TWINNING CONTRACT

Support to the State and Entity Statistical Institutions, phase V

Bosnia-Herzegovina

Final
MISSION REPORT

on

Extended household budget survey (EHBS)

Component no 2, Activity 2.3

Mission carried out by
Markku Lindqvist, Statistics Finland
Vesa Kuusela, Statistics Finland
Marco Di Zio, Istat
Teresa Buglielli, Istat

20 – 24 February 2012

Version: Final

		STATISTICS DENMARK

IPA 2008

Author's name, address, e-mail (keep the relevant information)

Markku Lindqvist
Statistics Finland
Verkstadsgatan 13
FI-00022 Helsinki
Finland
Tel:+358 9 1734 3418
Email: markku.lindqvist@stat.fi

Vesa Kuusela
Statistics Finland
Tel: +358
Email: vesamatti.kuusela@elisanet.fi

Marco Di Zio
Istat
Via Cesare Balbo 16
00186 Rome
Italy
Tel:+39 06 4673 2871
Email: dizio@istat.it

Teresa Buglielli
Istat
Tel:+39 06 4673 2809
Email: bugliell@istat.it

Table of contents

Executive Summary 3
1. General comments 3
2. Assessment and results 4
3. Conclusions and recommendations 5
4. What to do before the next mission for the BC Counterpart……6
5. Topics for the next meeting…… 6
Annex 1. Terms of Reference 5
Annex 2. Persons met .....**Fejl! Bogmærke er ikke defineret.**

List of Abbreviations

BiH	Bosnia and Herzegovina
BHAS	Agency for Statistics of Bosnia and Herzegovina
CONCORD	Controllo e Correzione Dati
DEI	Data Editing and Imputation
EHBS	Extended Household Budget Survey
ESS	European Statistical System
FIS	Institute for Statistics of Federation of Bosnia and Herzegovina
HBS	Household Budget Survey
RSIS	Institute for Statistics of Republika Srpska
Istat	Italian National Institute for Statistics
ToR	Terms of Reference

Executive Summary

If report-core text- exceeds 4 pages

Include information to Project Leaders and the RTA.
Main conclusions and highlights from findings.

1. General comments

This mission report was prepared within the Twinning Project „Support to the State and Entity Statistical Institutions, phase V". It was the second mission to be devoted to EHBS within Component 2 of the project.

The concrete objectives of the mission were:

- Overall revision of the DEI methods which were developed under the CONCORD program in 2007. Adjustment of the DEI procedures to the EHBS 2011
- Establishing and clarification of coherence and relationships between the DEI procedures, programs and input and output files
- Checking and improving of the DEI phases which were already done in 2011 by Bosnian team
- Finalizing of first part of the DEI procedures
- Scheduling of next steps in the DEI work, advisory work for preparing following procedures

- Presentation of alternative procedures for data cleaning, editing and imputation
- Preparation of the list of activities to be done before the next mission
- Preparation of the list of topics for the next mission

The consultants would like to express his/her thanks to all officials and individuals met for the kind support and valuable information which he/she received during the stay in Bosnia-Hercegovina, and which highly facilitated the work of the consultant.

This views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of EU, BHAS / FIS / RSIS or Statistics Finland.

2. Assessment and results

Introduction

The main objective of this mission was to assist and train the BHAS, FIS, RSIS staff to solve the existing problems concerning HBS 2011 data editing and imputation related to the CONCORD software that have been used in the previous HBS surveys (2004 and 2007). In addition demonstration was given how alternative software (SPSS; Blaise) can be used for editing (see ANNEX I).

Current situation in data editing and imputation

Data editing and imputation in two previous HBS surveys was performed using the tailor-made software for data editing and imputation (it consists of CONCORD and additional programs developed in C++) developed in ISTAT. This software was used for data editing and imputation of both categorical and continues variables. Software uses logical relations among variables (edit rules) for detecting errors in variables and for detecting missing variables as well.

