Proceedings of MSP2009
Modernisation of Statistics Production 2009

Editing at Statistics Sweden – Yesterday, today and tomorrow

Anders Norberg[1]

An internal survey at Statistics Sweden in 2004 revealed that about one third of all resources among business surveys are spent on data editing. In order to analyse the possibility to reduce this proportion, nine case studies were conducted on editing. The business surveys that spend most resources on editing and related processes, in absolute terms, were included.

The main purpose of the case studies was to learn about similarities and differences between the surveys with regard to editing. Knowledge and experiences gained from the studies was meant for use for development of generic tools for editing.

Efficient editing means that we, with a high hit rate, find data errors that have a noticeable, or even decisive, effect on the statistics. A potential for efficiency gains showed to be present in at least seven surveys, ranging from about 25 to 50 percent.

The expectations on generic tools for editing are that (1) fewer IT-tools means decreased system maintenance cost and (2) more flexible distribution of the work among the editing staff due to well-known interfaces. Efficient methods means (3) fewer records to follow-up by re-contacts and this, again, means (4) better work environment for editing staff, not having to re-contact respondents who have delivered *correct* data. A coordination of process data will be used for improvement of the measurement process (5).

A systematic work is going on to construct generic tools at Statistics Sweden. A paper is produced which serves as the methodology documentation and requests on the IT-solution, all in one. The method for selective/significance editing is outlined; defining methods, parameters and ways to asses best choices. A prototype of IT-tools has been built as part of the development, the SELEKT prototype software. It is put into practice and test in a couple of surveys 2008-2009.

Key words: Resources; Editing; Generic tools, Confidence.

1. Yesterday – Too much editing

Editing is a resource-demanding process for statistical products with businesses and other organizations as information providers (respondents). A study of 62 statistical products at Statistics Sweden (2004) showed that 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.

[1] Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden, anders.norberg@scb.se

Table 1 Average proportions of costs of sub-processes 2004

Process	Proportion of total cost (%)		
	All products	Short-period surveys	Annual and periodic surveys
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 study showed that there were no homogeneous editing methods in use, despite the CBM-type of handbook in use, Statistics Sweden (2002). The editing process was at that time decentralised to each and every survey, the "stove-pipe production model".

2. Yesterday and Today – The role of editing

2.1. If...

- If we only want information from businesses that we know they have,
- and we ask for that information so they understand what we want,
- and we motivate them to deliver as good quality in data as possible,
- and we help them to avoid accidental errors in answering our questionnaires;
- then editing would be a minor process!

2.2. The new role

A new role of editing is slowly being implemented at statistical institutes. It focus the editing on identifying and collecting process data on errors, problem areas and error causes in the measurement process to provide a basis for a continuous improvement of the process and the whole survey vehicle in general. The old paradigm – the more and tighter the edit checks and re-contacts, the better the quality – should be replaced (Granquist 1997).

The entire set of the query edit checks should be designed to focus on errors influencing the estimates, and be targeted on existing error types. The effects of the edit checks should be continuously evaluated by analysis of performance measures and other diagnostics, which the process should be designed to produce, i.e. process data is also used to improve on the editing process itself.

When editing primarily is quality control of the measurement process, it is still needed to contribute to quality declaration and to adjust *(change/correct)* significant errors in the current survey round to avoid bias.

2.3. Editing staff debriefing

A tool for collecting information on respondents problem is the editing staff debriefings. The editing staff that work with a particular survey meet and discuss their experiences, in presence of a moderator from the Unit for Cognitive Methods. The purpose is to find out how the respondents understand the questions, which questions are problematic and what kind of error indicators that turns up in the editing process. As such, although editing staff debriefing is a qualitative technique in nature, it can provide ideas about how common certain problems are. (Hartwig 2009).

3. Today – Development of generic tools

Data collection and editing of data from businesses and other organisations is now centralised at Statistics Sweden to one department. After the move, there is a potential for efficient spread of workload among the editing staff. But, there is a heavy demand for common tools to make this possible.

