30 May 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

REPORT OF THE APRIL 2008 WORK SESSION ON STATISTICAL DATA EDITING

Prepared by the UNECE secretariat

1. The Work Session on Statistical Data Editing was held in Vienna, Austria, from 21 to 23 April 2008 at the invitation of Statistics Austria. It was attended by participants from: Austria, Belgium, Canada, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, Japan, Lithuania, Netherlands, New Zealand, Norway, Republic of Korea, Slovenia, Spain, Sweden, Switzerland, United Kingdom, and the United States of America. Representatives of the United Nations Industrial Development Organization (UNIDO), United Nations Office on Drugs and Crime (UNODC) and the Food and Agriculture Organization of the United Nations (FAO) also attended.

2. The agenda contained the following substantive topics:
   (i) Editing of data acquired through electronic data collection;
   (ii) Editing administrative data and combined sources;
   (iii) Improvement of quality through data editing;
   (iv) New and emerging methods;
   (v) Editing based on results (post-editing);
   (vi) Censuses.

3. Mr. Reinhold Schwarzl, Deputy Director General of Statistics Austria opened the meeting and welcomed the participants. He explained that Statistics Austria endeavours to build a bridge with the academic community to increase knowledge and improve the quality of their products, thereby ensuring a certain level of quality in their figures for users. Official statistics is faced with challenges such as speeding up the data production process, reducing response burden and delivering more metadata. Editing and imputation play a decisive role in this process. Data editing is an area of growing importance and international cooperation is very important for research. He thanked Mr. John Kovar, the Steering Group, the UNECE secretariat and the International Department of Statistics Austria for their roles in motivating colleagues in this work and in organizing the meeting. He thanked authors and presenters for their interesting contributions and acknowledged the importance of these meetings in allowing participants to network. He wished the participants fruitful discussions for the broad and ambitious programme they have before them.

4. Mr. John Kovar (Canada) acted as Chairman. The Chairman, in his introductory remarks, thanked Statistics Austria for their hospitality and the excellent work conditions for the participants.

5. The following persons acted as Discussants/Session Organizers: Topic (i) – Mr. Pedro Revilla (Spain) and Ms. Paula Weir (United States of America); Topic (ii) – Ms. Natalie Shlomo (University of Southampton) and Ms. Vera Costa (New Zealand); Topic (iii) - Ms. Orietta Luzi (Italy) and Mr. Dan Hedlin (Sweden); Topic (iv) - Messrs. Jeroen Pannekoek and Ton de Waal (Netherlands); Topic (v) - Ms. Maria Garcia (United States of America) and Mr. Daniel Kilchmann (Switzerland); Topic (vi) – Ms. Heather Wagstaff (United Kingdom) and Mr. Thomas Burg (Austria).

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. Philippe Brion (France), Elmar Wein (Germany), Jeffrey Hoogland (Netherlands), Rudi Seljak (Slovenia) and Dale Atkinson (United States). 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 there are many issues that would deserve consideration at an international forum like the present Work Session. They recommended, therefore, that a future meeting on statistical data editing be convened in about 18 months time, subject to the approval of the Conference of European Statisticians and its Bureau.

8. The following substantive topics were recommended for the study programme of the future work sessions:

(i) Automated editing and imputation and software applications (possible contributions by: Canada, Germany, Norway, Austria, United Kingdom, Slovenia)
(ii) Editing near the source (possible contributions by United Kingdom, Slovenia, Spain, Norway, United States of America)
(iii) E/I administrative and census data (possible contributions by: Canada, Norway, New Zealand, France, United Kingdom, Spain)
(iv) Best practices for enterprise/population statistics (possible contributions by: Austria, France, Finland)
(v) Successful strategies for implementing new E/I- methods: (possible contributions by: Canada, New Zealand, France)
(vi) New and emerging methods (possible contributions by: Finland, Norway, Germany, United States)
(vii) Indicators for measuring the quality impact of data editing and imputation: (possible contributions offered Austria, New Zealand, United Kingdom, Slovenia, United States)
(viii) Selective and macro editing (possible contributions offered by Italy, USA, Spain)

9. The Knowledge Base on Statistical Data Editing (K-Base) is hosted on the UNECE website: (www.unece.org/stats/k-base). The future maintenance of K-Base needs an editorial team. Volunteers wishing to serve on the team are invited to express their interest. The suggestion was made to maintain K-Base in a wiki mode, with a free read access and password protected editing access. A reviewing team may be also envisaged.

