



Economic and Social Council

Distr.
GENERAL

CES/1999/29
9 June 1999

Original: ENGLISH

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Forty-seventh plenary session
(Neuchâtel, Switzerland, 14-16 June 1999)

REPORT OF THE WORK SESSION ON STATISTICAL DATA EDITING

1. The Work Session on Statistical Data Editing was held in Rome, Italy from 2 to 4 June 1999. It was attended by participants from: Austria, Bosnia and Herzegovina, Canada, Denmark, Finland, France, Georgia, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Netherlands, Poland, Portugal, Romania, Russian Federation, Slovenia, Spain, Sweden, United Kingdom, and the United States. A representative of the Food and Agricultural Organization (FAO) was also present. At the invitation of the secretariat, representatives of the Centre for Sociological Research of Madrid (Spain), Statistical Solutions Ltd. (Ireland) and the University of York (United Kingdom) participated as observers.
 2. The provisional agenda was adopted.
 3. Mr. John Kovar (Canada) was elected Chairperson. Mr. Giulio Barcaroli (Italy) was elected Vice-Chairperson.
 4. The meeting was opened by Mr. Paolo Garonna, Director General of the National Statistical Institute of Italy. In his opening address he highlighted the importance to link the Work Session on Statistical Data Editing with the activities going on in the framework of the Fifth Research Programme of the European Commission.
 5. The following substantive topics were discussed at the meeting:
- GE.99-

- (i) Measuring the impact of editing in various phases of statistical survey processing;
 - (ii) Generalized software packages for statistical data editing, their evaluation;
 - (iii) New methodological and technological developments in statistical data editing.
6. The following participants acted as Discussants: Mr. Ton de Waal (Netherlands) for topic (i); Mr. William Winkler (United States) for topic (ii); and Mr. Leopold Granquist (Sweden) for topic (iii).
7. The discussion was based on papers and demonstrations prepared by Canada, Czech Republic, Denmark, France, Germany, Ireland, Italy, the Netherlands, Romania, Slovenia, Spain, Sweden, United Kingdom, United States.
8. The Work Session recommended that the document "Strategies for improving statistical quality", after its finalisation based on the suggestions made by the Work Session, be reproduced by the secretariat and distributed to interested statistical offices as methodological material.
9. The Work Session recommended that the Conference convene a future Work Session on Statistical Data Editing in 2000/2001 and that the following items be on the agenda:
- (i) Management and evaluation of editing and imputation procedures;
 - (ii) Propagation of knowledge to users;
 - (iii) New techniques and tools for editing/imputation.
10. The following countries expressed interest in contributing papers on these topics: Denmark, Germany, Italy and United Kingdom in topic (i), Canada, France and Italy in topic (ii), and United Kingdom in topic (iii).
11. The United Kingdom offered to host the next Work Session on Statistical Data Editing in the fall of 2000.
12. The participants expressed high appreciation and gratitude to the Italian National Statistical Institute (ISTAT) for hosting this meeting.
13. The main conclusions the participants reached in their discussions are presented in the Annex (in English only).

ANNEX

SUMMARY OF THE MAIN CONCLUSIONS REACHED AT THE MEETING

I. Measuring the impact of editing in various phases of statistical survey processing

1. The discussion focused on the measurement of the quality of data editing in different phases of the statistical production process and different statistical areas. There was general agreement that quality of data editing and imputation is difficult to measure. An effort should be made, however, to develop methods to solve this problem.

2. The Work Session considered the draft methodological material "Strategies for improving statistical quality". The paper provides a theoretical foundation for systematic consideration of the evaluation of the data editing process and proposes several indicators for measuring the impact of editing on data quality. The meeting recommended that upon completion, it should be reproduced as methodological material within the framework of the Conference of European Statisticians and distributed to interested NSIs.

3. The goal of this methodological material is to find the best editing strategy for a given market situation, subject to available architectures (process designs) and resources. The market value of the quality of a statistical product should exceed its cost. The proposed strategy model aims to assist the statistical producer to investigate if a feasible design (editing structure) exists, given the architectures and resources that are available.

4. The relations between data editing and different aspects of data quality were considered. Important factors determining the quality of a statistical product are: product relevance, timeliness, accuracy, accessibility, interpretability and coherence. National and international users need data about quality to evaluate if the supplied statistics are suitable for their needs. The producers need data on quality to analyse alternative production strategies and to allocate resources for improving production. It therefore is important to develop tools supporting data quality control.

