UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

7-9 October 2024, Vienna

Application of the MissForest algorithm for imputing income variables in the Survey on Income and Living Conditions

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

blandine.bianchi@bfs.admin.ch

I. ABSTRACT

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. As income variables belong to the main survey objective variables, an appropriate processing is crucial to improve the quality and to increase the reliability of published results derived from the distribution of incomes.

2. Here we test the use of the MissForest algorithm, a non-parametric imputation method based on random forests and allowing the use of mixed data type (categorical and quantitative), to impute missing values across the SILC 2020 income variables. We first imputed individual income variables, followed by household income variables. Socio-demographic variables and household characteristics were used as auxiliary variables for imputing individual income variables. The imputed individual income variables were then added to auxiliary variables on the household level for the imputation of household variables. We ran simulations to evaluate the accuracy of the imputations obtained with this approach.

3. Studying the non-response model, we simulated partial non-response on survey respondents. During the simulations, we first imputed without the boundary variables and then with them. The use of boundary variables proved to be particularly beneficial leading to the reduction of imputation error by up to an order of magnitude for the variables of interest with sufficient observations. When imputing variables for true missing values, we checked that they were within the bounds set to assess the quality of the imputation. Our results show that MissForest hold great potential in improving our ability to fill gaps in SILC survey data.

II. INTRODUCTION

1. One of the most important issues in all surveys is the risk of bias due to non-response Lohr [2009]. The problem is that these missing data reduce the relevance of the results, can affect their credibility, and might even make analysis impossible. It is therefore important to avoid missing data,

and if they exist, they should be amended appropriately. The treatment of these missing data can use up a lot of resources but can improve the quality of the survey and increase the reliability of the published results. The data processing method varies according to the type of non-response: complete non-response is generally treated by reweighing, partial non-response (missing values) is treated by imputation. In general, partial non-response occurs when at least one question has been answered. With imputation, missing values are replaced by a (new) value.

2. In this paper, we show an application of the MissForest algorithm to impute missing values Stekhoven and Buhlmnan [2011] due to the partial non-response in the SILC, with the aim to reduce potential biases and facilitate subsequent analysis of the data. MissForest is a machine-learning algorithm that can process categorical and quantitative variables simultaneously and makes no specific assumptions about the structure and distribution of the data performing imputation of missing data based on random forests. A first application of the MissForest algorithm in the SILC has been presented at the last expert meeting on Statistical Data Editing Bianchi [2022].

3. MissForest provides by default out-of-the-bag (OOB) error estimates to assess the accuracy of the imputations. The error is quantified as the Normalized Root Mean Square Error (NRMSE) for imputed quantitative data and as the Proportion of Falsely Classified (PFC) entries in the categorical data. Here, we additionally validated the performance of this algorithm through simulation, which allowed us to also quantify the improvement linked to the inclusion of boundary variables that were available for this particular survey and for some of the variables. Finally, through impact analysis and plausibilisation, we evaluated the impact of the imputation on the distribution of the original data (with special focus on low income) and checked that the imputed values were inside the boundary variables.

III. DATA PREPARATION AND SIMULATION

A. Dataset

1. The EKL section (Income, Consumption and Living Conditions) provided us with all data of the 2020 survey. The dataset includes household, individual and auxiliary (socio-demographic) variables as well as boundary variables (179 variables). There are 8 156 households and 15 177 individuals considered as respondents.

2. The list of variables to be imputed is provided in the Appendix Table 5. In Table 1 and in Table 2 we show number of full respondents (Resp.), of partial non-respondents (Non-Resp.), the number of units unconcerned by the related question (N. of -3) and the percentage of missing values without the unconcerned ones. This provides an overview of the variables and the extend of the needed imputation. There were no consistency rules stated except from the already mentioned boundary variables. These boundary variables are an interval range, provided by the respondents in alternative to the variable of interest. If the person do not answer to the variable of interest or do not provide any ranges, the upper boundary variable should correspond to the p95.

B. Analysis of the non-response

1. Analysing the non-response mechanism is crucial to be able to deduce a non-response model that is as close as possible to this mechanism observed during the survey, in order to simulate missing

TABLE 1. List of the person variables with the number of (1) Non-missing units: number of repondents or non-cocerned units, (2) Non-respondents, (3) Unconcerned or zero values, (4) Not-respondents over total number of units, (5) Uncocerned over total and (6) Units providing a positive value among the 15 177 units inverviewed.

Variable	(1) Non-Missing	(2) Non-resp	(3) Uncon	(4) Non-resp/ Tot [%]	(5) Uncon/ Tot [%]	(6) Pos Values
P_HY050G_30	13286	1891	13265	12.5%	87.4%	21
P_HY050G_40	15177	0	15160	0.0%	99.9%	17
P_HY060G_30	13284	1893	13236	12.5%	87.2%	48
P_HY060G_40	15177	0	14993	0.0%	98.8%	184
P_HY080G_21	13267	1910	13015	12.6%	85.8%	252
P_HY081G_12	13275	1902	12953	12.5%	85.3%	322
PY010G_30	13262	1915	12272	12.6%	80.9%	990
PY010G_AG12	14793	384	4973	2.5%	32.8%	9820
PY020G_11	13628	1549	13146	10.2%	86.6%	482
PY050G_30	14794	383	14696	2.5%	96.8%	98
PY050G_AG11	14865	312	13868	2.1%	91.4%	997
PY080G_10	14615	562	14395	3.7%	94.8%	220
PY090G_10	15167	10	14750	0.1%	97.2%	417
PY100G_11	15158	19	11399	0.1%	75.1%	3759
PY100G_20	14367	810	12490	5.3%	82.3%	1877
PY100G_30	13277	1900	13071	12.5%	86.1%	206
PY110G_10	15172	5	15016	0.0%	98.9%	156
PY110G_20	14809	368	14643	2.4%	96.5%	166
PY110G_30	13291	1886	13279	12.4%	87.5%	12
PY120G_11	15176	1	15073	0.0%	99.3%	103
PY130G_10	15165	12	14750	0.1%	97.2%	415
PY130G_20	13601	1576	13527	10.4%	89.1%	74
PY130G_30	13287	1890	13284	12.5%	87.5%	3
PY130G_40	13229	1948	13085	12.8%	86.2%	144
PY140G_10	13648	1529	13551	10.1%	89.3%	97

TABLE 2. Number of respondents, missing values, unconcerned units for household variables to be imputed.

Variable	(1) Non-Missing	(2) Non-resp	(3) Uncon	(4) Non-resp/ Tot [%]	(5) Uncon/ Tot [%]	(6) Pos Values
HY090G_10	5455	2701	230	33.1%	2.8%	5225
HY090G_20	7134	1022	5131	12.5%	62.9%	2003
HY140G_10	7141	1015	1	12.4%	0.0%	7140
HH070_AG20	7387	769	4412	9.4%	54.1%	2975
HY050G_AG10	7558	598	5844	7.3%	71.7%	1714
HY5040	7578	578	0	7.1%	0.0%	7578
HH070_AG30	7550	606	3744	7.4%	45.9%	3806
HY100G_10	7748	408	4830	5.0%	59.2%	2918
HY060G_10	7865	291	6688	3.6%	82.0%	1177
HY040G_10	8000	156	6269	1.9%	76.9%	1731
HY130G_10	8083	73	6232	0.9%	76.4%	1851
HY131G_10	8125	31	6773	0.4%	83.0%	1352
HY070G_10	8141	15	8083	0.2%	99.1%	58
HY110G_AG10	8152	4	8032	0.0%	98.5%	120
HY060G_20	8155	1	7988	0.0%	97.9%	167

values for the 'respondents' and to assess the quality of the imputations. Certain combinations of non-respondent characteristics may be rare or even absent among respondents, which limits the exact reproduction of the non-response mechanism among respondents.

2. Using the imputation flags constructed for each variable, we analyse the different combinations of partial non-response present in the original data: we classify them from 1 to 8, with class 1 being the one with the largest number of observations and class 7 being the last, with more than 100 observations. All other classes with less than 100 observations are grouped in class 8 and the non-response profiles within this class are randomly simulated.

3. To calculate the probability of each non-respondent belonging to one of the 8 classes, we opted for a multinomial logistic regression, using the SAS logistic procedure.

4. Among all the auxiliary variables, we first used the `glmselect` (backward) procedure to identify the variables that could be the most explanatory. The model selected age, activity status, Need for dental consultation: yes/no, Current main job: self-employed with employees: yes/no and the Current main job: supervisory position: yes/no. Hence, these variables are used in the multinomial logistic procedure with link = glogit to calculate the non-response probabilities according to this model.

5. To simulate non-response for the sample of respondents (12 617 units), the above calculated non-response probabilities assigned to the respondents were used to choose randomly a non-response class (y_cl from 1 to 8).

C. **Simulation**

1. Simulations of missing values were generated based on the initial distribution of missing values. The simulation of 17% partial non-response was repeated 20 times for respondents. At each iteration, for each class from 1 to 8, the individuals to be placed in a non-response class are selected without replacement. They are selected randomly proportionally to the probability of belonging to the non-response class, obtained by the logistic regression, except for class 8 (class P_y_cl8) for which the exact non-response pattern was selected randomly among the ones present in this class.

2. We used the MissForest algorithm for imputation, with 100 trees and 10 iterations, which are the default parameters. The algorithm converged always with less than 10 iterations, the number of trees can be choosen between 50 and 100. If we halve the number of trees to 50 in order to halve the computation time, the errors increased slightly, therefore, we preferred to be conservative. We proceeded in four stages: at each stage, the individual income variables were imputed using household and socio-demographic auxiliary variables. We first run the algorithm without the boundary variables and then with them to quantify the improvements. We did not generated cases of non-response for which there is not even a range variable available in the simulation.

(1) Imputation of unconcernedness, value -3, of all variables together. Binary variables (-3, 0) have been created where all values different from -3 and not missing were grouped as 0. At the end of the imputation, the values imputed at this stage as -3 are kept, while those imputed as 0 were reset to missing.
(2) Imputation of zeros of all variables together. Multinomial variables with the following modalities were created:
    - -3: the -3 values at the end of the previous step.
    - 0: originally 0 values.
    - 1: grouping of all values above 0. The values imputed as 0 in this step are kept, while those imputed as -3 or 1 are reset to missing.
(3) Imputation of positive values of all variables together. We created new variables taking the original value or 0 which groups the -3 and 0 values. When we run the algorithm including the boundary variables, if the variables of interest have boundaries, we also impute quotient

variables based on these boundaries. If the person has boundaries, then the imputed value must be within those boundaries. If the imputation is outside the limits, the variable of interest takes the value of the limit. This is also guaranteed by the post-imputation processing (regardless of what was imputed in the first and second steps).

(4) In this step, the variables PY010G_AG12, the gross annual salaried income, and PY100G_20, the annual old-age pension 2nd pillar, are re-imputed separately, as these variables contain a large number of missing values expecting a positive value. Thus, first the values of the variable PY010G_AG12 from step 2 alone are re-imputed with all the other imputed or original variables from step 3 as auxiliary/explanatory variables, except PY010G_AG12. Afterwards, the values of the variable PY100G_20 from step 2 are re-imputed with all the other imputed or original variables from step 3 (expect PY100G_20) and also PY010G_AG12 from step 4 as auxiliary/explanatory variables.

IV. VALIDATION

1. For validation purposes, the imputations from the simulations were compared with the original values and the following performance metrics were calculated:

a. Accuracy: The common occurrences of "-3" were calculated.
b. Error1: The rate of occurrences when "-3" was imputed instead of a value different from "-3" among the imputed values.
c. Error2: The rate of occurrences where a value different from "-3" was imputed instead of "-3" among the imputed values.
d. RAE: If both the imputed and the original values are different from "-3", the Relative Absolute Error (RAE) can be calculated. The RAE is expressed as a ratio comparing an average (residual) error with the errors produced by our prediction model.

$$RAE = \frac{\sum_{id} \mid y_{id} - y^*_{id} \mid}{\sum_{id} \mid y_{id} - \overline{y} \mid} \tag{1}$$

where $id = \cap_{i,j}\{i, j|(y_j \neq -3, y^*_i \neq -3)\}$, y_{id} is the true value, y^*_{id} is the predicted value and $\overline{y}$ is mean of the actual values y_{id}. A reasonable model will give a ratio of less than one, Hill [2012].

e. RAE_{adj}: The RAE multiplied by a factor equal to the number of missing values (excluding units imputed in non-concerned) divided by the sum of the number of missing values (excluding units imputed in non-concerned) and the number of values greater than or equal to zero (from the original set).

$$RAE_{adj} = \frac{n(y^*_{id})}{n(y^*_{id}) + n(y_i \geq 0)} * RAE \tag{2}$$

2. Our simulations showed that the imputation error of -3 values was low across all imputed variables ranging from 0% to 0.16% and the accuracy ranging from 89.58% to 99.99% has been judged quite good.

3. Tables 3 and 4 shows the validation results. We observe that the imputation of the unconcerned units "-3" for the variable gross annual salary income (PY010G_AG12) is among the lesser efficient on average (1) with a precision of 0.9667. On the other hand, the imputation errors for positive values (4) and (5) is among the lowest.

TABLE 3. Average imputation performance across 20 simulations without boundary variables: (1) Accuracy, (2) Error1, (3) Error2, (4) Relative Absolute Error, (5) Relative Absolute Error Adjusted, (6) Average number of units imputed, (7) Average number of units imputed positively and (8) Average number of units imputed unconcered. Performance metrics are described in the Validation Section.

