United Nations ECE/CES/2014/15/Add.9

Economic and Social Council

Distr.: General
6 March 2014

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 8 (a) of the provisional agenda
Reports on the work of the Conference of European Statisticians, its Bureau and Teams of Specialists

Report of the work session on statistical data confidentiality

Note by the Secretariat

Summary

This document presents the outcomes of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, held in Ottawa, Canada, on 28-30 October 2013. The main objectives of the work session were to facilitate the exchange of experience and identify best practices in dealing with technical issues related to statistical data confidentiality in national statistical offices. The meeting targeted experts of national and international statistical offices as well as invited academics dealing with statistical disclosure limitation.

This document is submitted to the Conference of European Statisticians for information.

I. Participation

1. The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality was held in Ottawa, Canada, from 28 to 30 October 2013. It was attended by participants from: Australia, Canada, Finland, France, Germany, Japan, Mexico, Netherlands, Norway, Republic of Korea, Russian Federation, Slovenia, Sweden, United Arab Emirates, United Kingdom and United States of America. The European Commission was represented by Eurostat. Representatives of the World Bank and International Monetary Fund (IMF) also attended. Participants from numerous universities and research institutes attended the work session at the invitation of the UNECE secretariat.

II. Organization of the meeting

2. The agenda of the work session consisted of the following substantive topics:

(a) New methods for protection of tabular data or for other types of results from table and analysis servers;

(b) New methods for protection of microdata;

(c) Modes of access to microdata;

(d) The trade-off between quality, utility and privacy;

(e) Confidentiality issues and case studies;

(f) Panel discussion on transparency and/or privacy concepts (such as differential privacy);

(g) Software developments and demonstrations.

3. Mr. P-P. de Wolf (Netherlands) was elected as Chairman. He expressed his gratitude to Statistics Canada for hosting the meeting in Ottawa.

4. Ms. R. Bender, Assistant Chief Statistician, Analytical Studies, Methodology and Statistical Infrastructure Field at Statistics Canada opened the meeting and welcomed participants to the 8th work session in this field. She was pleased to see the large number of countries participating which included the remote participation of the U.S. Bureau for Labor Statistics via WebEx. She noted the continuous changes in technology, data availability and means of communication, which change how we address data issues. Balancing data access and confidentiality remains critical, and methods must evolve to meet new challenges. Academic researchers are key players in the field of confidentiality, and this work session is an ideal opportunity for researchers and national statistical institutes to exchange information and ideas. She wished participants a very good and fruitful work session.

5. The provisional agenda was adopted.

6. The following persons acted as Session Organizers/Discussants: Topic (i) – Ms. S. Giessing (Germany); Topic (ii) – Ms. M. Simard (Canada); Topic (iii) – Mr. P-P. de Wolf (Netherlands); Topic (iv) – Mr. L. H. Cox (United States of America); Topic (v) – Mr. E. Schulte Nordholt (Netherlands); Topic (vi) – Mr. J. Domingo Ferrer (University Rovira i Virgili, Spain); Topic (vii) – Mr. P-P. de Wolf (Netherlands).

III. Recommendations for future work

7. The participants reviewed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Ms. A. Nissenen (Finland), Ms. S. Giessing, and Mr. J. Drechsler (Germany).

8. The participants considered it useful to continue the exchange of experiences in the field of statistical data confidentiality, and recommended that a future work session on statistical data confidentiality be convened in 2015. Statistics Finland offered to host that work session in Helsinki. The following topics were proposed:

(a) Data access on an international level;

(b) Output checking: pros and cons;

(c) Remote access;

(d) Remote analysis and table servers;

(e) New methods for tabular data / microdata;

(f) Case studies and software demonstrations;

(g) Risk and utility measures;

(h) Consistent data release;

(i) Confidentiality as follows;

  (i) in the era of big data;

  (ii) for social networks;

  (iii) for geospatial data;

  (iv) for linked data;

  (v) across different forms of data release;

(j) Transparency;

(k) Privacy models considering data utility.

IV. 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: www.unece.org/stats/documents/2013.10.confidentiality.html

10. On behalf of the participants, Mr. S. Vale (UNECE) expressed his great appreciation to Statistics Canada for hosting this meeting and providing excellent facilities for the work.

V. Adoption of the report

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

Annex

Summary of discussions and main conclusions

A. Topic (i) New methods for protection of tabular data or for other types of results from table and analysis servers

1. Session A

1. The Australian Bureau of Statistics (ABS) has recently developed the TableBuilder and DataAnalyser remote server systems with automated confidentiality routines that allow users to build their own custom tables or undertake for example regression analyses on secured ABS microdata. The presentation described the statistical methodology behind the perturbation and other protection methods used in these systems. The perturbation routines applied in TableBuilder and DataAnalyser are applied not at the unit record level, as is the case with confidentialised unit record files, but at a level of aggregation relevant to the analysis. This results in lower levels of information loss by tailoring the perturbation both to the type of analysis requested and the nature of the underlying data. DataAnalyser is developed in R, so can be shared, whereas TableBuilder uses proprietory software from Space-Time Research.

