**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE** 1 October 2015
**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Budapest, Hungary, 14-16 September 2015)

## REPORT OF THE WORKSHOP

1. The Work Session on Statistical Data Editing was held in Budapest, Hungary, from 14-16 September 2015. It was attended by representatives from the statistical offices of Austria, Bosnia and Herzegovina, Canada, Chile, Denmark, Finland, France, Germany, Hungary, Italy, Japan, Latvia, Lithuania, Netherlands, New Zealand, Norway, Russian Federation, Slovenia, Spain, Sweden, Switzerland, United Kingdom and United States as well as by representatives from the Eurasian Economic Commission and World Health Organisation (WHO).

2. Mr Steven Vale, Head of the UNECE Statistical Management and Modernisation Unit, opened the workshop and welcomed participants. Mr Claude Poirier (Statistics Canada) and Daniel Kilchmann (Swiss Federal Statistical Office) were elected as co-chairs of the Work Session.

3. The keynote presentation was given by Mr. Zoltan Vereczkei, Deputy Director of Methodology in the Hungarian Central Statistical Office. He outlined the importance of this meeting to exchange best practices in the area of the statistical editing and imputation and evaluation. It is very important to have more efficient practices, especially due to shrinking resources in national statistical organisations.

4. The agenda included the following substantive topics, the outcomes of which are documented in the annex:

Session 1 Software tools and international collaboration
Session 2 Managing and supporting changes related to editing and imputation
Session 3 Selective editing and macro editing
Session 4 Report of the Task Team on a Generic Process Framework for Statistical Data Editing
Session 5 Evaluation and feedback
Session 6 Emerging methods and data revolution

5. The following persons participated in the Organizing Committee and/or acted as Discussants/Session Organizers: Topic (i) – Pedro Revilla (Spain) and Emmanuel Gros (France); Topic (ii) – Marco Di Zio (Italy) and Claude Poirier (Canada); Topic (iii) – Rudi Seljak (Slovenia) and Claude Poirier (Canada); Topic (iv) – Daniel Kilchmann (Switzerland) and Sander Scholtus (Netherlnads); Topic (v) - Jeroen Pannekoek (Netherlands) and Li Chun Zhang (Norway); Topic (vi) - Emmanuel Gros (France), Marco Di Zio, Ugo Guarnera, Orietta Luzi (Italy), Saara Oinonen, Pauli Ollila, Marjo Pyy-Martikainen (Finland), Li Chun Zhang (Norway), Jeroen Pannekoek (Netherlands), Steven Vale and Tetyana Kolomiyets (UNECE).

6. All background documents and presentations for the work session are available at http://www.unece.org/index.php?id=37497#/

7. The future work group drew up a list of possible topics that was modified slightly during the discussion, resulting in the following proposal for the topics to be discussed at the next Work Session. The countries mentioned in brackets expressed a possible interest in contributing to these topics:

(i) Machine learning (New Zealand, France)

(ii) New and emerging methods (Netherlands, Italy)

(iii) Shared software tools and CSPA services (Netherlands, Germany, Austria, UK, Slovenia and Canada)
- Demonstrations and implementation experiences

(iv) New data sources (Canada, Netherlands, USA and Italy)
- Big Data, multi-source statistics

(v) Standards and international collaboration (Finland, Germany, Slovenia and Netherlands)
- Including implementation of the new and emerging standards: VTL, GSDEMs, CSPA

(vi) Census 2021 (Germany, Norway, Italy and Canada)

(vii) Managing change (Canada, Denmark, New Zealand, USA and Finland)

8. It was proposed to add the following work formats for the next Work Session in addition to the traditional presentations:

- Break-out sessions
  - The participants can discuss topics in small groups, then report back to the plenary session.

- Market posters and software demos
  - The participants can show the software that they are using and present posters on different topics.

- Lightning talks
  - 5 minutes per talk and slides rotate automatically 15 seconds per slide. Good to present many short presentations.

9. Mr. Sander Scholtus, on behalf of Statistics Netherlands, offered to host the next Work Session on Statistical Data Editing in The Hague in spring 2017.

## ADOPTION OF THE REPORT

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

11. The co-chair of the Work Session, Mr Claude Poirier, thanked Hungarian Central Statistical Office for the excellent facilities and organisation, the Organising Committee for preparing the content of the Work Session, the participants and paper authors for their contributions, and the UNECE Secretariat for their support.

