UNITED
NATIONS

E

Economic and Social
Council

Distr.
GENERAL

CES/2002/24
3 June 2002

Original : ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Fiftieth plenary session
(Paris, 10-12 June 2002)

REPORT OF THE MAY 2002 WORK SESSION ON STATISTICAL DATA EDITING

1. The Work Session on Statistical Data Editing was held in Helsinki, Finland from 27 to 29 May 2002. It was attended by participants from: Austria, Belgium, Bulgaria, Canada, Czech Republic, Estonia, Finland, France, Germany, Israel, Italy, Netherlands, Norway, Poland, Russian Federation, Slovenia, Spain, Sweden, United Kingdom, and the United States. A representative from Australia participated under Article 11 of the Terms of Reference of the Economic Commission for Europe. A representative of the University of Southampton (United Kingdom) participated as an observer at the invitation of the Secretariat.

2. The provisional agenda was adopted.

3. Mr. John Kovar (Canada) acted as Chairperson.

4. The meeting was opened by Ms. Jeskanen-Sundström, Director General, Statistics Finland. Her address highlighted the importance of statistical data editing in conducting statistical surveys as well as when deriving data from administrative registers and records.

5. The following substantive topics were discussed at the meeting:
   (i) Planning and management of statistical data editing;
   (ii) Measuring and evaluating data editing quality;
   (iii) Editing of administrative data;
   (iv) Impact of new technologies on statistical data editing.

6. The following participants acted as Discussants: Messrs. Leopold Granquist and Svein Nordbotten (Sweden) for topic (i); Messrs. John Kovar (Canada) and Pascal Rivière (France) for topic (ii); Mr. Claude Poirier (Canada) for topic (iii); and Mr. Bill Winkler (United States) for topic (iv).

7. The participants expressed their great appreciation to Statistics Finland for hosting this meeting.

8. A summary of the main conclusions reached by the participants on the four substantive agenda items is presented in the Annex (English only).

RECOMMENDATIONS FOR FUTURE WORK

9. The Work Session considered the proposals for future work put forward by the Task Force composed of Ms. Paula Weir (United States), Ms. Pam Tate (United Kingdom), Mr. Pedro Revilla (Spain) and Mr. Elmar Wein (Germany).

10. The Work Session recommended summarizing in a publication "Statistical Data Editing, Vol. 3" the outcomes of recent and future meetings with special regard to evaluation methods and quality indicators for statistical data editing.

11. The Work Session also encouraged the further extension of the existing Web knowledge base (K-base) for data editing (http://amrads.jrc.cec.eu.int/k-base) towards making available on the website the evaluations and experiences of various data editing processes and establishing a steering group. The group would be responsible for identifying and encouraging researchers to share the results of their work through the site.

12. The Work Session recommended that a future meeting on statistical data editing be convened subject to the approval of the Conference of European Statisticians and its Bureau. The session also recommended that the following items be included on the agenda:

(i) Development and use of data editing quality indicators (indicators for different users; how much and how often?; monitoring and evaluation of the performance of a given editing and imputation methodology; metadata for evaluation of both processes and results);

(ii) Developments related to methods and techniques (new imputation methods; outlier detection, impact of IT developments; (meta)data handling; selective editing vs. increase demand fro microdata / user specific analysis; Web surveys and "real time statistical results");

(iii) Data editing processes and survey processing (structure of data editing processes; typical data editing processes; interfaces to other survey processes; feedback to survey design and other processes; optimisation of survey operations; management of the editing process; demands of data editing methodologists on other survey processes and demands on data editing);

(iv) Data editing by respondents and data suppliers (how much editing? which types of edits? – recommendations; strategies/best practices to convince suppliers of secondary data to perform edits; information on the reliability of "foreign" edits; incorporating data from administrative sources);

13. The representative of Spain informed the meeting, that the National Statistical Institute of Spain (INE) would like to host the next meeting.

ANNEX

Summary of the main conclusions reached by the participants at the May 2002 Work Session on Statistical Data Editing

The background documents referred to in this summary, were made available on the Secretariat's website: http://www.unece.org/stats/documents/2005.05.sde.htm

Topic (i): Planning and management of statistical data editing

Documentation: Invited papers by Canada (WP. 2), Germany (WP.3) and United States (WP.5). Supporting papers by Azerbaijan (WP.6), Bulgaria (WP.7), Finland (WP.8) and United States (WP.10).
Discussants: Messrs. Leopold Granquist and Svein Nordbotten (Sweden)

1. The Work Session considered that for editing to be truly effective, the process must be part of a continuing improvement cycle of the whole survey process, and to this end the editing steps must be properly and systematically planned and managed. The discussion considered various ways for quantifying the effect of data editing and information management on planning, monitoring and fine-tuning the editing process.

