UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Application of the "SwissCheese" algorithm for the imputation of partial non-response in the Survey on Income and Living Conditions

Michael Leuenberger, Federal Statistical Office, Switzerland

michael.leuenberger@bfs.admin.ch

I. INTRODUCTION

1. In this work, a balanced imputation method implemented in the SwissCheese algorithm is presented. Developed by the University of Neuchâtel [Eustache et al., 2021], this method can handle Swiss cheese non-response, i.e. in the case where all variables of a survey may contain missing values without a particular pattern. Applied on the Survey on Income and Living Conditions in Switzerland, this contribution shows the adaptations and optimisations made on the algorithm in order to improve the imputation.

2. A simulation framework is used to evaluate the quality of the imputation and to compare the results with other algorithms. In order to improve the imputation quality, several adjustments are explored, such as: selection of auxiliary variables, construction of range variables, optimisation of the SwissCheese parameters, and others.

3. The main objectives of the proposed approaches are to highlight the procedure of use and its adaptations for a specific survey. To this regard, a comparison with the MissForest algorithm [Stekhoven and Bühlmann, 2012] is provided. Finally, further potential improvements are exposed and discussed.

II. Data

1. The data used in this study come from the Survey on Income and Living Conditions (SILC). Focused on the data from 2020, which is still an experimental dataset, it contains several variables of interest, named VI later in the paper. The list and the description of the variables of interest are as follows:

- HV010: Wealth - owner of main residence: value.
- HV020: Wealth - property or land without main residence: amount.
- HV070: Debts - total mortgages on main residence: amount.
- HF5030: Wealth - balance on bank and postal accounts: amount.
- HF5040: Wealth - value of shares, bonds, investment funds etc.: amount.
- HF5050: Wealth - other valuables: amount.
- HF5060: Debts - total mortgages except on main residence: amount.

TABLE 1. Number of observation per dataset, number of missing values and the ratio of missing values in the dataset.

Dataset names	Number of observations	Number of NA in VI	Ratio of NA in VI
HV010	3445	299	9%
HV020	1658	218	13%
HV070	2995	311	10%
HF5030	6663	1270	19%
HF5040	2532	477	19%
HF5050	3048	338	11%
HF5060	1648	227	14%

2. From the dataset SILC2020, several datasets have been extracted. In particular, one dataset per variable of interest was generated due to different filters applied to each variable. In Table 1, these datasets are presented with their respective variable of interest, the total number of units expected to provide a value for the variable of interest, the number of missing values in the variable of interest and the ratio between the total number of units and the number of missing values of the total number of units. The number of units can vary greatly from one dataset to another. This is due to the fact that the questions in the SILC survey do not necessarily concern all persons participating in the survey. For example, a person who is not a home-owner will not answer the questions about property, creating a number of people who are "eligible" for the question that may vary. In turn, the proportion of missing values in the variables of interest also varies considerably between the different datasets, from 9% for dataset HV010 up to 19% for datasets HF5030 and HF5040.

3. In addition to these variables of interest, each dataset is composed of auxiliary variables. A total of 127 auxiliary variables are used in order to improve the quality of the imputation. These variables are composed of general information about the household like household composition and financial status among others.

4. The interest in obtaining correct imputation for missing values of these variables of interest is high. Mainly because the results of the first quintiles are later used in a poverty-related analysis. Therefore, the disposal of a complete dataset and, as far as possible, with conservation of correlations between variables is our main objective.

III. Method

A. SwissCheese

1. For this outlined development, we used the SwissCheese R package developed by the University of Neuchâtel [Eustache et al., 2021]. Consider a dataset S with k variables and n observations. This algorithm first separates the dataset S into two parts S_r and S_m, with S_r being the response part of the dataset and S_m being the observations with at least one missing value. We will later call S_m the non-respondent part. For the case of total non-response it should be handled by reweighting, but this aspect is out of the scope of this article.

2. The two main properties of this algorithm are as follows:

- missing values are imputed by real values.

- relationships between variables are preserved.