2. The representative of the University of Kentucky presented a joint work with Oklahoma State University and CSIRO Mathematics on a General Methodology for Masking Output from Remote Analysis Systems. Remote analysis systems are being considered as an effective approach for providing individuals the ability to perform analysis on data held by statistical agencies and to receive the results of such analysis. One of the major obstacles to the wider implementation of these systems is the lack of a masking mechanism that will ensure that the masked response will be both useful and prevent disclosure. It is also desirable that this mechanism is easy to implement and automate so as to minimize human intervention. The presentation suggested a bootstrap mechanism to satisfy these requirements.

3. The representative of the University of Manchester presented work on the development of a flexible table generating server and demonstrated an application of measuring risk-utility comparing different disclosure control methods for census data. They propose measures for disclosure risk and data utility that are based on Information Theory. For flexible table generating, the server has to measure the disclosure risk in the table, apply the disclosure control method and then reassess the disclosure risk. Methods may be applied either to the underlying data used to generate the tables and/or to the final output table generated from original data. Besides disclosure risk, the server should provide measures of information loss comparing the perturbed table to the original table.

4. The German Federal Statistical Office proposed a flexible rounding strategy for ratios also based on stochastic noise. Traditionally, many statistical agencies protect magnitude tabular establishment data by cell suppression. Typical risk concepts, rules and techniques for cell suppression apply to data of an additive nature, i.e. sums of a quantitative variable. For tables presenting means, ratios or other indicators, it is often considered enough to suppress a cell only if it relates to just a single unit. If, however, cell suppression is replaced by a flexible rounding strategy based on stochastic noise, a perturbation might be applied directly to ratios which would reduce the noise variance of the ratio compared to the naïve approach of taking the ratio of the perturbed data. Encouraging first test results were reported, and observed with real life data from German

Tourism Statistics. However, testing of the methodology is still work in progress, especially to prove the concepts suggested for establishing suitable parameters for the method. Another important issue for future work is to test, or eventually extend the method for special, important types of ratio data, like means, i.e. a magnitude variable divided by a count.

5. The CSIRO Computational Informatics presented a paper about their concerns with protecting confidentiality in statistical analysis outputs from a virtual data centre. The presentation described a proposal for a two-stage process involving dataset preparation by the data custodian before loading the data into the virtual data centre and confidentialisation of the analysis outputs by the researcher on removal from the secure environment. The second stage makes use of a checklist developed to assist researchers. However, it would be essential to provide researchers with training in disclosure control as part of the virtual data centre researcher and project approval process so that they could conduct the output confidentialisation themselves.

2. Session B

6. The representative of Statistics Norway presented their work in confidentiality protection of large data cubes. In 2014 all European Union and associated countries will have to submit 60 hypercubes from their 2011 censuses to Eurostat. All the cubes are frequency counts and in most of them the units are individual people. Each country is responsible for disclosure control of their own cubes according to their own legal definitions and with their own choice of disclosure control method. Norway has decided to use a form of balanced rounding of small counts as the preferred disclosure control method. The next step will be to test more advanced methods for the balancing mechanism, like e.g. balanced sampling or mixed integer linear programming.

7. The representative of the US Census Bureau presented lessons his office learned in building a linear programing system prototype of the modernized cell suppression system for his office. The cell suppression problem is mathematically sophisticated, so the development of the methodology for an application built to solve it poses some interesting design problems. By far the most pressing of the problems is a need for overall speed in producing a solution for a particular set of production tables.

8. Controlled Tabular Adjustment (CTA) is a recent approach to compute the closest safe table to the original data, using some distance. Sensitive cells are adjusted either upwards or downwards (binary decision), and the resulting cells have to be accordingly (and minimally) modified to preserve marginal totals. The binary decisions are modelled as disjunctive constraints, CTA resulting in a difficult mixed integer linear problem. The representative of the Universitat Politècnica de Catalunya presented a variant of CTA without binary variables and a software package they have implemented to solve a resulting four-objective optimization problem.

9. The representative of the Statistical Office of the Republic of Slovenia presented a metadata-driven procedure and application for aggregation and tabular protection, as well as the dilemmas and trade-offs of such an approach. The application will be metadata-driven, meaning that there is one general program code which is then parameterised for the particular survey using (process) metadata. The metadata approach causes some additional work for the first time, but for future iterations the metadata can be re-used. For tabular protection the German SAS-Tool which involves Tau Argus will be implemented and used whenever possible. In cases of tables that are very complex, the metadata-driven application will not be a suitable solution, however it should help to construct the input files for Tau Argus. At the same time precision requirements will be implemented and used as an additional rule for primary sensitivity of aggregated data.

10. During the discussion, the following points were raised:

(a) It is difficult to fully protect large linked tables such as those required for the Eurostat Census Hub. All countries protect their tables according to their own rules but there can be inconsistencies and differences in European totals. The system will be evaluated next year based on lessons learned;

(b) Participants discussed the extent to which differential privacy can be seen as a noise addition method;

(c) There is no single "one size fits all" solution for managing disclosure risk. Solutions should be context and problem driven;

(d) The challenges of managing secondary disclosure risks in the context of generating multiple tables on the fly;

(e) The possibilities for greater international collaboration, both in the development of sharable code and tools, and in the specification of additional features that statistical organizations would like to see added to commercial software such as SAS. In this context, the Tau Argus package is being released as open source software;

(f) Where it is not currently possible to share software tools, it may still be possible to share methods.

B. Topic (ii): New methods for protection of microdata

11. Duke University presented their work in developing a fully Bayesian, joint modelling approach for categorical data based on the nested Dirichlet process to protect respondents' privacy in surveys on household data. It induces a two-level clustering structure in modelling the household dataset, trying to simultaneously cluster household members within households, and borrow information across households that have similar clusters. The model is applied to a subset of the Current Population Survey March 2011 household dataset synthesis. The results demonstrate its abilities to preserve marginal and multivariate distributions of all nominal variables (dependence structures among variables), and within household relationships, such as difference in race, age and education.

12. Université Laval presented the problem of releasing microdata from categorical variables in the form of large confidentialized contingency tables. Several algorithms have been designed for this purpose, most of which are based on the idea of multiple imputation. While some offer some strong guarantee of privacy protection, such as differential privacy, others merely rely on the published observations not corresponding directly to real respondents. They took into account the difficulty of offering a confidentiality guarantee when creating synthetic datasets by sampling from posterior predictive distributions, and proposed a possible general mechanism for releasing differentially-private synthetic datasets using this method.

13. The representative of Universitat Rovira i Virgili proposed a framework to attain *k*-anonymity, based on micro-aggregation of confidential data that seeks the lowest possible variability for the confidential attributes, thereby maximizing utility. The proposal can be combined with *k*-anonymity refinements such as *l*-diversity and *t*-closeness, hence yielding simultaneous utility and privacy guarantees. ε-Differential privacy is another privacy model that is often opposed to *k*-anonymity like models. *k*-Anonymity is usually presented as a model that preserves data utility to a good extent but offers only limited privacy guarantees. In contrast, ε-differential privacy provides strong privacy guarantees but only limited data utility. For microdata releases, ε-differential privacy can be seen as a kind of *t*-closeness with a specific distance measure. The proposal to minimize the variability of the confidential attributes can therefore also be applied for ε-differential privacy.

14. The University of Manchester presented an evaluation study to compare statistical disclosure limitation methods for spatial outliers in microdata jointly undertaken with the Artificial Intelligence Research Institute (IIIA) and the Spanish National Research Council (CSIC), Spain. The test dataset for this evaluation study is based on the transportation products of the 2006-2008 combined public use microdata set from the American Community Survey of the United States. The spatial variables are trajectories defined as vectors of coordinates where the first component is the coordinate of place of residence (origin) and the second component is the coordinate of workplace (destination). Variables based on geographical spatial coordinates are particularly prone to disclosure risks since they can easily be visualized when disseminating statistical information through the use of maps. The first stage of the study is an outlier detection algorithm, which accounts for multivariate data and is robust to deviations from normality assumptions. Once outliers are identified, the second stage of the study is to recommend a targeted disclosure limitation method to confidentialize the outliers. The perturbative methods of record swapping and hot-deck with respect to disclosure risk and data utility were compared.

15. The German Institute for Employment Research and Universitat Rovira i Virgili presented their work on evaluating the potential of differential privacy mechanisms for Census data. The large number of available records coupled with a limited set of only a few (often categorical) variables will ensure that most of the cells defined by cross-classifying the different attributes still contain an ample number of records. Thus, the noise that needs to be added to fulfil the differential privacy requirements might have only minor effects on data quality. To enable the release of detailed geographical information they propose a differentially private procedure based on a micro-aggregation algorithm with a fixed minimal cluster size. They evaluated whether meaningful results can be obtained with this approach using administrative data gathered by the German Federal Employment Agency. Detailed geocoding information has been added to this database recently and plans call for making this valuable source of information available to the scientific community. They expect that the proposed micro-aggregation algorithm will enable them to release detailed geocoding information while offering strong differential privacy guarantees.

