Working Paper 8

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iii): Software tools and international collaboration

Editing and Imputation in Household Based Surveys-Case of Household Budget Survey in Bosnia and Herzegovina

Prepared by Edin Šabanović, Agency for Statistics of Bosnia and Herzegovina, Bosnia and Herzegovina

I. Introduction

Non-sampling errors are present in every statistical survey, regardless of whether it is based on full coverage or sample. As a consequence of such errors we deal with missing data in our surveys.

There are two main types of missing data: unit and item non-response. While the problem of unit non-response is usually solved by weighting techniques, the problem of item non-response is a subject of imputations methods. Editing and imputation are very important phases in the generic statistical business production model (GSBPM). Methods of editing and imputation that were used in the process phase of the GSBPM have an impact to survey estimates in terms of their accuracy, reliability and timeliness.

This paper gives an overview of methods of editing and imputations, as well as software used in household based surveys in Bosnia and Herzegovina in last fifteen years with a specific focus on Household Budget Survey. It shows an improvement in editing and imputation approaches from applying very simple to more advanced methods.

II. Data Editing and Imputation in Living Standards Measurement Survey

A. Background

Living Standards Measurement Survey (LSMS) was the first survey conducted in the post-war period in Bosnia and Herzegovina. In addition to physical destruction and loss of life, the war has had a very huge impact to the worsening of social status and living standard of the population. From the other hand, statistical offices did not have any reliable statistical data on population and households in the country. Population census data from 1991 were useless because of significant population moving and housing destructions during the war. There was a lack of relevant demographic and socio-economic data and it was the reason for designing the LSMS as a very complex household survey. It consisted of 12 modules: Demographic characteristics of households, Housing, Education, Health, Labour activity, Loans, Vouchers/Certificates, Migration, Social welfare, Household consumption, Non-agricultural activities and Agricultural activities. The survey main goal was to collect data on basic demographic and socio-economic characteristics of population and households, which should be used for creation of different policies at the level of state and entities.

B. Data editing and imputation in LSMS BiH

LSMS was conducted in 2001 and its data were analysed and published in 2002. Data was processed in SPSS, both editing and imputation and data analysis. In that period, this software was the only statistical software in use in statistical offices in Bosnia and Herzegovina.

In the first phase of the data processing, methods of descriptive data analysis were applied and all survey variables were checked in terms of their completeness, frequency distribution, measures of central tendency, measures of variability and skewness, as well as consistencies with other variables according the questionnaires. This analysis was the base for error detection and a number of edit rules was created in order to localise item non-responses, which were flagged for the imputation.

Imputation of the categorical variables were made by strictly adjustments of the data to skip patterns in the questionnaires and by implementing simple "if-then" rules in SPSS, which imputed missing or erroneous items. Imputed values of these variables were selected by choosing most logic and/or most frequent (Mode) answers from available data set.

Missing or erroneous data on continuous variables were imputed by simple median imputation method. Median was selected as a measure for imputation because it is not sensitive to extreme values and because it destroys less variability of data in comparison to the mean. The median imputation rules were designed in a hierarchical (bottom-up) way: the median of available data on the analysed variable at the level of enumeration area was used as a imputed value if at least five observations were recorded at the level of enumeration area. If less than five observations were available, median at the level of municipality was used under the same condition: existence of at least five observations at that level. If this condition was not satisfied, the median at the entity level was used. And finally, if there were less than five observations available at entity level, median of the analysed variable at the state level was used for the imputation.

All imputation rules were programmed and performed in SPSS and the complete data bases were prepared for the final data analysis.

II. Data Editing and Imputation in Household Budget Survey

A. Background

Household Budget Survey (HBS) in Bosnia and Herzegovina was not conducted in regular periodicity. Up to now, it was conducted three times: in 2004, 2007 and 2011. The newest round is in process in this year. The main goal of this survey is to collect data for the purposes of consumer price statistics, national accounts statistics, as well as to measure living standard and to analyse poverty.

The HBS is also a very complex survey and its complexity is expressed in the burden of respondents and the scope and the number of questionnaires and modules. Data collection period is the whole calendar years and it is done through following instruments:

(a) Diary of Purchase (two week self-completion questionnaire)
(b) Self-consumption Diary (two week self-completion questionnaire)
(c) Final Interview (face-to-face interview)
(d) Social inclusion questionnaires: Social Inclusion and Migration Module and Health Module (face-to-face interview)

Considering the length of the data collection within the diaries, very complex survey questionnaires and a variety of variables in them, it is evident that a large amount of data on many categorical and continuous variables is collected. Of course, there are non-sampling errors in data collected because no one statistical survey is immune from such errors.

