STATISTICS SWEDEN

TOWARDS EFFICIENT STATISTICAL DATA EDITING: THE SWEDISH EXPERIENCE

Anders Norberg*

Statistical data editing is a resource-demanding process in business surveys. A 2004 study at Statistics Sweden demonstrated that around one third of resources were spent on editing (somewhat more for annual and periodic surveys than for monthly and quarterly surveys). Most resources were 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. Output (macro) editing is another sub-process that has the potential to be improved and to be more important. Output editing can detect errors introduced in the production and compilation processes. When resources can be released from the large micro editing process, some of these same resources should be invested into these two types of editing.

The role of editing

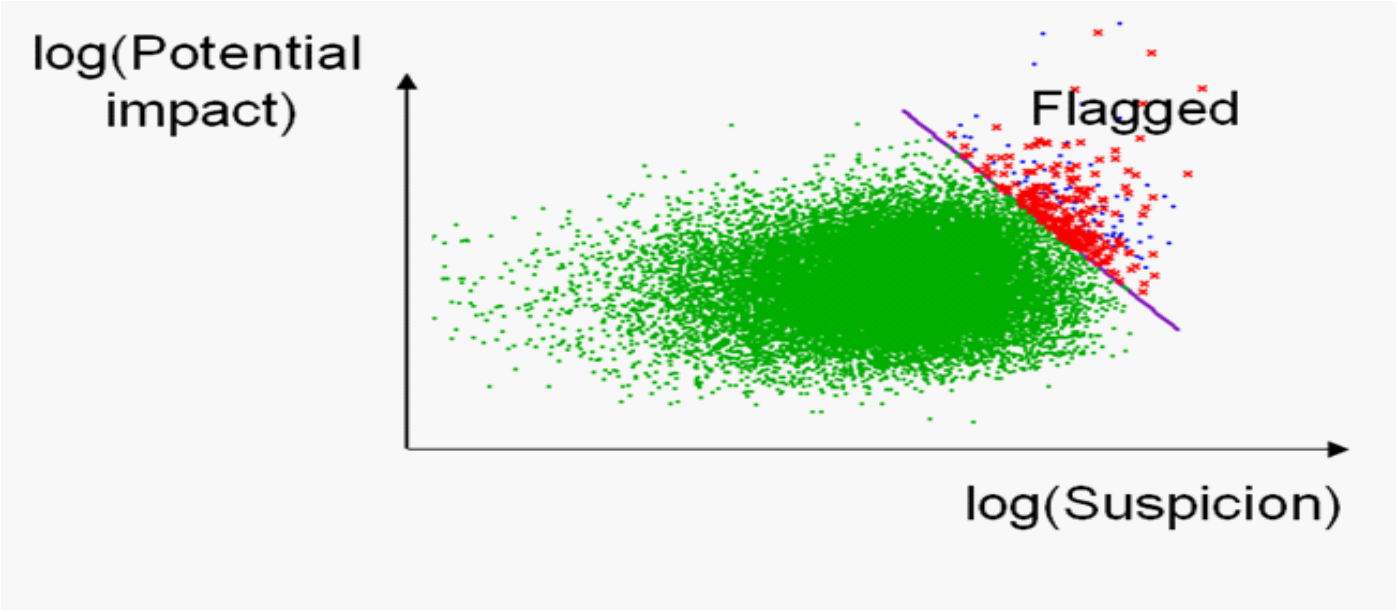
A new role of editing is slowly being implemented at statistical institutes. Its focus is on collecting process data on problem areas and causes of errors in the measurement process. These data will provide a basis for a continuous improvement of the measurement process and the whole survey vehicle in general. The old paradigm, "…the more and tighter the edit checks and re-contacts, the better the quality", should be replaced [Granquist, L. (1997) "The New View on Editing" International Statistical Review]. However, when editing primarily is quality control of the measurement process, it is still needed to adjust (change/correct) significant errors in the current survey round and to contribute to quality declarations.

The role of the query edit checks should be designed to focus on errors influencing the estimates. The effects of the edit checks should be continuously evaluated by analysis of performance measures, which the editing process should be designed to produce. The software SNOWDON-X developed jointly by the UK Office for National Statistics and Southampton University of the UK, is a good tool for this.

