UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE 28 April 2017
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

REPORT OF THE WORK SESSION

1. The Work Session on Statistical Data Editing was held in The Hague, Netherlands, from 24-26 April 2017. It was attended by representatives from the statistical offices of Albania, Austria, Bosnia and Herzegovina, Canada, Denmark, Finland, France, Germany, Hungary, Israel, Italy, Kazakhstan, Malta, Mexico, Netherlands, Norway, Poland, Russian Federation, Serbia, Slovenia, Spain, Switzerland, United Kingdom and United States as well as by representatives from Eurostat and Tilburg University (Netherlands).

2. Ms Therese Lalor, Head of the UNECE Statistical Management and Modernisation Unit, opened the workshop and welcomed participants. Mr. Daniel Kilchmann (Swiss Federal Statistical Office) was elected as chair of the Work Session.

3. Mr Wim van Nunspeet (Senior Director of Resources, IT and Methodology) welcomed participants on behalf of Statistics Netherlands. He highlighted the importance of data editing in the modernization of official statistics, and in the context of new data sources. He wished participants a successful Work Session.

4. Mr Steven Vale (UNECE) presented an update on UNECE modernization initiatives, under the High-Level Group for the Modernisation of Official Statistics. Mr Barteld Braaksma (Netherlands) presented the new "Blue-Skies Thinking Network", an initiative to encourage collaborative innovation.

5. The agenda included the following substantive topics, the outcomes of which are documented in the annex:
   Topic 1, theme (i): New and emerging methods
   Topic 1, theme (ii): New data sources and Census
   Topic 2, theme (iv): Standards and international collaboration - Including implementation of the new and emerging standards: VTL, GSDEMs, CSPA
   Topic 2, theme (iii): Shared software tools and CSPA services - Demonstrations and implementation experiences
   Topic 2, theme (v): Managing change
   Topic 3, mini-sprint: How to foster the implementation, within statistical offices, of good practices from other organizations, as well as areas for international collaboration, and priorities for joint work.

6. The following persons participated in the Organizing Committee and acted as Discussants/Session Organizers:
   Topic 1, theme (i) – Li-Chun Zhang (Norway), Simona Rosati (Italy), Pedro Revilla (Spain) and Jeroen Pannekoek (Netherlands)
   Topic 1, theme (ii) – Alexander Kowarik (Austria) and Thomas Deroyon (France)
   Topic 2, theme (iv) – Sander Scholtus (Netherlands), Agnes Andics (Hungary) and Simona Rosati (Italy)
   Topic 2, theme (iii) Jeroen Pannekoek (Netherlands), Thomas Deroyon (France) and Pedro Revilla (Spain)
   Topic 2, theme (v) - Sander Scholtus (Netherlands) and Agnes Andics (Hungary)

Topic 3 - Alexander Kowarik (Austria), Daniel Kilchmann (Switzerland), Steven Vale (UNECE) and Li-Chun Zhang (Norway).

7. All background documents and presentations for the work session are available at: http://www1.unece.org/stat/platform/display/WSSDE/Work+Session+on+Statistical+Data+Editing+2017

8. The outcomes of the Work Session, including proposals for future work, are detailed in Section F of the Annex. In summary, three topics were proposed for international collaboration activities:

   - Method library/architecture: Classification and inventory of methods
   - Consolidation and update of methodological guidelines on statistical data editing
   - Improving understanding and communication of CSPA and other relevant standards

9. The following topics were proposed for discussion at a future Work Session on Statistical Data Editing:

   - Use of administrative data, including efficient error detection and data editing in the context of an integrated data set
   - Quality of editing and imputation processes
   - Editing and imputation processes for Big Data/Geospatial data
   - Standardisation of editing and imputation methods and systems (including practical demonstrations)
   - Reducing cost, and increasing efficiency of data editing
   - Development of new data editing applications and methods (including practical demonstrations)
   - Privacy and data access implications of data editing
   - Machine learning in the context of data editing

ADOPTION OF THE REPORT

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

11. The chair of the Work Session, Mr Daniel Kilchmann, thanked Statistics Netherlands 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. Topic I: New perspectives for data editing in the context of new data sources and data integration,
Theme (i): New and emerging methods

12. This theme was organized by Li-Chun Zhang (Norway), Simona Rosati (Italy), Pedro Revilla (Spain) and Jeroen Pannekoek (Netherlands). It included the following presentations:

- Netherlands - Correcting for misclassification under edit restrictions in combined survey-register data using Multiple Imputation Latent Class modelling (MILC)
- Netherlands - Simplifying constraints in data editing
- Austria - An automatic procedure for selecting weights in kNN imputation
- Netherlands - Imputation Methods Satisfying Constraints
- Netherlands - Evaluating the quality of business survey data before and after automatic editing
- Netherlands - Computational estimates of data-editing related variance
- France- A comparison of Kokic and Bell winsorisation and conditional bias methods for outlier treatment

13. In addition to several clarification questions, the following points were raised in the discussions:

- The extent to which algorithms and their implementations in languages such as R can be shared
- Possible extensions to some of the methods presented
- Conditions under which it is more appropriate to use a specific outlier treatment method
- How methods presented could be implemented in practice
- How important is it to have an accurate estimation of variance in the context of new techniques such as machine learning?

