**UNITED NATIONS**

**E**

**Economic and Social Council**

Distr.
GENERAL

CES/2001/27
19 January 2001

ORIGINAL : ENGLISH

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Forty-ninth plenary session
(Geneva, 11-13 June 2001)

# REPORT OF THE OCTOBER 2000 WORK SESSION ON STATISTICAL DATA EDITING

Note prepared by the secretariat

1. The Work Session on Statistical Data Editing was held in Cardiff, United Kingdom from 18 to 20 October 2000. It was attended by participants from: Austria, Canada, Denmark, Finland, France, Germany, Hungary, Ireland, Israel, Italy, Netherlands, Poland, Russian Federation, Spain, Sweden, Switzerland, The former Yugoslav Republic of Macedonia, United Kingdom, and the United States. The European Commission was represented by Eurostat. At the invitation of the secretariat, a representative of the University of Southampton (United Kingdom) participated as an observer.

2. The provisional agenda was adopted.

3. Mr. John Kovar (Canada) was elected Chairperson. Mr. Paul Smith (United Kingdom) was elected Vice-Chairperson.

4. The meeting was opened by Mrs. Susan Linacre, Director, Methods and Quality Directorate, Office for National Statistics, United Kingdom.

5. The following substantive topics were discussed at the meeting:

(i) Management and evaluation of editing and imputation procedures;
(ii) Propagation of knowledge to users;
(iii) New techniques and tools for editing imputation;
(iv) Evaluation of efficiency of statistical data editing.

6. The following participants acted as Discussants: Mr. Leopold Granquist (Sweden) for topic (i); Mr. Giulio Barcaroli (Italy) and Mr. Claude Poirier (Canada) for topic (ii); Mr. William Winkler (United States) for topic (iii); and Mr. John Kovar for topic (iv).

7. The discussion was based on papers and demonstrations prepared by Canada, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Russian Federation, Spain, Sweden, Switzerland, United Kingdom, United States and Eurostat.

8. The Work Session recommended that Canada create a knowledge base on statistical data editing on the Internet. It requested the secretariat to create a link from the UN/ECE web-site to this knowledge base. Furthermore, the Work Session requested the secretariat to explore the possibility of translating selected methodological materials on statistical data editing prepared within the framework of the Work Session into French and Russian.

9. The Work Session recommended that the Conference convene a future Work Session on Statistical Data Editing in May 2002. Finland offered to host this meeting. The following items will be on the agenda:

(i) Planning and management of statistical data editing (papers to be contributed by Canada, Germany and Switzerland);
(ii) Measuring and evaluating data editing quality (papers to be contributed by Canada, Germany, Italy, Sweden, Switzerland and United Kingdom);
(iii) Editing of administrative data (papers to be contributed by France, Israel and Italy);
(iv) Impact of new technologies on statistical data editing (papers to be contributed by Austria, Finland, Poland, Switzerland and Eurostat).

10. Furthermore, the Work Session agreed to continue work on the preparation of good practices on the evaluation of efficiency of data editing. It was agreed that Canada, France, Germany and Spain will further develop their contributions for the next Work Session.

11. The participants expressed their great appreciation and gratitude to the Office for National Statistics of the United Kingdom for hosting this meeting.

12. The more detailed summary of the discussion that took place at the meeting on the four substantive agenda items is presented in the Annex (English only).

**ANNEX**

**SUMMARY OF THE MAIN CONCLUSIONS REACHED AT THE MEETING**

## A. Management and evaluation of editing and imputation procedures

1. The meeting considered the management and evaluation of editing and imputation from the viewpoint of how to improve the whole survey production process. Editing should be effective, which would facilitate measuring the impact of editing on data quality. Furthermore, an important aim of editing is to identify error sources so that errors can be prevented for future surveys. New developments that need to be explored are the editing of data from administrative sources and electronic data reporting both by businesses as well as individuals.

2. Data editing may be rendered inefficient because of poor planning. Planning of data editing is greatly influenced by survey specific aspects. There was general agreement that the planning of data editing needs a more systematic approach in statistical offices. Subject matter specialists should gain knowledge about the main tasks and relevant conditions of the data editing to be planned. In addition, the planning activities should take into account the new developments in data collection from administrative registers and directly from enterprise information systems.