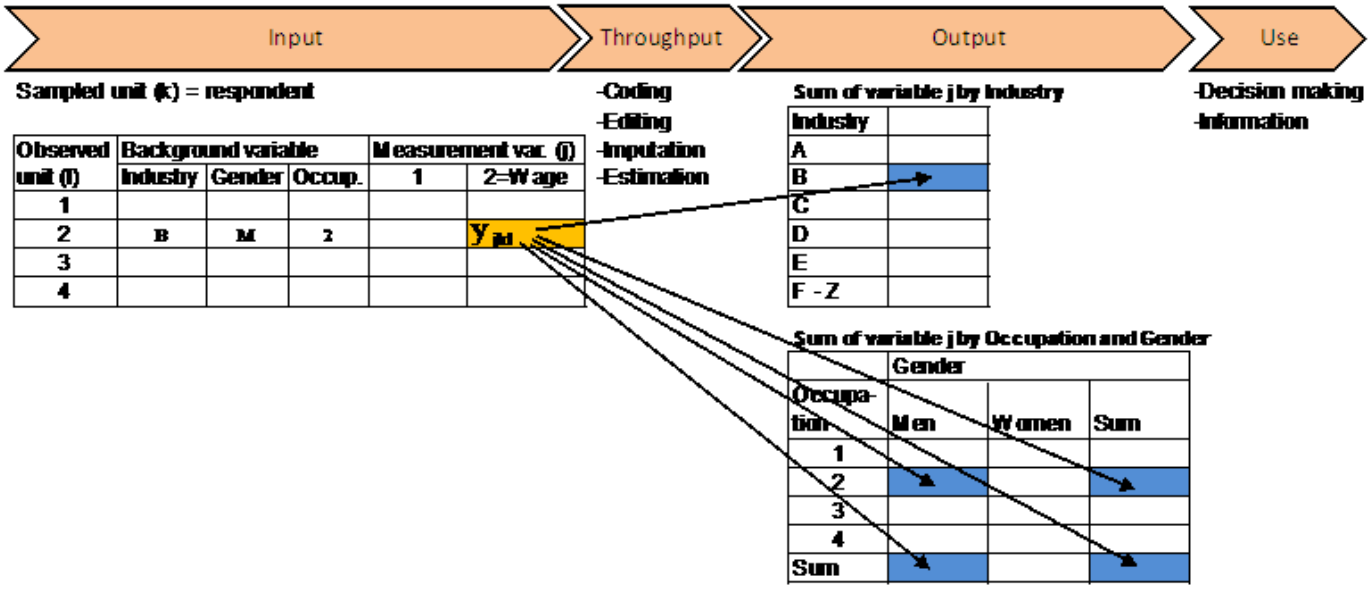
Editing staff debriefing is a method for collecting and analysing information on problems the respondents report from notes and contacts. At Statistics Sweden editing staff working on a particular survey meet and discuss their experiences in presence of a moderator from the Unit for Cognitive Methods. The purpose is to find out how the respondents understand the questions, which questions are problematic and what kind of error indicators are turning up in the editing process. As such, although the editing staff debriefings are a qualitative technique in nature, they can provide ideas about how common certain problems arise.

Good methods, generic IT-tools and a structured collection and analysis of process data have the potential to give better meaning to the editing job and provide for a better working environment for the editing staff. This manpower-demanding work will be easier to plan for in the business surveys area

Simultaneous significance editing of foreign trade statistics by suspicion and potential impact



Input data, production of statistics and use of statistics



as staff can work with and on several surveys collectively once they are well acquainted with the new IT editing tools. The editing staff will not have to contact or re-contact as many respondents (some annoyed) that have considered their delivered data (questionnaire) to be correct in the first place. This will happen when high hit rates of edit checks become part of the quality process.

Selective / significance editing

Traditional edit checks often only focus on "suspicion" towards a unit's value for a single variable. Flagged data are suspected whereas un-flagged data are accepted. There is a dichotomisation (in other words, the data have been divided into two opposing groups) of suspicion. Selective editing is a procedure which targets only some of the flagged variables or records that failed at least one edit check for manual review. This selection is based on the potential impact on estimates from the suspected error.

We see the dichotomisation as a waste of information when it is possible to measure suspicion on a continuous scale. The suspicion grows with the distance from the expected a priori distribution of the un-edited variable value. Suspicion and potential impacts can be treated simultaneously to form a score in significance editing. In foreign trade statistics the statistics produced are values of imports and exports and consequently the potential impact of a suspected error is expressed as an error in transaction value. Suspicion for the record is based on price per quantity, and as these two have a poor correlation both dimensions are important.

Respondents, producers and customers

One erroneous input data value can have an impact on several output statistical values. This is so when output is spread by more than one variable, for example when wages are presented by industrial sector, gender and occupation. Here, as in design in general, it is necessary that the national statistics institute can assess the quality demands of each output table from the users point of view.

Suspicion on a data value yjkl can be estimated by a variety of robust methods and from the saved edited "cold" data. The potential impact on statistical output, if input data is erroneous, is the difference between the received data value and an expected/ predicted value, weighted according to the estimation formula.

We have adopted the concept of relative pseudobias (RPB) to evaluate the quality of editing. This bias of an estimate is due to the follow-up of only a selected subset of input data (assuming that there are some errors left in the output data): the bias is analysed relative to the standard error of the estimate. A 20 percent RPB has little contribution to the total error in most statistics.

Generic tools

Tests of selective/significance editing were performed for nine of the most editing intensive business surveys in 2007 at Statistics Sweden. We saw likely efficiency gains and likely cost reductions. We also realised that the introduction of new methods demand intensive testing in every specific survey because of the variation between the surveys regarding data structure, users demands of the statistics, etc. Generic tools for editing must therefore be very flexible to be able to deal with these different situations.

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 variation on cold deck data to be used in the edits. AUTO-SELEKT calculates scores according to the parameter table, indicating the expected impact on all important output. In a laboratory environment and supported by modules in PRE- and AUTO-SELEKT, tests will be undertaken before implementation and up-dates to surveys to evaluate the earlier production rounds.

Expected/predicted values and variation are computed for homogenous groups. These may, but need not, correspond to strata or domains of study. In SELEKT, the 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. Estimation of totals, functions of totals and their estimated standard errors is done by a generic tool, outside the 'selekt' family. Several different types of software can undertake this calculation, Statistics Sweden uses its own software: CLAN.

EDIT is the tool editing staff use to follow-up flagged items. To be generic it must be flexible for different types of survey data. In this sense EDIT will have a standard interface, a windows look with a lot of tabs, functionality that presents all of the information needed such as previous data and analysis thereof, register data look-up ability, etc. It must also be possible to ask SELEKT to check a specific batch of data and this needs to be undertaken quickly.

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.

The first version of the SELEKT software is in place to begin the implemented phase. Prototype versions have been implemented and tested in a few surveys to date. Experience will bring us forward to efficient editing. A new project, TRITON, aiming to make all tools for several current processes in data capture and data processing communicate with others has just started at Statistics Sweden.

*Anders Norberg is a Senior Statistician at Statistics Sweden.

Data flow and software