10. The delegation of Switzerland offered to host the next meeting on statistical data editing during the second half of 2009.

FURTHER INFORMATION

11. 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/2008.04.sde.htm**).

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

ADOPTION OF THE REPORT

13. 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 of data acquired through electronic data collection

Discussants: Pedro Revilla, Spain and Paula Weir, United States of America
Documentation: Invited papers by Netherlands and United States of America; Supporting papers by United Kingdom, Canada, Norway and Spain.

1. The issues considered under this topic covered population censuses and business surveys. The presentations brought up the following points:

- While electronic data collection provides solutions to some problems of traditional data collection modes, it can create some new problems.
- Electronic data collection generally has fewer edit failures post submission as compared to other modes of data collection. There is also a visible decrease in item non-response. Experience shows that the original electronic responses have a higher error rate, but errors are corrected before submission. In general, it is believed that electronic data collection results in data of a better quality than paper questionnaires.
- The presentations also touched on the issues of defining the point at which to invoke an edit.
- It was suggested that a common look and feel might increase the familiarity of respondents with electronic questionnaires for repeated business surveys, and ultimately lead to an improved quality.
- It was suggested to perform usability testing prior to implementing electronic data collection, to identify issues that depend on the behaviour of respondents.

2. The participants discussed how to reconcile the differences between the edits and the edit process in electronic collection versus paper collection to reduce mode effects introduced in response and processing. The following opinions were expressed:

- Some countries tried to make the electronic questionnaires very similar to paper questionnaires, with practically no edits. However, these countries also plan to move cautiously towards more built-in edits in the future.
- Electronic questionnaires need more experimentation in order to learn more about mode effects and finding ways to limit them.
- If mode effects of the electronic data collection result in a better quality of data, it does not need to be viewed as a problem.
- In the case where there is a significant mode effect, the data should then be imputed separately for electronic and paper questionnaires based only on data sets obtained through a respective data collection mode.

3. In discussing the problem that the objectives of improving data quality and improving response rates in electronic collection appear to be in conflict, the following points were made:

- Automated editing, using the respondent's previous data and keeping the visible editing burden low was recommended as a way of promoting electronic data reporting.
- Several participants warned against building too much editing into the electronic data collection, but also to be careful not to over-edit. Otherwise respondents may feel frustrated and resist a further increase in electronic data reporting.
- Experiences showed that, at present, Internet data collection does not always represent significant cost savings. It is seen rather as a mode for obtaining higher data quality.

4. The participants discussed whether and how are common/standard edit types available in generalized edit system being implemented in the electronic collection systems. In this connection they also discussed the past period data, and the following points were raised.

- To some participants it appears that data editing rules implemented in electronic questionnaires should be simpler than those used in post editing.

- In using the previously reported data, it is important to make sure not to feed to respondents the imputed data. Several offices have experiences and also concerns related to the use of previously reported data.
- There was also a warning that messages, automated skips, and radio buttons that allow only a single response, promote a perception of intelligence that can lead to unwanted behaviour by the respondent, such as providing an answer that is really unknown and therefore, is incorrect but "valid" in the sense that it passes the edits.
- This issue also makes a case for more usability testing in order to learn more about respondents' perception and behaviour.

5. The participants also discussed metrics that should be maintained for electronic collection to determine how much editing to perform within the collection instrument to balance data quality with response burden and risk of non-response:

- In order to evaluate the editing process, at least all cases when an edit fails should be recorded.
- It was suggested keeping a complete history of all answers, and this approach was actually used in one practical application. This can also be useful in making the application more friendly by addressing areas of common difficulties.

