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 FSO / Data Science, AI and Statistical Methods / Statistical Methods

UNECE Conference of European Statisticians
Expert meeting on Statistical Data Editing | 3-7 October 2022

Introduction

A joint work with the *Institute of Statistics*, *University of Neuchâtel*, *Switzerland*. Many thanks to *Esther Eustache* and *Arnaud Tripet*.

Main Objectives *:*
Highlight the procedure of use and the adaptations of the algorithm to a specific survey :

- *SILC* dataset : Survey on Income and Living Conditions in Switzerland.
  - 7 fortune variables of interest with a missing rate ranging from 9% to 19%.
- *SwissCheese* algorithm : balance based imputation algorithm.

SILC Datasets

- 7 datasets extracted from the 2020 SILC in Switzerland.
- Due to filters applied to each variable of interest 7 datasets were extracted, the amount of available data and the missing rate can vary greatly from one dataset to another.
- The simultaneous imputation of the entire dataset is therefore not possible.

SILC Variables

This presentation will focus on :

- *HF5050* : Wealth - oher valuables : amount.
  - Number of observation : 3048
  - Number of missing value : 338 (11%)
- *HV070* : Debts - total mortgages on main residence : amount.
  - Number of observation : 2995
  - Number of missing value : 311 (10%)
- A total of 127 auxiliary variables are used (such as information about household composition and financial status).

SwissCheese algorithm

Developed by the University of Neuchâtel, this algorithm can handle multivariate imputation, and has the following properties :

- missing values are imputed by real values by selecting one donor from the respondents.
- relationships between imputed variables are preserved.

The choice of the donor is based on a trade-off between selecting a response unit in the neighbourhood and the balancing of auxiliary variables.

Simulation framework

From a first preliminary study based on the SILC 2015 dataset, the process has been adapted to the SILC 2020 dataset. It consists of the following steps :

- determination of homogeneous response groups based on the terminal leafs of a regression tree model.
- generation of missing data in each homogeneous group based on the observed missing rate.
- selection of auxiliary variables based on the correlation with the variable of interest.
- application of the SwissCheese algorithm on the simulated data.

Assessment

Comparisons with the MissForest algorithm was made with the following tools and measures :

- confusion matrix based on quintiles.
- boxplots based on deciles.
- measure of accuracy based on the confusion matrix and the root mean square error (RMSE) based on the original values.

Results

Dataset names	Accuracy SwissCheese	Accuracy MissForest	RMSE SwissCheese	RMSE MissForest
HV070	45.3%	61.7%	0.61	0.38
HF5050	43.4%	46.7%	1.01	0.83

The other five variables are mostly similar.

Results - HF5050

Imputed values \ Original values	Q1	Q2	Q3	Q4	Q5
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

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

Results - HV070

Imputed values \ Original values	Q1	Q2	Q3	Q4	Q5
Q1	0.71	0.38	0.08	0.10	0.00
Q2	0.05	0.19	0.15	0.00	0.09
Q3	0.14	0.31	0.38	0.25	0.17
Q4	0.05	0.12	0.27	0.35	0.17
Q5	0.05	0.00	0.12	0.30	0.57
Tot	1.00	1.00	1.00	1.00	1.00

Confusion matrix of the datset HV070 with the imputation of the SwissCheese algorithm.

Conclusions

- Encouraging results of the SwissCheese algorithm.
- At the extremes the SwissCheese algorithm obtains better results than the MissForest algorithm.
- Important improvement in the imputation results with the addition of range variables as auxiliary variables.

Potential improvements :

- Trying other procedure of variable selection.
- Adding further auxiliary variables.

References

- E. Eustache, A.-A. Vallée and Y. Tillé. Balanced Donor Imputation Handling Swiss Cheese Nonresponse. *Accepted paper in Statistica Sinica*.
- Daniel J. Stekhoven and P. Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112-118, 2012.

SwissCheese package available on :
https ://github.com/EstherEustache/SwissCheese