

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Neuchâtel, Switzerland, 5-7 October 2009)

**REPORT OF THE OCTOBER 2009 WORK SESSION ON STATISTICAL DATA EDITING**

**Prepared by the UNECE secretariat**

1. The Work Session on Statistical Data Editing was held in Neuchâtel, Switzerland, from 5 to 7 October 2009 at the invitation of the Swiss Federal Statistical Office. It was attended by participants from: Australia, Austria, Canada, Finland, France, Germany, Hungary, Israel, Italy, Mexico, Netherlands, New Zealand, Norway, Republic of Korea, Russian Federation, Serbia, Slovenia, Spain, Sweden, Switzerland, United Kingdom, and the United States of America. The European Community was represented by Eurostat. Representatives of the United Nations Industrial Development Organization (UNIDO) and the European Central Bank (ECB) also attended.
2. The agenda contained the following substantive topics:
  - (i) Automated editing and imputation and software applications;
  - (ii) Editing near the source;
  - (iii) Editing and imputation of administrative and census data;
  - (v) Successful strategies for implementing new editing and imputation methods;
  - (vi) New and emerging methods;
  - (vii) Indicators for measuring the quality impact of data editing and imputation;
  - (viii) Selective and macro editing.
3. Mr. Philippe Eichenberger, Head, Statistical Methods Unit of the Swiss Federal Statistical Office opened the meeting and welcomed the participants. Switzerland is pleased to be able to host this meeting. He highlighted the importance of the work of data editing which could help make processes more efficient. The Swiss Federal Statistical Office has launched an ambitious project on statistical information systems the aim of which is to standardize and update statistical production processes. He thanked the Chairman and members of the Steering Group for all their work in organizing the meeting and wished participants a successful meeting.
4. Mr. John Kovar (Canada) acted as Chairman. He thanked the Swiss Federal Statistical Office for their hospitality and creating excellent work conditions for the participants.
5. The following persons acted as Discussants/Session Organizers: Topic (i) – Messrs. Jeroen Pannekoek (Netherlands) and Rudi Seljak (Slovenia); Topic (ii) – Mr. Pedro Revilla (Spain); Topic (iii) – Ms. Heather Wagstaff (United Kingdom) and Ms. Maria Garcia (United States of America); Topic (v) – Mr. Thomas Burg (Austria); Topic (vi) – Ms. Natalie Shlomo (University of Southampton); Topic (vii) – Ms. Orietta Luzi (Italy) and Mr. Daniel Kilchmann (Switzerland); Topic (viii) – Ms. Paula Mason (United States of America) and Ms. Felipa Zabala (New Zealand).

## RECOMMENDATIONS FOR FUTURE WORK

6. At the last Work Session on Statistical Data Editing held in April 2008, the suggestion was made to update the Knowledge Base on Statistical Data Editing (K-Base). It has been transferred to a wiki, with public read access and password protected editing access. K-Base provides a repository for information and resources related to data editing. The UNECE secretariat gave a presentation on how it can now be maintained and further developed. A wiki is a website that allows the easy creation and editing of interlinked web pages (e.g. Wikipedia). They are used to create collaborative websites, in corporate intranets and in knowledge management systems. The UNECE currently has 22 live wikis using “Confluence” software.

7. If K-Base is to have a future, it will need new content (as well as updates to existing material). A new editorial team is also needed. Statistics New Zealand volunteered to help the UNECE secretariat with this work, and other volunteers wishing to serve on the team are welcome (contact Steven Vale or Diane Serikoff, [steven.vale@unece.org](mailto:steven.vale@unece.org); [diane.serikoff@unece.org](mailto:diane.serikoff@unece.org)).

8. The wiki approach provides an alternative to producing traditional paper publications, and can therefore serve as an output. Access to the wiki is via the following link: <http://www1.unece.org/stat/platform/display/kbase>. Participants should feel free to update K-base, passwords are available on request from the UNECE secretariat ([steven.vale@unece.org](mailto:steven.vale@unece.org)).

9. Participants discussed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Messrs. Eden Brinkley (Australia), Pauli Ollila (Finland), Philippe Brion (France), Sander Scholtus (Netherlands), and Peter Gidlund (Sweden). When preparing the proposal, the working group took into account suggestions made by other participants in side discussions during the meeting.

10. 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.

