2009 European Establishment Statistics Workshop - ESSW09 (Stockholm, 7-9 September 2009)
Session 6: Editing

Editing at Statistics Sweden – Yesterday, today and tomorrow

Prepared by Anders Norberg, Statistics Sweden 2009-07-10

Yesterday – Too much editing

Editing is a resource-demanding process for statistical products with organizations as information providers (respondents). In a study (2004) at Statistics Sweden of the total costs of 62 statistical products, one third of the resources were used for editing. This figure, although in accordance with experiences from other countries, was deemed too high by the management. The proportion of resources invested in editing is larger for annual and periodic surveys than for monthly and quarterly surveys.

Average proportions of costs of sub-processes 2004.

Process	Proportion of total cost (%)		
	All products	Short-period	Annual surveys and periodic
Respondent service	3.3	3.3	3.4
Manual pre-editing	4.4	3.9	5.1
Data-registration editing	5.6	5.1	6.5
Production editing	15.3	12.7	18.9
Output editing	3.9	3.4	4.8
Total editing cost	32.6	28.3	38.6

The 2004 study showed that there were no homogeneous editing methods in use, despite the handbook CBM 2002:1 "Guide till granskning". The process was decentralised to each and every survey.

Today – Development of general tools

Today data collection and editing of data from organisations and enterprises is centralised to one department. After the move, there is a potential for efficient spread of workload. But, there is a heavy demand for common tools.

In order to reduce the amount of editing and the associated costs a series of projects were started in 2006. The main purpose was to analyze which modules for methods that should be used and to build the neses-sary general tools for editing. Benchmarking had given us the information that there were no system at other national statistical institutes that would yield the properties we wanted.

The project "Nine case studies"

As a first step a number of studies were conducted focusing on if and how selective editing with score functions could be used. Other purposes of this project were to learn about similarities and differencies between the surveys with regard to editing and to see if something could be done quickly to improve the individual survey under the present production system. Nine of the most edit intensive surveys at Statistics Sweden were included in the project. The surveys included in the project differ in many respects, which is of significance for how editing is performed.

Periodicity: 1) One-off surveys and surveys that are conducted so seldom that there is no information from earlier observations that could provide a basis for finding reasonable edits. Here, the role of editing is to find significant errors rather than to contribute to survey improvement for the future. 2) Annual

surveys and also intermittent surveys. 3) Monthly and quarterly surveys that in most cases have data from many previous rounds of surveys. It is important to notice that even in a monthly survey some objects are new when a new sample is drawn. These objects lack earlier data.

Survey design: Distinction is made between 1) sample and 2) population surveys. In the case of samples, weighting is always involved. This means that the observed objects have different impacts on the output which must be considered during editing. The sampling method, whether it is stratified SRS or sampling with unequal probabilities, is of little concern for editing. Strata, however, can be used as homogenous groups in the estimation of good expected values, which are used in selective editing. One- or multi-stage samples make a difference in complexity.

Types of objects: In principle, type of object – individuals, enterprises, products, etc. – have no significance in terms of editing. Nevertheless, it is a fact that business populations generally show a much more skewed distribution on economic and other quantity variables than individual data. Surveys involving individual data with attitudinal questions cannot, for practical reasons, be edited retrospectively by means of re-contact.

Expected values: In a specific survey there might be hard to find proper expected values for some or all measurement variables. The gained efficiency of selective editing is very much depending on the quality of the expected values. In some surveys data are gathered on several variables that are not reported individually in the output statistics, but rather as a derived variable. If there is no interest in the individual variables themselves it is recommended to calculate scores only for the derived variable. Calculations are based on edited data, not using raising factor, for homogenous groups. All objects are included in the groups for cross-sectional data no matter if they belong to an old outgoing sampling panel or the present one.

Output: A survey may have a few clearly defined users and limited output or extensive statistical reporting to a general (public). It is natural to focus the editing process on impacts within the principal reporting.

Empirical data: Data from previous survey rounds were needed to define edits with efficient threshold values. A precondition for being able to introduce and also adjust already established methods and parameters for effective editing is that unedited data are available from previous survey rounds. Data can be used both in cross-sectional and time-series analysis. In each survey a choice must be made whether or not to utilize imputed values and whether or not to utilize flagged but accepted data. It seems to be a good idea not to use artificial or highly suspected data, but it is easier not to make a difference. An alternative to consider is good data with generated errors, which gives you full control over the search for errors.

