Deliverable Number: 2.12

Deliverable Title: Flexible Stata Code for Hot-Deck Imputations of Non-Monetary Variables in SHARE

Work Package: WP2 - Representing the population

Deliverable type: Other

Dissemination status: Public

Submitted by: SHARE ERIC

Authors: Giuseppe De Luca (MEA)

Date Submitted: June 2018

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey of Health Ageing and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: De Luca, G. (2018) Flexible Stata Code for Hot-Deck Imputations of Non-Monetary Variables in SHARE. Deliverable 2.12 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: www.seriss.eu/resources/deliverables

Deliverable 2.12: Flexible Stata Code for Hot-Deck Imputations of Non-Monetary Variables in SHARE

Giuseppe De Luca

Summary

This report describes the flexible algorithm used to produce multiple hot-deck imputations (MHDI) of the non-monetary variables in the first six waves of the Survey on Health, Ageing and Retirement in Europe (SHARE), excluding the retrospective wave in 2008. After discussing the rationale and the key logical steps of the SHARE algorithm, we focus on its implementation in the statistical software Stata (version 13.1) by providing a detailed description of the programs used for handling a number of issues in the construction of the SHARE public-use imputation dataset.

1. Introduction

Nonresponse is a serious problem that affects most survey datasets, especially in the medical, social and economic sciences. At the outset, it is useful to distinguish between two types of nonresponse. The first – unit nonresponse – occurs when eligible sample units fail to participate in a survey because of noncontact or explicit refusal to cooperate. The second – item nonresponse – occurs when responding units do not provide useful answers to particular items of the questionnaire. The potential implications of the two types of nonresponse are similar, namely selectivity bias and loss of precision. The key difference is that for unit nonresponse all items of the questionnaire are missing, while for item nonresponse missing values are confined to specific items of the questionnaire. This distinction has therefore relevant implications on the auxiliary information which is available for ex-post adjustment procedures. For unit nonresponse, the auxiliary

information is usually confined to that obtained from the sampling frame or the data collection process, whereas for item nonresponse one can also exploit the additional information collected during the interview process.

This report focuses on item nonresponse in the Survey of Health, Ageing and Retirement in Europe (SHARE) and the multiple imputation (MI) approach adopted to fill-in the missing values in the key variables of interest. The key idea of the MI approach is to use the distribution of the observed data to predict a set of plausible values for the missing data. The goal is not to get the most accurate predictions of the missing values, but to replace them by plausible values such that inference about population parameters is still valid. Unlike other methods, this approach allows the ultimate users to analyze the observed and imputed data by standard statistical procedures for complete data. Furthermore, combining the set of estimates obtained from repeated applications of the same statistical procedure over multiple complete datasets results in MI inference that properly reflects the uncertainty generated by the imputation of missing values.

Despite recent advances in the survey imputation methodology, the imputation of missing values in large-scale surveys is never a trivial task. Each survey has its own characteristics and presents unique challenges for statistical and economic data analysis. The imputation strategy adopted in SHARE relies on different multiple imputation methods depending on the prevalence of missing values. More precisely, we use multiple hot-deck imputations for non-monetary variables that are usually affected by low/moderate fractions of missing values (lower than 5 percent for the entire sample and lower than 10 percent at the country level). Monetary variables with fractions of missing values above these thresholds are instead imputed jointly by the fully conditional specification (FCS) method (van Buuren et al., 1999, 2006; van Buuren 2007; Raghunathan et al., 2001). In this deliverable of the SERISS project, Work Package 2, Task 2.4, we provide a detailed description of the flexible Stata algorithm used to produce multiple hot-deck imputations (MHDI) of the non-monetary variables. The Stata algorithm used to produce MI imputations of the monetary variables by the FCS method is described in the SERISS deliverable D2.14 (De Luca et al, 2018). Since our focus is on the Stata algorithm used to construct the SHARE public-use MI dataset, we shall assume that the reader has some basic knowledge of both SHARE (http://www.share-project.org/) and Stata.

The remainder of the report is organized as follows. Section 2 gives an overview of the SHARE MI algorithm. Section 3 describes a number of preliminary data-management operations needed to create an intermediate working dataset that imposes a common structure to all variables involved in the imputation process. Section 4 describes the setup of the imputation procedure, while Section 5 describes the implementation of MHDI algorithm. Section 6 concludes.

The Stata do-files used to carry out the imputations described below as well as for the imputations of the monetary variables by the FCS method (De Luca et al, 2018) are made available alongside this report to be adapted as required by other surveys whilst bearing in mind (as noted above) that each survey has its own characteristics – including potentially different survey items and patterns of missing data – and presents unique challenges for statistical and economic data analysis.

