Implementing Selective Editing at Statistic Finland
Case Studies for Four Statistical Surveys

Marjo Pyy-Martikainen
Statistics Finland, Helsinki, FINLAND
e-mail: marjo.pyy-martikainen@stat.fi

Abstract

The traditional approach to editing in the production of official statistics is to aim at detecting and correcting all errors in the data. This approach leads to an editing procedure that is highly resource-demanding. In some surveys conducted at Statistics Finland, over 50% of the working time is used for editing. Moreover, the traditional approach to editing often lacks a clear strategy. Typically, a large number of errors that have little impact on results are corrected. Selective editing helps target editing to errors that have a significant impact on results. Both the likelihood of errors and their impact on results are estimated in order to calculate score values for survey units. These score values help prioritize units subjected to manual editing. This presentation will describe the implementation of selective editing at four statistical surveys conducted at Statistics Finland. International trade in services, Finance of housing companies, Quarterly statistics on the finances of municipalities and Register-based statistics on buildings and dwellings were chosen as pilot surveys in a project that aims to harmonize and improve editing practices in the framework of an editing model developed at Statistics Finland. Selective editing is one of the new tools to be implemented in the project. The usefulness of selective editing in the pilot surveys is evaluated by the decrease in the number of units requiring manual editing.

Key Words: editing, official statistics, prioritization of units, score values,

1. Introduction

In 2009, a project was launched whose aims were to survey current editing practices at Statistics Finland and develop a process model for editing. A survey on editing practices revealed that in roughly one out of five statistical surveys, over 50% of working time is used for editing (including error detection, correction and imputation), see Ollila (2010). Moreover, the editing practices turned out to be heterogeneous and lacking a means to prioritize units subjected to manual editing.

The process model for editing (Figure 1) forms a backbone for a new project whose aim is to harmonize and improve editing practices in surveys conducted at Statistics Finland. Selective editing is an important new tool to be implemented in the project. In fact, the launching of selective editing in statistical surveys is one of the strategic goals of Statistics Finland (Statistics Finland 2012).

International trade in services, Finance of housing companies, Quarterly statistics on the finances of municipalities and Register-based statistics on buildings and dwellings were chosen as pilot surveys in the project. These surveys represent different domains of study, different units and different sources of data. Common features for these surveys are a need for improving editing practices as well as the applicability of selective editing.

This paper describes the pilot surveys of the project, their current editing practices as well as practical issues related to the implementation of selective editing in them.

As regards Register-based statistics on buildings and dwellings, the work of the project has not started yet. Therefore, it is not included in the following.

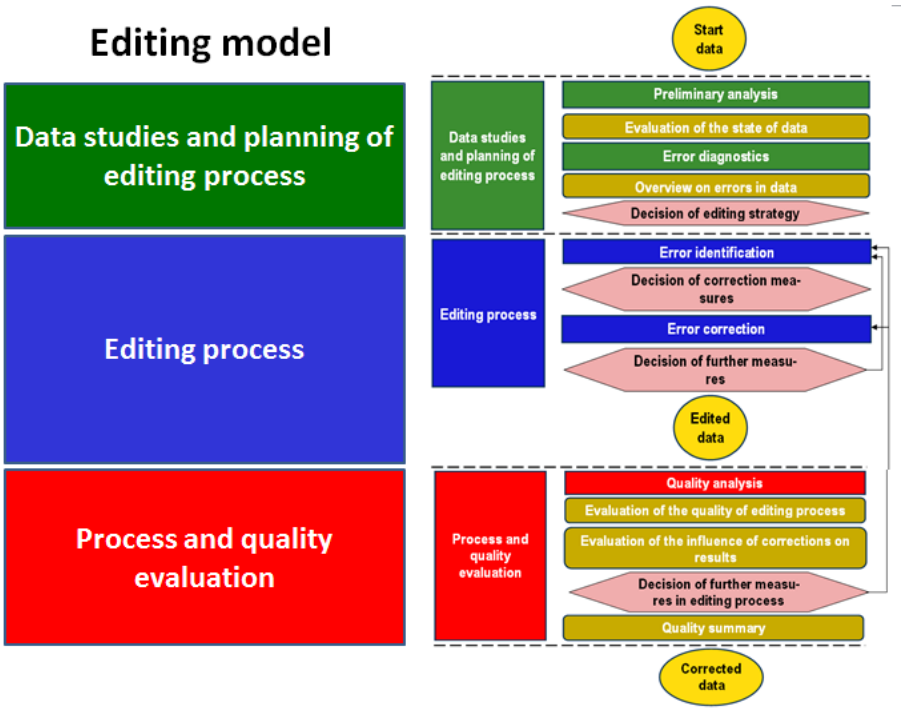

FIGURE 1: Process model for editing

2. Pilot surveys of the editing project

2.1 International trade in services

Statistics on international trade in services describe Finnish enterprises' international sales and imports of services both by service type and country of the foreign trading partner. Data is collected from approximately 3200 enterprises by quarterly (200 enterprises) and yearly (3000 enterprises) web inquiries. Most of the respondents are asked to provide data on the level of service type and country. Data on international trade in services is used by Bank of Finland in the compilation of the balance of payments. It is also used in the production of national accounts. Statistics on international trade in services are published yearly. The main output consists of total value of exports and imports by service type and country.

