UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (iii): Getting the support of all people when implementing data editing

Renewal of Editing Practices at Statistics Finland

Prepared by Marjo Pyy-Martikainen, Statistics Finland

I.   Introduction

1.   In the production of official statistics, the traditional approach to data editing has been to aim at detecting and correcting all errors in data. This approach leads to an editing procedure that is highly resource-demanding. The approach has been followed also at Statistics Finland.

2. Facing scarce resources for statistics production and an aim towards the standardization of processes, the work towards renewing of editing practices received a strong mandate from the management. The work was organized in subsequent projects. The first project is now finished and the second one is going on until the end of May 2014. The third project is being planned and will start in the beginning of September 2014. Projects at Statistics Finland are set by director general and they have a steering group consisting of managers from different levels of the organization.

3.   This paper describes the steps taken at Statistics Finland towards renewing editing practices in statistics. Section II describes the aims and results of the first two projects as well as some challenges faced along the way. Future steps of the renewal work will be sketched in Section III. Section IV contains a short summary of the paper and some concluding comments.

II.   Conducted renewal work

A.   Project 1: Editing project

4.   The renewal work started in July 2009 with a project somewhat generically named as Editing project. The aims of the project were
    (a) to survey current editing practices at Statistics Finland
    (b) to study current best methods and tools for editing and
    (c) to develop a process model for editing.

5.   The participants in the project were experts in statistical methods, IT experts and representatives from different statistics production departments. The project was led by a senior expert in statistical methods. As a first step, current editing practices were surveyed by a Digium questionnaire targeted to all statistics produced at Statistics Finland. There were 201 different sets of statistics conducted at the time of the survey and 185 of them were defined as relevant for the survey. The questions concerned types of data used for producing statistics, methods and tools of data collection, methods for error detection, correction and handling of missing data. There were also questions on working time used for editing as well as on controlling and documentation of editing.

DEPARTMENT	Missing	0–5	6–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	All
Population Statistics	2	16	7	7	1	4	4	1	0	9	51
Social Statistics	0	6	3	4	1	2	0	2	0	0	18
Prices and Wages Statistics	1	8	3	3	2	3	1	1	2	0	24
Economic Statistics	8	5	3	4	2	0	3	3	0	3	31
Business Trends Statistics	0	8	3	3	2	2	0	1	5	0	24
Business Structures Statistics	2	5	8	1	5	2	1	0	1	12	37
ALL	13	48	27	22	13	13	9	8	8	24	185

Table 1. Number of statistics by percentage of working time used for editing and by department (Ollila 2010).

6. The results from the survey revealed that in roughly one out of five statistics, over 50% of the time used for statistics production was used for editing (including error detection, correction and handling missing data), Table 1. In addition to being highly resource-demanding, there were other problems with the editing practices as well. The editing practices in different statistics were often somewhat outdated, heterogeneous and mostly lacking means to prioritize units subjected to manual editing. Moreover, only few statistics were evaluating the editing process or its effect on the quality of statistics. There was thus a clear need for modernizing and standardizing editing practices and making the editing process more efficient.

7. The second step in the project was to get familiarized with state-of-the-art editing methods and tools. This phase resulted in a report with several recommendations. Statistical software tools Banff (Statistics Canada 2007), SELEKT (Statistics Sweden 2011) and JMP (SAS Institute 2007) were recommended as tools for editing. Regarding editing methods, one of the recommendations from this step was the use of selective editing when applicable. The idea of selective editing -making the editing process more efficient by targeting manual editing to errors that have a substantial impact on key estimates - appealed to the management. As a result, the launching of selective editing by 2015 was set as one of the strategic goals of Statistics Finland (Statistics Finland 2012).

8. Results from these two steps were used to develop a generic process model for editing (later called editing model), which was the main goal of the project. Developing a generic editing model turned out to be challenging as there were very few examples available. The editing model describes the phases of editing as an ordered sequence of actions, decisions and assessments (Figure 1). The model divides the editing process into three main phases. The first phase contains a preliminary analysis of the variables to be edited as well as some preliminary error diagnostics. The results are used to make a decision on the appropriate error identification strategy. The second phase consists of Error identification and Error correction. Error identification results in information about fatal or suspected errors at the level of observations. An evaluation of error identification actions and a decision on correction measures is made. Error correction is conducted and the correction actions evaluated. An editing process may contain several sequential error identification and/or error correction actions. The influence of error correction actions on the quality of data is evaluated in the third phase of the editing model. On the basis of this assessment, a decision is made to take further error identification or error correction measures or to regard data as acceptably corrected.

