Distr.
GENERAL

CES/AC.71/2004/22/Add.2
5 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (v): Review and follow-up to the activities of the Conference of European Statisticians

REPORT OF THE WORK SESSION ON STATISTICAL DATA EDITING

Note by the UNECE secretariat

PARTICIPATION

1. The Work Session on Statistical Data Editing was held in Madrid, Spain, from 20 to 22 October 2003 at the invitation of the National Statistical Institute of Spain (INE). It was attended by participants from: Austria, Belgium, Canada, Finland, France, Germany, Greece, Hungary, Israel, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, Switzerland, United Kingdom, and the United States. The European Union was represented by the Statistical Office of the European Communities (Eurostat). Representatives of the United Nations Food and Agricultural Organization (FAO) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) also attended.

AGENDA AND PROCEDURE

2. Mr. John Kovar (Canada) acted as Chairman, Mr. Pedro Revilla (Spain) acted as Vice-Chairman.

3. The meeting was opened by Ms. Carmen Alcaide Guindo, President of the National Statistical Institute of Spain. Her address highlighted the demand on national statistical agencies to provide more data according to tighter deadlines and of greater quality. The professionalism and competence of statisticians would thus contribute to gaining the confidence of users. The role of data editing is crucial in meeting these goals.

4. The following substantive topics were discussed at the meeting:
   (i) Development and use of data editing quality indicators;
   (ii) Developments related to new methods and techniques;
   (iii) Data editing processes within survey processing;
   (iv) Data editing by respondents and data suppliers.

5. The following participants acted as Discussants: Messrs. Leopold Granquist, Svein Nordbotten (Sweden) and Thomas Burg (Austria) for topic (i); Messrs. Willam E. Winkler (United States) and Jean-Marc Museux (Eurostat) for topic (ii); Messrs. Claude Poirier (Canada) and Ton de Waal (Netherlands) for topic (iii); and Ms. Orietta Luzi (Italy) and Natalie Shlomo (Israel) for topic (iv).

6. The participants expressed their great appreciation to the National Statistical Institute of Spain for hosting this meeting.

RECOMMENDATIONS FOR FUTURE WORK

7. The Work Session considered the proposals for future work put forward by the Task Force composed of Claude Poirier (Canada), Carsten Kuchler (Germany), Jeffrey Hoogland (Netherlands), and John Charlton (United Kingdom).

8. The Work Session recommended having an in-depth discussion on long-term goals and expected achievements of the work programme of the Conference of European Statisticians in the field of statistical data editing and imputation, and to discuss new developments related to evaluation and quality, improving the edits and usefulness of data.

9. The Work Session also recommended to pursue the work on the methodological publication "Statistical Data Editing, Volume 3" devoted to the impact of editing and imputation on data quality.

10. The following possible substantive topics of common interest were recommended by the Work Session for further study:

(i) Editing administrative data and combined data sources – surveys, censuses;
(ii) Implementing editing strategies and links to other parts of processing;
(iii) Electronic data reporting (EDR) – editing nearer source and multimode collections;
(iv) New and emerging methods, including automation through machine learning, imputation, evaluation of methods.

11. The Work Session recommended that a future meeting on statistical data editing be convened in 2004-2005, subject to the approval of the Conference of European Statisticians and its Bureau, to consider the topics outlined in paragraphs 9 to 11 above. The delegation of Canada informed the Work Session, that Statistics Canada would like to host the next meeting.

FURTHER INFORMATION

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

13. Presentations, detailed summaries prepared by the discussants on individual topics and all background documents are available on the website of the UNECE Statistical Division (http://www.unece.org/stats/documents/2003.10.sde.htm).

ADOPTION OF THE REPORT

14. This report, as well as the attached summary of main conclusions (see Annex), were adopted by the participants at the Work Session before it adjourned.

SUMMARY OF THE MAIN CONCLUSIONS REACHED

A. Data editing by respondents and data suppliers

Documentation: Invited papers by: Finland (WP.10), Spain (WP.11) and Sweden (WP.12); Supporting papers by: Belgium (WP.34), Germany (WP.35) and United States (WP.36 to WP.39); Summary by the Discussants is provided in CRP.2 (electronic data reporting) and CRP.3 (use of administrative records).
Discussants: Orietta Luzi (Italy) and Natalie Shlomo (Israel)