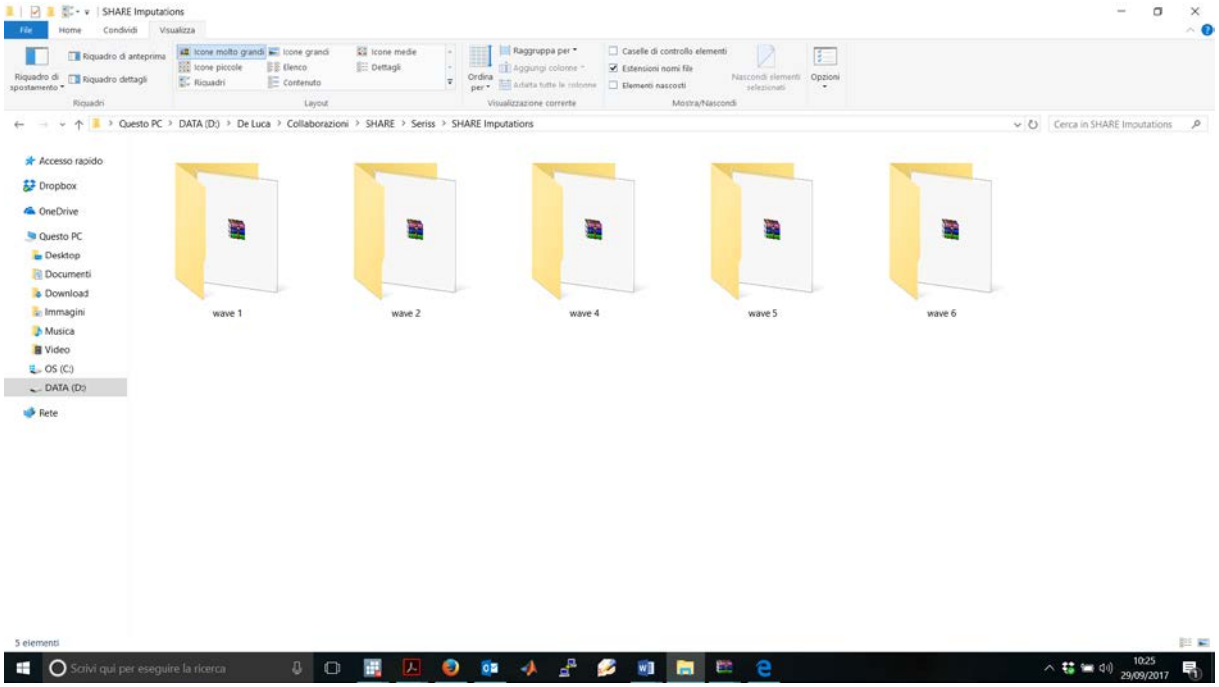
2. Overview of the SHARE imputation algorithm

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a unique panel dataset with rich micro-level information on people aged 50 and older in – as of wave 6 – 19 European countries (Austria, Belgium, Czech Republic, Denmark, Estonia, France, Germany, Greece, Hungary, Ireland, Italy, Luxemburg, Netherlands, Poland, Portugal, Sweden, Slovenia, Spain and Switzerland), plus Israel. SHARE is also harmonized with the US Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA) and has become a model for several ageing surveys worldwide (JSTAR in Japan, CHARLS in China, ELSI in Brazil, KLOSA in South Korea, and LASI in India).

To date, the SHARE data include 5 regular panel waves (2004, 2006, 2011, 2013, 2015) on current living circumstances and one retrospective wave (2008-2009, SHARELIFE) on life histories. The 2017 regular wave will be delivered in 2019, and additional regular and retrospective waves are planned every two years until 2024. The data collected in each wave give a broad picture of life after the age of 50 years, measuring physical and mental health, biomarkers, cognitive abilities, economic and non-economic activities, income and wealth, consumption and health expenditures, expectations, transfers of time and money within and outside the family, as well as life satisfaction and well-being. The scientific value of SHARE rests on three key pillars: (i) its panel design, which captures the dynamic nature of the ageing process and the need of controlling for unobserved heterogeneity; (ii) its multidisciplinary approach, which delivers an integrated picture of individual and societal ageing; and (iii) its cross-nationally harmonized design, which permits international comparisons of health, economic and social outcomes.

Like most other sample surveys, all SHARE cross-sections suffer from unit nonrespose that may lead to biased representations of the target population of interest. To cope with the selection effects due to unit nonresponse, the public-use SHARE data include a set of cross-sectional weights based on the calibration methodology of Deville and Särndal (1992). SHARE is also affected by substantial item nonresponse on key variables, such as income, wealth, and expenditures, collected through a number of open-ended and retrospective questions that are sensitive and difficult to answer. To cope with the loss of precision and the potential selection effects due to item nonresponse, the public-use SHARE data include MI of key survey variables.

Figure 1. Screenshot of the SHARE imputations folder.



As illustrated in Figure 1, the imputation algorithm contains one folder for each regular wave of the SHARE data collection. Two remarks are worth noticing. First, SHARE does not provide imputations of the data collected in SHARELIFE because the static setup of its imputation model is likely to be inconsistent with the dynamic nature of the process that generates missing values on people's life histories.

Second, despite the static setup of the imputation model (i.e. missing values are imputed separately by wave without using lagged variables from previous waves as predetermined predictors), the algorithms of different waves are not completely separated one from each other due to the presence of missing values on follow-up questions that are skipped to save interview time. As discussed in Section 3.4, most missing values on time-invariant variables can be recovered from the data collected in the previous waves of the panel.

In what follows, we shall focus on the MI algorithm of wave 6 because the algorithms of the other waves have a similar structure.

Figure 2. Screenshot of the wave 6 folder.

