INF.1
18 October 2016

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

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

INFORMATION NOTICE No.1*

Statistics Netherlands
will host the Work Session in The Hague, Netherlands
from 24 to 26 April 2017

I. Purpose of the Work Session

1. At its 2016 plenary session, the Conference of European Statisticians included the Work Session on Statistical Data Editing in its 2017 meeting programme. This Work Session will be held from 24 to 26 April 2017 in The Hague, Netherlands, at the kind invitation of Statistics Netherlands. This Work Session aims to progress work on statistical editing in the context of the wider work programme on Statistical Modernization. In particular, the Work Session will:

- Identify the most promising new methods that can be used to manage statistical quality risks arising from the use of new data sources (including Big Data).
- Develop approaches for standardizing and implementing statistical editing functionalities.
- Consider how methodologists can contribute to the wider Modernization work programme.

2. The target audience of this Work Session includes statisticians who deal with the editing and imputation of statistical data derived from surveys, censuses, administrative and external sources, spanning various subject-matter areas, both social and economic.

II. Agenda of the meeting

3. The programme of the Work Session will consist of the following substantive topics:

(i) New perspectives for data editing in the context of new data sources and data integration.
- Big data, data blending, modelling, geo-spatial data

(ii) International collaboration on standards and tools for data editing
- Common Statistical Production Architecture (CSPA) services
- Demonstrations, peer reviews

* A detailed agenda for the Work Session will be issued at a later date.
Information about the venue and local arrangements in The Hague (hotel accommodation, transportation etc.) will be provided in Information Notice No 2. See also Section VI of this Information Notice.

- Implementation of standards – Generic Statistical Data Editing Models, Validation and Transformation Language (VTL), etc.

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

4. Detailed explanatory notes on the nature and expected outcomes of topics (i) – (iii) are provided in Section V of this Information Notice.

III. Participation and accreditation

5. Representatives of all Member States of the United Nations and of interested intergovernmental organizations are welcome to participate in the Work Session. Participants representing non-governmental organizations in a consultative status with the United Nations Economic and Social Council may also attend. **All participants must be accredited by the competent authorities of their country or international organization.**

6. All participants attending the Work Session are requested to have a valid passport and, if required, a visa. Applications for visas should be made as soon as possible to the embassy of the Netherlands in the country in which the participant resides, with a reference to the UNECE Work Session on Statistical Data Editing. A letter to facilitate obtaining a visa can be requested from Statistics Netherlands (see also section VI of this Information Notice).

7. Statistical offices and international organizations should inform the UNECE secretariat **before the end of November 2016 if** their organization intends to participate at the Work Session and/or submit a contribution on the topics mentioned above, by submitting their details via the following website: http://www1.unece.org/stat/platform/display/WSSDE/Your+Participation+and+Contribution

8. Participants should also register **by 17 February 2017** by completing the on-line registration form available at: https://www2.unece.org/uncdb/app/ext/meeting-registration?id=9aua5b

IV. Documentation, methods of work and official languages

9. The working language of the Work Session is English. All documents should, therefore, be submitted in English only. The following deadlines and requirements apply:

- A short abstract of the paper should be submitted **by the end of November 2016**, via the following website: http://www1.unece.org/stat/platform/display/WSSDE/Your+Participation+and+Contribution
- Those whose abstracts are accepted will be asked to produce papers in MS Word or PDF formats, around 10 pages, **by 17 February 2017**.

10. All participants are welcome to submit a paper on any topic of the agenda. For each topic, authors, or a selection of them, depending on the number of papers received, will have the opportunity to give a short presentation of their paper. The Session Organizers may ask a few of these speakers to give a longer presentation.

11. About 20 minutes will be allocated for long presentations while short presentations will be given about 5-10 minutes (time permitting) to simply highlight the main issues raised in their papers. Presenters may use PowerPoint or Adobe Acrobat full screen for their presentations. The UNECE secretariat cannot provide translation of the presentations.