Statistical techniques of data editing and imputation were used to localise and correct erroneous or missing data. Generally, all items with errors were considered as item non-responses and consequently were treated by means of imputation, i.e. through the replacement of the missing item with a most plausible value.

Quantitative and qualitative checks were performed in order to detect and correct all errors in data. This process is described in following flow-chart:

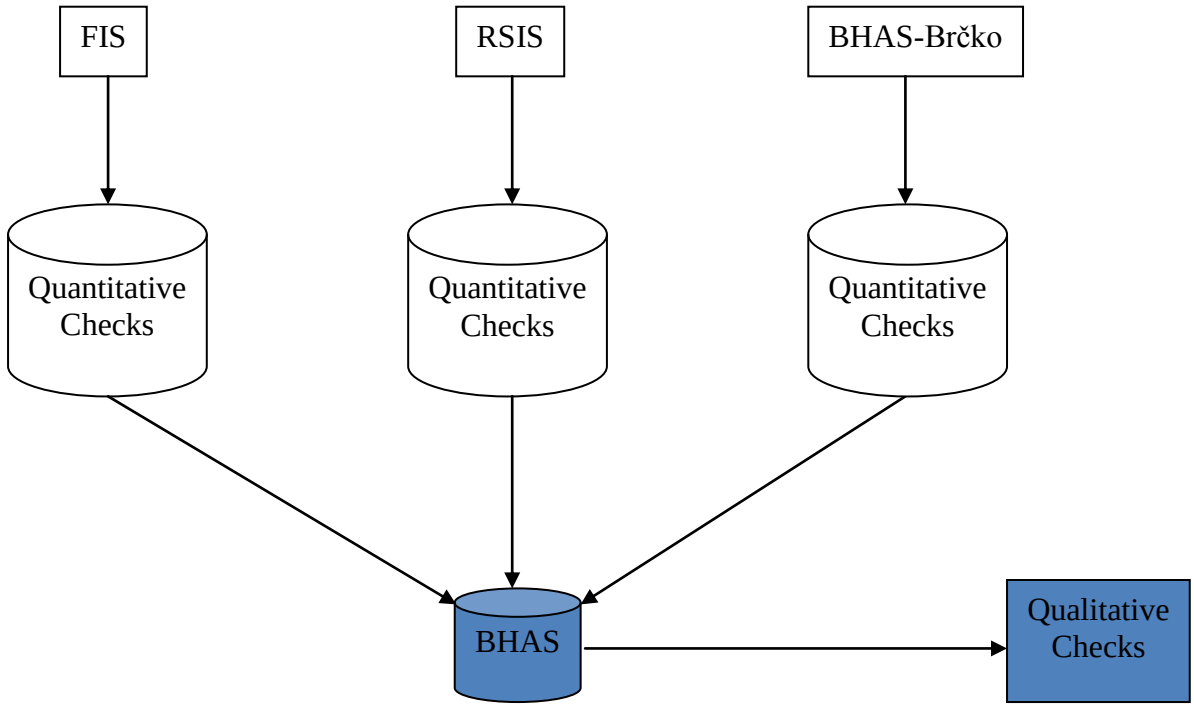
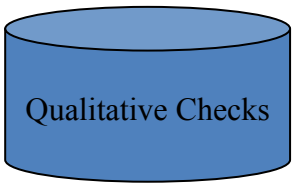


Figure 1. Flow-chart of editing and imputations procedures in HBS BiH

Quantitative checks were related to following issues:

(a) Verification of units to be sampled and units observed
(b) Units observed and units recorded (after data entry in Blaise).
- Verify if the household code is duplicated or not;
- Verify if the household code is out of range;
- Verify if the main parts of the questionnaire are not empty.