3. The description of risks should highlight the conditions which may cause the planned data editing to fail. A need for pilot studies was highlighted. They should provide information about: error descriptions; error frequencies; effort needed for the correction of errors; recommendations concerning tested edits; and information about the usefulness of error descriptions, instructions and work manuals.

4. A structural plan of data editing was presented. As the planning of data editing consists of many different activities, the assessment of detailed plans would be facilitated by benchmarks, which should enable comparisons between surveys. Various measures to evaluate the effect of editing were proposed and discussed, such as bias in the whole population and in groups, average effect of data editing on the whole population and on groups, effect of data editing for each individual, etc.

5. User demands for data quality should be collected as well. With regard to the planning of data editing the focus may be put on the commonly used quality criteria, such as timeliness, accuracy and clarity. Clarity is an important criterion for the planning of data editing because data editing must in many

cases provide information for informative and user friendly quality indicators. Continuous contact with users should be maintained and relevant measures for data quality used in interaction with users.

6. The discussion explored different indicators for the evaluation of editing based on input- and output-oriented approaches with the aim of finding out where to improve the survey process. It could be efficient to make a distinction between two kinds of (quality) indicators: a set of indicators to be used within the office, and another to be used as output to external users. The difficulty of evaluating data quality when the "true" values are not known was considered.

7. At present, there is often not much possibility in changing data editing because of the tight interaction with other phases of statistical data collection, production, archiving, etc. The aim is to make editing systems flexible enough to allow changes without affecting other phases of data production. Furthermore, implementation of new editing methods takes a long time which emphasises once more the need for proper planning.

8. A EUREDIT project funded by Eurostat was presented. Its objectives are the development of new methods for automatic editing and imputation and the comparative evaluation of the newly developed methods together with established methods. The main focus is on imputation for the types of data collected by national statistical agencies.

9. An important aspect of the assessment of an editing and imputation method is its operational efficiency, i.e. the ease with which it can be implemented, maintained and applied to data sets. The following criteria can be used: resources needed to implement and maintain the methods; expertise needed for implementation; hardware and software requirements; data limitations; the kind of feedback that it provides to "tune" the process to improve its efficiency; resources to modify the method; and its transparency.

10. The opinion was expressed that it would be desirable to concentrate on a small number of indicators which provide most of the information about data quality. The use of principal components analysis to find the measures which best capture the underlying variation in the data quality was proposed.

11. Statistical surveillance and data evaluation must include a standard by which the particular process of operation is measured. Such standards require estimates of error rates that are derived from quality control systems design. The quality control programme should fulfil the following functions: monitoring data operations; evaluation of the survey (assuring an acceptable quality level; on-going review to ensure that the measurement equipment and processes are under control); reviewing statistical reports to ensure the adequacy of these; conducting experimental studies to obtain estimates of the components of measurement error and costs; researching quality control systems to improve their efficiency and effectiveness; and training of statisticians.

## B. Evaluation of efficiency of statistical data editing

12. The evaluation of efficiency should take into account all aims of editing, not focusing purely on correcting erroneous data. The information on the editing process can play an important role in the improvement of data quality, feedback to the data collection and feed forward to the data processing and dissemination process, reduction of response burden, measuring what has been done in editing and its effect on data etc.

13. The concepts and framework of the methodological material "Evaluation of efficiency on statistical data editing: General framework" were considered useful. The proposed method for evaluation is based on a market perspective. However, it might be difficult to establish a market value of statistical information in a monopolistic situation where the statistical office is the only producer of official statistics or when the statistical products are targeted towards one specific user. The situation concerning the statistical market in countries can differ: in some countries, statistical data cannot be considered as a market commodity; in others, the development is towards everyone (including government agencies) having to pay for the data, thus helping to determine the market value of the data.

14. User demands on quality of statistics differ. While some sophisticated users request information on quality and how the quality can affect their interpretation of data, others do not need to be provided with the full documentation on all different aspects of quality. Quality can be used as a trademark and a marketing tool for statistical output. It is, however, essential to ask users about their requirements for quality as it is a multidimensional phenomenon. An alternative way of evaluating quality could be, for example, user auditing.

15. One of the problems with measuring the quality of statistical data is that statistical surveys do not present one but many results and their quality cannot be measured with one indicator. There is a need to develop statistical methods for calculating the accuracy of the statistical products.

