WP.26
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iv): Micro editing – methods and software

THE EDIT

Invited Paper

Prepared by Anders Norberg, Statistics Sweden

I. Introduction

1. Statistical data editing is a resource-demanding process in business surveys. The editing of data in statistical surveys takes place at several stages of the production process and communicates with several of the sub-processes. Most resources are spent on the traditional editing of micro data. The use of web-questionnaires makes it possible to include some form of editing for respondents at the point of data capture. In fact, many respondents today expect to meet "intelligent" communication via the web. So far, most such systems lack techniques to store process data (paradata) from the response process.

2. Output (macro) editing is another sub-process that has the potential to be improved and to be more important in the survey process. Output editing can also detect errors introduced in the production and compilation processes. When resources can be released from the large micro editing process, some of these resources should be invested into respondent´s editing in web-questionnaires and output editing.

3. Statistics Sweden has developed a generic tool for significance editing, the SELEKT system. It is a set of SAS -programs and SAS-macros developed to support methods for significance editing. The SELEKT system provides the user with options for computing expected values and normal dispersion on cold deck data from previous survey rounds. These estimates are used in SELEKT-type edits as well as for the computation of impacts on statistics. The present system does not support the user with a refined interface for registering edit rules, it is just a blank program editor sheet. Here is a potential for improvement of the system.

4. Statistics Sweden has the ambition to describe edits systematically as information objects. The purpose is to:

- Yield a common understanding of the terminology within "editing"
- Store meta data on edits used in micro data editing in various environments
- Construct edit expression scripts by generic tools in various environments, for example SAS, when the metadata already is at hand.

II. Standard terminology

5. The "Glossary of Terms on Statistical Data Editing" was prepared by the participants of the UN/ECE Work Sessions on Statistical Data Editing. Here there is no single term 'Edit', but

> "EDIT RULE SPECIFICATION
> CHECK RULE SPECIFICATION
> A set of check rules that should be applied in the given editing task."

6. In "Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys" (August 2007) the "Edit" is defined as "A logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct. Also known as edit rule or checking rule". In the following we use "Edit" synonymous to Edit rule and Check rule.

III. Editing processes

A. Edits in web-questionnaires

7. SIV is a generic tool developed at Statistics Sweden for designing and building web-questionnaires and for presenting these via Internet for respondents. The tool satisfies several purposes; increased coordination and standardisation of format and functionality in web- questionnaires and decreased dependence of IT-support. SIV is designated for surveys concerning enterprises, individuals, households, authorities and the public sector.

8. SIV has options for designing edits associated to each question/variable. These edits are directed to be used by the respondent and/or the editing staff in the micro data editing process at Statistics Sweden. The scripts of the edit rules are stored as part of the SIV-application.

B. The implementation of significance editing

9. The method and the IT tools for flagging of incorrect or suspected data values through traditional, selective and significance editing at Statistics Sweden is called SELEKT. Necessary parameters, several of these can be set to the default values, are stored in a table with the module PRE-SELEKT and need to be maintained on a regular basis. PRE-SELEKT also computes expected/predicted values and measures of dispersion on cold deck data to be used in the SELEKT-type edits. AUTO-SELEKT calculates scores according to the parameter table, indicating the anticipated impact on all important output.

10. Expected/predicted values and dispersion measures are computed for edit groups. These must be homogenous and may, but need not, correspond to strata or domains of study. In SELEKT, the edit groups can be formed by a set of auxiliary variables, the detail of classification (number of digits) and a fixed minimum number of observations required for the computation. Totals, functions of totals and their estimated standard errors are estimated "outside" the SELEKT. Several different types of software can undertake this estimation, Statistics Sweden uses its own software CLAN.