To guarantee these properties the SwissCheese algorithm passes through with the following steps. For each non-respondent in S_m it selects a donor from the observations in S_r. The donor is selected among the K nearest neighbours of the observation with missing value. Moreover, if the observed auxiliary values of the non-respondents were imputed, the total estimator of each of these variables should remain unchanged.

3. It is worth mentioning that when a donor is selected, it provides values for each of the missing values in the non-respondent observation. This last aspect ensures a coherence among the imputed values and with respect the observed values of the non-respondent. More details on this algorithm can be found in the following article [Eustache et al., 2022].

B. Simulation framework and Assessment

1. The evaluation of the imputation quality is done by means of a simulation framework where missing values are randomly generated among the respondents. The generation of missing values in S_r is performed by following the ratio of missing data in the original dataset, and by generating the same ratio of missing data in pre-constructed homogeneous response groups based on the terminal leaf of a regression tree model. This generation of missing data allows a comparison between the imputed values and the real values. Based on the SILC data, this procedure of generating missing data is applied on each dataset related to a variable of interest, allowing the comparison of different imputation methods and the optimization of the SwissCheese parameters.

2. In this development all comparisons are made with the following tools and measures:

- a quintile confusion matrix, comparing the original values with the imputed values,
- a boxplot by decile,
- and a measure of accuracy based on the confusion matrix and the root mean square error (RMSE) based on the original values.

3. The results are compared with those obtained with the MissForest algorithm [Stekhoven and Bühlmann, 2012] on the same datasets. Moreover, in order to improve the results of the SwissCheese algorithm on this type of dataset, a selection of variables, using the calculation of the correlation between each variable of interest and the auxiliary variables, is carried out on the datasets. The correlation must then exceed a certain threshold in order to retain a variable as an explanatory variable in the imputation. In our case, several thresholds have been tested and a threshold value of 0.3 was kept for our datasets. This selection of variables allows to remove unnecessary variables, which can disturb the results of the balanced imputation and add excessive extra calculation time. On the other hand, all variables were retained for the MissForest procedure because it can handle such situations.

IV. Results and Discussion

1. In order to compare the different algorithms with each other, several methods and measures are used, namely confusion matrices and boxplots, as well as a measure of accuracy based on the confusion matrix and the RMSE based on the original values. Since we are working with simulated NA data, we have the original values available as a basis for evaluating the accuracy of the different imputations.

2. The confusion matrix is based on the imputed values per quintile and on the original values per quintile, the disaggregation aimed by the analysts of the SILC data. Therefore, the greater the

proportion of data on the diagonal, the more correctly the data are imputed in their respective quintile. An overall accuracy percentage or "Accuracy" is also calculated in order to give a general idea of the imputation quality. It is also important to look in detail at the proportion of correctly imputed data for each quintile, as a high overall accuracy can be strongly influenced by a very good imputation in one quintile only.

3. The boxplots are divided by decile (based on the original data) giving a more detailed assessment of the imputation quality than the confusion matrix. They represent the difference between the logarithm of the original values and the logarithm of the imputed values, with a red line indicating 0. It shows the quality of the imputation for each decile, how large the difference is or how close to 0 it is. Extreme values can also be observed. We can also relate the difference between the different deciles, for example between the more central deciles and the extreme deciles.

TABLE 2. Confusion matrix of the datset HF5050 with the imputation of the SwissCheese algorithm.

		Original values				
		Q1	Q2	Q3	Q4	Q5
Imputed values	Q1	0.60	0.20	0.15	0.09	0.02
	Q2	0.04	0.07	0.15	0.01	0.02
	Q3	0.16	0.41	0.40	0.16	0.02
	Q4	0.13	0.29	0.30	0.42	0.38
	Q5	0.07	0.02	0.00	0.31	0.57
	Tot	1.00	1.00	1.00	1.00	1.00

TABLE 3. Confusion matrix of the datset HF5050 with the imputation of the MissForest algorithm.