16. Continuous international cooperation is required to overcome the problems with the evaluation of efficiency of data editing since it is not possible to take the necessary research steps at the national level. The evaluation of editing has to be incorporated into the evaluation of the data quality in general. Further research has to be carried out on the description of the editing process, ways to minimise over-editing, synthetic data to evaluate the editing methods, international knowledge base, alternative ways of evaluating data editing and quality which are not based on a market perspective.

17. Several good practices in statistical offices on how to measure quality of statistical products already exist. In some countries (e.g. Canada and Sweden), special groups are set up to work on quality (quality secretariats). It is not possible to develop standards and guidelines that would be useful for all statistical offices, for all kinds of surveys and all products. However, it would be desirable to exchange information on good experiences and to distribute the documentation of good practices. Such practical experiences cannot be replaced by research or academic work. A need for further international work in

this area was emphasised. A good way to exchange experiences could be to set up a web-site for this purpose.

## C. Propagation of knowledge to users

18. The propagation of knowledge to users was considered at three major levels: between statistical offices on the international arena, within statistical offices, and between the statistical offices and their clients. Users can be considered to be individuals who are in touch with statistical products, whether it is in their production or in their analysis. This includes survey methodologists, survey statisticians, survey managers and end-users.

19. The prototype knowledge base on data editing methods and techniques proposed by Statistics Canada addressed the need for an exchange of knowledge and experience in this domain internationally. The prototype is composed of four different items: the glossary, the evaluation of existing systems, experiences, and technical papers.

20. An important part of the knowledge base is sharing experiences on specific techniques and products. User manuals do not give indications on how difficult and expensive it is to implement the applications. Both good and bad experiences should be reported. Information could be included concerning the expertise needed for implementation, money and human investment required, number of employees involved in developing the applications, required training, sources and volume of data that the system can handle, changes that had to be done to other systems, methods that had to be modified because of the available functionality, turn-around time, difficulties encountered and comparison with former systems.

21. It was requested to make a link to the knowledge base from the UN/ECE web-site. The importance of permanent maintenance of such a knowledge base was highlighted. It was proposed to create an editorial committee responsible for content and organisation of the knowledge base. Canada and United Kingdom volunteered to be members of the committee. It was also pointed out that it would be desirable to make such material available in French and Russian languages in addition to English. The UN/ECE secretariat was requested to explore the possibility of translating key documents into French and Russian. In this respect, bilingual contribution submitted in English and French are very appreciated.

22. The propagation of knowledge within the statistical office was analysed from the viewpoint of information flows among internal users of editing and imputation methods and researchers/developers. The analysis was based on a functional model, the "data editing method life cycle", as a conceptual framework.

23. There is something to gain from having a good mechanism for information exchange. It ensures that everybody understands concepts, it increases the users' interest, and enables brainstorming discussions on principles, techniques, methods and systems. Some offices reported difficulties

encountered when the information flow within the office was insufficient: it makes it difficult to implement new methods, to promote the use of best practices within the office, there is no overview about the efficiency of the methods, etc. The transfer of knowledge on data editing in a user-friendly and easily understandable format should be better supported by the top management.

24. Exploring the propagation of knowledge between producers and users of statistics was based on the assumption that statistics become commercial commodities. Although in many countries there is not yet a developed statistical market at the national level, there can be much competition in producing statistics on international and regional levels. As such, the determination of the price of a single statistical piece of information could be based on the demand related to its accuracy, and on the offer related to production costs. However, other dimensions of quality, like timeliness, can be more important for the user than accuracy. Many participants pointed out that, in their experience, few users ask for accuracy measures.

25. It is important to know the needs of the end users for metadata about the quality of statistics and statistical data editing, and how a statistical producer should provide these metadata. Producers should try to get feedback from users on the information about product accuracy and price setting in order to adjust adequately to needs. Research is needed on which data is needed by end-users and how to collect the data, in which form the metadata about accuracy should be disseminated. A marketing policy for providing metadata about the quality of the products and advice on how this information should be interpreted could be useful.

26. It was pointed out that in a real market situation, quality is assessed by the user rather than by the producer who might not be willing to inform users about bad quality. The official producers of statistics could provide information on the accuracy of their product by a kind of code of conduct rather than for economic reasons or for achieving competitiveness on the market.