B. Topic 1, Theme (ii): New data sources and Census

14. This topic was organized by Alexander Kowarik (Austria) and Thomas Deroyon (France). It included the following presentations:

- Austria - Improving the census core topic occupation by using new administrative data sources
- Germany - Possible imputation procedures for the Census 2021
- Israel - Editing and Imputing Income Data in Integrated Census: lessons learned from the 2008 Census - toward the 2020 Census
- USA - Creating an Initial Donor Pool for New Questions in the Census of Agriculture

15. In addition to several clarification questions, the following points were raised in the discussions:

- The possibility of using parallel processing and Big Data tools to increase the speed of computationally heavy processes such as multiple imputation
- How to incorporate CSPA principles in new census processes
- The challenge of multiple values for the same variable when several data sources are used
- Whether censuses are “special” in terms of editing and imputation, or whether standard approaches can be used. The main feature of censuses is the size of the resulting data sets and the level of detail of outputs, which limit the application of methods used for other data sets
- The extent to which imputation error contributes towards total error for censuses

- Do we all have to process large datasets in our own environments, or could we pool resources to create a shared high specification environment for heavy processing? Confidentiality constraints would need to be addressed
- The challenges of moving from research to production environments

C. Topic (ii): International collaboration on standards and tools for data editing

Theme (iv): Standards and international collaboration - Including implementation of the new and emerging standards: VTL, GSDEMs, CSPA

16. This topic was organized by Sander Scholtus (Netherlands), Agnes Andics (Hungary) and Simona Rosati (Italy). It included the following presentations:

- Eurostat - The modernisation of validation in the ESS – a multidimensional approach
- Norway - Method library as part of the modernization of the statistical production in Norway
- Poland - Technical aspects of VTL to SQL translation
- Norway - A new GSDEM for multisource statistics
- Germany - Validation, Shared Services and Enterprise Architecture: How it fits
- Germany - The ESSnet Validat Integration

17. In addition to several clarification questions, the following points were raised in the discussions:

- Whether components presented can be shared
- Possibilities to translate validation rules between different languages such as VTL, R, SAS and SQL
- Possible extensions of the Generic Statistical Data Editing Models (GSDEMs) to cover statistical registers and other scenarios, and the extent to which they can encourage greater standardisation
- How to cover many to many relationships between units in the GSDEMs
- The extent to which re-use of tools and methods is actually happening so far
- The desirability of having multiple tools for a single function. In such cases, the best performing solutions are likely to become the most used, in a sort of "survival of the fittest" evolutionary scenario
- The different layers of the CSPA catalogue, accessible through the UNECE wiki platform, provide a central list of statistical services and other resources that are already available and planned
- The extent to which validation activities can be centralised, and whether there remains any need for domain-specific validation
- Purely open-source technology, and hands-on demonstrations of tools, are factors that can facilitate sharing

D. Theme (iii): Shared software tools and CSPA services - Demonstrations and implementation experiences

18. This topic was organized by Jeroen Pannekoek (Netherlands), Thomas Deroyon (France) and Pedro Revilla (Spain). It included the following presentations:

- Spain - Software implementation of optimization-based selective editing techniques at Statistics Spain (INE)
- Canada - Future Development of Statistics Canada's Edit and Imputation System Banff

- USA - Automated Data Editing and Imputation for Surveys of Multinational Enterprises, a Banff Implementation
- UK - The role of rules based editing for maintaining quality in the drive for efficiency
- Slovenia - General tool for macro editing at SURS
- Presentation of Edwin de Jonge (Netherlands) 'CSPA with R'

19. In addition to several clarification questions, the following points were raised in the discussions:

- There are some similarities with recent developments in the R community in terms of defining inputs, throughputs and outputs
- Some statistical organisations have a strategy to move from proprietary software tools, and towards open source platforms such as R
- Possible integration of VTL with data editing tools such as Banff
- How shareable are the tools presented, and how to know what other organisations are developing? More discoverability of shareable services is needed
- What it means to be CSPA-compliant, and on-going work to elaborate this
- The extent to which technical choices should be, or can be, prescribed in CSPA, or whether a (limited) range of options should be specified
- Finer granularity of services can increase flexibility, but the level of granularity should not go beyond what can be understood as statistical functionality
- Consistency and standards within the chosen programming environment can be more important than whether that environment is proprietary or open source
- This topic is multi-disciplinary, and communication between methodologists, statisticians and IT experts is essential
- Production and research environments have different requirements, CSPA is more oriented to production. Transfer of functionality between research and production environments is a challenge, and more precise guidelines may be needed for this

E. Theme (v): Managing change

20. This topic was organized by Sander Scholtus (Netherlands) and Agnes Andics (Hungary). It included the following presentations:

- UK - Improvements in editing methods and processes for use of Value Added Tax data in UK National Accounts
- Denmark - Usage of data editing process data at Statistics Denmark
- Switzerland - Data preparation process analysis of the structural survey of the Swiss population census
- Finland - An information model for a metadata-driven editing and imputation system.

