UNECE

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing 8 November 2022

3-6 October 2022, Online

REPORT OF THE EXPERT MEETING

1. The expert meeting was organized as part of the Conference of European Statisticians' work programme for 2022, within the context of the High Level Group for the Modernisation of Official Statistics. It was held online from 3-6 October 2022.

2. There were 157 participants, including representatives from the following 27 countries: Australia, Austria, Brazil, Bulgaria, Canada, Chile, Denmark, Estonia, Finland, Germany, Hungary, Iceland, Ireland, Italy, Mexico, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Türkiye, United Kingdom of Great Britain and Northern Ireland, and the United States of America.

3. In addition, there were representatives from the United Nations Economic Commission for Europe, and the Economic Commission for Latin America and the Caribbean. There were academic participants from Technische Universität Dortmund and the Užice School of Economics, as well as a keynote presentation from the University of Ottawa.

4. The expert meeting was organised under the responsibility of the High-Level Group for the Modernisation of Official Statistics. The Steering Committee consisted of Alexander Kowarik (Austria), Darren Gray (Canada), Agnes Andics (Hungary), Simona Rosati (Italy), Sander Scholtus (Netherlands), Pedro Revilla (Spain), David Salgado Fernandez (Spain), and Daniel Kilchmann (Switzerland). The meeting was jointly chaired by Daniel Kilchmann, Alexander Kowarik, and David Salgado Fernandez.

5. The agenda included the following substantive topics, each comprising one or more sessions within the meeting:

- Modernisation of data editing and statistical production
- New and emerging methods
- Machine Learning /Artificial Intelligence for editing and imputation
- Use of administrative data for editing and imputation
- Quality

6. There was also a keynote presentation by Prof. David Haziza from the University of Ottawa, as well as a "lightning talks" session that featured 5-minute long presentations that were not accompanied by any paper.

7. Twenty-six substantive presentations were made on within these sessions, including the keynote presentation. (For reference, the timetable is included as Annex 1.)

8. This expert meeting included an online brainstorming exercise to determine future work priorities for this body. This resulted in a number of suggestions from participants (as well as initial

suggestions from the organizing committee and members of the Executive Board of the High-Level Group for the Modernisation of Official Statistics).

9. Towards the end of the meeting, the submitted suggestions were grouped into the following topic groupings, according to the nature of the suggestions, and this was presented before the end of the meeting.

- *Methods* for new Data Sources
  This includes data from sources such as registers, web-scraping, telecoms operators, and financial transactions, as well as trusted smart statistics. Some of this data may be non-probabilistic in nature.

- Synthetic Data
  This involves the use of synthetic data with similar distributional properties to real data for assessing data editing methods.

- Software Tools, including for Machine Learning
  Building upon earlier efforts to develop a list of official statistics software for data editing, this topic focuses on improving the suite of available software, including for the purposes of machine learning and artificial intelligence, for which a standardised methodology within official statistics is lacking.

- Possible new frameworks/common approaches for dealing with Machine Learning and New Data Sources
  This would explore whether Generic Statistical Data Editing Models and standardisation can be applied in the context of new data sources (web-scraped, telecoms, etc.), and approaches to data editing (and to determining its quality) from an input perspective.

- Production
  This focuses on identification of efficiency improvements in data editing processes, and ways to implement new approaches, such as machine learning and artificial intelligence.

- Organizational Aspects
  Adapting organizational culture and training to embrace new tools and methods.

- Miscellaneous
  E.g., Data editing in the context of falling survey response rates, etc.

10. A full list of suggestions submitted, as well as some replies to these suggestions, are available in Annex 2. All papers, presentations, and other output from the meeting are available at the UNECE website (https://unece.org/statistics/events/SDE2022).

Annex 1: Timetable of the UNECE Expert Meeting on Statistical Data Editing 2022, 3-6 October, Online

DAY 1 – Monday, 3 October

Time	Item	Speaker
12:30	Connect to meeting	
13:00	**Opening of the meeting by chair**	
13:05	Instructions for participants	
13:10	**Presentation on the UNECE High-Level Group for the Modernisation of Official Statistics**	**Taeke Gjaltema**
13:20	**Session 1a: Modernisation of data editing and statistical production (Part 1)** Session Organizers: Darren Gray (Statistics Canada) and Pedro Revilla (INE, Spain)	
13:25	**Multiple software systems for the editing and imputation process of the 7th General Census of Agriculture**	**Simona Rosati**
13:40	**Towards a new integrated uniform production system for business statistics at Statistics Netherlands: quality indicators to guide top-down analysis**	**Frank Aelen & Anita Vaasen-Otten**
13:55	*Questions to the authors*	
14:05	**The SCIA system implementing Fellegi and Holt methodology compared to the recent R packages**	**Simona Rosati**
14:20	**Towards a new integrated uniform production system for business statistics at Statistics Netherlands: automatic data editing with multiple data sources**	**Wilco de Jong & Sander Scholtus**
14:35	*Questions to the authors*	
14:45	***Break***	
15:00	***Introduction of the Keynote***	
15:05	**Keynote presentation: David Haziza**	
15:45	Questions to keynote presenter	
16:00	**Session 2: Lightning talks** Session Organizer: Alexander Kowarik (Statistics Austria)	
16:05	**A modern statistical production process based on administrative registers**	**Ewelina Wójcik**
16:10	**e-invoice time series nowcasting with R**	**Bruno Lima**
16:15	Questions to the authors	
16:25	**Concluding remarks for day 1**	
16:30	**Reminder about the future work discussion**	
16:35	**End of Day 1**	