9. The first project was finished in the end of 2011. The project produced information on editing practices and problems with them, suggestions for improving them and, as a main output, an editing model. With this information at hand, the management decided that it is time to start piloting the editing model as well as the new methods and tools. The planning of the piloting project started in spring 2012.
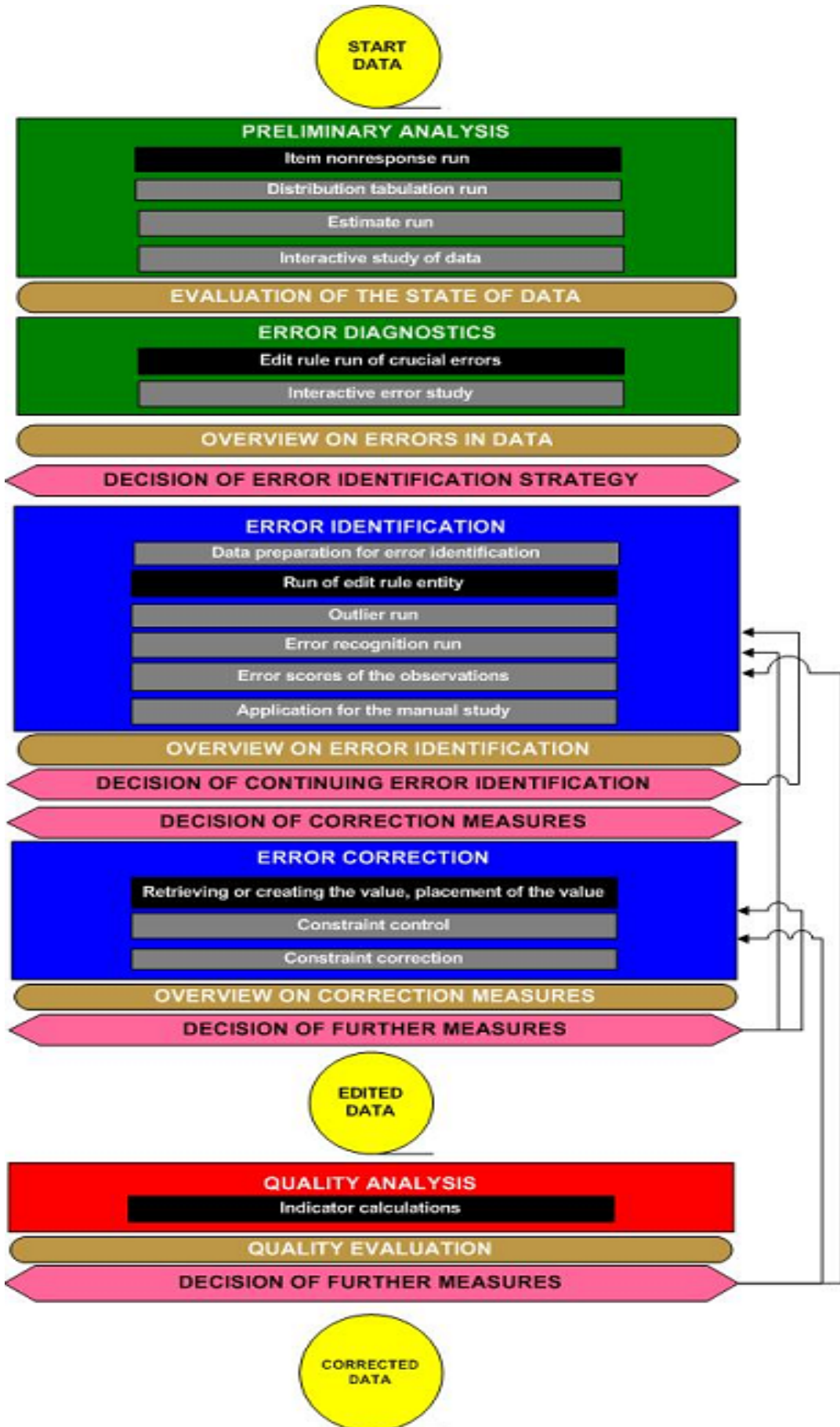
START DATA