5. Quality depends on the kind of survey and the level at which the data are examined. At the highest level of aggregation, the quality of edited and imputed data may be sufficiently high. But at the detailed levels the quality may be quite low.

6. In some cases, quality predictors may be successfully used. The predictors measure the quality of edited and imputed data. It was recommended to use such quality indicators whenever possible. These indicators can be used, for example, if re-edited data are available for a sample of the edited

and imputed data, or if high quality administrative data are available for part of the survey.

7. Considerable research is still needed, however, for the further evaluation of editing procedures. It is necessary to develop a conceptual framework for describing the editing processes. More investigation is required to give answers to the following questions: which data need to be recorded, what is the best way to obtain the required data, can a causal model be derived explaining the relation between data quality and costs, are the computed indicators reliable, etc.

8. It was pointed out that it could be difficult to create general quality indicators that would be valid for different areas of statistics and respond to different users' needs. Some participants stressed the importance of "true" statistical data. Others, who did not agree with the notion of "true" data, highlighted the need to orient towards target values. The observed values depend on how these values were measured (e.g., different questions that are supposed to receive the same answer often lead to different final results).

9. It was also discussed where in the survey process and at what level of detail it is more efficient to obtain information on data quality. Pre-testing questionnaires and pilot studies of surveys were mentioned as one possibility. On the other hand, it might be necessary to collect such information continuously over regular surveys. Concerning the level of detail of the data on the editing process, it is sometimes better to concentrate on summary statistics of the editing process (e.g., frequency of edit actions per question, more probable error causes). It is often more important to examine the changes in the values of indicators than to examine the actual values.

10. Editing processes have to be described in a uniform way, making it possible for a statistical agency to compare the effectiveness of editing between surveys. It is important to store all the computed quality indicators about both data and process quality in order to compare the indicators with each other and their modification over time. It might also be useful to compare the level of quality indicators between similar surveys, or between different surveys on the same population, which use different editing and imputation techniques. The need for relevant metadata was highlighted.

11. An information system for survey documentation (SIDI) that is designed to support quality control was presented. SIDI aims to monitor the survey production process, to document data production and quality control and also to disseminate suitable information on data quality to the end-users. The system manages qualitative information (metadata related to the survey production process) and quantitative information (quality measures) in an integrated way.

12. Concrete examples were presented on measuring the impact of editing and the effectiveness of different editing procedures, e.g. in business surveys,

Labour Cost Survey, CATI and mail surveys. Editing during the data collection/capture and follow-up processes proved to be efficient and provided high quality data. Automatic editing procedures were demonstrated using data from many different sources, including administrative registers.

II. Generalized software packages for statistical data editing, their evaluation

13. The comparative advantages and disadvantages of several widely used generalized data editing and imputation packages were considered. The discussion touched upon the working principles underlying edit and imputation methods, hardware requirements and the usability of the systems in statistical agencies. Special attention was drawn to different criteria that should be considered when evaluating the feasibility of a generalized system for a statistical agency.

14. The following systems were considered: the Generalized Edit and Imputation System (GEIS) and the New Imputation Methodology (NIM) developed by Statistics Canada, the Standard Economic Processing system (StEPS) developed by the U.S. Bureau of the Census, the Agriculture Generalized Imputation and Edit System (AGGIES) developed by the U.S. Department of Agriculture, and the SOLAS software for multiple imputation developed by the Statistical Solutions Ltd. (Ireland). Several contributions dealt with the applications and new developments of the Blaise system, e.g. a module for Statistical Localization, Imputation and Correction of Errors (SLICE) developed by Statistics Netherlands.

15. When considering the methods and software to be used for edit and imputation, statistical offices need to identify the best methods for the existing statistical environment and to find suitable methods for evaluating the available methods.

16. Identification of the best methods depends on the statistical environment and types of surveys carried out in a statistical office. It is not possible to provide a general valid comparison of the systems against each other. Which system performs better depends on several factors, such as the type, complexity and size of the survey, existing computer environment, in-house expertise, cost of the commercial products, etc. Often the systems with a more complete functionality are not user-friendly and require additional training and auxiliary skills (e.g., knowledge of SAS, specific programming languages etc.). The choice can depend also on the final goal of the edit/imputation process, whether it is to reproduce the original true data or to improve a series of aggregate estimates.

17. Apart from the associated technical aspects, there are also human resource problems to consider when evaluating the feasibility of introducing a generalized software package. The skills of available individuals need to be determined, training needs identified, and whether individuals in the agency are willing to accept the new methods.

