UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Expert Meeting on Statistical Data Confidentiality 6 November 2023
26-28 September 2023, Wiesbaden, Germany

REPORT OF THE EXPERT MEETING

1. The expert meeting was organized as part of the Conference of European Statisticians' work programme for 2023, within the context of the High-Level Group for the Modernisation of Official Statistics activity. It was held from 26-28 September 2023 in Wiesbaden, Germany, hosted by the Federal Statistical Office of Germany, in cooperation with the RheinMain University of Applied Sciences.

2. There were 105 participants, including representatives of national statistical offices, central banks and government agencies of the following 23 countries: Austria, Chile, Czechia, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Mexico, Netherlands (Kingdom of the), North Macedonia, Norway, Poland, Spain, Switzerland, United Arab Emirates, United Kingdom of Great Britain and Northern Ireland, and United States of America.

3. In addition, there were representatives from Eurostat, Food and Agriculture Organization, International Monetary Fund, United Nations Office for the Coordination of Humanitarian Affairs, and the World Bank Group.

4. The meeting was also attended by academic participants from Chuo University (Japan), German Cancer Research Center, GESIS-Leibniz Institute for the Social Sciences (Germany), RheinMain University of Applied Sciences (Germany), Utrecht University (Netherlands), Swiss Data Science Center (Switzerland), University of Applied Sciences and Arts Northwestern Switzerland, Universitat Rovira i Virgili (Spain), University of Edinburgh (United Kingdom), University of Manchester (United Kingdom), University of the West of England (United Kingdom), and University of Oklahoma (United States of America). There were also participants from Cancer Research United Kingdom, Centre d'accès sécurisé aux données (CASD), NTT DOCOMO (Japan), and The Sensible Code Company.

5. The expert meeting was organized under the responsibility of the High-Level Group for the Modernisation of Official Statistics. The Organizing Committee consisted of Annu Cabrera (Statistics Finland), Sarah Giessing (Destatis, Germany), Eric Schulte Nordholt and Peter-Paul de Wolf (Statistics Netherlands), Tomasz Klimanek (Statistics Poland), Aleksandra Bujnowska (Eurostat), Krish Muralidhar (University of Oklahoma), and Josep Domingo-Ferrer (Universitat Rovira i Virgili). Peter-Paul de Wolf chaired the meeting.

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

- Innovative approaches in granting access to microdata;
- Producing useful microdata files;
- Challenges in publishing safe tables and maps;
- Risk assessment: Privacy, confidentiality, and disclosure vs utility;
- Output checking in research data centres; and
- Other emerging issues, including a discussion on the topics for future work.

7. Thirty-two substantive presentations were made within these sessions. The timetable, papers, presentations, and other output from the meeting are available at the UNECE website https://unece.org/statistics/events/SDC2023.

8. The discussions resulted in the following main themes and issues being highlighted:

Innovative approaches in granting access to microdata.

- How different modes of access can be provided for analysing different anonymised data products – going from remote execution to remote access.
- At Eurostat, Covid 19 gave momentum to initiatives to provide remote access to microdata, and in January 2022, a remote access system for data was made operational.
- There is variation among low- and middle-income countries and high-income countries in terms of the solutions, challenges, needs and specificities of data access.
- Safe locations and accredited access points tend to be preferred by data providers, as it can ensure they have better control of access to data by third parties.

Producing useful microdata files

- Various organizations and institutions are grappling with the challenge of safeguarding privacy while simultaneously ensuring the usefulness of data, particularly in the context of georeferenced, individual-level, and microdata files.
- Exploring innovative approaches offers the potential to strike a balance between data confidentiality and the research potential of data.
- In the United Kingdom, Secure Data Environments (SDE) unlock significant research potential for securely stored sensitive data, which are accessible to accredited researchers. However, this approach poses challenges, requiring researchers to invest time and resources in complex datasets without prior access, and burdens SDEs with administrative responsibilities.
- The German Federal Statistical Office is collaborating with universities on the "Anonymization for integrated and georeferenced Data" (AnigeD) project, to improve availability of such data, and to explore the potential of synthetic datasets for research, while ensuring data confidentiality.
- The Basque Statistics Office described their process for providing protected microdata files for research purposes, using the example of their Labour Force Survey.