		Original values				
		Q1	Q2	Q3	Q4	Q5
Imputed values	Q1	0.45	0.02	0.00	0.00	0.00
	Q2	0.27	0.34	0.21	0.03	0.05
	Q3	0.27	0.61	0.78	0.51	0.20
	Q4	0.00	0.02	0.01	0.38	0.41
	Q5	0.00	0.00	0.00	0.09	0.34
	Tot	1.00	1.00	1.00	1.00	1.00

4. In the two confusion matrices presented in Tables 2 and 3 we can see that the results from the SwissCheese algorithm perform better in the first and the fifth quintiles. However the MissForest algorithm shows better results in the middle quintiles. Nevertheless, the measures of accuracy and RMSE for this dataset (HF5050) presented in Table 4 are relatively close.

5. The same observations can be made for the Figure 1 for the dataset HF5050. The deciles at the ends show a better behaviour for the SwissCheese algorithm than the MissForest algorithm. However, the precision of the MissForest is improved in the middle deciles.

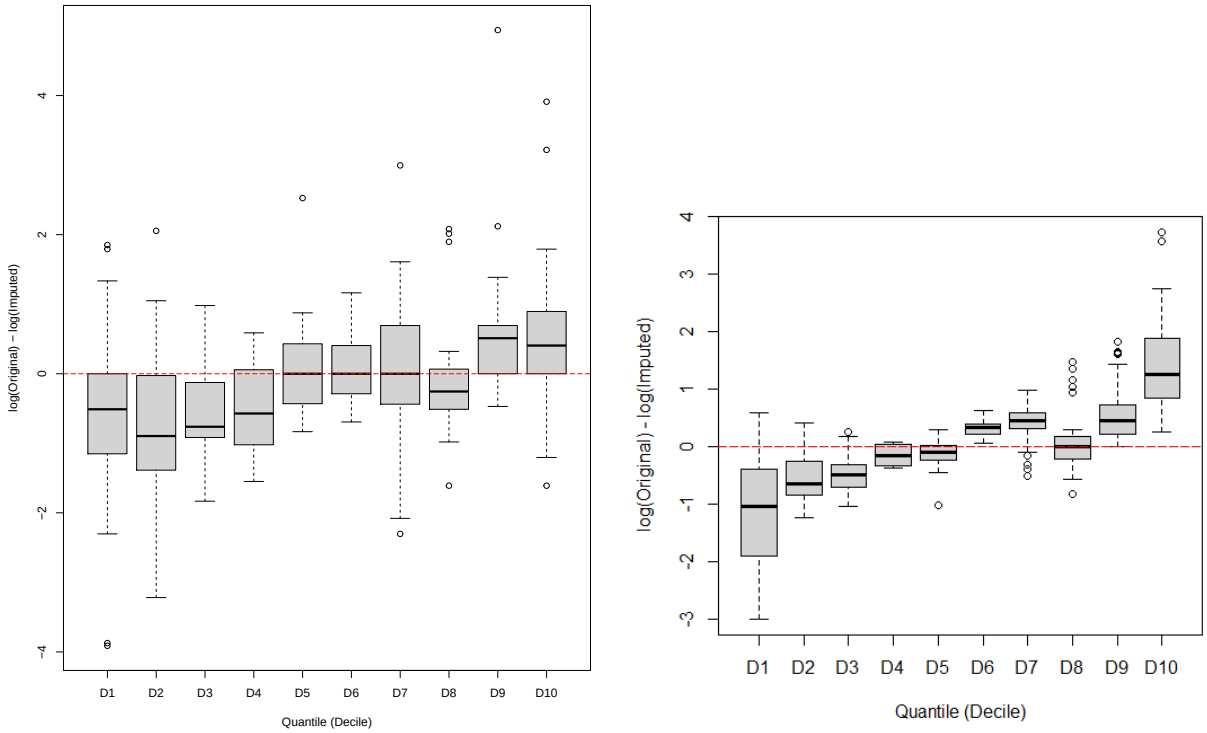

FIGURE 1. Boxplot of the dataset HF5050 with the imputation of the SwissCheese algorithm on the left, and with the imputation of the MissForest algorithm on the right.

