{mis,p}^{(d,t)}, \theta_2^{(d,t+1)}\right) \\ & \vdots \\ Y_{mis,p}^{(d,t+1)} & \sim p\left(Y_{mis,p} \mid Y_{obs}, Y_{mis,1}^{(d,t+1)}, Y_{mis,2}^{(d,t+1)}, \cdots, Y_{mis,p-1}^{(d,t+1)}, \theta_p^{(d,t+1)}\right) \end{cases}$$

[11] See Schafer (1997) for more details on JM.

[12] See van Buuren and Groothuis-Oudshoorn (2010) for even more names used in the literature.

3. Repeat steps 1 and 2 t times. As t tends to infinity, this sequence is expected to converge to an approximation of a draw from the posterior predictive distribution of Y_{mis} and an approximation of a draw from the posterior distribution of θ;

4. Repeat steps 1-3 D times to obtain D multiple imputations.

Note that for all $j = 1, \ldots, p$ no information about θ or about $Y_{mis,j}$ is used to draw θ_j, and that for all $-j = 1, \ldots, j-1, j+1, \ldots, p$, the θ_{-j} are omitted from the conditional density of $Y_{mis,j}$, which differs from the Gibbs' sampler in JM (see expressions 1). Thus, the FCS approach can be seen as an approximation of the Gibbs sampler. The advantage is that the conditional density of each $Y_{mis,j}$ can now be easily specified by a regression model that depends upon the variable type for $Y_{mis,j}$ (continuous, binary, ordinal or nominal).

As already mentioned, a disadvantage of the FCS approach is that it is less formally rigorous and, therefore, it is theoretically possible that a sequence of draws based on the above conditional densities may not converge to a (implicitly assumed) stationary distribution, because these conditional densities may not be compatible with $p(Y_{mis} \mid Y_{obs})$ or $p(\theta \mid Y_{obs})$ (see Arnold and Press (1989)).[13] However, simulation studies provide evidence that the approach works quite well in many applications and yields estimates that are unbiased (see Van Buuren *et al.* (2006)).

There are different practical implementations of the FCS approach. The software implementation we use is ice (Royston (2004), Royston (2005a), Royston (2005b), Royston (2007), Royston (2009)) in STATA, which is itself an implementation of MICE (van Buuren *et al.* (1999)) in R. A slightly different implementation, but following the same idea, is successfully being used in other wealth surveys, such as the Federal Reserve's SCF, or the Banco de España's EFF. See the Appendix for a comparison of the SCF/EFF imputation algorithm with the HSHW algorithm.

3.2. Specification of the imputation model

Choice of variables to be imputed

A necessary step before starting to build up the imputation model, is the choice of $Y_{mis,1}, \cdots, Y_{mis,p}$, the set of p variables with missing values that are going to be imputed. Depending on ones imputation strategy, this set need not always be equivalent with the set of all variables with missing values in the data set. For example, if the strategy is to only impute a small set of key variables which are most necessary for the future analyses of the data set, Y_{mis} will be a very small subset of all the variables with nonresponse in the data set. Such a strategy might be tempting, because it reduces considerably the size of the imputation model (i.e. the number of regression equations), but it has some important drawbacks, too. First, it might not always be clear which analyses are going to be done in the future. Second, although this strategy reduces the size of the imputation model, it does not necessarily mean that imputation becomes easier. Especially in data sets with high and frequent nonresponse as in the HSHW, the smaller the set of Y_{mis}, the smaller the set of predictors that can be used for imputations, as all those predictors that are not going to be imputed and have missing values on the same observations as the variable we want to impute cannot be used as predictors and must be discarded. Thus,

[13] Rubin (2003) gives an example of incoherent models for which no joint distribution exists.

contrary to the hope of simplicity behind such a strategy, it might even become harder to impute because of the difficulty of finding good predictors. Finally, a further drawback of an imputation strategy aiming at a reduction of the size of the imputation model is that it contradicts points (1) and (2) of the three general imputation requirements by Little and Rubin (2002)) mentioned in the Introduction. These points say that imputations should generally preserve association between (1) missing and observed variables, and (2) missing variables. But by restricting imputations to a small subset of all the variables with nonresponse in the data set, we would violate these requirements because we are excluding missing variables from the regressions and, hence, ignoring their correlations with the included (observed and missing) variables.

For the above reasons our imputation strategy in the HSHW is to impute the biggest possible set of variables Y_{mis}, which in our case consists of $p = 165$ variables out of all the 183 variables with missing values in the data set. We excluded 18 variables from imputation because of their lack of observations which makes it impossible to run a regression due to insufficient degrees of freedom. These variables correspond to items asked only to a very small number of households who had them (e.g. the amounts of the 4th to 9th inherited house, details about the 4th mortgage of the main residence, the purchase price of the hotel or restaurant owned by a household).