Challenges in publishing safe tables and maps

- The increased use of geo-referenced data in official statistics has enabled more precise population estimates, but it also poses privacy challenges. Differential Privacy (DP) is being used in a Swiss pilot study to protect individual privacy, while providing detailed poverty statistics. DP has specific application to sub-national statistics and represents a framework for controlling disclosure risk.
- The application of Targeted Record Swapping and the Cell Key Method were presented for protecting census data, particularly at the 1km x 1km grid cell level, as well as the impact of these methods on safeguarding census hypercubes and their effects on population statistics publications.
- Norwegian census data provided a case study for comparing Eurostat's recommendations on using the cell-key method and targeted record swapping, with alternative methods like small count rounding, to protect dissemination tables from their 2021 population census.
- How a standardized "census-like" dataset was shared across multiple countries to assess disclosure risk in grid data, by evaluating and comparing the risk before and after applying spatial Statistical Disclosure Control (SDC) methods (such as kernel density smoothing and

quad tree aggregation). This allowed analysis of resulting loss in data utility, while facilitating cross-country comparisons.
- The examination of how plausible it might be to potentially re-identify 52 to 179 million respondents from 2010 U.S. Census Bureau data, arguing that such claims stem from a misunderstanding of disclosure definitions.
- Investigating the effectiveness of differential privacy for 2015 Japanese Population Census data, based on international practice, and considering data usability when various differential privacy methods are applied to create statistical tables from census microdata.

Risk assessment: Privacy, confidentiality, and disclosure vs utility

- A range of experiences related to privacy protection and disclosure risk assessment were discussed, including the use of synthetic data, geo-masking methods for location data, differential privacy, and various techniques for safeguarding privacy in different data dissemination contexts.
- Providing synthetic data for research purposes as a means of reducing disclosure risk requires ensuring that the synthetic data distribution closely matches the data distribution in the real data. Comparing densities of real versus synthetic data to assess synthetic data quality could be a relatively easy way to interpret utility and could provide insights to identify discrepancies.
- An example using 2021 UK Census data applied various techniques to safeguard the confidentiality of respondents (record swapping, the cell key method, and disclosure rules in data publication). Using an intruder test involving 30 participants to assess the methods found that more than half of the claims made by intruders were incorrect, demonstrating the effectiveness of such measures.
- Applying Aggregation Equivalence Level (AEL) and Differential Correct Attribution Probability (DCAP) to assess the privacy impact of attribute information in synthetic data, allows non-statisticians to make informed decisions about privacy risks, especially in datasets with relevant attribute information.
- A GPT-based transformer model for generating synthetic microdata, called REaLTabFormer, was described, together with its performance based on a benchmark dataset. That model can create a synthetic census dataset to assess disclosure risk measures and to suggest alternative approaches using a synthetic superpopulation.
- Discussion of whether samples generated from a synthetic population exhibit the same risk and utility relationship as samples from the original population, to provide insights into risk assessment models for real data using synthetic data.
- Addressing concerns regarding reconstruction attacks on census data protected by differential privacy, it was argued that the current confidence-ranking method fails to effectively assess privacy risks and to identify the most vulnerable records for re-identification, as it tends to favour highly repeated records over outliers, which are at greater re-identification risk.
- The use of differential privacy as a privacy model for releasing microdata was explored, comparing it with k-anonymity and discussing its applications in practice, especially in contexts where database queries are not feasible, and evaluating the feasibility of differential privacy for microdata dissemination.
- The impact of noise bounds for protecting statistical confidentiality was examined, particularly in tabular population statistics outputs, highlighting that while bounding noise can control additional disclosure risks, it also offers specific utility benefits.
- The dissemination of geo-referenced agricultural survey data is increasingly requested, but it poses privacy risks. The use of geo-masking methods for anonymizing location data and exploring alternative approaches to mitigate disclosure risks by disseminating spatial variables instead of anonymized coordinates was discussed, along with a framework for assessing spatial signature disclosure risk in agricultural surveys.

Output checking in research data centres.

- Japan's National Statistics Centre shared experiences in checking the output of official statistical microdata for research purposes, with a particular focus on quantiles, and insights and plans for introducing new rules in practice.
- Collaborative efforts in the UK to provide an integrated analysis of output Statistical Disclosure Control (SDC), including theoretical concepts and a classification model for guidelines, aimed at advancing and sustaining this area, were discussed.
- A semi-automated approach to output checking of research data from statistical institutes that combines machine learning models and human expertise (COACH) was presented, which improved the reliability and interpretability of model decision-making in presented use cases.
- The Semi-Automated Checking of Research Outputs (SACRO) project, which aims to develop a versatile and semi-automatic output checking system, was outlined. It's benefits include automating the verification of statistics across various research environments and providing support to researchers using major analytical languages, thus ensuring usability in different secure environments.
- An attempt to automate output checking in the context of a research data centre was presented by CASD, where certain output-document features were automatically extracted and fed to several machine learning models that estimate the risk of disclosure of the concerned output.