For categorical variables SCIA program was used implementing both the Fellegi-Holt algorithm for localizing random errors in data and nearest neighbour hot-deck technique for imputation. The application of this step was divided into different hierarchical procedures. Continues variables checks mainly were based on the idea that variable values (or ratio variable values) must be located in certain pre-defined intervals. The boundaries for these intervals were built by local experts by observing data. Once these intervals have been defined, values that were not in the acceptance region were considered erroneous and thus replaced by imputed values. Nearest neighbor method was used for imputation in two previous surveys.

Rules for data editing, i.e. logical relations among variables and edit rules as well as boundaries for pre-defined intervals for variables were built by BHAS, FIS and RSIS staff. Such rules and boundaries for pre-defined intervals for EHBS 2011 have been prepared.

What concerns data editing and imputation in HBS 2004 and HBS 2007, it was performed with the help of the tailor-made software for data editing and imputation mentioned above and developed within a technical assistance projects. Staff in BiH statistical institutions has not enough knowledge in using this software.

HBS team in statistical institutes in BiH opinion is that above mentioned software is the only way to get data cleaned and edited for the EHBS 2011. Time schedule is too tight to implement alternative ways and alternative software for data editing.

During this mission, coherence and relationships between the DEI procedures have been established, and the organization of the programs and input and output files has been shown. The DEI methods developed under the CONCORD program in 2007 has been revised. In particular all the rules were verified and adpated to the HBS 2011, and some other logical rules were added. It has been assessed the work about the DEI phases done by the Bosnian team. It has been finalized almost all the DEI procedure, apart the analysis of otuliers, through the construction of acceptance ranges) because for this task all data are needed, and at this stage they are not available.

The activities that the Bosnian team shoul do before the next mission are the following

1. as soon as data are available and registered, the Bosnian team should analyse the distribution and in particular they have to look for outliers in order to understand if they are acceptable or not because of some mistakes. For this task it is important to have an interactive analysis and whenever possible to recontact the unit. In any case the general recommendation to follow is to leave data as much as they are. Once anomalous values non acceptable are found is needed to create acceptance region. The file with the acceptance range will be used in the last step of the procedure.
2. Concerning the two new modules on helath and social inclusion, do some statistics about missing items, and think about some rules involving the main variables. These are critical modules because they gather subjective information, and it is difficult to check the validity of such a kind of information.

3. Conclusions and recommendations

The meeting has been very fruitful especially because of the work done by the Bosnian team on the DEI before the meeting. Actually it was expected to finalize the first part of the DEI, but in fact almost the DEI procedure was completed. Only a part concerning the analysis of outliers has not completed beacause of the the lack of all data at this stage. It is expected that the Bosnian team will not have any particular problems concerning this task because this analysis was carried out in the last edition (HBS 2007) completely by the Bosnian team.
The recommendations are:

1. as soon as data are available and registered, the Bosnian team should analyse the distribution and in particular they have to look for outliers in order to understand if they are acceptable or not because of some mistakes. For this task it is important to have an interactive analysis and whenever possible to recontact the unit. In any case the general recommendation to follow is to leave data as much as they are. Once anomalous values non acceptable are found is needed to create acceptance regions. The file with the acceptance range will be used in the last step of the procedure.
2. Include the ranges in the DEI procedure that was finalized during this meeting and make the first run of the DEI procedure in order to see whter there are still technical problems.
3. Concerning the two new modules on helath and social inclusion, do some statistics about missing items, and think about some rules involving the main variables. These are critical modules because they gather subjective information, and it is difficult to check the validity of such a kind of information.

4. Topics for the next meeting

The next mission in this component will be mission 3 from the plan. This mission will take place in the period April 23- April 27, 2012. The main activities during this mission will be

1. *Going through the draft timetable for activities regarding the 2011 HBS survey.*

2. *Assess the analysis of outliers made by the Bosnian team*
3. *Finalize the DEI procedure for the HBS 2011.*
4. *Apply the DEI to the HBS 2011 data*
5. *Assess the results of the application of the DEI procedure to the HBS 2011 data*

Annex I. Vesa Kuusela presentation about alternative data editing procedure

Overview on statistical editing

Vesa Kuusela
Social Survey Unit
Statistics Finland

Outline of the presentation

- ***What for is Statistical Data Editing done***
- ***Data editing methods***
- ***Outlier detection***
- ***Automatic error localisation***
- ***Some imputation methods***

Goals of Statistical Data Editing (SDE)

- ***1. Localize errors and correct them***
- ***2. Identify error sources in order to provide feedback on entire survey process, e.g.***
  - **For questionnaire design**
  - **For interviewer training and instructions**
  - **For data entry**
- ***3. Provide information about the quality of incoming and outgoing data.***
  - **Quality reporting**
  - **Interviewer performance**

Providing feedback and information on quality

- ***Not only errors but also the*** *sources of errors* ***in data collection and data handling should be located and reported***
- ***The lessons learned in data editing of one survey can be used to reduce errors in the forthcoming surveys***
  - **Mostly, the errors that are found are caused by deficiencies of data collection, data handling and data entry**
- ***In addition, the edits should be classified and tabulated and documented as a part of the quality reporting.***
  - **In some international surveys it is required to produce a protocol about editing**

The importance of data editing

- ***Data editing is one of the most labour consuming tasks in the whole survey process. In some cases, 40% of total costs are caused by data editing and corrections!*** *(Biemer & Lyberg 2003)*
  - **Careful planning of data collection, including questionnaire design, is the best way to reduce the costs of data editing.**

- ***Cost can also be reduced if the needs of data editing and data entry are taken into account already in the (paper)questionnaires.***
- ***Statistical data editing should have a predefined plan***
  - **What will be checked and in what phase and by whom**
    - **All cases are not as significant**
    - **All variables are not as significant**
  - **Organisation of SDE procedures has a great impact on total costs and on the quality of data**
- ***N.B. All errors in data cannot be corrected!***

Edit types in SDE

- ***Deterministic edits***
  - **Deviation from the rule is an error with certainty**
- ***Stochastic edits***
  - **Deviation from the rule is probably an error**
    - **Focus on errors that have considerable impact on results**
- ***Critical edits***
  - **Errors that must be corrected, otherwise the observation is useless.**
    - **Errors in variables that are the basic background variables in reporting**
    - **Errors in technical variables that are used in case identification and file merging**
  - **The errors in central variables in reporting**
- ***Different error types require different rules and methods***

Micro editing

- ***Errors in data are localized either interactively, observation by observation (interactive editing) or in a batch run***
  - **In both cases, error localisation is done by rules and algorithms**
    - **If AGE < 15 then MARITALSTATUS = NeverMarried**
    - **EXPENDITURE < TOTAL_INCOME**
- ***Error localisation in micro editing is done with specialized sofware or statistical software (e.g. SAS, SPSS, Blaise)***
- ***In interactive editing, the found errors are evaluated and corrected immediately***
  - **An specialized user interface is needed for interactive editing. That can be done by specialized software (e.g. Blaise) or with statistical software**
- ***A batch run produces an error report. Errors are usually analysed separately and all corrections are entered at the same time***
- ***Error correction often requires checking from the original data source***
  - **An easy access to original data source is necessary**
- ***One data set may be checked several times with different criteria***

Selective editing

- ***Split the data set into two separate data sets:***
  - **1. Critical errors data set**
    - **Contains records (observations) that are most likely to have influential errors**
      - ***For example, the observations that have big sampling weights***
    - **These records are edited traditionally - manually.**
  - **2. Non-critical errors data set**
    - **Records that are unlikely to contain** influential **errors.**
    - **Records are either not edited at all, or edited automatically**

Macro editing

- ***The last part of editing process. It examines potential impact on survey estimates to identify suspicious data in individual records.***
- ***This is an important step in data editing, because macro editing can find such errors that remain unnoticed by previously mentioned editing***
- ***There are two macro editing methods:***
  - **Aggregation method - verify whether the figures to be published seem to be plausible.**