PRELIMINARY ANALYSIS
- Item nonresponse run
- Distribution tabulation run
- Estimate run
- Interactive study of data

EVALUATION OF THE STATE OF DATA

ERROR DIAGNOSTICS
- Edit rule run of crucial errors
- Interactive error study

OVERVIEW ON ERRORS IN DATA

DECISION OF ERROR IDENTIFICATION STRATEGY

ERROR IDENTIFICATION
- Data preparation for error identification
- Run of edit rule entity
- Outlier run
- Error recognition run
- Error scores of the observations
- Application for the manual study

OVERVIEW ON ERROR IDENTIFICATION

DECISION OF CONTINUING ERROR IDENTIFICATION

DECISION OF CORRECTION MEASURES

ERROR CORRECTION
- Retrieving or creating the value, placement of the value
- Constraint control
- Constraint correction

OVERVIEW ON CORRECTION MEASURES

DECISION OF FURTHER MEASURES

EDITED DATA

QUALITY ANALYSIS
- Indicator calculations

QUALITY EVALUATION

DECISION OF FURTHER MEASURES

CORRECTED DATA

Figure 1. Process model for editing.

B.　　Project 2: Renewal of editing practices in statistics

10.　　The editing model forms a backbone for the second project, Renewal of editing practices in statistics. The aim of the project is to renew editing practices of four pilot statistics and change them to follow the editing model. The renewal consists of the following three phases:
- (a) Planning phase: the current editing practices are surveyed and new practices designed and tested
- (b) Implementation phase: the necessary changes in the production process are made, documentation and training is provided
- (c) Launching phase: The new practices are launched in the production process of statistics

The second project started in September 2012 and continues until end of May 2014. The project participants are experts in statistical methods, IT experts and representatives from the pilot surveys. As in the first project, the project leader is a senior expert in statistical methods.

11.　　To implement the editing model in practice, a generic SAS EG tool for editing was developed. The software packages BANFF and SELEKT were incorporated in a SAS EG project and complemented with SAS macro programs developed at Statistics Finland. BANFF is a collection of SAS procedures for editing and imputation developed at Statistics Canada. SELEKT is a SAS application for selective editing developed at Statistics Sweden. Various SAS macro programs were developed at Statistics Finland for preparing data and for defining edit rules and parameters of editing. The idea of the SAS EG tool is to create an environment where all the necessary information is given as parameter definitions that are used throughout the editing process. In this way, the programming by the user is minimized and the control of the process is turned into defining parameter values. For a more detailed description of the SAS EG tool for editing, see Oinonen (2014). The launching of the SAS EG tool for editing is expected to result in savings in costs and time due to systematization of the editing process and the decrease in the amount of manual editing. Moreover, having a common tool helps to standardize the editing phase of the statistics production processes.

12.　　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 statistics in the project. These statistics represent different domains of study, different units and different sources of data. Common features for these statistics are a need for improving editing practices as well as the applicability of selective editing. In the following, a short description of the pilot statistics is provided.

13.　　*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.

14.　　*Statistics on the finance of housing companies* measure the cost of housing in housing companies, and analyze 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 2,500 housing companies. 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.

15.　　*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.

16.   *Register-based statistics on buildings and dwellings* is an entity comprising two different statistics: statistics on buildings and free-time residences and statistics on dwellings and housing conditions. They describe the number of buildings, free-time residences and dwellings on the last day of the year. The most important classifications used are the intended use of the building and region. Both statistics are released annually, compiled from register data and cover the whole population.

17.   The heads of the pilot statistics were mostly keen on having the editing practices of their statistics renewed. A problem encountered in practice with most pilot statistics was the lack of time of the persons involved in the production of statistics. A lot of subject-matter information is needed when new editing practices are designed for statistics and this information can only be provided by the persons working for the statistics. Quarterly statistics on the finances of municipalities and Statistics on international trade in services have been undergoing major changes in their production processes alongside the renewal of editing practices (for changes in the production process of the latter, see Saarikivi 2013) which has demanded a lot of resources from the statistics producing staff. A consequence of this are some delays in the renewal of editing practices project.