Qualitative checks were performed separately for categorical and numerical variables as described in following flow-chart:



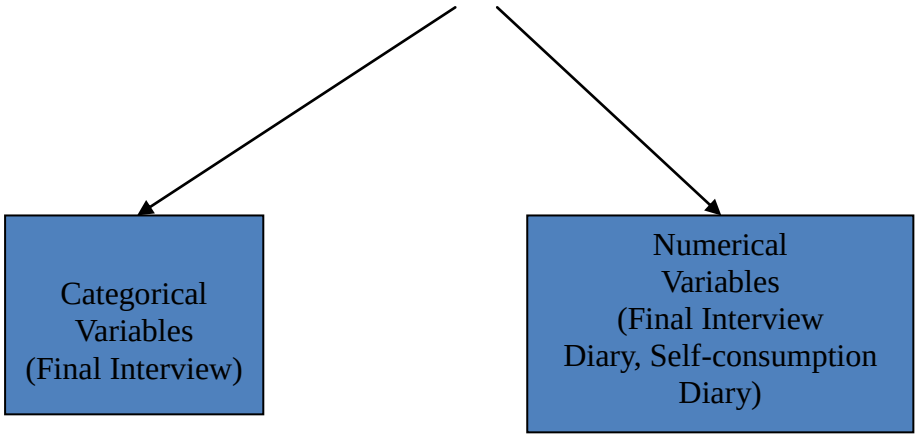


Figure 2. Flow-chart of the quantitative checks in HBS BiH

Two different sub-processes composed the overall editing and imputation procedure. Methods used for editing and imputation of categorical and continuous variables are described in following paragraphs.

B. Data editing and imputation of categorical variables

Data editing and imputation of categorical variables in HBS consisted of two steps:

(a) Deterministic step
(b) Probabilistic step

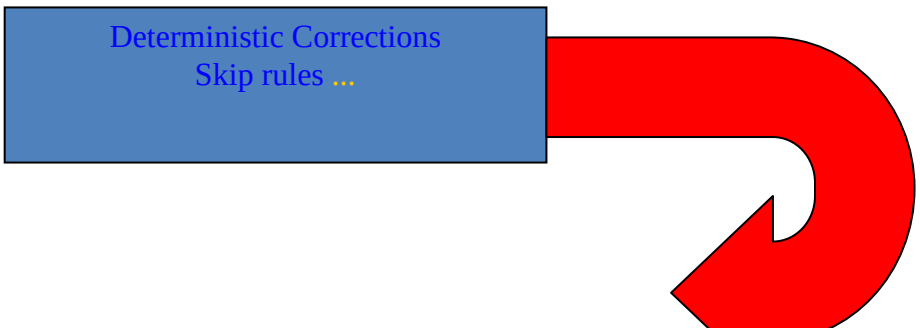
The whole process is described in the flow-chart in figure 3.

The goal of the deterministic step was to adjust key variables of the sampling unit (e.g. the identification code, some skip rules variables, etc.) or to deal with systematic errors. All deterministic checks were performed in developed tailor made C++ codes and they were the basis for the probabilistic step.

Deterministic corrections were related to following items:

(a) Sum of the 14 daily expenses in an only expense
(b) Deterministic rules (IF …THEN rules):

- If sex of spouse (not head) is equal to sex of head, change the first one
- Insert random building year of the house where this information is wrong or not available
- If appliance to make hot water is available, then hot water must be available
- If there are electric equipment, electric power must be available
- Members aged less than 15 years cannot have current activity status
- Members aged 15 years or over must have current activity status, etc.