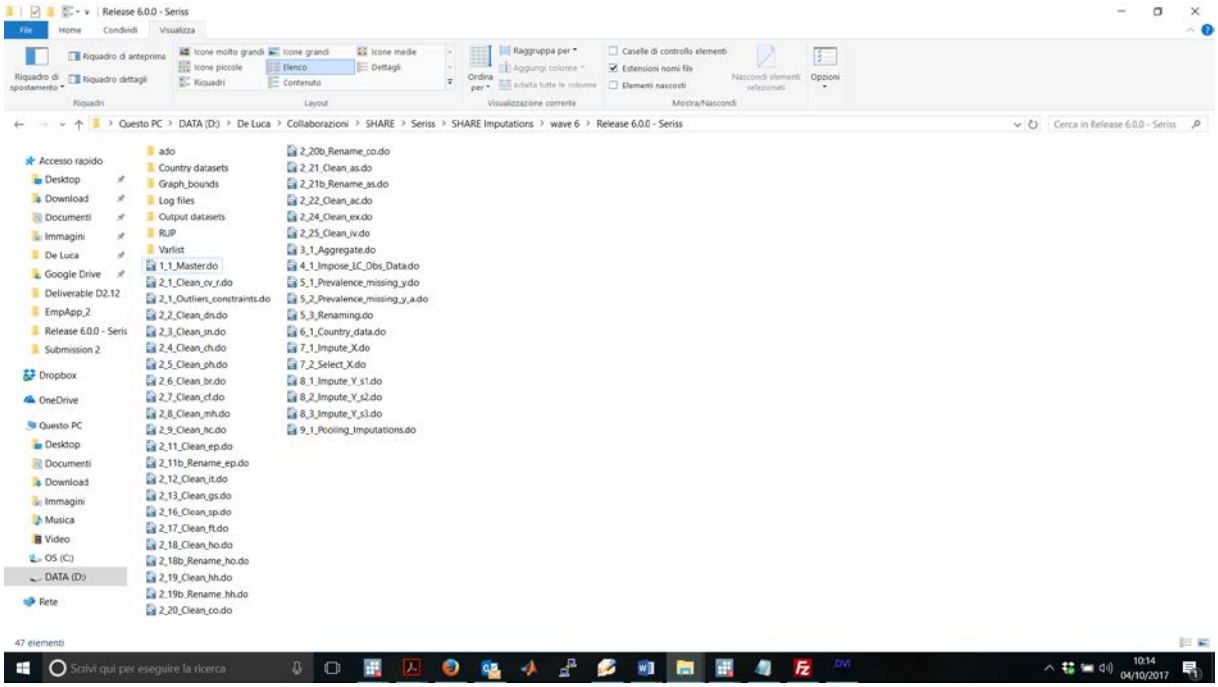


Figure 2 illustrates a screenshot of the wave 6 folder, which contains 40 do-files and 7 subfolders used to store additional do and ado-files (i.e. Stata programs), log-files for tracing the output of imputation algorithm, graphs, and intermediate datasets. The SHARE MI algorithm has a modular structure. In total, there are 9 logical steps (indexed by the first digit in the name of each do-file) and each of them may consist of various sub-steps (indexed by the second digit in the name of each do-file). In the following sections we provide a detailed description of the Stata do-files used to impute the subset of non-monetary variables. The do-files concerned with the imputation of monetary variables are described in the SERISS deliverable D2.14.

3. Working dataset

The first step of the imputation algorithm aims to transform the raw data collected in a given wave of SHARE to create a working dataset that imposes a common structure to all variables involved in the imputation process. Of course, this step of the algorithm is survey-specific as it depends crucially on the structure of the original raw data. Even for SHARE, we had to implement several adjustments to the associated set of do-files for taking into account changes occurred in the interview process of different waves. For other studies, similar or even larger adjustments are necessary that require very good knowledge of the respective datasets and the underlying structures. The creation of a working dataset requires in general a number of data-management operations such as recoding and renaming of the original variables, implementation of specialized routines for dealing with currency conversion issues, branching and skip-patterns, proxy interviews, and partial information on missing values, definition of criteria for outliers, and creation of flag variables for various item nonresponse errors. To our experience, the coding of these operations can be time-consuming and prone to coding errors. The key ingredient for this step of the imputation algorithm is a good knowledge of the raw data and the data management capabilities of Stata.

3.1. General principles to program in Stata

In SHARE, we have implemented a separate do-file for each module of the interview process. We have neglected only few modules (e.g. the Blood Sample and the Peak Flow modules) concerned with the collection of experimental data that do not fit the methodological setup of our imputation model. Few general principles for programming with Stata revealed to be very helpful to manage this step of the imputation algorithm:

a) We never overwrite the raw data. Data-management operations are applied only to backup copies of thereof.
b) We always trace the effects of the data-management operations implemented by each do-file within an associated text file (i.e. a Stata log-file stored in "Log files" folder). For each relevant operation, we show comparisons between the raw data and the transformed data that help identify possible coding errors. Of course, to be useful, the log-files must have a readable format. In that respect, it is important that the do-files make a proper usage of the Stata commands *quietly* and *noisily*.
c) We always comment relevant pieces of do-files and indent each single line of code with a consistent style to improve as much as possible the interpretation and the readability of our programs. Subsequent revisions of the do-files will benefit a lot of the initial effort putted in these operations.
d) We never take for granted the correctness of the raw data. In addition to standard checks performed during the coding stage, our do-files perform some automatic checks on the coherency between the raw data and the instructions of the questionnaire. In SHARE, this helped us to discover several inconsistencies in the raw data. Instructions for correcting these data inconsistencies are usually placed at the beginning of each do-file (after loading the raw data of a given module) until when they are not approved and internalized by the SHARE data-production team.
e) Whenever possible, we try to avoid unnecessarily long do-files by coding similar operations with the aid of loops and auxiliary do/ado-files (such as those stored in "*RUP*" and "*ado*" folders).

3.2. The Master do-file

The file *1_1_Master* contains a distinct section for each step of the SHARE MI algorithm. After setting the current Stata working section (e.g. version of the software, amount of memory, number of processors, seed for random-number generator, current working directory, and directory for user-written ado-files), we define in the first section some flexible features of the algorithm that the researcher may want to change by modifying a single line of code. Among other things, we define in a flexible way the name of various depository folders, the number of multiple imputations, the number of maximum iterations and burn-in iterations for the Gibbs-sampling algorithm discussed in the SERISS deliverable 2.14. In the remaining sections of the file *1_1_Master*, we simply execute in a sequential order all other do-files of the SHARE MI algorithm. This structure allows us to execute the entire MI algorithm by running the file *1_1_Master*. Moreover, it is possible to execute the algorithm on different workstations by changing only one line of code for the path of the "*SHARE imputations*" folder.