Types of models

The next step is to define a regression model for each variable $Y_{mis,j}$ we want to impute. The choice of such a model determines the functional form of the conditional posterior distribution of the regression coefficients and residual variance θ_j (Step 1 in Section 3.1), and the conditional posterior predictive distribution of $Y_{mis,j}$ from which we are going to draw the values used to impute the missing observations (Step 2). For example, if we chose a linear regression model for $Y_{mis,j}$, then $Y_{mis,j}$ would follow a Normal distribution by assumption, and it can be shown that both its posterior predictive distribution and the distribution of θ_j would be Normal.[14]

We choose each regression model depending upon the variable type for $Y_{mis,j}$. There are four basic variable types in our data set: continuous (e.g. income), binary (e.g. gender), ordinal (e.g. education) and nominal (e.g. occupation) variables. The choice of the regression models goes as follows: we use a logit model for the binary variables, an ordered logit model for the ordinal variables and a multinomial logit model for the nominal variables. The fact of using logit and multinomial models to impute ordinal and nominal variables allows us to condition on a wider set of covariates than when using hotdeck, as is done in other wealth surveys, such as the Federal Reserve's SCF, or the Banco de España's EFF.

For the continuous variables we use an interval regression model[15] because all our continuous variables are bounded either from above, or from below, or from both above and below. See Section 3.2.4 for more details on bounds.

In case of continuous variables we usually assume that the regression coefficients distribution is Normal. However, in some of these cases we relax this assumption by doing

[14] Given that the priors are non-informative, as we assume in our imputation model.

[15] The interval regression model is a generalization of the Tobit model to account for censoring from below and/or above. See Cameron and Trivedi (2005).

bootstrapping[16] because otherwise we have convergence problems with the imputations of these variables. Furthermore, this also has the advantage of robustness since the distribution of the regression coefficients is no longer assumed to be multivariate normal. The disadvantage is the cost of a longer computation time. The cases where bootstrapping is used are typically variables with very few observations like, for example, the purchase year and value of the fourth residence owned by a household, or the interest rate of the second outstanding mortgage.

Predictor selection

As mentioned in the Introduction and in the section about the choice of the variables to be imputed, one of the main goals of imputation is to preserve association between missing and observed variables, and also between missing variables. Therefore, when choosing predictors for the imputation model, it is not enough to select the most accurate predictors for each outcome variable. Such an approach could bias the correlation structure between the outcome variable and the excluded variables. Furthermore, ignoring variables that are determinants of non-response of the outcome variable makes the ignorability assumption on which our imputation model relies (see Theoretical Framework) less plausible.

Thus, we choose the number of predictors as large as possible (broad conditioning approach): the more predictors, the lower the bias and the higher the certainty of our imputations. However, there is a limit, of course. In such a large data set as in the HSHW with several hundreds of variables, it is not feasible to include all of them. On the one hand, multicollinearity problems can arise, on the other hand, computational problems. Similarly to van Buuren *et al.* (1999) or Barceló (2006), we adopt the following strategy for selecting predictor variables:

1. Include the variables that are determinants of non-response. These are necessary to satisfy the ignorability assumption, on which our imputation model relies. According to the ignorability assumption, the distribution of the complete data (including the unobserved values) only depends on the observed data, conditional on the determinants of item-nonresponse and other covariates. Determinants of nonresponse are found by inspecting their correlations with the response indicator of the variable to be imputed (see e.g. the logit regression in Section 2). For example, variables included as determinants of nonresponse in the HSHW imputation model are the following: variables describing the household (household income, household size, number of children), variables describing the household members (age, education, sex and occupational status of the household head and partner, whether the person who answered the questions was the household head or not), stratification variables (province, city size), information provided by the interviewers (standard of living, type of neighborhood, type of building, interview atmosphere, number of people participating in the interview, whether documents were used or not as a help to answer some questions, sex of interviewer).

2. In addition, include variables that are very good at predicting and explaining the

[16] Bootstrapping in this context consists in taking bootstrap samples of the non-missing observations and then obtaining the posterior predictive distribution of Y_{mis} by running regressions on these bootstrap samples.

variable of interest we want to impute. This is the classical criterion for predictors and helps to reduce uncertainty of the imputations. These predictors are identified by their correlation with the target variable. Concerning the HSHW data, when the target variables are the outstanding amount of different types of loans, we usually use as predictors the initial loan amount and the years elapsed since the loan was taken, since they turn out to explain a considerable amount of variance in most regressions. Or when we impute the market value of different types of real assets, we usually include their purchase value, the years of ownership of the corresponding asset, and the total value of real estate properties owned by the household (estimated by the household itself).

3. In addition, remove the predictor variables from above that have too many missing values within the subsample of missing observations of the variable to be imputed and substitute them with more complete predictors of these predictors. As a rule of thumb, predictors with percentages of observed cases within this subsample lower than 50 per cent are removed and substituted by more complete predictors. This criterion contributes to make imputations more robust. Typical such predictors of predictors are essential household characteristics like household size, number of children, region, age, employment and marital status of household head)

4. In addition, include all variables that appear in the models that will be applied to the data after imputation. In other words, think about different economic theories that might be tested with the data and include the variables as predictors that are expected to affect or explain according to these theories the variable to be imputed. Failure to do so will tend to bias results of potential users of the data when testing the hypothesis of one particular model. For example, the HSHW data offers information on the parents of the household head, like whether they still live, whether they are/were homeowners, and which education they have/had. This information is used when doing intergenerational transfer analysis[17]. Therefore, we include these variables when imputing the education level of the household head or the value of real estate inheritances of the household, so that we do not bias empirical evidence on intergenerational transfers.