**Annex: Summary of discussions on substantive topics**

## A. Session 1: Software tools and international collaboration

12. This topic was organized by Rudi Seljak (Slovenia) and Claude Poirier (Canada). It included the following presentations:

- ValiDat Towards a generic approach to validation: the ValiDat foundation project
- ValiDat The ValiDat foundation project: Survey on the different approaches to validation applied within the European Statistical System
- Austria - Flash estimates for Short Term indicators - data cleaning with X12 Arima
- Netherlands - A formal typology of data validation functions
- Hungary - Integrated data entry and validation system in HCSO
- Slovenia - Usage of external software tools at SURS - experiences and lessons learned so far
- Bosnia and Herzegovina - Editing and imputation in Household Based Surveys - case of Household Budget Survey in Bosnia and Herzegovina

13. The following points were raised in the discussions:

- The extent to which existing solutions can be wrapped to make them compliant with the Common Statistical Production Architecture (CSPA)
- The relative maturity of different software solutions offered. Software developed in collaborative projects is generally easier to implement as it is usually better documented
- The possibility to start software development in a bottom-up approach by collaborative coding in a language such as R. This can reduce the time needed for initial development, but a more formal approach is likely to be needed when the software is put into production
- Open-source software is preferable for sharing, but if it is very complex, it still may be difficult for others to modify
- The challenges of providing support to those who use your software. A community approach seems preferable
- The need for a test environment and test data for new software and upgrades. The UNECE sandbox is being expanded to provide this functionality
- The extent to which multiple sources and modes add complexity to the editing process. Modelling and machine learning techniques may provide solutions. Mapping models of different modes to a single logical model may also help, though data sources with limited metadata are difficult to model
- Although there are differences between business and social data during the design and collect phases, those differences are much less for the process phase

## B. Session 2: Managing and supporting changes related to E&I

14. This topic was organized by Marco Di Zio (Italy) and Claude Poirier (Canada). It included the following presentations:

- Canada - Redefining roles and responsibilities in a new harmonized statistical production process: opportunities and challenges
- Italy - Managing changes in the E&I strategy of the Italian structural business statistics
- Finland - Implementation of selective editing methods at Statistics Finland using innovative and efficient team work methods
- Hungary - Improvement of the quality of statistics by Mezo-validation
- Hungary - Data collection optimization – first attempt
- USA - Imputation at the National Agricultural Statistics Service
- New Zealand - Getting commitment to a new editing strategy

- UK – Managing and supporting changes related to editing and imputation in the United Kingdom

15. The discussants highlighted seven key issues from the papers and presentations:

- Transition phase
- Measuring benefits and costs
- Commitment from top-management
- Reference model for introducing innovation in editing and imputation (E&I)
- Spreading E&I culture
- A model for ‘design’ of E&I
- Emotional aspects

16. The following points were raised in the discussions:

- Change is increasingly seen as a continuous process rather than a one-off event. It is necessary to accept change as the norm
- Quality and efficiency impacts should be measured, but this is not easy, particularly during a change
- The changing role of methodologists in many organisations: Less direct involvement in statistical production, and more of a consultative role
- Whether changes in organisation and methodology can be applied to all subject-matter areas
- The need for a consistent vocabulary to communicate editing and imputation methods, building on the existing data editing glossary
- “Agile” project management techniques can help to drive change and engage colleagues
- The importance of comparing variables from multiple sources to improve data quality
- How to manage and communicate discontinuities caused by process changes
- It is important to understand the value added of editing. If this is high, it is easier to convince colleagues to change the editing procedure. The concepts of “fit for use” or “good enough” are useful for communication purposes
- It is difficult to define a generic transformation plan as all organisations are starting from different points with different constraints
- The impact of change on people (particularly methodologists) should not be ignored
- An atmosphere where people are encouraged to innovate, even to make some mistakes, with full support of top managers, is an important facilitator of change
- Good communication and marketing of change and its benefits are essential to convince colleagues
- Shared e-learning resources and wiki-based information on good practices would help to spread ideas and share innovation more effectively

**C. Session 3: Selective editing and macro editing**

17. This topic was organized by Pedro Revilla (Spain) and Emmanuel Gros (France). It included the following presentations:

- Italy - Selective editing of business investments by using administrative data as auxiliary information
- France - Output editing based on winsorization in the French structural business statistics multisource system Esane
- Spain - Developing a theoretical framework for selective editing based on modelling and optimization
- Netherlands - Changes in macro-editing and score functions for Dutch short-term statistics
- Italy - Model-based selective editing procedures for agricultural price indices

- USA - Selective and macro-editing of a large business based administrative data set
- UK - Method for reviewing selective editing thresholds at the Office for National Statistics, Retail Sales Inquiry pilot study