2.2 Finance of housing companies

Statistics on the finance of housing companies measure the cost of housing in housing companies, and analyse its composition and reasons for differences in it. The statistics also show how housing corporations finance their expenditure i.e. what their income is composed of and what residents have to pay for housing in them. The data on housing companies are based on a web inquiry made once a year to around 2,400 housing companies and to about 1,000 government-subsidised rental units. One third of the sample changes yearly. The data concern accounting periods. Statistics on the finance of housing companies are published yearly. The main output consists of mean expenditures and incomes in domains defined by building's location, age and type.

2.3 Quarterly statistics on the finances of municipalities

Quarterly statistics on the finances of municipalities describe development in the finances of Finnish municipalities (including from the beginning of 2013 also municipal enterprises and federations of municipalities) in the whole country by

quarter according to the execution of budgets. Data are collected by quarterly web inquiries. Data collection for quarters 1-3 is based on a sample of approximately 150 municipalities. The fourth quarter is a census of all municipalities, approximately 300 municipalities. Data are collected on municipalities' profit and loss accounts and investments, as well as their liabilities and certain receivables by quarter. Data are used for compiling quarterly statistics on municipalities. The results are published as totals over all municipalities. Data are also used for quarterly monitoring of municipal finances, for compilation of quarterly national accounts statistics, for preparation of quarterly statistics of the EU on general government income and expenditure, and for the EU's quarterly financial accounts of general government.

3 Principles of selective editing shortly

The traditional approach to editing in the production of official statistics is to detect and correct manually all errors in the data. This approach leads to an editing procedure that is highly resource-demanding. Selective editing is based on the idea that manual editing should be targeted to errors that have a substantial impact on key estimates. Selective editing splits the data into two streams. The critical stream consists of observations that are likely to contain influential errors. Observations in the critical stream are subjected to manual editing while observations in the non-critical stream are edited automatically or left unedited. Methods for selective editing have been developed by Lawrence and McKenzie (2000), Hedlin (2003), Granquist (1995) and Granquist and Kovar (1997).

4 Implementing selective editing at Statistics Finland

The use of selective editing is tested in four pilot surveys in a project whose aim is to harmonize and improve editing practices in statistical surveys conducted at Statistics Finland. The project has a planning phase and an implementation phase. At the planning phase, current editing practices of the survey are documented and changes made to them planned. The implementation phase consists of implementing the new editing practices in the production of statistics, as well as documenting the new methods and training persons in charge of the pilot surveys to use them.

The SAS application SELEKT developed at Statistics Sweden (Statistics Sweden 2011) is used as a tool for selective editing. For each unit k, variable to be edited j and domain of study d, SELEKT calculates a score value. The score value is a product of three terms: the likelihood of error (Susp(j,k)), its impact on the domain level estimate (Potimp(d,j,k)) and a CELLO(d,j) parameter that is used to adjust the importance of variables to be edited, and domains of study. These local scores are then aggregated into unit-level global scores that are used to prioritize units subjected to manual editing.

The SELEKT programs are incorporated at Statistics Finland in a SAS EG project together with BANFF (a generalized edit and imputation system developed at Statistics Canada) and SAS macros developed at Statistics Finland. SAS macros were developed for the preprocessing of data before the SELEKT implementation. Also, several test functions needed for the SELEKT type edits, including the Hidiroglou-Berthelot method (Hidiroglou and Berthelot 1986), were programmed as SAS macros.

4.1 International trade in services

Current editing practices

The editing of quarterly data starts during the data collection phase and continues until data are transmitted to Bank of Finland, around 6 weeks after the end of the quarter. Data are aggregated by service type and country into an Excel table and enterprises contributing to suspicious cells are subjected to manual editing. Suspicion is based on

comparison with previous year's data. Most enterprises of the quarterly survey undergo manual editing. Errors are corrected by contacting the respondent.

The editing of yearly data starts right after the end of the data collection phase and continues until beginning of December, around two weeks before publishing of the statistics. The edit rules are defined at enterprise level. Enterprises whose data are suspected are checked manually. Suspicion is based on comparison of reported data with register data (business register, tax administration's registers) and with previous year's reports. Enterprises subjected to manual editing are prioritized by degree of suspicion. Around one third of enterprises of the yearly survey undergo manual editing. Errors are corrected by contacting the respondent.

Implementing selective editing in Statistics on international trade in services
It was decided to concentrate on renewing the editing practices of the yearly data collection. The implementation of selective editing with an easy-to-read and prioritized listing of enterprises subjected to manual editing was set as a first priority.