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 necessary generic tools. Benchmarking had given us the information that there were no system at other national statistical institutes that would yield the properties we wanted.

3.1. The project "Nine case studies"

As a first step nine case 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 differences 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 editing intensive business surveys were included in the project. The surveys included in the project differ in many respects, some of which is of significance for how editing is performed.

Periodicity: It makes a great difference if one has access to data from previous surveys when estimating expected values etc. to compare unedited data with. There are three situations. 1) One-off surveys and surveys that are conducted seldom have no information from earlier observations that could provide a basis for finding reasonable edit checks. 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. Time series analysis can be used to produce forecasts. It is important to notice that even in a monthly survey some units are new when a new sample is drawn.

Survey design: Distinction is made between 1) sample and 2) population surveys. In the case of samples, weighting is always involved, which must be considered. 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 expected values. One- or multi-stage samples make a difference in complexity.

Types of units: In principle, type of unit – 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 followed up 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. Calculations are based on edited data, not using raising factor, for homogenous groups. All units are included in the groups for cross-sectional data no matter if they belong to an old outgoing sampling panel or the present one.

Empirical data: Data from previous survey rounds are needed to estimate expected values and to define edit checks with efficient threshold values. A precondition for being able to introduce and also adjust already established methods and parameters for efficient 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 quality data with generated/simulated errors, which allows full control over the search for errors.

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. 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 impact on output only for the derived variable.

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. Generic 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 often lead to more editing, but this cannot 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.

3.2 Generic tools

3.2.1 Expectations on new methods and generic tools

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

- Standardized, common, generic tools lead to:
  - Less maintenance of IT systems, reducing a large cost for the NSI.
  - Easier planning of the manpower-demanding editing work for the total set of business surveys as individuals of the staff can work with several surveys when they are well acquainted with the tools.
  - Better work environment for the editing staff, when being familiar with an efficient tool.
  - 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 having to re-contact so many respondents that consider their delivered data to be correct. This is so when high hit rates of edit checks is a quality of the process.
- 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.

3.2.2 Selective / significance editing

Selective editing can be used as a complementary second stage 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 editing. In this latter case, we propose simple continuous propensity measures for suspicion.

Figure 1a Two-stage selective editing; data are first flagged by suspicion and thereafter the list of flagged data is reduce by potential impact

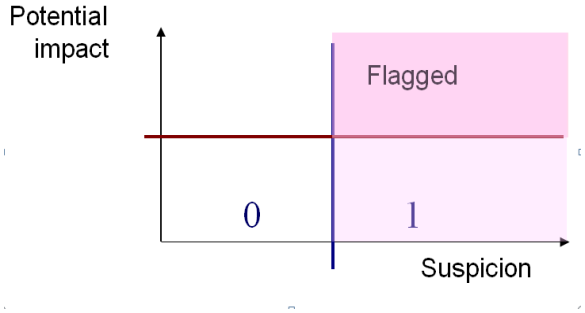
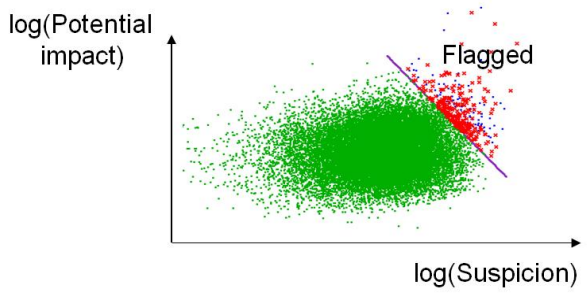


