UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Online, 31 August - 4 September 2020)

REPORT OF THE WORKSHOP

1. The Workshop on Statistical Data Editing was held online from 31 August to 4 September in 2020. It was attended by 209 participants from Australia, Austria, Azerbaijan, Belgium, Brazil, Bulgaria, Canada, Chile, Czech Republic, Estonia, Finland, France, Germany, Hungary, India, Indonesia, Ireland, Israel, Italy, Mexico, Netherlands, New Zealand, Norway, Philippines, Portugal, Slovakia, Spain, Lithuania, Poland, Sweden, Switzerland, Ukraine, United Arab Emirates (U.A.E), United Kingdom (U.K), Unite States of America (U.S.), Organisation for Economic Co-operation and Development (OECD), Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC) as well as from Landmark University, Tsuda University, University of Geneva and University of Neuchatel.

2. The workshop aimed to progress the work on statistical editing in the wider context of the High-Level Group on Modernisation of Official Statistics (HLG-MOS) work programme, in particular, to:
   - Identify new methods that can improve the quality and efficiency of editing and imputation (E&I);
   - Investigate the statistical quality risks arising from using new methods and data sources and the ways to address them;
   - Develop approaches to standardising and implementing statistical editing functionalities;
   - Facilitate the sharing of experiences, ideas and tools for modernising E&I process.

3. The opening remarks were given by Mr. Daniel Kilchmann, Deputy Head of the Statistical Methods Unit of the Swiss Federal Statistics Office (FSO) and Mr. Taeke Gjaltema, Chief of the Statistical Management and Modernisation Unit of the UNECE Statistics Division. The programme of the workshop was organised with following substantive topics:

   - Quality: assessing data quality and indicators, chaired by Mr. Sander Scholtus (Statistics Netherlands) and Mr. Pedro Revilla (INE, Spain);
   - Imputation Methods: machine learning and new / emerging methods, chaired by Mr. Sander Scholtus (Statistics Netherlands) and Mr. Li-Chun Zhang (Statistics Norway);
   - Methods: for machine learning and time series data, and new/emerging methods, chaired by Mr. Darren Gray (Statistics Canada) and Mr. Daniel Kilchmann (Federal Statistics Office, Switzerland);
   - Processes: editing in a generic process, standardisation and meta-data driven processes, chaired by Ms. Agnes Andics (Central Statistical Office, Hungary) and Ms. Simona Rosati (Istat, Italy);
   - Data: 2021 Census, administrative data, geospatial data, big data and other alternative data, chaired by Ms. Fern Leather (Office of National Statistics, UK) and Mr. Li-Chun Zhang (Statistics Norway).

4. The workshop included two keynote presentations given by Mr. Yves Tille (University of Neuchatel) and Mr. Mark van der Loo (Statistics Netherlands) on the recent advances in imputation methods and the statistical data cleaning for official statistics with R respectively, as well as a poster session.

5. All background documents and presentations for the workshops are available at UNECE webpage: https://unece.org/statistics/events/workshop-statistical-data-editing

6. Due to the constraint of online workshop, the discussion for future work and topics for the next workshop could not be held through small group discussion format as usual. Instead, a survey questionnaire was conducted prior to the workshop and a plenary discussion was held at the end of the workshop. The work areas proposed include:

   - Development of e-learning courses (e.g. for GSDEMs, validation, influence observation, and more complex methods);
   - Comparison with standard tools (e.g. CANCEIS, BANFF);
   - Design of E&I, guidelines when "no intervention" or "simple methodology" would be fit for purpose;
   - Methodologies for imputation and estimating its variance for register based / combined census model.

7. The following topics were proposed for the future Workshop on Statistical Data Editing:

   - Machine learning, Artificial Intelligence for E&I;
   - Imputation (e.g. for 'Covid-19 mass non-response) and variance;
   - Modernisation of data editing and statistical production;
   - Use of administrative data for E&I;
   - SDE and pre-processing of new digital data for statistical purposes;
   - Quality.