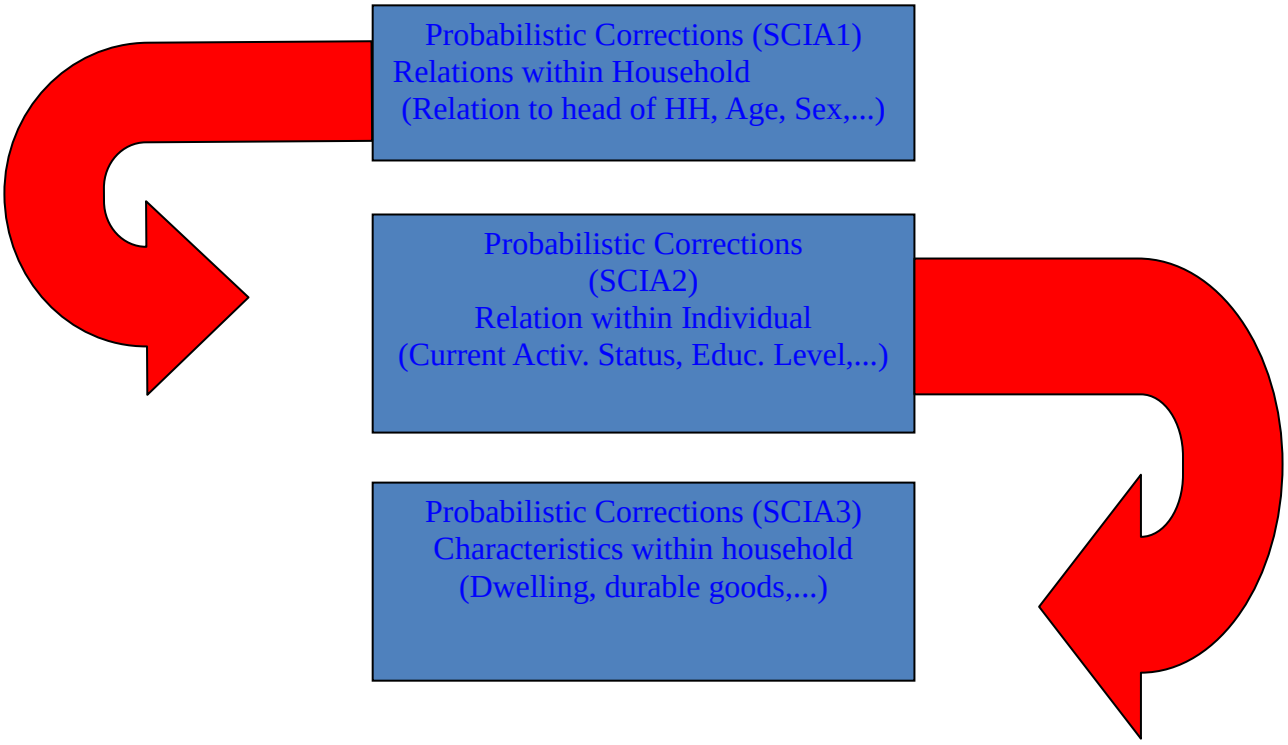


Figure 3. Flow-chart of the qualitative checks of categorical variables in HBS BiH

Within the probabilistic step, the error localisation of categorical variables was performed by using Fellegi-Holt approach, while for the treatment of missing items the hot-deck imputation procedure was adopted. This approaches had three goals[1]:

(a) The data in each record should be made to satisfy all edits by changing the fewest possible variables (fields);
(b) Imputation rules should derive automatically from edit rules;
(c) When imputation is necessary, it should maintain the joint distribution of variables.

The implementation of the probabilistic step procedures was done in SCIA, a software developed in ISTAT-Italian National Statistical Institute and implementing both, the Fellegi-Holt algorithm for error localisation and the nearest neighbour hot-deck technique for imputation. The application of this step was divided into three different hierarchical procedures.

In the first procedure, household typology variables were a subject of data editing, i.e. all the variables (generally demographic variables) characterising the household, and for which edit rules specified relationships among household members (e.g. relationship to household head, marital status, etc.). The goal of this phase was to make the members coherent with respect to the household composition. The edited variables were: Relationship to the head of household, Sex, Year of birth and Marital status. This part of the program is called SCIA1.

The second phase of the probabilistic step was concerned to editing of all variables concerning the characteristics of the single units (person-wise level, i.e. household member) and for which edit rules specified relationships among variables within the units. The variables are: Education level, Current activity status, Branch of economic activity, Occupation, Professional status, Type of work contract or

[1] William E. Winkler and Bor-Chung Chen (2001), " Extending the Fellegi-Holt Model of Statistical Data Editing", Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001, p. 1.

activity, Property sector and Branch of economic activity. Of course all the variables corrected in the first phase could not be modified. The records of this phase were obtained by a fragmentation of the household: in other words a household with two members was decomposed in two units. This is done in SCIA2.

Finally, a third phase was performed. At that stage, the editing and imputation of the remaining variables of the "Final Interview" questionnaire were carried out. These variables were related to dwelling, durable goods, expenditure habits, etc. The target unit in this case was the same of the first phase, i.e. the household. Thus, between the second and the third phase, a further code for passing from individual record to household record was required. Following the previous example, two members belonging to the same household must be considered as a unique record (household). These procedures were performed in SCIA3.