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

- (i) Editing of administrative and Census data (expressions of interest from: Canada, France, Slovenia, United States, United Kingdom, Norway, Italy, New Zealand, Netherlands, Austria, Germany)
  - Using administrative data to improve the data editing strategy, or the quality of the estimates;
  - Using the administrative data for direct substitution;
  - Editing of statistical registers and auxiliary information;
  - Advances in editing of population censuses.
- (ii) Editing for electronic collection (expressions of interest from: Canada)
  - Balancing editing at the source versus editing in the office;
  - Impacts of mode effects;
  - Editing and incorporating past responses;
  - Developments of software and tools.
- (iii) Macro editing methods (expressions of interest from: Spain, Netherlands, New Zealand, Canada, United Kingdom, France)
  - Data visualization methods;
  - Developing expected values for use in editing (e.g. for the global financial crisis);
  - Selective macro techniques;
  - Quality measures to support macro editing (e.g. relative standard error, imputation rate);
  - Software and tools to support macro editing;
  - Approaches to annual and sub-annual data.

- (iv) Micro editing – methods and software (expressions of interest from: Austria, Sweden)
  - (v) New and emerging methods
    - Multiple imputation;
    - Mass imputation;
    - Imputation from longitudinal collections; etc.
  - (vi) Changing organizational cultures (expressions of interest from: Finland, France, Canada, Eurostat)
    - Engaging top-level management and each of the levels below
    - Training and education to change the culture, both processes and methods
    - Evaluation of editing strategies – quality, cost and timeliness
    - Understanding our users and fitness for purpose.
  - (vii) Panel session – International collaboration (expressions of interest from: Canada, Austria, and possibly several international statistical organizations including the UNECE)
    - How can we share tools and methods more effectively?
    - Can we collaborate on tool development?
    - The role of international organizations: UNECE, Eurostat, OECD, World Bank etc.
    - Should we identify case studies (e.g. selective editing tools)?
    - Can we reactivate past initiatives?
12. The representative of the Statistical Office of the Republic of Slovenia offered to host the next meeting on statistical data editing during the spring of 2011.

#### **FURTHER INFORMATION**

13. 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/2009.10.sde.htm>).
14. The participants expressed their great appreciation to the Swiss Federal Statistical Office for hosting this meeting and providing excellent facilities for their work.

#### **ADOPTION OF THE REPORT**

15. 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. Automated editing and imputation and software applications**

**Discussants/Session Organizers:** Jeroen Pannekoek (Netherlands) and Rudi Seljak (Slovenia)

**Documentation:** Invited papers by Germany and Slovenia; Supporting papers by Germany, Netherlands, New Zealand, Norway and Spain.

1. The Discussants highlighted the complexity of the statistical data editing process as it contains many sub-processes with often multiple data sources and choices for models and algorithms. Designing effective software applications can be challenging. The presentations discussed tailor-made applications, generic systems applications, in several cases there is a clear move towards the latter. Pressures to reduce costs and response burdens increasingly require applications designed to work with administrative and combined data sources.
2. Points raised in the general discussion included:
  - Improvements to edit processes inevitably lead to breaks in time series – users may be reluctant to accept such breaks even if the data quality is improved. This requires careful management and clear explanation in the metadata.
  - Persuading data producers of the need for change takes time; a top-down approach may be helpful.
  - Immediate benefits of standardization of systems are mostly related to IT support. Quality related benefits may take longer to realize.
  - Deterministic imputation approaches may be more relevant in cases where there are many logical links between variables.
  - Unit-level feedback to administrative data suppliers about data quality is generally limited to quality issues discovered before the administrative data set is merged with other sources, in order to ensure statistical confidentiality. Some statistical organizations have permission from tax authorities to contact businesses directly with queries about data.
  - As systems become more complex, and particularly when they offer the user multiple options, there is a need to ensure users are suitably trained to operate them correctly.
  - The Teide 2 software has so far only been used with survey data within Spain. It has proved difficult to find suitable data sets to test it on, due to data confidentiality concerns. Several other countries have downloaded the software, but would need the documentation translated to be able to use it in practice. The author asked for support to do this translation and to continue improving Teide 2.
  - The general question on the future of the open source application was raised. Despite the big effort to encourage this kind of cooperation between countries, not many examples of usage of the open source has been detected so far.

#### **II. Editing near the source**

**Discussant/Session Organizer:** Pedro Revilla (Spain)

**Documentation:** Invited paper by United States of America; Supporting papers by Norway, Slovenia, Spain and United Kingdom.