2. The meeting considered the methods to evaluate editing and imputation procedures as a prerequisite for the efficient management of data editing. Good practices were discussed on using information from the editing stage in optimising resources devoted to editing and on improving other stages of the survey process.

3. Another important issue under consideration was defining necessary data for obtaining knowledge of error sources for improving survey processing. Different ways of communicating experiences from planning and collection of data on error sources to survey managers were outlined. The communication of this information to data users was also discussed –experience shows that while some of the information is made available within the statistical agencies, it is often not provided to the external users of statistical data.

4 It was stressed that choosing the data to be used for evaluation is probably the most difficult part of any evaluation study. The use of simulated data to evaluate data editing was also considered. It was suggested, in this respect, to use both the historical data and simulated data.

5. The importance of administrative and other auxiliary data for planning, editing and imputation was stressed. The levels of accuracy of such data are often variable and the impact of such differences was discussed. The use of all possible sources of such data was considered as especially important. Evaluating alternative methods and criteria for that was also under consideration. Some participants stressed that not only the quality of administrative data, but also their processing by statistical agencies has an impact on the result. It was pointed out during the discussion that it is important that statisticians have access to different data at the moment of data capture and other phases of editing, which should be possible with increasing interactive access to auxiliary data.

6. In concluding the discussion on this topic, participants emphasized the need for: (i) the development of a typology and conceptual frameworks for the description of editing processes; (ii) the collection of empirical processing information from design end execution phases of survey processes; (iii) routines and comparison and evaluation of process data; (iv) a systematic exchange of process knowledge gained; and (v) routines and processes for obtaining better knowledge of the quality of the input, and suggested that some of these topics be considered in future work on statistical data editing. The Work Session also confirmed the interdependence between the data editing and other phases of survey processing.

7. The participants suggested that it would be useful if the conceptual frameworks and various methodologies be included on the respective website (see the section “Recommended Future Work”).

Topic (ii): Measuring and evaluating data editing quality
Documentation: Invited papers by Canada (WP.11), Italy (WP.12), Spain (WP.13) and University of Southampton (WP.14). Supporting papers by Austria (WP.15), France (WP.16), Spain (WP.17) and Sweden (WP. No.18).
Discussants: Mr. John Kovar (Canada) and Pascal Rivière (France)

8. The emphasis of the discussion under this agenda item was on the measurement of data editing quality and therefore concentrated on the indications allowing its evaluation. The questions of defining data editing quality indicators were addressed both from the point of view of accuracy and various other quality dimensions as related to user needs such as timeliness, relevance, consistency, completeness, etc. There was general agreement that the basic goal of editing is to improve quality. While accuracy was considered as one of the key aspects, participants pointed out that the evaluation of data editing should also be linked to other quality aspects.

9. The importance of the practical value and usefulness of proposed indicators and their ease of implementation and understanding were emphasized. Particular attention was paid to the impacts on process quality as well as to the quality of outputs. Different kinds of necessary metadata for constructing the required indicators were also discussed.

10. The work session emphasized that the quality indicators should provide sufficient information to managers, analysts, researchers and other data users. This information should cover the impact of various kinds of errors (e.g. sampling and non-sampling errors, non-response bias, coverage errors, measurement errors, processing errors, etc.).

11. The participants considered the possible impact of quality indicators on choices between different methods, their comparison and benchmarking. The possibility of using quality indicators for budget allocation was also considered at the meeting along with their effect on continuous improvement (e.g. process optimisation, acquisition of tools, staffing, collecting the best practices, etc.).

12. Some participants stressed that while it is necessary to evaluate the quality of both error localisation and imputation, it is desirable that the quality indicators allow that a distinction be made between these two phases of the editing process, as well as the distinction between the indicators related to the process and to the output quality. Concerning the way in which the indicators were obtained, it is important to distinguish between the observational and experimental evaluation studies. A concern was expressed, that the evaluation activity should not end with collecting the quality indicators, but that these should be considered as an impulse for action – to improve survey processing.

