Working Paper.
ENGLISH ONLY

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Paris, France, 28-30 April 2014)

Topic (v): International collaboration and processing tools

# SAS Enterprise Guide project for editing and imputation

Prepared by Saara Oinonen
Statistics Finland, Helsinki, Finland saara.oinonen@stat.fi

## I. Introduction

1. Editing and imputation practices have been in specific inspection in Statistics Finland for past five years. The first project about the context was lead in 1.7.2009 - 31.12.2011, and it surveyed editing and imputation practices in statistics and produced a process model for editing and imputation (Ollila & Rouhuvirta 2011). The project did also broad research on the international development. At this time it was decided at Statistics Finland that selective editing should be an important part of the process of editing in statistics. The practice of selective editing will be implemented in all suitable statistics gradually in forthcoming years.

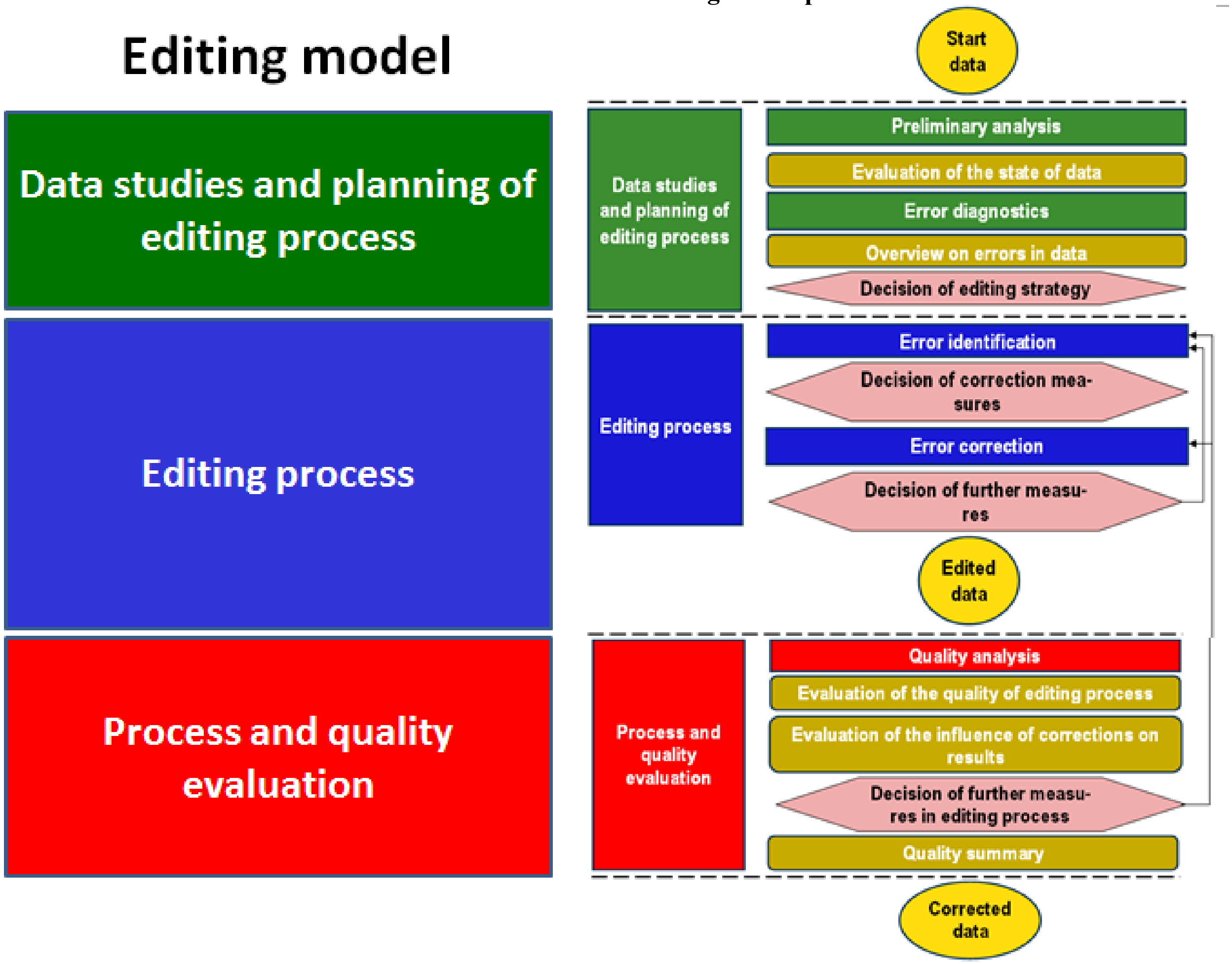
2. A new project was launched in 2012 with an intention of piloting the editing model and renewing editing methods in four statistics. For piloting purposes two software were chosen: SELEKT developed in Statistics Sweden (Nordberg et al. 2010) and BANFF developed in Statistics Canada (2007) which both operate in the SAS environment. During the process, the piloting statistics suggested some improvements for usability of editing and imputation tools. These two programs together with SAS macro language based modules provided a great foundation for developing a parameterized generic editing and imputation process operated in SAS Enterprise Guide. The key factor is that almost no programming is needed: the parameterisation guides processing of the data sets, conducting methods, calculating indicators and providing results. The data structure created supports this unified process.