3. The papers presented under this topic considered the challenges faced by statistical organizations and the fact that new technologies such as IT tools and statistical methods could greatly improve quality and efficiency in the editing processes by allowing editing closer to the point of data collection. Specific points included:
  - Strategies and methods to move editing to the point of data collection;
  - Editing of data acquired through electronic data reporting;
  - Editing multi-mode data collections;

- Editing strategies in new production processes.
4. During the discussion, participants asked for clarification on the following points:
- The use of decision logic tables to determine edits and their sequence;
  - Questionnaire testing and mode effect from a editing point of view;
  - Errors observed as a result of using scrolling drop-down menus where the bottom item was erroneously selected;
  - The extent to which web collection is replacing other methods;
  - The measurement of mode effects;
  - Whether donors for imputation should be taken only from the sub-set of units whose data were collected using the same mode;
  - Distinguishing between errors based on their level of severity;
  - Whether the same set of edits should be applied across all collection modes.

### **III. Editing and imputation of administrative and census data**

**Discussants/Session Organizers:** Heather Wagstaff (United Kingdom) and Maria Garcia (United States of America)

**Documentation:** Invited paper by Canada; Supporting papers by Israel, Netherlands and United Kingdom.

5. The aim of the session was to focus on methodological advances of editing and imputation techniques applied to both administrative and census data:
- using administrative data in the statistical production process raises significant challenges and opportunities; and
  - currently many NSIs are researching and testing editing and imputation methods for 2010 census round.
6. Other aspects that were highlighted in the papers were:
- Lessons learned from earlier censuses help to improve future processes (Canada);
  - Applying generalised systems provide opportunity to deal with complex issues (UK);
  - Integrating administrative data with census data leads to complex cleaning processes (Israel);
  - Importance of modelling to take account of differences in data definitions and periodicity when combining distinct data sources (UK, Netherlands).
  - Risk of dependence on suppliers of administrative data (Netherlands).
7. Points highlighted in the discussion included:
- Imputation should be controlled to prevent distorting numbers for special population groups that attract user interest.
  - Imputation and disclosure control each seek to achieve different goals.
  - There are significant benefits in using generalized systems such as CANCEIS as opposed to developing a bespoke system.
  - In order to maximize data quality, it may be preferable to use complete administrative data sets rather than focusing on a specific sub-sample.
  - Are survey data more reliable than administrative data as a basis for comparing quality?
  - The different logical approaches taken by different editing and imputation software (e.g. CANCEIS and Banff) depend on the purpose for which they were developed and is therefore acceptable to have more than one generalized system
  - There is the potential for significant benefits from the large-scale use of administrative data, such as tax returns, especially for smaller businesses.

## **V. Successful strategies for implementing new editing and imputation methods**

**Discussant/Session Organizer:** Thomas Burg (Austria)

**Documentation:** Invited paper by France and Germany; Supporting papers by Canada, United States of America, European Central Bank and Eurostat.

8. This session covered the planning editing process; evaluation of strategies, implementation of new methods, historical evolution, and localizing improvement potentials.
9. Specific issues developed in the papers were:
  - Combining and editing multiple-source data and the use of common variables for evaluation of coherence.
  - Application of multiple imputation in practice.
  - The evolution of editing and imputation systems and lessons learned along the way.
  - Practical ways to improve the edit process.
  - Issues for developing editing and validation strategies for intergovernmental organizations.
10. There was a lively discussion about the applicability and merits of multiple imputation methods compared to more traditional single imputation approaches.
  - Although multiple imputation deals with imputation variance, imputation bias may still be an important issue, depending on the quality of the model used.
  - If there is a problem with imputation for one variable, this can be magnified by sequential multiple imputation.
  - Multiple imputation may be more appropriate for data sets with many users, as it makes it easier for users to estimate imputation variance.
  - Multiple imputation has been difficult for very large data sets due to the storage space and computing power required, though this limitation is reducing as technology advances.
  - Multiple imputation is most appropriate for missing values, its use for imputing inconsistencies is more controversial.
  - It is a good idea to complete simple edits before starting multiple imputation.
  - Multiple imputation relies less on the assumption that donor data are similar to the missing data, compared to more traditional donor imputation methods.
  - Donor imputation is not always optimal for numeric variables, particularly in business surveys.
  - Multiple imputation of qualitative data poses particular difficulties.

## **VI. New and emerging methods**

**Discussant/Session Organizer:** Natalie Shlomo (University of Southampton, United Kingdom)

**Documentation:** Invited papers by Italy and Norway; Supporting papers by Austria, Italy, Netherlands and Netherlands/University of Southampton.

11. This session covered the use of model-based techniques in selective editing and the detection and treatment of outliers; new developments in the automation of edit and imputation processes; imputation under edit constraints and benchmarking; and specialized imputation techniques. Several presentations also included initial results from applying these methods in practice.
12. The following issues were raised during the discussion:
  - How to account for variance estimation under mass imputation for cells where there is no sample available. The sampling process generates the sampling error, so if there is no sample available, a model for the sample distribution is required.
  - The use of robust regression models as opposed to transforming the data to obtain normality assumptions. The use of robust regression in the contamination model described in the Italian paper

is not straightforward since the model assumes that the data are normally distributed. The Italian authors discussed future plans to extend their model-based approach to selective editing, by using more robust methods and methods that take into account more complex and multivariate models. The main goal is not to model the data but to model the expected errors. Through this model, thresholds can be established for determining influential data and outliers.

