WP. 32
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (v): Editing based on results (post-editing)

CONDUCTED CASE STUDIES AT STATISTICS SWEDEN

Supporting Paper

Prepared by Chandra Adolfsson and Peter Gidlund, Statistics Sweden

I. INTRODUCTION

1. Editing is a resource-demanding process for statistical products with organizations as information providers (respondents). In a study at Statistics Sweden of the total costs of 62 statistical products in 2004, one third of the resources was 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. The short-period surveys are subjected to relatively less production editing.

Average proportions of the costs of sub-processes in the total cost of statistical products during 2003 and 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

Note: 9 non- responses out of 62, weighted as non-respondent cost of 9 percent

2. Over editing has proven costly for Statistics Sweden. In order to reduce the amount of editing and the associated costs an editing project was started in the end of 2006. The main purpose of the project was to analyze which modules for methods that should be included in a general tool for editing. As a first step on the way of constructing the tool a number of studies were conducted focusing on how and if selective editing with score functions could be used at Statistics Sweden. Other purposes of the 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.

- Structural business statistics

- Short term employment, private sector and Job openings and unmet labour demand, private sector
- Short term statistics, wages and salaries, private sector
- Wage and salary structures in the private sector
- Price indices in producer and import stages
- Foreign trade
- Business activity indicators
- Rents for dwellings
- Swedish national and international road goods transport

The surveys included in the project differ in many respects, which is of significance for how editing is performed. The most important differences are described next.

Periodicity

3. Distinction is made between three types of surveys in terms of how often they are conducted, which entails variable access to earlier data for the construction of edits.

A. 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.

B. Annual surveys and also intermittent surveys that, by contrast with A, have data of value from one or more previous rounds of surveys.

C. 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

4. Distinction is also made between sample and population surveys. In the case of samples, weighting is always involved. This means that the observed objects have different impacts on the outcome 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 also be used as homogenous groups in the estimation of good expected values which are used in selective editing. Another consideration that must be done is if the design of the sample is one- or multi-stage.

Primary and secondary sampling units

5. It is of significance if a provider is responsible for one observed object or a cluster of objects. An example of the latter is Wage and salary structures in the private sector where the providers supply one record for each employee (secondary sampling unit) in their enterprise (primary sampling unit). In surveys that include secondary sampling units the initial scores are computed at this level.

Types of objects

6. 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 recontact.

Variables

7. In a specific survey there might be hard to find proper expected values for some or all measurement variables. Job openings and vacancies in *Job openings and unmet labor demand* are

variables that often take on the value zero and are therefore hard to handle in the editing process. The expected values are used in the calculation of the scores in selective editing with score functions. The gained efficiency of selective editing is very much depending on the quality of the expected values. If the quality of the expected values is too low selective editing with score functions is not an appropriate method for editing.

8. In some surveys data are gathered on several variables that are not reported individually in the 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.

Output

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

Empirical data

10. In the case studies data from previous survey rounds were needed to set checks with effective threshold values to trace major deviations in observed values in each study. 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. Also, editing codes that show which checking rules have generated error signals from unedited data are used in such analysis.

11. Data of interest in an analysis like this might be:

- Data on objects gathered so far in the current survey round
- Edited data for the same sample or observation on the latest occasion
- Edited data for the same sample or observation on one or many previous occasions
- Edited data for another sample or observations from a previous occasion
- Registry data, e.g. from a sampling frame

12. Data can be used both in cross-sectional and time-series analysis. In the latter type of analysis more information can be extracted from the data, since time also becomes a variable. Trends, seasonal patterns, etc. can be estimated and used in forecasts for the period in question. In each survey a choice must be made whether or not to utilize imputed values.

II. SOME RESULTS FROM THE CASE STUDIES

Short-term employment and Job openings and unmet labour demand

13. In the surveys *Short-term employment* and *Job openings and unmet labour demand* about 14-16 percent of the incoming objects contain at least one fatal- or suspected error. Today all of these errors are handled manually by the editing staff at Statistics Sweden. The result of the case study revealed that introducing selective editing with score functions can reduce the number of errors that require manual attention by about 60 percent. Notice that *Short term business statistics on sick pay*, which is a part of the *Short-term employment* survey, was not included or analyzed in the case study. In the present *Short term business statistics on sick pay* requires a substantial amount of editing. There are two conditions that must be fulfilled before the present amount of editing can be reduced and it is appropriate to implement selective editing with score functions in the *Short-term employment*. These conditions are 1) *Short term business statistics on sick pay* also needs to be considered and analyzed since it is an integrated part on the *Short-term employment* questionnaire and 2) all the obvious measurement issues must be taken care of. These actions are necessary to be able to make a significant reduction of editing regarding this survey. The current editing method being used for *Job openings and unmet labor demand* is very inefficient, the hit rate is very low. Job openings and vacancies are variables that seldom take on a positive value and are therefore hard to handle in the statistical process and certainly in the editing process.