18. It was agreed that the use of generalized systems can improve efficiency in statistical agencies. This is achieved through the commonality of interfaces, similarity of training for personnel, and consistency of techniques across surveys. Often it is not efficient for the statistical offices to develop their own tailor-made software. However, the cost- and time-savings from use of the new system and the benefits to the statistical agency need to be analyzed.

19. Modularity and integration were noted to be desirable features of any generalized system. It was suggested that the generalized systems be built in self-contained modules that are easily embeddable in commonly used environments (e.g., SAS). In this way, generalized modules can be combined with customized ones to produce an integrated customized system.

20. The importance of promoting internationally the use of generalized editing systems was emphasized. This can be done through creating international user-groups, organizing conferences for users of a specific software, etc. A good example of this is the Blaise software which is used by several statistical offices. Statistics Netherlands organizes conferences regularly for Blaise users that provide user feedback for further Blaise development. The next conference for Blaise users will take place in May 2000 in Cork, Ireland.

21. Some participants expressed the opinion that it would be desirable to create and maintain a "knowledge base" on standard data editing software. The components of such a knowledge base could also be the glossary of data editing terms, bibliography and frequently asked questions about the capabilities of general software packages.

22. In order to obtain comparable evaluation data for different edit and imputation systems, these have to be tested on the same data. Three different data sets are needed: *raw data* (with errors), *true data* (without errors) and *clean data* (result of the application of the edit and imputation procedure). A generalized Editing System Standard Evaluation (ESSE) software was demonstrated that can be used for this purpose. The software is based on a simulation approach and creates the required data sets allowing for the evaluation of the quality of edit and imputation procedures. Metrics for comparing the effects of different edit/imputation systems were proposed. The described method is universal and should be straightforward for other agencies to implement.

III. New methodological and technological developments in statistical data editing

23. The Work Session discussed the "new editing paradigm" and its implications on the methodological development of data editing. It is focused on identifying and collecting data on errors and their causes in order to provide a basis for a continuous improvement of the whole survey process.

This role of editing is expected to significantly improve data quality and reduce the cost of editing.

24. Related issues discussed were the collection of data on causes of error, the need for and the requirements of a Process Data Subsystem, and standardisation of the editing process by developing and implementing Current Best Methods. It was stressed that editing will improve data quality only if the editing process statistics are used to improve aspects of the entire survey process, not only the editing process. The importance of carefully designing and pre-testing the questionnaires was highlighted. It was pointed out that often the information on errors coming from the editing process is not taken into account in the questionnaire design.

25. The challenges of successful interaction between methodologists, programmers and users were highlighted in order to assure successful implementation and continued use of the new proposed methodologies. To this end, it was suggested that methodologists/statisticians should acquire sufficient programming background and vice versa, to enable better communication between the various groups within the agency and between the agencies.

26. An approach for combining macroediting and selective editing to detect influential outliers in sample surveys was discussed. This method uses aggregate data from previous surveys in order to locate suspicious variations of survey variables. Thus the selective editing approach can be extended to cross-sectional surveys where information on the sampled units is available only for one survey. This is a typical situation in many households and business surveys.

27. Another solution for outlier detection was demonstrated making use of ARIMA models for time-series' data analysis. The method allows the use of the whole set of data from previous surveys in an optimal way. If the observed data differ considerably from the ARIMA forecast, the data can be erroneous. The method enables to use probabilistic data editing, and to take into account the different variability of the economic sectors, products, etc.

28. The use of graphical editing techniques was highlighted as a powerful and efficient tool to examine large amounts of data. It was viewed as a useful technique for developing inlier and outlier edits. The method enables the data to be edited interactively, and the effect of edits on graphs, distributions, etc. to be noted immediately.

29. The discussion examined the new developments for data imputation using the New Imputation Methodology (NIM), a model-based imputation and neural networks. The use of these technologies is often being considered for the forthcoming Census 2000.

30. The meeting considered current research trends in improving the Fellegi-Holt (FH) systems of editing. Different solutions are being developed to minimize the workload needed to generate implicit edits for large complicated

surveys. The participants concluded that further research is required especially in the following areas: (i) the algorithms increasing the speed of the software, (ii) optimization of the conversion of data structures enabling the use of FH systems, and (iii) optimizing the use of FH systems to discrete and continuous data simultaneously and the conversion of discrete data into continuous data allowing the use of the FH methodology.