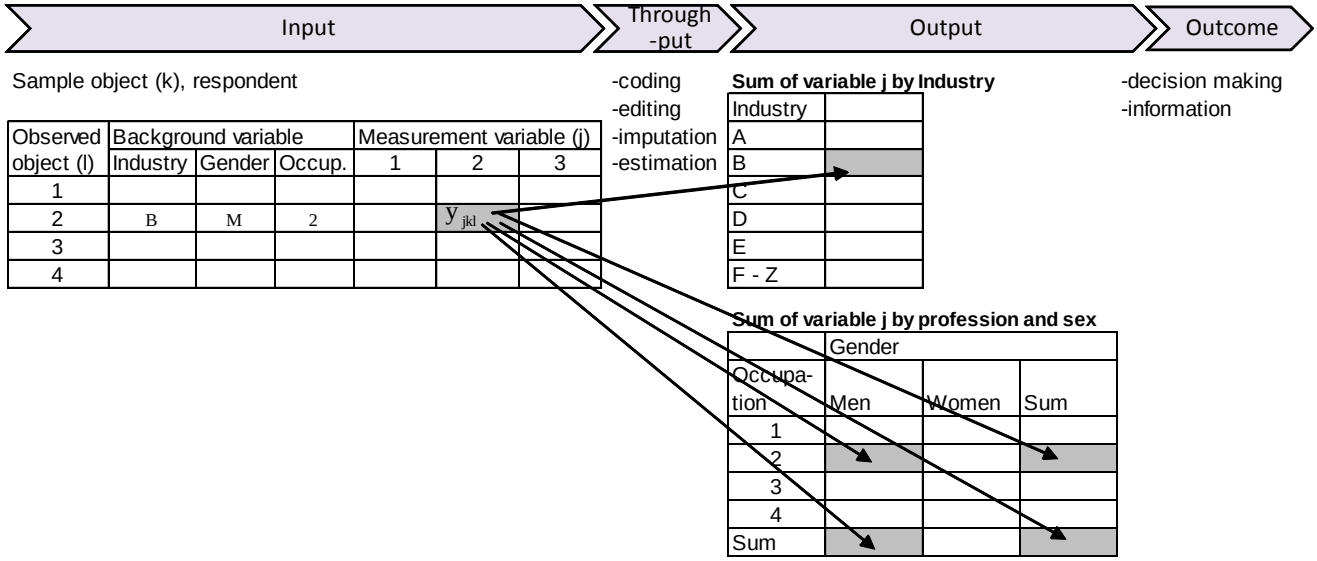
Figure 1b Simultaneous significance editing by suspicion and potential impact. Jäder and Norberg (2005)



3.2.3 Input, Output and Outcome

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 table 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 ideally provide a description of what is most important, presented in relative numbers. This information will be used to adjust parameter settings etc. in order to obtain high quality where it is most needed.

Figure 2 Input data, production of statistics and use of statistics. Suspicion on a data value y_{jkl} can be estimated by a variety of methods and historical data . The potential impact on statistical output, if input data is erroneous, is the difference between received data value and expected value weighted according to the estimation formula.



