UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert meeting on Statistical Data Editing
3-7 October 2022, (virtual)

Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions

BIANCHI Blandine, Swiss Federal Statistical Office, Espace de l'Europe 10, 2010 Neuchâtel

blandine.bianchi@bfs.admin.ch

I. SUMMARY

1. The Survey on Income and Living Conditions (SILC) is a yearly household survey. The household and its members, based on a random sample, are yearly interviewed and are followed for four years for longitudinal analysis. Unsurprisingly, non-response is a key issue and an appropriate treatment is crucial to improve the quality of the survey and increase the reliability of the published results. Among other methods, we tested the MissForest algorithm, a non-parametric method based on random forests to impute missing values that allows the use of mixed data types (categorical and quantitative). We apply the MissForest algorithm to different datasets containing about 10000 persons and several variables of interest, having up to 20% of missing values and multiple auxiliary variables. We use a simulation framework to evaluate the performance of MissForest thus allowing estimating the imputation accuracy and error.
In Section II we present the issues associated with the analysis of missing data and introduce the MissForest approach for data imputation. In Section III we present SILC data used in this study, while in section IV we describe the simulation setup and our validation framework. In Section V we show the results of the imputation with the MissForest algorithm and evaluate their accuracy and robustness. In Section VI we discuss the potential and limitations of our imputation framework using MissForest applied to SILC data.

II. INTRODUCTION

A. Objective

1. One of the major issues in surveys is the risk of bias due to non-response Lohr [2009]. The non-response can be partial, when not all variables are missing, or complete when the whole unit is missing total otherwise. In this paper we show the application of a non-parametric algorithm called MissForest to impute missing values Stekhoven and Bühlmann [2012] with the aim to reduce potential non-response bias and to facilitate analysis of the data. MissForest is a Machine Learning algorithm that allows categorical variables and quantitative variables to be processed simultaneously, and that does not make specific assumptions about the structure and distribution of the data. Although MissForest provides Out-of-the-bag (OOB) error estimates to assess the accuracy of the imputations (obtained

by computing the normalized root mean squared error), we further validated the performance of this algorithm through simulation. The simulation showed that the errors in the imputed variables ranged between 5.69% and 0.68%. Through an additional impact analysis we also showed that the imputation had a minor impact on the distribution of the original data and that the algorithm was efficient in terms of the required computational resources.

B. MissForest Algorithm

1. The MissForest algorithm, developed by Stekhoven and Bühlmann [2012], uses a Machine Learning algorithm called Random Forest Breiman [2001] to make predictions or to impute missing data. In an initialization step, the algorithm replaces the missing values with the most frequent value for the categorical variables and with the average for continuous and discrete variables.
The imputation procedure is performed sequentially starting with the variable with fewest missing values. At each iteration a random forest is trained on the available observations and used to make predictions for the missing values. A series of random forests are generated until the stopping criterion is met or a maximum number of iterations is reached.
Formally, we define a matrix M $((m+n)\times p)$, with $m+n$ observations and p variables. We indicate with Y_k with $k \in 1,...,p$ the variables containing missing values. The initial dataset is split into four parts:

1) We denote $\mathbf{y}_k^{obs}$ the vector of observed values of the variable to be imputed $\mathbf{Y}_k$.
2) We denote $\mathbf{y}_k^{mis}$ the vector of missing values of the variable to be imputed $\mathbf{Y}_k$.
3) We denote $\mathbf{X}_k^{obs}$ the matrix of the variables $\mathbf{X}_i$ with $i = 1,...,k-1,k+1,...,p$ without missing values in $\mathbf{Y}_k$.
4) We denote $\mathbf{X}_k^{mis}$, same as $\mathbf{X}_k^{obs}$, for observations with missing values in Y_k.

$\mathbf{X}_1$	$\mathbf{X}_2$	...	$\mathbf{X}_p$	$\mathbf{Y}_k$
x_{11}^{obs}	x_{11}^{obs}	...	x_{1p}^{obs}	y_{1k}^{obs}
x_{21}^{obs}	x_{22}^{obs}	...	x_{2p}^{obs}	y_{2k}^{obs}
...	...	...	...	
...	...	...	...	
x_{m1}^{obs}	x_{m2}^{obs}	...	x_{mp}^{obs}	y_{mk}^{obs}
x_{11}^{mis}	x_{21}^{mis}	...	x_{1p}^{mis}	y_{1k}^{mis}
...	...	...	...	
x_{n1}^{mis}	x_{n2}^{mis}	...	x_{np}^{mis}	y_{nk}^{mis}

TABLE 1. Example of matrix with observed values and missing values.