16. The University of Minnesota presented a joint work on a tripartite collaborative experiment using a ten per cent household sample from the 2011 census of Ireland to estimate risk, mask the data using controlled shuffling, and assess analytical utility by comparing the masked data against the unprotected source microdata. Controlled shuffling exploits hierarchically ordered coding schemes to protect privacy and enhance utility. With controlled shuffling, the lesson seems to be more detail means less risk and greater utility. Overall, despite substantial perturbations of the masked dataset, they find that data utility is very high and information loss is slight, almost imperceptible even for fairly complex analytical problems. IPUMS-International disseminates more than two hundred integrated, confidentialized census microdata samples to thousands of researchers world-wide at no cost. The number of samples is increasing at the rate of several dozen per year, as the process of integrating metadata and microdata is completed. Protecting the statistical confidentiality and privacy of individuals represented in the microdata is vital for the IPUMS project. For the 2010 round of censuses, even greater protections are required, while researchers are demanding ever higher precision and greater utility.

17. The following points were made during the discussion:

(a) Structural zeros are problematic. Further investigations should be made on treating these cases appropriately while creating synthetic datasets;

(b) Differential privacy is defined as participation disclosure, i.e. whether an individual is present or not in a particular data set. However, if participation is mandatory such as in a population census, differential privacy may still be of interest with regard to ensuring confidentiality of outputs;

(c) How to assess goodness of fit of different methods for specific sub-groups? Researchers could help to determine this, particularly in the context of international research data centres;

(d) There remain various legal issues with international research data centres. The Data without Boundaries project is addressing this point;

(e) It is important to ask what the usages are and who are the users when starting research on confidentiality topics. Often one answer is not suitable for all uses or users;

(f) The user community is changing enormously. We are seeing an explosion in terms of users and usage. The challenge is how to provide the different types of data sets needed, whilst maintaining confidentiality. Different balances between utility and privacy may be needed;

(g) As differential privacy was developed for output perturbation, can it be applied to microdata inputs? However, microdata is increasingly being seen as an output, and interactive queries that cannot always be predicted in advance, mean that a differential privacy approach could be relevant;

(h) There is increasing demand for metadata on confidentiality procedures, and transparency can be seen as part of utility, however, the risk to privacy could be increased. An appropriate balance needs to be found.

C. Topic (iii): Modes of access to microdata

18. Papers presented in this session covered remote access within the European context and national modes of access. Two papers were tabled but not presented: Istat on their experience on releasing multiple microdata files stemming from the same survey, and Statistics New Zealand on their Integrated Data Infrastructure (IDI) which is a new feature for the microdata Statistics New Zealand provides access to.

19. The representatives of the German Federal Statistical Office and the Institute for Employment Research, Germany presented their paper on the need to develop a Safe Centre network in order to enrich European research. On the national level, sophisticated approaches to access microdata have been developed, but transnational access and international comparative research are still not easy to do. Two European projects, the ESSnet "Decentralised and Remote Access to Confidential Data in the ESS" (DARA) and "Data without Boundaries" (DwB) are working on improving transnational microdata access in the field of social science research. At the moment, the most common solution is for researchers to travel to the location where the data are stored and work within a Safe Centre. There is, however, a possibility to combine the advantages of Safe Centres and remote access in a single approach of a transnational secure network of Safe Centres. In that case, researchers can access international data from a Safe Centre in their own country. The presentation highlighted the need for an organizational network of Safe Centres bound together by secure remote access connections.

20. Eurostat presented new legal provisions for access to confidential data at the EU level focusing on the modes of access enabled in the new Commission Regulation 557/2013. The advantages and disadvantages of different access systems were analysed from the point of view of their suitability for cross border access to microdata. Due to various changes in technical, organizational and legal environment, the legal basis for access to confidential data for scientific purposes has recently been revised. The main objectives of the new regulation establishing conditions of access to EU confidential data were: to offer more datasets, to allow for new modes of access, to improve access procedures and to adapt them better to the broader legal framework.

21. The representative of the German Federal Statistical Office presented the ESSnet project "Decentralised and Remote Access to Confidential Data in the ESS" (DARA) to improve access to European microdata for scientific purposes. The aim of DARA is to establish a secure channel from a safe centre within a National Statistical Institute (NSI) to the server at the safe center of Eurostat so that researchers can use confidential EU microdata in their own Member States without travelling to Luxembourg. The ESSnet DARA-project team has defined a concept of technical implementation and safety requirements for a remote access system. The concrete task of participating NSIs is to provide a secure channel to guarantee access for data users to the central node and also to provide service and IT-support for the researchers on the local national level. The project team has drafted a handbook with descriptions and guidelines for NSI staff and researchers and an accreditation system for access facilities. For a proof of concept and feasibility, the project team has implemented a remote access pilot with 5 access points in 4 countries, and a central node in France.

22. The German Institute for Employment Research presented the main components of the proof of concept pilot project for a European Remote Access Network (Eu-RAN) within the Data without Boundaries project funded by the European Commission. Eu-RAN will be built around a single point of access that allows connecting researchers from different locations with data sources located in different European countries via secure remote access solutions. Additional tools like a sophisticated user account management, a Virtual Research Environment (VRE) and a Microdata Computation Centre (MiCoCe) complete the services offered by the Eu-RAN. While the goal of this work package is to build and proof a technical concept, technical aspects cannot be addressed without taking into account legal, organizational and financial issues.

