



**Economic and Social
Council**

Distr.
GENERAL

CES/1998/3
30 October 1997

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Forty-sixth plenary session
(Paris, France, 18-20 May 1998)

OCTOBER 1997 WORK SESSION ON STATISTICAL DATA EDITING

Note prepared by the Secretariat

Introduction

1. The meeting was held from 14 to 17 October 1997 in Prague, Czech Republic. It was hosted by the Czech Statistical Office. It was attended by participants from Austria, Canada, Croatia, Czech Republic, Denmark, France, Hungary, Ireland, Israel, Italy, Latvia, Lithuania, the Netherlands, Norway, Poland, Slovakia, Slovenia, Spain, Sweden, United Kingdom, United States and Eurostat.
2. The welcoming address was delivered by Mr. Sujan, Vice-President of the Czech Statistical Office.
3. The provisional agenda was adopted.

4. The following substantive topics were discussed at the meeting:
 - (a) Designing a complete editing strategy;
 - (b) Organizational issues related to the implementation of data editing methods.
 - (c) Evaluation and monitoring of the editing process;
 - (d) National reports;
 - (e) Panel discussion on evaluation of data editing.

5. The discussion was based on papers and demonstrations prepared by Austria, Canada, Czech Republic, Denmark, France, Israel, Italy, Lithuania, the Netherlands, Norway, Slovakia, Slovenia, Spain, Sweden, United Kingdom, United States, Eurostat and UN/ECE Secretariat.

6. Mr. John Kovar (Canada) was elected Chair. Mr. Giulio Barcaroli (Italy), Leopold Granquist (Sweden) and Bill Winkler (USA) were elected Vice-Chairs.

7. The Work Session recommended that the Conference convene a future Work Session on statistical data editing in 1998/99 and that the following items should be on the agenda:
 - (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;

8. The other conclusions which the participants reached at the meeting on the substantive items of the agenda are summarised (in English only) in the Annex.

ANNEX

Other conclusions reached at the work session on the substantive items of the agenda

A. Designing a complete editing strategy

1. Presentations considered under this agenda item demonstrated a general trend to combine different data editing and imputation methods with the common aim to achieve savings in time and cost while improving the overall data quality. The design of data editing strategies in various statistical applications was discussed (e.g. in demographics, agriculture, business surveys, enterprise and foreign trade statistics).

2. The discussion showed that the aim of editing is not only to detect and correct errors but also to monitor the quality of the different phases of survey processing and to give indication for removing the causes of errors. In order to introduce this approach in practice, two main issues were discussed: **methodological aspects** and **implementation aspects**.

3. Concerning the **methodological** viewpoint, the Work Session noted that many edit procedures involve outlier detection and correction, and can be efficient in improving the quality of aggregates. However, there can also be erroneous data inside the distribution of variables (so-called inliers). Those can be caused by systematic errors or can arise when data are merged from different sources. In the case of many inliers, statistical analysis of microdata can be seriously affected. One way to handle this problem could be the use of auxiliary data that allows to apply outlier-detection methods.

4. The Work Session also highlighted the use of graphical methods because of their capabilities to visualise the editing process.

5. Furthermore, it was discussed how to select the relevant editing and imputation methods, e.g. to what extent to use the automatic imputation, how to balance between the probabilistic and deterministic approach. Those methods are, however, used for different kinds of statistical data (qualitative and quantitative data). Therefore, the analysis of comparable advantages of both systems is often not available. One system, NIM (Canada), does provide the opportunity to automatically impute qualitative and quantitative data simultaneously.

6. An opinion was expressed that although data editing methods and techniques have progressed, a certain slow-down in their improvement can be observed. This can be caused by a lack of clear data editing concepts and definitions, as well as their evaluation methods. In particular, without quantification of the evaluation process, it is not possible to rank alternative methods or to optimize the quality/cost trade-off for the editing process.

