CRP.1
ENGLISH ONLY
7 October 2015

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Helsinki, Finland, 5-7 October 2015)

DRAFT REPORT OF THE MEETING

Prepared by the UNECE secretariat

PARTICIPATION

1. The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality was held in Helsinki, Finland, from 5 to 7 October 2015. It was attended by participants from: Albania, Austria, Canada, Croatia, Denmark, Finland, France, Germany, Hungary, Israel, Italy, Japan, Latvia, Lithuania, Mexico, Montenegro, Netherlands, Norway, Romania, Russian Federation, Serbia, Slovenia, Spain, Sweden, The former Yugoslav Republic of Macedonia, Turkey, Ukraine, the United Kingdom of Great Britain and Northern Ireland and the United States of America, as well as by representatives from the European Central Bank, Eurostat, International Monetary Fund, UNECE, UNMIK, and World Bank. Participants from numerous universities and institutes attended the work session at the invitation of the UNECE secretariat.

ORGANIZATION OF THE MEETING

2. The agenda of the work session consisted of the following substantive topics, the outcomes of which are documented in the annex::

(i) Identifying and measuring risk of statistical disclosure
(ii) New Methodologies for Protecting Data (Disclosure Limitation)
(iii) Preserving Data Quality and Usability in Disclosure-Limited Data
(iv) Access to Statistical Data for Scientific Purposes
(v) Practicum: Case Studies and Software
(vi) Emerging Challenges to Data Confidentiality
(vii) Closing Panel Discussion

3. Marjo Bruun, head of Statistics Finland, opened the workshop and welcomed the participants. Peter-Paul de Wolf (Netherlands) was elected as Chairperson.

4. The provisional agenda was adopted.

5. The following persons acted as Session Organizers/Discussants: Topic (i) – Josep Domingo Ferrer (University Rovira i Virgili, Spain); Topic (ii) – Sarah Giessing (Germany), Topic (iii) – Lawrence H. Cox (United States of America), assisted by Eric Schulte Nordholt (Netherlands); Topic (iv) – Aleksandra Bujnowska (Eurostat) and Annu Cabrera (Statistics Finland), Topic (v) – Eric Schulte Nordholt and Peter-Paul de Wolf (Statistics Netherlands), Topic (vi) – Michelle Simard (Statistics Canada) and Peter-Paul de Wolf, Topic (vii) – Annu Cabrera and Faiz Alsuhail (Statistics Finland).

RECOMMENDATIONS FOR FUTURE WORK

6. The participants reviewed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Mr. Maurice Brandt (Germany), Ms. Aleksandra Bujnowska (Eurostat), and Mr. Krish Muralidhar (Univ. of Oklahoma).

7. The participants considered it useful to continue the exchange of experiences in the field of statistical data confidentiality (SDC), and recommended that a future work session on statistical data confidentiality be convened in 2017. The following topics were proposed:

- Data Sharing and related issues
    - Linkage of administrative data
    - Harmonized methodology for microdata protection - between datasets and countries
    - Access to international and national microdata
- Best practices
    - Best practices for public use files
    - Best practices in remote access systems
    - How users respond to different release mechanisms (public use, scientific use etc.)
    - Practical implementation
    - User perspective – feedback
    - Reproducible research
- Technical issues – Organizational perspective
    - Utility and quality measures for microdata
    - Task force for international guidelines for SDC
    - Use of register and social sciences
    - Census 2021 - SDC issues
- Technical issues – Methodological perspective
    - Geocoding
    - Visualization of geo-data
    - Output checking
    - SDC methods for Big Data
    - Confidentiality for high dimensional tables
    - SDC for (qualitative) administrative data and different data types (social media etc.)
    - Open data / mosaic effect - future availability of information

8. Ms. Lidija Kostovska, Director General of the State Statistical Office offered to host that work session in The former Yugoslav Republic of Macedonia

FURTHER INFORMATION

9. The conclusions reached during the discussion of the substantive items of the agenda are contained in the Annex. All background documents, presentations and the final report for the meeting are available on the website of the UNECE Statistical Division:
http://www.unece.org/stats/documents/2015.10.confidentiality.html

10. On behalf of the participants, Mr. de Wolf expressed his great appreciation to Statistics Finland for hosting this meeting and providing excellent facilities for the work.