6. The discussion also reviewed complex coding frames in on-line interfaces. Traditional data collection modes face problems with coding related to consistency between different coders and of a single coder over all questionnaires. These problems can be even more amplified when the coding is done directly by respondents.

7. The discussants also identified other questions that need further consideration:

- What are the ways to define and use metadata on feedback from respondents to improve the electronic questionnaire design and the edits in the electronic questionnaire, as well as those used for the paper questionnaires?
- Do respondents report less accurately when filling in the entries with no associated validation rules in an electronic questionnaire?

II. Editing administrative data and combined sources

Discussants: Natalie Shlomo, University of Southampton and Vera Costa, New Zealand
Documentation: Invited papers by Italy, New Zealand and Norway; Supporting papers by Austria, France, Italy, Netherlands and Norway.

8. Statistical Offices rely on administrative data to improve the quality of statistics, reduce costs and response burden. Administrative data are not originally designed for use as statistical data and need to undergo extensive processing and editing. In recent years, more emphasis has been put on the use of tax data to for business statistics and register data to for social and economic data. Combining multiple sources of data presents new challenges: ensuring quality in line with statistical standards and coherence across different sources.

9. There is growing pressure on statistical offices to move from using administrative registers data as auxiliary towards administrative registers as a source of data. Administrative registers and records are created for non-statistical purposes. Therefore, the quality of data obtained from administrative sources may not be of a desirable level. Presentations considered methods for adjusting data from administrative sources to statistical use:

- Use of time series model was suggested for creating estimates in order to overcome seasonal effects. It was stressed that time series methods for macro editing on aggregated administrative data may have their advantages, but also disadvantages.
- Availability of metadata and metadata systems is important for improving quality of data from administrative sources. In particular, metadata systems should record changes in concepts and definitions.

- Use of effective editing and imputation strategies was suggested for the construction of quality statistical databases. Such strategies should ensure correct coverage, consistent and clean records.
- Most methods recommended comprised a mass imputation instead of weighting, micro edits and selective editing based on scores. Weight adjustment through benchmarking was also recommended.
- There is a need to consider practicality and feasibility for large scale production systems when analyzing imputation methods.

10. Issues relevant to data from administrative sources that lead to the need to use imputation are the following:

- Timeliness of administrative sources is not always in line with the needs of the statistical business process cycle.
- Many administrative sources generally provide only annual data, while there is a need for monthly data. This is partially addressed by the existence of VAT data, but these are often not sufficiently timely.
- Businesses often provide only data for a higher aggregation level, such as the enterprise level. There is a need for an establishment level.
- The administrative sources do not have a desirable disaggregation level. For example the administrative sources may not cover all units of interest to statisticians (e.g. only legal units); it may not cover all activities, etc.
- Depending on the legislative settings, data may be available from the administrative source only for enterprises above a certain threshold.

11. The use of administrative registers and records leads to integration of multiple sources – multiple administrative sources, or administrative registers with surveys.

- Quality considerations for integrated data sources should put the emphasis on errors in linking data and inconsistencies in addition to quality considerations for single source data.
- There are advantages and disadvantages in incorporating administrative data at different stages of the survey process.
- Multiple administrative registers may be used to construct a single statistical registers. For example, statistical business registers may be constructed from the legal business register, tax register, social security register and various industrial chamber registries. While this is usually the best method to obtain comprehensive information, this creates the problem of inconsistent units and timeliness between different administrative sources.

12. When discussing editing of administrative data, the participants made the following points:

- Organizational issues are important. One of them is timeliness. It is impossible to call back an enterprise after a long period.
- Some statistical offices try to overcome the timeliness issues by using forecasts as preliminary values. This method anticipates stability and having a sufficient amount of past data to allow forecasting. However, it was stressed that such stability is unlikely for large enterprises that undergo permanent restructurings, acquisitions and sales.
- Keeping track of changes in administrative sources is difficult. This represents a difficulty for constructing longitudinal time series. One possible solution might be a metadata system within the statistical office that comprises information about reporting units and their changes in time.
- The quality of information obtained from administrative sources, also depends on the relationship with the administration(s) concerned.
- Automation of macro editing when a large number of series are produced is used by some statistical offices.