7. As for the **implementation** viewpoint, the importance of generalized data editing systems was emphasized. It allows standardization and relative independence of subject-matter statisticians from IT specialists.

8. The Work Session noted that the editing systems used have developed towards more generality and re-usability. They are designed to integrate many processing needs, including data collection, interactive editing, survey management,

tabulation and meta-data management. The introduction of interactive editing makes it possible to place an edit check in the processing step where it is most beneficial.

9. In this respect, the importance of integration of one generalized data editing system with other editing tools (e.g. Blaise, SPEER, graphical packages) and/or integration of data editing systems with other statistical software packages (e.g. SAS) was highlighted. Experiences with developing general-purpose editing systems, like CherryPi (the Netherlands), "Plain Vanilla" (US Bureau of the Census), and the generalized edit system used in the US National Agriculture Statistics Service were reported.

10. Some participants mentioned the importance of linking survey sampling with data editing. Coordination between samples in order to avoid double contacts and reduce response burden is needed. If the results received from sample editing are affecting fundamental variables those would require manual revision.

11. Furthermore, it was mentioned that restructurings of enterprises can have a strong effect on aggregate data. Therefore keeping track of the aggregations' history has to be taken into account when implementing data editing procedures.

12. The growing need for faster and more accurate statistical information are often in conflict with the developer's possibility to complete and test the survey application in time. Some possibilities of redefining the survey after implementation were presented at the Work Session.

13. It was also mentioned that use of historical data can contribute to formulate a better data editing strategy. Data warehousing technology was mentioned as a possible tool for how to solve this problem. Another use of analysis of time series for designing a data editing strategy was demonstrated using ARIMA models.

14. It was also pointed out that one way to implement an efficient data editing system in a statistical office is to develop and use a "Current Best Methods" document for vital processes in the production cycle. This document should give guidelines for designing efficient editing processes and include the checklist of factors that have to be addressed in the course of work. The document should be maintained continuously to ensure that best practices are followed.

B. Organizational issues related to the implementation of data editing methods

15. The contributions under this topic highlighted the importance of effective strategies for the implementation of new data editing techniques. The suggested solutions have to be compatible with the IT strategy of the statistical office. Analysis of risk factors should be an integral part of this strategy that should be transparent to both IT specialists and subject-matter statisticians.

16. The Work Session drew attention especially to the question how to introduce new theoretical solutions of data editing into practice. The importance of evaluating the operational feasibility of new methods was underlined. The need for trials in an operational environment was highlighted. Many critical factors, like how voluminous data sets should be processed, how time consuming the process could be, how much storage capacities would be needed etc., should be considered.

17. It was pointed out that in many cases, simply showing the advantages of new techniques is not sufficient to convince statisticians to use them. It must go hand-in-hand with easy-to-use practical applications, and organizing relevant training. Such training should be based to the maximum extent possible on practical examples with real data. It was also mentioned that it would be desirable to establish user-groups working on similar problems and to rotate experts on new methods.

18. The Work Session considered some feedback with implementation of winsorisation techniques (reducing the effect of outlying values on survey estimates). Experience showed that although the users understood the principles of this technique, the implementation was in some cases not very successful and the method could be misused because the offices concerned have insufficient practical experiences. It was also mentioned that a similar situation can arise with other new (especially visual) tools.

19. Another important issue raised in the discussion was the need for systematic management support to ensure the implementation of new data editing methods. It was pointed out that it is essential to empower editors with relevant authority to involve them more in this process.

20. Some participants mentioned that users sometimes resist introducing generalized systems and they discussed how to overcome this obstacle. The opinion was expressed that the best way could be to demonstrate the advantages of those systems using practical examples. Furthermore, the additional work needed for changing to the new system should be taken into consideration. The ideal situation for implementing new methods seems to be to completely redesign the survey.