18. The following points were raised in the discussions:

- How to share selective editing software
- How to conduct clerical reviews of selective editing processes: Using surveys or re-contacting respondents. How to measure the costs and benefits of different approaches
- The need to review the process should be factored into the business case for selective editing, and can probably be funded from efficiency savings
- The impact of different types of systematic errors on selective editing
- The extent to which current selective editing techniques can be applied to administrative and other data sources
- Application of selective editing to classification variables: How to select units to edit
- Selective editing is not suitable for treating systematic errors, but can free resources to work on detecting and correcting such errors

**D. Session 4: Report of the Task Team on a Generic Process Framework for Statistical Data Editing**

19. This topic was organized  by Emmanuel Gros (France), Marco Di Zio, Ugo Guarnera, Orietta Luzi (Italy), Saara Oinonen, Pauli Ollila, Marjo Pyy-Martikainen (Finland), Li Chun Zhang (Norway), Jeroen Pannekoek (Netherlands), Steven Vale and Tetyana Kolomiyets (UNECE). It included a presentation by Task Team members on the Generic Statistical Data Editing Models (GSDEMs - version 0.5)

20. The following points were raised in the discussions:

- Participants congratulated Task Team and expressed their high appreciation for the GSDEMs
- There was general agreement with the categorization of products and methods presented, as well as on the proposed treatment of granularity of process steps
- There are no major omissions in the set of generic flow models presented
- In terminology the aim of the GSDEMs is to confirm with Generic Statistical Information Model (GSIM) and the Generic Statistical Business Process Model (GSBPM)
- A formal standard for modelling notation could be used for the flow models
- It was suggested to provide a few concrete examples of implementation of the GSDEMs
- The GSDEMs can be seen as a conceptual extension and revision of the Edimbus handbook
- The GSDEMs are useful for the communication between methodology and IT departments, and subject-matter experts
- Revisions to the GSDEMs will be coordinated by the Modernisation Committee on Production and Methods, and it will be up to the future sessions on SDE to review it. Proposed revision period is every 5 years
- It is planned to finalise GSDEMs and release version 1.0 by the end of 2015.

**E. Session 5: Evaluation and feedback**

21. This topic was organized Daniel Kilchmann (Switzerland) and Sander Scholtus (Netherlands). It included the following presentations:

- Netherlands - Editing Big Data: an holistic approach
- Finland - Editing process and its quality regarding design and production phases using process metadata and calculation modules

- Switzerland - Analysis of the data preparation process of the structural survey of the federal population census
- Norway - Editing and evaluation of statistics based on administrative microdata - example by Norway
- Germany - Evaluation of Census 2011 survey estimates
- France - Using the CURIOS algorithm to manage the prioritization of CAPI surveys

22. The following points were raised in the discussions:

- The applicability of the Big Data “noise and signal” approach to “small data”
- Whether filtering approaches can be used to treat intermittent errors – more work is needed
- The extent to which current quality indicators are useful in practice: They say little about process quality
- How to organise feedback from an edit process to improve the statistical process as a whole
- How to define mandatory indicators for editing and imputation
- What information about edit processes should be given to users of statistical data?
- What information about edit processes should be recorded to produce useful indicators
- Quality indicators should be transmitted alongside data; This may require enhancements to SDMX (Statistical Data and Metadata eXchange)
- Current quality indicators like those from Eurostat might not be applicable to administrative data or multisource surveys with administrative data as a component

### F. Session 6: Emerging methods and data revolution

23. This topic was organized Jeroen Pannekoek (Netherlands) and Li Chun Zhang (Norway). It included the following presentations:

- New Zealand - Let the data speak: Machine learning methods for data editing and imputation
- Italy - Estimation and editing for data from different sources. An approach based on latent class model.
- UK – An assessment of the feasibility of editing and imputing administrative tax return data to provide a substitute for survey data
- Japan - Multiple ratio imputation by the EMB algorithm
- Netherlands - New results on automatic editing using hard and soft edit rules

24. The following points were raised in the discussions:

- Whether machine learning methods are really so different to existing statistical methods: Differences in terminology may increase the perception of more substantive differences
- The dangers of applying new techniques in statistical production that are not fully understood or supported by existing methodologists
- It would be useful to share case studies on applying machine learning techniques, including both successes and failures
- A review of the information on machine learning compiled in the context of the Euredit project was proposed
- Are statistical organisations sufficiently confident to release statistics compiled purely using model-based estimation? This may depend on the uses of the statistics, and may not, in reality, be much different to extensive use of techniques such as regression or nearest-neighbour imputation
- Effective marketing and communication of statistics based on new methods are essential, and may help to catch the imagination of users
- There is a strategic trade-off between the cost savings associated with new sources and the additional expense of processing multi-source data. Some of the costs may not be easy to measure.