15. The presentations showed that the widespread use of information and communication technologies (ICT) and the increasing use of electronic data storage, management and data transmission, invite statistical agencies to develop as much as possible ICT-based approaches to data collection. This is a general trend, which has, naturally, its impact on statistical data editing. The goal is to optimize the effectiveness of both data capture and survey processes and obtain information from data suppliers (persons, enterprises, private agencies, public administrations) with greater guarantees for improving data quality; improving timeliness; reducing organizational costs (questionnaire delivery, coding, data entry, data editing, etc.); security; reducing respondents burden; reducing non-response by increasing cooperation and offering benefits to data providers. The use of ICT for data capture concentrates the costs, for the statistical agencies, in the design phase of the questionnaire and the need to manage and improve the relationship with respondents. The gains are made in the reduction of processing costs, improvement of data quality and timeliness of data transmission and further data processing. Participants noted that, for example in the EU context, timeliness has become a crucial element because of regulations imposed on member countries to provide both preliminary and final results by pre-defined deadlines. This puts pressure on the data editing phase.

16. The advancement of ICT is a very important motivation for moving the data editing phase to an earlier stage, before the data are transmitted from the reporting unit to the statistical agency – including editing by respondents, suppliers of administrative records or interviewers. Electronic data reporting (EDR) offers the possibility of using built-in edits in electronic questionnaires, while this was not possible in the days of paper questionnaires. The advanced level of development in Internet technologies invites further in-depth consideration of how it can be better used for statistical data editing.

17. One of the expectations of introducing ICT (Internet, etc.) technologies is that enterprises would prepare data in a form needed also for their own purposes (decrease of the reporting burden). However, some participants shared their experience that enterprises do not always keep all the information needed for reporting in an electronic format, while the original expectations were the opposite. In this connection, it was highlighted that statistical agencies should seek how data editing embedded into EDR can help respondents and data suppliers, and efforts should be made to learn more about data providers' needs and possibilities. It is important that this EDR is not perceived as a new form of burden on data providers. Cognitive studies (usability testing) were recommended.

18. The use of incentives (to encourage provider-side editing) needs further research with respect to their impact on response rates. Participants raised the issues of the type and amount of incentives that could be offered and their advantages and disadvantages.

19. Several papers touched on the subject of increasing the use of data from administrative registers and records, and thus specific issues of editing the administrative or external data. No administrative data are ready for use in the statistical process, and statistical classifications and concepts are usually inconsistent with the needs of a particular survey. New research is needed to adapt administrative or external data to the statistical processes, and develop methodologies for improving the edit and

imputation procedures. Several participants also expressed their interest in looking further into the issue of identifying systematic errors of administrative data.

20. During the discussion, the participants expressed their interest in knowing more about the impact on overall data quality of editing on the supplier's side as well as evaluating the amount of edits that can be made by data providers. Further study is needed to examine the quality issues.

21. There was also a discussion on pre-filling of questionnaires by data obtained from administrative sources, as it can improve the timeliness. In some cases the questionnaires were pre-filled by previously reported data. Some participants stressed that these techniques might present some risks as the responses are influenced by pre-filled values, and there are also confidentiality issues.

22. In concluding their discussion on this topic, the participants agreed that several issues related to provider-side editing in EDR and the use of administrative data, require further research. Some participants stressed the importance of using focus groups when determining the most appropriate strategy for a specific survey context using electronic data reporting with embedded edit checks. In addition to detailed suggestions for further work mentioned in the previous paragraphs, the main research topics were identified in the area of moving editing as close as possible to data providers:

- Integrating the use of Internet, other forms of electronic data transmission and Computer Aided Interviews (in general CATI, CAPI)
- Measuring the effects of mixed modes of data collection on data quality;

and the following main research topics in the area of using administrative data in survey processes:

- Making use as much as possible of external data for reducing costs/burden and increasing the quality of the data;
- Integrating different sources of information in the survey processes (linkage, imputation, modelling);
- Increasing cooperation with data suppliers of administrative and external data to conform with the requirements of the statistical agency.

B. Data editing processes within survey processing

Documentation: Invited papers by: Spain: (WP.8) and United Kingdom (WP.9); Supporting papers by: Canada (WP.28 and WP.29), Israel (WP.30), Netherlands (WP.31) and United States (WP.32)
Discussants: Claude Poirier (Canada) and Ton de Waal (Netherlands)

23. The discussion covered the aspects of the editing infrastructure and the role of editing and imputation in improving the statistical survey program as a whole. The diagnostic reports from the editing and imputation process represent a valuable source of information to fine-tune other survey processes.