3.3. Selection of core variables

The first object of the do-files associated with this step of the SHARE MI algorithm is to select a rather large number of core variables expected to be relevant for either the key purposes of the survey or the missing data mechanism. Notice that the core variables can have various formats: binary, categorical, count, semi-continuous and continuous. Moreover, the subsample of respondents who are eligible for each variable varies across different items of the SHARE questionnaire due to branching, skip patterns, and proxy interviews. For our purposes, this implies that the raw data contain the so-called "missing values by design" (i.e. missing values generated by the design of the interview process). Since this type of missing values requires special attention in the subsequent steps of the imputation algorithm, we have created a binary eligibility indicator for each core variable that takes value 1 for the subsample of eligible respondents and value 0 otherwise. Conventionally, we name the binary eligibility indicator of a generic variable *y* as *y_r*.

3.4. Editing of non-monetary variables

In SHARE, most of the editing operations required for binary, categorical and count variables are straightforward. Sometimes, we have used a slightly more complex syntax for some specific purposes. Below, we discuss two examples taken from the do-file *2_2_Clean_dn.* The first example concerns cases where we need to recover the missing information due to follow-up questions on time-invariant variables.

Figure 3. Syntax of the editing operations for citizenship.

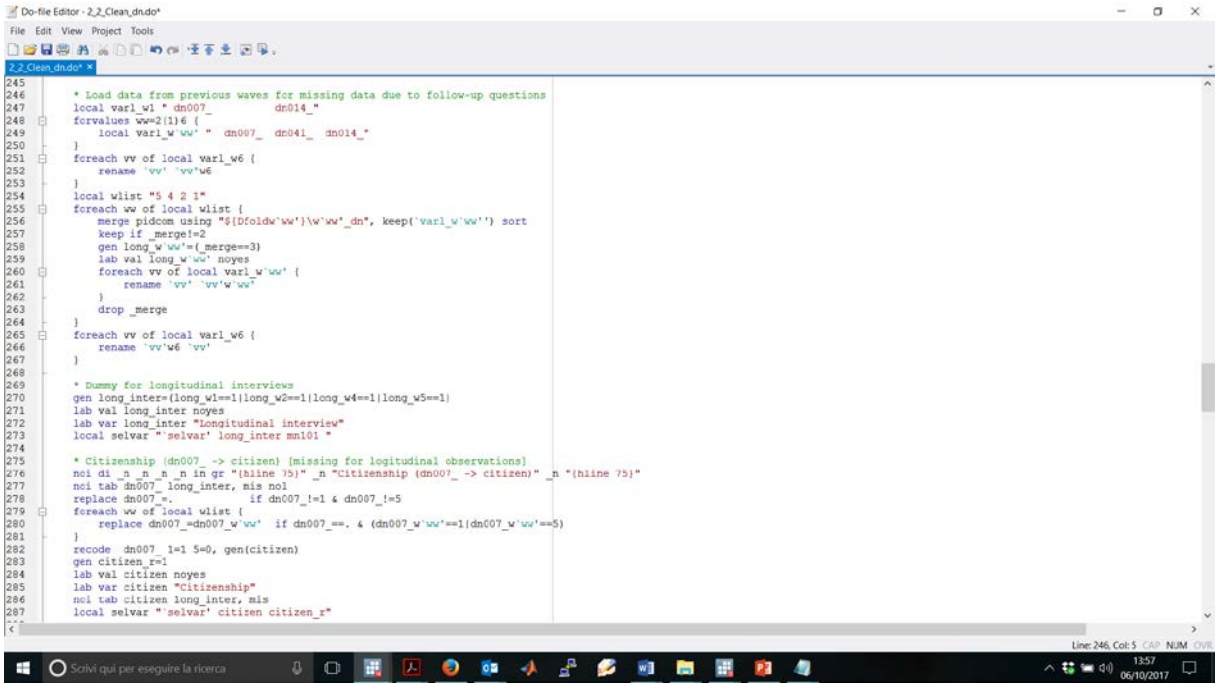


Figure 3 illustrates the syntax used for dealing with this issue in the data-management operations for the citizenship dummy variable. There are three logical stages here. First, we load data from previous waves. Second, we generate a binary variable for longitudinal interviews. Third, we recover the missing information in the variable *dn007_* by using the time-invariant data collected in the previous waves.

Figure 4. Output of the editing operations for citizenship.