Once all the categorical variables had been edited and imputed, numerical variables were checked.

C. Data editing and imputation of continuous variables

Continuous variables were present in all survey questionnaires. The editing procedure of these variables (mainly related to consumption expenditures) was based on the idea that variable values (or ratio variable values) must be located in certain pre-defined and acceptable intervals. The bounds of these intervals were estimated both, by data analysis and by subject matter experts. An observation was classified as outlier if it did not belong to an opportune pre-defined interval. Data editing was concerned to the identification and investigation of missing data and outliers, i.e. observations reporting suspicious high/low values of the variables inspected. Once an outlier was classified as erroneous, it was treated as a missing item. Moreover, there was the additional problem that it was important to distinguish between missing item and zero expenditure. The missing items that must be imputed were identified through the flags recorded in the questionnaire. It is worth mentioning that the flags were checked in the categorical editing phase previously described.

Once the editing phase was completed, missing items were imputed by using two methods:

(a) Donor imputation (RIDA module of CONCORD)
(b) Model based imputations

Imputations were performed according to the following scheme:

(a) Numerical variables in the Final interview and some of them in the Diary of Purchase and in the Self-consumption diary, for which only the data related to the consumption expenditure was reported, were imputed through the nearest neighbour hot-deck method;
(b) Variables in the Diary and Self-consumption diary, for which both, the prices and quantities purchased or consumed were reported, were imputed according to the two strategies described in the following. They were based on the use of nearest neighbour hot-deck techniques and random draws from a specific constrained distribution family.

CA. Numerical variables in the Final interview where only the consumption expenditure amount is requested

As already said, these variables were imputed by means of nearest neighbour hot-deck technique. This imputation method consisted of choosing, for a record to be imputed (recipient record), the most similar correct record (donor record) that will donate the values of the variables needed to complete the recipient record. It was useful to create homogeneous strata (imputation cells), where both the recipient and donor must belong. Furthermore, it was necessary to define a set of variables (matching variables), on which the similarity through an opportune distance will be computed. Hence, it was natural to state that the stratification and matching variables must strongly characterise the observations. They have been

chosen both, following data analysis and using the knowledge acquired in the Italian Household Budget Survey.

CB. Numerical variables in the Diary of purchase where both, the consumption expenditures and the purchased quantities are requested

Focusing on a single item, let us define with S_i, i = 1, …, 14 the amount of consumption expenditure of the i-th day, and with Q_i, i = 1,…,14 the quantity purchased of the variable under analysis.

Let S and Q be the total amount of expenditures and quantities related to 14 days:

$S_1 + S_2 + \ldots + S_{14} = S$
$Q_1 + Q_2 + \ldots + Q_{14} = Q$

Moreover, let S_L and S_U the lower and upper limits of the acceptance intervals for the variable S, and let P_L and P_U be the lower and upper limits for the acceptance intervals for the price per unit P = S/Q.

Denoting with RS the rule "the variable S belongs to the acceptance interval $S_L < S < S_U$" and with RP the rule "the variable P belongs to the acceptance interval $P_L < S/Q < P_U$", the flow of the procedure is described in the following figure[2]:

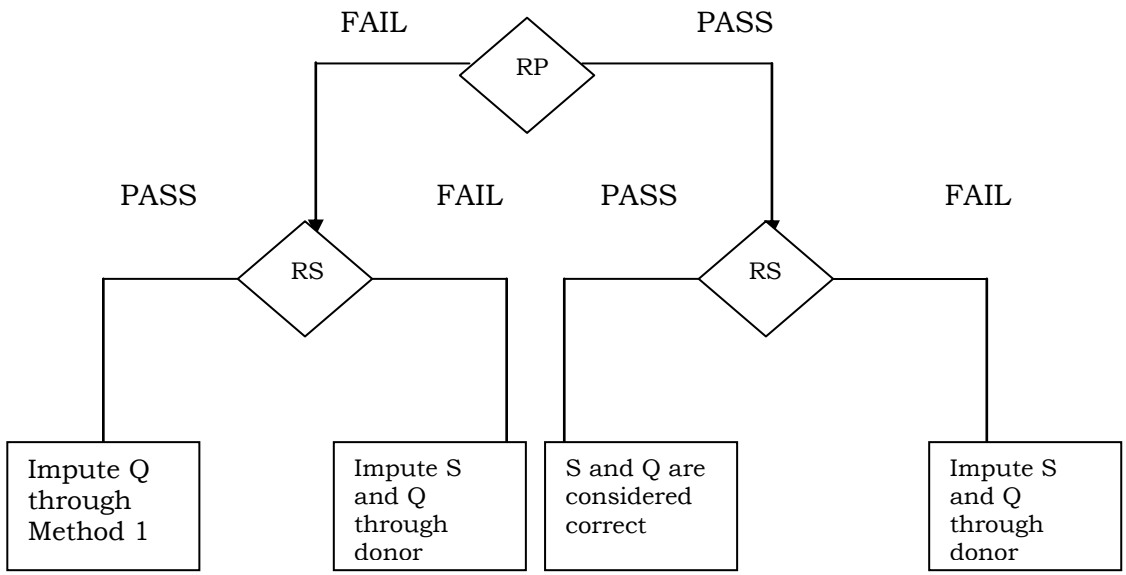


Figure 4. Flow-chart of the imputation procedure for numerical variables in the Diary of purchase

The imputations of such variables is made by Method 1, which is described by following procedures:

(a) Compute the 1st (q_1) and 3rd (q_3) quartile of the frequency distribution of S/Q;
(b) Generate a random draw (Q*) from the uniform distribution on the interval [S/q_3, S/q_1];
(c) Impute the missing item with Q*.

[2] Agency for Statistics of Bosnia and Herzegovina, Institute of Statistics of the Federation of Bosnia and Herzegovina, Institute of Statistics of the Republika Srpska, District Brcko Branch Office (2005), " The BiH Household Budget Survey 2004 – Methodology", Sarajevo, 2005., p. 38.

This imputation has the property that S/Q* is included in the interval (q_1, q_3).

CC. Variables in the Self-consumption diary where both, the expenditure and the consumption quantities are requested

In this case we made the hypothesis that the most reliable variable is the quantity. Furthermore, since quantities consumed in the reference period were recorded in this diary, it made sense to introduce the per capita quantities variables.

Focusing on a single item, let us define with S_i , i = 1,…,14 the amount of consumption expenditure of the i-th day, and with Q_i , i = 1,…,14 quantity consumed of the variable under analysis.

Let S and Q be the total amount of consumption expenditures and quantities related to 14 days:

$S_1 + S_2 + \ldots + S_{14} = S$
$Q_1 + Q_2 + \ldots + Q_{14} = Q$

Moreover, let QP_L and QP_U the bounds of the acceptance intervals for the variable Q/ncomp (ncomp = number of household members) and let P_L and P_U the bounds of the acceptance intervals for the variable price per unit P = S/Q.

Denoting with RQ the rule "the variable Q/ncomp belongs to the acceptance interval QP_L < Q/ncomp < QP_U" and with RP the rule "the variable P belongs to the acceptance interval P_L < S/Q < P_U", the flow of the procedure is described in the following figure[3]:

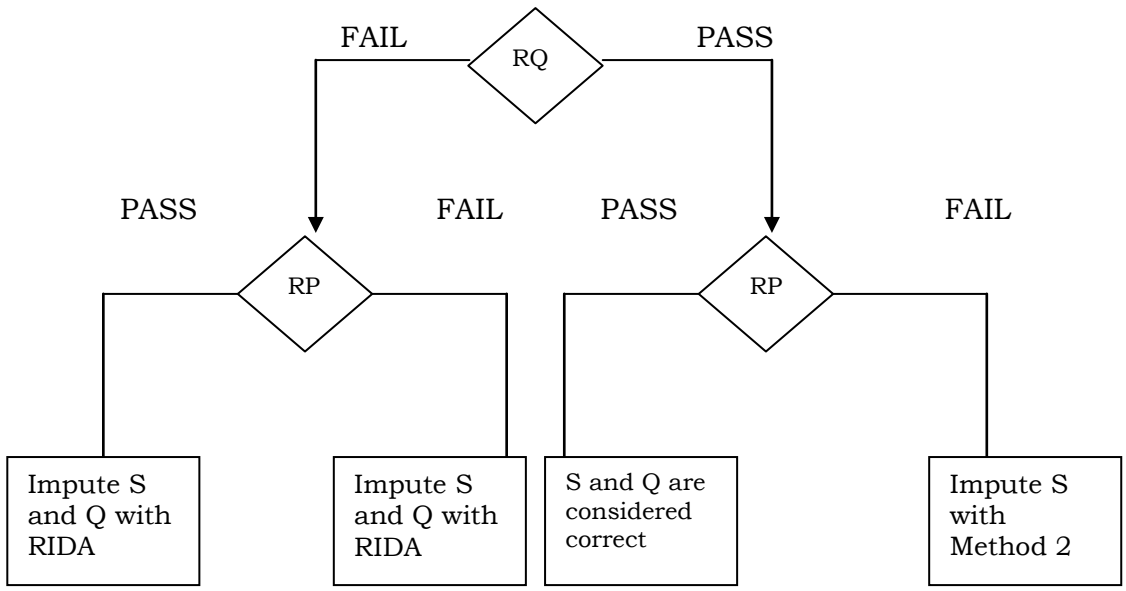


Figure 5. Flow-chart of the imputation procedure for numerical variables in the Self consumption diary

The imputations of such variables are made by Method 2, which is described by following procedures:

(a) Compute the 1st (q_1) and 3rd (q_3) quartile of the frequency distribution of S/Q;

[3] Agency for Statistics of Bosnia and Herzegovina, Institute of Statistics of the Federation of Bosnia and Herzegovina, Institute of Statistics of the Republika Srpska, District Brcko Branch Office (2005), " The BiH Household Budget Survey 2004 – Methodology", Sarajevo, 2005., p. 39.

(b) Generate a random draw (S*) from the uniform distribution on the interval ($Q \times q_1$, $Q \times q_3$);
(c) Impute the missing item with S*.

This imputation has the property that S*/Q is included in the interval (q_1, q_3).

A final important remark concerns the treatment of the variables income and the number of persons related to that income. These variables were edited following limits related to the per capita income (analogously to the previous steps). However, it is important to underline that, because of the nature of these variables and the main objectives of the HBS survey, this first editing check should be considered as a first editing step to be further deepened. Thus results on these variables should be carefully evaluated taking in mind that a further revision could be needed.

III. Software used for Data Editing and Imputation in Household Budget Survey

Statistical offices in Bosnia and Herzegovina have had a very close and productive co-operation with ISTAT-Italian National Statistical Office in the past years. One of the products of this work is the use of tailor made software which is created in ISTAT, as well as lessons learned in different statistical areas. Consequently, the CONCORD[4] (CONtrol and CORrection of Data) software was used in data editing and imputations phases in all household budget surveys in Bosnia and Herzegovina and it became a standard tool for these purposes.

The CONCORD is generalised software devoted to data editing and imputation developed in ISTAT. It consists of three modules[5]:

(a) SCIA module (Automatic Imputation and Checking System), which facilitates integral application of the Fellegi-Holt methodology for the location and imputation of errors, to a limited extent for categorical variables;
(b) GRANADA module (management of rules for data analysis), which enables the deterministic location of errors through application of IF-THEN type rules to be carried out on variables of both categorical and continuous type;
(c) RIDA module (Information Reconstruction with Automatic Donation) which enables imputation of both categorical and continuous variables through donor.

All program codes were prepared by using both, Italian experiences and expert knowledge of statisticians from Bosnia and Herzegovina and they were adjusted to all variables, survey instruments and flow of the questions.

IV. References

1. William E. Winkler and Bor-Chung Chen (2001), " Extending the Fellegi-Holt Model of Statistical Data Editing", Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

2. Agency for Statistics of Bosnia and Herzegovina, Institute of Statistics of the Federation of Bosnia and Herzegovina, Institute of Statistics of the Republika Srpska, District Brcko Branch Office (2005), " The BiH Household Budget Survey 2004 – Methodology", Sarajevo, 2005.

[4] G. Barcaroli, S. Bergamasco, M. Jouvenal, G. Pieraccini, L. Tininini, (2008) , " Generalised software for statistical cooperation", Contributi e Documenti Istat 2008, n. 16/2008

[5] Idem, p. 19.

3. G. Barcaroli, S. Bergamasco, M. Jouvenal, G. Pieraccini, L. Tininini (2008), " Generalised software for statistical cooperation", Contributi e Documenti Istat 2008, n. 16/2008