18.   Statistics on international trade in services, Statistics on the finance of housing companies and Quarterly statistics on the finances of municipalities are at the time of writing in the implementation phase of the project. Statistics on international trade in services will be launching the new editing practices in the production process during spring 2014. For Statistics on international trade in services, the first priority was the implementation of selective editing with an easy to read list of enterprises prioritized for manual editing by both suspicion and impact on estimates. Tests conducted showed that there is a potential for reducing markedly the amount of manual editing. While the old practice was to edit manually roughly 1/3 of enterprises, it seems that by using selective editing, it suffices to edit manually only 15-20 percent of enterprises. The work with Register-based statistics on buildings and dwellings started in the end of 2013 and currently, new editing practices are being planned.

19.   The SAS EG tool cannot be just blindly run by the persons producing the statistics. Instead, some knowledge of the underlying statistical methodology is required. For this purpose, a database for concepts and definitions is being developed and training in editing methods provided. Hands-on training sessions in the use of the SAS EG tool are also provided. For the editing staff, the new methods seem to be somewhat difficult to understand. The challenge for the experts in statistical methods is to explain the ideas underlying the new methods in a non-technical way.   The editing staff have also expressed some concerns about the new tool taking over their tasks. Responding to these concerns is a challenge that should be taken up by the management.

20.   The project will be finished in the end of May 2014. By that time the new editing practices, including selective editing, should be in use in the production of the pilot statistics. The strategic goal of Statistics Finland is however to implement selective editing in *all* statistics where the method is applicable. To help meeting this ambitious goal, a roadmap for the implementation of selective editing was written (Ollila 2013). The roadmap indicates in which statistics selective editing is applicable and prioritizes these statistics on the basis of expected benefit from the launching of selective editing. The expected benefit is assessed by the amount of working time used for editing. Those statistics that use a lot of working time for editing should be prioritized in the launching of selective editing. The roadmap finds 36 statistics where selective editing is strictly applicable or might be considered after more detailed study.

III.   Future work

21.   A third project, Implementation of selective editing, is currently being planned. The project starts in the beginning of September 2014 and concentrates on meeting the strategic goal of launching selective editing by 2015 in all statistics where the method is applicable. The overly ambitious timetable has however been loosened somewhat by extending the duration of the project until the end of 2015. The roadmap as well as the editing model, the SAS EG tool for editing and knowledge gathered from earlier work will be used to achieve the ambitious aims of the forthcoming project.

IV.   Summary

22.   A survey on editing practices of statistics produced at Statistics Finland revealed that the practices were heterogeneous and often somewhat outdated. Moreover, only few statistics were evaluating their editing practices or their effect on the quality of statistics. A renewal of the editing practices was needed. The renewal work received strong support from the management as they recognized the potential for cost and time savings and for the standardization of production processes.

23.   A process model for editing and a SAS EG tool for editing were developed and tested with four pilot statistics. Most of the persons working for the pilot statistics recognized the need to modernize and systematize the editing processes of their statistics and accordingly, were in favor of the renewal. The editing staff were not in all cases welcoming the new practices. They found the new methods somewhat difficult to understand. Training in the new methods and tools is hoped to help in gaining support of all persons working for the statistics.

24.   The implementation of selective editing is set as one of the strategic goals on Statistics Finland and the management strongly supports achieving this goal. A necessary condition to achieve this goal is full commitment and sufficient resources from the statistics whose practices are being renewed.

V.   References

Oinonen Saara (2014). SAS Enterprise Guide project or editing and imputation. UNECE Work Session on Statistical Data Editing, Topic (v): International collaboration and processing tools, Paris, France, 28-30 April 2014.

Ollila Pauli (2010). A survey on editing practices at Statistics Finland (in Finnish).

Ollila Pauli (2013). A roadmap for the launching of selective editing (in Finnish).

Ollila Pauli & Rouhuvirta Heikki (2011). The process model for Editing (in Finnish), Statistics Finland.

Saarikivi Sami (2013). Revision Project of the Business Register and Business Statistics in Finland. 59th World Statistics Congress of the International Statistical Institute, Hong Kong, China, 25-30 August 2013.

SAS Institute (2007). JMP User Guide, Release 7.

Statistics Canada (2007). Functional description of the BANFF system for edit and imputation, version 2.02.

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

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