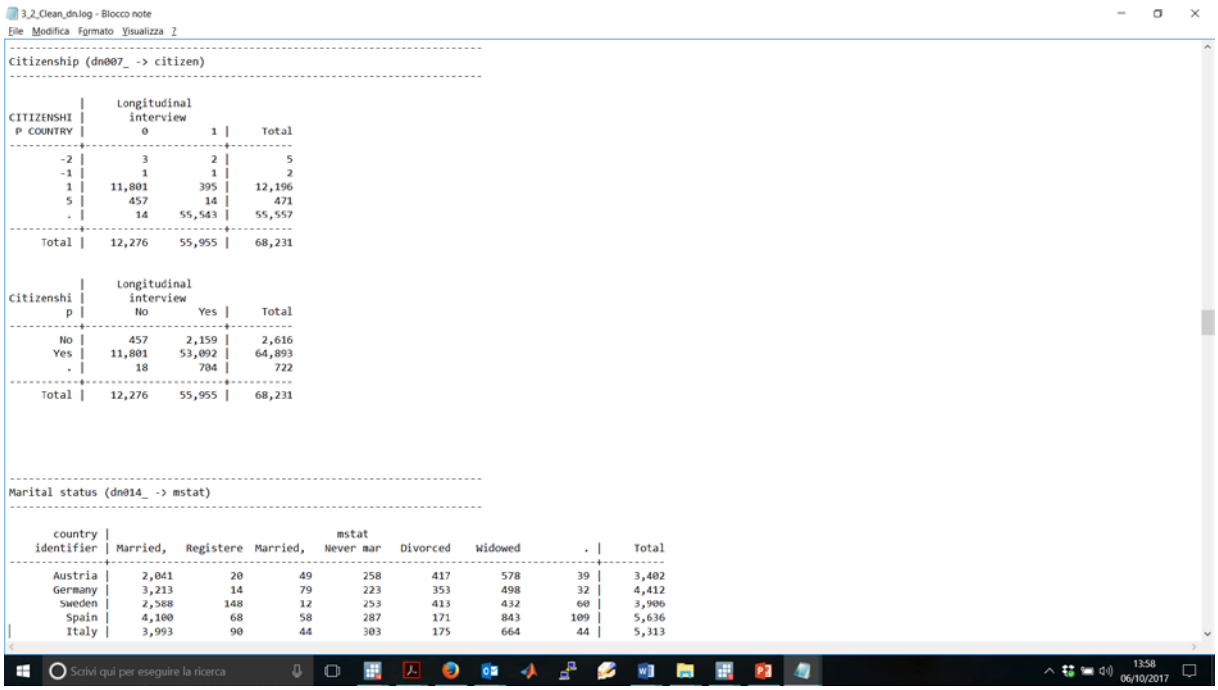


Figure 4 illustrates the output of this code, as it appears in the log-file associated with this do-file. The first tabulation shows that the variable *dn007_* has apparently a huge amount of missing values. This problem arises because the CAPI instrument of wave 6 does not ask the citizenship question to most people interviewed in some previous wave of study. By exploiting the data collected in the previous waves, our code allows us to recover most of these missing values. Finally, we have only 722 missing observations (i.e. about 1% of the sample).

Our second example is about the syntax of the editing operations for ISCED coding of education, for which the SHARE data-production team have already implemented a specialized set of Stata do-files for mapping the country-specific educational categories into international comparable data. Here we avoid reproducing this complex set of operations because this strategy could be subject to coding errors.

Figure 5. Syntax of the editing operations for ISCED coding of education.

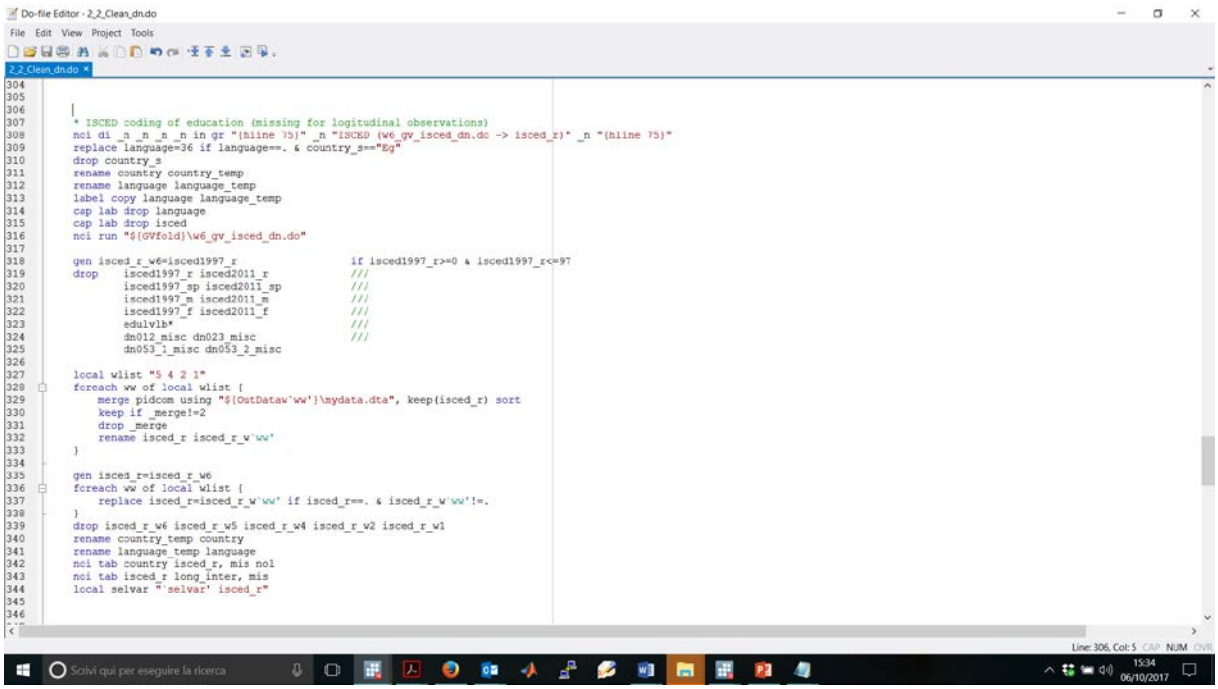
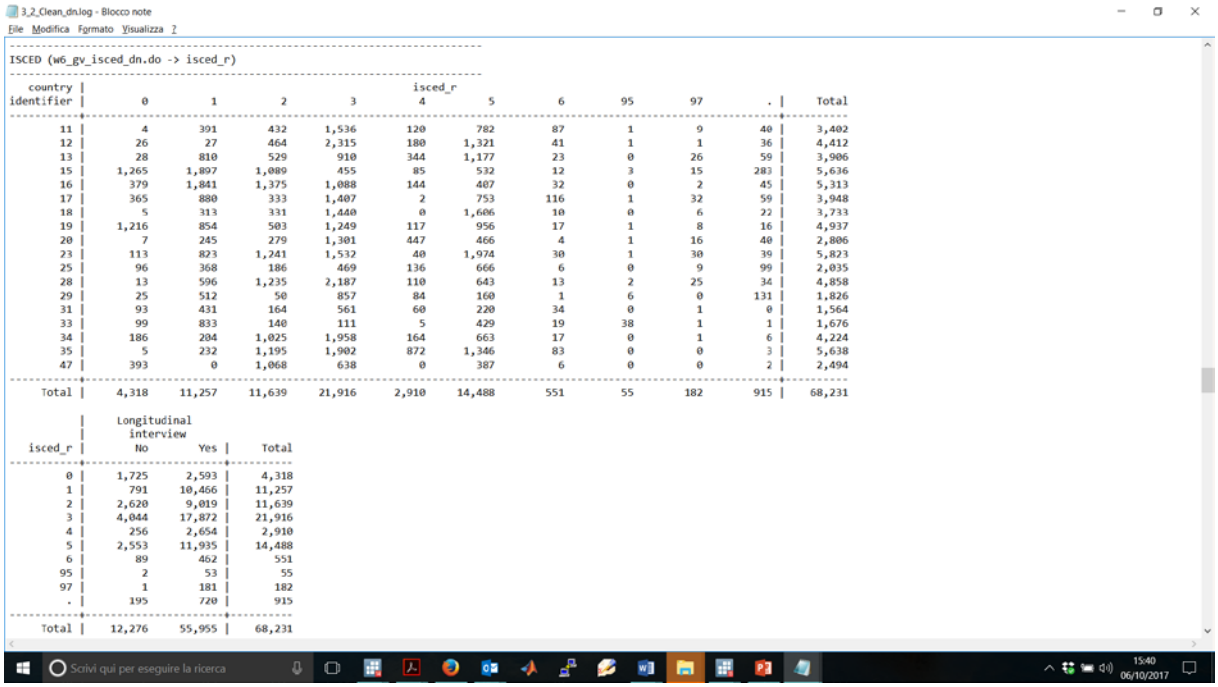