Please note, that many variables in the survey fulfill more than one criterion at the same time, like e.g. income, age, or education.

In all regression models we also include an interaction term and a main effect dummy for each one of the above predictor variables that was not asked to every household to which the variable to be imputed was asked. In these cases, we substitute each such predictor variable with both a dummy indicating whether the question was asked to the household or not (first-order head variable) and an interaction term multiplying the predictor with this dummy. Thus, the interaction term equals the predictor if the household arrived to this question or zero otherwise. In case that a predictor has higher (than first) order head variables, we also include a dummy for each higher order head variable indicating whether the higher order head question was asked to the household or not. Ignoring this type of predictors leads to biased estimates because information concerning certain characteristics of the households would be omitted that determine whether a question is posed to a

[17] See Fessler, Mooslechner, and Schürz (2010) for an analysis of intergenerational transfers in Austria.

household or not. For example, suppose that we want to impute household income using mortgage amount as one of our predictors. While household income was asked to every household in the sample, mortgage amount was not. If for those households who do not have any mortgage we just set mortgage amount to zero, the estimates would be biased because of omitting the information of whether the household has a mortgage or not. This information is the first order head variable of mortgage amount and should be included as a dummy in the regression. But again, having a mortgage or not was not asked to every household, just to homeowners. Thus, we also should include a homeowner dummy in the regression, which is the second order head variable of mortgage amount.

Finally, of course, the number of predictors is restricted by the size of the subsample over which the regression is estimated. In cases where the subsample size is smaller than the number of predictors selected according to the above strategy, we use the Akaike information criterion to choose the subset of predictors which best fits the data, given that each one of the above four predictor categories is still represented in each regression equation. In the rare cases where the sample size is still smaller than the number of predictors, we just choose the predictors with the best fit, without taking into account that each one of the above predictor categories is represented in the equation. Typically, the number of predictors used for each regression model is around 20 percent of the number of observed cases of the variable to be imputed for small subsamples. For large subsamples, the number of predictors usually lies between 5 and 10 percent. For more details on the specification of subsamples, see the corresponding section below.

Bounds

In order to avoid the imputation of values that are either not defined, very unrealistic, or inconsistent with other variables in the survey we impose lower and/or upper bounds on the imputed values of each continuous variable. A useful aid for finding such bounds was provided by the consistency checks done both during the interview and also afterwards during the editing procedure previous to imputation of the HSHW. We use two types of bounds: general bounds that are the same for all households and individual bounds that take different values depending on each household. General bounds are usually employed to avoid imputing values that are not defined or that are very unrealistic. Examples for this type of bounds are non-negativity constraints on quantitative or count variables (income, age). The lower bound for these variables is zero for all households. Furthermore, for each quantitative variable we use the following rule: for every household set the half of the smallest observed value of the variable as the lower bound and the double of the largest observed value as the upper bound. Our aim with this rule is to carefully avoid the imputation of very unrealistic values without manipulating results. More examples for general bounds are share variables (e.g. share of homeownership), where we set the lower bound to zero and the upper bound to 100, or some year variables (e.g. purchase or inheritance year of the real asset owned by the household, year of parents' death), where the upper bound equals 2008, the year when the last interviews of the survey were done.

The second type of bounds, the ones that vary across households, usually ensures consistency with other variables of the same household. Most of the HSHW bounds are of this type. For example, when imputing total household income we set as a lower bound the sum of the different income sources of the household head (personal income of the other household members is not asked in the HSHW). On the other way round when imput-

ing the individual income of the household head, we set as upper bound total household income.[18] Another very useful implementation of individual bounds is done when the household provided information about the range of the value that is missing. In most of the quantitative questions of the HSHW, such ranges where asked when households denied the answer to a question. More examples for individual bounds in the HSHW are when imputing rents (with gross rent as an upper bound for net rent and vice versa), aggregate amounts (e.g. with total housing wealth as an upper bound for house value and vice versa, total inheritances as an upper bound for each individual inheritance and vice versa, initial amount of loan as an upper bound for outstanding amount of loan and vice versa), or when imputing several count variables (e.g. birth year of the oldest household member as a lower bound for year of acquisition of the real asset, age of the household head minus 1 as a lower bound for the age at which his father died, age of the household head as a lower bound for the age at which his mother died). In case that an observation has more than one lower or more than one upper bound (e.g. general and individual bounds) we take the lower and/or upper bound that is most restrictive among all.

Subsamples specification

The subsample over which each regression of the variable we want to impute is estimated simply consists of all the households to which the corresponding question was posed. In particular, when questions were asked separately about each particular asset within an asset type, or each particular loan within a loan type, we estimate each item within a type separately over the subsample of households that have this item. For example, if a household has two mortgages and we want to impute the outstanding amount of the second mortgage, then we impute this missing value by regressing over the subsample of households that have at least two mortgages. If we also included the households that only have one mortgage to impute the second mortgage amounts we would ignore systematic differences between the first and the second mortgages. Especially, we would ignore the fact that the first mortgage is higher than the second one because households order mortgages after their importance, which will introduce a bias in our estimates. Of course, in such a case, we could introduce a lot of interaction terms in our model to reduce the bias, but there still might be unobserved differences between both groups. When imputing question by question, as we do, the bias will be very small, although at the cost of precision, because the sample size will also be very small to condition on a wide set of covariates.