ADOPTION OF THE REPORT

11. The participants adopted the present report before the Work Session adjourned.

Annex: Summary of discussions on substantive topics

A. Topic (i): Identifying and measuring risk of statistical disclosure

12. This topic was organized by Josep Domingo Ferrer (University Rovira i Virgili, Spain). It included the following presentations:

- France – Microdata protection: A method that mixes subsampling and calibration
- Univ. Oklahoma & Univ. Rovira I Virgili – Microdata Masking as Permutation
- Universitat Rovira I Virgili – Assessing Disclosure Risk via Record Linkage
- Univ. Skövde – Transparency in microaggregation
- Univ. Manchester – Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables
- Univ. Edinburgh / Univ. Bristol / Glasgow Centre for Population Health – Running an analysis of combined data when the individual records cannot be combined.

13. The following points were raised in the discussions:

- It was emphasised that the permutation method described in the second presentation was a way of assessing disclosure risk, which facilitated comparison between different classes of masking methodologies. However, it was not in itself a means of optimising utility subject to disclosure risk, and it was based upon the worst case scenario, that an intruder possessed both the original and permuted data sets, but lacked only the mapping between them.
- It was further commented that permutation would not affect mean values. However, for large samples, the variance between consecutive elements in ranked data sets was close to zero, and the results are comparable to those encountered when using microaggregation.
- For a multivariate perturbation, although a *linear* inverse relationship exists between information and disclosure risk, the relationship between utility and disclosure risk is not necessarily *linearly* inverse.
- It was clarified that the Data SHIELD programme described in the last presentation was for use with horizontally-partitioned datasets.

B. Topic (ii): New Methodologies for Protecting Data (Disclosure Limitation)

14. This topic was organized by Sarah Giessing (Germany). It included the following presentations:

- Univ. Edinburgh - Utility of synthetic microdata generated using parametric and tree-based methods
- Duke Univ., U.S. Census Bureau & Cornell Univ.– Formal privacy protection for data products combining individual and employer frames
- Universitat Politècnica de Catalunya / Netherlands – Using BCD CTA for difficult tables: a practical experiment with a real Eurostat table
- United States of America – Techniques to apply cell suppression to large sparse linked tables and some results using those techniques on the 2012 (US) Economic Census

15. The following points were raised in the discussions:

- How to ensure the consistency of disclosure-controlled data with national-level published figures: It is possible to freeze figures at the national level.
- When freezing cells, whether to take into account information that has been made public from non-official (e.g., commercial or private sources).
- The advantages and disadvantages of using machine learning methods: CART tends to reproduce

the general structure better than parameter-based methods.

- Comparison of additive versus multiplicative noise methods: For sensitivity queries, Laplace noise often performs well compared to multiplicative noise.
- When Controlled Tabular Adjustment users should use the heuristic Block Coordinate Descent method, rather than optimisation methods. One advantage of the heuristic method is that it is faster.
- The potential for the wide application of approaches presented by the U.S. Census Bureau.
- Whether techniques to apply cell suppression to large sparse linked tables can be:
  - Used for tables with more than 3 dimensions
  - Applied using other applications, such as Tau Argus.

C. Topic (iii): Preserving Data Quality and Usability in Disclosure-Limited Data

16. This topic was organized by Lawrence H. Cox (United States of America), assisted by Eric Schulte Nordholt (Netherlands). It included the following presentations:

- Germany / Duke Univ. – Generating synthetic geocoding information for public release
- Saarland State Univ. Applied Sciences – Anonymization of longitudinal surveys in the presence of outliers
- U.S. Bureau of the Census & Cornell Univ. – Using partially synthetic microdata to protect sensitive cells in business statistics
- Chuo Univ. / Japan – Quantitative Assessment of Data Confidentiality and Data Utility to Create Anonymized Census Microdata in Japan