13. The following points were made in the discussion on assessing quality:

- It seems that the situation is easier when combining registers with surveys. Purely register-based statistics are not sufficiently transparent to allow recognizing systematic errors.
- Sample surveys may help to estimate the quality of administrative sources, and determine the future editing and imputation approach to data originating from individual administrative registers.

14. The last part of the general discussion dealt with mass imputation, and there were two contradictory opinions expressed:

- Some participants warned of mass imputation created microdata that (i) will be analyzed by econometricians at a detailed microlevel for which it was not designed, and (ii) aggregated at levels and details for which the mass imputation did not control for. The main problem is that mass imputed data give a false sense of accuracy/quality particularly because of the injected, but unrealistic, level of consistency.
- Others defended the principle that mass imputation may help maintain the consistency of data obtained through administrative and/or combined data sources, and more opportunities for statistical analysis by researchers due to the availability of a data set with complete unit record data. This position was defended mostly by countries with a tradition in register-based statistics.
- In concluding the discussion, participants agreed that this issue requires further work.

III. Improvement of quality through data editing

Discussants: Orietta Luzi, Italy and Dan Hedlin, Sweden
Documentation: Invited papers by Germany, and New Zealand; Supporting papers by: Italy, Switzerland, Finland and United Kingdom

15. The presentations on this topic covered the following themes:

- Monitoring and improving the statistical business process;
- Balancing quality components;
- Developing supporting tools (statistics, recommended practices, information systems, etc.);
- Automatic editing of unit errors.

16. The participants discussed the usability of corrections for improving quality. The following observations were made:

- Relative corrections provide very helpful information concerning the improvement of a questionnaire. This information indicates an under- or over-coverage during the data collection process. Relative corrections are defined as an amount of change between the raw and the plausible value.
- Relative correction can be easily calculated and explained. It can deliver hints concerning measurement issues. It can be used for the optimization of the data editing process.
- Relative corrections may be used for distributing costs of the editing process to variables.
- Information based on relative corrections should be complemented by additional information on the underlying data editing process.

17. The creation of a data editing framework within a statistical office is a necessary pre-condition for a sustainable improvement of data quality through the editing and imputation process:

- The objective of data editing and imputation is to provide users with plausible data and information on data quality, as well as to improve the end-to-end statistical business processes.
- The editing and imputation strategy should cover the business plan, standards and guidelines and training.
- The governance of editing and imputation projects should include methodologists as well as senior managers and subject-matter statisticians.
- The cultural issues are particularly important. They can be approached through a communication strategy aimed at convincing subject-matter statisticians about the value that the editing and imputation process brings to improved data quality. There is a need to provide results in a relatively short time. Selective editing may help avoid the perception of a huge editing and imputation process.
- In addition to convincing subject-matter statisticians, it is necessary to provide them with appropriate guidelines indicating tasks to be performed and powerful IT tools. Otherwise, editing and imputation risks staying at the level of a theoretical discussion.

- It was recommended that only validated methods that fit into the business architecture should be used. However, the question was raised what "validated" means, because from the scientific viewpoint most available methods were already validated. The applicability of a method to a statistical business process depends on the concrete context. It is necessary to preserve the possibility for innovative approaches, and the ultimate goal should be to use the best methods.

18. The work session also discussed reports on the implementation of the European Union's project on Editing and Imputation in Cross-sectional Business Surveys (EDIMBUS) and the following points were made:

- The Recommended Practices Manual developed within the framework of this project is a valuable basis for developing a statistical data processing framework. The manual contains guidelines and a list of standard indicators.
- The future work of the EDIMBUS project will focus on:
    - Dissemination of the Recommended Practices Manual within the European Statistical System and at the national level;
    - Promotion of the implementation of the Recommended Practices Manual within pilot surveys;
    - Training on editing and imputation methodologies and processes and on principles defined in the manual;
    - Development, by individual national statistical offices, of high-level strategies for editing and imputation in business surveys;
    - Periodic update of the Recommended Practices Manual by expert groups at appropriate levels.
- The published manual should be considered as a starting point. It should be tailored to the needs of individual statistical offices as part of the implementation process.
- The electronic version of the Recommended Practices Manual is available on the Internet at http://edimbus.istat.it .