Figure 5 illustrate our implementation of the editing operation for ISCED coding of education. More precisely, in line 316, we execute the do-files written by our experts in this field on the raw data of wave 6. Like the data-management operations used to construct the citizenship dummy variable, we then recover (lines 318-339) the missing values for ISCED coding of education using the data from previous waves.

Figure 6 presents the output of this code. Notice that ISCED coding of education is one of the core variables used in the subsequent steps of our imputation algorithm. In this step of the algorithm, we do not want to impute the missing values of this variable. We only notice that non-substantial answers (i.e. “Don’t know” and “Refusal”) to the items required for the ISCED coding of education represent a negligible fraction of the sample (less than 1.5%). Moreover, our tabulations do not show evident signals of coding errors.

Figure 6. Output of the editing operations for ISCED coding of education.



3_2_Clean_dn.log - Blocco note

File Modifica Formato Visualizza ?

ISCED (w6_gv_isced_dn.do -> isced_r)

country identifier	isced_r 0	1	2	3	4	5	6	95	97	.	Total
11	4	391	432	1,536	120	782	87	1	9	40	3,402
12	26	27	464	2,315	180	1,321	41	1	1	36	4,412
13	28	810	529	910	344	1,177	23	0	26	59	3,906
15	1,265	1,897	1,089	455	85	532	12	3	15	283	5,636
16	379	1,841	1,375	1,088	144	407	32	0	2	45	5,313
17	365	880	333	1,407	2	753	116	1	32	59	3,948
18	5	313	331	1,440	0	1,606	10	0	6	22	3,733
19	1,216	854	503	1,249	117	956	17	1	8	16	4,937
20	7	245	279	1,301	447	466	4	1	16	40	2,806
23	113	823	1,241	1,532	40	1,974	30	1	30	39	5,823
25	96	368	186	469	136	666	6	0	9	99	2,035
28	13	596	1,235	2,187	110	643	13	2	25	34	4,858
29	25	512	50	857	84	160	1	6	0	131	1,826
31	93	431	164	561	60	220	34	0	1	0	1,564
33	99	833	140	111	5	429	19	38	1	1	1,676
34	186	204	1,025	1,958	164	663	17	0	1	6	4,224
35	5	232	1,195	1,902	872	1,346	83	0	0	3	5,638
47	393	0	1,068	638	0	387	6	0	0	2	2,494
Total	4,318	11,257	11,639	21,916	2,910	14,488	551	55	182	915	68,231

isced_r	Longitudinal interview No	Yes	Total
0	1,725	2,593	4,318
1	791	10,466	11,257
2	2,620	9,019	11,639
3	4,044	17,872	21,916
4	256	2,654	2,910
5	2,553	11,935	14,488
6	89	462	551
95	2	53	55
97	1	181	182
.	195	720	915
Total	12,276	55,955	68,231

Scrivi qui per eseguire la ricerca

15:40 06/10/2017

4. Setup of the imputation process

So far, we have implemented only a number of preliminary editing operations aimed to standardize as much as possible the structure of our working dataset. However, the variables included in the file "*mydata_final.dta*" still contain missing values. The setup of the SHARE imputation process starts in the do-file *6_1_Country_data*, where we split the original SHARE sample in nineteen country-level datasets with an expanded number of observations to initialize the MI datasets. These datasets are stored in the country-specific subfolders of the "*Country datasets*" folder under the name "*imp_data*".