C. Evaluation and monitoring of the editing process

21. The Work Session considered the need for evaluation of different layers of data editing. Besides evaluating the whole editing process it is important to monitor its individual phases and evaluate specific editing procedures. The evaluation of editing should give answers to many related questions, such as whether the use of a certain technique is justified; What is the overall effect of editing and imputation on the estimates; Is editing cost-efficient; Is it feasible operationally; and What is its impact on data quality?

22. The participants recognized that editing-phase monitoring, evaluation, and cost reduction, while improving the quality of processed data, are crucial survey performance measures. Such measures enable managers to allocate resources more efficiently in both the broad/long-term sense and more specific/short-term sense. They provide information that can prevent data errors in future periods or surveys. To be able to develop efficient performance measures, the whole process should allow the identification of all potential errors and correction processes.

23. It is necessary to evaluate performance at the macro-level as well as at the micro-level of data. Both at macro- and micro-level the summaries should reflect the frequency of changes, the extent of changes, and the distribution of changes by edit type etc. This information should be maintained and analysed over time for periodic surveys. Calculating the measures for groups of respondents and reporting categories provides feedback on the editing rules and the parameters, as well as the respondent type and data analyst.

24. The Work Session discussed the required theoretical background for establishing the system of indicators for evaluation of data editing methods. A formalized approach was presented by Italy. It represents a standard methodology that can be applied in a variety of situations. A set of indicators is defined that allows to assess the capability of the editing procedure to detect errors and to correct them by imputing the true values.

25. The problem of evaluating a given editing procedure can be solved using two different approaches: by re-doing more carefully all or some steps of the process, or by simulating sets of "raw" and "true" data. In theory, the first method is preferable as it better reflects the real situation. However, in practice, it is not always possible because of the high costs for the statistical office and the high response burden. The second method strongly depends on models that are chosen to generate raw and true data.

26. Another theoretical approach for predicting the quality of editing and imputation was presented by Sweden. It is based on a probabilistic approach, as the quality of statistical estimate is determined by its deviation from the target value which is usually unknown.

27. Some countries reported that an efficient method to evaluate data editing could be the use of administrative records for this purpose. This method can be used, for example, in evaluating population coverage or response quality in censuses. Using administrative registers, requires the evaluation of not only the editing process and results, but also the quality of the register itself.

28. The Work Session considered the use of neural networks for imputation. The evaluation of using neural networks for the United Kingdom census variables' imputation has shown that in some cases the variable distribution was not preserved. Although the hot-deck method was not superior to neural networks concerning statistical evaluation, it was the best method operationally. The need to better understand the workings of the neural networks "black box" was emphasized. The United Kingdom volunteered to continue investigating the use of neural networks for imputation.

29. In order to implement the evaluation of data editing techniques, the necessary tools have to be incorporated into all data editing processes, and into survey design. In many cases, carrying out the evaluation requires proper documentation of all processes. The documentation itself can serve as a good link between survey statisticians and IT specialists.

D. Panel discussion on the evaluation of data editing

30. The panel discussion concentrated on possibilities of how to measure and evaluate the efficiency of data editing methods and techniques. The participants agreed that data editing has to serve three goals: to gather information about the quality of the data, to form the basis for future improvement of the survey process, and to clean up the data. It was generally felt that in implementing statistical data editing techniques, most emphasis is put on achieving the third goal, while gathering information about the data quality and further improving the survey process based on this information should be the primary reasons for editing.

31. The discussion revealed that there is a lack of methodological materials concerning the evaluation of statistical data editing. It was highlighted that

internationally prepared recommendations, such as a set of standard practices, would be highly appreciated.

32. The Work Session agreed that methodological recommendations on how to construct indicators for measuring the data editing process could contribute significantly to improve the evaluation of the impact of data editing methods in statistical production. It was pointed out that standardization of evaluation methods can provide an objective basis for comparing different data editing applications. There was general understanding that the whole survey processing method should be monitored also. Some participants mentioned the need for benchmarks of data editing indicators allowing offices to better assess the quality of a specific data editing process (Italy, Sweden, US Department of Energy and US National Agricultural Statistics Service volunteered to prepare papers on this issue).