19. The general discussion brought up the issue of understanding who the users are and what their needs are:

- It is necessary to specify who the users are with respect to editing and imputation. While this term is frequently used, its meaning is very wide, as it can cover the subject-matter statisticians validating the data and applying the editing and imputation methods. Other categories of users are final users of official statistics who need information (metadata) on quality of released statistics. However, respondents and other data providers are also users in a specific context.
- In general, users need good documentation and sufficient metadata. The documentation should be created at various appropriate levels, such as the European Statistical Systems, national level or internal documentation of a single statistical office.
- Understanding the users' needs would help to identify the minimum set of documentation so as to avoid an overload.

IV. New and emerging methods

Session Organizer: Jeroen Pannekoek, Netherlands
Discussant: Ton de Waal, Netherlands
Documentation: Papers by: Italy, University of Southampton/Netherlands, United Kingdom, Germany/University of Michigan, Austria, Netherlands, Spain and University of La Laguna, Spain

20. The purpose of this topic was to bring to the attention of participants innovative methods and techniques. Contributions highlighted progress in data editing theory and new methodologies, as well as new techniques and software tools, and in exploring organizational implementation issues and impacts. The papers presented under this topic covered a wide variety of subjects:

- Automatic editing (presentations by Spain, University of La Laguna and Netherlands);
- Statistical methods for error detection (a presentation by Italy);
- Impact of editing (a presentation by United Kingdom);

- Imputation (presentations by Austria, Germany/University of Michigan and University of Southampton/United Kingdom/Netherlands).

21. The discussion on automatic editing raised the following questions:
- Which errors are more important - random errors or systematic errors?
- Do we need new methodology to detect random errors based on edit rules or for statistical outlier detection?
- What are the experiences with respect to detecting and correcting systematic errors at other agencies?
- What kinds of systematic errors can we distinguish?

The following answers were suggested:
- Random errors may be less important, because they are likely not to cause a bias. However, the importance of random errors may be amplified when working with small domains. The impact of random errors on variance estimates should be further considered.
- Setting a threshold for how much to edit, is necessary when correcting random errors.
- Presence of systematic errors suggests that there are many similar errors in the data set. In this case the solution may be found in a corrective action with respect to the design of the questionnaire or other parts of the survey process, rather than continued automated correction of such errors.

22. Questions discussed concerning imputation included:
- Is imputation under edit restrictions an important new topic?
- Is imputation preserving known totals an important new topic?
- Is multiple imputation the way of the future to smooth variability of single imputations and to estimate (co)variances?

The discussion made the following suggestion:
- The time series model could be used for predictions. However, experiences by an office that has tried to use the time series method were not very promising. This may be due to the limited length of the available time series.

23. The meeting also reviewed the various methods of developing software such as open source, commercial off-the-shelf, finding a commercial vendor to develop software for us, developing software together in international projects, agencies developing their own software. The following issues were raised during the discussion:
- It is not very likely that commercial vendors will produce software covering all the needs of official statistics. In particular the editing and imputation market is not large enough to be commercially attractive.
- Ideally the software should be shared in the Open Source Software (OSS) mode. OSS would facilitate the portability of the software and components between different offices and different IT platforms. Unfortunately, there are legal issues in some countries that prevent statistical offices from making the in-house developed software available as open source. By contrast, other countries oblige government agencies to produce all software with open source.
- A Task Force was created under the auspices of the UNECE/Eurostat/OECD Steering Group on Management of Statistical Information Systems (MSIS) aiming at identifying the scope and models for sharing statistical software between statistical offices. The Task Force comprises representatives from the following countries and organizations: Canada, Italy, Netherlands, Norway, the United Kingdom, Eurostat, UNIDO and UNECE. Mr. Marton Vucsan (Netherlands, mvcn@cbs.nl) is the Coordinator of the Task Force. The UNECE serves as the Secretariat of the Task Force (steven.vale@unece.org).