DAY 2 – Tuesday, 4 October

Time	Item	Speaker
12:30	Connect to meeting	
13:00	**Opening of day 2**	
13:05	**Session 3: New and emerging methods** Session Organizers: Simona Rosati (Istat, Italy) and Sander Scholtus (Statistics Netherlands)	
13:10	**Machine learning Imputation for Social Surveys – Random forest imputation of ONS' Household Financial Survey.**	**Mark Edward**
13:25	**Application of the "SwissCheese" method for the imputation of partial non-response in the Survey on Income and Living Conditions.**	**Michael Leuenberger**
13:40	*Questions to the authors*	
13:50	**Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking datasets**	**Andrea del Monaco**
14:05	**Discover the hidden validation rules in your data with 'validatesuggest'**	**Olav ten Bosch**
14:15	- Software demonstration	
14:25	*Questions to the authors*	
14:35	**Session: General discussion and conclusions**	
14:50	***Break***	
15:05	**Session 1b: Modernisation of data editing and statistical production (Part 2)** Session Organizers: Simona Rosati (Istat, Italy) and David Salgado (INE, Spain)	
15:10	**Validation rule management**	**Mark van der Loo**
15:20	- Software demonstration	
15:30	**Banff's next step: an open-source data editing system for advanced tools and collaboration**	**Darren Gray**
15:45	*Questions to the authors*	
15:55	**Growing a Modern Edit and Imputation System**	**Darcy Miller & Megan Lipke**
16:10	**Automatic Data Editing and Imputation Experience in 2020 Mexican Census**	**Edgar Vielma**
16:25	*Questions to the authors*	
16:35	**Session: General discussion and conclusions**	
16:55	**Concluding remarks for day 2**	
17:00	**End of Day 2**	

DAY 3 – Wednesday, 5 October

Time	Item	Speaker
12:30	Connect to meeting	
13:00	**Opening of day 3**	
13:05	**Session 4: Machine Learning /Artificial Intelligence for editing and imputation** Session Organizers: Alexander Kowarik (Statistics Austria) and Daniel Kilchmann (Federal Statistics Office, Switzerland)	
13:10	**Improving statistical data editing with Machine Learning: some use cases in Statistics Spain (INE)**	**Sandra Barragán**
13:25	**Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions.**	**Blandine Bianchi**

Time	Item	Speaker
13:40	**Robust regression, MissForest and calibration combined with non-linear optimization to impute VAT turnover**	**Jacques Saliba**
13:55	*Questions to the authors*	
14:10	**Univariate and multivariate goodness (of fit) of imputation**	**Maria Thurow & Florian Dumpert**
14:25	**The imputation of the "Attained Level of Education" in the base register of individuals through Neural Networks using sampling weights.**	**Fabrizio De Fausti**
14:40	*Questions to the authors*	
14:50	**Session: General discussion and conclusions**	
15:05	***Break***	
15:20	**Session 5: Use of administrative data for editing and imputation** Session Organizers: Ágnes Andics (Central Statistical Office, Hungary) and Pedro Revilla (INE, Spain)	
15:25	**Data imputation for the purposes of statistical research with the use data from administrative registers**	**Paweł Murawski**
15:40	**Producing admin-based property floor area statistics for England and Wales: methods, data and quality**	**Stephan Tietz & Emily Mason-Apps**
15:55	**The use of administrative records for data imputation in Mexico's Economic Censuses**	**José Luis Mercado Hernández**
16:10	*Questions to the authors*	
16:25	**Session: General discussion and conclusions**	
16:40	**Concluding remarks for day 3**	
16:45	**End of Day 3**	

DAY 4 – Thursday, 6 October

Time (Geneva)	Item	Speaker
12:30	Connect to meeting	
13:00	**Opening of day 4**	
13:05	**Session 6: Quality** Session Organizers: Darren Gray (Statistics Canada) and Sander Scholtus (Statistics Netherlands)	
13:10	**Experimental Short-Term Statistics based on Data Imputation Methods**	**Jan Ditscheid**
13:25	**Variance estimation for the mass imputation of the "Attained level of education" in the Italian Base Register of individuals: A comparison between analytical and MonteCarlo estimates.**	**Romina Filippini**
13:40	**Automatic selective editing approach using machine learning: an application to VAT data**	**Benjamin Vasquez**
13:55	*Questions to the authors*	
14:10	**Session: General discussion and conclusions**	
14:20	**Results of the discussions on future work**	
14:40	Conclusion of the meeting	
15:00	**End of the meeting**	

Annex 2: Suggestions made in the future work brainstorming activity

1. This expert meeting included an online brainstorming exercise to determine future work priorities for this body. This resulted in a number of suggestions from participants (as well as initial suggestions from the organizing committee and members of the Executive Board of the High-Level Group for the Modernisation of Official Statistics).