23. The National Institute of Statistics of France (INSEE) presented their policy in allowing researchers to access confidential business data and confidential household data. The various ways of accessing data include specific tables made for researchers and access through a safe data centre. Access for foreign researchers, access to fiscal data when linked to statistical data and the use of output checking in a safe data centre were also discussed. The presenter outlined his personal view that output checking is not effective, because researchers could simply commit to memory some specific numbers, or hide some numbers in a specific text. Output checking is also resource intensive, and would focus the blame on the statistical organisation rather than the researcher if confidential data are disclosed.

24. The presentation by the Netherlands described the various options they provide for access to microdata from a historical point of view, including public use files or microdata under contract. On site access is provided, but remote access is becoming increasingly popular with researchers. In addition, co-operation projects in which Statistics Netherlands works closely together with one or more external partners are becoming more common. Different statistical disclosure control rules are applied for the different access methods, and overall responsibility for microdata access is centralised in one unit.

25. The United Kingdom Data Service introduced and shared some practical issues and experiences encountered with everyday output checking requests. They also suggested how practical/contextual and data specific issues may affect decisions on whether to release an output or not. This in turn may raise more questions than answers given the complex nature of statistical disclosure control. They also discussed the mandatory training of researchers to minimize problems.

26. The following issues were raised during the discussion:

(a) How to ensure consistency in output checking. Some organizations have a policy of double checking of outputs, whilst others have an approach based on standards and guidelines;

(b) Linked data can augment problems, as it is increasingly the case that many sources are used for many outputs, rather than the traditional one-to-one relationship. The 2009 UNECE publication on Principles and Guidelines on Confidentiality Aspects of Data Integration was cited as useful guidance in this area. It was said that the linkage itself is not the cause for additional confidentiality issues. The fact that the linked datasets are enriched however, might increase the risk;

(c) The consistency of metadata relating to microdata sets was raised. Eurostat provide standard quality reports to researchers with the microdata;

(d) Remote access is likely to become the main mode of access in future for the research community;

(e) At the European Union level, legal and technological constraints are diminishing, but resource constraints are growing. Funding models are needed for microdata access services;

(f) "Campus files" should be considered as public use files to be used for educational purposes. Depending on the educational goal, the requirements may be different to "ordinary" public use files. It will be difficult if not impossible however, to produce real public use files with enough scientific content to satisfy researchers;

(g) Remote access may offer lower risks than scientific use files because the raw data do not physically leave the organization;

(h) The European Statistical System develops guidelines and organizes training on output checking, to improve consistency.

D. Topic (iv) The trade-off between quality, utility and privacy

27. The mission of national statistical offices is to release high-quality statistical data, without bias or censorship. However, national statistical offices are also responsible to preserve the confidentiality of data pertaining to individuals collected and used for statistical purposes. This is called disclosure limitation and often involves censoring (suppression, topcoding), modifying (perturbation, tabular adjustment) or replacing (synthesizing) original data prior to releasing or providing access to a final data product. Ideally, the final product both exhibits low to acceptable disclosure risk and is a suitable substitute for original data in terms of usability and outputs from statistical analysis of the released data. These two objectives are often in conflict; resolving this conflict is referred to as balancing data confidentiality and data quality. Six contributions to this Session represented a variety of countries and disclosure limitation methodologies.

28. The representative of the University of Kentucky presented their joint work with Oklahoma State University/University of Alicante on the need for a common set of performance metrics to maintain consistency of output from remote access servers. Government agencies release information in the form of tabular data, microdata, and increasingly in recent years, through remote access servers. The different forms of release are intended to satisfy the needs of different users. While this flexible approach for releasing data is a welcome sign, it also creates new issues that must be addressed carefully. Failure to maintain consistency across the different forms of data release could result in losing user trust in the information that is provided. While there are performance metrics associated with each of the individual forms of data release, they are tailored to the specific form of data release. Adopting common performance metrics is necessary in order to maintain consistency across all forms.

29. Minimum distance controlled tabular adjustment (CTA) is a recent perturbation approach for statistical disclosure control in tabular data. CTA looks for the closest safe

table, using some particular distance. The representative of the Universitat Politècnica de Catalunya, Spain, provided empirical results to assess the disclosure risk of the method and summarized the results reported in the paper. A set of 33 instances from the literature and four different attacker scenarios are considered. The results show that, unless the attacker has good information about the original table, CTA has low disclosure risk.

30. The representative of the French National Institute for Statistics and Economic Studies presented the advantages and disadvantages of three methods of dealing with a secondary cell suppression problem in Tau-Argus in terms of quality and usability, and examined the time needed by algorithms to converge against the loss of information contained in suppressed cells. Typical cases are tested on real data sets: two hierarchical variables, sets of linked tables etc. Some guidelines were then proposed to find the best trade-off between quality of released data and time consumption.