17. The following points were raised in the discussions:

- For disclosure control methods used with geographical variables, it is important to consider the fact that some research questions involve short-distance factors (e.g., traffic pollution). Synthetic data approaches may not capture such factors unless these variables are included within the data set.
- In the context of geographical variables, an advantage of synthetic data is that they can deny an intruder certainty of the genuineness of any match. However, this depends on the proportion of non-genuine data contained within the data set.
- Regarding the values chosen for those suppressed cells that are replaced by synthetic values, different statisticians have different preferences for selecting these synthetic values, depending on:
  - Whether they wish to preserve the properties of such tabulations compared to other sets of published statistics.
  - Whether these should be chosen to be different to the suppressed values of those cells.
  - The complexity of deriving the values from a set of equations.
- In some countries, legal distinctions may exist between real and synthetic data, even if the synthetic values are quite similar to the real data.
- Clear documentation of the methodology used to reduce disclosure risk can help to reduce the risk that users may misinterpret or misuse such data. Training of users can also be helpful.
- Those who generate and publish research using data should bear some of the responsibility for the correctness of its content.

D. Topic (iv): Access to Statistical Data for Scientific Purposes

18. This topic was organized by Aleksandra Bujnowska (Eurostat) and Annu Cabrera (Statistics Finland). It included the following presentations:

- Eurostat – Access to EU microdata for research purposes

- Finland – Creating a National Remote Access System for Register-based Research
- Denmark – New Nordic model for researchers joint access to data from the Nordic Statistical Institution
- Germany – Circle of Trust for International Microdata Access
- Germany – Virtual Research Environments (VREs) to enable access to confidential data for scientific purposes
- Univ. Edinburgh – Micro, remote, safe settings (safePODS) – extending a safe setting network across a country
- Montenegro – Access to Statistical Data for Scientific Purposes
- Saarland State University of Applied Sciences, University of Dortmund, University of the West of England, UK Data Service - Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use.

19. The following points were raised in the discussions:

- A number of different systems were described for allowing researchers to access or tabulate from microdata, along with some of their advantages and disadvantages, e.g.:
  - The management of hashed keys for matching data from different sources, and whether the use of the same hash key for multiple organisations poses a security risk.
  - Whether to allow researchers to see (but not remove) microdata
  - Whether national (and European Union) laws allow remote access to unaffiliated foreign institutions, or to nationally-affiliated researchers, who are living outside of the European Union. Such laws are a prerequisite for data sharing.
  - Whether to grant access to commercial companies, and private individuals.
- There was debate about whether statistical institutions could take a more open, user-focused stance on data access, based on trust in the users. -
  - The changes that statistical institutions make to data to protect confidentiality may not always damage the data, depending on its intended purpose of use. – Such approaches can be cheaper to implement than sophisticated microdata access systems, which may be prohibitive for some organisations.
  - Statistical Institutions are not always rewarded for openness, and may be sharply criticized for disclosures, so may naturally take a conservative view of the balance between openness and caution.
  - It was suggested that users and respondents should be included within the circle of trust model that was described in Germany's presentation. It s important for respondents to have confidence in the way their data is managed, and for this to be transparent. Indeed, public attitudes to the use of one's data varies by national culture, and by age group.
  - Big data poses a challenge for greater openness, given the possibility of linking data from different sources.
- There was debate over whether utility was an objective or a constraint. – From a mathematical optimisation perspective, there is no difference, however human decision making is often influenced by psychological factors, especially for uncertainties which are not amenable to calculation.

E. Topic (v): Practicum: Case Studies and Software

20. This topic was organized by Eric Schulte Nordholt and Peter-Paul de Wolf (Statistics Netherlands). It included the following presentations:

- Hitotsubashi Univ. & Chuo Univ. – Creating synthetic microdata from official statistics: Random number generation in consideration of Anscombe's quartet
- Germany – A Graphical User Interface to Manage Cell Suppression on Sets of Linked Tables

Using SAS and t Argus

- Netherlands – Public Use Files of EU-SILC and EU-LFS data
- Albania - Statistical data confidentiality and microdata in Albania
- Slovenia – The application for statistical processing at SURS
- Univ. Edinburgh – synthpop: An R package for generating synthetic versions of sensitive microdata for statistical disclosure limitation
- Romania – New online services for accessing PHC microdata for any user
- Norway – The RAIRD Project: Remote Access Infrastructure for Register Data
- Univ. Essex – An Introduction to the Administrative Data Research Network
- Ukraine - Statistical confidentiality assurance framework in State Statistics Service of Ukraine