24. Several statistical agencies try to standardize the technical infrastructure, methodologies and tools for editing and imputation. This can be done in conjunction with a modernization process of the whole office. The modernization project undertaken in the Office for National Statistics in the United Kingdom, drew a lot of attention. The aim of editing and imputation methods is to assemble cost-effective, standard editing and imputation tools that could be used for all data sources, which would incorporate best practices and could be applied in a standard fashion with minimal human intervention. Information about the program is on the ONS website (http://www.statistics.gov.uk/). More details will be added as the work progresses.

25. The meeting considered how to set realistic aims when designing a new survey processing system. There is a risk of being too ambitious: taking account of various kinds of output, aiming at a

high level of automation, and developing a too complex infrastructure. Such a change requires the investment of a lot of resources and time – the system can become overly complex and therefore prone to failure. It was noted that the main question is not ambition but flexibility and the ability to learn from the mistakes on the way toward developing a new system. There should be a vision of what the ideal system should be like and then be realistic in its implementation.

26. The pros and cons of changing the whole survey system at once versus implementing gradual changes were considered. Participants discussed several examples of developing a completely new survey system. The process of making a radical change has always been difficult, it can succeed only over time requiring several adjustments. Budget constraints have a major impact. A limited budget makes it difficult to change everything at the same time, therefore it can be easier to introduce changes gradually. There is an additional advantage in people learning to use the environment in one survey and then using the obtained knowledge in other surveys. It is more difficult to change the mentality to convince people to use new radical general solutions.

27. A generalized system can grow out of small software packages by adding bit-by-bit the required functionality. However, it can be a big problem to integrate the small software packages into the overall editing process. By developing modules one at a time there is a danger of having resulting products that are not always compatible, nor on a common software platform.

28. There was discussion on how to improve the links between editing and imputation and other survey processes in practice. Communication and transfer of information is a key issue. Being able to use the results from editing and imputation depends very much on creating the right metadata and adequate interfaces between the editing and other processes. The knowledge sometimes does not move between different departments and persons. Different approaches have been tried in different offices and over time. Some offices organize surveys in teams, which contributes to a good exchange of information about the survey. But in this case it is difficult to take advantage of the synergies between different surveys and to develop a general methodology suitable for different surveys. The communication and the subject expertise should be present at the same time as the functional expertise.

29. The systems to evaluate the impact of non-response and imputation on the quality, GENESIS and SEVANI, presented by Statistics Canada generated much interest as they might also be applicable in other offices. The systems are implemented in SAS.

30. The use of administrative data in the editing process was discussed. When using administrative registers as a primary source of data, extensive editing and cleaning must be undertaken to improve its quality. Incorporating administrative data as a secondary source can help to understand the underlying mechanisms of non-response and improve estimation procedures. By incorporating different sources of administrative data, edit failures can automatically be corrected, data can be obtained for non-respondents and target variables can be imputed based on more complete covariates and better imputation models.

31. The question was raised of the usefulness of the feedback generated by editing and imputation. At present, the experiences with quantifying the gains achieved by using the information from the editing and imputation process for improvements in output data quality and in making the survey process more efficient are not well documented. It was agreed to encourage such documentation.

C. Developments related to new methods and techniques

Documentation: Invited papers by: Spain (WP.5 and WP.6) and United States (WP.7); Supporting papers by: Canada (WP.18), Germany (WP.19), Greece (WP.20), Italy (WP.21), Netherlands (WP.22 and WP.23), Switzerland (WP.24), United Kingdom (WP.25) and United States (WP.26 and WP.27)
Discussants: William E. Winkler (United States) and Jean-Marc Museux (Eurostat)

32. Methods and algorithms aimed at automated editing and imputation, for example, the application of REG-ARIMA models and Integer Linear Programming were presented, along with software applications and some comparative studies. The presentations and following discussion were spread over a wide range of issues related to methods, software and the impact of information and communication technologies (ICT) developments. However, the participants expressed their interest in following the progress in data editing and imputation and suggested to return to several of the presented new methods and techniques (e.g. NIM, new methods included in Euredit, etc.) when the work is progressed.

33. The discussion proposed that it is important for the researchers to have a benchmark library of data sets for test and comparison purposes. Such a library may include both real data and artificially generated data. Some participants emphasized the usefulness of (randomly) generated artificial data, which may better cover the various situations to be tested. A method to take a true data set with intentional errors did not gain general acceptance, but would deserve further consideration, and more work should be done.

34. Some of the comparative studies showed that various systems may have different advantages and shortcomings, and they can also differ with technological requirements (more or less need for changing the code, need of large or small computing resources, etc.) so the choice has to follow a concrete situation.