8. More details about proceeding and discussion can be found in the Annex.

Annex: Summary of proceeding and discussions

Topic - Quality: assessing data quality and indicators

10. This topic was organised by Mr. Sander Scholtus (Statistics Netherlands) and Mr. Pedro Revilla (INE, Spain). It included the following presentations:

- Evaluating Imputation Methods using ImpACT: First Case Study, by Mr. Darren Gray (Statistics Canada);
- Variance estimation after mass imputation with an application to the Dutch population census, by Mr. Sander Scholtus (Statistics Netherlands).

11. In addition to several clarification questions, the points raised during the discussions include:

- Imputation strategies used by the national statistical organisations are often complicated, involving several methods and parametrisations applied in sequence. Reproducing this complexity in a simulation study can be challenging;
- There might be a belief that complex method would produce more accurate results, but this has to be thoroughly explored and tested via comparing with results from simpler methods to check gain and loss in terms of accuracy;
- It would worth discussing questions regarding communication with end-users: how to report the quality of E&I to end-users; whether they would understand the steps taken during E&I process, how these affect quality and uncertainty associated with them.

Topic - Imputation Methods: machine learning and new / emerging methods

12. This topic was organised by Mr. Sander Scholtus (Statistics Netherlands) and Mr. Li-Chun Zhang (Statistics Norway). It included the following presentations:

- Wage Imputation with Deep Learning in the French Labor Force Survey, by Mr. Damien Babet (Insee, France);
- Bayesian Estimation of Linear Dynamic Panel Models with Missing Values, by Mr. Marcel Preising (Federal Statistical Office, Germany);
- Outlier detection and imputation using ML, by Ms. Susie Jentoft (Statistics Norway);
- RBEIS: A robust nearest neighbour donor imputation system implemented in SAS, by Ms. Fern Leather (Office for National Statistics, UK).

13. In addition to several clarification questions, the points raised during the discussions include:

- Issue of ML interpretability is drawing increasing attention from NSOs. Users might want some level of interpretability, but it may not be in the form of coefficients as in the traditional parametric models. Also, the concept might be interpreted differently depending on people and organisations;
- Interpretability is important when quality of methods is checked. Often subject-matter experts have difficulties in figuring out the relationship between the input variables and the predictions which leads to scepticism and makes it difficult the methods to be accepted;
- When there are rare outcomes (predicted value) in the training dataset, re-balancing techniques such as bootstrapping are often mentioned as a solution, but it is not clear what exactly is expected to achieve;
- While machine learning methods are good at predicting individual records, they do not perform well in keeping the distribution of the data, hence there is a trade-off between individual predictability and distribution predictability.

Topic - Methods: for machine learning and time series data, and new / emerging methods

14. This topic was organised by Mr. Darren Gray (Statistics Canada) and Mr. Daniel Kilchmann (Federal Statistics Office, Switzerland). It included the following presentations:

- The UNECE High-Level-Group for the Modernization of Official Statistics Machine Learning Project: A report of the Editing & Imputation Group, by Mr. Florian Dumpert (Federal Statistical Office, Germany);
- Editing of Social Survey Data, by Mr. Claus Sthamer (Office for National Statistics, UK);
- ML to identify patterns behind errors in STS statistics, by Ms. Fabiana Rocci (Istat, Italy);
- Two-Phase Learning, by Ms. Tatsiana Pekarskaya (Statistics Norway).

15. In addition to several clarification questions, the points raised during the discussions include:

- Machine learning models learn from data, hence the models become outdated when data is outdated. Creating a new ground-truth data and re-training the ML models could be a massive burden to keep the system running. This issue could be a bottleneck for ML to be integrated in the production. There needs an operational procedure checking if the environmental changes (e.g. change of tax rules) affect data substantially and if so, how to update the old models;
- Given that ML models learn from training data, there is risks of reproducing the limitations of the legacy systems (e.g. mistakes by human workers) on which training data is created. Legacy system may not always be the gold standard, when ML model makes mistake compared to legacy system, this case can occur because ML is not good as legacy system, but it could be also mean that legacy system was wrong;
- Due to the strong driver in improving timeliness of the product, the willingness to accept new technologies such as ML is high among external customers. Internally in the statistical organisations, there are different stakeholders such as methodologists, statisticians and subject matter experts. We have to clearly show what the gains and losses are, in terms of accuracy, interpretability and assumptions when comparing ML methods with traditional methods.