TABLE 4. Accuracy and RMSE calculated on the imputation with SwissCheese and MissForest.

Dataset names	Accuracy SwissCheese	Accuracy MissForest	RMSE SwissCheese	RMSE MissForest
HV010	19.3%	60.4%	0.68	0.32
HV020	38.4%	50.8%	1.76	0.96
HV070	45.3%	61.7%	0.61	0.38
HF5030	47.4%	56.3%	1.43	1.07
HF5040	52.1%	53.1%	1.35	1.08
HF5050	43.4%	46.7%	1.01	0.83
HF5060	34.5%	52.3%	1.68	1.03

6. Finally, in Table 4 the global results of the imputation for the SwissCheese and the MissForest algorithms are depicted for all datasets. In particular, we can see that in most of the cases the results of the MissForest are slightly better than those of the SwissCheese algorithm. In the case where the results are close, it is always the same pattern, the SwissCheese shows better results in the ends of the distribution, while MissForest performs better in the centre of the distribution. However, in some cases the results of the SwissCheese show a relatively low performance (i.e. for HV010, HV020 and HF5060). After investigation of these cases, it appears that the number of selected variables is low, which may explain a part of this effect. For future improvements, further analysis on the method of variable selection, as well as the addition of more relevant auxiliary variables, could greatly improve the outcome of these cases.

7. Another important improvement in the imputation results of SILC2020, but not directly related to the SwissCheese algorithm, is the addition of range variables in the auxiliary variables. Indeed, during the SILC2020 survey, a new possibility of response was given. For example in the question related to the variable HF5030 (balance on bank and postal accounts), when the respondent is not able or not willing to give an integer value, he/she has the opportunity to provide a categorical answer based on predefined intervals of interest such as:

- $[0, 5'000[$
- $[5'000, 25'000[$
- $[25'000, 100'000[$
- $[100'000, +\infty[$

This new information resulted in a significant improvement of the imputation quality. For example, on a simulation framework based on the variable HF5030 of the survey SILC2015, a decrease in the measure of RMSE was observed, moving from 1.82 to 1.33. Now with the SILC2020 survey and by having this range variable at hand, we are able to reach the same level of quality as expected with the simulation based on SILC2015.

8. Other ideas for improvement are still under development. Among these we can cite the use of additional auxiliary variables. However, attention must be paid to the relevance of the added variable. As it was mentioned, a variable selection is required in order to improve the results and the calculation time. In this regard, the SwissCheese algorithm is more sensitive to this aspect than is the MissForest algorithm. Adding new variables would be counter-productive and time consuming if the relevance of these new variables is low in relation to the variables of interest. However, as outlined before with the addition of range variables, a good auxiliary variable can have a significant impact on the imputation quality, especially with the SwissCheese algorithm.

V. Conclusions

1. From the different results, we see that the SwissCheese algorithm seems to work particularly well on the values placed at the extremes and a little less well for the values in the middle of the distribution, although we can see a improvement when a variable selection is performed.

2. It is also important to mention the results of the MissForest algorithm, which are not only quite good but also subject to improvement. Indeed, where the SwissCheese algorithm does better at the extremes, the MissForest algorithm imputes the middle quintiles particularly well. Moreover, the MissForest algorithm does not require variable selection, as it can itself discard variables of little interest, and obtains fewer large errors than the SwissCheese algorithm.

3. In conclusion, the SwissCheese algorithm achieves encouraging results in the imputation of SILC data, which can still be improved. However, care should be taken to ensure that the selection of variables is efficient and that the use of additional auxiliary variables are consistent with the variables of interest.

References

E. Eustache, A. A. Vallée, and Y. Tillé. *Balanced Imputation for Swiss Cheese Nonresponse*. R package SwissCheese, 2021. URL https://github.com/EstherEustache/SwissCheese.

E. Eustache, A. A. Vallée, and Y. Tillé. Balanced donor imputation handling swiss cheese nonresponse. *Accepted paper in Statistica Sinica*, 2022.

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