UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

**Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality**

27 January 2022

1-3 December 2021, Poznań, Poland

# REPORT OF THE EXPERT MEETING

1. The expert meeting[1] was organized as part of the Conference of European Statisticians' work programme for 2021, within the context of the High Level Group for the Modernisation of Official Statistics. It was held from 1-3 December 2021, Poznań, Poland, hosted by Statistics Poland, in cooperation with the Poznań University of Economics and Business.

2. There were 168 participants, attending either in person or remotely. These participants included representatives from the following 33 countries: Albania, Australia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Iceland, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Mexico, Netherlands, North Macedonia, Norway, Poland, Portugal, Slovenia, Sweden, Turkey, United Kingdom of Great Britain and Northern Ireland, United States of America, and Viet Nam.

3. In addition, there were representatives from Bank for International Settlements, European Central Bank, Eurostat, Food and Agriculture Organization, Statistical, Economic and Social Research and Training Centre for Islamic Countries, United Nations Economic Commission for Europe, United Nations High Commissioner for Refugees and the World Bank Group.

4. There were also academic participants from Chuo University, Delft University of Technology, Duke University, Georgia Southern University, IPUMS at the University of Minnesota, Poznań University of Economics and Business, Scottish Centre for Administrative Data Research, Università di Bologna, Universitat Rovira i Virgili, University of Edinburgh, University of Manchester, University of Minnesota, University of Oklahoma, University of Ottawa, and the University of the West of England. There were also participants from Cancer Research UK, Centre d'accès sécurisé aux données, Cybernetica AS, DataFirst, Knexus Research Corporation, and SBA Research.

5. The expert meeting was organised under the responsibility of the High-Level Group for the Modernisation of Official Statistics. The Steering Committee consisted of Steven Thomas (Statistics Canada), Janika Tarkoma (Statistics Finland), Sarah Giessing (Destatis, Germany), Eric Schulte Nordholt and Peter-Paul de Wolf (Statistics Netherlands), Andrzej Młodak (Statistics Poland), Aleksandra Bujnowska and Wim Kloek (Eurostat), Krish Muralidhar (University of Oklahoma), and Josep Domingo-Ferrer (Universitat Rovira i Virgili). Peter-Paul de Wolf was the overall chair of the meeting.

6. The agenda included the following substantive topics, each comprising its own session within the meeting:

- Access to microdata;
- Risk assessment: Privacy, confidentiality, and disclosure;
- Software tools for statistical data confidentiality;
- Microdata protection;
- Tabular Data; and

[1] Until 2021, it was called the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

- Other Emerging issues.

7. Thirty-three substantive presentations were made on within these sessions. (For reference, the timetable is included as Annex 1.) In addition, there was an interactive brainstorming to identify issues to be considered in future, the results of which were presented near to the end of the meeting. This resulted in the following main themes being highlighted:

*Risk and utility*

- Standardising and developing better measures of disclosure risk and utility or information loss, and possibly relating them to parameters used, and using them to:
  - Automate the way they are balanced; and
  - Compare original and synthetic data.
- Examining how different SDC techniques impact utility at the level of population or geographic subgroups within the dataset.
- Studying risk and utility when SDC methods are combined (e.g., Targeted Record Swapping and Cell Key Method).
- Assessing risk by mimicking real-life attacks in an ethical way.
- Whether to Standardise attack / risk scenarios, and whether the DP standard and current strategies are suitable.
- The relationship between formal risk measures and perceived risk.

*Output checking*

- Whether the automation of output checking is realisable, or whether algorithmic approaches could at least augment the output checking process.
- Whether cell-suppression rules should be reconsidered (suppressing/perturbing values equal to 1 or 0).
- The relative advantages of suppression versus perturbation techniques.
- Output checking of research results to avoid risks to officially-released figures based on the same data that may have employed perturbation as a SDC safeguard.

*Inference and bias*

- How to relate the purpose of use of synthetic data (for inference versus prediction) to the choice of method used to generate it (for example statistical versus non-statistical, such as machine learning).
- The pros and cons of developing multivariate regression methods that can deal with missing, suppressed, generalised or quantised values, compared to methods using synthetic data (possibly multiple copies of the dataset) to impute these values.

*Input Privacy Preservation Approaches*

- Using input privacy-preserving techniques for remote secure computation and privacy-preserving analytics (such as federated machine learning) in a multi-party context.