With the Random Forest we aim to find f so that

$$\left(\mathbf{y}_k^{obs} - f(\mathbf{x}_k^{obs})\right)^2 = 0 \tag{1}$$

and we predict the $\mathbf{y}_k^{mis}$ with $f(\mathbf{x}_k^{mis})$.

III. DATA

1. At the Swiss Federal Statistical Office the unit income, consumption and living conditions (EKL) is responsible for the Survey on Income and Living Conditions (SILC) with its module on material and social deprivation. Thanks to auxiliary variables (both categorical and continuous) and the use of missforest to impute missing data we were able to keep all units of the survey even when the non-response was complete.
The EKL unit provided the SILC net sample for 2019 covering 7341 households and 16662 persons. The list of variables concerning material and social deprivation to be imputed is provided in Table 2.

Person variables	
PD020	Replace worn-out clothes by some new (not second-hand) ones
PD030	Two pairs of properly fitting shoes (including a pair of allweather shoes)
PD050	Get-together with friends/family (relatives) for a drink/meal at least once a month
PD060	Regularly participate in a leisure activity such as sport, cinema, concert
PD070	Spend a small amount of money each week on yourself
PD080	Internet connection for personal use at home
Household variables	
HD080	Replacing worn-out furniture
HS011	Arrears on mortgage or rent payments
HS021	Arrears on utility bills
HS031	Arrears on hire purchase instalments or other loan payments
HS040	Capacity to afford paying for one week annual holiday away from home
HS050	Capacity to afford a meal with meat, fish (or equivalent) every second day
HS060	Capacity to face unexpected financial expenses
HS110	Do you have a car?
HH050	Ability to keep home adequately warm

TABLE 2. List of person and household variables concerning material and social deprivation to be imputed.

2. The variables to be imputed were categorical $(1, 2, 3$ or $1, 2)$. For instance the variable HD080 containing the answer to the question if the household could afford to replace worn-out furniture: 1 yes, I can afford it, 2 no, I cannot afford it, 3 no, I cannot afford it, but for non-finantial reasons. Variables PB020, PD030, PD050, PD060, PD070, PD080, HD080, HS011, HS021, HS031, and HHS110 had the same three kind of modalities adapted to the question referred to. The modalities for variables having only two modalities (HS040, HS050, HS060, HHS110 and HH050) were: 1 yes, I can afford it and 2 no, I cannot afford it. As shown in Table 4 and Table 5 the non-response for households variables is almost negligible while for the person variables the complete non-response is around 18%. The need of auxiliary variables for their imputation is crucial. Here, we used 72 auxiliary variables for the imputation: some socio-demographic variables of the person like age, sex, nationality, marital status, children, to list some of them, but also some variables related to satisfaction with the health status, financial status, hobbies, or political orientation, education, social security benefits and the surface of the lodging. These variables improved significantly the quality of imputation as shown in Section 5.

Variable	Number of Units	Number of units requested to respond	Units with missing values	Values to impute [%]
PD020	16662	13861	2552	18.48 %
PD030	16662	13861	2553	18.42 %
PD050	16662	13861	2563	18.49 %
PD060	16662	13861	2565	18.51 %
PD070	16662	13861	2559	18.46 %
PD080	16662	13861	2556	18.44 %
HD080	7341	7341	29	0.40 %
HS011	7341	6867	26	0.00 %
HS021	7341	7341	13	0.18 %
HS031	7341	1354	12	0.01 %
HS040	7341	7341	16	0.22 %
HS050	7341	7341	7	0.10 %
HS060	7341	7341	28	0.38 %
HS110	7341	7341	3	0.04 %
HH050	7341	7341	10	0.14 %

TABLE 3. List of variables of the module on material and social deprivation to impute. Person variables form the first block, and household variables the second one. The definitions of the acronymes are provided in Table 2. For the person variables the non-response is almost complete, that is the full units are missing.

	Observations	[%]
Complete answer	11267	81.29%
Partial non-response	42	0.02%
Complete non-response	2552	18.41 %

TABLE 4. Analysis of non-response for person variables.

	Observations	[%]
Complete answer	7218	98.32%
Partial non-response	112	1.66%
Complete non-response	1	0.01 %

TABLE 5. Analysis of non-response for household variables, used as auxiliares variables.

IV. SIMULATION FRAMEWORK

1. To validate the MissForest algorithm with our data we applied a classical simulation procedure for the person variables, whereas no simulation has been done for the household variables due to their limited rate of missings:

a) For each imputation variable, consider only its respondents.
b) Missing values were then randomly simulated for all variables jointly (about 18%, see Table 3).
c) Perform the imputation of the artificially created missing values with MissForest.
d) The imputed values are compared with the original values to quantify the error. There are not many missing values for the household variables and in its simulated missing values were correctly imputed.