31. The representative of the Artificial Intelligence Research Institute presented their work with Stockholm University on the Hellinger distance for measuring information loss in microdata. Literature presents different approaches to measure information loss in data protection. The probabilistic information loss measure is based on comparison of means, variances, covariance, correlation coefficients and quantiles. In a recent paper, the Hellinger distance has also been used to measure information loss. This distance has been used to compare univariate probability distributions and to measure information loss. A set of experiments using numerical files protected using a few data protection methods was presented.

32. The International Monetary Fund presented the policies, procedures, and IT implementations put in place to safeguard confidential data, and reduce the possibility of accidental disclosure, with specific examples from the ongoing data management procedures within the IMF Statistics Department. The topics covered were: Levels of Data Confidentiality - with respect to the commitments with National Authorities, and internal Fund policies; Procedures – omissions and suppression of data to limit residual disclosure, and safeguard information derived only for the purpose of internal analysis; Technology – data management and dissemination system and user-level restrictions to limit the access to confidential data; Challenges – in terms of balancing transparency of data dissemination and the need to safeguard dissemination of unauthorized data.

33. The representative of the United States National Institute of Statistical Sciences explored the use of circuits for analysis of suppressed tabular data. Theoretical methods and software are available for performing optimal complementary cell suppression (CCS) in tables. The released resulting suppression patterns comprise algebraic circuits which define alternative tables for the original table while controlling variation between original and alternative cell values. For an important class of statistical tables including two-way tables, these circuits are simple alternating (+/-) cycles. Suppressed tables are notoriously difficult to analyse statistically. A user with sufficient resources could construct the full set, a subset or a probability sample of alternative tables and analyse these tables, resulting in bounds for or estimates of analytical outcomes for the original table.

34. The following issues were raised during the discussion:

(a) Methodologists don’t necessarily have the power to make decisions on the optimal trade-off between quality, utility and privacy, but should try to influence the decision makers;

(b) We have been creating piecemeal solutions to specific problems for many years. This may not be sustainable. We should look for different and more consistent ways to address confidentiality problems from a wider perspective;

(c) Lack of consistency can be a significant disclosure risk for some methods. This risk can be lower for perturbation methods;

(d) When agencies share data, if they apply different disclosure protection methods there can be a real risk of accidental disclosure;

(e) Why make the user reconstruct the table when we can release one reconstructed table at random? It can achieve consistency and utility at the same time;

(f) The aim of the researcher is to provide the most useful information and at the same time ensure privacy;

(g) The need for perturbation in regressions, particularly for units with high leverage.

E. Topic (v): Confidentiality issues and case studies

35. The representative of the Office for National Statistics of the United Kingdom presented their paper describing an "intruder test" where ONS staff were supplied with actual disclosure-controlled 2011 census tables for their local area and under secure conditions, attempted to identify people or households and consequently discover attributes. He summarised the results and showed how they indicated sufficient uncertainty to validate the effectiveness of the statistical disclosure control policy for 2011 UK population and housing census data.

36. The representative of Sweden presented their experiences of implementing statistical disclosure control with the Bifrost application. Their experiences, both positive and negative, were presented and a solution to the problem of integrating consent control into the process of disclosure control was described. To be able to publish tables where information about businesses can be disclosed, Statistics Sweden has to ask the businesses for written consent. Statistics Sweden has recently put a lot of effort into developing standard tools for different parts of the statistics production process. For statistical disclosure control, they have developed "Bifrost", which includes SAS2Argus, a collection of SAS macros that facilitates the use of the Tau-Argus program via SAS. Copies of SAS2Argus are available for anyone that is interested.

37. Statistics Finland produces around 200 different sets of statistics per year, and thus has a variety of statistical disclosure control (SDC) practices in use. In 2012, the office's internal guidelines on the protection of tabulated personal and enterprise data needed updating. Parts of the old guidelines were too open to different interpretations on the necessity and appropriate practices of SDC. There was strong demand for more harmonised methods and practices inside the office, and also opening the black box of SDC to customers and users of statistics. New guidelines were prepared during 2012 and they came into force in February 2013, an English version is available. The next step towards more harmonised SDC is to complete the adoption of the new guidelines in practice.

38. The representative of Meikai University, Japan, presented an approach to assessing the effectiveness of disclosure limitation methods for census microdata. Following the revision of the Statistics Act, anonymized microdata from official statistics have been released in Japan since April 2009. Six types of anonymized microdata from Japanese official statistics are available. In addition, there are plans to release anonymized microdata from the population census. Several empirical studies on the effectiveness of disclosure limitation methods such as micro-aggregation, additive noise, and data swapping for official microdata have been conducted by the National Statistics Center, Japan. The effectiveness of different disclosure limitation methods including non-perturbative methods

and perturbative methods, and their potential for the creation of anonymized official microdata in Japan, were discussed.

