16.10.03

CRP.4
ENGLISH ONLY

DEVELOPMENT AND USE OF DATA EDITING QUALITY INDICATORS

Comments by Leopold Granquist and Svein Nordbotten

1. Reflections

As a background for our discussion of the invited and contributed papers, we want to start by reflecting about the interpretation of the title of this session. The main topic is quality indicators for statistical data. Quality indicators are not the primary products of NSI, but belong to the domain of meta data describing properties of the statistical data.

The session concerns data editing domain in the general sense, i.e. including both editing and imputation. There is no restriction to *automatic* editing issues only, and the title opens for contribution related to *development* as well as *application*, which welcomes views also including indicators in manual and mixed editing, methodological considerations as well as experience from implementation and use of quality indicators.

The specific topic aims at data editing quality indicators, not on quality indicators for the statistical production in general. The term indicator means that we need not be concerned with measurements of the ideal concept of quality, but rather with measuring substitute variables, which we assume are varying in the same direction as quality. As a matter of fact, we shall never be able to measure quality itself, but must be satisfied by measuring indicators more or less successfully

Independent of how quality is indicated, a reference basis is required to which we can relate our current process and/or process observations. The reference basis may be a historical, a manually edited or an artificial set of output/performance data set. Such a set of data is often referred to as ‘true’, ‘cleaned’ and ‘edited’ data.

In our opinion, there are two categories of indicators, which are interesting in connection with the data editing process:

- indicators of the data editing output quality
- indicators of the data editing performance quality

The purpose of establishing editing output quality indicators is to describe to which extent the process performs as expected.

Who can benefit from the indicators? Two broad groups are the main interested parties:

- producers of statistics
- users of statistics

The producers' needs are connected to their task to select and improve methods, allocate resources to processes and declare the quality of their products. The users of statistics want to evaluate how well available statistics are suited for their particular needs.

2. Invited papers

Three papers were invited for the topic (i). The first, WP. 2: ***PROCEDURES TO IMPROVE THE DATA-CLEANING PROCESS BASED ON QUALITY INFORMATION*** was prepared by Thomas Burg, Statistics Austria. The paper describes experience of collecting data about the data editing process in a quality report system. The paper outlines how the data from the data base is used to improve statistics, and presents model prospects on using the collected and saved quality data to optimize the editing process. An approach for measuring the quality of the data editing process over time is outlined.

The paper raises several questions which the Work Session participants may want to discuss:

- What is the overall design of the Austrian Quality Report system including the Quality Database?
- Which are the main variables collected and stored in the system, and how are they observed/measured?
- From the paper we can read that the system was established in 2001-2002. What were the development and implementation costs?
- Do other NSIs have implemented similar systems, and is the experience reported in Mr. Burg's paper compatible with experience from other NSIs?
- Which are the comments of the participants of the Work Session to the list of possible questions for a checklist for managers on evaluation of the data editing?

Paper WP.3: ***EVALUATING, MONITORING AND DOCUMENTING THE EFFECTS OF EDITING AND IMPUTATION IN ISTAT SURVEYS*** was prepared by Giorgio Della Rocca, Orietta Luzi, Emanuela Scavalli, Marina Signore and Giorgia Simeoni, ISTAT. ISTAT has during the last decade developed one of the more competent groups in research and development of editing and imputation. Following an introduction to evaluation of quality, the paper describes the tool *Indices for Data Editing Assessment* (IDEA) and the more general *Information System for Survey Documentation* (SIDI). The performance criteria from the EUREDIT project were adopted, and the indicators implemented in IDEA are discussed. In the third part of the paper, the documentation of the 23 editing and imputation indicators in and the implementation of the SIDI system are explained. An impressive number of some 50 people have been trained as *quality facilitators* to provide survey meta data including quality indicators for SIDI. The paper end by operational aspects of IDEA, and comments on future work.

The paper is a continuation of work presented for the Work Session in Rome in 1999 and included in reports within the EUREDIT project recently completed with new ideas and improvements.