5. Multiple hot-deck imputations of non-monetary variables

In the do-file *7_1_Impute_X*, we produce multiple hot-deck imputations of the non-monetary variables that are affected by negligible amounts of missing values. The hot-deck method involves replacing missing values of one or more variables for a nonrespondent (called the recipient) with observed values from a respondent (the donor) that is similar to the nonrespondent with respect to a set of observable characteristics and is selected randomly from a set of potential donors (see, e.g., Andridge and Little 2010). Specifically, we apply an ordered sequence of univariate hot-deck imputations separately by country. For each variable, we always create multiple imputations of its missing values using independent draws of the donors in the various hot-deck classes.

In this step of the algorithm, we first impute few missing values for basic socio-demographic characteristics (namely, gender, age, education and self-reported health status) which are then used as conditioning predictors when performing hot-deck imputations of the other variables. For some non-monetary variables, we use a slightly larger conditioning set. For

example, our set of conditional predictors also includes the number of children when imputing the number of grandchildren and an indicator for being a patient in a hospital overnight during the last year when imputing other health-related variables. For variables that are logically related to each other (e.g. such as weight, height and body mass index), we create multivariate hot-deck imputations by drawing the imputed values of two more variables simultaneously from the same donor.

Figure 7. Hot-deck imputations for age and years of education.

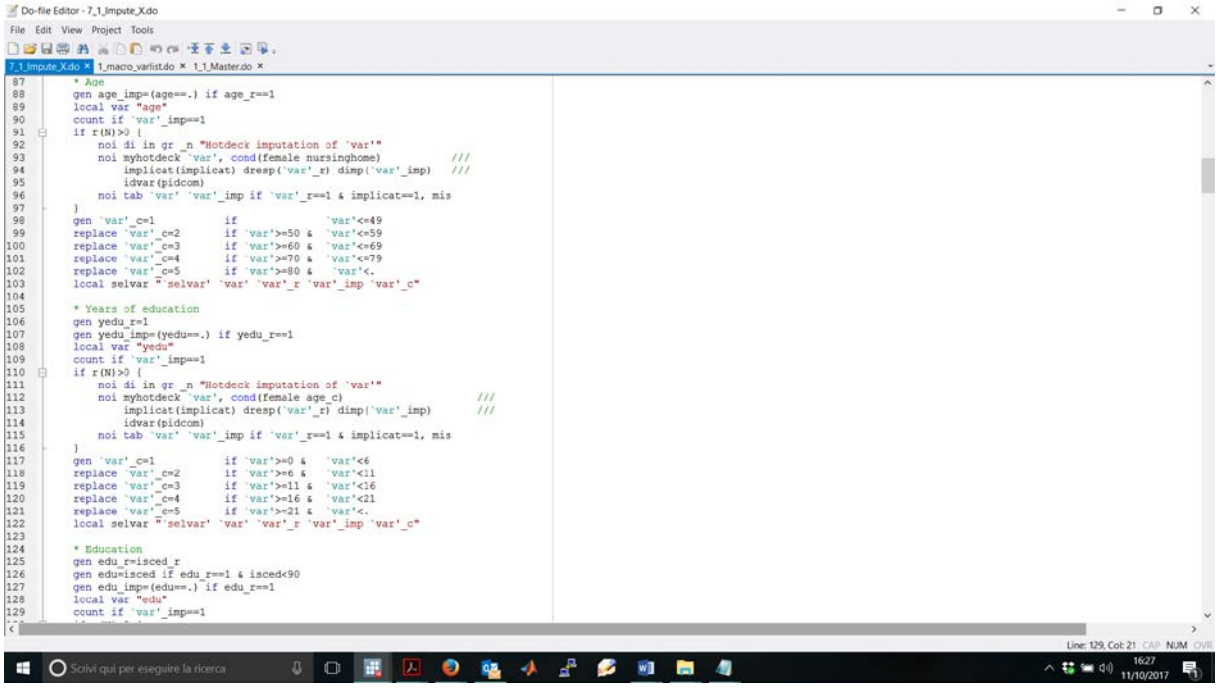


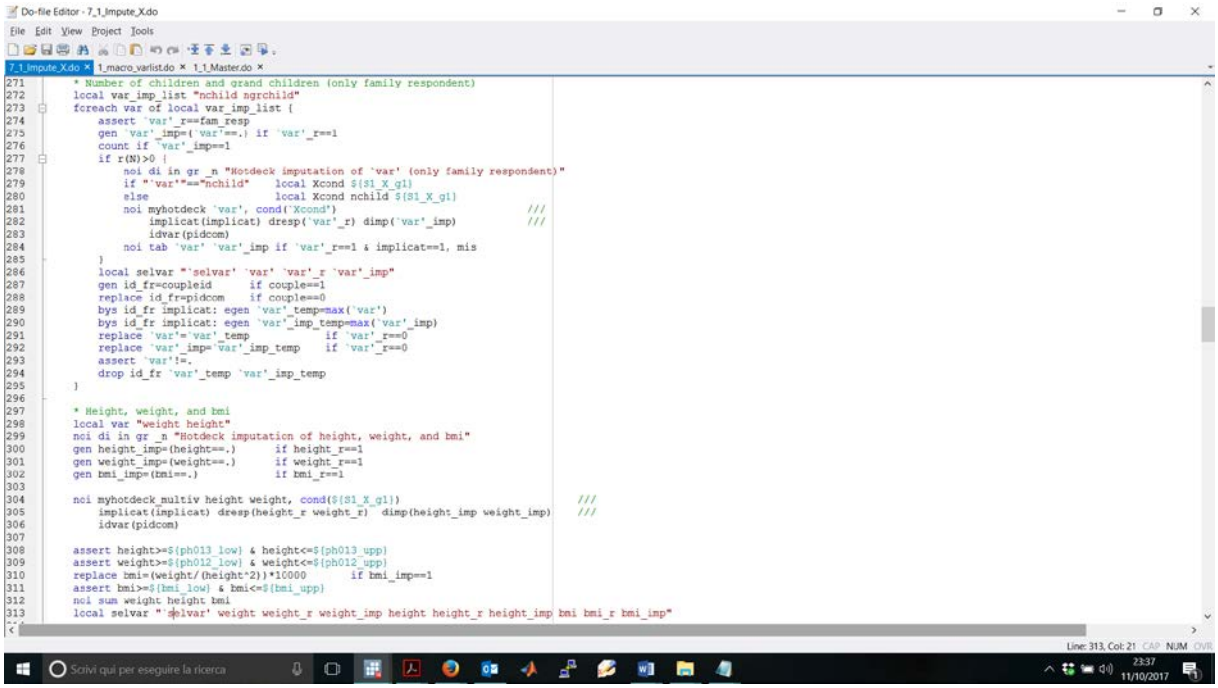
Figure 7 shows the syntax used to create hot-deck imputations of the few missing values on the age variable. Conditional on eligibility, we first construct a binary missing data indicator that takes value 1 for the missing values and value 0 for the complete cases. Provided the age variable contains some missing values, we generate multiple hot-deck imputations by the ado-file *myhotdeck* stored in the "*ado*" folder. This command allows us to specify as flexible arguments the list of variables used to construct the conditioning set, the count variable used to index multiple imputations, the binary eligibility indicator, the binary missing data indicator, and the respondent's identifier. After imputing the missing values, we generate an age-group indicator that is used as conditional predictor in the hot-deck imputations of the other variables. For example, when imputing years of education, our conditioning set includes the binary indicator for being female and the age-group indicator.

Figure 8 shows the syntax used to create hot-deck imputations for number of children, number of grandchildren, weight, height, and body mass index. There are three worth noticing aspects. First, eligibility to the first two variables is restricted to the subsample of household respondents only. When creating hot-deck imputations, we take into account this restriction by the binary eligibility indicator associated with each variable. Notice however that, after imputing the missing values for the household respondents, we also extend the imputed values of each replicate to the other household members.

Second, the conditioning set usually consists of the list of variables specified in the global macro *S1_X_g1* (see the do-file *1_macro_varlist* stored within the *RUP* folder). The only

exception is the conditioning set used to impute the number of grandchildren, which includes the number of children as an additional predictor.

Figure 8. Hot-deck imputations for number of children, number of grandchildren, weight, height and body mass index.



Third, for the two variables regarding respondent's weight and height, we create multivariate hot-deck imputations by the ado-file *myhotdeck_multiv*. This command has the same syntax of the *myhotdeck* command, but it allows us to impute jointly two or more variables. After imputing these two variables, we also impute passively body mass index (i.e. we use the imputed values of respondent's weight and height to determine the imputed values of body mass index).

In total, the do-file *7_1_Impute_X* allows us to create multiple hot-deck imputations for 69 variables. For each missing value, we generate *M=5* independent imputations. The resulting datasets are stored in the country-specific subfolders of the "*Country datasets*" folder under the name "*imp_data_x*".

6. Conclusions

In this deliverable of the SERISS project, Work Package 2, Task 2.4, we have described the Stata algorithm used to create the public-use MI datasets in the first six waves of SHARE (excluding the retrospective wave in 2008). Although the modular structure of the programs can theoretically be adapted to a variety of other datasets, some features of the algorithm are unavoidably tailored to the specific structure of the SHARE data. First, we have transformed and recoded a large number of variables of varying type (e.g., binary, ordinal, nominal, count, semi-continuous, and continuous) using flexible and efficient Stata programs that account for possible skip patterns, partial information on the missing values, and currency conversion issues. Then, we have created multiple hot-deck imputations of the

missing values of non-monetary variables that suffer from moderate item nonresponse. This imputation method does not rely on model fitting for the variables to be imputed and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model, such as regression imputation. A disadvantage is that some assumptions are implicit in the choice of the metric to match donors to recipients and the observable variables included in this metric. Moreover, sequences of univariate hot-deck imputations may not preserve the correlation structure of the imputed variables. In the SHARE MI algorithm, we address this challenging issue only when imputing the missing values of the monetary variables that suffer from substantial item nonresponse. For a description of the iterative FCS algorithm used to impute this type of variables, we refer the reader to the SERISS deliverable D2.14 (De Luca et al, 2018).

References

Andridge, R. R., R. J. A. Little (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* 78: 40-64.

De Luca, G. (2018) *Flexible Stata Code for Multiple FCS Imputations of Monetary Variables in SHARE. Deliverable 2.14 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: www.seriss.eu/resources/deliverables

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85–95.

Rubin, D. B. (1996). Multiple imputations after 18+ years. *Journal of the American Statistical Association* 91: 473-489.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219-242.

Van Buuren, S., J. P. L. Brands, C. G. M. Groothuis-Oudshoorn and D. B. Rubin (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 76: 1049-1064.

Van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681–694.