39. Statistics Canada presented the two methods used by the Canadian Centre for Justice Statistics (CCJS) to protect tabular data: the scoring approach and additive controlled rounding. The disclosure method for tables uses a scoring approach whereby all survey variables are assigned a sensitivity score. A table's score is the sum of the scores of the contributing variables. If the score falls above a threshold value, then the table cannot be released. This approach is applied to tables that CCJS produces for its own publications as well as for custom requests that it receives from external clients. CCJS is also making data available through the Real Time Remote Access (RTRA) tool which allows individuals to submit SAS code to produce tables. The RTRA uses controlled rounding to mask sensitive data.

40. The session chair summarised the key points raised:

(a) Linking statistical software packages can improve efficiency, and could be a topic for future meetings;

(b) Asking consent to disclose information – more experiences would be interesting;

(c) Sharing experiences on the implementation of new confidentiality rules – are these rules known by producers and accepted by users?;

(d) Protection methods – how effective are they in terms of balancing risk and utility?;

(e) How to manage statistical disclosure for very sensitive data – can such microdata be released?

41. During the discussion, the following issues were raised:

(a) Maintaining a record of data releases to see if new requests are close to previous ones;

(b) The application of intruder testing to census microdata – some experiments are under way in the United Kingdom;

(c) Noise in data can disrupt correct identification of individuals, which could be an argument for lower levels of aggregation in outputs, e.g. single year age bands rather than 5-year bands;

(d) Disclosure control guidelines need to be very precise and mandatory to avoid diverging interpretations. Different guidelines are needed for personal and business data;

(e) Documenting different disclosure control practices can be a useful step towards harmonisation, as it provides a basis for discussion;

(f) Perceived disclosure can be as harmful as real disclosure, particularly for sensitive data;

(g) Eurostat has recently completed a survey on disclosure control tools and methods in member countries. In most countries the application of tools and methods is the responsibility of the subject-matter units. Clear guidance and simple procedures are therefore important;

(h) The possibilities for international collaboration in the development of tools and methods, for example based on the SAS2Argus macros used in several countries.

F. Topic (vi): Panel discussion on transparency and/or privacy concepts (such as differential privacy)

42. Different views on the issue of transparency versus privacy were presented by the panellists:

(a) As discussed in the panel, transparency refers to the information about SDC methods applied to protect the data; transparency should also be ensured to explain microdata access conditions and requirements;

(b) From the users (researchers) point of view transparency helps to interpret the data correctly. If the information about SDC methods is not provided to the users, a wrong statistical conclusion can be drawn. If the information about SDC method is transparent the users can:

(i) Evaluate the limitations of the released information;

(ii) Correct for the SDC mechanism in the analysis phase;

(iii) The risk of re-identification of statistical unit is very low; moreover, there are not known cases of malicious identification; the known cases are rather due to lack of data security;

(c) From the statisticians' and data owners' point of view:

(i) Transparency may lead to additional disclosure risk; the SDC protection may be redone if parameters of the SDC methods are known to the users;

(ii) In the process of SDC protection of the data statisticians should ensure that the SDC does not change drastically the structure of the data and that data utility is preserved as much as possible;

(iii) There are technological solutions that should help to find out who has breached statistical confidentiality; provision of more data in a more transparent way should be considered ensuring at the same time that confidentiality breaches are appropriately sanctioned;

(iv) The few (zero?) malicious cases of disclosure are due to adequate SDC protection; if the SDC protection were reduced there would be more confidentiality breaches.

43. J. Bambauer – Tragedy of the De-Identified Data Commons: An appeal for transparency and access. Most policy discussions about research data conclude that the privacy risks are too great, and that research data cannot be released safely. The presentation challenged the dominant wisdom, arguing that properly de-identified data is not only safe, but of unmatchable social utility claiming that:

(a) First, many people have misinterpreted the relevant literature from computer science, and thus have significantly overstated the difficulty of anonymizing data;

(b) Second, the available evidence demonstrates that the risks from anonymized data are hypothetical - they rarely, if ever, materialize;

(c) Third, the alternative approach taken with Differential Privacy has extremely limited application. Outside a narrow range of circumstances, it fails to deliver either privacy or utility or both;

(d) Finally, de-identified data is crucial to beneficial social research, and constitutes a public resource - a commons - under threat of depletion.

44. A-S. Charest – Statistical Need for Transparency: Statistical Disclosure Control (SDC) methods always involve a trade-off between privacy protection and data accuracy. I will argue that transparency of the SDC mechanism should be an important part of the data accuracy evaluation. It is necessary not only to provide the analyst with data that is as accurate as possible but also with an appropriate measure of the accuracy of this data. Ignoring the additional variability introduced for the purpose of confidentiality may lead to inaccurate inference. Agencies should thus either provide data which can be correctly analysed in the format in which they are given, or provide enough information for users to take the mechanism into account in their analysis. Examples of both approaches were discussed.

