ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

REPORT OF THE SEPTEMBER 2006 WORK SESSION ON STATISTICAL DATA EDITING

Prepared by the UNECE secretariat

1. The Work Session on Statistical Data Editing was held in Bonn, Germany, from 25 to 27 September 2006, at the invitation of the German Federal Statistical Office. Participants from the following countries attended the meeting: Austria, Azerbaijan, Bulgaria, Canada, Croatia, Estonia, France, Germany, Hungary, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Poland, Romania, Slovenia, Spain, Sweden, Switzerland, 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 nearer the source;
   (ii) Editing data from multiple sources;
   (iii) Editing microdata for release;
   (iv) Macro-editing;
   (v) New and emerging methods.

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

4. Mr. Walter Radermacher, Vice President of the German Federal Statistical Office opened the meeting and welcomed participants. He emphasized the important contribution of data editing and imputation methods in improving data quality. Mr. Radermacher encouraged further research in editing of data originating from administrative registers and records and in ensuring coherence of data collected from combined data sources. He also highlighted the growing importance of electronic data capture on the respondent side and the consequences for data editing.

5. The following persons acted as Discussants/Session Organizers: Topic (i) - Pedro Revilla (Spain) and Heather Wagstaff (United Kingdom); Topic (ii) - Ton de Waal (Netherlands) and Thomas Burg (Austria); Topic (iii) - Natalie Shlomo (United Kingdom) and Rainer Lenz (Germany); Topic (iv) - Leopold Granquist, Dan Hedlin and Svein Nordbotten (Sweden); Topic (v) - Paula Weir and Maria Garcia (United States) and Orietta Luzi (Italy).

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 Manfred Ehling (Germany), Thomas Burg (Austria), Gunnar Arvidson (Sweden), Caren Tempelman (Netherlands), Vera Costa (New Zealand) and Dolores Lorca (Spain). When preparing the proposal, the working group took into account suggestions made by other participants in side discussions during the meeting.

7. Participants noted that there are many issues that would deserve consideration at an international forum like the present Work Session. They, therefore, recommended 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 session:

(i) Editing of data acquired through electronic data collection:
- Data editing by respondents;
- General guidelines for editing web surveys, e.g. balance between editing on the client side and the statistical agency side;
- Experiences from systems already in use (process data, information about respondents);
- Comparison of data collection modes;
- Editing data from decennial population and housing censuses;
- Possible contributions: Canada, Germany, Norway, Poland, Spain, United Kingdom, United States;

(ii) Editing administrative data and combined sources:
- Improvement of usability of business registers;
- Editing of combined sources;
- Results of ongoing projects;
- Editing large data sets (censuses)
- Possible contributions: Austria, Canada, Netherlands, Norway, Sweden;

(iii) Improvement of quality through data editing:
- Ways of documenting data editing (accessibility and clarity);
- Trade-off between accuracy and coherence when editing;
- Monitoring and improving the survey process through editing;
- Developments of recommended practices;
- Impact of cultural differences and training of statisticians and users;
- Quality indicators and metadata on editing processes (for users and producers);
- Impact of efficiency on data editing and imputation and the need for simplicity;
- Possible contributions: Austria, Germany, Italy, Netherlands, New Zealand, Spain;

(iv) New and emerging methods:
- Special estimation situations;
- Multiple imputation;
- New error localization methods;
- (Synthetic) microdata for testing and comparison of new methods and tools;

(v) Editing based on results (post-editing):
- Macro-editing;
- Graphical analysis;
- Selective data editing vs. microdata analysis (scientific community).

9. The delegation of Austria offered to host the next meeting on statistical data editing. The participants recommended that the next meeting be held in spring 2008.

FURTHER INFORMATION

10. 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/2006.09.sde.htm .

11. The participants expressed their great appreciation to the German Federal Statistical Office for hosting this meeting and providing excellent facilities for their work.

ADOPTION OF THE REPORT

12. 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 nearer the source

Discussants: Pedro Revilla, Spain and Heather Wagstaff, United Kingdom
Documentation: Invited papers by Germany, Italy and Spain; Supporting papers by United Kingdom

1. This topic reviewed strategies or methods aiming at moving editing closer to respondents. One of the enabling factors is the increased use of electronic reporting and data capture. Technological and methodological progress in this area offers the opportunity for re-engineering statistical production processes. As a result, respondents will be able to play a more active role in data editing.

