UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2005/41
31 May 2005

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-third plenary session
(Geneva, 13-15 June 2005)

REPORT OF THE MAY 2005 WORK SESSION ON STATISTICAL DATA EDITING

Prepared by the UNECE secretariat

1. The Work Session on Statistical Data Editing was held in Ottawa, Canada, from 16 to 18 May 2005 at the invitation of Statistics Canada. It was attended by participants from: Australia, Austria, Canada, Denmark, Finland, France, Germany, Hungary, Israel, Italy, Latvia, Netherlands, New Zealand, Norway, Poland, Republic of Korea, Slovenia, Spain, Sweden, United Kingdom, and the United States of America. A representative of the United Nations Educational, Scientific and Cultural Organization (UNESCO) also attended.

2. The agenda contained the following substantive topics:
(i) Editing administrative data and combined data sources;
(ii) Implementing editing strategies and links to other parts of processing;
(iii) Electronic data reporting – editing nearer source and multimode collections;
(iv) New and emerging methods, including automation through machine learning, imputation, evaluation of methods;
(v) Quality indicators and quality reporting.

3. Mr. John Kovar (Canada) acted as Chairman and Mr. Claude Poirier (Canada) as Vice-Chairman.

4. Dr. Ivan Fellegi, Chief Statistician of Canada opened the meeting. He highlighted the twenty-year involvement of the Conference of European Statisticians in the field of statistical data editing. The statistical data editing working group originating from the UNECE/UNDP Statistical Computing Project was the only one that survived to the present. Over the years, the group has proved that it was still possible to come up with new and improved methods, and to align them with progress in statistical computing and the subject matter areas. Dr. Fellegi highlighted the importance of the substantive topics to be discussed at the meeting, such as editing data originating from administrative sources and specific issues related to electronic data reporting. He praised the group for the very active exchange of experience as documented by the numerous publications and the website on best practices in data editing and imputation (maintained by Statistics Canada with input from other participants). He thanked everyone involved in the preparatory work for the meeting – the authors of papers, the Organizing Committee, session discussants, UNECE secretariat and special thanks to Claude Poirier and his team for the organizational arrangements.

5. The following persons acted as Discussants and Session Organizers: Topic (i) - Ms. Natalie Shlomo and Ms. Heather Wagstaff (United Kingdom); Topic (ii) - Ms. Orietta Luzi (Italy) and Mr. Carsten Kuchler

(Germany); Topic (iii) - Mr. Pedro Revilla (Spain) and Ms. Paula Weir (United States); Topic (iv) - Mr. Ton de Waal (Netherlands) and Ms. Maria Garcia (United States); and Topic (v) - Messrs. Leopold Granquist and Svein Nordbotten (Sweden).

RECOMMENDATIONS FOR FUTURE WORK

6. Participants discussed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Messrs. Eric Rancourt (Canada), Philippe Brion (France), Carsten Kuchler (Germany) and Jeffrey Hoogland (Netherlands). When preparing the proposal, the working group took into account suggestions made by other participants in side discussions during the meeting.

7. Participants considered that the intensive developments on the emerging topics that have appeared in recent data editing meetings, for example editing by respondents and editing of data from administrative and combined sources, justify the future continuation of the international exchange of data editing and imputation experience. The Work Session, therefore, recommended that a future meeting on statistical data editing be convened in 2006, subject to the approval of the Conference of European Statisticians and its Bureau, with a study programme as suggested below.

8. The following substantive topics were recommended for the study programme of the future work session. The quality related issues were not retained as a specific topic, but they are expected to be present within each topic.

- Editing nearer the source;
  - Electronic acquisition and processing of data (EDI, EDT, EDR);
  - Editing by the respondent;
  - Measuring the response/editing burden;
  - Quality and reliability;
  - etc.

  (possible contributions from Germany, Spain and United Kingdom)
- Editing data from multiple sources;
  - Data from multiple sources;
    - Surveys;
    - Censuses;
    - Admin;
    - Matched data;
    - …
  - Re-editing of combined data;
  - Assessment of data quality;
  - etc.

  (possible contributions from Austria, Canada, Finland, Germany, Israel, Netherlands and Norway)
- Editing microdata for release;
  - Provision of microdata;
  - User demands;
  - Quality of data;
  - Post-confidentiality editing;

  (possible contributions from Austria, Denmark, Germany, Israel and Norway)
- Macro-editing;
  - Time series;
  - Data confrontation;
  - Consistency;
  - Quality;
  - Selective editing;
  - etc.

  (possible contributions from Canada, Germany, Spain, United Kingdom and United States)
- New and emerging methods;
  - Testing and benchmarking;
  - Spatial dimension;
  - Data and imputation models;
  - New software (other than the software demonstrations);
  - etc.

- Software presentations.