        - **Results are comparing with the same results in previous publications, with register data, or related data from other sources.**
    - **Distribution method - find distribution of the variables and then compare separate values with that distribution.**
        - **The records containing values that are uncommon to be found in distribution are considered as potentially erroneous.**
- ***It is also possible to use graphical techniques in macro editing method.***

Macro editing methods

- ***In macro editing the data set is checked as a whole***
    - **Macro and micro editing supplement each other**
- ***For macro editing, specialized software are available but statistical software (e.g. SPSS, SAS) are usually sufficient and good***
- ***Checking should focus only on critical and key variables, and only occasionally on (few) other variables***
- ***Distributions of variables are the best tools***
    - **One variable distributions**
        - **Range, minimum, maximum**
        - **Plausibility of means and medians**
        - **Plausibility of cell frequencies**
    - **Conditional one variable distributions and cross tabulation**
        - **Range, minimum, maximum**
        - **Logicality of cell frequencies**
        - **Plausibility of cell frequencies**
        - **Cross tabulation reveals errors in question order and errors on skipping**
- ***Currently, also some sophisticated methods are available***

Outlier detection

- ***Outlier is an observation or value that does not ”belong” to the data set***
    - **Outlier can be in single variable or multivariate**
- ***Often, the tool is the distribution***
    - **With categorical variables outlier detection is simple**
    - **With measurement variables, the observations in both ends of the distribution are checked**
        - **Visual methods are often very efficient (e.g. box plot)**
- ***Top-down-editing***
    - **The observations of significant variables are sorted in ascending (e.g.) order and the largest and/or smallest cases are checked**

Automatic error localisation

- ***Usually, automatic error localisation is multivariate and based on outlier detection techniques***
- ***To detect erroneous data values, some sort of statistical model for the variables is made (implicitly or explicitly).***
    - **Regression models are common.**
- ***To measure the ”outlyingness”, a metric needs to be defined (e.g. Mahalanobis distance)***
- ***Other possibilities suggested in the EUREDIT project: Transformed rank correlations, Epidemic algorithms, Robust tree modelling, and Forward Search Algorithms.***

What to do with found errors?

- ***Sometimes, the correct value can be deduced from other answers on the same record***
- ***Often, it is necessary to look at the original data source (questionnaire)***
    - **If the value in the questionnaire is correct (data entry error) then the correct value is entered**
    - **The error in the questionnaire also is incorrect. The reason for the error needs to be localized and the correct value is entered on questionnaire (the original marking should remain visible), and in data set**
    - **N.B. The value in the questionnaire may be correct!**
- ***If the error cannot be corrected, it should be marked as a missing value***
    - **Later, the missing values may be replaced by imputed values.**
        - **N.B. Imputations do not make the data set correct but more usable**

Simple imputation methods

- ***Imputation is a method to fill in missing data with plausible values to produce a complete data set.***
- **Deductive methods *impute a missing value by using logical relations between variables and derive a value for the missing item***
- **(Unconditional) mean imputation *imputes the overall mean of a numeric variable for each missing item within that variable.***
  - **A variation of this method is to impute a class mean, where the classes may be defined based on some explanatory variables**
- **Deterministic regression *or* conditional mean *imputation, involves the use of one or more auxiliary variables, of which the values are known for complete units and units with missing values in the variable of interest.***
- **Hot Deck *Imputation is a simple way is to impute for each missing item the response of a randomly selected case for the variable of interest. Alternatively,* imputation classes *can be constructed, selecting donor values at random within classes.***
- **Nearest-neighbour imputation*, or* distance function matching*, is a donor method where the donor is selected by minimising a some 'distance'***

Component leader, BHAS

_____

Component leader, FIS

_____

Expert, RSIS

_____

Expert, Statistics Finland

_____

Expert, Statistics Finland

_____

Expert, Istat

_____

Expert, Istat

_____

RTA