CRP.1
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

REPORT OF THE DECEMBER 2009 WORK SESSION ON STATISTICAL DATA CONFIDENTIALITY

Prepared by the UNECE secretariat

PARTICIPATION

1. The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality was held in Bilbao, Spain, from 2 to 4 December 2009. It was attended by participants from: Australia, Austria, Bulgaria, Canada, Finland, France, Georgia, Germany, Italy, Japan, Luxembourg, Mexico, Netherlands, Norway, Poland, Portugal, Republic of Korea, Russian Federation, Singapore, Slovenia, Spain, Sweden, United Kingdom and United States of America. The European Commission was represented by Eurostat. Representatives of the Organisation for Economic Co-operation and Development (OECD) and European Central Bank (ECB) also attended. Participants from numerous universities and research 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:

(i) Harmonization of statistical data confidentiality – legal and methodological aspects;
(ii) Synthetic and hybrid data;
(iii) Research data centres and virtual labs;
(iv) Tools and software improvements;
(v) Statistical disclosure control methods for the next census round;
(vi) Case studies;
(vii) Risk/benefit analysis and new directions for statistical disclosure limitation.

3. Mr. Anco Hundepool (Netherlands) acted as Chairman.

4. The representative of the Basque Statistical Office (EUSTAT) opened the meeting and welcomed participants. He stressed the importance of the confidential treatment of data. The information society places increasing demands for data but at the same time protecting confidentiality. Eustat has undertaken several research and development projects to address this.

5. The representative of Eurostat thanked EUSTAT for their hospitality and for facilitating the organization of the Work Session. The framework for statistical work at the level of the European Statistical System has been restructured considerably by the Regulation on European Statistics (Regl. (EC) No. 223/2009) and the Vision on the production methods of EU statistics (COM404/2009). To address better subject-matter and governance issues of methodological work, a new Directors of Methodology (DIME) group has been established that will work closely with the IT Directors group to ensure that good methodological work has reliable IT support.

6. The representative of UNECE thanked the Basque Statistical Office and the Steering Group in undertaking the organization of the meeting. He also drew attention to a publication presenting a new set of guidelines on confidential aspects of data integration which underlines the importance of statistical confidentiality.

7. The following persons acted as Session Organizers/Discussants: Topic (i) – Ms. Aleksandra Bujnowska and Mr. Rainer Muthmann (Eurostat); Topic (ii) – Mr. Josep Domingo Ferrer (University Rovira i Virgili, Spain); Topic (iii) – Mr. Maurice Brandt (Germany); Topic (iv) – Mr. Anco Hundepool (Netherlands); Topic (v) – Mr. Eric Schulte Nordholt (Netherlands); Topic (vi) – Ms. Luisa Franconi (Italy) and Ms. Sarah Giessing (Germany); and Topic (vii) – Mr. Lawrence H. Cox (United States of America).

PUBLICATION OF PAPERS

8. Statistics Netherlands will publish the proceedings of the conference, containing a wide selection of the papers presented. Selected papers will also be invited for publication in the journal "Transactions on Data Privacy". Authors are requested to send their final versions to Anco Hundepool (aj.hundepool@cbs.nl) by 15 January 2010 at the latest. The papers should be sent in both pdf format and the original in either Word or LaTex.

RECOMMENDATIONS FOR FUTURE WORK

9. The participants reviewed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Larry Cox (United States of America), Sarah Giessing (Germany), Josep Domingo-Ferrer (Rovira i Virgili University of Tarragona) and Jean-Marc Museux (Eurostat).

10. The participants considered it useful for national and international statistical offices to continue the exchange of experiences in the field of statistical data confidentiality. The Work Session, therefore, recommended that a future meeting on statistical data confidentiality be convened in 2011, subject to the approval of the Conference of European Statisticians and its Bureau, with a study programme taking into account actual issues of statistical data confidentiality and disclosure control. The following issues were suggested:

- Output checking;
- Output perturbation;
- Statistical disclosure control for municipality level data;
- Transparency:
    - Information loss metrics;
    - User education / data archives and statistical disclosure control;
- Census;
- Software and applications;
- Research and development directions:
    - Open-source statistical disclosure control tools;
- Case studies on "intrusion events";
- Data linking;
- Rules, disclosure risk assessment, differential privacy;
- The effect of pre-release SDC of microdata on complex/multivariate analyses.
- Legal framework comparison of different countries will be very helpful;
- Blending confidentiality and quality.