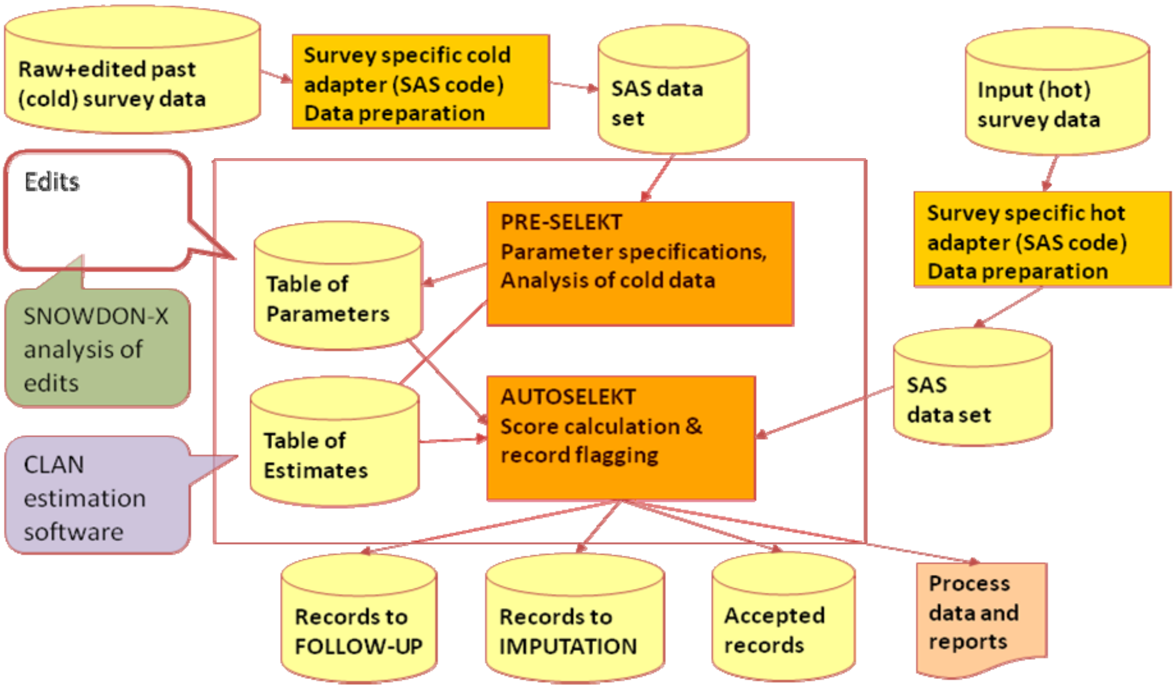
11. The system does not support the user with a refined interface for registering "traditional edits", it is just an blank program editor sheet. Here is a potential for improvement of the system. The optimal would be to use some kind of metadata for edits, stored in a system available for various IT-systems.

12. Items failing edits are attached an error flag. In the SELEKT system the flag takes the values 1, 2, 3 or 4, thus bearing information how the flagged data is to be further treated. The flag 4 means that the record shall pass thru SELEKT outright to the error list. The flag 3 means that the record shall be tested for impact on output and be given a global score to decide if it will be on the error list or not. The flag 2 means that the variable is given a specified suspicion and that the local scores are computed for the variable using that suspicion. The flag 1 is assigned to the SELEKT-type edits that are optional.

13. EDIT is being developed to be the tool for the editing staff to follow-up flagged items. EDIT will have a standard interface, functionality that presents all of the information needed such as previous data and analysis thereof, register data look-up ability, etc. It can trig SELEKT to check a specific batch of data.

14. Process data are generated in an ongoing process. They can be used both for continuous monitoring and for analysis and evaluation in order to improve the production cycle and reach an optimal resource allocation.

Figure 1 Data flow and SELEKT software


Raw+edited past (cold) survey data
Survey specific cold adapter (SAS code) Data preparation
SAS data set
Input (hot) survey data
Edits
PRE-SELEKT Parameter specifications, Analysis of cold data
Survey specific hot adapter (SAS code) Data preparation
Table of Parameters
SNOWDON-X analysis of edits
SAS data set
AUTOSELEKT Score calculation & record flagging
Table of Estimates
CLAN estimation software
Records to FOLLOW-UP
Records to IMPUTATION
Accepted records
Process data and reports

15. The ambition is to use the Office of National Statistics´ software SNOWDON-X to improve the traditional edits. Prototype versions have been implemented and tested in a few surveys to date. Experience will bring us forward to efficient editing. A new project aiming to make all tools for several current processes in data capture and data processing communicate with each other is in progress to build the TRITON-system. SELEKT will then be a piece of TRITON.

IV. The elements of an edit

A. Concepts

16. In the annual survey Structures of salaries, it is important to test salaries against different acceptance regions for occupation groups at some level of detail, and perhaps combined with some other variable decisive for salaries.

Example 1:

if Occupation = ‘Doctor’ and not (29000 < Salary < 71000) then Errcode_A01 = ‘Flag’

We say that

“Persons with Occupation = ‘Doctor’” is an *Edit group*.

Salary is the *Test variable.*

{29000, 71000} is the *Acceptance region.*

The logic of an edit is that if the value of a survey unit´s test-variable is not within the acceptance region, specified for an edit group to which the unit belongs, the value is flagged to be a fatal or suspected error.