The project showed that it is possible to implement selective editing in at least seven of nine surveys, two failed because of lack of unedited data. Selective editing will lead to efficiency gains and likely cost reductions. The experiences from the case studies reveal that the introduction of new methods demands intensive testing in every specific survey where selective editing is supposed to be implemented. The reason for this is the variation between the surveys regarding data structure, use of the statistics etc. General tools for editing must therefore be very flexible to be able to deal with these different situations.

Besides implementation of selective editing the efficiency gains can be increased even more by dealing with the existing measurements issues. It is important that the questionnaires are adjusted to what the respondents are capable of delivering and it is equally important that the questions asked are well defined. If this is not fulfilled it will lead to more editing to compensate for low data quality. The results of the project show that several of the nine included surveys suffer from measurement issues concerning at least some variables. The case studies have not only delivered results according to the project plan, but also improved the competence of the participants of the project. This is very important for implementations and evaluations of the editing process ahead.

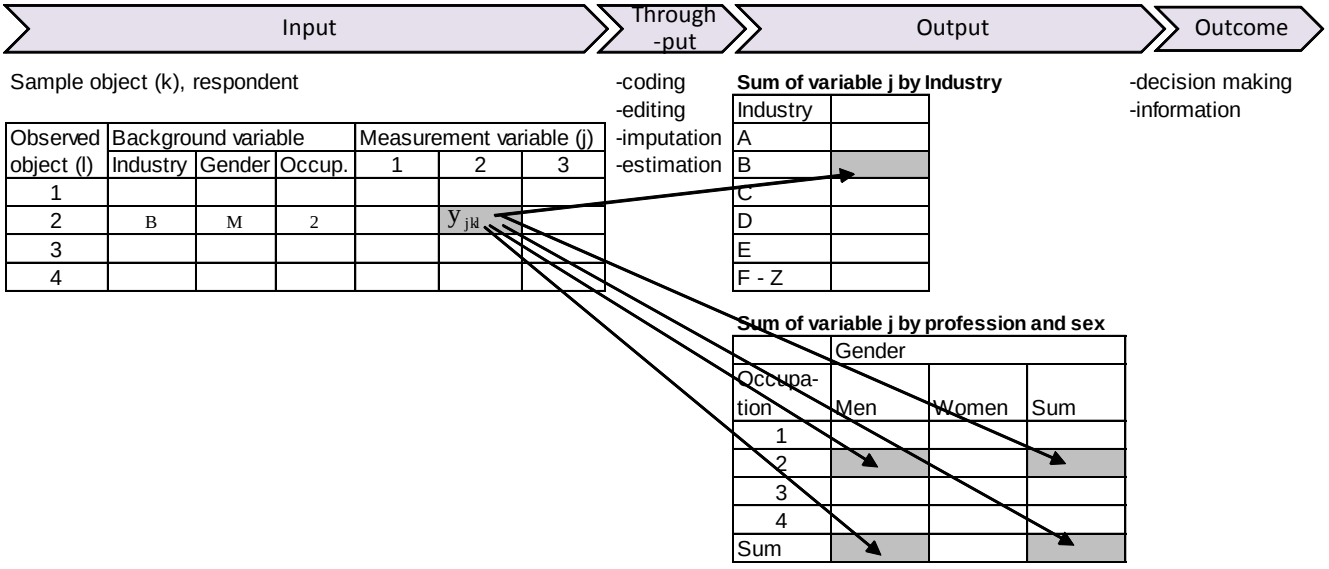
General tools

We think that common general tools would be the best way to get acceptance for the change of methods. These are the expectations for the editing process.

- Standardized, common, general/generic tools lead to:
  - Less maintenance of IT systems, reducing a large cost for the NSI.
  - Easier planning of the manpower-demanding editing work as members of the staff can be replaced when they are well acquainted with the tools.
  - Better work environment for editing staff.
  - Methodology studies are facilitated; studies of methods are possible when pre-requisites are comparable.
- Efficient editing methods (selective editing, significance editing) lead to:
  - Smaller volumes of follow-up, cheaper for Statistics Sweden and a smaller burden for respondents.
  - Better work environment for editing staff, not so many annoyed respondents to meet.
- Structured collection and analysis of process data lead to:
  - Systematic improvement of data collection.
  - More efficient application of the editing methods and tools.
  - Better quality in statistics.
  - Information for quality declaration of statistics.

Selective editing can be used as a complement to "regular" editing to reduce the list of flagged data that are identified by suspicion only. Suspicion and potential impacts can also be treated simultaneously in an integrated procedure for significance. We propose simple continuous propensity measures for suspicion.