45. L. Cox – Transparency Issues for Tabular Data: Statistical disclosure limitation (SDL) is necessary to protect the confidentiality of information pertaining to subjects of surveys and other statistical inquiries. SDL methods typically remove, modify or replace original data, thus distorting data, total mean square error and, potentially, inference and analytical outputs such as regression coefficients. Transparency in SDL refers to equipping analysts with information about the SDL method and its effects on statistical analysis--provided that this information does not lead to inferential disclosure. Thus, for example, if the SDL introduces bias, inflates variance or attenuates estimated regression coefficients, then the analyst may be provided with estimates of bias, inflation or deflation for the purpose of correcting computed estimates. (Nearly) everyone would agree that transparency is a good thing, but like truth and beauty, can be difficult to assess, obtain and preserve. This is true for tabular data subject to suppression as these data involve complicated relationships between true and alternative values that in some situations may be deconstructed or even undone.

46. During the discussion, the following points were made:

(a) Transparency can therefore be considered as an opportunity (as it increases data utility) and a danger (as it increases the disclosure risk);

(b) The catalogue of SDC methods assigned with the risks related to the transparency could be helpful in taking the decision how much information could be provided to the user;

(c) The information available to the user should be taken into account when evaluating the risk of disclosure;

(d) Particular values of parameters of the SDC methods should be rather kept secret, only general protection methods should be transparent;

(e) In the process of protecting the data the statisticians should ensure that correct inferential analysis can be performed with the protected data;

(f) Effects of the SDC methods on utility should be measured and adequately balanced with the level of protection.

G. Topic (vii): Software developments and demonstrations

47. This session provided an opportunity for organizations to show recent developments in their software. The following presentations and demonstrations were made:

(a) Open source software Argus (Netherlands) – The “Argus-twins” are two packages that have been developed within several European projects (funded by Eurostat). The maintenance of the software has been in the hands of a small group of people, mainly at Statistics Netherlands. Since one of them has reached the age to retire, a shortage of developers/maintainers threatens. Moreover, the Argus software was developed for the

Windows operating system. These two arguments have led to a new project (partly funded by Eurostat) in which the Argus software will be ported to an Open Source version. We will show the ideas behind this porting, discuss the expected advantages and show some preliminary results;

(b) G-Confid: Turning the tables on disclosure risk (Canada) – New business survey confidentiality software G-Confid, created by Statistics Canada was based on methodology used in the old Confid system for cell suppression in tabular data. It can handle any table size & number of dimensions subject to SAS & hardware limitations. This software is user-friendly, has the same look and feel as Statistics Canada's other generalized systems and can incorporate new approaches. The main objective of G-Confid is to provide the appropriate level of protection for confidential cells while minimizing the loss of information. To achieve this objective, G-Confid uses an automated linear programming procedure to optimize residual suppression. The presentation covered the functionality and characteristics of G-Confid, focusing on new features including the MINRESP and SCALECOST functions, as well as new auxiliary functions for analysts;

(c) G-Tab: Generalized Tabulation Tool (Canada) – In the social statistics field in Canada, there is no common process for disseminating tabular information. Standards for table production, confidentiality and quality vary. Several tools already exist to manage the table creation process, while Statistics Canada's new dissemination strategy is putting pressure on divisions to disseminate more information to the public through CANSIM and other on-line tools. The increase in tabular information being offered to the public increases the risk of potential breaches of privacy. Due to the complex nature of many of the surveys, the creation of estimates and quality indicators is complicated and prone to human errors which affect quality. G-Tab resolves the inconsistencies in standards in applying confidentiality and updating confidentiality rules to enable better protection in the changing dissemination environment at Statistics Canada. G-Tab has a fully automated set of confidentiality rules that will help protect the information published via free CANSIM;

(d) Innovative Microdata Access - Confidentialising on the fly (Australia) – The Australian Bureau of Statistics (ABS) has recognised the need to 'rethink' microdata access to meet increasing user demand for access to more detailed unit record data, across a wider array of datasets. Until now, ABS data access strategies have relied predominantly upon the confidentialisation of the microdata prior to their release. ABS has recognised that this approach is not sustainable, due to the considerable burden it places on resources and the protracted length of time to release. In addition, this one-size-fits-all confidentialisation solution has to cover all possible users and all possible analytical uses of a dataset, which therefore results in additional loss of information for users. The ABS has developed a remote execution environment for microdata consisting of two different but complementary services; Survey Table Builder (STB) and Analysis Service (for statistical analysis). STB includes an innovative approach with a routine that dynamically confidentialises tabular output using perturbation techniques prior to it being returned to the user. The primary advantage of this technique is that the impact on overall data quality of the confidentialisation routine is minimised by applying the necessary modifications only to the final output and not to the underlying data source. The Analysis Service functionality includes data transformations and manipulations, summary tables, exploratory data analysis and regression models with graphical and diagnostic outputs. As for the STB, all outputs are confidentialised 'on the fly' to protect data confidentiality, using perturbation methods and other protections.