Variable	(1) Acc.	(2) Error1	(3) Error2	(4) RAE	(5) RAE_{adj}	(6) Units imp	(7) Units imp pos	(8) Units imp unconc
P_HY050G_30	0.9982	0.002	-	0.100	-	1'506	0	1506
P_HY050G_40	NaN	NaN	NaN	0.100	-	-	0	0
P_HY060G_30	0.9974	0.003	-	0.100	-	1'500	0	1500
P_HY060G_40	NaN	NaN	NaN	0.100	-	-	0	0
P_HY080G_21	0.9759	0.024	0.000	0.399	0.002	1'497	0	1497
P_HY081G_12	0.9754	0.022	0.002	1.094	0.036	1'499	5	1494
PY010G_30	0.8926	0.107	0.001	0.346	0.001	1'499	0	1498
PY010G_AG12	0.9556	0.011	0.034	0.429	0.012	308	205	103
PY020G_11	0.9507	0.049	0.000	0.100	0.000	1'154	0	1154
PY050G_30	0.9926	0.007	0.000	0.124	0.000	291	0	291
PY050G_AG11	0.9610	0.030	0.009	0.764	0.012	247	12	236
PY080G_10	0.9776	0.022	0.000	0.100	0.000	479	0	479
PY090G_10	0.9576	0.042	-	0.100	-	36	0	36
PY100G_11	0.9884	0.006	0.006	0.592	0.001	29	3	26
PY100G_20	0.8749	0.057	0.068	0.842	0.082	715	134	581
PY100G_30	0.9893	0.010	0.000	0.889	0.012	1'484	1	1483
PY110G_10	0.9945	0.006	-	2.584	0.042	36	0	36
PY110G_20	0.9948	0.004	0.001	1.105	0.007	331	0	330
PY110G_30	0.9994	0.001	-	0.100	-	1'478	0	1478
PY120G_11	1.0000	-	-	0.100	-	8	0	8
PY130G_10	0.9774	0.018	0.004	0.095	0.000	35	0	35
PY130G_20	0.9935	0.006	0.000	0.265	0.004	1'187	0	1187
PY130G_30	0.9998	0.000	-	0.100	-	1'478	0	1478
PY130G_40	0.9880	0.012	-	0.100	-	1'506	0	1506
PY140G_10	0.9846	0.015	0.000	0.100	0.000	1'139	0	1139

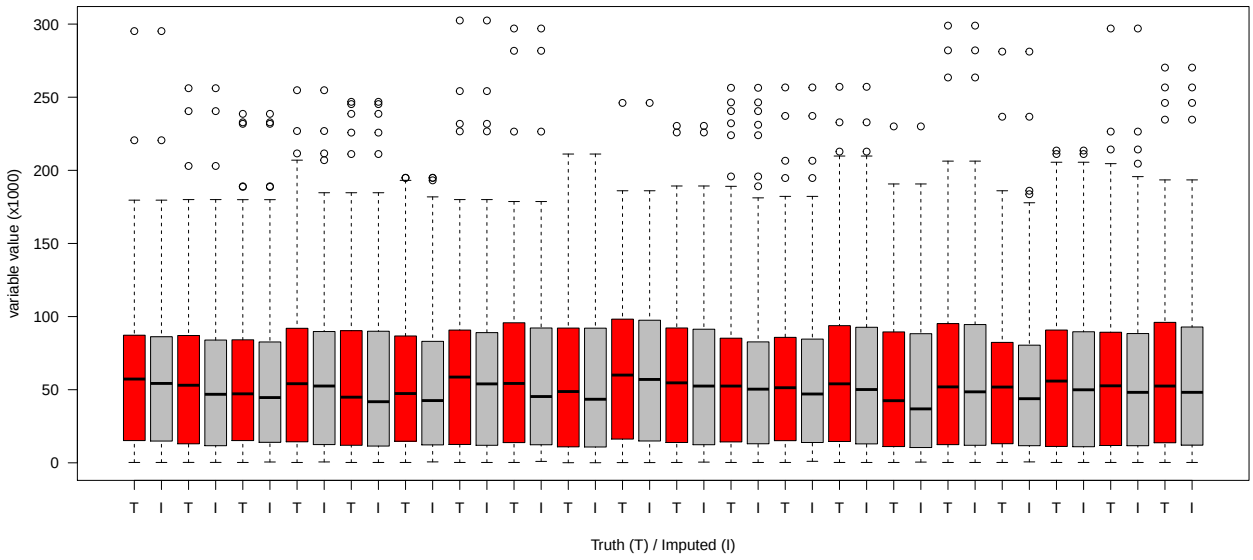
TABLE 4. Average imputation performance across 20 simulations using boundary variables. Metrics as in table 3.