## II. Editing and imputation practices compiled

### A. Process model turned into SAS EG project

3. Statistics Finland has a definition of policy of using SAS Enterprise Guide as main operating system, so it must be favoured when new practices and software are implemented. The editing team consisting of four persons of Statistical Methods unit (Ollila, Ahti-Miettinen, Oinonen, Pyy-Martikainen) formed a vision of an overall package of programs which included all functions presented in the process model for editing and imputation. The main functions of process model for editing and imputation are displayed in picture 1. SAS EG provided good foundation for implementing this vision since programs and codes can be easily parsed and they can be operated in different hierarchical levels. By producing smaller blocks of programs for different purposes and linking them in SAS EG the whole process becomes easier to digest and use. It is also possible to add notifications and instructions which can be read without opening a program code file. With co-operation of SAS team of Statistics Finland it would also be possible to create a user interface, but that would be finalizing development in the future.

**Picture 1. Process model for editing and imputation**


Editing model
Start data
Data studies and planning of editing process
Data studies and planning of editing process
Preliminary analysis
Evaluation of the state of data
Error diagnostics
Overview on errors in data
Decision of editing strategy
Editing process
Editing process
Error identification
Decision of correction measures
Error correction
Decision of further measures
Edited data
Process and quality evaluation
Process and quality evaluation
Quality analysis
Evaluation of the quality of editing process
Evaluation of the influence of corrections on results
Decision of further measures in editing process
Quality summary
Corrected data

4. The process model for editing and imputation was divided in three phases in SAS EG: Editing practices, imputation and testing. Currently the imputation phase is in development and not in use. Editing practices include selective editing methods offered by SELEKT. BANFF add-ins offer tools for error localisation, outlier detection and edit rule study and later for imputation practices. On top of these programs SAS EG project is supplemented with macros by Pauli Ollila and Max Niemelä (Statistics Finland) which carry out e.g. indicator calculations. All these phases and programs are run by parameters, which are given in defining programs at the beginning of the SAS EG project.

### B. Including SELEKT and BANFF to SAS EG project

5. The idea of modifying SELEKT and BANFF procedures into one generic editing program originated from piloting statistics. The main idea for revising editing practices is to make the editing phase of statistical production process more efficient and less time consuming. If introducing new editing methods requires that the statistics production staff needs to learn programming and new statistical methods concerning editing, it is unlikely that major benefit is gained. After parameterisation the statistics producers still need to know what methods are applied to their data, but there is no need to learn programming and in-depth structures of the programs. Also adding data sets from the new rounds to the program does not require any programming skills.

## III. Generic editing tool on SAS Enterprise guide

### A. Structure of the EG project

6. SAS Enterprise Guide project for editing does not yet include imputation practices, but several tests have been carried out. Currently, the project has been divided to 12 process flows which parse the editing practices and actions. Two of the process flows are for testing the settings of selective editing.