FURTHER INFORMATION

11. The conclusions reached during the discussion of the substantive items of the agenda are contained in the Annex. All background documents and presentations for the meeting are available on the website of the UNECE Statistical Division (**http://www.unece.org/stats/documents/2009.12.confidentiality.htm**).

12. The participants expressed their great appreciation to the Basque Statistical Office (EUSTAT) for hosting this meeting and providing excellent facilities for their work.

ADOPTION OF THE REPORT

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

ANNEX

SUMMARY OF MAIN CONCLUSIONS REACHED AT THE JOINT UNECE/EUROSTAT WORK SESSION ON STATISTICAL DATA CONFIDENTIALITY

Bilbao, Spain, 2-4 December 2009

Topic (i): Harmonization of statistical data confidentiality – legal and methodological aspects
Session Organizer: Aleksandra Bujnowska, Eurostat (Aleksandra.Bujnowska@ec.europa.eu)
Discussant: Rainer Muthmann (rainer.muthmann@ec.europa.eu)
Documentation: Invited papers by Germany, Eurostat, ECB and UNECE; supporting papers by Republic of Korea and OECD

1. This session considered the issues of harmonization of statistical data confidentiality, taking into account legal and methodological aspects.

2. The UNECE reported on the principles and guidelines on confidentiality aspects of data integration developed by the Task Force composed of experts from different countries and different organizations that highlighted various considerations that should be taken into account when different sources of data are integrated. These considerations relate mostly to the legal requirements, prerequisites and objectives of the statistical data integration. In this respect, the confidentiality aspects are particularly important as the disclosure risk is higher when different sources of data are combined together.

3. Eurostat aims to put for discussion a vision for the harmonization of methods for statistical confidentiality in the European Statistical System (ESS). Dissemination of ESS statistics has always been hampered by the lack of harmonization of SDC methods and practices. The new Regulation on European Statistics and the Vision on the production method of EU statistics constitute an important basis for the reformulation of current practices with regard to the treatment of confidentiality in the ESS. Various legal, administrative and methodological aspects were also reviewed.

4. The methodological aspects of harmonization of statistical data confidentiality were highlighted by ECB. The problem of ensuring the application of harmonized anonymisation measures to an international survey is addressed. As it is often solved by selecting the strictest anonymisation procedures in each country, the resulting data are comparable depending on the level of detail or the amount of information available. A procedure using partially synthetic data is described where differing country practices in terms of data reduction techniques (recoding of continuous and categorical variables) are respected while preserving the information content as much as possible.

5. In Germany, the problem of harmonizing confidentiality methods between Federal Statistical Office (FSO) and the statistical offices of the 16 German federal states requires common guidelines in which statistical methods are defined and brought in line to guarantee a consistent standard of data collecting, processing, publication and dissemination to third parties. A division of labour between the Federal Statistical Office and the statistical offices of the federal states is necessary.

6. The OECD paper presents the project on harmonization of labour force and migration statistics based on microdatasets from the OECD member countries, and the setting-up of a remote access interface. The feasibility of the project is discussed by studying the legal and technical issues that could potentially arise.

7. The paper submitted by Statistics Korea presents the current situation in this country regarding access to microdata. The procedural and methodological issues are discussed in the paper.

8. The following points were raised during the discussion:

- How to ensure consistency between the “federal”, national and international contexts.

- Practices for the release of data sets to different groups of users; scientific use files may contain more information than public use files, to allow research on specific topics, however their use needs to be managed more carefully.
- The need for harmonization of practices and a common legal framework for "federal" statistical systems (as in the case of Germany). However, the federal approach also allows individual agencies to coordinate on issues in specific methodological areas. This approach is also being used within the European Statistical System under the new "sponsorship" model (for example, the Sponsorship on Quality is coordinated by Statistics Norway and Eurostat).
- In a "federal" context, there is a need for both a forum for discussion and a governance framework for methodological work on statistical disclosure control.
- The need to agree with researchers on how to maximize utility: different researchers will have different requirements, but all will want to maximize data availability.
- Removing auxiliary variables may reduce options for future data integration, but this can be mitigated by the use of number keys as unique identifiers.

Topic (ii): Synthetic and hybrid data
Session Organizer/Discussant: Josep Domingo-Ferrer, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain (josep.domingo@urv.cat)
Documentation: Invited papers by Germany, Italy, Spain (2 papers), and the United States of America (2 papers).

9. The papers presented under this topic covered the areas of new methods and related software, differential privacy and applications.

10. Differential privacy was discussed in the two papers from the United States. The underlying principle is that removing a record from a dataset should not affect the statistical properties of that dataset. Perturbation techniques should compensate for the missing record, although if the missing record is an outlier (as is often the case) care is needed to ensure that this is done effectively. Domain pruning, where a proportion of the sparsest data cells are removed, can improve analytical validity, and the quality of synthetic data is strongly affected by the degree of protection of differential privacy. While differential privacy is an interesting new concept, the relevance and applicability of this concept in a statistical disclosure limitation context is still unproven and further investigation is required to evaluate disclosure risk and information loss.

11. Methods for the use of partially synthetic data for scientific use files based on establishment surveys have been developed in Germany. In this approach, only the more sensitive variables (including those with a high disclosure risk) are replaced. The confidence intervals of the original and synthetic variables are compared, with the aim of maximising the overlap, and keeping the ratio of their lengths close to one.

12. The generation of hybrid data using micro-aggregation methods was proposed in the first Spanish paper. Mixing original and synthetic data in a hybrid data set is seen as a way to have a low disclosure risk, whilst maintaining practical utility, particularly concerning higher order moments. This approach partitions records into logical clusters before applying synthetic data generation.

13. The second Spanish presentation concerned how to deal with logical edit constraints in the context of microdata protection. It described a method to ensure the preservation of logical relationships between variables, concluding that micro-aggregation is an appropriate method for cases with edit constraints.

14. The Italian presentation addressed the derivation of artificial data through calibration and empirical copulas, in what was described as a model-free approach, using simple statistical tools. It concluded that in this approach the trade-off between accuracy and protection against disclosure is controlled through compression of univariate supports. Calibration weights obtained by means of moment conditions accommodate the information loss and empirical copulas ensure a growth of computational burden which is

only linear with respect to the number of variables. The ability to satisfactory reproduce position indices, cumulants, correlations and mixed moments involving five variables (at least up to the 8th order) was discussed w.r.t. two extreme simulation settings and a global protection indicator related to simulated data was proposed.

15. The main points raised in the discussion were:

- In differential privacy there are two different research strategies, the cryptography approach assumes that even if intruders know the algorithms they can not crack them without the secret parameters, whereas a more "statistical" approach assumes that information about the data set and the protection applied should not generally be released. The value of releasing this sort of information should be assessed;
- There is some evidence that these strategies may be starting to move a little closer to each other;
- Differential privacy approaches are constrained more by sampling zeros than by structural zeros because every point in the support of the posterior must have positive probability. Structural zeros remove points of support from the posterior without reference to the confidential data. Sampling zeros are part of the confidential data, so if there are many of them--as in very sparse tables--the differentially private release will have a great deal of perturbation unless supplemental tuning is done.
- If the data protector knows the sub-domains relevant to the user, he may microaggregate them separately (as described in the second Spanish paper). The first Spanish paper tries to cater for the case when sub-domains are not known by the data protector (random sub-domains).
- The German approach is based on perturbation of only a proportion of the variables, but the true values of variables are unlikely to be available to intruders, therefore they are confident the data are protected.

Topic (iii): Research data centres and virtual labs
Session Organizer/Discussant: Maurice Brandt, Federal Statistical Office, Germany (maurice.brandt@destatis.de)
Documentation: Invited papers by Canada, Germany (2 papers), Netherlands, United Kingdom and University of Essex; supporting paper by United States of America.

16. All contributions in this session dealt with the improvement of microdata access for the research community, though they focussed on different aspects. Some of the approaches presented are more concrete, for example the implementation of a real time remote access solution in Canada. Others concerned work at earlier, purely methodological stages.

17. Four presentations dealt with the access to national microdata, whereas one paper by Germany and the paper by the Netherlands discussed issues relating to cross national datasets. These two papers described work to produce harmonised guidelines related to decentralized access and for output checking. Improved methods for output checking are a key requirement to improve microdata provision via remote access. In this context, the work described by Canada to automate checking procedures for simple analysis (which can make a big part of research projects) is of great interest.