Variable	(1) Acc.	(2) Error1	(3) Error2	(4) RAE	(5) RAE_{adj}	(6) Units imp	(7) Units imp pos	(8) Units imp unconc
P_HY050G_30	0.9984	0.002	-	0.100	-	1'500	0	1500
P_HY050G_40	NaN	NaN	NaN	0.100	-	-	0	0
P_HY060G_30	0.9973	0.003	-	0.100	-	1'496	0	1496
P_HY060G_40	NaN	NaN	NaN	0.100	-	-	0	0
P_HY080G_21	0.9757	0.024	0.000	0.486	0.002	1'496	0	1496
P_HY081G_12	0.9755	0.022	0.003	1.003	0.031	1'494	6	1489
PY010G_30	0.8958	0.103	0.001	0.250	0.001	1'494	0	1494
PY010G_AG12	0.9667	-	0.033	0.035	0.001	315	214	102
PY020G_11	0.9496	0.050	0.000	0.100	0.000	1'155	0	1155
PY050G_30	0.9920	0.007	0.001	0.148	0.001	290	0	290
PY050G_AG11	0.9893	-	0.011	0.079	0.002	249	19	230
PY080G_10	0.9997	-	0.000	0.410	0.024	478	13	465
PY090G_10	0.9672	0.033	-	0.100	-	36	0	36
PY100G_11	0.9944	0.001	0.004	0.663	0.001	29	4	26
PY100G_20	0.9339	-	0.066	0.073	0.008	716	186	530
PY100G_30	0.9889	0.011	0.000	10.871	0.158	1'481	1	1480
PY110G_10	0.9935	0.006	-	0.095	-	35	0	35
PY110G_20	0.9980	-	0.002	0.455	0.010	327	2	325
PY110G_30	0.9993	0.001	-	0.100	-	1'474	0	1474
PY120G_11	0.9900	0.010	-	0.100	-	6	0	6
PY130G_10	0.9902	0.010	-	0.090	-	36	0	36
PY130G_20	0.9999	-	0.000	0.422	0.038	1'188	7	1181
PY130G_30	0.9999	0.000	-	0.100	-	1'475	0	1475
PY130G_40	0.9881	0.012	-	0.100	-	1'502	0	1502
PY140G_10	0.9836	0.016	-	0.100	-	1'139	0	1139

4. Looking the results of these two tables we can also notice that, for instance for the variables gross annual salaried income (PY010G_AG12) and 2nd pillar (PY100G_20), we improve the quality of the imputation including the boundary variables in the imputation : we have an increase in accuracy (units correctly imputed when not concerned by the variable) and a decrease in relative absolute error of an order of magnitude.

5. We also observe that the performance metric (4) is greater than 1 for variables P_HY081G_12 and PY100G_30. We note that the number of units imputed as unconcerned is very high (above 97%) affecting this metric highly due to the small number of imputed positive values. This measure is therefore not well appropriated for such circumstances. The REA_{adj} provides a clearer estimate of the model's imputation errors, since it takes into account the number of missing values to be imputed and the number of positive values imputed.

FIGURE 1. Box-plots of the variable gross annual salaried income (between 0 and 300 000 CHF) - In red original values and in grey the values imputed during the 20 imputations. Units not involved are not plotted.