Short-term statistics, wages and salaries, private sector

14. The fraction of incoming objects with at least one fatal- or suspected error is about 60-65 percent in the *Short-term statistics, wages and salaries, private sector*. Less than half of the respondents are recontacted, for the most part the errors are edited manually by the editing staff at Statistics Sweden whom are able to edit the errors with help from other sources of information. The number of errors needing manual attention can be reduced by approximately 20-40 percent using selective editing with score functions, but the amount of the reduction of the manual editing has not been estimated.

Wage and salary structures in the private sector

15. The fraction of employed with at least one fatal- or suspected error is 31 percent in the *Wage and salary structures in the private sector* survey. This results in that 89 percent of the businesses are error flagged. About half of the error flagged businesses are recontacted. If a small decrease in quality is accepted, selective editing can reduce the portion of manually handled businesses with almost 25 percent. This fraction is though a little bit uncertain because of some simplifications that had to be done to be able to finish the case study in time.

Business activity indicators

16. The hit rate is very low in the *Business activity indicators* survey. The editing staff at Statistics Sweden look into and then accept most part of the error flagged objects manually, they find most of their support for their actions in historical data. If the number of editing controls is reduced and if selective editing is implemented the numbers of error flagged object will be reduced to half. Today 16 percent of the incoming objects are error flagged and this fraction will be reduced to 8 percent using selective editing. The amount of the reduction of editing can not be estimated using the available data material.

Foreign Trade

17. In the *Foreign Trade* survey´s enterprise-editing about one percent of the incoming enterprises are manually edited judgmentally. The hit rate is low. In the case study a SAS-script has been created that generates a ranked error list including about one percent of the respondents. The production editing can then be focused on respondents instead of enterprises.

18. In the *Foreign Trade* survey´s price-editing selective editing was implemented in 2004. Only 0,3 percent of the commodity items are error flagged and the hit rate is about 60 percent. In the case study the method was evaluated and the outcome was that five adjustments of parameters and some minor additions were proposed. The efficiency gain is at least five percent.

Structural business statistics

19. In the *Structural business statistics* survey selective editing is already implemented in one part of the survey. The method being used need to be evaluated and selective editing should be implemented in the remaining parts of the survey. Lack of useful data at the time for the case study made it impossible to estimate the efficiency gain from using selective editing in the whole survey.

Price indices in producer and import stages

20. In *price indices in producer and import stages* many seasonal commodities are error flagged but accepted without contact with the respondent. The prioritization of the error flagged objects is judgmental. In the case study selective editing including time-series analysis was tested. The quality of the production editing would be improved with selective editing, but a possible cost reduction has not been estimated.

Rents for dwellings

21. Recontact with the respondents in *Rents for dwellings* is taken for about 30 percent of the dwellings. The case study could not successfully use unedited data from a back-up file because the material was incomplete. One conclusion of the study is that the production editing should be focused on net rents instead of uncorrected rents. The overall conclusion of the study is that selective editing would work well in the survey.

Swedish national and international road goods transport

22. The only editing of *Swedish national and international road goods transport* is manual pre-editing. The manual pre-editing includes coding of commodity groups and manual imputation. This means that there are no unedited data available in the survey, because of this no evaluation of selective editing could be performed.

III. SUMMARY

23. The results show that implementation of selective editing is possible in many surveys. The use of selective editing will lead to efficiency gains and likely cost reductions. The experiences from the different 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 and so on. Tools for editing must therefore be very flexible to be able to deal with these different situations. The project team is working on a documentation of methods which defines and explains the method demands on a general tool for selective editing. This document will be input to coming projects focusing on developing general tools.

24. 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 suffers from measurement issues concerning at least some variables. Previous experiences from using efficient methods in the editing are limited at Statistics Sweden. A consequence of this is that 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.