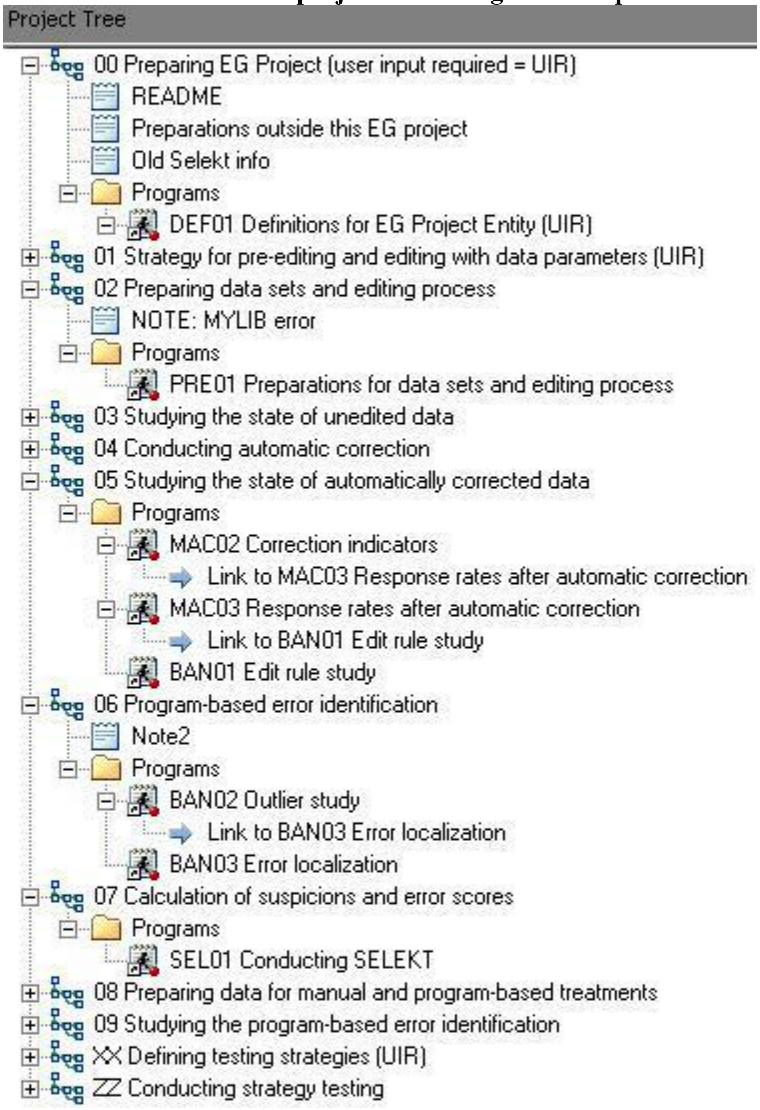
The structure of the project on SAS EG environment can be seen from picture 2. Some of the programs and information notes are hidden under the topic of process flow and in SAS EG they can be revealed by clicking the + icon in front of the name of the process flow. Full structure of the project is presented in the annex. The process flows that include programs which need parameters from the user are marked with letters 'UIR' (user input required).

7. One process flow specifies one type of action to be concluded. It may include one or more programs to run and notifications. Process flows are:

- 00 Preparing EG project (UIR)
- 01 Strategy for pre-editing and editing with data parameters (UIR)
- 02 Preparing data sets and editing process
- 03 Studying the state of unedited data
- 04 Conducting automatic correction
- 05 Studying the state of automatically corrected data
- 06 Program-based error identification
- 07 Calculation of suspicions and error scores
- 08 Preparing data for manual and program-based treatments
- 09 Studying the program-based error identification
- XX Defining testing strategies (UIR)
- ZZ Conducting strategy testing

Process flows for testing at the end of the project are not yet numbered since it is not certain how many additional process flows are inserted before them (e.g. imputation).

**Picture 2. SAS EG project for editing screen capture**



Project Tree

- ⊟ 00 Preparing EG Project (user input required = UIR)
  - README
  - Preparations outside this EG project
  - Old Selekt info
  - ⊟ Programs
    - ⊟ DEF01 Definitions for EG Project Entity (UIR)
- ⊞ 01 Strategy for pre-editing and editing with data parameters (UIR)
- ⊟ 02 Preparing data sets and editing process
  - NOTE: MYLIB error
  - ⊟ Programs
    - PRE01 Preparations for data sets and editing process
- ⊞ 03 Studying the state of unedited data
- ⊞ 04 Conducting automatic correction
- ⊟ 05 Studying the state of automatically corrected data
  - ⊟ Programs
    - ⊟ MAC02 Correction indicators
      - Link to MAC03 Response rates after automatic correction
    - ⊟ MAC03 Response rates after automatic correction
      - Link to BAN01 Edit rule study
    - BAN01 Edit rule study
- ⊟ 06 Program-based error identification
  - Note2
  - ⊟ Programs
    - ⊟ BAN02 Outlier study
      - Link to BAN03 Error localization
    - BAN03 Error localization
- ⊟ 07 Calculation of suspicions and error scores
  - ⊟ Programs
    - SEL01 Conducting SELEKT
- ⊞ 08 Preparing data for manual and program-based treatments
- ⊞ 09 Studying the program-based error identification
- ⊞ XX Defining testing strategies (UIR)
- ⊞ ZZ Conducting strategy testing

## B. Different types of macro programs

### B1. Defining and preparing macro programs

8. All functions in the SAS EG project for editing operate by parameters. Macro programs which require parameters has initials 'DEF' (definitions) and also letters 'UIR' at the end of the name. At the moment there are 10 defining macro programs and four of them are included in process flow XX (Defining testing strategies). All other defining macros are included on process flows 00 (Preparing EG project) and 01 (Strategy for pre-editing and editing with data parameters). Additional information is attached to all defining macro programs by notifications so user will have instructions at hand when parameter definitions are under consideration.

9. Preparing macro programs have initials 'PRE' and they modify data to the suitable form for executive macro programs. There are two preparing macro programs in SAS EG project, one in process flow 02 (Preparing data sets and editing process) and one in 08 (Preparing data for manual and program based treatments). These programs do not produce any results but it is advisable to check from log information that these programs have finished without errors. Preparing macro programs do not require any input from user.