2. For the imputation of the person variables of material and social deprivation we additionally used the household variables as auxiliary variables. The person variables include many more missing values, 18.4% of the net sample never responded to any of the questions. To reproduce the observed frequency of missing values

in the data, we simulated 18.4% of missing values (randomly) for all person material and social deprivation variables (2073 in total, excluding the 42 observations that had an incomplete missingness pattern) among the 11267 observations where all questions on material and social deprivation had been answered. We then imputed the missing values with MissForest using the person and household material and social deprivation variables. For the imputation we used 300 trees and setting the maximum number of iterations to 100, but the models were generally optimized more rapidly meeting the stopping criterion after four to five iterations.

V. RESULTS

A. Validation

1. The person variables are significantly better imputed if additional auxiliary variables are included with an imputation error dropping from more than 20% to around 5% Table 6. Table 6 summarizes the imputation errors of the person variables with the auxiliary variables: (1) only with household variables, (2) with household and socio-demographic variables (3) with household, socio-demographic and historical variables. On average the addition of all auxiliares variables improved the accuracy of the imputations by 10%.

2. At the request of the EKL unit, the additional historical variables from previous years were added to the model, as the information from the current year can be highly correlated to the information from previous years. Indeed, using this approach strongly increased the imputation accuracy. A constant has been attributed to units not in the sample the previous years or which did not respond in the past. The inclusion of historical variables was carried out in two stages: firstly, behond the household and socio-demografic variables, only the individual historical variables were used. In a second time also the historical household variables were included. The personal historical variables brought a significant improvement to the model. Figure 1 show the frequencies of original and imputed variables using historical variables. The imputation did not create any new combinations.

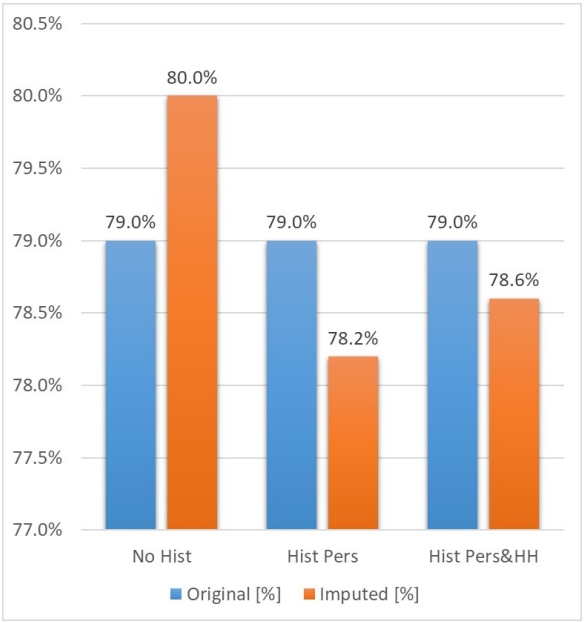

FIGURE 1. Frequencies of original (blue) and imputed missing values (orange) of the class '111111', corresponding to no material and social deprivation, by the amount of auxiliary variables used. Classes 111111 and those containing less than two of the persons. The imputation did not create any new combinations and well represents the distribution of the actual values.

Auxiliary variables: (1) only household variables

Variable imputed	[%] Total Error	[%] Main Error	N. wrongly imputed
PD020	24.17	23.83	494
PD030	4.78	4.73	98
PD050	26.05	25.18	522
PD060	26.39	24.51	508
PD070	23.64	22.72	471
PD080	17.70	17.51	363

(2) household and socio-demografic variables

Variable imputed	[%] Total Error	[%] Main Error	N. wrongly imputed
PD020	2.89	2.89	60
PD030	0.92	0.92	19
PD050	6.03	5.93	123
PD060	12.11	11.38	236
PD070	6.17	6.13	127
PD080	6.03	5.93	104

(3) household, socio-demografic and historical variables

Variable imputed	[%] Total Error	[%] Main Error	N. wrongly imputed
PD020	1.16	1.06	22
PD030	0.87	0.87	18
PD050	2.75	2.46	51
PD060	5.69	4.97	103
PD070	2.12	2.03	42
PD080	0.68	0.63	13

TABLE 6. Imputation errors for person variables with auxiliary variables: with the household variables (top left), with additional socio-demografic variables (top right) and with additional historical variables (bottom). The column "Main Error" and "N.wrongly imputed" is restricted to errors between "yes or no for other reason" and "no for financial reason". The column "Main Error" and "N.wrongly imputed" is restricted to errors between "yes" or "no for other reason" and "no for financial reason".