4. Tomorrow – The SELEKT and EDIT systems

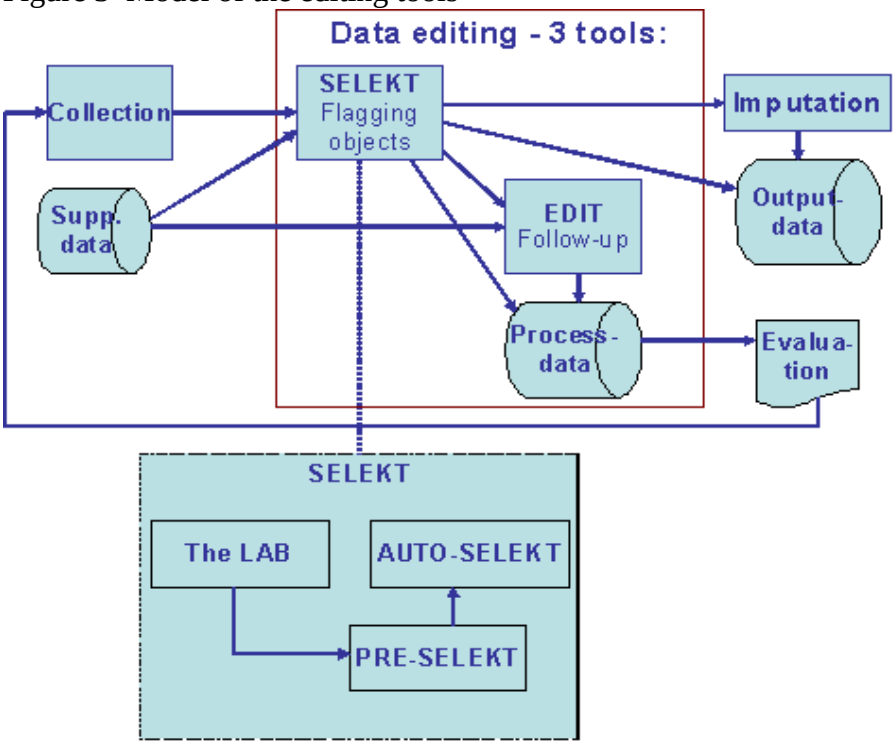
The generic tools are of three kinds.

A. The method and the IT tools for flagging of incorrect or suspect data values through selective/ significance editing is called SELEKT. Primary selected units, observations and variables are flagged to go to manual follow-up, imputation or acceptance by 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 calculate according to the settings made in PRE-SELEKT, by reading the parameter table. A so called laboratory environment, LAB, is a third production tool box in the SELEKT. The LAB will be used before implementation in surveys. The LAB-modules are 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 also used later in order to now and then adjust parameters with 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 of process data are generated in the editing process. A cohesive investigation concerning process dates and analysis of these is required for the editing process.

Figure 3 Model of the editing tools



5. Tomorrow – The SELEKT system

By the end of January 2010 the first version of a set of generic tools is planned to be at place. A first prototype was implemented in production in October 2008. The tools have 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.

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 for 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.

5.1. A continuous measure of suspicion in SELEKT

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.

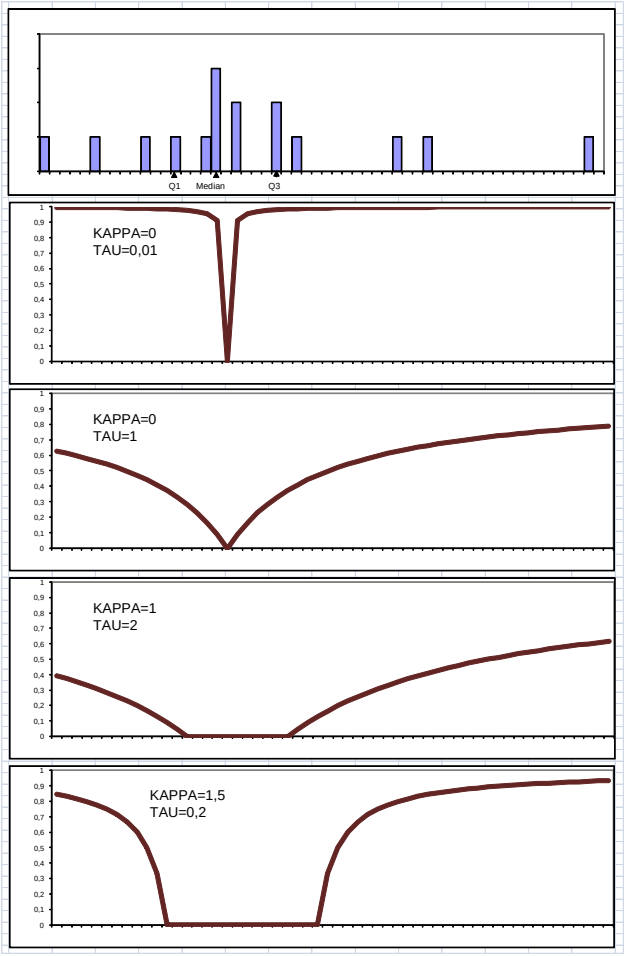
KAPPA =1 gives suspicion=0 for unedited data between the lower and the upper quartiles of the cold deck data.