### B2. Executive macro programs

10. Currently SAS EG project for editing has three types of executive macro programs with initials 'MAC', 'BAN' and 'SEL'. 'BAN' and 'SEL' refer to BANFF and SELEKT programs. 'MAC' macro programs execute indicator calculations. At the moment there are four 'MAC' programs, three 'BAN' programs and two 'SEL' programs in the SAS EG project, but especially the number of MAC and BAN programs might increase as the development proceeds. If there are no definitions for some executive macro programs, they do not carry out any operations. One can also omit selective editing with an option. By this matter the SAS EG project for editing is suitable also for those statistics for which some editing operations are not applicable.

11. 'MAC' programs can be found in process flows 03 (Studying the state of unedited data), 05 (Studying the state of automatically corrected data) and 09 (Studying the program-based error identification) and they carry out most of the indicator calculations. The process model for editing and imputation divides indicators in three groups: Indicators describing the state of the data, indicators related to error identification and indicators related to error correction actions (Ollila, Ahti-Miettinen & Oinonen, 2012). Indicator macros follow this same structure in SAS EG project for editing. Some of the indicator calculations are operated by 'BAN' macro programs because BANFF includes also that property for edit rules.

12. 'BAN' macro programs run operations of BANFF system. Currently the SAS EG project for editing includes BANFF procedures Proc Editstats, Proc Outlier and Proc Errorloc. BANFF procedures for imputation are intended to be added to the project. At the moment there are three 'BAN' macro programs in the project and they are found on process flows 05 (Studying the state of automatically corrected data) and 06 (Program-based error identification).

13. 'SEL' macro programs execute SELEKT operations. There are two process flows including these macro programs in SAS EG project for editing and one of them are in testing section (ZZ Conducting strategy testing). On the actual editing program SELEKT operations can be found on process flow 07: Calculation of suspicions and error scores. All the parameters for SELEKT are given in definition macro program DEF06: Parameters of SELEKT.

# IV. Operating SAS Enterprise Guide project for editing

## A. Data preparations

14. SAS EG project for editing requires multiple data sets to function and those data sets should be formed correctly. Minimum requirements for running EG project are current unedited data and previous edited data. Other possible data sets are current and previous sample and frame, current edited data and previous unedited data. Sample and frame information can also be given by parameters. Sample and frame information can be used for example in SELEKT for calculating weights and estimates in order to find the influential observations. The more data sets there are available the more possibilities SAS EG project for editing offers. There are information notes right in the beginning of the project about preparing data sets e.g. where they should be located and in what form these data sets should be.

15. The system does not alter original data sets, but creates new data sets in a generic form for the use throughout the process, directed by the defined parameters. In addition, there are some data sets prepared for further use in later phases, e.g. edit rules in BANFF and SELEKT form and status information after operations. The realisation of a new round of statistics is based on constant-form adapter programs, which need only data name and time point update at a minimum. If structural changes have appeared (e.g. variable content, new classifications) the adapter programs need additional information.

## B. Defining parameters

(a) Data and process parameters;
These parameters include for example information about folder directory, time parameters and operating system. Operating system information is important since SAS EG project for editing uses program CLAN developed in Statistics Sweden (1998) which functions with 32 bit data sets. In newer operating systems data sets are 64 bit and program ETOS (2012, Andersson, Statistics Sweden) must be used instead of CLAN. Data parameters can also include information about sampling and stratification, weights, frame, classifications and parameter estimation.

(b) Edit rules;
Edit rules are given either via specified macros or in written form. Edit rules are used by BANFF (error localization) and SELEKT (fatal and query edits). Fatal edits direct observations to manual inspection and query edits, with defined suspicion, for score calculations. Before performing edit rule study it is also possible to carry out automatic corrections which are also based on defined rules. Automatic corrections apply only for those situations where there is certainty of an error and definite correction for it. Edit rules can be targeted to specific subgroups of observations. In addition, one can exclude non-erroneous variables appearing in edit rules (e.g. register variables) from the error identification process as variables to be edited. Calculating response rates is one important function of SAS EG project and it is also possible to take structural missingness into account in the item response rates by using restriction conditions.

(c) Parameters for outlier study;
These parameters are aimed for BANFF and its procedure PROC OUTLIER. It has three methods available: Current data outlier detection, ratio outlier detection and historical outlier detection. One or more can be used depending on situation and available datasets. It is also possible to define different scenarios with different parameters.

(d) Parameters for SELEKT;
In addition to the edit rules, SELEKT allows the use of so called test variables for defining functions in order to score variables and observations (e.g. Hidiroglou-Berthelot). SELEKT parameters here refer to variable specifications, thresholds, time series and cross sectional parameters and parameters for test variables and for CLAN/ETOS estimation. All these functions are optional and can be applied to the situation.