Variable Transformations

Certain transformations of variables in our imputation model turned out to be extremely helpful in terms of improving the plausibility of their imputed values and, hence, of the imputed values in general. This was the case with the logarithmic transformation. We check the distribution of each continuous variable in our model that we want to impute and take the logarithm when the distribution is highly skewed. not range variables! During the imputation procedure we maintain this transformation, even when the variable is used as a predictor for another variable. Only after imputations are finished we transform back the variables into their original measure.

Another very helpful transformation consists in imputing durations instead of years. For

[18] Our imputation model is not able to use imputed values as bounds for imputing other variables. Thus, we cannot set as bounds observations which are missing. In these cases we have to use general bounds as nonnegativity constraints or smallest/largest-observed-value type of bounds instead of individual bounds.

example, instead of imputing the purchase year of the house we impute the time elapsed since the house was purchased. In these cases the above mentioned logarithmic transformation was done on the durations and not on the years and again it is kept even when the variable changes to a predictor during the imputation process.

Last but not least, another transformation that we employ for improving the plausibility of the imputed values is splitting some quantitative variables into head and branch variables if they are not already splitted. For example, suppose that we want to impute income of the household head. The distribution of this variable shows a small peak around zero, because there are some cases where the household head is not working. Therefore, instead of imputing this heterogeneous variable and probably bias the results, it is better to split the variable into a dummy head variable indicating whether the household head has income or not, and a continuous branch variable without zeroes with all the positive income quantities of all the household heads that have income. Subsequently, we first impute the dummy based on a logit regression model and, afterwards, if the household head has been imputed as having income, we impute the continuous income variable based on an interval regression model. Another example where splitting is useful are multiple-response variables: we split each multiple-response category into a dummy variable indicating whether the category applies or not. Then each dummy is imputed separately. However, in most of the cases splitting in the HSHW is not necessary because the survey structure already consists of head and branch variables.

Imputation order

As we mentioned in the Introduction and in the Theoretical Framework, a weakness of the FCS approach is that the conditional densities in step 1 and 2 may not converge to a stationary distribution. In practice, however, choosing a particular ordering of the variables often aid convergence. In the HSHW we start imputation by the variables with the least missing values, and so on. Variables with the same amount of missingness are processed in an arbitrary order, but always in the same order. The imputation order of head variables is not arbitrary and is done always before their corresponding branch variables. For example, whether the household has a mortgage or not is always imputed before imputing the mortgage amount, even if missingness is the same for both variables.

Number of iterations

The number of iterations t determines how often the imputation procedure cycles through the variables to be imputed, replacing variables that are being conditioned in any regression by the observed or currently imputed values. As t tends to infinity, the sequence of parameters and predicted values should converge to a draw from the posterior distribution of θ and a draw from the posterior predictive distribution of Y_{mis}. However, according to van Buuren *et al.* (1999) in practice convergence in these models usually occurs very fast during the first few iterations. This is because the posterior distributions of the regression coefficients already absorb a lot of uncertainty in the predictors and because the procedure creates imputations that are already statistically independent. Given the large computational effort required for the HSHW imputation model and following the number of iterations used in other similar surveys (like SCF (Kennickell (1991) or EFF (Barceló (2006))), we set the iteration number for the HSHW imputation model to $t = 6$.

Typically, we graphically check convergence by plotting the mean of the imputed values