One erroneous input data value can have impacts on several output statistical values. This is so when output is spread by more than one classification variable, for example when wages are computed and presented by industrial sector, gender and occupation. Here it is necessary that the producer of statistics can assess the importance of each output tables and estimate the consequences of lack of quality on the statistics, from the user´s point of view (outcome). The manager of the survey should provide a description of what is most important, evaluated in relative numbers. This information will be used to adjust parameter settings etc. in order to obtain high quality where it is most needed.



Tomorrow – The SELEKT and EDIT systems

By the end of September 2009 the first version of a set of general tools is planned to be at place. A first prototype was implemented in production in October 2008. The tools are of three kinds.

A. Method - and IT tools for **flagging** of incorrect or suspect data values through selective/significance editing. We call the toolbox SELEKT. Observations and variables are flagged to go to manual follow-up or computerised imputation or acceptance, this is decided in SELEKT. Necessary parameters are set with the boundary intersection PRE-SELEKT for each current survey and are seen over regularly by a process- and system-expert. The parameter values are stored in a table. AUTO-SELEKT is a module that does calculations according to the settings made in PRE-SELEKT, by reading the parameter table. A so called laboratory, LAB, is a third production tool box in the SELEKT. The LAB will be used before implementation in surveys. The LAB-module is used in order to evaluate the earlier production rounds to find best values on parameters, i.e. threshold values etc. To a large extent the code for AUTO-SELEKT is used, but the LAB requires some extra functionality. The LAB is used also later in order to now and then adjust parameters after more current data.

B. EDIT is the tool for the editing staff to use for follow up of error flagged items. It is here very important with a standard interface, correct functionality that present all information needed and a layout that gives good working conditions. It must be possible to ask SELEKT to check whatever batch of data from EDIT, for example those just adjusted, and this must go quickly.

C. A lot process data are generated in the editing process. A cohesive investigation concerning process dates and analysis of these is required for the editing process.

To summarize, the characteristics of SELEKT are:

- Selective and significance editing
- Potential impact is estimated for all important output
- Modelling of suspicion and potential impact separately
- Continuous suspicion measures can be computed by an integrated analysis pack
- Standard methods for analysis of cold-deck data are available
- Option for constructing homogeneous groups by hierarchical use of explanatory variables
- A SAS-program, PRE-SELEKT, is the user interface for delivery of parameter settings. This is a rough environment for a non-methodologist, but it is familiar for the process- and system-expert.

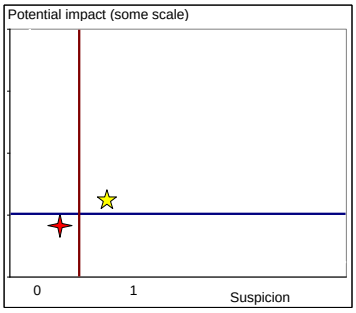
A tool for selective editing according the specifications we have made has many parameters to be set. Several of these can be set to the default values. We see today no other option than empirical studies within this concept to reach best performance, while fulfilling the demand on maximum 20 percent relative pseudobias or a similar demand.

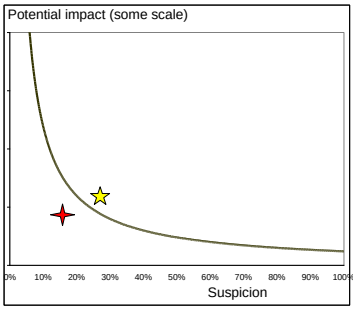
For one-time-surveys we can possibly wait until half of all the records have arrived and been entered. Use what is available, divided into homogeneous groups and compute measures of central tendency and dispersion as above. This can be done continuously as more data are stored.

Expected values and other estimates are computed in homogenous groups. These may, but need not, correspond to strata or domains of study. In SELEKT, the groups can be formed in a generalized way by a set of classificatory variables, the detail of classification (number of digits) and a parameter stating the minimum number of observations required for the computation.

Selective editing and significance editing

Figures Selective editing after flagging with traditional edits (left) and significance editing with a combination of a continuous measure of suspicion and potential impact (right)





A continuous measure of suspicion

We are modelling suspicion with two parameters KAPPA and TAU and a homogenous group of cold deck data. Here we exemplify by setting expected value as the median and the quartiles as the measure of dispersion in the cold deck data.