21. In addition to several clarification questions, the following points were raised in the discussions:

- Change is more successful if all interested groups are involved. The "Agile" approach can help in this respect
- Is there scope for a service to control / steer the production process, and for process indicators to continuously monitor process performance?
- A standard way of structuring methods can facilitate a more "plug and play" approach
- How to manage loops when modelling methods within process flows?
- How to provide more information about methods (and edit rules) in a meaningful way, in the context of process modelling? A link to a method library is a possible solution, possibly using semantic tools?
- A classification of methods would be useful

F. Mini sprint for Topic (iii): How to foster the implementation, within statistical offices, of good practices from other organizations, as well as areas for international collaboration, and priorities for joint work.

22. This topic was organized by Alexander Kowarik (Austria), Daniel Kilchmann (Switzerland), Steven Vale (UNECE) and Li-Chun Zhang (Norway).

23. After an introduction to the working methods and expected outcomes from the mini sprint, by Daniel Kilchmann (Switzerland), participants identified potential future activities through a mixture of small-group and plenary discussions.

24. The first round of the mini-sprint identified three areas for international collaboration activities between meetings. The second round resulted in outline proposals for these three activities, as follows:

(i) Method library/architecture: Classification and inventory of methods

- Description of activity:
  - Define the scope of the inventory, and whether it should be at the level of methodological functions or individual methods
  - Agree a classification, starting from existing work in the GSDEMs, and on-line statistical software catalogues and lists, including consideration of semantic approaches and the use of tags, e.g. to GSBPM sub-processes
  - Develop the inventory in a way that links methods to implementation tools
  - Develop strategies for maintenance and communication
- Benefits:
  - It will support re-use of methods and tools, reducing duplication of effort and improving efficiency
  - It will help to standardise terminology
  - It will be a practical implementation of modernisation standards and models, demonstrating their value to methodologists
- Expected effort:
  - A group of methodologists, IT experts and statisticians, including expected users
- Interested people:
  - The expected users are those developing technical and methodological solutions in statistical organisations

(ii) Consolidation and update of methodological guidelines on statistical data editing

- Description of activity:
  - Prepare a list of existing guidelines and handbooks
  - Create an indexed and searchable inventory of papers from Statistical Data Editing Work Sessions
  - Add references to implemented tools
  - Assess the usefulness (and use) of existing guidelines
- Benefits:
  - It will promote common understanding and standardisation
- Expected effort:
  - A group of experts including methodologists, statisticians and IT staff
  - Feedback and review by the wider statistical community
  - Possible translation of final materials
- Interested people:
  - Methodologists
  - IT specialists
  - Subject-matter experts

(iii) Improving understanding and communication of CSPA and other relevant standards

- Description of activity
  - Formulate questions to CSPA experts to improve knowledge of CSPA for methodologists
  - Prepare explanations on key points such as how to move from a method to a CSPA service
  - Create practical examples of building CSPA services from R packages, SAS macros and existing tools such as jDemetra+
  - Develop and promote communication tools, e.g. CSPA video
- Benefits
  - Improved understanding (and use) of CSPA by methodologists
- Expected effort:
  - A group including methodologists, statisticians, IT staff and CSPA experts
  - Participation in different activities for different groups at the CSPA Workshop in Wiesbaden in July, extending the target audience to non-IT staff
- Interested people:
  - Methodologists, subject-matter specialists, IT staff, business / information architects

25. The following topics were proposed for discussion at a future Work Session on Statistical Data Editing, based on discussions during the mini-sprint and in the plenary session. Countries and organisations that provisionally indicated an interest to provide contributions on these topics are shown in brackets:

- Use of administrative data, including efficient error detection and data editing in the context of an integrated data set (France, Germany, Netherlands, UK, Norway)
- Quality of editing and imputation processes (Germany, Canada, Switzerland, Albania, Netherlands)
- Editing and imputation processes for Big Data/Geospatial data (Netherlands, France, Denmark, Norway)
- Standardisation of editing and imputation methods and systems (including practical demonstrations) (Austria, Canada, Poland, Slovenia, Spain, Malta, UK, Germany, Finland, Netherlands)
- Reducing cost, and increasing efficiency of data editing (Germany, Eurostat)
- Development of new data editing applications and methods (including practical demonstrations) (France, Canada, Poland, Netherlands, Germany)
- Privacy and data access implications of data editing, (Canada, Germany)
- Machine learning in the context of data editing (Norway, Netherlands, France, Spain, Austria)

26. Presentations in future events should include a final slide on what other organisations can gain from the methods, tools or experiences presented.