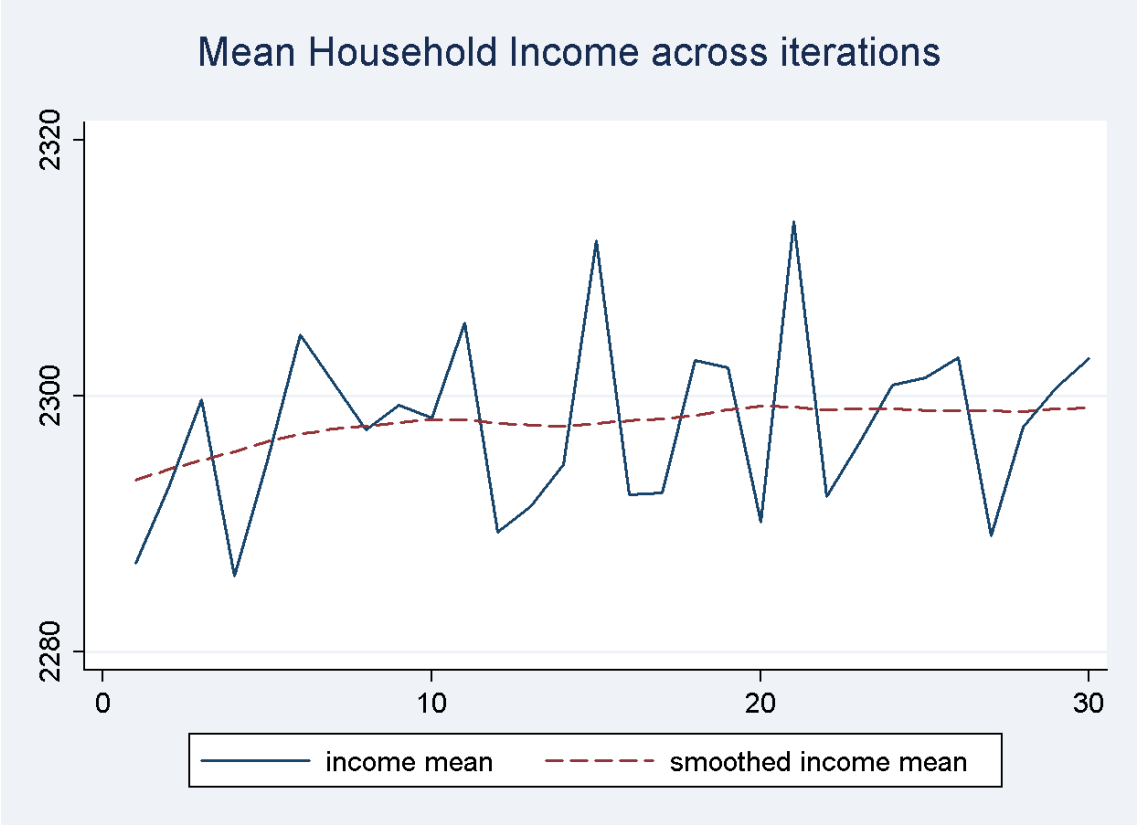

Figure 1: Monitoring the convergence of imputations of household income in the HSHW

against the iteration number t. As an example, Figure 1 shows this plot for the income variable. Convergence is judged to have occurred as soon as the pattern of the imputed means turns to be random. In Figure 1 this seems to be the case very soon: at the latest from the fifth iteration forward no trend in the smoothed curve of the imputed means of income can be recognized any more. Furthermore, Figure 1 shows that the fluctuation range of the imputed means is around 20 euros and, thus, very small, which is a further indicator of convergence. Of course, these kind of checks can never confirm convergence (like any other check in the FCS approach), but they can highlight weaknesses in the imputation model or other unusual outcomes that could be indicators of non-convergence.

Number of imputations

Finally, we choose the number of realizations D that we want to have from the posterior predictive distribution $p\left(Y_{mis} \mid Y_{obs}\right)$ or, in other words, the number of multiply imputed data sets. Setting D too low leads to standard errors of the estimates that are too low and to p-values that are too low. Schafer and Olsen (1998) show that the gains of efficiency of an estimate rapidly diminish after the first few D imputations. They claim that good inferences can already be made with $D = 3$ to 5. However, Graham, Olchowski, and Gilreath (2007) show that another important quantity such as statistical power can vary more dramatically with D than is implied by efficiency. They claim that good inferences can be made with $D = 20$ to 40. It seems unlikely that a single correct value for D will be established in the literature because, like sample size, the number of imputation that are necessary depends on features of the individual data set and analysis model. In the HSHW imputation model, given the substantial increase in computational effort for every

further imputation and following other similar surveys like the SCF or EFF we set the number of imputations to $D = 5$.

4. Some results

Estimating the HSHW imputation model with the above presented specifications is computationally very intensive and takes around 8 days.[19] Table 2 summarizes the resulting imputations for some variables, similarly as Kennickell (1991) and Kennickell (1998b) does.

The first two columns show the weighted sum of all imputed values of a given item in percent of the weighted sum of all values of this item, distinguishing between imputations that used range information and imputations that did not. For example, 64.8 percent of the total amount of real estate in the sample is imputed, with 29.1 percentage points of that amount imputed using range information provided by households. In case of total income, 36.5 percent of euros are imputed with 23.5 percentage points based on ranges. In most of the other cases reported, the proportion of the total value imputed based on ranges is higher than for completely missing variables which clearly shows that ranges provide very valuable information that greatly helps in improving the precision of imputations. The most extreme case is the tenants' deposit to the household association, where 54.8 percent of total euros are imputed with an amazing 50.9 percentage points constrained by range estimates.[20] The reported variable with most missing information due to nonresponse is the outstanding mortgage amount for the primary residence with 72.9 percent euros imputed.

The rest of the columns in Table 2 display the coefficients of variation (CV) for the mean and median values of the imputations.[21] These help us to measure the performance of our imputation model. The CV describes the precision of the estimated mean and median values due to imputation in a way that does not depend on the variable's measurement unit. The higher the CV, the lower the precision of the estimate. For comparison, coefficients of variation of the observed sample and of the complete sample are also provided. Table 2 shows that the model performs better in predicting higher order aggregated variables than individual assets. For example, while the variation for current value of other real estate is 145 percent, the variation in total household income is only 3.3 percent. Of course, the reason is the smaller sample size in case of individual assets. In most cases, the CV of the imputations is higher than the CV in the observed sample, what makes sense since, in general, imputations are probably less accurate than observed values. Finally, the median based CV are almost always higher than the mean based CV which reflects the fact that the distributions of the variables are very skewed and that the euro amounts are concentrated in small groups of households.

5. Conclusions

[19] We use a computer with 3.4 GHz CPU and 1 GB RAM.

[20] Another reason why the proportion of the total value imputed based on ranges is higher than that for completely missing observations is that the fraction of nonrespondent households giving a range is higher than the fraction of completely nonrespondent households. This is particularly important in the case of tenants' deposit to the housing association (27.8 percent vs 4.2 percent in Table 1).