17. For the editing to be efficient the edit groups must be homogenous with respect to the variable in focus, here Salary. The formation of edit groups benefit from multivariate analysis of data. The edit itself does not tell if it is a fatal edit or a query edit. A query edit identifies suspicious data items that may be erroneous. A fatal edit flags data items that are known with certainty to be in error.

18. An edit could consist of all the acceptance regions for a set of the non-overlapping edit groups that have been defined for the edit. The test-variable is one and the same. We use the logical operator ‘OR’ to combine simple edits to one combined edit.

Example 2:

```
If   Occupation = ‘Doctor’ and not (29000 < Salary < 71000)
or  Occupation = ‘Nurse’ and not (23300 < Salary < 43800)
then Errcode_A02 = ‘Flag’
```

19. There are reasons for saying that one edit must not be composed of several edit groups and acceptance regions;

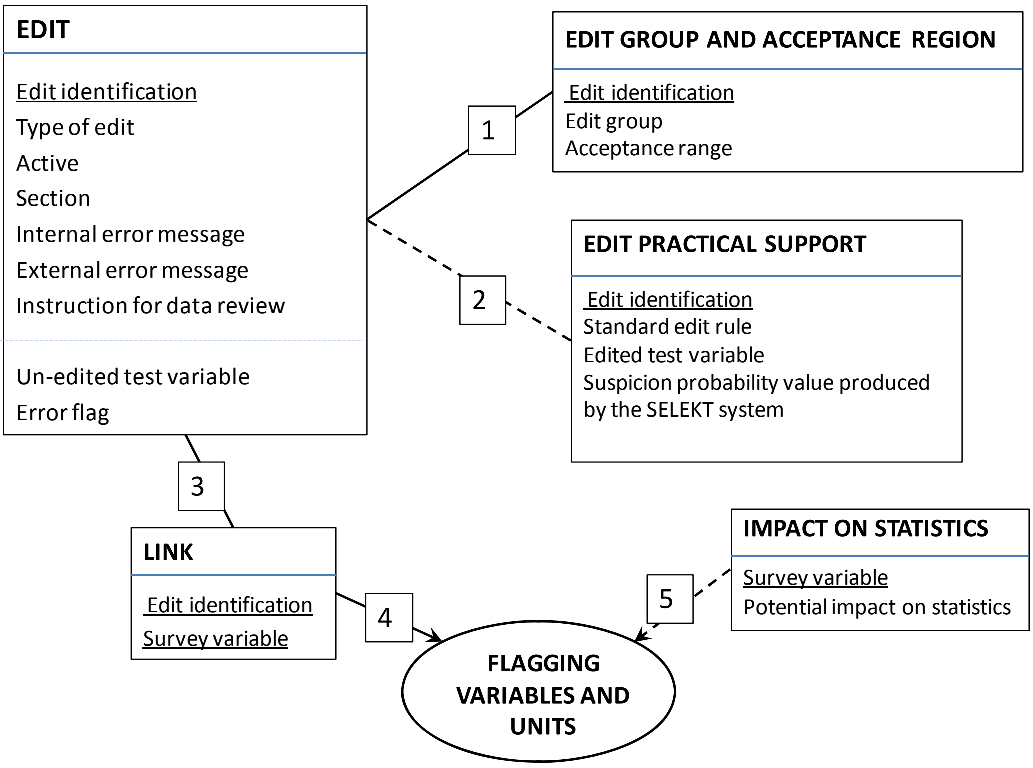
- Simplicity and understanding of edit conditions - it is much easier to understand and maintain simple edits than a rather complex series of conditions;
- Impact - it can be easier to monitor impact of the various separate edits by assigning each one a separate edit number (identification), and then being able to compute process indicators/statistics for each of them.

20. Our argument for allowing one edit to be composed of several edit groups, each with an acceptance region, is that we will have fewer edits and thus fewer error codes to administrate. Process indicators can be computed for the two-dimensional space of edit and edit group as well as for the entire edit. Taking this point of view is no obstacle for defining single-group edits in a survey.

B. Information object model

21. Most of “traditional” edits can be described in a common way. Figure 1 gives an overview of the connections within the edit object. In the tables below the contents of the objects is specified.

Figure 2. The EDIT object and related information objects.



Explanations of the relations in figure 1:

1 There can be one or many edit groups and acceptance regions in one edit.

2 The optional parameters *Standard edit rule*, *Edited test variable* and *Suspicion probability by SELEKT*, are supporting the edit object for practical use. One intention is to produce edit scripts with all necessary information in these tables.

The concept *Standard edit rule* can facilitate the construction of edit scripts by providing a wealth of standard edits by tested code stored in a library. Examples: Check for numeric value, check digit test for 12-digit PIN.