35. The discussion reviewed issues to be resolved when designing data editing and imputation processes. There are particularities depending on the statistical subject matter to which editing and imputation are applied. Differences were mentioned between population and business statistics. For example, it was mentioned that it is difficult to determine exact bounds for economic statistics data.

36. There was an extensive discussion of software tools and their portability. The participants noted that it is unlikely to have commercial and generalized software solutions. It was mentioned that some statistical agencies made an effort to make available a source code to other users. Several examples were mentioned where the software was reused between different countries. Delegates stated that a useful trend is toward open source code.

37. Sharing of the software tools is a “grey area”. The examples quoted allowed free-of-charge use and without the guarantee of support. Re-using the software by other agencies had advantages in testing and tuning the tools. However, some delegates stressed that a correct relationship between cooperating statistical agencies is very important. Publicizing the bugs found by third parties may be damaging to the credibility of the originating agency when ignored.

38. It was stressed that differences between national statistical agencies may limit the portability and some modifications are required before reusing the software. The agency that wants to reuse the software also needs the necessary level of skills. A suggestion was made to have the source code, but also the underlying methodological descriptions and notes in some form of a library, consistent with the policies of the statistical agency.

39. The discussion underlined that editing and imputation should be considered jointly. Examples of good and bad practices were quoted.

40. In the general discussion, participants also considered general issues of new methods and techniques in statistical data editing. Many new ideas are acceptable. It was suggested that it is important to describe these ideas in a manner that relates them to other existing ideas in statistical data editing and imputation.

- **Is the method described a new concept, a novel extension of existing ideas, or an application of concepts from some other field?** How does the method relate to other concepts in statistical data editing and imputation? Are the ideas primarily theoretical or are they practical in the sense of having been implemented?
- **How practical are the ideas?** Have they been implemented in a research system or in a production system? To what types of data (discrete, continuous, combination of discrete and continuous) are the ideas applicable? If the method is part of an edit/imputation system, do the output data satisfy probabilistic distributional constraints? If the methods are in a production system, how good is the user interface of the software?
- **Do the concepts rely on expert knowledge?** Some methods require the use of experts to develop and train other individuals in their use. Are the methods described those that can be easily understood by potential appliers and users of the technology? Will the potential users need extensive training?
- **Do the concepts rely on training data?** How were the training data created? How representative is the training data of representative situations in which statistical data editing and imputation are applied?
- **How was testing performed?** What quantification of the performance of the new methods has been done? Were standard evaluation metrics applied in the analyses?

D. Development and use of data editing quality indicators

Documentation: Invited papers by: Austria (WP.2), Italy (WP.3) and Netherlands (WP.4); Supporting papers by: Canada (WP.13) and Spain (WP.14); Summary by the Discussants is provided in CRP.4
Discussants: Leopold Granquist (Sweden); Svein Nordbotten (Sweden); and Thomas Burg (Austria)

27. The discussion on this topic focused on the development and use of data editing quality indicators for different uses and different purposes. The indicators can focus on the quality of the data, and on the quality of the performance of a given editing and imputation methodology. Both qualitative and quantitative information is needed. The editing indicators cannot improve the quality of input data directly but they can pinpoint problems in the collection process and, therefore, improve the incoming data quality in future surveys. For example, if a particular variable is frequently changed as a result of editing checks, the formulation of the question in the questionnaire might need improvement. The usefulness of data can be improved if additional information (e.g. data from other sources) is taken into account.

28. The indicators should be developed with the clear aim in mind of what they are supposed to measure and why. Questions that need to be answered are, for example, for whom and for what purpose the quality indicators are needed, which indicators to choose, how to collect data on them, and how to use the collected information to improve the survey process. It was agreed that the aims of the editing quality indicators are: (i) to obtain knowledge of measurement errors or severe problems for respondents to provide answers, in order to have a basis for improving ingoing data quality of the survey; and (ii) to have a basis to take measures to increase the efficiency of the data editing process. In addition, the indicators should be useful not only for producers but also for the users of data. It is necessary to produce indicators that can inform the users about data quality in general and to have a strategy on how to communicate this information to users.

29. Several possible indicators were presented in the papers. It was noted that often a special study is required to produce and analyse the edit process. For routine needs, some indirect, proxy measures are required that would be easy to compute. Indirect indicators that relate to the impact of editing can be produced at a later stage after the editing process has finished.

30. It is important to have a good structure for recording the collected process data. Some offices have developed databases and information systems for that purpose. Implementing and maintaining the information system is a demanding and time-consuming task. Therefore, it is important to have a strategy for populating and keeping the information systems updated. A high degree of standardization of both metadata and the quantitative indicators can be considered desirable. Generalized software and tools to calculate the required indicators help to introduce the computation of these indicators as a regular procedure and to speed up the process.