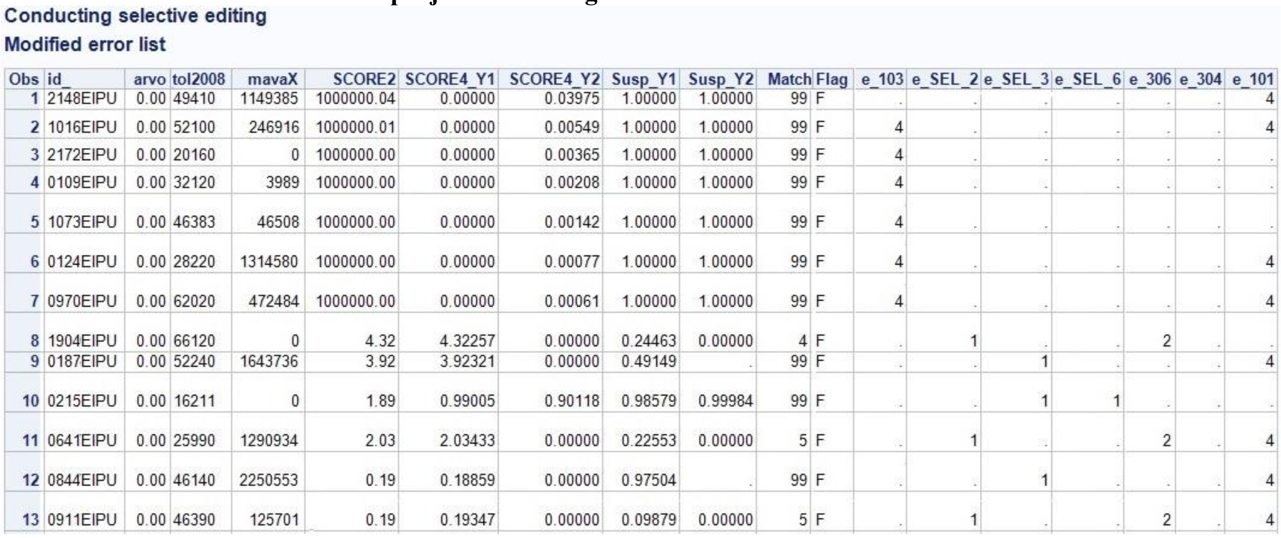
(e) Parameters for testing;
The idea for testing different editing schemas is adopted from SELEKT section LAB. On the testing section it is possible to modify edit rules, create new rules or exclude edit rules and find out if the editing process performs differently. Different collections of edit rules and testing strategies are called schemas. Schema structures are parsed to match the structure of SAS EG project for editing. Parameters are given separately in three categories: test schemas, edit rule alternatives and methodological parameters for testing.

## C. Results

(a) Error identification, status information;
Error information originates from four different sources: fatal edits and query edits via error localization, outlier study and finally by identification based on variable error scores from SELEKT. Mainly this means status information: erroneous or missing observations of variables and outliers. It also contains information of how this status was created by attaching the information of matching edit rule to the status. The principle of information about which data fields should be corrected (FTI, Field to be imputed) is adopted from BANFF.

(b) Scores from SELEKT;
Information coming from SELEKT is mainly error scores and suspicions. In this application it is also possible to define cut values for suspicion and the variable is marked erroneous when cut value is exceeded. Combining status information with scores from SELEKT there are various possibilities to structure lists of faulty variables. Lists can be modified to suit the situation and wishes of statistics. In the piloting phase it appeared that the needs of statistics vary a lot in that sense, and flexibility is needed.
Example of modified error list which compiles error information from several sources is presented in picture 3 (example only, not equivalent for real error list). Scores, suspicion values and edit rules including initials *e_SEL* are from SELEKT. First column contains id and three following includes data variables which are described to be helpful for the staff of statistics. Flag column indicates if observation has been identified erroneous. Each column with initial *e_* represents one edit rule. *Match* column refers to SELEKTs match method, which indicates if the error was found on historical or cross sectional inspection. The development of creating structurally demonstrative and usable error lists without being too large is still going on, e.g. the edit rule and SELEKT test variable code system should be opened to the reader in the most informative way.