13. It was suggested to consider the following steps in the evaluation of editing and imputation:
(i) verifying the statistical properties of a given approach for a given problem; (ii) choosing the best set of techniques for a given survey application; (iii) monitoring and optimising the performance of a given editing and imputation methodology; (iv) obtaining information on non-sampling errors sources;
(v) measuring the impact of editing and imputation on the original raw data.

Topic (iii): Editing of administrative data

Documentation: Invited papers by Finland (WP.19), France (WP.20) and Israel (WP.21). Supporting papers by Canada (WP.22), Czech Republic (WP.23) and Finland (WP.24).
Discussant: Mr. Claude Poirier (Canada)

14. The emphasis of this topic was placed on the major differences between editing of data originating from statistical surveys and obtained from administrative registers and records, such as the amount of information, the difference in concepts and the intended use of the data. Problems related to data definitions and various reference periods were outlined and discussed. Constraints emanating from the respective legislation were also mentioned in the discussion.

15 Typical methods may be used, when applying editing and imputation on data coming from administrative registers and records, depending on the available information: manual editing, historical, ratios, models, balance editing, outlier analysis, etc. The discussion considered recommendations that depend on the target use: primary source of information for which statistical inferences can be made, and development of a statistical infrastructure for the purposes of sampling frames, sampling information, editing other sources of data, estimation, analysis, etc. It was pointed out that automated methods may be preferred to manual methods.

16. Administrative data may over-cover the population due to duplicates, miss-linkage of units, or dead units that are not reflected in the administrative data. It may also under-cover the population of interest. The participants discussed the impacts and how to deal with this. Participants also discussed the issues related to the integration of microdata from different registers and records.

17. There was agreement that data obtained through administrative registers and records represent a valuable source of information, and there is a growing interest of statisticians to combining various data sources. Therefore, the use of administrative registers and records brings with them new challenges for all statistical processes including data editing and imputation.

18. When discussing potential improvements, standardization and harmonization (of concepts and definitions, as well as interfaces, data models, data formats, unit identifiers, etc.) was mentioned as a key for success. In concluding their discussion, participants agreed that this topic is very broad, which was demonstrated by the wide variety of issues covered by the background documents, and suggested to return to this topic at a future meeting.

Topic (iv): Impact of new technologies on statistical data editing

Documentation: Invited papers by Canada (WP.25), Italy (WP.26) and Netherlands (WP.27). Supporting papers by Belgium (WP.28), Finland (WP.29), France (WP. 30), Kyrgyzstan (WP.31), Netherlands (WP.32 and 33), Poland (WP.34), United Kingdom (WP.3) and United States (WP.36 - 39).
Discussant: Mr. Bill Winkler (United States)

19. This topic covered a wide variety of issues. Two of the invited presentations related to the editing an imputation of data from population and housing censuses. A system using a mixture of hot-deck and deterministic imputation was presented. Examples of demographic variables with highest frequency of errors were discussed (e.g. misreporting of relationship to Person 1, marital status, etc.). This system (CANCEIS) has been shown to be a highly efficient editing and imputation system which can be used by censuses and in various types of surveys to handle minimum change hot-deck imputation. Further enhancements are envisaged, such as adding the ability to perform deterministic imputation and a graphical user interface. The system has been used so far with social and household surveys, but it is believed to have a potential for business surveys. More study of the requirements of these surveys and some extensions to the system may be required.

20. Another presentation related to a new software currently being developed for the editing and imputation of hierarchical demographic data. This system (DIESIS) was compared with CANCEIS and the results suggested that there were no significant differences in quality. The evaluation has been performed by computing, for each variable, accuracy indicators of preservation of individual original values as well as of preservation of the marginal distributions. Therefore, the development will continue aiming at obtaining a tool for simultaneous editing and imputation of qualitative and quantitative variables.

21. The editing system presented is to be used for processing of structural business statistics. It is based on selective editing. Testing and evaluation is under way, and only the preliminary results could be presented at the Work Session.

22. Other presentations (contributed papers) raised a number of issues, such as evaluation metrics (Euredit), general systems (Blaise, Impect StEPS), selective editing (WP.30 and 33), automation (LANS, PCs, standardized packages, handhelds, etc.), difficulties (WP.30, 36 and 37), research directions (SOMs, Euredit, linear programming/MCMC, imputation) and training methods (basic skills and advanced development skills.

- - - - -