*Support for SDC practitioners*

- Whether extra training needs to be given to SDC practitioners, and how to improve the access of low- and middle-income countries to SDC procedure best practices, which may have been developed mainly in high-income countries.

- Clearer support for users of new SDC methods on the fixing of parameters employed by such methods.

*Communicating with and educating data users*

- How to communicate the principles of SDC techniques to those who use such data (e.g., perturbation methods, noise addition, non-additivity, etc.).
- Educating data users (and perhaps regulators also) to not over-interpret small/insignificant numbers (e.g., discrepancies in cell-key adjustment, or apparent class disclosure in rounding).

8. All abstracts, papers, presentations, and other output from the meeting are available at the UNECE website (https://unece.org/statistics/events/SDC2021).

# Annex 1 Timetable of the joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality 2021, 1-3 December, Poznań, Poland

### DAY 1 – Wednesday 1 December

| Time | Item |
|---|---|
| **08:45** | *Registration of those arriving in Poznań & Connection for remote participants* |
| **09:30** | **Opening of the meeting by hosts:**<br>- Maciej Żukowski, Rector of Poznań University of Economics and Business<br>- Dominik Rozkrut, President of Statistics Poland<br>**Introduction to HLG-MOS and Statistical Data Confidentiality (Taeke Gjaltema)**<br>- Taeke Gjaltema (UNECE) |
| **10:00** | **Introducing the procedure for questions and answers** |
| **10:05** | **Topic: Access to microdata**<br>Session Organizers: Aleksandra Bujnowska (Eurostat) and Eric Schulte Nordholt (Statistics Netherlands) |
| **10:15** | **Access to different kinds of Statistics Netherlands' microdata**<br>Eric Schulte Nordholt (Statistics Netherlands) |
| **10:30** | **Creating ready-made research datasets from national administrative registers**<br>Päivi Kankaanranta (Statistics Finland) |
| **10:45** | **Fingerprinting relational data**<br>Tanja Šarčević (SBA Research) |
| **11:00** | **Questions to the authors** |
| **11:15** | ***Break (20 minutes)*** |
| **11:35** | **Shedding light on the legal approach to aggregate data under the GDPR & the FFDR**<br>Emanuela Podda (Università di Bologna) |
| **11:50** | **Transnational access to confidential microdata: Progress and impact for research**<br>Maria Alkhoury (Centre d'accès sécurisé aux données / Secure Data Hub) |
| **12:05** | **Microdata access where we are and where we need to go**<br>Elizabeth Green (University of the West of England) |
| **12:20** | **Questions to the authors** |
| **12:35** | ***Lunch break*** |
| **13:50** | **Topic: Access to microdata (contd)** |
| **13:55** | **Microdata access services coping with COVID-19 lockdown**<br>Natalia Volkow (National Institute of Statistics and Geography) |
| **14:10** | **Questions to the author and general discussion** |
| **14:30** | **Topic: Risk assessment: Privacy, confidentiality, and disclosure**<br>Session Organizers: Josep Domingo-Ferrer (Universitat Rovira i Virgili), Krish Muralidhar (University of Oklahoma) |
| **14:40** | **Statistical disclosure control for machine learning models**<br>Felix Ritchie (University of the West of England) |
| **14:55** | **The trade-off between the risk of disclosure and data utility in SDC – a case of data from a survey of accidents at work**<br>Andrzej Młodak (Statistical Office in Poznań) |
| **15:10** | **Using machine learning to assist output checking**<br>Josep Domingo-Ferrer (Universitat Rovira i Virgili) |
| **15:25** | **Questions to the authors** |
| **15:40** | ***Break (20 minutes)*** |

| | |
|---|---|
| **16:00** | **Disclosure metrics born from statistical evaluations of data utility**<br>Devyani Biswal (University of Ottawa) |
| **16:15** | **Risk assessment procedures for the 2020 U.S. census**<br>David Van Riper (University of Minnesota) |
| **16:30** | **Questions to the authors** |
| **16:45** | **Proposal for a risk assessment scale for privacy risks in the disclosure of statistical information**<br>Jesús González López (National Institute of Statistics and Geography) |
| **17:00** | **Database reconstruction is very difficult in practice**<br>Krishnamurty Muralidhar (University of Oklahoma) |
| **17:15** | **Questions to the authors and general discussion** |
| **17:40** | **End of Day 1** |