2. Towards the end of the meeting, the submitted suggestions were grouped into a number of topic clusters, according to the nature of the suggestions, and this was presented before the end of the meeting. These suggestions, together with their group headings are as follows. There were also some replies to the suggestions, which are also displayed.

Methods for new Data Sources

- Editing methods for New Data sources (web-scraped data, telecoms operators, financial transaction data, etc.). *(Received 10 supporting votes.)*
- Data editing for trusted smart statistics (TSS). The structure of TSS is different from classic surveys. What is the impact of this difference to the editing phase? Should existing methods be adapted or new ones developed? What is the best data editing process in this context?
- Imputation methods for non-probabilistic data. The use of non-probabilistic data sources is mainly characterised by the risk of selection bias. This also affects the imputation procedures that are mainly based on Missing-At-Random assumption. For this reason, imputation methods for Not-Missing-At-Random cases should be studied. *(Received 6 supporting votes.)*
- Machine learning for imputation of statistical registers. Since statistical registers are a representation of a population, and are updated constantly, they are characterised by longitudinal information at unit level. Machine learning can be particularly promising to deal with this type of data. *(Received 2 supporting votes.)*
- Use of machine learning methods in time series, and especially the detection of outliers in these time series. *(Received 1 supporting vote.)*

Synthetic Data

- Use of synthetic data for editing methods assessment. *(Received 5 supporting votes.)*
    - After listening to David Haziza's (keynote) talk, I believe we should produce synthetic/artificial populations with similar distributional properties to real ones, e.g., using Generative adversarial networks or any other ML technique, for the research community.

Software Tools, including for Machine Learning (ML)

- The awesome list of official statistics software was created during the 2017 meeting. From then, it grew, for the current list. What are the thoughts of this year's participants on how it could further improve? Any ideas or suggestions are welcome. *(Received 8 supporting votes.)*
    - I suggest to begin considering the use of ML/AI tools, which is frightening, because no clear official methodology around ML exists (and a rule of that list is to be actually used in production).
    - So, you suggest we need standardized methodology for ML in official statistics and on the awesome list we would have the tools(s) / packages (or maybe even models) that are fit for the job?
    - Yes, something like that. I'm beginning to see tasks in ML application which shall need standardisation, like the validate package ecosystem. Feature engineering (methods, tools, and repositories) and hyperparameter optimization are two tasks which will consume a lot of resources.

- The approach in the validate package ecosystem, in my opinion, should be exported to more production activities.
- Agree, interesting to think about the generic ML parts that are proven, reliable and accepted (and explainable), and catch those in generic packages.
- There are plenty of implementations: *sklearn* and *keras* in Python, and mlr3 and recipes in R. If we could have some criteria and recommend one or the other, that would help.
- Yes, but I think this means there could be a statistical layer on top of those.

Possible new frameworks/common approaches for dealing with Machine Learning and New Data Sources

- Editing and quality from an input standpoint. *(Received 1 supporting vote.)*
  - Could you specify this a bit more?
  - The main idea is that so far quality in Official Statistics is indeed an output quality conception. For example, selective editing or macro editing is already using preliminary forms of the estimators. With new data sources, this may not be the case anymore.
  - A given admin register may be used in many different ways. A whole new view on input data quality should be furnished, i.e., new editing and imputation techniques.
- Generic Statistical Data Editing Models (GSDEM) and standardisation in the context of New Data sources (web-scraped data, telecoms operators, financial transaction data, etc.). *(Received 3 supporting votes.)*
  - Based on today's discussion, maybe just GSDEM and standardization in general?

Production

- Reviews with the objective of identifying potential for efficiencies in data editing processes (cost/timeliness, quality aspects, etc), suggestions for experimentations in editing, and modernization of processes using ML and AI. *(Received 1 supporting vote.)*
- Putting into production machine learning for data editing.
- Transfer of the (production and) editing and imputation process into newly designed processes: How to do it, as an entire simulation might not always be possible.

Organizational Aspects

- Adapting organizational culture and training to embrace new tools (open-source software) and methods (ML, geostatistics, state space models, etc.). *(Received 1 supporting vote.)*

Miscellaneous

- Editing and imputation in the time of declining survey response rates. *(Received 3 supporting votes.)*