**Picture 3. Error list from SAS EG project for editing**



Conducting selective editing

Modified error list

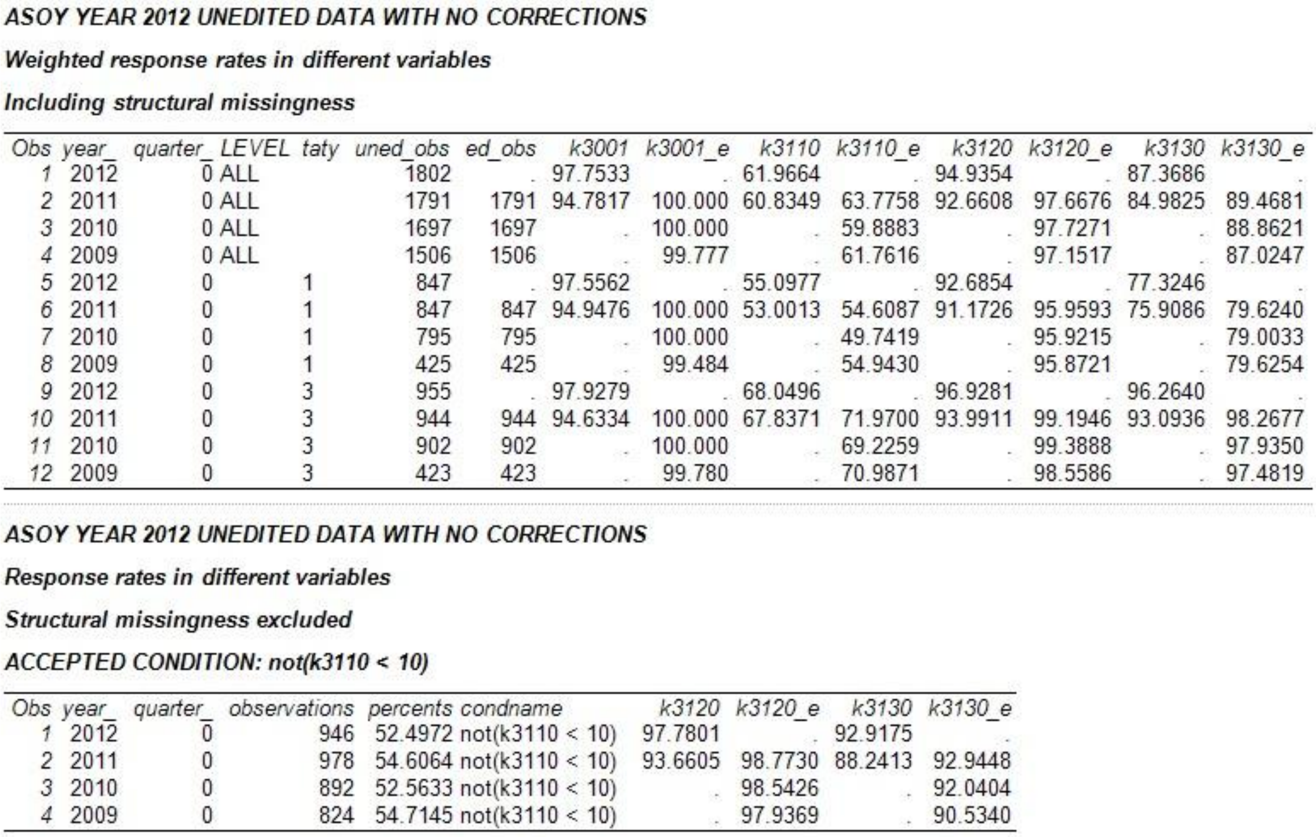
| Obs | id_ | arvo | tol2008 | mavaX | SCORE2 | SCORE4_Y1 | SCORE4_Y2 | Susp_Y1 | Susp_Y2 | Match | Flag | e_103 | e_SEL_2 | e_SEL_3 | e_SEL_6 | e_306 | e_304 | e_101 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2148EIPU | 0.00 | 49410 | 1149385 | 1000000.04 | 0.00000 | 0.03975 | 1.00000 | 1.00000 | 99 | F | . | . | . | . | . | . | 4 |
| 2 | 1016EIPU | 0.00 | 52100 | 246916 | 1000000.01 | 0.00000 | 0.00549 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | 4 |
| 3 | 2172EIPU | 0.00 | 20160 | 0 | 1000000.00 | 0.00000 | 0.00365 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | . |
| 4 | 0109EIPU | 0.00 | 32120 | 3989 | 1000000.00 | 0.00000 | 0.00208 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | . |
| 5 | 1073EIPU | 0.00 | 46383 | 46508 | 1000000.00 | 0.00000 | 0.00142 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | . |
| 6 | 0124EIPU | 0.00 | 28220 | 1314580 | 1000000.00 | 0.00000 | 0.00077 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | 4 |
| 7 | 0970EIPU | 0.00 | 62020 | 472484 | 1000000.00 | 0.00000 | 0.00061 | 1.00000 | 1.00000 | 99 | F | 4 | . | . | . | . | . | 4 |
| 8 | 1904EIPU | 0.00 | 66120 | 0 | 4.32 | 4.32257 | 0.00000 | 0.24463 | 0.00000 | 4 | F | . | 1 | . | . | 2 | . | . |
| 9 | 0187EIPU | 0.00 | 52240 | 1643736 | 3.92 | 3.92321 | 0.00000 | 0.49149 | . | 99 | F | . | . | 1 | . | . | . | 4 |
| 10 | 0215EIPU | 0.00 | 16211 | 0 | 1.89 | 0.99005 | 0.90118 | 0.98579 | 0.99984 | 99 | F | . | . | 1 | 1 | . | . | . |
| 11 | 0641EIPU | 0.00 | 25990 | 1290934 | 2.03 | 2.03433 | 0.00000 | 0.22553 | 0.00000 | 5 | F | . | 1 | . | . | 2 | . | 4 |
| 12 | 0844EIPU | 0.00 | 46140 | 2250553 | 0.19 | 0.18859 | 0.00000 | 0.97504 | . | 99 | F | . | . | 1 | . | . | . | 4 |
| 13 | 0911EIPU | 0.00 | 46390 | 125701 | 0.19 | 0.19347 | 0.00000 | 0.09879 | 0.00000 | 5 | F | . | 1 | . | . | 2 | . | 4 |