21. The following points were raised in the discussions:

- When synthetic data is created, it may sometimes be desirable to preserve relationships between household members.
- Post-editing imputation methods (such as nearest-neighbour imputation), were also suggested as way in which intra-household relationships could be represented in synthetic data.
- How to manage cell suppression for tables which are constantly being updated.
- How to model correlations in synthetic data.
- Sub sampling is one method for minimizing disclosure risk that is suitable for use with census data
- It is important to calibrate and test software for generating synthetic data. One method for ensuring that synthetic and real data generate comparable results is to fit generalized linear models to the original and synthetic data, and to then examine any differences in the coefficients derived from each data set.

- It can be helpful to develop criteria for deciding whether aggregate data was "safe enough".
- It would be useful to be able to quantify the benefits delivered to users by remote access systems. Whilst this is difficult to do, one potential measure is the number of analyses undertaken that would not otherwise be possible.
- There was some discussion about the benefits of allowing remote access systems to display microdata, or intermediate files generated in the analysis of the underlying microdata. Although most of the requirements of researchers could be met with the aggregate data, allowing researchers to see the microdata could be helpful for consistency checking, and for having a better intuitive understanding of the data.
- Researchers may wish to use both "public use" as well as "scientific" versions of survey microdata, depending on their requirements, and whether they wish to access data without entering into a detailed user agreement.

- There was discussion of whether separate software applications needed to be developed for remote access systems, as opposed to shared solutions, which would facilitate greater convergence of methods and practices
- Whilst it is possible to share different organizational experiences, in developing systems, these may be implementation-specific, and there may be more value in sharing frameworks and standards, to which different implementations are compatible. It is easier to share software than to share data.
- It was suggested that a generic language or typology for confidentiality would facilitate international cooperation, and sharing of software components between organizations.
- It is also important to have management support from the start when developing projects to share more data with users, and to engage IT departments.

F. Topic (vi): Emerging Challenges to Data Confidentiality

22. This topic was organized by Michelle Simard (Statistics Canada) and Peter-Paul de Wolf (Netherlands). It included the following presentations:

- Canada - Confidentiality on the fly
- Sweden – Protection of frequency tables – current work at Statistics Sweden
- Canada - Development of rules from administrative data

23. The following points were raised in the discussions:

- The conditions of access to the confidentiality on the fly system.
- Whether the "confidentiality on the fly" approach could be extended to general analyses
- The move to open source software, such as R packages facilitates the development and testing of new methods

G. Topic (vii): Closing Panel Discussion

24. This topic was organized by Annu Cabrera and Faiz Alsuhail (Statistics Finland). The panellists were Mr. Mark Elliot (University of Manchester), Prof. Josep Domingo-Ferrer (Universitat Rovira i Virgili) and Dr. Felix Ritchie (University of the West of England). The panellists were asked to consider two questions:

- What are the challenges that new and different forms of data pose to confidentiality?
- How should statistical disclosure control methods, tools and practices be developed in order to tackle these challenges?

25. The following points were raised in the discussions:

- A number of privacy principles are set out in European Union legislation, including the need for consent from data providers, transparency and accountability.
- Finding the right level of anonymization is a challenge. Too little leaves the data insecure, whilst too much inhibits data integration. Anonymization over time is particularly difficult to maintain when longitudinal data sets are linked
- The issues are more related to data management than to statistical production
- The more statisticians try to find new, more interesting data outputs, the higher the risk to privacy
- We are moving to an era characterised by an increasingly intense relationship between people and their data.
- Will preserving anonymity still be relevant in 20 years?
- Data brokers are missing from this discussion, and should be invited to the next work session.
- Statistical organisations have more incentive than private companies to take a conservative view on confidentiality, because they rely more on public trust
- Risk assessment should be based on realistic scenarios, not worst case scenarios
- Researchers tend to be good at data management, and are not malicious hackers
- The notion of reasonable rather than absolute protection should be explicit in new statistical laws.
- Statistical organisations should not be liable for attacks based on information people have shared freely on social media
- It is important to educate people about how to manage the access to their data, particularly on social media