- From the paper, it seems that the development and implementation of data editing quality indicators in ISTAT is focused on the needs of the production staff. To which extent can the development be used for product quality declarations, and serve the needs of the consumers of statistics?
- The introduction and training of some 50 editing facilitators is an interesting novelty. Which are the prerequisite training and experience for the recruits to the facilitator staff?
- Specialists like the facilitators can be organized in different ways. How has ISTAT organized their facilitators, and how is their relationship to the staff of expert editors?

The Work Session will certainly be looking forward to learning about the results of the future work on indicators from ISTAT announced in the last paragraph of the present paper.

Jeffrey Hoogland and Eugène van der Pijll, Statistics Netherlands are authors of WP. 4: ***EVALUATION OF AUTOMATIC VERSUS MANUAL EDITING OF PRODUCTION STATISTICS 2000 TRADE & TRANSPORT.*** This paper is reporting on automatic and manual editing of the Dutch production statistics data for year 2000. Based on a plausibility indicator, the records are divided into 2 groups, plausible and implausible records. For the investigation, 4162 records of 12 publication cells of both groups in the sectors Trade and Transport were edited manually as well as automatically by SLICE 1. Even when the border for the plausible records is moved to include 80% of the records, the differences between the completely manually and the selectively edited results are insignificant for most publication cells.

This paper on selective editing represents an approach to evaluating realistic editing architectures. The organizers would like to raise the following questions for consideration:

- The authors conclude that adding edits may be an improvement. To which extent do the set of specified instructions issued to the human editors determine the conclusions of a study like the reported?
- One crucial component is the plausibility indicator. The paper demonstrates the effect by moving the borderline between the 2 groups of records. How sensitive are the results for the selection of partial plausibility indicators?
- Is it possible to establish any guidelines for specifying preferable PPIs?
- Another component having a decisive effect on the result is of course the editing program SLICE 1. SLICE is program based on the Fellegi-Holt principle and on manually pres-specified edits. To which extent do the participants of this Work Session believe that other editing methods might have resulted in significantly different conclusions?

3. Contributed papers

This topic attracted 2 contributed papers for which the organizers are very pleased, and therefore would like to comment on as if they had been invited.

The first is ***Impact of Data processing on Unified Enterprise Survey Micro Data – Route to the Final Data Point*** by Fred Hazelton, Statistics Canada. This paper discusses the impact of the data processing on the micro data of the Unified Enterprise Survey including the steps data capture, post data capture correction, edit and imputation, and post edit and imputation correction. The idea pursued was a data point map concept providing a visual presentation of how the data points moved through the four steps.

The graphical approach of the data point map is fascinating, and it might be interesting to hear more about the reception by the survey managers of the viewing the 'big picture'.

- Did the managers accept the method of displaying the impact of the different steps in data point maps?
- How did they eventually use the information the perceived from the graphical maps?

Felix Aparicio-Pérez and Dolores Lorca from our host NSI, INE, have contributed the paper ***Performance of Bootstrap Techniques with Imputed Survey Data.*** The paper reports on the calculated relative bias and RMSE of the variance estimator and the coverage rate of the bootstrap confidence interval compared with previously calculated jack-knife variance estimator, both on imputed survey data. After explaining the method used, a Monte Carlo study based on data from the Industrial Business Survey is described, and the results reported. Even though the results do not point out any winner, the paper adds to our collection of tools which is important for evaluating the quality of surveys with imputed data.

The organizers think this is an important field of study because it is the bases for predicting the data quality of imputation processes, and would like to ask:

- Are similar studies carried out at other NSIs?
- Do there exist any the practical applications demonstrating the use of these techniques?

4. Concluding remarks

It has been a pleasure to read both invited and contributed papers received for topic (i). Different aspects have been presented and discussed in the papers, and a most aspects have been relevant within the announced topic title. There is, however, one aspect which we cannot see any paper has discussed: *The development and use of data*

quality indicators for the users of official statistics, which we hope will be focused on in a future work session.