[21] $CV_{S_{\overline{x}}}(\%) = \frac{S_{\overline{x}}}{\overline{x}} \cdot 100$, where $S_{\overline{x}}$ is the standard error of the mean $\overline{x}$, and $CV_{S_{\tilde{x}}}(\%) = \frac{S_{\tilde{x}}}{\tilde{x}} \cdot 100$, where $S_{\tilde{x}}$ is the standard error of the median $\tilde{x}$. The standard error of the median is approximated as $S_{\tilde{x}} = 1.253 \cdot S_{\overline{x}}$.

Table 2: Performance of imputations in selected items of the HSHW (weighted)

Item	% of total value imputed with range	% of total value imputed without range	Coefficients of variation in % (relative standard errors) Imputed sample Mean	Imputed sample Median	Observed sample Mean	Observed sample Median	Complete sample Mean	Complete sample Median
HOUSING WEALTH								
Current value of the primary residence	11.7	18.0	7.6	9.7	4.6	5.8	3.9	5.3
Current value of the first additional house/condominium	0.0	23.0	26.6	44.1	15.1	27.0	13.3	22.8
Current value of the second additional house/condominium	0.0	35.2	29.5	39.6	16.0	17.1	14.5	16.4
Current value of plots of land/building lot	0.0	21.5	59.9	192.3	12.0	26.6	16.0	38.4
Current value of agricultural or forestry real estate, fields, forest etc.	0.0	22.1	39.8	112.7	19.1	78.7	17.2	58.8
Current value of office, business premises, company site	0.0	50.8	46.8	58.3	16.7	18.7	25.5	33.5
Current value of other real estate	0.0	40.5	145.0	380.4	50.3	236.1	65.3	208.7
Estimated total value of real estate	29.1	35.7	13.6	34.6	7.5	10.7	9.7	23.0
HOUSING DEBT								
Outstanding amount of loans for acquiring the primary residence	36.9	36.0	7.0	9.3	8.8	14.3	5.8	8.6
Outstanding amount of loans for acquiring additional houses	0.0	53.6	43.4	116.4	26.0	36.6	26.4	60.4
Outstanding amount of loans for acquiring plots of land/building lots	0.0	45.3	24.0	36.0	41.7	55.3	32.3	62.5
Annual repayment for acquiring the primary residence	39.3	21.2	9.7	13.5	5.7	7.1	6.5	8.5
INTERGENERATIONAL TRANSFERS								
Value of the properties given away as a gift	0.0	39.4	43.0	75.6	15.0	24.4	19.5	34.5
Value of the properties inherited	0.0	38.0	29.8	101.5	13.1	32.5	13.9	38.4
OTHER CHARACTERISTICS OF THE HOUSEHOLD								
Total monthly net household income	23.5	13.0	3.3	4.1	2.6	3.3	2.1	2.6
Household head's monthly net income from paid employment	15.4	11.0	3.6	3.6	1.7	1.6	1.5	1.5
Homeowners' imputed rent	11.8	14.1	3.9	4.1	1.8	2.1	1.7	1.9
Tenants' monthly rent paid	13.5	9.1	3.7	4.2	1.9	2.1	1.7	1.8
Tenants' deposit to the housing association (Genossenschaftsbeitrag)	50.9	3.9	7.9	6.4	8.2	17.0	6.4	13.1
Annual total of rental or leasing income	0.0	21.1	108.0	825.3	18.0	41.2	27.3	73.1
Time since the household head's father passed away	0.0	11.8	7.4	9.2	2.5	3.0	2.4	2.8
Time since the household head's mother passed away	0.0	11.6	12.5	15.6	3.3	4.1	3.2	3.9

This paper presents the HSHW multiple imputation model and its implementation. After justifying the choice of the fully conditional specification approach in the context of several other missing data methods, we show that nonresponse in the HSHW is not random and that it fluctuates a lot depending on the question posed.

We then present the theoretical framework of the model and subsequently its specifications. In comparison with other imputation models of similar surveys like the SCF or the EFF, our implementation allows to impute ordinal variables using an ordered logit model and nominal variables using a multinomial logit model and, thus, to condition on many more variables than when using hot deck for the imputation of such variables.

Finally, we summarize the resulting imputations for various items by using two statistics: the proportions of euros imputed and the relative standard errors of imputations which both try to measure the performance of the imputations in terms of their precision. We see that higher order aggregate variables and ranges improve a lot the precision of our imputations, but there are still some cases, especially some individual asset categories, where the reliability of imputations is rather low. However, in such cases we prefer to have the cost of a small increase in variance for less bias in order to avoid distorted results being wrongly considered significant too often.

Nevertheless, to improve the reliability of imputations of such variables, we could increase the number of imputations, but, of course, at the cost of a higher computational effort. We hope to be able to reduce this cost for the imputation of the upcoming Austrian Household Finance and Consumption Survey by having improved our technological resources by then. Another way to improve imputations of such variables is, of course, to introduce possibilites to reduce their item nonresponse. For example, by incentivating range answers when no exact amount is provided in a euro question. This could be done by additionally allowing the household to indicate individual ranges, additionally to the possibility of choosing a predefined range.

One interesting analysis that goes beyond the purpose of this paper, but is left for future research, is to evaluate our imputations in more depth by developing additional evaluation criteria like, for example, distributional or bias criteria and then comparing them with other imputation methods with the help of simulations.