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 centre of the distribution in the cold deck data.

Figure 4 Modelling suspicion on cold deck data



5.2. 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 the 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.

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

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 (5) is

$$\text{Score5} = \text{Suspicion} \cdot \left|\text{Potential Im pact}\right| \cdot \text{CELLO}$$

Scores are aggregated from output domain cells (5) via variables (4) and observed units (3) to primary selected unit (2) by

$$\text{Score}[x-1] = \left(\sum [\max(0, \text{Score}[x] - \text{Treshold}[x])]^{\text{LAMBDA}_{[x]}} \right)^{1/\text{LAMBDA}_{[x]}}$$

The thresholds and the powers LAMBDA are parameters to be set.

The indexation leaves the level (1) to respondent. In some surveys one respondent can respond for several primary selected units.

5.3. Quality measure

We have adopted the concept of relative pseudobias (RPB), the bias of an estimate that is due to follow-up of only a selected subset of input data, assuming there being some errors left in the data, relative to standard error of estimate. The goal is a maximum 20 percent relative pseudobias or a similar demand in most output statistics.

5.4. Characteristics of SELEKT

To summarize, the characteristics of SELEKT are:

- Selective and significance editing
- Standard methods for analysis of cold-deck data are available in an integrated analysis pack
  - Option for constructing homogeneous groups by hierarchical use of explanatory variables
  - Time series and cross section analysis
- Potential impact is estimated for all important output
- Modelling of suspicion and potential impact separately
- Continuous suspicion measures can be computed
- 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.

5.4. Confidence in National Statistical Institutes

It can have a negative effect on respondents and personnel if too many erroneous items of information pass through without any reaction. A reputation that *the Statistical Institute will accept whatever data is supplied* hardly promotes the will to supply high quality data. In order to maintain confidence it is desirable to identify and re-contact respondents who have supplied data that are strongly suspected to be erroneous. The use of efficient editing in order to maintain low

costs for production may conflict with this quality. This is especially so if there are data with minor impacts on the output.

Item non-response can have its roots in bad measurement tools but also in some respondents easy way to keep away from respondents duty. Item non-response is one type of fatal error, sometimes easy to identify and sometimes impossible to discriminate from implicit zeros. Either we make a re-contact with all respondents with item non-response, even if this is not justified from a short-term resource perspective, or we treat them in selective editing with full suspicion and an arbitrary estimated impact on the statistics. If, in short term statistics, it is a repeated procedure by a respondent not to deliver information, the impact on statistics should be estimated not only for one survey occasion but for a series for which the respondent is supposed to be in the sample.