(c) Indicators;
Indicators of editing and imputation can be divided in three groups by their purpose: Indicators describing the state of the data, indicators related to error identification and indicators related to

error correction actions. All these are calculated automatically on SAS EG project for editing. Some indicators require parameters from user e.g. from which variables are response rates calculated. Indicators are essential for evaluating the editing process. Indicator information are presented in a form of a table and, if there are no major changes on data during time, the comparison of indicator tables between statistics cycles is very easy.

Example of indicator tables is presented as response rates in picture 4. On fist table there are weighted response rates for four variables (k3001, k3110, k3120, k3130). Suffix *_e* at the end of variable name refers to the edited values of the variable. If data includes structural missingness it can be defined in parameters and program produces tables excluding structural missingness. This is done for two variables (k3120, k3130) on second table.

**Picture 4. Response rates of unedited data**



```
ASOY YEAR 2012 UNEDITED DATA WITH NO CORRECTIONS

Weighted response rates in different variables

Including structural missingness

Obs year_ quarter_ LEVEL taty uned_obs ed_obs   k3001  k3001_e   k3110  k3110_e   k3120  k3120_e   k3130  k3130_e
  1  2012        0 ALL           1802      .  97.7533        .  61.9664       .  94.9354       .  87.3686        .
  2  2011        0 ALL           1791   1791  94.7817  100.000  60.8349 63.7758  92.6608 97.6676  84.9825  89.4681
  3  2010        0 ALL           1697   1697        .  100.000        . 59.8883        . 97.7271        .  88.8621
  4  2009        0 ALL           1506   1506        .   99.777        . 61.7616        . 97.1517        .  87.0247
  5  2012        0          1     847      .  97.5562        .  55.0977       .  92.6854       .  77.3246        .
  6  2011        0          1     847    847  94.9476  100.000  53.0013 54.6087  91.1726 95.9593  75.9086  79.6240
  7  2010        0          1     795    795        .  100.000        . 49.7419        . 95.9215        .  79.0033
  8  2009        0          1     425    425        .   99.484        . 54.9430        . 95.8721        .  79.6254
  9  2012        0          3     955      .  97.9279        .  68.0496       .  96.9281       .  96.2640        .
 10  2011        0          3     944    944  94.6334  100.000  67.8371 71.9700  93.9911 99.1946  93.0936  98.2677
 11  2010        0          3     902    902        .  100.000        . 69.2259        . 99.3888        .  97.9350
 12  2009        0          3     423    423        .   99.780        . 70.9871        . 98.5586        .  97.4819
```

```
ASOY YEAR 2012 UNEDITED DATA WITH NO CORRECTIONS

Response rates in different variables

Structural missingness excluded

ACCEPTED CONDITION: not(k3110 < 10)

Obs year_ quarter_ observations percents condname            k3120  k3120_e   k3130  k3130_e
  1  2012        0          946  52.4972 not(k3110 < 10)   97.7801        .  92.9175        .
  2  2011        0          978  54.6064 not(k3110 < 10)   93.6605  98.7730  88.2413  92.9448
  3  2010        0          892  52.5633 not(k3110 < 10)         .  98.5426        .  92.0404
  4  2009        0          824  54.7145 not(k3110 < 10)         .  97.9369        .  90.5340
```

## V. Outline

16. Currently four piloting statistics are testing the SAS EG project for editing with their own data and editing strategies. These statistics were chosen from different subjects and data types so that testing would be as thorough as possible. Statistics are International trade in services, Finance of housing companies, Quarterly statistics on the finances of municipalities and Register-based statistics on buildings and dwellings. First three are sample surveys and fourth one uses register data. Two of the statistics have own software for receiving and editing data and estimating statistics so SAS EG project for editing must be incorporated to existing systems.