33. Concerning the subject to be evaluated, the discussion highlighted automated data editing systems. It was recommended to compare the efficiency of some of the most widely used data editing software packages on different statistical surveys (Canada, Italy and Eurostat volunteered to coordinate this work).

34. Furthermore, the use of administrative registers for evaluation of data editing was identified as one of the subjects requiring further studies, and that many countries would probably be interested in the result (Israel, Denmark and the Netherlands volunteered to contribute papers on this subject).

35. The Work Session discussed the impact of the reduction of resources used for data editing (it is estimated that 20-40 % of overall costs are devoted to this). It was pointed out that there are no objective criteria for estimating this impact, and it was recommended to study this problem in more detail. It was also mentioned that it is very difficult to estimate the impact of the findings from the data editing process on the improvement of the whole survey processing operation (Sweden agreed to prepare a paper on this issue).

36. The role of data editing in quality assurance process was discussed. It was pointed out that quantitative measures are required that would allow to document data quality for users. In this respect, the need for measuring the impact of non-sampling errors on the estimates was also mentioned. Eurostat informed the meeting about its projects on the quality of business statistics, and proposed to continue informing the Work Session about the progress it makes in this area in the future.

37. The Work Session considered the need to improve the "Glossary of terms used in statistical data editing". There was a joint understanding that such material could be very useful for statistical practice. Eurostat expressed willingness to coordinate this activity in cooperation with the UN/ECE Secretariat. Sweden and the United Kingdom agreed to cooperate in this work. The representative of Eurostat said that Eurostat could link this initiative with the preparation of its methodological manual on statistical editing.

E. National reports

38. The Work Session was informed about the progress in national data editing projects via written reports. Denmark and the Netherlands made an oral presentation. Based on the reported experiences, the Work Session discussed the

effect of imputation on second-order estimates (variance, correlation). It was stressed that the variance component due to imputation procedure should not be ignored.

F. Future work

39. The UN/ECE secretariat informed the meeting that in order to increase the efficiency of work in the framework of the programme of work of the Conference of European Statisticians there is a need to decrease the frequency of meetings from 12 to 18 months.

40. There was a general understanding that it would be highly desirable to continue activities under this project between the meetings. Participants were encouraged to make greater use of Internet facilities, and particularly the discussion group on Internet set up by Statistics Netherlands. The e-mail address of that group is **statedit@krypton.vb.cbs.nl**. All messages sent to this address will be automatically forwarded to the subscribers. It is possible to subscribe by sending a message: **subscribe statedit** to the address **statedit-request@krypton.vb.cbs.nl**. The secretariat informed the participants that the UN/ECE Statistical Division's Web page is **<http://www.unece.org/stats>**.

41. The Work Session considered the possibility of maintaining the mutual exchange of information concerning the progress achieved in national data editing projects. In the light of the fact that there will be no special agenda item dealing with national reports, it was recommended to continue reporting in written form and to seek the possibility to give the floor to those countries who would like to report orally in the framework of the next Work Session's work programme provided that there was sufficient time and this does not detract from the main agenda items of the meeting.

42. Taking into account that the duration of the next Work Session will be 3 days instead of 4, the opinion was expressed that the discussion during the meeting should concentrate more attention on selected items. The Work Session recommended to hold a future Work Session on statistical data editing in May/June 1999.

G. Other business

43. The Czech Statistical Office presented an interesting demonstration on the software system ProjektMan that supports the design and implementation of statistical surveys.

44. The Work Session expressed its appreciation to Chair and Vice-Chairs and to all authors of contributions and the demonstration for their excellent work.

45. The Work Session expressed its gratitude to the Czech Republic for having hosted this meeting and for the excellent organization and working atmosphere.