9. The delegation of Germany offered to host the next meeting on statistical data editing, and invited the participants on behalf of the President of the Federal Statistical Office. The meeting will be held in Bonn in autumn 2006.

STATISTICAL DATA EDITING, VOLUME NO. 3 – IMPACT ON DATA QUALITY

10. The Chairman provided some background information to the Statistical Data Editing publication series. Volume 1 defined what is data editing; Volume 2 focused on the "how" of data editing, while Volume 3 intends to describe the "how well". The editorial group that was set up to coordinate the work on Volume 3 was led by Mr. John Kovar (Canada) and composed of Mr. Leopold Granquist (Sweden), Mr. Carsten Kuchler (Germany), Mr. Pedro Revilla (Spain), Ms. Natalie Shlomo, Ms. Heather Wagstaff (United Kingdom) and Ms. Paula Weir (United States). It has prepared an outline of the future publication comprising introductory texts and references to the papers expected to form the content of individual chapters. The outline, together with the revised versions of some papers presented at previous work sessions, was made available to the participants prior to the discussion. It is planned to finalize the publication by September 2005.

11. There was a proposal to include references to the K-Base and need for benchmark data. The benchmark data were originally mentioned in Madrid (2003), and it slipped from the publication mainly because there were no suitable papers available. The Chairman will look into papers referring to this issue or will build it into the introductions. The Chairman also asked the participants to propose additional papers for inclusion into the publication.

FURTHER INFORMATION

12. The conclusions reached during the discussion of the substantive items of the agenda are contained in the Annex. All background documents and presentations for the meeting are available on the website of the UNECE Statistical Division (**http://www.unece.org/stats/documents/2005.05.sde.htm**).

13. The participants expressed their great appreciation to Statistics Canada for hosting this meeting and providing excellent facilities for their work.

ADOPTION OF THE REPORT

14. The participants adopted the present report before the Work Session adjourned.

ANNEX

SUMMARY OF THE MAIN CONCLUSIONS REACHED AT THE WORK SESSION ON STATISTICAL DATA EDITING

I. Editing administrative data and combined data sources

Discussants: Natalie Shlomo and Heather Wagstaff, United Kingdom
Documentation: Invited papers by Canada, Finland and New Zealand; Supporting papers by Canada, Denmark, Israel, Norway and United States; Tabled paper by Israel

1. The discussion on this topic covered three sub-groups of issues: (i) use of administrative data for business surveys and economic statistics; (ii) combining data sources for social and demographic statistics and (iii) other processes supporting data editing and imputation.

2. In connection with the administrative sources the presenters demonstrated several practical implementations, for example the use of tax and customs registers for economic data and population registers, employees and self-employed registers for social statistics. The discussion referred to the fact that administrative records are originally not intended for statistical purposes. For example, the concern was expressed that the metadata concepts in administrative registers and records do not fully meet the needs of statistics, or that the metadata are not available in an electronic format. This may have an impact on the final quality of data as it may cause problems in interpreting correctly the data. On the other hand, it was felt that in general, administrative data improves the quality of statistical data through error localization, imputation models, outlier detection, and selective editing techniques. It may also reduce the need for edit and imputation.

3. The procedures applicable to administrative sources may change, and statisticians may have very little influence on these, if any. It is, therefore, very important that the statistical offices follow all the changes in the administrations, analyse these changes and adapt their editing and imputation systems accordingly. Methods for improving interaction and cooperation with data suppliers were also discussed. It was mentioned that the stability of administrative sources might also depend on the legislative situation. However, when the use of administrative sources is mandatory according to law, there are still administrative processes beyond the control of the statistical offices.

4. The discussion also mentioned the importance of setting a threshold to identify significant information for the purpose of selective editing. While it was considered a useful method, some participants emphasized that the understanding of significant information may be relative according to the final use of statistics. For example, economic and financial analysts require a different level of detail. Another issue linked with selective editing concerned the elimination of small business units, as these are influential in some survey processes where survey data of small units are directly replaced by administrative data.

5. Linking different administrative sources was mentioned as a way to improve the editing and imputation processes. Similarly, some countries shared their experience in combining administrative and survey data, provided that the results are used exclusively for statistical purposes. In this connection some participants suggested that trying to consolidate administrative data in common statistical registers is one way to improve the completeness and reliability. The possibility of matching records from different registers may be subject to the legal situation in individual countries.

6. Statistical metadata represent a key in linking together data from different sources. Some offices experienced problems in building unified metadata systems fully consistent with the UNECE recommendations. On the other hand, some other countries are attempting to share the statistical metadata systems with the administration, thus attempting to achieve the methodological consistency of data from administrative sources.