Appendices

A. Ignorable nonresponse

The missing-data mechanism is ignorable for Bayesian inference if:

1. the missing data are missing at random (MAR): $f\left(M \mid Y_{obs}, Y_{mis}, \psi\right) = p\left(M \mid Y_{obs}, \psi\right)$ for all Y_{mis}; and
2. the parameters θ and ψ are *a priori* independent, that is, the prior distribution has the form $p\left(\theta, \psi\right) = p\left(\theta\right) p\left(\psi\right)$.

B. Comparison of imputation algorithms

The imputation of the Federal Reserve's SCF or Banco de España's EFF and the one of the Oesterreichische Nationalbank's HSHW are based on the same approach. Neither SCF/EFF nor HSHW specify *explicitly* a joint distribution of the data, but they do it *implicitly* by specifying separately the conditional distribution of each variable having missing values. Both implementations are less formally rigorous than the joint modelling approach, but they are easier to implement and much more flexible, being able to account for the numerous nonlinear relationships in the data.

However, there are still some differences between the algorithm of the implementation of this imputation approach in the SCF/EFF and the one in the HSHW:

Starting values. The starting values for the first iteration are different. In the HSHW, a draw of $Y_{mis}^{(0)}$ is needed and is obtained by filling the incomplete entries of each variable with random draws from its observed values. In the SCF/EFF, a draw of $\theta^{(0)}$ is needed and is obtained by sequentially estimating the imputation model of each variable, using the subsample of both observed data and the values of the missing data previously imputed within the first iteration.

Order of steps. As a consequence of the different starting values in the two implementations, a different order of the imputation step and the posterior step is needed, too. In the HSHW implementation, within each iteration, first the parameters θ_j are drawn and then, conditional on them, the missing values $Y_{mis,j}$ are drawn. In the SCF/EFF it is the other way round: first the missing values $Y_{mis,j}$ are drawn and then given these values the corresponding parameters θ_j are drawn. Both algorithms should be equivalent, because in the limit, the order of the sequences should not matter.

Posterior step. While the HSHW implementation does not take into account missing information of the outcome variable $Y_{mis,j}$ to estimate the parameters θ_j of their own imputation models, the SCF/EFF implementation *does*.

Imputation step. Unlike the HSHW implementation, the SCF/EFF implementation do not use values $Y_{mis,1}^{(t-1)}, \ldots, Y_{mis,j-1}^{(t-1)}, Y_{mis,j+1}^{(t-1)}, \ldots, Y_{mis,p}^{(t-1)}$ imputed in the previous iteration of the imputation process to reimpute missing values $Y_{mis,j}^{(t)}$ in the current iteration of the imputation step, but just use observed values and values $Y_{mis,1}^{(t)}, \ldots, Y_{mis,j-1}^{(t)}$ imputed so far in the current iteration.

C. Probit regression

Table C.1: Determinants of nonresponse on value of primary residence

Variables	Coeff. (SE)
OWNER'S CHARACTERISTICS	
Female	0.381**
	(0.193)
Age	-0.0894**
	(0.0363)
Age squared	0.000772**
	(0.000326)
Highest educational level completed	
Apprenticeship, vocational school/Intermediate or higher technical/vocational school	-0.311
	(0.240)
High school (Matura)	-0.750**
	(0.343)
College, university, university of applied sciences, academy	-0.897**
	(0.382)
Occupational status	
White-collar worker	0.240
	(0.358)
Civil servant	-0.142
	(0.499)
Farmer	1.292**
	(0.512)
Blue-collar worker	0.336
	(0.402)
Other occupation	0.394
	(0.530)
Retired	0.294
	(0.394)
Out of labor force	0.547
	(0.433)
HOUSEHOLD'S CHARACTERISTICS	
Number of children in household	-0.148
	(0.117)
Number of adults in household	-0.0927
	(0.111)
Household has to debt service some housing loan	-0.477*
	(0.261)
Spouse/partner in household	-0.0647
	(0.396)
INTERVIEWER'S ASSESSMENT	
Size of municipality	
Up to 5,000 inhabitants	0.131
	(0.213)
Up to 20,000 inhabitants	-0.382
	(0.267)
Up to 50,000 inhabitants	0.171
	(0.457)
More than 50,000 inhabitants	0.922**
	(0.425)
Impression of apartment/house	
Good, medium standard of living	-0.681***
	(0.189)
Rather basic standard of living/Poor standard of living	-0.557**
	(0.275)
Unpleasant atmosphere during the interview	1.398***
	(0.322)
Number of persons who provided information during the interview	-0.0617
	(0.186)
No documents consulted during interview	0.703***
	(0.216)
Questions were not answered honestly and seriously	0.586
	(0.544)
Female interviewer	0.832***
	(0.174)
Constant	0.402
	(1.296)
Observations	1,085

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Note: Other variables included in the regression are dummies for the marital status of the owner, being the household head, province, neighborhood and type of building.

References

Arnold BC, Press SJ (1989). "Compatible Conditional Distributions." *Journal of the American Statistical Association*, **Vol. 84 No. 405**, 152–156.