- The United Kingdom suggested that simpler methods for outlier treatments may work just as well. For Italy, however, simpler methods may not be satisfactory and the overall aim is to avoid subjective thresholds.
- The lack of harmonization in the selective editing techniques for European short-term business statistics is problematic for European statistical systems.
- Routines developed in R by the Netherlands for automatic editing and correcting of violated balance edits due to miscoding may be made available on an open-source basis. The flexibility of R makes it relatively easy to apply new functions and disseminate methods.
- The Netherlands proposed to test their new approaches for automatic editing on real data. Other necessary edit algorithms may be identified by studying the results.
- The framework proposed by Austria for imputation of compositional data could possibly be extended to ordinal variables, and other models were suggested for the imputation of numerical compositional data.
- Several countries (Norway, Netherlands and United Kingdom) proposed to collaborate on a study on the combination of different approaches for mass imputing a statistical dataset with a combination of numerical and categorical variables whilst preserving edit constraints and benchmarking totals.

## **VII. Indicators for measuring the quality of data editing and imputation**

**Discussants/Session Organizers:** Orietta Luzi (Italy) and Daniel Kilchmann (Switzerland)

**Documentation:** Invited papers by Canada and Italy; Supporting papers by Austria, Finland, Switzerland and United States of America.

13. Monitoring and documenting are crucial parts of data editing and imputation with the aim to assess the quality impact of these parts of the data processing phase and provide feedback to the other parts of the survey process. However, the identification of appropriate measurements in different contexts, taking into account the different users needs, is a complex task. Furthermore, the interpretation of statistical measurements and indicators may not be straightforward in every situation.

14. The discussion of this topic covered the following points:

- The availability of documentation and quality indicators to different groups of users, and how to monitor their use.
- The relationship between the proposed quality metadata and international standards.
- The use of an audit approach to monitor clerical edits.
- Graphical analysis of the impact of editing and imputation should be used to underpin the evaluation of E&I impact for numeric data.
- Making micro data available to external users raises questions about the impact of imputed data. How much information should be made available to these users. If they are given raw un-imputed data they may produce results which are quite different to official statistics.
- It is difficult to apply more than a minimal set of standard quality indicators given the variety of editing and imputation purposes, strategies and application contexts.
- How to measure and include user needs when defining quality indicators? It is necessary to categorize users taking into account their level of competence in statistics.
- Whether to retain edit rules that have little or no impact on the raw data - such rules may still be needed to control imputations.
- Ratios of values before and after edits are useful tools to help prioritize the investigation of possible errors in the editing process.
- How to measure improvements to data quality due to editing and imputation, and how to feed back quality issues to the collection process?

### **VIII. Selective and macro editing**

**Discussants/Session Organizers:** Paula Mason (United States of America) and Felipa Zabala (New Zealand)

**Documentation:** Invited papers by Austria, France and United Kingdom; Supporting papers by Canada, France, Germany, Sweden and United States of America.

15. This topic focused on various aspects of macro editing and selective editing techniques used to make editing and imputation processes more efficient while maintaining statistical output quality. These techniques may be applied to data prior to release, used to complement micro editing, or even applied after data release. Areas include, for example, the development, application or evaluation of local, global, or multivariate score functions, selection of data sources for defining score functions, effects of sample change on the use of score functions, and stochastic optimization in selective editing; aggregate outlier detection methods, incorporation of measures of accuracy or edit history in macro editing, and use of aggregate graphical or numerical based models in macro editing.

16. The main points from the discussion included:

- Selective editing for business surveys stratified by size needs to take sampling weights into account, as smaller businesses may have a high impact if their weights are high, and may have a higher probability of error.
  - Prioritizing follow-up activities can have an impact on response probabilities, and may subsequently undo sampling design work.
  - The use of robust estimates for aggregates.
  - Selective editing techniques were thought to be more suitable for datasets with uneven rather than even distributions.
  - It is important to incorporate feedback from processing clerks to refine selective editing processes, site visits can help to verify data edits and resolve issues with key contributors.
  - The measurement of editing bias is relatively easy on a comparative basis (between two methods), but it is not really possible to calculate the absolute editing bias.
  - The use of robust estimates to detect outliers.
-