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

Edin Šabanović, Assistant Director
Sector for Statistical Methodology, Standards, Planning, Quality and Coordination
Agency for Statistics of Bosnia and Herzegovina

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

• Introduction
• Data Editing and Imputation in Living Standards Measurement Survey
• Data Editing and Imputation in Household Budget Survey
• Data editing and imputation of categorical variables
• Data editing and imputation of continuous variables
• Numerical variables in the Final interview
• Numerical variables in the Diary of purchase
• Variables in the Self-consumption diary
• Software used for Data Editing and Imputation in Household Budget Survey
• Conclusions and Future Steps
Introduction

• Non-sampling errors are present in every statistical survey, regardless of whether it is based on full coverage or sample
• Two main types of missing data: unit and item non-response
• Editing and imputation are very important phases in the generic statistical business production model (GSBPM)
• 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
Data Editing and Imputation in Living Standards Measurement Survey-LSMS (1)

• LSMS was the first survey conducted in the post-war period
• Very complex household survey which consisted of 12 modules
• 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
• Methods of descriptive data analysis were used
• Imputations 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
Data Editing and Imputation in Living Standards Measurement Survey (2)

- 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
- The median imputation rules were designed in a hierarchical (bottom-up) way: EA, Municipality, Entity, Country level
- Precondition: existence of at least five observations at the level used
- All imputation rules were programmed and performed in SPSS and the complete data bases were prepared for the final data analysis
Data Editing and Imputation in Household Budget Survey (1)

Three main survey instruments:

- Diary of Purchase (14 days recording period)
- Self-consumption Diary (14 days recording period)
- Final Interview (face-to-face interview)
- Social inclusion questionnaires: Social Inclusion and Migration Module and Health Module (face-to-face interview)
- Quantitative and qualitative checks were performed in order to detect and correct all errors in data
Data Editing and Imputation in Household Budget Survey (2)

- FBiH
- RS
- Brcko

Quantitative Check

BHAS

Qualitative Check
Quantitative checks

(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

Categorical Variables (Final Interview)

Numerical Variables (Final Interview Diary of Purchase, Self-consumption diary)
### Data editing and imputation of categorical variables (1)

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deterministic Corrections</strong></td>
<td>Skip rules ...</td>
</tr>
<tr>
<td><strong>Probabilistic Corrections (SCIA1)</strong></td>
<td>Relations within Household (Relation to head, Age, Sex...)</td>
</tr>
<tr>
<td><strong>Probabilistic Corrections (SCIA2)</strong></td>
<td>Relations within Individual (Current Activ. Status, Educ. Level...)</td>
</tr>
<tr>
<td><strong>Probabilistic Corrections (SCIA3)</strong></td>
<td>Characteristics within Household (Dwelling, durable goods,..)</td>
</tr>
</tbody>
</table>
Data editing and imputation of categorical variables (2)

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.
Data editing and imputation of categorical variables (3)

• 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:
  • SCIA1 for variables within HH (demographic module)
  • SCIA2 for variables within Individuals
  • SCIA3 for remaining variables within HH (Final Interview)
Data editing and imputation of continuous variables (1)

• 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

• Data editing was concerned to the identification and investigation of missing data and outliers

• 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
Data editing and imputation of continuous variables (2)

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:

• 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

• 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
Data editing and imputation of continuous variables (3)

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

- These variables were imputed by means of nearest neighbour hot-deck technique
- Homogeneous strata (imputation cells), where both the recipient and donor must belong, were created
- Set of variables (matching variables), on which the similarity through an opportune distance will be computed, was created
- Stratification and matching variables must strongly characterise the observations
- Data analysis and the knowledge acquired in the Italian Household Budget Survey were used
Data editing and imputation of continuous variables (4)

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:

\[
S = S_1 + S_2 + ... + S_{14} \\
Q = Q_1 + Q_2 + ... + Q_{14}
\]

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 $P = S/Q$

Denote 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 $P_L < S/Q < P_U$"
Data editing and imputation of continuous variables (5)

PASS

FAIL

PASS

FAIL

PASS

FAIL

Impute Q through method 1

Impute S and Q through donor

Impute S and Q through donor

S and Q are considered correct
Data editing and imputation of continuous variables (6)

Method 1:

1) Compute the 1\textsuperscript{st} ($q_1$) and 3\textsuperscript{rd} ($q_3$) quartile of the frequency distribution of $S/Q$
2) Generate a random draw ($Q^*$) from the uniform distribution on the interval $[S/q_3, S/q_1]$
3) Impute the missing item with $Q^*$

- This imputation has the property that $S/Q^*$ is included in the interval $(q_1, q_3)$
Data editing and imputation of continuous variables (7)

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:

\[
S_1 + S_2 + ... + S_{14} = S \\
Q_1 + Q_2 + ... + Q_{14} = Q
\]

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 $P = S/Q$

Denote 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 $P_L < S/Q < P_U$"
Data editing and imputation of continuous variables (8)

Impute S and Q with RIDA

Impute S and Q with RIDA

S and Q are considered correct

Impute S with Method 2
Data editing and imputation of continuous variables (9)

Method 2:

1) Compute the 1st ($q_1$) and 3rd ($q_3$) quartile of the frequency distribution of $S/Q$
2) Generate a random draw ($S^*$) from the uniform distribution on the interval ($Q \times q_1$, $Q \times q_3$)
3) Impute the missing item with $S^*$

- This imputation has the property that $S^*/Q$ is included in the interval ($q_1$, $q_3$)
Software used for Data Editing and Imputation in Household Budget Survey

• The CONCORD is generalised software devoted to data editing and imputation developed in ISTAT. It consists of three modules:
  
• 1) 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;

2) 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;

3) 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.
Conclusions and Future Steps

• The experiences in data editing and imputation of statistical data are still modest in Bosnia and Herzegovina
• There was a progress in the last decade, but there is a lot of room for improvement
• The procedures of data editing and imputations mostly developed for Household Budget Survey and Population Census and need to be applied also in other surveys
• CONCORD is generalised software which offers a lot of opportunities for the implementation of statistical data editing and imputation procedures
Thank you for the attention!

Edin Šabanović
edin.sabanovic@bhas.ba