2. The expected benefits and goals of embedding editing in the data capture phase are:
- Drawing benefits from new technologies (ICT tools and statistical methods);
- Decreased burden on respondents, mainly businesses, through seamless and full automation of editing processes;
- Increased efficiency of statistical systems;
- Improved data editing and imputation processes;
- Preventing non-sampling errors and improved consistency of data;
- Retaining the design and monitoring phases of data editing within statistical offices.

3. To optimize the workload burden and above-mentioned benefits, national statistical offices have to develop approaches in implementing respondent-side editing:
- It is necessary to adapt electronic (e.g. Web) questionnaires to respondents' habits;
- Statistical offices should offer incentives to respondents to encourage them to assume responsibility for the editing burden;
- How to encourage the use of electronic reporting as experience in some countries has shown that the willingness to use the electronic response option is still relatively low. Other experiences has shown that advanced preparation, marketing and incentives may increase the interest in Internet reporting;
- Selective editing may help in balancing the two targets: error reduction and work reduction. Examples presented at the meeting showed reduced editing of Web questionnaires from 67% to less than 10%, while the error remaining is in the order of 0.01;
- Ideally a dynamic method for editing rule parameters would allow relaxing editing rules so that less data would have to be edited. However, such a method has yet to be developed and pragmatic solutions are being used instead;
- Electronic response options can make the survey definitions available to the respondent. The question was raised as to how well these are understood by respondents and how their understanding can be improved;
- Similarly, it is important to ensure that the answers, notably textual answers, are understood by statisticians;
- In order to decrease the reporting burden, statistical data reporting usually represents part of a broader relationship between citizens/businesses and public services.

4. A specific example of data collection techniques, when data editing and imputation can be also embedded, is computer-assisted telephone and personal interviews (CATI/CAPI). The following points were made:
- The goal is to decrease the respondent's burden and make the interviewer's job easier;
- The lessons learned from CATI and CAPI can be applied to computer-assisted self interviewing (CASI);
- Depending on the circumstances of the survey, editing can be implemented more or less rigorously. The downside of the rigorous application of editing rules during the interview is that the interview may be interrupted without completion.

5. Performance measures evaluating the number and characteristics of errors that occurred during the interviews can be used for further adjustments to the data collection process. It was emphasized that statistical data editing is not only about detecting and correcting errors, but also about learning from them and improving statistical processes with a view to preventing further errors.

II. Editing data from multiple sources

Discussants: Ton de Waal (Netherlands) and Thomas Burg (Austria)
Documentation: Invited papers by Canada, France and Netherlands; Supporting papers by Canada, Norway and Sweden

6. The growing use of administrative registers and records as sources of data for statistical purposes is motivated by the commitment of public services to decrease the burden on persons and businesses. These data were not originally intended for statistical use. Moreover, it is necessary to combine data from multiple sources. This represents a new challenge for statistical data editing, namely ensuring quality in line with statistical standards and coherence across different sources. All papers presented under this topic outlined strategies of national statistical offices for editing data from administrative and multiple sources.

7. The following general issues concerning the use of administrative and multiple sources were brought up during the discussion:

- Linking often concerns not only different administrative sources, but also combining administrative sources and statistical survey data;
- Parallel editing and imputation processes are often applied on each of the data sources. Two approaches were discussed:
    - Re-engineering aiming at merging the two processes into a single one;
    - Further editing for consistency in the combined file.

  Combinations of the two approaches are also possible (editing the first of the two sources and combined file only);
- In order to preserve the reputation of statistical agencies, it is important to assure citizens that the whole process is managed by the statistical agency and used exclusively for statistical purposes;
- Combining the statistical survey with administrative sources raises a number of questions:
    - What is the difference in quality?
    - How to deal with the grey areas (e.g. when combining tax and survey data)?
- Metadata should comprise a sufficient description of procedures used for editing. This is important, for example, when owners of administrative registers question statistical results that differ from the content of their registers;
- The feedback provided to holders of administrative registers and records is treated differently:
    - Feedback, not violating the principle of confidentiality was found to be beneficial in some cases;
    - Other statistical offices do not provide any feedback, because they are not allowed to by law;
- It is unlikely that questions asked in registration procedures and statistical surveys would be harmonized, taking into account their different purposes. This can be resolved as long as there is a possibility of mapping the concepts between those used in surveys and registers.

8. At the more detailed level there are a number of problems that statisticians encounter in using administrative registers and records and combining them among themselves with statistical survey data. Some of these were mentioned during the discussion:

- In some countries, linking of records from multiple administrative sources is facilitated through individual identification numbers (businesses or physical persons). This depends on national legislation;
- The timing of different data sources may differ significantly;
- Different registers may have variable outliers and missing values. The weights for individual registers have to be defined according to quality considerations;
- There are also differences between registers that are maintained at the national and sub-national (regional, municipal, etc.) levels;

- Reporting units in different registers (e.g. tax registers and business registers) may be different, and this discrepancy has to be reconciled;
- Businesses may change their name and the register may change their identification numbers within the administrative records due to the changed ownership, while they represent the same entities from the statistical viewpoint.

9. Some of the requirements on editing data from multiple sources are contradictory and raise questions requiring further consideration, for example:

- There is a logical requirement for consistency of final data across all sources used. However, accuracy is compromised when the sources are combined and re-edited to ensure consistency. How to reconcile this and/or find an optimal compromise?

III. Editing microdata for release

Discussants/Session Organizers: Natalie Shlomo (United Kingdom) and Rainer Lenz (Germany)
Documentation: Invited papers by the Netherlands/University of Southampton, Norway and Germany; Supporting papers by the Netherlands

10. The focus of this topic is on post-editing and imputation of microdata prior to its release. This is a relatively new topic, in the context of editing and imputation and was discussed at the May 2005 Work Session on Statistical Data Editing in Ottawa. In the light of present discussions on confidentiality aspects of statistical microdata[1] it gains interest and importance.

11. The presentations and discussion focused on two major issues:

- Perturbing microdata prior to its release in order to protect the confidentiality of statistical units and its impact on the edit and imputation strategies;
- Reconciliation of microdata when it undergoes further processing for different purposes other than its original intent.

12. Statistical disclosure techniques do not usually take edit constraints into account. Standard perturbation methods can be extended to do so. The following methods were discussed:

- Additive noise for a single variable using correlated noise;
- Additive noise for multiple variables and linear programming;
- Additive noise for multiple variables using correlated noise;
- Micro-aggregation, additive noise and linear programming;
- Rounding based on random rounding and preserving totals;
- Rank swapping and minimizing bias;

The choice of a method to protect data is subject to concrete circumstances, according to the nature of data and needs of users. It must also provide protection with a tolerable risk of disclosure according to the nature of the data and the statistical disclosure policies of the statistical offices.

13. The following general issues concerning editing and imputation of microdata for release were brought up at the Work Session:

- From the viewpoint of the statistical office, there is a need to aim for coherence and consistency across the chain of related sub-processes for processing microdata, often undertaken by different units;
- National accounts, in particular quarterly national accounts, have some editing and imputation experiences that may offer lessons for other short-term statistics;
- Some experts consider issues related to microdata in a broader context of editing statistical products for release;
- Synthetic data accessed over the Web can be used for prior exploratory analysis, before researchers have access to raw data in research data centres, since researchers can perform only limited operations that are in-line with their approved research plan.

[1] For example, in the context of the Guidelines Managing Statistical Confidentiality and Microdata Access that are being drafted by the Task Force established by the Conference of European Statisticians.

14. Microdata originating from administrative registers and from multiple data sources raise new issues that should be taken into account when editing such data:

- The following methods for protecting confidentiality take into account editing and imputation issues:
    - Standard perturbative methods for continuous variables;
    - Fully synthetic data;
    - Selective multiple imputation of key variables;
- If multiple imputation is used (for example to create a synthetic data set), then it is necessary to convince users that they have to use multiply imputed data sets for using synthetic data, while they would usually prefer a single data set;
- Some further lessons can be learned from the experiences of the Nordic statistical offices such as linking a large number of data sources, data checking for inconsistencies and the use of paradata and metadata in the edit and imputation processes;
- The general e-government policies, in some countries, allow reporting each information item only once. As a result, data editing and imputation methods have to be adapted to the availability of a single data source;
- Original registers and survey records can often be linked requiring editing and imputation to correct for inconsistencies. This represents a confidentiality threat and requires anonymisation before release.

IV. Macro-editing

Discussants/Session Organizers: Leopold Granquist, Dan Hedlin and Svein Nordbotten (Sweden)
Documentation: Invited papers by Canada, France and Spain; Supporting papers by United States

15. This topic reviewed selective editing. The aim is reducing manual follow-up work while maintaining quality in statistical results. This reduction can be achieved by prioritizing responding units according to their contribution to the aggregated statistics. Some of the experiences presented combined the selective editing with other methods.