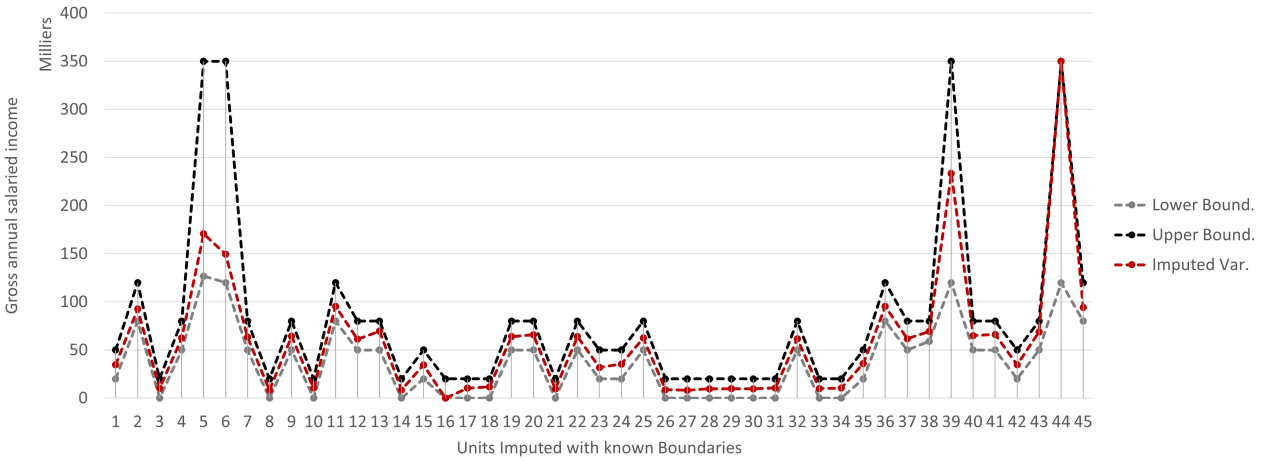
6. In figure 1 we show the box-plot of the variable of interest gross annual wage income (between 0 and 300 000 CHF). The original values are shown in red, while the values imputed during the 20 simulations (using the boundary values) are shown in grey. The distribution of original and imputed values are highly consistent, even though the median of the imputed values is a slight underestimate. The higher incomes have also been correctly imputed, although we note that we used here the boundaries conditions set to the percentile 0.95 to ensure that extreme values do not skew the results and provide a more robust measure of the low and central tendency.

A. **Plausibilisation**

1. We checked that the values imputed for the units with missing values were within the known boundaries. In figure 2 we can see for instance imputed values for the gross annual salaried income missing but with known boundaries, the values are always imputed within the interval range and when they are not the post-treatment imputes them on the boundaries.

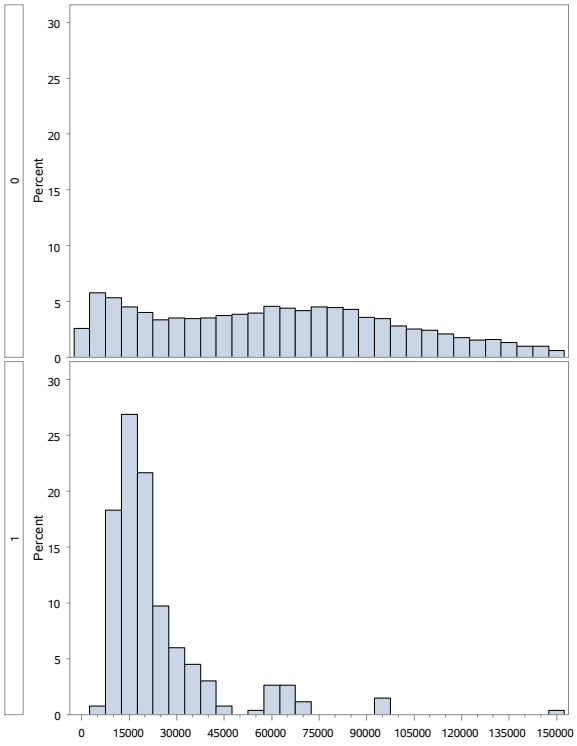
2. We also checked that the imputed values for the household-level variables were within the limits that was the case for all imputed values with known boundaries.

FIGURE 2. In red, imputed values of the gross annual salaried income variable and the known boundaries variables used as auxiliary variables (in grey and black).



B. Impact Analysis

FIGURE 3. Distributions of the original (flag = 0) and imputed (flag = 1) values for the variable gross annual salaried income. Note that for clarity the x-axis was truncated at 150 000 CHF and only positive values are shown.



1. As impact analysis to evaluate the effect of imputing the missing values for each variable of interest, we compared the distribution of reported values against the distribution of imputed ones. In

the absence of a bias of the non-response, we would expect the two distributions to be similar. However, in our dataset we found that the distribution of imputed values substantially differed from that of the known values in several variables, as shown in figure 3 for the variable gross annual salaried income (PY010G_AG12). We observed a similar pattern when we performed the impact analysis on the household variables. These results highlight the importance of carrying out imputations, particularly when including auxiliary variables and boundaries, to correct for the potential bias due to non-random non-response.