7. While the term "administrative sources" is usually understood as using registers and records belonging to the public administration, there are also attempts to use the administrative systems of enterprises. These enterprise administrative systems would be used, applying electronic data reporting techniques, for supplying data to both public administration and national statistical offices.

8. The harmonization of concepts and definitions is a necessary precondition for the use of multiple administrative sources within the survey process. In addition, the importance of quality checks on final outputs was emphasized.

II. Implementing editing strategies and links to other parts of processing

Discussants: Orietta Luzi (Italy) and Carsten Kuchler (Germany)
Documentation: Invited papers by Canada, Netherlands/United Kingdom and Sweden; Supporting papers by Finland, Germany, Israel, Italy, Netherlands, Spain, United Kingdom; Tabled paper by the Netherlands.

9. The discussion on this topic was organized under four sub-themes: (i) editing and variance estimation, (ii) editing and data dissemination, (iii) (re-)designing editing processes, and (iv) using external data sources.

10. The discussion on the first sub-theme stressed that estimating the variance proportion due to editing and imputation is necessary for assessing data quality in terms of measuring the overall quality of survey estimates as well as the reliability of the involved data editing processes. In particular, the computation of variance proportions may lead to the decision of whether to increase sample size or data editing efforts to obtain high quality survey data. Different kinds of simulation approaches were discussed with which the computation task can be accomplished under different model assumptions. Tools for variance estimation and simulation were presented.

11. With regard to statistical dissemination, the participants discussed the relationship between editing and imputation and statistical disclosure control. In particular, the perturbation methods applied to microdata may cause failure of edits and may affect the statistical consistency of the data (distribution, variables associations, etc.). Editing and imputation under statistical disclosure control is different, as it has to balance the minimized disclosure risk with the editing and imputation objectives. An adoption of the Post-Randomisation Method for balancing consistency and confidentiality purposes was presented.

12. Redesigning editing and imputation in survey processes has an impact on overall survey planning and organization. Not only the data editing methodologists are responsible for the efficiency of the editing phase, but this also requires a broad approach involving many organizational and methodological aspects. A part of the discussion focused on the issue of selective editing, in particular on using quality criteria (bias, variability, marginal/joint distributions) based on the level of detail/aggregation of the published data.

13. The awareness of managers and subject-matter specialists on the utility of the editing and imputation methodology applied to the data should be ensured within their field of responsibility. For example, it is necessary to explain how missing values for particular reporting units were imputed.

14. A general remark made in the discussion was that it is necessary to take into account the multipurpose character of statistical surveys, in order to ensure that estimators made for various purposes meet the required criteria.

15. When speaking about integrating multiple data sources, participants highlighted the necessity to balance the increased costs of complex data editing and imputation methods with the reduction of costs by using already existing data.

III. Electronic data reporting – editing nearer source and multimode collections

Discussants: Pedro Revilla (Spain) and Paula Weir (United States)
Documentation: Invited papers by Austria and United States; Supporting papers by Canada, Italy, Latvia, Poland, Spain and United States

16. Electronic data reporting (EDR) offers the possibility of using built-in edits in electronic questionnaires, thus moving editing closer to respondents. This was not possible at the time of paper questionnaires or other modes of data collection. The discussion on this topic identified the following issues of interest: (i) the impact of EDR on the editing strategy; (ii) the optimization of the effectiveness of both editing at data capture and at post-collection survey processing; (iii) performance measures and indicators of editing at data capture and post-collection processing as it affects the overall survey quality; (iv) challenges and issues such as security and confidentiality, respondents' burden, response rates, timeliness, and incentives; (v) the use of focus groups, usability or cognitive testing of data providers with respect to editing at data collection.

17. Several participants stressed that in spite of the possibility of including checks in electronic questionnaires, there was still a way to go before achieving a true respondent-side editing. The balance of the respondent burden was identified as one of the issues to resolve, in particular, the fatigue from error messages. Using unique identifiers for the checks may facilitate further reference by respondents.

18. Different practices were described with regard to balancing the respondent-side editing and post-processing. Concerning the question, whether to automatically replace the data as a result of edits in the presence of the respondents, the current practice in most of the offices is just to flag the cells to be verified by the respondent. At present, some offices do just a minimum editing, like flagging the non-response and other very elementary checks.

19. A transaction log identifying the original data and their subsequent changes may be useful from the methodologist’s viewpoint. More sophisticated editing tools in the EDR environment may take into account the respondent’s previous responses.

20. The opinion was expressed that users expect to have some edits when using the on-line electronic questionnaires and might be surprised if they did not observe any edits. Many surveys presently offer several options, and users have a choice between the Internet and traditional forms of data response. A comfortable respondent-side editing may be one of the incentives in choosing the Internet. While the error messages generated by the edits can be perceived on the one hand as negative, on the other hand they guide the questionnaire completion process and discharge part of the respondent’s responsibilities.