24. The discussants also raised the following questions:
- Concerning the impact of editing:
  - Should we focus more on measuring the impact of (changing) the edit strategy?
  - What are the experiences at other agencies in this respect?
  - Can we improve on the Snowdon-X approach?
- For data:

- Is the focus shifting to periodic data/time series?
- Is the focus shifting to longitudinal data?

V. Editing based on results (post-editing)

Discussants: Maria Garcia, United States of America and Daniel Kilchmann, Switzerland
Documentation: Invited papers by United States of America and Sweden; Supporting papers by Sweden (2) Finland and Spain

25. Papers presented under this topic covered the following issues:
- Macro editing;
- Selective editing;
- Post editing.

26. The participants discussed the use of score functions in selective editing:
- It was suggested that the use of global score functions may require standardizing variables with respect to robust estimators and variances.
- The covariance effect may be important when using multivariate score functions.
- Experiences show that in the case of the wholesale trade, the previous year value is not suitable for determining the plausible value, because enterprises react very dynamically to market demand.
- When samples change significantly between different instances of the survey, historical data do not exist for new units. The lack of historical data makes it difficult to define the score function.
- When using data from administrative sources for defining the score function (a distance between the value and the data from administrative sources) the score would be minimal for administrative data sources. However, when incorporating the use of administrative data in selective editing, high scores should also reveal possible errors in the administrative data. Some participants suggested defining score functions for administrative data sources through predictions based on statistical models.

27. The discussion brought up the following issues related to macro-editing:
- Macro-editing should be accompanied by other methods (e.g. graphical methods).
- Macrodata obtained from other agencies are often not accompanied by measures of accuracy which makes it difficult to incorporate them in the models. Use of lower level models may help target the problem.

28. Regarding post-editing:
- The already published results may change when applying post-editing, but this cannot be generalized, and some of the results may remain unchanged. This leads to possible revisions that need to be controlled.
- Changes in released data may also occur when data are used for a purpose that was not envisaged earlier.
- Recording of what was edited at previous stages would be very helpful in reconciling previously edited longitudinal data in recent data.

VI. Censuses

Discussants: Heather Wagstaff, United Kingdom and Thomas Burg, Austria
Documentation: Invited papers by Austria and Canada; Supporting papers by New Zealand, United Kingdom and United States of America.

29. Recent years have seen a movement away from conventional census taking and towards purely register-based censuses. Two key purposes for census taking are:

(i) to provide a solid basis for important political and/or economic decision making; and
(ii) to create a comprehensive set of reference data.

Hence, data editing is essential to ensure sufficient quality of the final outputs. This topic provided an overview of data cleaning methods and technologies used in the latest census round or planned for the near future.

30. The following points were made in the general discussion:

- The learning process using information from prior censuses seems to move us into the right direction. This might serve as a general line for developing editing and imputation strategies.
- Subject-matter statisticians should be provided with enough information concerning the data cleaning processes. This information would also be useful to the users of census figures.
- Considering the transition to register based censuses, there might be a need for further research to investigate the impact of record linkage to editing and imputation. Some experiences showed that in some areas the register-based approach works well, but there was a decrease of quality in other areas.
- It was suggested that ex-post studies (like census coverage surveys) can be a helpful tool for improving the processes and/or for evaluating the results. Some participants considered that editing and imputation should be applied to data from such surveys, like for other statistical activities.
- Linking census records to the present and past records may help in addressing the cluster non-response. However, some statistical offices are careful about linking census records, in order not to harm the perception of confidentiality.
- One example presented used a small scale labour force survey for imputing occupation into a register-based census. The value added of imputing the occupation was put into question by some non-European participants. However, the European Union regulation obliges reporting information on occupation combined with some other attributes.
- One country reported on their first experiences with the rolling census (about one fifth of the population surveyed every year). The first results are about to be released, and there should be follow-up work on assessing the quality.