Other emerging issues

- Statistics Poland is promoting public awareness of SDC methods and their importance in safeguarding privacy. This involves training statistical staff responsible for data protection, educating data users, and teaching statistics and data management professionals, with various initiatives such as training workshops, and courses designed to enhance understanding and application of SDC techniques.
- Federated Learning (FL) as an approach to decentralized statistical model training, especially useful in scenarios where data cannot be shared due to legal, commercial, or ethical reasons. The feasibility of using FL to generate synthetic microdata was explored, which can allow multiple organizations to contribute to synthetic datasets.
- To examine FL utility from the perspective of national statistical sffices, a Human Activity Recognition dataset was partitioned into subsets, using different aggregation strategies in a distributed environment, to identify the method with the best overall performance.
- Statistics Netherlands created a population-level network dataset, modelling connections between individuals in real-world contexts. An anonymity measure was developed, and a student-led hackathon aimed to assess an attacker's likelihood of acquiring prior knowledge about network structures around an individual. Results indicated varying difficulty in discovering different connection types, highlighting the importance of tailored privacy protection strategies, with social media links being relatively easier to find compared to links related to geographical proximity and household sharing.
- The COVID-19 pandemic accelerated access to medical records for research, leading to the development of platforms like OpenSafely and Trusted Research Environments (TREs) in the UK. However, challenges arise when dealing with highly detailed medical data, such as genomics or medical imaging, necessitating robust solutions for safeguarding both input and output data.

Topics proposed for future work.

9. A general discussion during the meeting and a survey among all meeting participants were organized to collect input for future work. The proposals that were received are presented below (by descending order of votes), and are grouped into themes.

- *Data Privacy vs Utility*
  - Utility measures for tables with suppressed cells;
  - Innovative methods and approaches/examples to ensure a balance between data privacy and data utility;
  - Frameworks for benchmarking/comparing SDC methods and approaches with respect to risk and utility;
  - The practical application of differential privacy methods, and research on their effectiveness in balancing utility and risk.

- *Enhancing privacy when data is connected (across time/space/etc.)*
  - SDC applied to other data sources (including network data, and data connected by space and time);
  - How SDC methods can be applied to time series data, where preserving the privacy of evolving datasets over time is crucial;
  - Showcasing practical examples where privacy protection, risk assessment, and utility evaluation are applied to georeferenced data with a temporal dimension.

- *Synthetic data for official statistics:*
  - Assessing the feasibility and advantages of using synthetic data to create official statistical tables for publication;
  - Analysing and providing critical assessment of disclosure measures applied to synthetic data;
  - Exploring governance challenges related to synthetic data, and whether governance should adapt based on the level of risk and data fidelity;
  - Addressing the synthesis of data with varying structures, such as time-series data or clustered datasets;
  - Presenting real-world examples and projects where synthetic data plays a vital role in preserving privacy while delivering valuable insights;
  - Users' perspectives concerning synthetic data: how useful it is, what can it be used for, whether sufficient protection is perceived, etc.

- *How to communicate SDC to different audiences*
  - What we tell the public about SDC (especially examples of experiences);
  - Discussing communication strategies and experiences related to educating the public about statistical disclosure control methods;
  - Delving into the different audiences who need to be informed about SDC, and exploring effective methods for disseminating knowledge.

- *Data sharing architecture and best practices.*
  - Effective data sharing methods, and what makes a good data sharing agreement;
  - Governance, responsibilities of stakeholders;
  - Discussing best practices and key components of effective data sharing methods and agreements, to promote data collaboration while preserving privacy.

- *Transferability and generalizability*
  - Examining the transferability and generalizability of privacy-preserving methods and models across different datasets and scenarios.

- *Input privacy methods*
  - Showcasing applications and challenges of fully homomorphic encryption in enabling secure data analysis across institutions.

- *Anonymisation of text and other data*
  - Potential practical examples of various techniques, including generalization, noise injection, etc. being applied to ensure that the data is de-identified effectively;
  - Machine Learning (ML) and generative Artificial Intelligence (AI) use to improve or attack SDC methods, and also to assess the privacy and utility of SDC methods applied to ML applications;
  - SDC/privacy aspects of ML and (generative) AI models themselves.