21. Subject matter specialists can be involved in several ways. One is the testing of electronic questionnaires and the built-in edits and the edit failure messages. Another is in setting the priorities for what checks to build-in, as these depend on the original purpose of the survey.

22. Users already have experience with applications similar to EDR such as e-commerce and e-banking. These applications are, by nature, very strict as to the precision of data entered in on-line forms. Therefore, they may accept if the Internet application refuses to submit statistical data due to failed edits. This is obviously only one of the possibilities and the final decision has to be taken with respect to the nature of respondents and the survey.

23. Electronic data collection tools in addition to traditional tools introduce a possibility of another kind of errors; attention has to be paid to biases that may be due to the collection mode effect.

IV. New and emerging methods, including automation through machine learning, imputation, evaluation of methods

Discussants: Ton de Waal (Netherlands) and Maria Garcia (United States)
Documentation: Invited papers by: Italy and United States; Supporting papers by: Australia, Canada, Finland, Germany, Italy, Netherlands, Slovenia, Spain, United Kingdom, and United States

24. In general, the participants expressed their appreciation for the opportunity to share their experience on the cutting edge issues of statistical data editing and imputation, and recommended continuing this exchange at future meetings. The new approaches are often based on a combination and/or extension of original methods.

25. This topic covered a very wide variety of emerging issues that could be grouped under five headings: (i) automatic editing; (ii) imputation; (iii) editing and imputation for demographic variables; (iv) selective editing and (v) software.

26. Experts in the fields of operations research and mathematical logic expressed the opinion that combining techniques from both areas will likely result in improved methods and software for automatic editing.

27. When comparing different editing and imputation methods, the impact on quality has a crucial importance. The discussion pointed out that there is a lack of benchmark data available for such comparison. Criteria for the comparison should also include the systematic errors.

28. In some cases, complex models for data imputation are used to take into account a wider set of donors for missing values. Reference was made to an example when the Bayesian network provided better results as a simple hot-deck imputation. It was felt that complex models for data imputation might lead to data of higher quality. The relative complexity of some methods may not be an issue in the longer perspective, as practical implementations may occur.

29. Social and demographic data have some specificity that should be taken into account. A combination of editing and imputation methods seems preferable due to the mixed character of social and demographic statistics. It is too early to say whether these specific methods that are currently being developed, may be used in other subject-matter areas, and more testing might be needed after the development is completed.

30. With respect to selective editing, participants felt that the current state of the theory is closer to a beginning than to the end. Although progress has been made on many fronts, there seems to be ample possibilities for further improvement.

31. A number of projects aim at developing automated data editing software. While some results may be expected in the future, the human interaction and involvement in the editing and imputation process will still be necessary. Therefore, automation cannot be understood as just a “push-button” approach.

V. Quality indicators and quality reporting

Discussant: Leopold Granquist (Sweden)
Documentation: Invited papers by France and Italy; Supporting papers by Canada and United States; Tabled papers by Germany, Italy and United Kingdom

32. The participants considered two objectives of the quality indicators. The first objective is that the indicators shall inform producers (statistical managers, planners) about the efficiency of resources spent in different areas and on different processes to attain optimal resource allocation. The second objective is that the indicators shall inform users about the risk (an inverted quality measure) involved in taking decisions based on the different statistical products, and, if special statistical processing implies a cost, does quality gained justify this cost. It is not sure whether the same set of indicators can meet both objectives.

33. The ultimate objective of the exercise is quality reporting, and quality indicators provide the basis for it. The most difficult aspect of quality reporting is to present it in an effective way to users. The development of communication and global access to multiple sources of data allows users to compare the relative quality of individual sources. A specific case of quality reporting is the feedback to the respondents, which can contribute to improved quality of reported data and better respondent cooperation.

34. While this topic focused on data quality, the discussion also covered the quality of the data editing process. One of the purposes of the quality measures is to determine whether any editing is needed. In an ideal case no editing would be needed, and the fact that editing has to be used signifies that there are problems at some of the stages of the survey process. One of the goals of quality reporting should be, therefore, to provide feedback to survey organizers in order to improve the survey process. In this connection it was suggested that the quality related information should be stored in the metadata systems.

35. Simulation is one of the ways to assess the impact of data editing methods on the quality of data. Some participants considered that it might be useful to simulate errors and non-responses based on a modelled probability distribution. The difficulty of the issue may be in finding the probability model, and some participants were of the opinion that there should be more study of the real data. The pattern of errors might change over years and with the gradual change from paper to electronic methods of data collection.

36. Some participants mentioned that there is ongoing work in Eurostat's work programme on quality indicators and quality reporting and an effort should be made not to duplicate that work. Therefore, they suggested that future work on the data editing project should study data quality in the context of concrete data editing issues.

- - - - -