Topic - Processes: editing in a generic process, standardisation and meta-data driven processes

16. This topic was organised by Ms. Agnes Andics (Central Statistical Office, Hungary) and Ms. Simona Rosati (Istat, Italy). It included the following presentations:

- Generic Statistical Data Editing Model (GSDEM), by Mr. Daniel Kilchmann (Federal Statistics Office, Switzerland);
- Implementing main types of International validation rules in national validation processes, by Mr. Olav ten Bosch (Statistics Netherlands);
- Modern, process oriented and metadata driven statistical production, by Ms. Anna Długosz (Statistics Poland);
- Automation of E & I Processes, by Ms. Kerstin Lange (Federal Statistical Office, Germany).

17. In addition to several clarification questions, the points raised during the discussions include:

- Reference architecture framework based on a process-oriented model of statistical production is of interest for many statistical organisations. A lot of work has been done from a theoretical point of view, but much more work still needs to be done in practice;
- Metadata is important for version control and reproducibility to make it possible to re-run a process exactly the same way it was run before;
- GSDEM needs to take into account new technologies such as ML to check if there is any function missed or if new flow models are needed;

- While ML methods show some promise, incorporating them into a larger scale production environment has challenges. Many ML packages are often designed all-in-one (e.g. performs review, selection and imputation) and this adds difficulties in incorporating them into process as one might want to have more control over each step. Standardisation for how metadata is transferred between steps and within flow would be helpful so that ML tools can incorporate these standards into their process.

Topic - Data: 2021 Census, administrative data, geospatial data, big data and other alternative data

18. This topic was organised by Ms. Fern Leather (Office of National Statistics, UK) and Mr. Li-Chun Zhang (Statistics Norway). It included the following presentations:

- Webscraped data for replacing and validating survey questions, by Mr. Johannes Gussenbauer (Statistics Austria);
- An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data, by Ms. Romina Filippini (Istat, Italy);
- Use of administrative data and alternative data for census when applying modern technologies, by Mr. Janusz Dygaszewicz (Statistics Poland);
- An overview of the editing and imputation process of the 2018 Italian Permanent census, by Mr. Francesco Scalfati (Istat, Italy).

19. In addition to several clarification questions, the points raised during the discussions include:

- For administrative sources, there could be two groups of quality criteria: one regarding quality of register (e.g. register stability in the future) and the other regarding quality of data set (e.g. coverage of population in the register, cost of data transformation);
- With new data sources, complexity of editing process is increasing. There is a challenging issue of communicating a trade-off between speed (i.e. high frequency products) and accuracy to customers;
- Metadata can play a key role in producing data in timely manner using new sources as it can inform the users about the original purpose of the data which can decrease the need of editing interventions.

F. Poster Session

- Robust Tools for Statistical Data Editing and Imputation, by Ms. Kazumi Wada (Tsuda University, Japan);
- Internal Information System. A possibility of low-cost data governance inside the National Statistical Offices, by Ms. Tania Garcia (INEGI, Mexico);
- The imputation of the "Attained Level of Education" in the base register of individuals: an experimentation using Machine Learning techniques, by Mr. Fabrizio DeFausti (Istat, Italy);
- Profile of Manufacturing Exports Enterprise, by Mr. Carlo Lopez (INEGI, Mexico);
- Territorial preparation in census 2021, by Ms. Ludmila Ivancikova (Slovakia);
- (Poster-only) Challenges and advancements in assessing data quality during the generation of criminal and justice statistics in Mexico, by Ms. Ines Arce (INEGI, Mexico).