31. Measuring and documenting the effects on data of any data processing activity, including editing and imputation has become a mandatory requirement in many NSOs. As a consequence of the introduction of the Total Quality Management approach, the whole production process should be fully documented and as well as the data editing steps. Proper documentation helps to improve the sharing of information and to maintain the knowledge when staff changes.

32. The management aspects of improving the data cleaning process were considered. It is a complex process and needs a management plan. In order to succeed with implementing major improvements, the changes must be supported by both staff and management. It can be viewed as a job enhancement and increases the qualification of staff.

33. The main focus of the discussion has been on the accuracy dimension of editing quality. However, the editing process also affects the timeliness, clarity, accessibility, comparability and coherence of the results. The cost-effectiveness of the editing process is as important. The indicators measuring the editing process itself have to be cost effective, the costs of producing the indicators have to be taken into account when considering their usefulness for improving the survey process.

34. Sometimes data edited manually by experts are used as “true data” in evaluations. The question is whether the evaluation of data from an editing process against manually edited data mean anything more than a comparison to another editing method. If the edited data are similar to the “true data”, the new method can be justified by saving resources as the output quality will be the same. A good idea might be to create benchmark data sets by creating synthetic data and use simulations for evaluating editing methods.

35. Other ways of obtaining information on editing quality are post-editing studies and simulation studies. Post-editing studies always require resources and the possibility of making them decreases with limited budgets. Therefore simulation studies are very important.

36. The question is also how to create realistic mechanisms simulating the errors. It is important to simulate dependent errors and error patterns. It was agreed that there should be more research into errors and the mechanisms that are producing errors. Finding answers to these questions contributes to the improvement of the survey process without concentrating only on quality. It also provides information to introducing checks into questionnaires.

37. Participants discussed how to guarantee that all important edits are included and how to identify unimportant edits that are included. Mathematical methods can be used to clean out the redundant ones from a given set of edits. Some participants had the view that the unimportant edits should be identified only if they have a negative effect on the editing process, e.g. by slowing it down. Moving unimportant edits can also be risky because these can be good for detecting errors in imputation. Information from the subject-matter area and from previous surveys and statistical models can also be used for this purpose.

38. An interesting visual tool was demonstrated showing how data values change while travelling through the processing stages. The tool presents a concise way of reviewing the big picture of what is happening to the data through the editing process. It allows the share of data changed through the editing process and the impact of these changes to be analysed. The results thus discovered lead to further questions that allow the overall survey processing steps and data quality to be improved. The tool highlights important questions about what is happening in the edit and imputation process as a whole and should bring people working on the different stages of one survey together in order to obtain a better overall result.

E. Brainstorming session on "Statistical Data Editing, Volume 3"

39. The participants recalled the two previous volumes of "Statistical Data Editing". Volume 1 focused on what is editing and Volume 2 on techniques and methods.

40. The work on Volume 3 was initiated by the Bureau of the Conference of European Statisticians in 1997, and will focus on evaluation methods and quality issues. Between 1997 and 2003 several papers were produced to develop this methodological material. During the discussion, it was pointed out that comparative studies presented at this Work Session also present good basic material for Volume 3.

41. An editorial group composed of Pedro Revilla (Spain), Carsten Kuchler (Germany), Natalie Shlomo (Israel), Leopold Granquist (Sweden) and John Kovar (Canada) was created to coordinate the work on Volume 3. The editorial group presented its proposal on the structure of the publication, which was further commented on by the participants. A preliminary structure was proposed during the brainstorming meeting, which will be further developed:

- Framework
- Quality Measures
  - Why / How / What
  - Evaluation of quality
  - Impact of editing and imputation process on the quality
  - Edits
  - Variance of estimates
  - Measures for users
- How to improve on the quality
  - Reaction to the feedback and reaction to other processes
  - Impact on new modes
  - Use of variance
  - Improving efficiency (including impact on timeliness)

42. The participants agreed to continue an exchange of views after the meeting, and to send their proposals by e-mail to kovar@statcan.ca .

F. Other Business

K-BASE

43. Claude Poirier presented the present status and improvements of K-BASE (knowledge base on statistical data editing), which can be found on

http://amrads.jrc.cec.eu.int/k-base/

Among other achievements, electronic versions of "Statistical Data Editing" Volumes 1 and 2 are available on K-base. Feedback on K-BASE may be sent to poircla@statcan.ca .

- - - - -