V. CONCLUSIONS

1. We applied the MissForest algorithm to impute the SILC 2020 income variables of interest. We first imputed individual income variables, using socio-demographic auxiliary variables and reported household variables. We studied the non-response mechanism and simulated missing values to evaluate imputations without and with the use of boundary variables. We found a significant improvement in the accuracy of the imputations with the use of boundary variables, with absolute relative errors decreasing by up to one order of magnitude. We then imputed the true missing values and reconstructed the total value of personal income. This variable was used as an auxiliary variable to impute true missing values for households using in addition other household socio-demographic variables. We then carried out an impact analysis of all the variables of interest that we had to impute, and a plausibility check to ensure that the imputed values were well within the boundary variables. Based on our simulations, the magnitude of the error is acceptable and is mainly related to the initial distribution of the variable of interest to be imputed: the fewer positive values there are in the variable of interest, the less likely it is that positive values will be imputed. However, we observed from both the simulations and the impact analyses that the most important variables in terms of salaries are those for which the imputation errors of positive values are among the lowest.

VI. APPENDIX

A. List of person and household variables

TABLE 5. List of person and household income variables to be imputed.

Person Variables (P)	
P_HY050G_30	Annual advances on maintenance and alimony payments
P_HY050G_40	Allowances for loss of earnings, maternity: amount
P_HY060G_30	Annual income from other institutions
P_HY060G_40	Allowances for loss of earnings, other than maternity: amount
P_HY080G_21	Annual income from other monetary transfers received from other households
P_HY081G_12	Annual income from monetary transfers received from other households
PY010G_30	Gross annual income from a secondary activity
PY010G_AG12	Gross annual salaried income
PY020G_11	Annual benefits in kind received from employers
PY050G_30	Annual benefits in kind received from self-employed persons
PY050G_AG11	Gross annual self-employed income
PY080G_10	Annual 3rd pillar pension
PY090G_10	Annual income from unemployment
PY100G_11	Annual old-age pension 1st pillar
PY100G_20	Annual old-age pension 2nd pillar
PY100G_30	Old-age pension from abroad
PY110G_10	Annual 1st pillar survivor's pension
PY110G_20	Annual 2nd pillar survivor's pension
PY110G_30	Survivor's pension from abroad
PY120G_11	Pension or daily sickness benefit
PY130G_10	Annual 1st pillar invalidity pension
PY130G_20	Annual invalidity pension 2nd pillar
PY130G_30	Invalidity pensions from abroad
PY130G_40	Pension or daily allowances from insurance other than sickness
PY140G_10	Annual income from study grants

Household variables (H)	
HY090G_10	Annual interest income from bank and post office accounts
HY090G_20	Annual amount of interest income or dividends from investment funds
HY140G_10	Tax: amount
HH070_AG20	Annual housing costs for homeowners (mortgage interest + other housing costs)
HY050G_AG10	Family and maternity allowances
HY5040	Annual amount of household income donated to the CATI
HH070_AG30	Annual housing costs for tenants, including service charges and ancillary costs
HY100G_10	Annual mortgage interest on main home
HY060G_10	Annual health insurance subsidies
HY040G_10	Annual income from property or land
HY130G_10	Annual payments to non-residents
HY131G_10	Annual maintenance payments to persons outside the household
HY070G_10	Annual amount from private or public housing assistance
HY110G_AG10	Annual income from children
HY060G_20	Public welfare benefits: amount

TABLE 6. List of socio-demographic auxiliary variables 1-59.