17. Building an efficient program for editing is not enough for making implementation of new editing and imputation methods easy and simple. Comprehensive documentation of the SAS EG project is under production and producing teaching material for statistics is brewing. With the help of undergraduate trainee there is *a concept library* of the terms concerning editing and imputation for whole staff at Statistics Finland to use. Terms and concepts can be found from internal web-pages and special program called *Concept database* managed by Metadata services. Currently concepts are only in Finnish. Similar development is planned for statistical methods concerning editing and imputation. The goal is

that methods can be found from wiki-style pages and linking between methods, terms, manuals and teaching material is possible.

18. SAS EG project for editing is in unfinished state at the moment and there is no demo version available yet. Its reliability is not high enough for public distribution and thorough documentation is not finished. Project is lacking imputation practices but development is under way. One objective is to have functional and fully documented program prior next project which is planned to initiate on September 2014. This is the project that should implement selective editing methods for all suitable statistics in Statistics Finland.

## References

Andersson, C. & Nordberg, L. (1998): A User's Guide to CLAN 97, Statistics Sweden

Anderssin, C. (2012): ETOS 2.0 User's guide, Statistics Sweden

Banff Support Team (2007): Functional Description of the Banff System for Edit and Imputation, Statistics Canada.

Norberg, A., Adolfsson, C., Arvidson, G., Brundell, P,. Elffors, C., Gidlund, P., Kraftling, A., Lindgren, K., Nordberg, L., & Tongur, C. (2011): User's Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing, Statistics Sweden.

Norberg A., Adolfsson, C., Arvidson, G., Gidlund, P. & Nordberg, L. (2010): A General Methodology for Selective Data Editing, version 1.0, Statistics Sweden.

Norberg A., Arvidson, G., Kraftling, A. & Nordberg, L. (2010): SELEKT - A Generic SAS™ System for Selective Data Editing, Statistics Sweden.

Ollila, P. & Rouhuvirta, H. (2011): Process Model for Editing (draft), Internal methodology paper (in Finnish), Statistics Finland.

Ollila, P., Ahti-Miettinen, O. & Oinonen, S. (2012): Outlining a Process Model for Editing With Quality Indicators, Work Session on Statistical Data Editing, Oslo, Norway, September 2012.

## **Annex:** A full structure of SAS Enterprise Guide project for editing

Version 25.2.2014

Process flows are numbered, uderlined and printed in blue colour. Macro programs have initials DEF, PRE, BAN, SEL or MAC and have sign [P] at the end of the name. Regular text titles without suffixes refer to information notification. Process flows and programs which require parameters from user are marked with letters UIR.

### 00 Preparing EG project **(UIR)**

- README
- Preparations outside this EG project
- Old SELEKT info
- **DEF01** Definitions for EG Project Entity **(UIR)** [P]
  - DEF01: Descriptions and instructions

### 01 Strategy for pre-editing and editing with data parameters **(UIR)**

- Purpose of these strategy definitions
- Information of statistics
- Description of the strategy definitions
- Edit rule functions (DEF02, DEF03)
- Automatic correction functions (DEF02)
- **DEF02** Automatic correction definitions for pre-editing **(UIR)** [P]
  - DEF02: Descriptions and instructions
- **DEF03** Edit rules and constraints **(UIR)** [P]
  - DEF03: Descriptions and instructions
  - DEF03: Edit rule functions
- **DEF04** Parameters of outlier study **(UIR)** [P]
  - DEF04: Descriptions and instructions
- **DEF05** Parameters of data preparations **(UIR)** [P]
  - DEF05: Descriptions and instructions
- **DEF06** Parameters of SELEKT **(UIR)** [P]
  - DEF06: Descriptions and instructions

### 02 Preparing data sets and editing process

- NOTE: MYLIB error
- **PRE01** Preparations for data sets and editing process [P]

### 03 Studying the state of unedited data

- **MAC01** Response rates [P]

### 04 Conducting automatic correction

- **PRE02** Automatic correction [P]

### 05 Studying the state of automatically corrected data

- **MAC02** Corrections indicators [P]
- **MAC03** Response rates after automatic correction [P]
- **BAN01** Edit rule study [P]

### 06 Program-based error identification

- **BAN02** Outlier study [P]
  - BAN02: Note
- **BAN03** Error localization [P]
  - BAN03: Note

### 07 Calculation of suspicions and error scores

- **SEL01** Conducting SELEKT [P]

### 08 Preparing data for manual and program-based treatments

- **PRE03** Error and score preparations of data [P]

### 09 Studying the program-based error identification

- **MAC04** Indicators of error identification [P]

### XX Defining testing strategies (UIR)

- Original LAB1 code
- Original schema structures
- **DEF__** SELEKT test schemas [P]
- **DEF__** Edit rule alternatives [P]
- **DEF__** Parameters of schemas [P]
  - Defining a cost function
  - Available schemas
- **DEF__** Testing strategies [P]

### ZZ Conducting strategy testing

- **SEL02** Strategy Evaluation [P]
  - Data Set Limit Reached