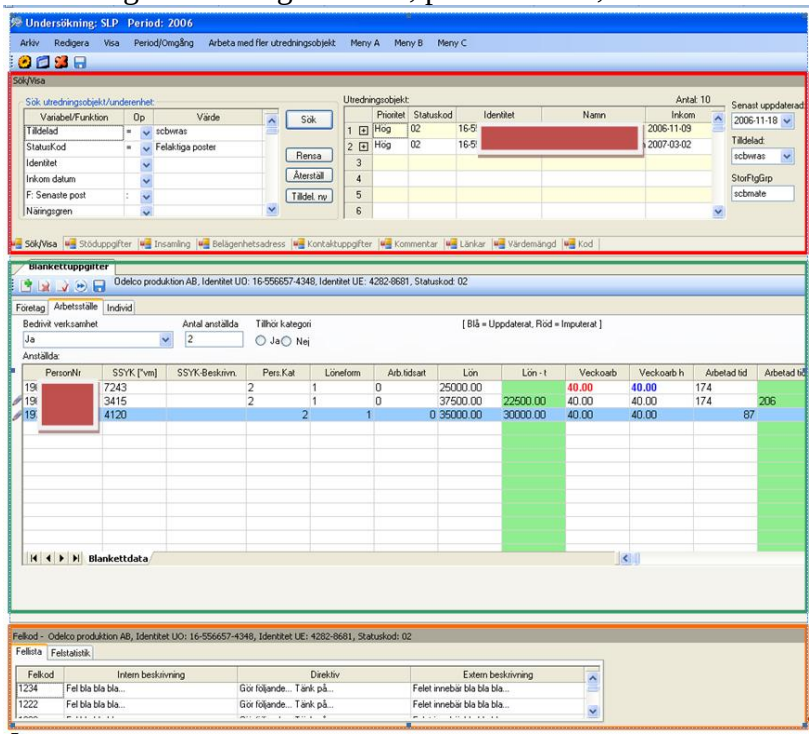
Data from electronic questionnaires, where edit checks are implemented and processed by the respondents, should only sparsely be flagged for follow-up in the production editing process. Repeated re-contacts hardly promotes trust from respondent to what the NSI is doing.

The general concept of selective/significance editing is not suitable for detection of inliers. Suppose that a respondent supplies exactly the same information on several survey occasions in a survey where this may not be reasonable. Making editing more efficient can release resources to be spent on measurement problems, identified by other means as inliers.

6. Tomorrow – The EDIT system

The EDIT system is the interface to data, the working environment for the editing staff. To be generic it must be very flexible for different types of survey data. It must give service to the editor as presentation of previous data and analysis, look-up for register data etc. The tool will have a windows look with a lot of tiles. The development of this tool is much more complex than that of the SELEKT tools.

Figure 5 The EDIT interface, a series of tiles giving support to the editing staff on register data, previous data, etc.



7. Tomorrow – Process data

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 and reach a more optimal resource allocation. Process data can also give some information of the quality of the final product. Uses:

A. Control of next process steps:
   AUTO-SELEKT points out what data to be managed in the manual follow-up (EDIT) and imputation procedures. Error flag codes are set for units and variables that show what to be tackled together with accompanying review instructions.

B. Analysis of the editing process:
   Examples: Number of items sent to manual follow-up, re-contacted and changed. Hit rate is an important indicator for reducing respondent burden.

C. Evaluation of the data collection process:
   Can indications be seen that any variable has shortcomings in quality?
   Examples: Number of changes, high item non-response rate.
   Respondents reasons for making error in the first place can be registered by editor.

D. Method - for editing the survey data:
   Are thresholds and other parameters good enough?
   Example: Hit-rate for edit checks, analysis of expected values

E. Method - General:
   Learning from experience - how does SELEKT work in different types of surveys?
   Examples: Which is the best expected value; mean or median?
   How are local scores best aggregated to global scores?

F. Quality indicators as a basis for quality descriptions of statistics
   Example: Number of changed values for variables.

G. Micro-data to researchers
   Process data are added to final observation files.
   Examples: The suspicions, indication whether contact has been made with respondent or not

References

Jäder, A. and Norberg, A. (2005) "A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics", Invited paper for the Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005

Granquist, L. (1995) "Improving the Traditional Editing process", in "Business Survey Methods" (ed. Cox, Binder, Chinnappa, Christianson, Colledge, and Kott)

Granquist, L. (1997) "The New View on Editing". International Statistical Review

Hartwig, P. (2009) "How to use edit staff debriefings in questionnaire design", paper presented at 2009 European Establishment Statistics Workshop, Stockholm.

Norberg, A. et. al. (2009), "A General Methodology for Selective Data Editing", preliminary version 002, SCB, not published.

Statistics Sweden (2002) "Guide till granskning", A Current Best Methods type of handbook, SCB 2002:1