**DAY 2 – Thursday 2 December**

| | |
|---|---|
| **08:45** | *Connection for remote participants* |
| **09:00** | **Reflections from day 1** |
| **09:15** | **Topic: Software tools for statistical data confidentiality**<br>Session Organizers: Peter-Paul de Wolf (Statistics Netherlands) and Andrzej Młodak (Poznań Statistical Office) |
| **09:25** | **Suppression of directly-disclosive cells in frequency tables**<br>Daniel Lupp (Statistics Norway) |
| **09:40** | **Introducing a graphical user interface for creating the metadata governing the secondary cell suppression process**<br>Michel Reiffert (Destatis) |
| **09:55** | **Assessing, visualizing and improving the utility of synthetic data**<br>Gillian Raab (Scottish Centre for Administrative Data Research) |
| **10:10** | **Questions to the authors** |
| **10:25** | **Private linear regression: Can we scale up with Big Data?**<br>Giuseppe Bruno (Bank of Italy) |
| **10:40** | **Automatic checking of research outputs**<br>Marco Stocchi (Eurostat) |
| **10:55** | **Questions to the authors and general discussion** |
| **11:15** | ***Break (20 minutes)*** |
| **11:35** | **Topic: Microdata protection**<br>Session Organizers: Josep Domingo-Ferrer (Universitat Rovira i Virgili), Krish Muralidhar (University of Oklahoma) |
| **11:45** | **Accounting for longitudinal data structures when disseminating synthetic data to the public**<br>Joerg Drechsler (Institute for Employment Research) |
| **12:00** | **AI-based privacy preserving census(like) data publication**<br>Johannes Gussenbauer (Statistics Austria) |
| **12:15** | **Generating tabular data using generative adversarial networks with differential privacy**<br>Giacomo Astolfi (European Central Bank) |
| **12:30** | **Questions to the authors** |
| **12:45** | ***Lunch break*** |
| **14:00** | **Topic: Microdata protection (contd)** |
| **14:05** | **Generative adversarial networks for synthetic data generation: A comparative study**<br>Claire Little (University of Manchester) |

| | |
|---|---|
| **14:20** | **Data access modernization in National Statistical Offices through synthetic data, the HLG-MOS guide**<br>Kenza Sallier (Statistics Canada) |
| **14:35** | **Extreme value protection adjustment for different subpopulations in complex data sets**<br>Anna Oganian (National Center for Health Statistics) |
| **14:50** | **Questions to the authors and general discussion** |
| **15:25** | ***Break (20 minutes)*** |
| **15:35** | **Topic: Tabular data**<br>Session Organizers: Steven Thomas (Statistics Canada) and Sarah Giessing (Destatis) |
| **15:45** | **Differential privacy and noisy confidentiality concepts for European population statistics**<br>Fabian Bach (Eurostat) |
| **16:00** | **Fair risk-utility comparison of tabular perturbation methods by post-processing to expected frequencies**<br>Øyvind Langsrud (Statistics Norway) |
| **16:15** | **Suppression or perturbation?**<br>Wim Kloek (Eurostat) |
| **16:30** | **Questions to the authors** |
| **16:45** | **Considerations to deal with the frozen cell problem in Tau-Argus Modular**<br>Sarah Giessing (Destatis) |
| **17:00** | **Increasing utility of economic statistical information**<br>Steven Thomas (Statistics Canada) |
| **17:15** | **Questions to the authors and general discussion** |
| **17:35** | **End of Day 2** |

### DAY 3 – Friday 3 December

| | |
|---|---|
| **08:45** | *Connection for remote participants* |
| **09:00** | **Reflections from day 2** |
| **09:15** | **Topic: Other Emerging issues**<br>Session Organizers: Peter-Paul de Wolf (Statistics Netherlands) and Janika Tarkoma (Statistics Finland) |
| **09:25** | **A proof-of-concept solution for secure processing of mobile network operator data for official statistics**<br>Fabio Ricciato, (Eurostat) |
| **09:40** | **Modelling data environments within PROV to assist anonymisation decision-making**<br>Mark Elliot (University of Manchester) |
| **09:55** | **Structural uniqueness in network data**<br>Marieke de Vries (Statistics Netherlands) |
| **10:10** | **Questions to the authors and general discussion** |
| **10:40** | **Updates on the HLG project on input privacy preservation**<br>Dennis Ramondt (Statistics Netherlands) |
| **10:55** | **Questions to the presenter** |
| **11:00** | **Results of the discussions on future work** |
| **11:30** | **Conclusion of the meeting** |
| **11:45** | **End of the meeting** |