16. The following general issues were discussed:

- Selective editing can be applied to any process requiring significant resources (editing, non-response follow-up, etc.)
- The question was raised of how to evaluate/compare the effectiveness of different score functions? Results and experiences from selective editing can be used to improve the score functions;
- In some applications, the importance of different variables was reflected in parameters of the score function;
- Some experiences showed that applying selective editing resulted in higher variances;
- Selective editing has to be viewed in relation to the estimates published. In particular, when publication cells are grouped in order to decrease their number, such a grouping influences the parameters of the selective editing;
- In foreign trade statistics the focus is on editing unit prices. The problem is that many commodity codes, delivered by the respondents are not on the list of valid codes, meaning that the codes (commodities) are certainly wrong. To improve the quality, the focus should be on improving the coding process.

17. The relationship between selective editing, micro editing and macro editing was discussed:

- There was an understanding that for macro editing, all data have to be available, while selective editing can be applied as soon as the first record becomes available;
- In some examples, micro editing was applied on all records, before subjecting the data set to editing;
- The main goal is to make the whole editing and imputation process more efficient, so the distinction between different types of editing is not so important;
- Data may be used for detailed analysis (e.g. macroeconomic research). In such cases the impact of selective editing on the quality of microdata requires further consideration. This may lead to the requirement to edit and impute data for all reporting units. Automated editing may be

applied on units that were not selected for a manual follow-up in order to restore consistency. However, care should be exercised not to apply automatic imputation blindly simply to make data conform to preconceived models.

18. The following outline of a selective editing procedure could be drawn from the presentations:
- Form a local score (freshness of data, prediction);
- Form a global score (reporting unit level)
  - The global score gives a mean for prioritization;
- Apply some selection mechanism. This raises the following questions:
  - How?
  - In what groups?
  - What should the global score reflect?

19. Experiences with concrete methods for selective editing were discussed:
- "Estimate-related" method;
- "Edit-related" method;
- Hidiroglou-Berthelot method and other methods.
  - What records to flag;
  - Effect of changes on totals.

V. New and emerging methods

Discussants: Paula Weir and Maria Garcia (United States) and Orietta Luzi (Italy)
Documentation: Invited papers by Italy/Netherlands/Switzerland, Italy, United States; Supporting papers by Israel, Netherlands, Spain, Switzerland and United States

20. Three main themes were discussed in the papers submitted under this topic:
- Current approaches and best practices at statistical agencies:
  - EDIMBUS project (editing and imputation in business statistics);
  - Outlier detection at the Swiss Federal Statistical Office;
- New approaches for editing and imputation:
  - Graphical editing;
  - Geospatial editing;
  - TEIDE (Techniques for Editing and Imputation of Statistical Data);
  - Editing and imputation for the US Census of agriculture (addressing less successful aspects);
- New imputation methodologies:
  - Predictive mean matching using GMM;
  - Regression imputation with linear equality constraints;
  - Sequential regression approach.

21. Two software presentations were also made for participants:
- TEIDE – Techniques for Editing and Imputation of Statistical Data, by University La Laguna, Tenerife, Spain;
- A graphical view for editing through E-Sphere, by the US Department of Energy, United States

22. The following issues were raised in the general discussion:
- To detect the outlier, it is necessary to use the whole series. If the outlier is not properly captured, it is difficult/impossible to detect the next outlier;
- One of the main concerns is the balance between costs and benefits. This requires refraining from editing too much, while maintaining quality at an acceptable level. In some cases, managers like to review all data, and statistical measures are often not used to optimize the editing and imputation. In other cases, such as in managing census field operations, the managers are concerned about the high costs;
- Timeliness is another very important factor that has to be taken into account in addition to quality and costs, in order to align editing with release calendars for estimates;

- Technology improvements are an enabling factor for new editing and imputation solutions. However, they cannot be left to the IT specialists, and experts in editing and imputation should be involved;
- The concept of quality needs to be clearly defined, because it is not unanimous among clients.

23. A specific issue arising in the discussion was the effect of new methods and techniques and their impact on daily practices in national statistical offices:

- Some participants felt that automation is not used to its full extent, and manual editing is still being used. It is, therefore, important to promote the new methods and techniques, so that they are accepted by other subject-matter experts and managers;
- It is important to pass on the data issues and solutions discovered in the editing and imputation to the collection process. The aim is to eliminate or at least reduce errors, and not just to correct them. Changing the data collection systems is a costly exercise, and therefore, the recommendations emanating from editing and imputation should be prioritized;
- Some subject-matter experts feel that by using advanced editing and imputation techniques, they lose the direct contact with respondents/clients.

* * * * *