Edited test variable is the expression for a test variable, based on edited data. These edited test variables are used in evaluation of the editing process.

The SELEKT software allows you to flag suspected data values and attach a measure of suspicion, a probability (propensity) for error, that the user set for the edit.

3 There is a many-to-many relation between the edits and the survey variables. Many survey variables can be involved in the test variable expression and consequently be tested by one edit. On the other hand several edit tests can be performed on each survey variable. The table Link consists of all relations.

4 In the editing process, suspicions are computed for each survey variable, based on edits. The primary choice is the maximum of the resulted suspicions related to each survey variable. In case of dichotomised suspicions this is saying that if the survey variable fails at least one edit, it is flagged.

5 Selective editing is a second step in the editing procedure, ranking units by scores. To each survey variable is attached a computed estimate of the potential impact on statistics, assuming that the variable value is erroneous. The SELEKT software has the option of computing suspicion as a continuous probability-like measure and combines suspicion and potential impact to anticipated impact as the basis of local score.

Remember that SELEKT uses a set of informative flag-values to manage the continuation of the editing process. For example, we might wish to send all fatal errors to manual follow-up whereas variable values that are only suspected to be erroneous are further processed by selective editing.

C. Object tables

The EDIT table		
Identification and messages:		
Column name:	Variable name	Comments/Description:
EditNumber	**Edit identification**	Edit identification, can be a number or a name. This name is also the error code.
EditType	**Type of edit**	**F**(atal) or **H**(ard) or **E**(rror) = Fatal/Obvious errors is identified by a Fatal edit **Q(**uery) or **S**(oft) or **W**(arning) = Query/Stocastic/Suspected error is identified by a Query/Stochastic or Statistical edit
EditActive	**Active**	Actual, 0=Not in action, 1=In action
EditSection	**Section**	Grouping of edits to make possible various analysis of process data
EditInternalMessage	**Internal error message**	Error message for internal use at National Statistical Institute
EditExternalMessage	**External error message**	Error message for respondents using web questionnaires and as metadata to researchers using data files.
EditInstruction	**Instruction for data review**	Instruction for data review/follow-up
Mappings and expressions:		
Column name:	Variable name	Comments/Description:
EditTest	**Un-edited test variable**	Test variable expression to be used in production: a) For many of the standard edits the survey variable name is declared b) For more general edits a test-variable is declared, being an expression consisting of survey-variables, variables in cold deck data, register variables, constants, functions etc.
EditFlag	**Error flag**	The character set as information of any kind, for example the direction for the following processes.

The EDIT GROUP AND ACCEPTANCE REGION table		
Column name:	Variable name	Comments/Description:
EditNumber	**Edit identification**	Edit identification, can be a number or a name.
EditGroup	**Edit group**	Homogeneous edit group of units for which the acceptance region is applicable for the test variable. The field is mainly blank for standard fatal edits.
EditAcceptanceRegion	**Acceptance region**	Acceptance region. The field is used in different ways depending of EditStandard. Examples: Lower and upper limits for an interval (ex: 95-105), a discrete code list (ex: {F, M})

The EDIT PRACTICAL SUPPORT table		
Column name:	Variable name	Comments/Description:
EditNumber	**Edit identification**	Edit identification, can be a number or a name.
EditStandard	**Standard edit rule**	Declaration of Standard Edit, as Date, Interval, Change etc., or “No standard”.
EditTestEdited	**Edited test variable**	Test variable expression to be used in evaluation of the editing process. See Un-edited test variable
EditSuspicion	**Suspicion probability value produced by the SELEKT system**	*(Only for SELEKT)* Suspicion probability, if EditFlag=**2**. A number in the interval {0,1}

V. QUESTIONS FOR DISCUSSION

22. As a benchmark for further development of generic tools for editing we would gratefully take part of similar systems or attempts to build such systems.

- Can most edits be described as consisting of the components
  - test variable,
  - edit group,
  - acceptance region?
- What types of edits can not?
  - Residuals from regression analysis?
  - Outliers detected by seasonal adjustment procedures?
- If the edits can, what arguments are there for saying that
  - one edit has only one edit group and one acceptance region or
  - one edit can be composed of many edit groups with one acceptance region each?
- Examples of modeling edits?
- Examples of a metadata storage for edits?
- Examples of edit scrip generators using a standard metadata storage for edits?

References

UN/ECE Work Sessions on Statistical Data Editing (2000), “Glossary of Terms on Statistical Data Editing”

EDIMBUS (2007), “Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys”, August 2007

Statistics Sweden (2011) “User´s Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing”, 2011-02-17