- Presentations in PowerPoint or Adobe Acrobat should be sent to the UNECE secretariat by **the end of March at the latest**. These will be installed on the conference room computer prior to the meeting.
- Participants intending to demonstrate software should contact the UNECE secretariat to ensure the necessary technical infrastructure will be available.

12. All papers and presentations will be made available on the website of the UNECE secretariat at the following location: http://www.unece.org/index.php?id=43887#/

13. Participants are encouraged to download the papers from the website and bring their own copies to the Work Session. Documents posted on the website before the Work Session will **not** be distributed in the conference room.

V. Explanatory notes to the agenda

Topic (i): New perspectives for data editing in the context of new data sources and data integration.

This topic could include contributions concerning the following themes:

Machine learning

14. Machine learning is a subfield of artificial intelligence that allows computers to learn to perform a certain task based on examples, and without being explicitly programmed for that purpose. The application of machine learning techniques to processes within the GSBPM (Generic Statistical Business Process Model) can contribute to the modernisation and improvement of the efficiency of statistical production.

15. This theme aims to bring together contributions addressing the use of machine learning and data mining methods (for instance algorithms for classification, clustering, tree based methods, random forests, neural networks ...) with a special focus on data editing and imputation.

16. Contributions may report, for instance, on machine learning methods that automatically find rules for error detection or correction, or methods to impute missing values in different types of variables (categorical, continuous, ordinal). Contributions that examine the ease and generality of application and the resulting data quality arising from such approaches, compared to the more traditional ones are also very welcome. Data mining methods that are developed to monitor the effects of data editing processes, in order to evaluate and improve quality and/or efficiency are also of interest.

New and emerging methods

17. This theme covers new and emerging methods for improving or optimizing the process of data editing and imputation. In recent years there have been major contributions based on probability models, robust statistical methods, time series analysis, Bayesian networks, data visualization, and machine learning techniques. This topic aims to bring together contributions that include innovative ideas and applications related to various aspects of editing and imputation.

18. More and more often, sample surveys are replaced by other data sources, or used in combination with them. Therefore, of special interest are methods for the detection and amendment of errors in these alternative or new data sources, including administrative and big data, either on their own or in combination with survey data.

19. Contributions could report on developments in theory and techniques, empirical comparisons and evaluations of methods, exploration of potential bias of given methods, and ways of combining different methods in an editing process. They could also highlight the expected impact that new methods might have on the statistical agency, including how they contribute to standardizing concepts, terminology, methods, data structures and the quality of its data products.

New data sources – Big Data, multi –source statistics and geo-spatial data

20. National Statistical Institutes are increasingly faced with the need to reduce costs of data collection, but also to benefit from the increasing availability of administrative data, as well as from the development of private owned « big data ». Sample surveys are therefore gradually replaced by other sources, used in combination with them, or collected through multimodal processes to address declines in response rates.

21. The development of new data sources (e.g. shop checkout scanner data, data captured by road censors, geo-spatial data generated by mobile phones, but also more common financial data, etc.) represents a great opportunity for NSIs, but also raises many new methodological questions, especially as far as data editing, control and imputation are concerned. For instance, how to best combine and integrate exhaustive administrative data and survey data only available on a sample, how to compare different data sources available on the same population or part of the population, how to edit or extract the relevant information from massive data for which the usual paradigm of selective editing is no longer bearable due to the mass of information to control?

22. Contributions could report on advancements in developing a theoretical framework to best describe the issues raised by the editing and imputation of big or multisource data. They could report on theories, practises and empirical comparisons on the best ways to deal with them. They also could highlight the expected impact that new data sources have on the statistical agencies, including how they contribute to changing their processes, with more and more statistical outputs depending on data that the agency does not control the production of, nor the standardizing concepts, terminology, methods, data structures and the quality of its data products.