B. Imputation Impact Analysis

1. In the SILC 2019 dataset there are 13861 observations supposed to answer the material and social deprivation questions at the person level of which 2594 observations have at least one missing value in these variables (2552 are completly missing). To explore the potenial impact that the imputations would have on the final dataset, we compared the distribution of the unmodified units with the distribution of the imputed values. Figures 2 and 3 show the distributions of the actual values without missing values (in blue) and after imputations (in grey). In Figures 2 and 3 we see from the joint distribution of the individual material and social deprivation variables that the impact due to the imputation is limited.

VI. CONCLUSIONS

1. For the individual material and social deprivation variables the partial non-response is quite high leading to a important challenge to reduce the risk of bias in the results. In this paper we show that MissForest provides

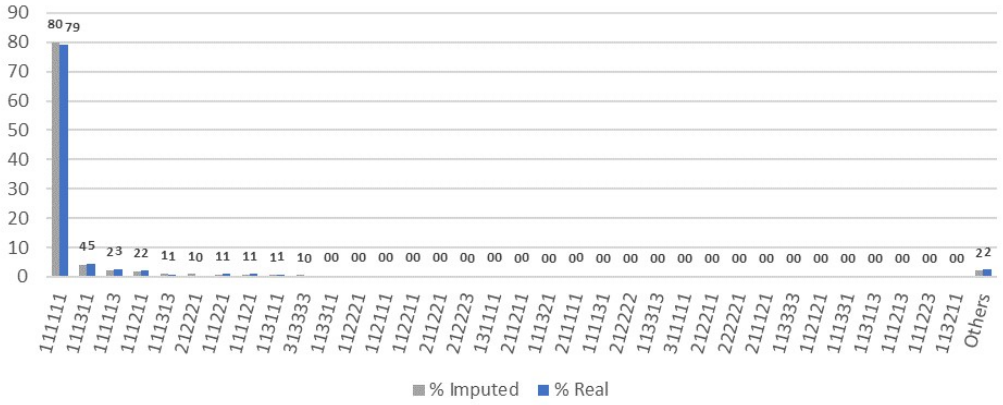

FIGURE 2. Joint relative frequency of original (blue) and imputed values (grey), using the additional auxiliary variables. The "Others" class contains the 2% of the observations (272 persons before imputation and 285 after imputations for 114 different response combinations).

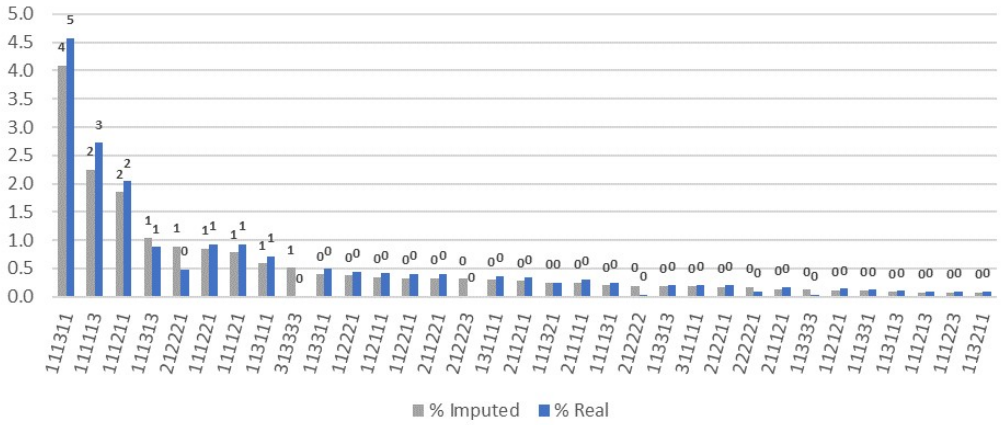

FIGURE 3. Frequencies of Figure 2 without the categories "111111" and "Others".

accurate imputation of the non-response in the variables of material and social deprivation of the SILC. This algorithm is non-parametric and can handle different data types (continuous and categorial) simultaneously, without making any assumptions about the structure and distribution of the data. A simulation study allowed the assessment of the accuracy of the imputation procedure in this contexte. The errors range from 0.92% for the variable PD030 (Two pairs of shoes of the appropriate size) to 12.11% for the variable PD060 (Regular participation in a leisure activity such as sport, cinema, concert) using household and socio-democgraphic variables. The imputation impact remains limited. The use of historical variables strongly improved the imputation quality, with errors ranging from 0.9% to 5.7%. Nevertheless, these findings remain to be verified on the basis of the indicators calculated on the basis of these variables by EKL.

References

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Sharon Lohr. *Sampling: Desingh and Analysis*. Brooks/Cole, 2009.

Daniel J. Stekhoven and P. Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.