N.	Name of the variable	Description
1	Age	Age
2	Age_18_64_cl3	Age classes
3	Age_cl3	Age classes
4	Age_GE16_cl4	Age classes
5	Age_GE16_cl6	Age classes
6	agemax_1	Age of oldest person in household 16-19
7	agemax_2	Age class of oldest person in household 20-29
8	agemax_3	Age group of oldest person in household 30-39
9	agemax_4	Age of oldest person in household 40-49
10	agemax_5	Age of oldest person in household 50-64
11	agemax_6	Age of oldest person in household 65-74
12	agemax_7	Age class of oldest person in household > 74
13	ARP60_cdc	At-risk-of-poverty status with threshold at 60% of median (from register)
14	arp60_LG1	At-risk-of-poverty status with threshold at 60% of median, year T-1
15	arp60_LG2	At-risk-of-poverty status with 60% median threshold, year T-2
16	arp60_LG3	At-risk-of-poverty status with threshold at 60% of median, year T-3
17	Atype	Household composition by nationality
18	G_LNG_10	Interview language grid
19	Gtype	Household composition by gender
20	HH_RES_GDET25_2012_1	Place of residence: 2012 commune types (25 types)
21	HH_RES_GDET9_2012_1	Place of residence: commune types 2012 (9 types)
22	HH_RES_REGCH_2011_2	Major regions of Switzerland
23	HH_RES_SPRGEB_2011_2	Language regions 2000
24	HH010_10D_C	Type of dwelling: number of apartments in building
25	HH010_10E_C	Housing type: number of apartments in the building
26	HH010_10L2_C	Housing type: type of house
27	HH021_10M	Housing: occupancy status
28	HH021_10N	Housing: occupancy status: tenant: market price
29	HH030_10	Housing: number of rooms available
30	HH031_10_C	Housing: owner: year of purchase - bound
31	HH031_20_C	Housing: tenant: year lease signed - consolidated
32	HH031_30_C	Housing: year of move-in - consolidated
33	HH060_X	Housing: tenant: current rent excluding utilities and ancillary costs: annual amount
34	HHTYPE_30	Household type
35	HQ5010_10	Satisfaction with household financial situation: scale
36	HS120_10	Financial ability to make ends meet
37	HS130_10	Minimum monthly income to make ends meet: amount
38	HS5010_10	presence of ASSMAL arrears
39	HS5020_10	tax arrears
40	HT_MaxAgeKID_2X3	Combination family with child(ren) 2CL and age of oldest child 3CL
41	HT_MaxAgeKID_2X4	Combination family with child(ren) 2CL and age of oldest child 4CL
42	HT_MaxAgeKID_2X5	Combination family with child(ren) 2CL and age of oldest child 5CL
43	HT_MinAgeKID_2X2	Combination family with child(ren) 2CL and age of youngest child 2CL
44	HT_MinAgeKID_2X3	Combination family with child(ren) 2CL and age of youngest child 3CL
45	HT_MinAgeKID_2X4	Combination family with child(ren) 2CL and age of youngest child4CL
46	HT_NbKID_2X3	Combination family with child(ren) 2CL and number of child(ren) 3CL
47	HY5010_10	Evaluation of income and expenses
48	max_educ_H	Maximum household education
49	natio_geo_cl4	Nationality in four classes
50	nbactifs	Number of working members of household
51	NBIND_GE18	number of persons in the household aged 18 and over
52	NBIND_GE65	number of household members aged 65 and over
53	NBIND_GE75	number of household members aged 75 and over
54	NBIND_LE55	number of household members aged 55 and under
55	NBIND_LT16	number of household members aged 15 and under
56	NBIND_LT25	number of household members under 25 years of age
57	nbpers	Number of people in household, plausibilized
58	occupa_cl4	type of occupation
59	PB190_10_C	Civil status: SILC OFS codes, consolidated variable

TABLE 7. List of socio-demographic auxiliary variables 60-91.

N.	Name of the variable	Description
60	PB5100_10	Residence permits from register, SILC quality codes
61	PE040_11BCons_cl3	Highest level of education attained, consolidated
62	PH010_10	General health
63	PH020_10	Chronic illness or long-term health problem
64	PH030_10	Limitation in daily activities due to health problems
65	PH060_10A	Need for dental consultation: yes/no
66	PH060_20A	Need for dental consultation: completed: yes/no
67	PH070_10	Dental consultation: main reason for not consulting
68	PL031_10	Current main activity: self-evaluation
69	PL040_10P	Current main job: type of job
70	PL040_10Q	Current main job: self-employed with employees: yes/no
71	PL051_2digits	Current or last main occupation
72	PL060_10	Number of weekly hours usually worked in main job: SFO filter
73	PL100_10	Total hours usually worked in secondary jobs: SFO filter
74	PL101_10	Number of hours worked per week in all jobs
75	PL111_10	Current company: NACE
76	PL150_10	Current main job: supervisory function
77	PL5040_10	Current main job: part-time or full-time
78	PP5010_10_cl2	Interest in politics
79	PP5010_10_cl5	Interest in politics
80	PP5030_10_cl5	Ideological stance: left-right
81	PW5020_10_cl5	General health: satisfaction
82	PW5060_10_cl5	Evaluation of satisfaction with financial situation
83	PW5150_10_cl5	General satisfaction with life
84	PW5170_10_cl3	Feeling of discouragement or depression
85	PW5180_10_cl3	Sense of happiness
86	PW5200_10_cl5	Satisfaction with amount of free time
87	PW5210_10_cl5	Confidence in the political system
88	PW5220_10_cl5	Confidence in the judicial system
89	PW5230_10_cl5	Confidence in the police
90	sex	Gender
91	wareas	Size of home GWS

References

B. Bianchi. Application of the missforest algorithm for imputation in the survey on income and living conditions. *UNECE*, 3-7 October, 2022.

A. Hill. *The Encyclopedia of Operations Management: A Field Manual and Glossary of Operations Management Terms and Concepts*. FT Press, 2012.

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

D. J. Stekhoven and P. Buhlmnan. Missforest non-parametric missing value imputation for mixed-type data. *BIOINFORMATICS*, 28:112–118, 2011.