18. The UK paper outlined a cooperative approach to data security and SDC that involves both the data provider and the data user. Training of researchers to make them more aware of and responsible for the confidentiality of their own results can help to prevent breaches of data confidentiality. On the one hand researchers should understand how data access in managed, and on the other hand they have to be aware of the confidentiality policy of the statistical agencies and that they share the duty to protect data confidentiality. Training of statistical agency staff is also important, particularly concerning automated remote access procedures and (partly automated) output checking.

19. The United States example concerning tax data shows that giving access a limited number of researchers is feasible, but if data requests increase, other ways to provide access will be needed. The research community is not so willing to accept general "public use files" any more.

20. The overall orientation from the presentations seems to promote development in direction of remote access. This is a comfortable way for researchers to access data and can lower the burden for employees in statistical agencies.

21. The following points were raised during the discussion:

- It is very difficult in practice to define rules for automatic output checking. Although automatic checking would significantly reduce the volume of outputs to be checked, it also introduces some degree of risk;
- Is it necessary to check all outputs of "trusted" researchers, or just a sample of them?
- In the experience of Germany, remote execution is more popular with users than "scientific use files", because it allows more disaggregation, particularly by region;
- As more outputs are generated, it becomes increasingly difficult, if not impossible, to ensure there is no disclosure by comparison of different outputs, a certain degree of risk management is necessary;
- It may be worth publicizing the positive results of microdata access in case there is a problem, and practices need to be defended;
- Private sector organisations often face similar disclosure control issues, it might be useful to see if any lessons can be learned from them.

Topic (iv): Tools and software improvements
Session Organizer/Discussant: Anco Hundepool, Statistics Netherlands (aj.hundepool@cbs.nl)
Documentation: Invited papers by Australia, Austria and Spain (2 papers).

22. The papers presented under this topic discussed the various challenges in the development and use of tools to manage statistical data confidentiality. Most of the presentations focussed on open-source and free software.

23. The Australian approach to automatic disclosure protection for flexible table generation is promising and provides solutions to the problems of repeated queries. Cell keys are generated to ensure that individual cells are always affected by perturbation in the same way. An additivity module can be used to preserve perturbed totals, though this can lead to the replacement of structural zeros with non-zero values.

24. The presentation and paper from Austria described their open-source project using a graphical user interface (GUI) which is a user-friendly object-oriented method for microdata protection. This approach uses the "R" programming environment and allows the use of all main methods for microdata perturbation. The application and GUI were demonstrated. Feedback and expressions of interest in cooperation for future development are welcome.

25. The first Spanish presentation reviewed by the University of La Laguna of the various types of open software. It defined the basic concepts of open-source, freeware, and different licensing arrangements. The approach of making software available for free and with open source code, under a "copyleft" type of license, such as GPL (General Public License) was strongly advocated. A case study comparing controlled rounding performance for several software tools was presented.

26. The second Spanish presentation discussed an application for controlled tabular adjustment (CTA), as an alternative to cell suppression. The application can preserve certain values, such as totals, and, using an optimization programme, can be constrained to produce the minimum necessary changes.

27. Points raised during the discussion included:

- Planned extensions to the Australian approach to cover non-frequency tables;

- The response of users to the use of perturbation techniques;
- "R" routines can be integrated with other programming languages;
- Free software is constantly improving, and is becoming a serious competitor to commercial software;
- Cell suppression techniques are difficult to apply to large and complex tables. In such cases controlled tabular adjustment methods may be more appropriate;

Topic (v): Statistical disclosure control methods for the next census round
Session Organizer/Discussant: Eric Schulte Nordholt, Statistics Netherlands (e.schultenordholt@cbs.nl)
Documentation: Invited papers by United Kingdom, United Kingdom/Australia/New Zealand and United States of America; supporting papers by Eurostat/Netherlands and United Kingdom

28. An overview was given of the strategy leading to the decision of record swapping in order to protect UK Census 2011 tabular outputs. This was done on the basis of a thorough evaluation on Census 2001 data considering three short-listed methods in terms of risk and utility. Further work is continuing to determine the specifics of the swapping methodology and the interaction with outputs, table design, thresholds and imputation. An important point is that the record swapping methodology to be used in 2011 will be different to that in 2001. It will be 'smarter', targeting risky records and using swapping levels appropriate to characteristics of the geographical area – size and level of imputation (non-response). So the additional protection that was afforded in 2001 by small cell adjustment is supplied in 2011 by undertaking record swapping in a 'smarter' way, retaining as much data utility as possible.