Census 2021

23. With the next census already around the corner and enough time since the last census, National Statistical Institutes have had the opportunity to make improvements to their statistical production process, and to advance their methodology in editing and imputation techniques.

24. Several countries make heavy (or exclusive) use of administrative data in the production of their census, which poses additional challenges in terms of methodological problems and appropriate solutions. This point in time is an ideal occasion for discussing results, ideas and tests, related to methodological developments that have been carried out in this area.

25. Papers discussing the results of studies on, and applications of, editing and imputation in the context of the production of the census would be welcome, especially those describing the methodologies developed or used, with particular attention to the evaluation of the benefits, risks and open problems deriving from the adoption of specific editing and imputation solutions.

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

This topic could include contributions concerning the following themes:

Shared software tools and CSPA services – Demonstrations and implementation experiences

26. The modernisation efforts of National Statistical Institutes include the international alignment of concepts and processes, to enhance the possibilities for international collaboration. This theme is specifically focused on cooperation in the sharing and development of software solutions for data editing. The common statistical production architecture (CSPA), and generic statistical data editing models (GSDEMs), greatly facilitate the development of shareable software tools for the various data editing functions (e.g. rule-checking, error localisation, imputation, etc.). Stronger international collaboration in this respect is considered a key opportunity for improving the efficiency of developing high quality statistical software.

27. Under this theme, papers are welcome on portable software solutions, on collaboration initiatives targeting software development, or on reuse of tools for data editing. Examples of relevant topics include:

- Software implementations of the Common Statistical Production Architecture (CSPA)
- Implementation experiences of generic software modules;
- Shareable data editing tools developed for national purposes;
- Development or (re-)use of portable (plug-and-play) software tools for editing and imputation functions; and
- Open source tools for data editing.

Standards in international collaboration – Including implementation of the new and emerging standards: VTL, GSDEMs, and CSPA

28. When designing and implementing modern data editing strategies, it is necessary to use suitable methods, processes and software tools. Development of these elements can be a complex, long-term and very costly job. For smaller organisations, development of state-of-the-art methods and software tools often exceeds their capability and budget constraints. The principle of sharing seems to be a natural solution in such cases.

29. Use of standards can support the sharing of knowledge, experiences and methodologies, as well as the sharing of tools, data, services and resources. Improvement of the efficiency and robustness of statistical processes through systematic collaboration requires agreed standards. Although the statistical community has shared knowledge and established concepts, methods and best practices, there is still a lack of successful collaboration in the field of exchange of standardisation and harmonisation of data editing strategies at the global architecture level.

30. Papers focusing on the following issues are especially welcome under this topic:

- Building bridges between local and global solutions;
- International initiatives and related activities;
- Modernisation and harmonisation by using Enterprise Architecture;
- Implementation of the Validation and Transformation Language (VTL);
- Implementation of the Generic Statistical Data Editing Models (GSDEMs); and
- Implementation of the Generic Statistical Information Model (GSIM).

31. Given the nature of this Work Session, we are only looking for papers that focus on applications and developments related to data editing and/or imputation.

Managing change

32. In recent years, standardised procedures and new methods for data editing have been studied for improving the harmonisation and efficiency of the editing process. These methods aim to combine a high level of quality in the statistical products with a low level of cost for the editing processes. Many generic improvements to processes and methods are by now well established. However, their introduction in the statistical production process requires additional efforts to ensure changes are smoothly implemented and lasting benefits from these changes are achieved.

33. Some important factors that can help with the successful introduction of changes to editing processes include:

- Involving and obtaining support and cooperation from top-level management and the editing staff as early, openly and fully as possible;
- Understanding where an NSI wants to be and determining the initiatives to get there; and
- Planning development work in appropriate, achievable and measurable stages.

34. Getting the necessary support requires making stakeholders aware of the importance of the new procedures/methodologies in terms of improvements in the production process. A demonstration of these improvements, using clear quality indicators, can be useful in this respect. It is of the utmost importance that changes to the editing and imputation process are smoothly introduced in the production process to facilitate transition. It is also important to monitor the impact of changes, to make valid conclusions on expectations that are met, or explain those which are not. If the data editing system is standardised and based on an industrialised statistical production process, changes will be easier to introduce, manage and accept.