27. A possible development direction for the future could be to set quality standards for statistical products. This could be done by the national statistical offices based on ISO quality standards. On the other hand, it was pointed out that when a statistical office has a good reputation for high quality products it might not need an external certification. The opinion was expressed that future close cooperation with Eurostat's Working Group on Statistical Data Quality would be highly desirable.

D. New techniques and tools for editing and imputation

28. New techniques for automatic editing, time series analysis, and recent developments in software were considered. New developments are often focused on efficient management of the whole data editing process and combining the existing techniques rather than developing new methods. An ideal edit strategy could be a combination of selective editing, automatic editing and (graphical) macro-editing. Their combined use can be an efficient way to clean data so that the data quality is maintained, while the necessary resources are reduced and the timeliness of releasing statistical data is improved.

29. Several new techniques and improvements on the existing ones are aimed at making the algorithms substantially faster, easier to understand and implement, being able to treat both quantitative and qualitative data, and developing more generic implementation that could be used in different statistical offices and for different types of data. Recent developments of automatic editing in Canada, Denmark and the Netherlands were considered in this respect.

30. The development of methods for the use of time series analysis in data editing was presented by Spain. The theoretical framework for editing is based on the sequences of variables using a time series model. Thus, the different behaviour in production activity in different months of the year, or days-of-the-week composition and holiday periods has to be taken into account. Another new imputation method presented by Russia allows to select estimates for each imputed value where entropy is used as a criterion of the regression modelling. The possibility to choose the imputed values results in a higher probability for the imputed record to pass all edits and the imputed values have more realistic standard errors.

31. An overview of Eurostat research projects concerning statistical data editing was given. There are two projects directly researching editing and imputation methodology: automatic imputation software for business surveys and population censuses (AUTIMP) and the development and evaluation of new methods for editing and imputation (EUREDIT). In addition, there are other projects indirectly related to data editing: new methods and software for time series analysis (TRAMO), and a project that has as its objective the generation of validation rules in the process of data exchange. There are also some projects investigating the effects of changes in survey methodology on data quality (CHINTEX), and developing an expert system computer package for data analysis including editing and imputation methods (X-STATIS).

32. Information was also given on Eurostat's upcoming activities in the area of information exchange and dissemination of best practices. A European Statistical Laboratory (ESL) is being set up together with the joint research centre ISPRA, a help desk for time series analysis in official statistics (TSAOS) is available on Internet, a software demonstration centre (SODECE) is being set up. Eurostat is also preparing two conferences: ETK (Exchange of Technology and Know-how) and NTTS (New Techniques and Technologies for Statistics) which will take place 18-23 June 2001 in Crete.

33. A general difficulty in maintaining the software applications developed in statistical offices was pointed out. It is mostly a management problem and the question of keeping the required human resources and expertise within the office. The software maintenance often depends on a few key persons who have been involved in the development of the software. International cooperation is needed to share the costs of maintenance and to make the softwares usable in different national environments.

34. To ensure the implementation of new techniques requires careful management. Systematic planning, testing, pilot studies, implementation first in selected areas and then spreading the knowledge and good experience can help in this respect. United Kingdom and the U.S. National Agricultural Statistics Service (NASS) presented new approaches to management of the data editing. The systems currently used or various survey programmes differ and lack the integration of the editing, imputation and analysis

models. The objective is to develop systems and procedures that will result in less manual editing, increased interactive editing and analysis capabilities, a more streamlined data analysis process, while maintaining or improving the data quality standards.

35. The new developments in data collection (electronic questionnaires, EDI, data collection via Internet, use of administrative registers) are expected to have a significant influence on editing techniques in the coming years. Electronic questionnaires can replace ordinary paper questionnaires and in some circumstances data can be collected directly from enterprises information systems. However, incentives are needed for users to promote electronic reporting, otherwise all the benefits from this action are on the data collector's side. A lot of work in this area is required in future concerning e.g. edits on the electronic questionnaire, improving the design of electronic questionnaires, technical capabilities and requirements to incorporate editing in Internet data collection, design of the whole process, etc. Internet allows the possibility for immediate feedback and monitoring on how people are filling in the electronic questionnaires. Usability testing is also very useful in this respect.

36. An administrative policy at government level is needed to ensure that electronic infrastructure in the public sector is open and secure (secure communications services for public services and EID cards). However, an NSI is dependent on major authorities such as the tax board and national social insurance board acting as the driving forces in development.