29. The United Kingdom, Australia and New Zealand all have long histories of carrying out respondent-based censuses, and each have plans for a Census of population and dwellings in 2011. Publishing aggregate or individual data carries the risk that individuals or entities could be identified and confidential information about them could be released. NSIs need to protect the confidentiality of census respondents for a number of reasons. The production and use of official statistics depends on the cooperation and trust of citizens. This trust cannot be maintained unless the privacy of individuals' information is protected. The first obligation of the NSIs in developing SDC solutions is to protect the confidentiality of respondents by complying with legal requirements. In order to meet these obligations, the NSIs need to choose the most appropriate method to use balancing the impacts on disclosure risk and data utility. Factors which influence the choice of SDC method include the lessons learnt from previous censuses, the attitudes and risk thresholds within NSIs, the impact on users and data utility; and the feasibility of approach in conjunction with other census systems and geographic classifications.

30. The U.S. Census Bureau collects its survey and census data under a pledge of confidentiality to the respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. As with most national statistical institutes, the goal is to release as much high quality data as possible without violating the pledge of confidentiality. Disclosure avoidance techniques are applied prior to publicly releasing the data products to uphold the pledge. While data swapping will still be the primary way of protecting Census 2010 and the American Community Survey (ACS) five-year tabular data products, some changes will be made to the Census 2000 disclosure avoidance procedures. These include the generation of partially synthetic data for Group Quarters respondents, an increase in the amount of data swapping and a number of variables in the key used to find households with a disclosure risk, and the development of a Microdata Analysis System.

31. The programme for data dissemination that Eurostat is implementing for the census round 2011 is based on an innovative approach. The basic data input is in the form of hypercubes, which are multidimensional tables. The size of these hypercubes has a relevant impact on confidentiality issues: while for a set of predefined common bi- or tri-dimensional tables the disclosure control for census data could be considered (relatively) feasible to implement, such control becomes a real challenge as more dimensions are added. In order to minimise the risk of disclosure, the size of the hypercubes has been kept as small as possible; nevertheless, it remains very likely that confidentiality methods will have to be applied. If these methods differ from one country to another, the comparability of the data might be

affected. It has thus been considered worthy to explore the margins of action for a common approach at EU level for disclosure control of census data.

32. An initial scoping study is described by the Census Statistical Disclosure Control (CENSDC) task force to assess applications of SDC on pre-defined hypercubes containing census counts with the aim of providing recommendations to Member States for a uniform SDC method. These protected hypercubes could then be used in the flexible table generating package of the Census Hub Project and would allow users to tailor and generate their own tables with data from several Member States. Based on the scoping study, it was clear that the hypercubes as defined by Eurostat were too large to handle most SDC methods and also had very skewed distributions of cell counts. The recommendation of the task force at this stage was for Eurostat to reduce the size of the hypercubes required. Applying ad hoc SDC rules and a post-tabular method within the flexible table generating package would relieve the NSIs of having to protect the hypercubes. Indeed, the Census Hub Project can be developed in such a way that the hypercubes never have to physically leave the NSI, rather the information is accessed remotely according to the definitions provided by the user for their table of interest. To rely solely on 'online' SDC methods for protecting the generated tables means that some form of random noise, i.e. stochastic rounding or perturbation, needs to be applied to final output tables. This approach would improve their quality since it does not exacerbate the SDC issues arising from aggregating perturbed building blocks. Further research is needed to improve the stochastic post-tabular methods and the additivity and consistency of the tables.

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

- How to ensure consistency between data sets released for national users, and those required under international agreements, and specifically how to avoid disclosure through comparison;
- The value of multi-national comparisons of methods: Germany is organising a workshop for German speaking countries in January;
- Overlap control in the American Community Survey.

Topic (vi): Case studies
Session Organizers: Luisa Franconi, Istat, Italy (franconi@istat.it) and Sarah Giessing, Federal Statistical Office, Germany (sarah.giessing@destatis.de)
Documentation: Invited papers by Germany, Italy and Spain; supporting papers by Germany, Italy, Republic of Korea and United Kingdom

34. Three presentations in this session concerned microdata and three tabular data.

35. The presentation by Germany on synthetic data structure files reported on first results of testing a method to generate a partially synthetic dataset using multiple imputation. First experiments were carried out using the standard imputation package IVEware and applying it to data from a monthly manufacturing survey. In a second stage the disclosure risk of the resulting partially synthetic data files was assessed empirically using a record linkage approach. For these matching experiments a real, commercial database was used.