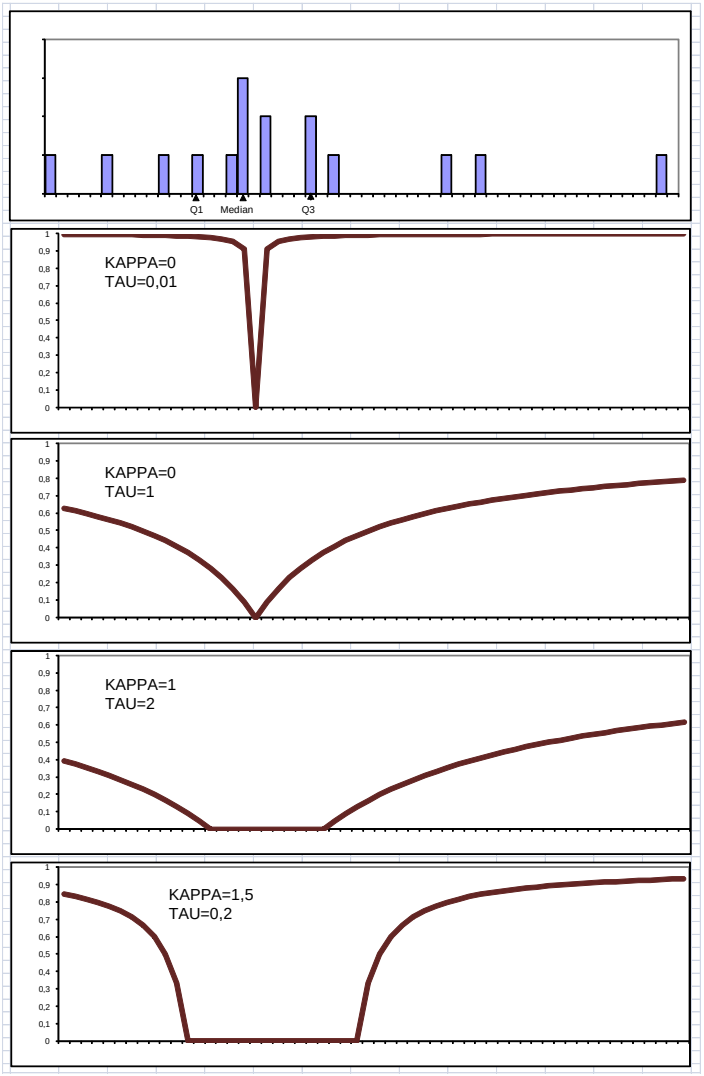
KAPPA=0 gives suspicions >0 when the observed unedited value differ from the median of the cold deck data., i.e. for practically all data.

KAPPA=1 gives suspicion=0 for unedited data between the lower and the upper quartiles of the cold deck data.

Larger KAPPA´s broaden the range where suspicion is set to zero.

TAU is decisive for the shape of the curve. For small TAU-values suspicion is close to 1,0 when an observed unedited value lies outside the interval decided by KAPPA.

A large TAU makes the suspicion almost proportional to the distance away from the center of the distribution in the cold deck data.



The Score function

The score function is combined of three parts; the suspicion of the unedited data value, the potential impact on the output table of the possibly erratic value and a weight. The weights are stored in a so called CELLO-matrix, computed by parameters regarding importance of classifications in output (Line of Business, Gender, Occupation etc.), importance of variables in output (Total salary, Salary per hour etc.) and the size of estimate in the output cell. Besides the importance parameters, CELLO is a function of the estimated totals from a previous period, $\hat{Y}_{t-1}$, their corresponding estimated standard errors, $\hat{\sigma}_{t-1}$, for the variable and domain, and a parameter ALFA.

$$\text{CELLO} = \frac{\text{Importance of classification} \cdot \text{Importance of variable}}{(\text{maximum}(\text{ALFA} \cdot \hat{Y}_{t-1}, \hat{\sigma}_{t-1}))^{\text{Importance of size}}} \qquad (1)$$

Suspicion multiplied with potential impact on output by a possibly erroneous unedited data, we call the anticipated (expected) impact. CELLO transforms the anticipated impacts for variables on different scales and variation to comparable levels. The score on the most detailed level is

$$L_1\text{Score} = \text{Suspicion} \cdot |\text{Potential Impact}| \cdot \text{CELLO} \qquad (2)$$

Scores are aggregated from output domain cells via variables and observed units to primary selected unit (or respondent) by

$$L_{x+1}\text{Score} = \left(\sum [\max(0, L_x\text{Score} - Tr_x)]^{GAMMA_X} \right)^{1/GAMMA_X} \qquad (3)$$

The thresholds Tr_x and the powers $GAMMA_x$ are parameters to set.

Documentation

The method and requirement for building the IT-tool is described in detail in SCB 2009-02-20, "A General Methodology for Selective Data Editing", preliminary version 002.