35. For this topic, contributions are invited concerning the introduction of changes to data editing processes. We particularly welcome contributions on concrete experiences with the following points:

- Engaging top-level management (and each of the levels below) in supporting changes in culture and to editing and imputation methods and processes;
- Training and education to support these changes; and
- Evaluating the quality (including efficiency) of editing and imputation processes, and monitoring the effects of changes made to these processes.

Contributions may also discuss the benefits of a standardised data editing system for managing and supporting changes.

Topic (iii): How to foster the implementation in NSOs of good practices from other organisations, areas for international collaboration, and priorities for joint work.

36. During the last decades, editing and imputation has been developed and integrated into most statistical processes in most statistical agencies, with the aim of reducing costs and managing resources more efficiently. International collaboration and the use of good practices were often the basis of these activities, driven by the experiences of the NSIs, and the results of international projects such as GSDEMs, MEMOBUST Handbook, EDIMBUS-RPM and EUREDIT.

37. The UNECE Work Session on Statistical Data Editing is organised under the activities of The High Level Group of Modernisation of Official Statistics (HLG-MOS) and its modernisation committee 'Production and Methods'. HLG-MOS has asked this Work Session for a tangible demonstration of its contribution to the implementation of good practices, and promotion and identification of priorities for joint work on international collaboration .

38. The participants of the Work Session will discuss topics in small groups, then report back to the plenary session. More information on the workshop will be provided to registered participants in Information Notice No. 3. Participants will then be asked to choose among the items and to prepare for the discussion. The organising committee in collaboration with the UNECE will produce a report summarising the findings of the discussions based on the following items:

a) Discussing which method presented at the Work Session or already well-established could/should be implemented as a CSPA service, how to do it, which are the benefits compared to sharing tools (SAS Macros, R packages, etc.) and whether an international collaboration might be beneficial;

b) Detecting missing domains and suggestions for future topics, including how this could be done and how they are related to the contributions of the Work Session;

c) Defining potential priorities and finding overlaps between countries; and

d) Revision and updating of standards and discussion of gaps and potential further development of standards, e.g. the UNECE glossary (http://www1.unece.org/stat/platform/display/kbase/Glossary), GSDEMs, etc., and how the adoption to these standards may be enhanced.

VI. Local arrangements

A. Meeting venue

39. The meeting will be held at:

Central Bureau of Statistics / Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague
Tel. + 31 70 3373800

B. Travel and Accommodation

40. Participants and/or their offices are requested to make their own travel arrangements and hotel reservations.

41. A letter to facilitate obtaining a visa can be requested from Statistics Netherlands. Please contact Jeroen Pannekoek **(**j.pannekoek@cbs.nl**)** if you need a special invitation letter to obtain a visa. Please note that, since the meeting is organised by the UN, the visas should be issued free of charge.

VII. Further information

42. For further information, please contact the following organizers:

Statistics Netherlands:

Mr. Jeroen Pannekoek
Senior Methodologist
Statistics Netherlands, P.O. Box 24500
NL-2490 HA THE HAGUE
Netherlands
Phone: +31-70 337-4919
Email: J.Pannekoek@cbs.nl

Mr. Sander Scholtus
Methodologist
Statistics Netherlands, P.O. Box 24500
NL-2490 HA THE HAGUE
Netherlands
Phone: +31-70 337-4926
Email: s.scholtus@cbs.nl

UNECE:

Ms. Tetyana Kolomiyets
Statistical Information and Methodology Unit
Statistical Division
United Nations Economic Commission for Europe
Palais des Nations
1211-GENEVA 10
Switzerland
Tel.: +41 22 917-4150
email: tetyana.kolomiyets@unece.org