36. The Italian presentation proposed a method for estimating a global measure of the re-identification risk in microdata that makes use of the concept of smoothing in contingency tables and penalized maximum likelihood approach. The issue of smoothing in 2-way sparse contingency tables was addressed and the cross ratio was proposed as a measure of smoothness that can be used also for non ordinal variables. The approach has been applied to the Italian 2001 Census and the Labour Force survey.

37. The Spanish presentation concerned an approach for perturbing the categorical variables of a microdata file together with its application to the Household Environmental Survey conducted by EUSTAT. Two perturbation techniques have been tested: the Invariant PRAM (Post-Randomisation Method) implemented in Mu-Argus software, and a Random Assign function, included in the SAS package, that

preserves statistical distributions and allows for treating the structural zeros of the survey. The presentation also reported on results on the impact of the perturbation applied in terms of number of changes.

38. Another presentation by Germany presented a concept of a program currently being developed in SAS. The goal is to automate looping the ARGUS Modular method for secondary cell suppression across a set of linked tables. When supplied with certain case specific parameters and input data files, this SAS procedure should be generally applicable to any set of linked tables without the requirement of case specific adaptation. The concept was illustrated by the example of tables of the German business tax statistics and those based on the German cattle register.

39. Another project undertaken by Italy was presented which aims at presenting the rationale followed to protect a set of non-nested hierarchical linked tables. The core of the problem is to disentangle a non-nested table into a set of nested ones taking into account the chosen disclosure scenario and release plan; then a general procedure can be defined and applied by a standard software package like Argus to perform the protection. The application to the set of tables stemming from Foreign Affiliates Trade Statistics supplied to Eurostat was presented.

40. The paper by Statistics Korea described the results of a comparison of various linear sensitivity measures and minimum frequency rules to assess primary sensitivity. Results are obtained for tabular data from the Korean survey on industry structure. It recommended employing a linear sensitivity measure rather than increasing the threshold for a minimum frequency rule when the goal is to improve privacy protection.

41. Points raised in the discussion included:

- The need to check the impact of suppression in one table on suppression in other tables in a linked series, and if this is too high, to adjust the table structure where possible;
- The need for data producers to be aware of the risk of residual disclosure where there are non-nested hierarchical variables, as it is not yet possible to fully automate this sort of checking;
- IVEware and standard SAS routines do not seem to cope very well with complex survey structures, but have been successfully applied in more simple cases;
- IVEware can cope with structural constraints, such as minimum employment size, through the bounds statement function;
- Information loss analysis is an area where further work would be useful.

Topic (vii): Risk/benefit analysis and new directions for statistical disclosure limitation
Session Organizer/Discussant: Mr. Lawrence H. Cox, U.S. National Center for Health Statistics
Documentation: Invited papers from the United States of America (3 papers) and the United Kingdom (2 papers).

42. The key themes of this session were how to assess data protection and how to assess the preservation of data quality/usability. This topic includes the preservation of data structures; the key statistics, parameters and relationships within a dataset. It raises questions of how to quantify the extent of preservation, and how much information about the data set can be given to users.

43. The presentation by OptTek Systems (United States) described an enhanced framework and decision system for protecting the confidentiality of tabular data. It presented an optimal method for switching microdata records in terms of maximizing quality, minimizing the number of switches and maintaining confidentiality.

44. The first presentation from the United Kingdom concerned the application of game theory to understanding statistical disclosure events, including how they might happen, what they are and what their consequences might be. Most statistical disclosure control methods focus on the probability of the event happening rather than the likely impact. Risk management should take both into account. Game theory provides a framework for analysing the interactions of agents participating in a disclosure event, assuming

that the agents will try to maximise their outcomes, and can thus help to determine the likely impact of a disclosure event.

45. The role of transparency in statistical disclosure limitation was discussed in the second presentation from the United States. Transparency concerns the release of information on statistical disclosure control processes, from both a risk and a utility perspective. There is a key distinction between regular users, who integrate this information to perform their analyses, and intruders, who try to maximise their knowledge of specific items in the dataset. An important consequence is that statistical disclosure control can be performed in ways that, in terms of computational complexity, are more burdensome for intruders than for legitimate users.

46. The presentation from the United States on a common index of similarity for numerical data masking techniques proposed an independent benchmark approach to assess different statistical disclosure control methods. A common index of similarity based on canonical correlation analysis can provide such a benchmark, is relatively easy to construct, and is widely applicable. However, it is not a measure of security or information loss.

47. The final presentation from the United Kingdom concerned how to assess disclosure risk under misclassification for microdata. It considered links between probabilistic disclosure control and probabilistic record linkage models, and discussed the use of deliberate misclassification as a data perturbation technique.

48. The discussant briefly outlined the main themes of a paper on a comparison of confidentiality protection methods for tabular data in terms of data quality and disclosure protection.

49. The main points raised during the discussion were:

- So far the game theory approach has been limited to considering costs and benefits in terms of rankings, cases with multiple costs and benefits need further research;
- Privacy has traditionally been a fixed line that you don't cross, quantified by policy and legal constraints, is it possible to envisage a different approach which would weigh considerations of quality and confidentiality separately for each output, or even each data value, treating statistical disclosure control as just another statistical function?
- Easing confidentiality restrictions may affect response rates, and could therefore do more harm than good in terms of the volume of data available;
- The context of the data is important when determining levels of confidentiality, as in some contexts, the impact of disclosure on data subjects may be very high;
- Should confidentiality limit the potential "public good" of providing the best possible information?
- Perturbation should not be a major problem for researchers, as data often contain a significant number of errors to start with;
- More could be done to prepare for disclosure events, both in terms of pre-empting them, and having strategies to deal with them;
- Assessments of data utility would benefit from more input from users;
- Research in statistical disclosure techniques is not necessarily reaching all of those who might benefit from it, for example sub-national agencies;

Panel discussion on suppression vs perturbation
Chair: Mr. Lawrence H. Cox, U.S. National Center for Health Statistics
Panelists: Larry Cox and Tanvi Desai (United Kingdom)

50. In view of the success of the panel discussions at the previous Work Session, it was decided to organize one under this topic. Larry Cox spoke on the question of perturbation versus suppression in statistical disclosure limitation, from a methodology perspective. He emphasized the importance of examining disclosure limitation methods in the framework of balancing data confidentiality with data

quality and usability. He categorized familiar methods as *perturbative*, *suppressive* or *substitution* methods, and went on to provide a thumbnail analysis of the disclosure limitation and quality preserving effects of these methods. He suggested that substitution methods, such as controlled tabular adjustment, synthetic data, and shuffling appear to offer a superior combination of protection and quality, and in addition control deviations from important statistical properties of original data and enable quantification and specification of their disclosure limitation and quality properties.

51. Tanvi Desai spoke on the question of perturbation vs. suppression from the analyst perspective. She pointed out that many analysts do not understand statistical disclosure control issues, and often do not thoroughly read documentation before beginning to use data. Therefore they either do not trust data that have been subjected to disclosure control, or they use the data without knowing that they have been adjusted. She pointed out that many analysts tend to reduce the question of suppression vs. perturbation to one of 'missing data vs. inaccurate data'. However it seems that analysts instinctively prefer suppression to perturbation, particularly for tabular output. This may be because it is easier to understand, or because suppressed cells provide some accurate information. She ended by posing questions to the floor on whether, with the increase in secure access facilities, pre-release anonymisation is still a good use of assets, and if it is, how the effects of statistical disclosure control can be better communicated to increase understanding and trust among analysts?

52. The main points raised during the discussion were:

- The problem of aggregation after pre-tabular methods (e.g. in Censuses);
- Some exceptional researchers understand and use perturbed data correctly, others do not;
- We have to show the results of perturbative vs. suppressive methods to the users – an education issue;
- Different kinds of users exist: some advanced users are happy with perturbed methods like PRAM, other users will prefer suppressed data from which they could derive intervals;
- Shall we publish intervals? This could help users, but also involves confidentiality risks;
- Suppression is often fine for tables, but not necessarily for microdata;
- Users are demanding numbers, even if they do not understand what they get;
- Record swapping is difficult to deal with, coarsening gives fewer problems;
- Is rounding a way out of the trouble with suppression applied on tabular data?
- Several numbers for one phenomenon could be a solution, but this will confuse users;
- When we educate people on statistical disclosure control, they may start to reconstruct the original data!
- What do users do to impute for suppressed data, or to predict original data values from perturbed data?
- What ancillary data should the NSI provide to assist analysis?