Baccini M, Cook S, Frangakis C, Li F, Mealli F, Rubin D, Zell E (2010). "Multiple imputation in the anthrax vaccine research program." *CHANCE*, **23**, 16–23. ISSN 0933-2480. 10.1007/s00144-010-0004-3, URL http://dx.doi.org/10.1007/s00144-010-0004-3.

Barceló C (2006). "Imputation of the 2002 wave of the Spanish survey of household finances (EFF)." *Banco de España Occasional Papers 0603*, Banco de España. URL http://ideas.repec.org/p/bde/opaper/0603.html.

Cameron A, Trivedi P (2005). *Microeconometrics: methods and applications*. Cambridge University Press. ISBN 9780521848053. URL http://books.google.at/books?id=Zf0gCwxC9ocC.

Fessler P, Mooslechner P, Schürz M (2010). "Real Estate Inheritance in Austria." *Monetary Policy & the Economy*, (2), 33–53. URL http://ideas.repec.org/a/onb/oenbmp/y2010i2b2.html.

Fessler P, Mooslechner P, Schürz M, Wagner K (2009). "Housing Wealth of Austrian Households." *Monetary Policy & the Economy*, (2), 104–124. URL http://econpapers.repec.org/RePEc:onb:oenbmp:y:2009:i:2:b:5.

Graham J, Olchowski A, Gilreath T (2007). "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science*, **8**, 206–213. ISSN 1389-4986. 10.1007/s11121-007-0070-9, URL http://dx.doi.org/10.1007/s11121-007-0070-9.

Kennickell AB (1991). "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple ImputationT̆ mimeo, Board of Governors of the Federal Reserve System." In *1991 Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association*.

Kennickell AB (1998a). "Analysis of Nonresponse Effects in the 1995 Survey of Consumer Finances." In *Prepared for the 1997 Joint Statistical Meetings, Anaheim, CA*.

Kennickell AB (1998b). "Multiple Imputation in the Survey of Consumer Finances." In *Proceedings of the Section on Business and Economic Statistics, 1998 Annual Meetings of the American Statistical Association*, pp. 63–74.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data, Second Edition*. 2 edition. Wiley-Interscience. ISBN 0471183865. URL http://www.worldcat.org/isbn/0471183865.

Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology*, **vol. 27 no. 1**, 85–95.

Royston P (2004). "Multiple imputation of missing values." *Stata Journal*, **4**(3), 227–241. URL http://ideas.repec.org/a/tsj/stataj/v4y2004i3p227-241.html.

Royston P (2005a). "Multiple imputation of missing values: update." *Stata Journal*, **5**(2), 188–201. URL http://ideas.repec.org/a/tsj/stataj/v5y2005i2p188-201.html.

Royston P (2005b). "Multiple imputation of missing values: Update of ice." *Stata Journal*, **5**(4), 527–536. URL http://ideas.repec.org/a/tsj/stataj/v5y2005i4p527-536.html.

Royston P (2007). "Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring." *Stata Journal*, **7**(4), 445–464. URL http://ideas.repec.org/a/tsj/stataj/v7y2007i4p445-464.html.

Royston P (2009). "Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables." *Stata Journal*, **9**(3), 466–477. URL http://ideas.repec.org/a/tsj/stataj/v9y2009i3p466-477.html.

Rubin DB (1974). "Characterizing the Estimation of Parameters in Incomplete-Data Problems." *Journal of the American Statistical Association*, **Vol. 69, No. 346**, pp. 467–474.

Rubin DB (2003). "Nested multiple imputation of NMES via partially incompatible MCMC." *Statistica Neerlandica*, **57**(1), 3–18. URL http://ideas.repec.org/a/bla/stanee/v57y2003i1p3-18.html.

Schafer J (1997). *Analysis of incomplete multivariate data*. Monographs on statistics and applied probability. Chapman & Hall. ISBN 9780412040610. URL http://books.google.com/books?id=3TFWRjn1f-oC.

Schafer JL, Olsen MK (1998). "Multiple imputation for multivariate missing-data problems: a data analyst's perspective." *Multivariate Behavioral Research*, **33**, 545–571.

van Buuren S, Boshuizen HC, Knook DL (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." *Statistics in Medicine*, **18**, 681Ű694.

Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB (2006). "Fully conditional specification in multivariate imputation." *Journal of Statistical Computation and Simulation*, **76**, 1049–1064(16). doi:doi:10.1080/10629360600810434. URL http://www.ingentaconnect.com/content/tandf/gscs/2006/00000076/00000012/art00002.

van Buuren S, Groothuis-Oudshoorn K (2010). "MICE: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **forthcoming**.

van Buuren S, Oudshoorn KC (2000). *Multivariate Imputation by Chained Equations. MICE V1.0 User's manual*. TNO Prevention and Health, http://web.inter.nl.net/users/S.van.Buuren/mi/docs/Manual.pdf (07.10.2005).

Wagner K, Zottel S (2009). "Methodologische Aspekte der Immobilienvermögenserhebung 2008." *Statistiken*, **Q4/09**, Wien: OeNB.

Affiliation:

Nicolás Albacete
Economic Analysis Division
Oesterreichische Nationalbank
A-1090 Vienna, Austria
E-mail: nicolas.albacete@oenb.at

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43
April 2014

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2013-12-01
Accepted: 2014-03-03