Most of the old edit rules were transformed into SELEKT editing functions where the acceptance regions are based on times series data or, if not available, cross-sectional data. The old edit rules were defined solely at the enterprise level. Several new editing functions as well as traditional edit rules were defined at the level of enterprise * service type.

To handle editing at two levels in SELEKT, data were aggregated into two levels: 1) value of total exports/imports for the enterprise and 2) value of exports/imports for service types. Each enterprise contributes a row for each service type and direction of trade and for total exports and for total imports. The edit rules and editing functions making use of register data are applied only at level 1). This was done by setting the values of register variables for the level 2) rows as missing. In order to check service types with no reported trade at the current year and reporting exceeding a threshold level at the previous year, service type rows with missing values had to be generated to the current year data.

4.2 Finance of housing companies
Current editing practices
Most of the editing takes place after the end of the data collection phase. The variables being edited are related to housing companies' profit and loss accounts. The edit rules consist of balance edits, logical and range checks as well as outlier checks. Errors are corrected either deductively or by contacting the respondent. Some of the fatal errors are corrected automatically. Approximately one fourth of housing companies are subjected to manual editing.

Implementing selective editing in Finance of housing companies
A revision of current edit rules was conducted including several new aspects utilizing historic values. The process of making edit rules was parameterized. The key variables (e.g. some overall expenses and problematic reconstruction expenses) form the basis for carrying out the SELEKT process for gaining the prioritization of units subjected to manual editing. The editing functions include mainly the Hidiroglou-Berthelot method and the ratio method. Now when writing this paper, the solutions for the best alternatives of selective editing variables are being sought.

Data are input to SELEKT in a column-oriented form with one row for each housing company and one column for each variable.

4.3. Quarterly statistics on the finances of municipalities

Current editing practices

Editing starts during the data collection phase and takes place in a time slot of three weeks. The variables being edited are related to municipalities' profit and loss accounts and investments, as well as their liabilities and certain receivables by quarter. The editing of municipalities is done in several phases and in the order defined by the inflow of data. The edit rules make use of other variables in the questionnaire, data from the previous quarter and from the same quarter a year ago. Municipalities whose data are suspected are listed and then manually edited. Data from previous quarters are used in order to assess the validity of reported data. Corrections are based on re-contacting the respondent. Virtually all municipalities undergo manual editing.

Implementing selective editing in Quarterly statistics on the finances of municipalities

The staff in charge of the production of statistics expressed a clear need for the prioritization of municipalities subjected to manual editing. The inclusion of municipal enterprises and federations of municipalities in the target population implies an enlargement in the contents of the questionnaire and, thus, many new variables to be edited. New edit rules need to be defined and tested for these variables. The threshold values for error flags should take into account the size of municipality.

The SELEKT process is a combination of municipality-level edit rules and editing functions. In this case no division of data to critical and noncritical streams will be done with the scores achieved. Instead, the strategy is to edit manually as many municipalities as time resources allow in the order defined by score values. A few pre-defined, most important municipalities are assigned very high scores are thus always subjected to manual editing. For those low score municipalities, which have only non-critical errors, automatic correction can be considered when there is no time for manual editing.

Data are input to SELEKT in a column-oriented form with one row for each municipality and one column for each variable.

5. Conclusions

The pilot surveys are at the moment in different phases in the implementation of selective editing. Work with Register-based statistics on buildings and dwellings has not started yet, while Statistics on international trade in services is in the SELEKT LAB testing phase, where parameters of selective editing are adjusted in order to find an empirically optimal editing setting (for the LAB environment, see Statistics Sweden 2011). Quarterly statistics on the finances of municipalities and Finance of housing companies are in the phase of defining new edit rules and editing functions. A close cooperation of experts in statistical methods, persons in charge of the production of statistics and IT persons has turned out to be vital in the work of the project.

References

Granquist, L. (1995). Improving the Traditional Editing Process. In: *Business Survey Methods*, B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge and P. Kott, eds. John Wiley and Sons, New York, pp. 385-401.

Granquist, L. and J. Kovar (1997). Editing of Survey Data: How Much is Enough? In: *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwartz and D. Trevis, eds. John Wiley and Sons, New York, pp. 415-435.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K.

Office for National Statistics. *Journal of Official Statistics* 19, pp. 177-199.

Hidiroglou, M. and J. Berthelot (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* 12, pp. 73-78.

Lawrence, D. and R. McKenzie (2000). The General Application of Significance Editing. *Journal of Official Statistics* 16, pp. 243-253.

Ollila, P. (2010). Survey on editing practices at Statistics Finland. Unpublished report (in Finnish).

Statistics Finland (2012). Strategy of Statistics Finland for 2015. Unpublished report (in Finnish).

Statistics Sweden (2011). User's Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing.