Report
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
(Manchester, United Kingdom, 17-19 December 2007)

REPORT OF THE DECEMBER 2007 JOINT UNECE/EUROSTAT 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 Manchester, United Kingdom, from 17 to 19 December 2007. It was attended by participants from: Austria, Bulgaria, Canada, Finland, France, Germany, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Republic of Korea, Singapore, Slovenia, Spain, Switzerland, United Kingdom and United States of America. The European Commission was represented by Eurostat. Representatives of the United Nations Economic and Social Commission for Asia and the Pacific (ESCAP) and the 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) Microdata;
(ii) Tabular data protection;
(iii) Applications (including practical implementation of SDC methods, actual issues within NSIs and software);
(iv) Panel discussion on microdata protection and remote access facilities;
(v) Panel discussion on balancing data quality and confidentiality.

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

4. The Dean of the Humanities Faculty of the University of Manchester, Director of Census and Social Methodology of the Office for National Statistics of the United Kingdom, Head of Unit B/5 of Eurostat and the Chief of User Services Section of the UNECE Statistical Division addressed the meeting. They emphasized the importance of international cooperation in the field of confidentiality protection and disclosure control between academics and practitioners. The University of Manchester has developed a tradition of research projects in this area, that are currently undertaken in cooperation with statistical offices of the United Kingdom, Australia and New Zealand. This cross-disciplinary research is crucial to research today. The Eurostat representative outlined current work and future plans of the office in the statistical confidentiality domain, such as on-going work on anonymized European microdata files for research and plans for decentralization of access to confidential European microdata files in safe centres of Member States. He also felt that this meeting can make a positive contribution towards a common definition of disclosure risk and make this definition more objective.

5. The participants also recalled the 6th principle of the Fundamental Principles of Official Statistics, adopted in 1992, that guarantees the protection of privacy with respect to statistical data. The UNECE

representative also informed participants about the recently published Guidelines on Confidentiality and Microdata Access adopted in 1996 by the Conference of European Statisticians and the current work on the Principles for ensuring confidentiality with respect to integrated data sets.

6. The following persons acted as Session Organizers: Topic (i) – Mr. Josep Domingo Ferrer (University Rovira I Virgili, Spain); Topic (ii) – Ms. Sarah Giessing (Germany); Topic (iii) – Mr. Eric Schulte Nordholt (Netherlands); Topic (iv) – Ms. Jane Longhurst (United Kingdom); and Topic (v) – Mr. Lawrence H. Cox (United States of America).

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 Mr. Peter-Paul De Wolf (Netherlands), Ms. Nina Jukic (Slovenia) and Mr. Stefan Bender (Germany).

8. 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 2009, 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 for the discussion:

- Tabular data protection and statistical software:
  - Is it possible to bring tabular data protection into SAS, SPSS or STATA?
  - Safe data centres.
- Microdata protection:
  - Hierarchical data sets (e.g. unified employee-employer data sets) requires two-dimensional protection of businesses and individuals.
  - Small area data sets, including challenges created by the use of GPS for data collection.
  - Biomedical data.
- Remote access;
- Disclosure risk and disclosure control of analytical output;
  - The work is on-going, but there is a need for a broader information flow, and possibly for standards ensuring comparability.
- Risk assessment;
- Methods and applications:
  - Combining different methods
  - Data coming from multiple sources
  - Multiple modes access and the effect on disclosure control
  - Driving criteria for comparisons between different techniques.
- Metadata and data protection:
  - Metadata represent one possible source of information for an intruder.
  - What is the relationship between data protection and metadata?
- Harmonizing disclosure control methods:
  - This may be an idea for a panel discussion: Is it possible to come up with guidelines a compatible data protection for different access channels under different legal settings?
- User perspective:
  - Input from the research community on what issues they are confronted with in disclosure control
  - How to get better cooperation between users and SDC practitioners?

9. It was recommended that an organizing committee of the future work session try to group these topics into a smaller number of themes. It was also recommended that the future work session should provide more time for discussion, and look at alternative ways of presenting numerous supporting papers (poster sessions, etc.). The organization of panel discussions was also recommended as a possible method of work.

10. The participants considered that the following topics may also be of interest, and possibly discussed at one of the future meeting on statistical data confidentiality or a related subject (in order of preference):

- Linking external data;
- Access to business microdata for analysis;
- Intruder aspects;
- Case study on legal aspects in different countries;
- Statistical data confidentiality aspects of population and housing censuses taking into account classical, register-based and continuous censuses;
- Building a standard set of test data sets;
- Statistical disclosure control methodology assessment;
- Safety aspects of analysis results, disclosure assurance of analytical outputs;
- Links coordination with the Comité du Secret;
- Reducing the burden/costs of statistical disclosure control;
- Editing/imputation/SDC

PUBLICATION OF PAPERS

11. Eurostat will publish a special volume in the series Monographs of Official Statistics devoted to this work session, containing a wide selection of the papers presented. The publication will be available in the first half of 2008.

12. The UNECE website set up for the work session will continue to be available at:
http://www.unece.org/stats/documents/2007.12.confidentiality.htm

* * * * *

ANNEX

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

Manchester, United Kingdom, 17-19 December 2007

Topic (i): Microdata

Session Organizer: Josep Domingo-Ferrer (University Rovira I Virgili, Spain)
Documentation: Invited papers by Italy, University of Malaga/Radboud University Nijmegen and PriceWaterhouseCoopers, Sabanci University, University of Kentucky,Duke University, IIIA-CSIC Catalonia and Universitat Rovira I Virgili Catalonia; Supporting papers by Finland, University of Applied Sciences Mainz, Germany, Italy and United Kingdom.

1. The papers presented under this topic focused on the following aspects of microdata:
   - crypto methods for database privacy:
     - D. Galindo and E.R. Verheul (University of Malaga and University Nijmegen) propose a use of cryptographic pseudonyms for microdata sharing systems. This would be useful to calculate correlations between data sources that cannot be otherwise linked.
     - T.B. Pedersen, E. Savas and Y. Saygin (Sabanci University) compare two methods for privacy preserving data mining. They suggest that the encryption base, secret sharing and anonymization approaches may be efficiently used, if they are tailored to specific use cases, like multivariate analysis, clustering, association rule mining, etc. The methodology suggested by the Sabanci University is believed to guarantee that nothing but the final result is revealed.
   - survey-specific anonymization:
     - L. Franconi and D. Ichim (ISTAT) stress that different approaches to ensuring confidentiality, emanating from differences in legislation, practices and cultural approaches, may severely limit the comparability within the resulting data sets. Similarly as harmonization of basic statistical concepts, it is important to harmonize among countries the regulations applicable to microdata release.
     - R. Lenz and H-P. Hafner (University of Mainz and Statistical Office of Hesse) study the risk of re-identification of linked employer-employee data sets. They present a strategy to measure dependencies between employer and employee data and to evaluate the impact of these dependencies on the re-identification risk. Further work will be carried other aspects, like sampling weights and the number of respondents.
     - D. Ichim and L. Franconi (ISTAT) illustrate the statistical disclosure paradigm on the anonymization of the Structure of Earning Survey. Using the global recording and a constrained regression model, only few variables had to be modified, and only those records that represented a high disclosure risk.
   - new methodology for microdata protection:
     - K. Muralidhar and R. Sarathy (University of Kentucky and Oklahoma State University) discuss two methods for data masking. Both the data shuffling and the sufficiency-based linear models (SBLM) approaches minimize the disclosure risk for all sub-domain and preserve the distribution of the data set. Unlike for the data shuffling method, the rank order correlation for SBLM is very different from the original rank order correlation. The SBLM preserves the product moment correlation.
     - J. Domingo-Ferrer, F. Sebé and A. Solanas (University Rovira i Virgili) present and evaluate two heuristics for *p*-sensitive *k*-anonymity which, being based on microaggregation, overcome most of the drawbacks of the current computational approach, while offering a smooth information loss increase as p and k grow.

- disclosure risk assessment:
  - To protect confidentiality, statistical agencies typically alter data before releasing them to the public. Interactive systems for assessing data quality may help the secondary data analysts assess the accuracy of their conclusions. However, J.P. Reiter, A. Oganian and F. Karr (Duke University and National Institute for Statistical Sciences) warn that releasing precise quality measure can threaten confidentiality.
  - J. Nin, J. Herranz and V. Torra (IIIA-CSIC) propose a new method that permits better risk assessment under condition of the record linkage.
  - J. Longhurst and P. Vickers study methods that make use of statistical models to estimate risk from the sample information in particular a probabilistic disclosure risk approach based on the Poisson Distribution and log-linear models (Elamir and Skinner (2006)). The results have shown that it is feasible to assess the risk of a microdata file at the individual level for a 6 and 8-variable key and that the results are robust.
- comparisons of methods:
  - J. Konnu (Statistics Finland) compares the microaggregation and the post-randomization method (PRAM) for protection of personal data. The usefulness of microaggregation lies primarily with numerical data. PRAM has a potential in future when researchers are more familiar with statistics and mathematics.
  - J. Drechsler, S. Bender and S. Rässler (IAB and University Bamberg) discuss advantages and disadvantages of two methods for synthetic data sets: Rubin (1993) that generates fully synthetic data sets and Little (1993) that uses imputation only for selected variables with a high disclosure risk. The partially synthetic data have, as compared to fully synthetic sets, a higher utility as the estimates have lower deviation and a higher confidence interval overlap with the original set. The price for this is a higher disclosure risk.

2. The discussion following the presentations raised further questions:
- What do crypto methods bring to microdata protection and, more generally, to database privacy?
- Who fills the gap between new methodology design and anonymization of specific surveys?
- Comparison of masking methods as to quality vs. specific data uses/surveys needed.
- There are a number of different techniques. How do the statistical offices make their choice? How to define objective measures for comparison of the disclosure protection methods and techniques? Some participants supported the notion of a competition that, in reality, exists in any scientific field.
- If the anonymization methodology or quality measures are published for a survey, disclosure attacks by intruders can be method-specific. How to deal with this?
- It is important to record the procedure for anonymization for the office's internal use. There was not a unanimous opinion, whether these methods can be made publicly available, as they can increase the risk of a malicious attack.
- The terminology emphasizing words "malicious" and "intruder" may not always be appropriate. The threats to disclosure protection can be unintentional, or those who may then consider their reasons as intentional.

Topic (ii): Tabular data protection
Session Organizer: Sarah Giessing, Federal Statistical Office, Germany (sarah.giessing@destatis.de)
Documentation: Invited papers by United States of America (3), University of Southampton, Microsoft Research and University of Manchester; Supporting papers by Netherlands, New Zealand, United States of America and University of La Laguna (Spain)

3. The topics covered in the presentations within this session centred on frequency tables, magnitude tables and web access.
- For frequency tables:

  - C. Dwork et al. (Microsoft Research) suggest a new concept for disclosure risk avoidance for frequency data - "differential privacy" – and techniques to ensure it. These techniques are based on adding noise to Fourier coefficients corresponding to a given contingency table.
  - A new method for assessing disclosure risk (i.e. risk of attribute disclosure) for tables of counts, the subtraction - attribution probability (SAP) method has been proposed by D. Smith and M. Elliot (University of Manchester).
  - N. Shlomo (University of Southampton) compares the performance of several techniques to protect population counts tables with respect to disclosure risk and information loss.
  - Statistics New Zealand uses Random Rounding, a mean cell size rule and a threshold rule for SDC of population count tables. M. Camden et al. calculate measures for utility and safety assessing the quality of this SDC concept.
  - J.J. Salazar (University La Laguna) explains advantages and disadvantages of the mathematical models for Controlled (Integer) Rounding vs.(continuous) Tabular Adjustment.
- For magnitude tables:
  - The U.S. Census Bureau adds noise to the underlying microdata prior to tabulation. The paper by L. Zayatz also addresses other SDC research areas at the USBC like synthetic microdata generation (also used to protect frequency tabular data) and a remote microdata analysis system.
  - L. Cox (National Center for Health Statistics, United States) examines and compares the effects on data quality and usability of two methods for Controlled Tabular Adjustment, one based on LP technology, the other on iterative proportional fitting.
  - Using tabular structures of EIA publications, and artificial microdata, R. Dandekar compares empirically the performance of various methods for tabular data protection, i.e. CTA, USBC's noise method and cell suppression.
  - P.P. de Wolf (CBS Netherlands) discusses a possible way to describe a simple class of linked tables that is often considered at NSIs.
- Regarding web access,
  - The U.S. Department of Agriculture's Economic Research Service has developed web-based data delivery tools for access to farm survey data (M. Morchart, C. Towe)

4. The discussion has raised further questions:
- With respect to "differential privacy" (WP.19):
  - Are there limitations with respect to the number of categories and variables? There is, for example, a limitation in processing social network data, as it is not yet known how to treat social networks. On the other hand the number of variables does not seem to represent a limitation, and can be used for tabulation from a survey with hundreds of variables.
  - Are there other statistical disclosure control methods that could ensure differential privacy?
- With respect to the Subtraction - Attribution Probability (SAP) method (WP.20)
  - How to embed it in the statistical disclosure control strategy? This seems to be a usable tool if the computational problems can be solved
- Is cell suppression the best method to protect population counts tables with respect to disclosure risk and information loss, since the method provides the least distortion (WP.18)? It is not a feasible method for Censuses and the disclosure risk remains high. However, this method is useful in many other situations, for example suppression of cells when releasing a very small number of tables.
- In connection with the comparison of integer controlled rounding vs. continuous adjustment for tabular data (WP.24)
  - Do integrality problems matter for magnitude tables? This is up to the users to decide.
  - Could variable controlled rounding be efficiently modelled? According to the author the controlled rounding provides flexibility, however, the computational burden is quite high.
- The models for controlled tabular adjustments and controlled rounding are closely related, and appeared like one was a sub-model of the other (WP.16 and WP.24).

- What are the user reactions to adding noise to microdata prior to tabulation/aggregation (WP.15)? While there are frequent concerns with users reactions to masking values in avoiding the disclosure, statisticians rarely ask users how they are satisfied with other aspects of the statistical survey, like sampling design.
- Are there any plans for incorporating linked tables cell suppression into τ-ARGUS (WP.21)?
- Further details related to the cell suppression approach within the web-based farms data delivery tools would be of the interest to participants (WP.23).
- Every method attempts to put limits on variance of the resulting data set. However, it seems that in some cases it is possible to have a better estimate of the variance than in others.

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSOs and software)

Session Organizer: Eric Schulte Nordholt, Statistics Netherlands (esle@cbs.nl)
Documentation: Invited papers by Germany/Netherlands, Netherlands, Republic of Korea/United States of America, United Kingdom (2); Supporting papers by Austria (2), Bulgaria, Germany, Spain, United Kingdom (3) and Universitat Autonoma de Barcelona/University of Minnesota

5. Presentations under this topic focused on development of new procedures, software application and their practical implementation in the statistical offices:

- Methods and applications for disclosure control of tabular data:
  - S. Giessing, A. Hundepool and J. Castro (Destatis, Statistics Netherlands and Universitat Politechnica de Catalunya) have developed a new rounding procedure based on restricted controlled tabular adjustment for disclosure control of European Structural Business Statistics (SBS). The τ-ARGUS rounding procedure is being used for the ProdCom statistics. However, due to the higher complexity of data structures, this procedure is not applicable to the SBS case. Authors also outlined a method based on controlled tabular adjustment.
  - A. Toniolo Staggemeier, P. Lowthian and Grant Lee (UK ONS and Space-Time Research Pty Ltd., Australia) present a link between Super CROSS and τ-ARGUS for the purposes of the disclosure control of tabular outputs for the Business Registers Unit of the Office for National Statistics. When implementing such a solution, it is important to understand principles of all components (τ-ARGUS, SuperCROSS, solver(s), etc.), because different vendors implemented specific procedures.
  - Experiments presented by (A.T. Staggemeier, A.R Clark, J. Smith and J. Thompson) show that the Greedy Randomising Adaptive Search Procedure (GRASP) is the preferred solution method, as Ant Colony Optimisation requires too much time for learning to take place. Even GRASP had to be simplified for run times to be reasonable for the larger datasets.
- Methods and applications for disclosure control of microdata:
  - E. Schulte Nordholt (Statistics Netherlands) presented methods developed for access to microdata originating from register-based censuses. The methods alter or suppress sensitive microdata and use μ-ARGUS and τ-ARGUS software.
  - J.J. Kim and D.M. Jeong (U.S. National Center for Health Statistics and Korea National Statistical Offices) presented two probability models quantifying disclosure risk for microdata files. The models assume that only unique persons in the sample files face disclosure risk. The measure of disclosure risk can depend on the number of population uniques, population twins or triplets. One of the two probability models applies to the population unique case and the other to population twins or triplets, etc.
  - B. Meindl and M. Templ (Statistics Austria and Vienna University of Technology) presented a procedure using global recording, local suppression and microaggregation for protection of microdata. The procedure preserves the data quality, while the disclosure risk is minimized. M. Templ also presents a new flexible R-package for the generation of anonymized microdata.

  - B. Todorova (National Institute for Statistics, Bulgaria) describes legislative and managerial measures aiming at implementing the EC regulation on dissemination of unidentified individual data. The NIS of Bulgaria cooperated with the UNESCO Chair on Privacy, in order to acquire the necessary expertise.
  - M. Brandt, M. Konold, R. Lenz and M. Rosemann (Germany) presented a project targeted to protection of longitudinal enterprise microdata. The project is expected to cover other economic statistics in the future.
  - A. Esteve, J. Garcia, J. Spijker (Universitat Autònoma de Barcelona) and R. McCaa (University of Minnesota) present the achievements in the trade-off in privacy and quality in providing access to anonymized microdata samples on the Integrated European Census Microdata (IECM) website. The work represents a joint venture of the IECM project with the Integrated Public Use Microdata Series (IPUMS).
- Development of strategies for statistical disclosure control:
  - J. Longhurst, N. Tromans, C. Young and C. Miller (UK Office for National Statistics) describe the strategy their Office adopted for statistical disclosure control of the 2011 Census. The strategy covers all expected outputs: pre-defined tabular outputs, microdata samples and possibly flexible user defined tabular outputs. The most desirable qualities for the strategy are: maximum data utility; minimum disclosure risk acceptable to users, simple to understand and transparent and easy to implement.
  - M. Mas and C. Prado (Technical Assistance, Vittoria-Gasteiz and EUSTAT) describe a comprehensible data protection policy for the Basque Statistical Office (EUSTAT). The approaches not only cover the published products, but also general data protection.
  - J. Longhurst, C. Abrahams, A. Blake, N. Dattani, M. Grinsted and G. Thomas (United Kingdom) present the outcomes of the review of the dissemination of health statistics. The review provides guidance to suppliers on applying confidentiality when publishing health statistics.
  - F. Ritchie (UK Office for National Statistics) presents ideas on how to combine security, consistency and efficiency in practical statistical disclosure control systems. The paper also provides a common framework for evaluating the disclosure control approaches.

6. Further questions can be raised:
- Many EU-aggregates are blocked and that implies a low usefulness. In the case of ProdCom statistics no aggregates over the products are needed. For the SBS data the situation is more difficult as aggregates over the NACE variable are necessary. Controlled rounding is in both situations the solution and if necessary a variable rounding base can be adopted. Already published national cells should not be rounded in EU publications. The question is whether this leads to too many restrictions. In other words: when is this approach feasible?
- It is interesting to see what probability models can be used for quantifying disclosure risk of a public use microdata file. The probability of a population unique is calculated in the US using the 100 percent census file. Clearly, in general more grouping leads to less uniques. The Census Bureau uses the 1 in 100,000 rule. The question is whether this rule needs to be adapted for use in other countries.
- In the last UK census, users were dissatisfied when the protection level lead to less usability. One of the options to overcome this problem in the future is to make use of the ABS cell perturbation method. However, can one (to the extent possible) guarantee the consistency of all the linked UK tables? How important is the one figure for one phenomenon idea?
- Austria has legal problems with remote access. Possibly also other countries have such legal problems. Could such problems be circumvented in a coordinated European approach?
- Germany has the so-called de facto anonymisation for microdata. How do the results of this technique compare with other countries?
- The UK is active in checking output from microdata analyses more efficiently. Could this be the start of a joint European effort to derive guidelines?

- The Integrated European Census Microdata (IECM) samples are very useful for research. If researchers want more detail than currently available in these samples, could remote access then be an option?

Topic (iv): Panel discussion on on microdata protection versus remote acess facilities

Chair: Jane Longhurst, Office for National Statistics, United Kingdom (jane.longhurst@ons.gov.uk)
Panel members: Paul Jackson, (United Kingdom), Luisa Franconi (Italy), Anco Hundepool (Netherlands)
Documentation: Summary paper by the Chair (WP. 39)

7. The panel discussion session concentrated on three principal issues:
   - Who are the users and what are their microdata needs?
   - How can different access options meet the needs of users? What are the barriers?
     - Masking techniques;
     - Licensing;
     - Remote access.
   - Where should future research be concentrated?

8. The panellists, in their introductory remarks, made the following points:
   - There is an increasing demand for microdata from a variety of users who require more details than National Statistics Institutes (NSIs) are willing to release. The aim of this discussion is to consider ways of improving access to these data while maintaining as much detail as possible.
   - ONS currently publishes a large number of tables and hopes that some of them are useful. However there are many academic researchers who have both skills and resources, in the form of money and facilities, and we should consider whether there are good reasons why they should not be enabled to access microdata and carry out their own small area and longitudinal analyses. There is a case for loosening restrictions for certain academic researchers e.g. releasing local geographic identifiers.
   - Under the UK Statistics and Registration Services Act (SRSA), which will become law on 1 April 2008, there will be "approved researchers" who should be bona fide and working on the correct project. We need to explore ways to supply them with data without incurring excessive risk. Possible methods are larger data archives, remote access, and on-site access.
   - Users of microdata in Italy include lawyers and market researchers. Different users have different needs and types of access. Microdata should be customized to meet user needs, including the provision of teaching aids. Data quality should be assessed from two perspectives: producers in terms of coherence and confidentiality, and users in terms of measure for distortion and data usability. For example, users need to know how much their results differ from the true data.
   - Although remote access is very important, the provision of anonymized microdata is also required. The initial investment needed to supply this needs to be balanced against the value of the service to researchers.
   - Traditionally many large tables are published but there are good reasons to provide microdata. PCs and good software allow for quality research. However there is a conflict between research needs and privacy. Microdata can be provided as heavily protected public use files (PUFs), files under contract or in research data centres (RDCs). Researchers complained they had to travel to use the data even when they are located in the same office! Researchers should be seen as people with whom we can work. Many want and need direct access, seven days a week, 24 hours a day. We have many rich databases and it is a waste of resources not to use them. Remote access is definitely needed. At Statistics Netherlands bona fide researchers get direct access but it is still necessary to check their output.

9. The discussion on managing access to data brought up the following points:
   - There is a wide distribution of where data can be accessed and processed.

- There are many different types of users:
  - The main users of microdata are government departments and the regional development agencies.
  - More people, such as students and schoolchildren, will want access over time to real data (as opposed to synthetic data).
  - Should everyone have the same level of access to data, or should there be a differentiation?
- Remote and licensed access reduces access as a whole.
- There are different practices, but in some countries public use files are placed on the Internet. In addition there is totally automated remote access, safe centres and the provision of bespoke analyses.
- Statisticians compile microdata well, but skills are needed to be able to handle the data.
- Good metadata is also important. But technical skills are needed and you need to be a researcher to do the analysis. Remote access is not free, and researchers generally have to declare to what use the data will be put.
- Most researchers do not want to publish tables, just the conclusions they have drawn from the analysis.

10. The following points were made when discussing disclosure risks related to data access:
- Work has been done on intruder scenarios. The criteria to determine user access should not be related to education or employer. Access should not rule out legitimate researchers and they should be given a measure of confidence. In contrast, it is much more likely that it will be professors or people with PhDs who are also the intruders.
- An assessment should be made of where the risk comes from.
- One opinion is to vet the proposal not the researcher. This implies that researchers would have to write proposals. These would be judged on their feasibility rather than who the researcher is or what their project is.
- Another opinion is whether the data would support the analysis (not the scientific value). In addition to feasibility, the administration should be checked. This is separate from the statistical world. One participant cited an example that a proposal was rejected from a renowned professor who wanted to look at fraud in NSIs.
- Research is not always benign. When estimating the risk, it is necessary not just to consider who the user is but who is paying for the research? If governments are paying, they can also be viewed as users, as long as the "output" doesn't lead to disclosure.

11. The participants also discussed advantages and disadvantages of different data access options:
- When different access options are considered, it is worth bearing in mind that sometimes the staff will do the work for the researcher. Decisions can be made on the appropriate level of competency and there should be a trust element. This takes a lot of resources.
- One possibility is to give everyone the masked data. Using the correct masking technique, co-variances can be maintained and no inflated variances - it is now thought to be viable to use synthetic data. However, researchers often need very detailed data and there always has to be this option. The general public is more likely to want tables and overall figures, though journalists may try to reveal confidential information in order to generate headlines.
- Researchers may find themselves more closely supervised than they are used to and may have privacy issues. An informed decision should be made about this.
- A statistical office can establish disabled or enabled access**:**
  - Disabled access means that the NSI will inspect or approve the program which the researcher can then run. This will usually preclude certain functions.
  - Enabled access means that a researcher defines the analysis required and this is then carried out by the staff of the NSI.
- An example was given of problems in licensing access to medical data as there is so much and they are held in so many places. Researchers have to go through a safety check before they can have their proposal accepted.

- A better definition of remote access is needed. It may mean seeing the actual data on the screen or it may mean submitting a model to be processed.
- A synthetic dataset could be used while waiting for approval.

12. Training of researchers is one method of disclosure prevention:
- Training must be given against accidental or unintentional disclosure. This includes mistakes such as leaving a CD out on a desk or passing data to a graduate student. Academics don't want any perturbation done to the data. There is a problem with public use files giving consistent answers with published tables.
- Training is essential. While masking for social data is good, remote access is better for business data. The type of data needs to be taken into account. Health data is a big problem. Training should be in place earlier on - for undergraduate students. Definitions for such terms as remote access / remote execution can be found in the UNECE glossary.

13. In concluding the panel discussion, the following points were made by the panellists:
- Remote access and remote execution are both used in the Netherlands. Checking users' programs is not practical. Users should definitely be trained in confidentiality. The future of data analysis is with remote access and remote execution. (Anco Hundepool).
- Attitudes differ between cultures. Public use files need to be heavily perturbed and Italian researchers do not want perturbed data. Also different users have different needs. Licensing is the answer for business microdata. (Luisa Franconi).
- If the information produced by an NSI is not relevant then it will be ignored. As a result, the NSI will be ignored leading to budget cuts etc. The UK Data Archive will continue to provide microdata for researchers. The Virtual Microdata Laboratory (VML) at ONS will also continue to be used, and this may be developed to provide remote access in the future. (Paul Jackson).

Topic (v): Panel discussion on balancing data quality and confidentiality

Chair: Lawrence H. Cox, National Center for Health Statistics, United States (lcox@cdc.gov)
Panel members: Krish Muralidhar (Univ. Kentucky), Jerome P. Reiter (Duke University)
Documentation: Summary paper by the Chair (WP. 40)

14. The panel discussion session concentrated on three principal issues:
- Quality/Confidentiality for Tabular Data:
  - Local data quality;
  - Global data quality.
- Masking numerical microdata.
- Fully synthetic vs. partly synthetic data sets.

15. The panellists highlighted the following points in their introductions:
- Data quality can be broken down into 2 classes:
  - Local data quality is defined by characteristics of a small number of cells (individual cell values, time trends).
  - Global data quality refers to characteristics of a dataset or subset (distributional parameters, quantiles, distributional shape).
- There is an overlap, for example, in measurement error and correlation. Quality in a multivariate sense requires relationships between variables such as covariance and regression to be maintained.
- Complementary (secondary) cell suppression preserves local quality for unsuppressed cells but obviously not for suppressed cells. It inhibits analysis as there are holes in the data. For multivariate data the problems are magnified.

- Random rounding and perturbation along with controlled tabular adjustment can preserve well local and global quality for both univariate and multivariate data while there is insufficient information to judge about the strengths and weaknesses of perturbing the underlying microdata.
- Masking techniques can be used to prevent the disclosure of confidential information while at the same time providing the highest level of data utility. Responses using the masked data should be identical to those using the original data. These are usually carried out by empirical measures such as univariate characteristics of the confidential variable, linear and non linear relationships between the variables and the ability to reach the same inferences with the masked and original data.
- Synthetic data can be partially synthetic or fully synthetic. The fully synthetic data can include new units sampled and there is a zero identity disclosure risk. The disadvantages of fully synthetic data are based on the full reliance on a custom model. They also require a large sample size. The partially synthetic data follow the original design and are not fully manipulated by a custom model. A smaller sample size is needed. However, there is still a risk of matching the identity for partially synthetic data.
- The quality of a synthetic dataset needs to be able to lead to valid inferences, allow for calculation of features other than the first 2 moments and replicate the entire model building process.

16. The following issues were raised by the participants during the discussion:

- The USA Census Bureau has produced a synthetic dataset released in October 2007. There are only a small number of users (2 or 3) at the moment but they are providing feedback. A team will perform analysis on a real dataset and a synthetic dataset and compare the results.
- There is opposition from European NSIs to the use of synthetic data. They are more willing to perturb the microdata by techniques such as PRAM. However ISTAT reports that their users are not keen on perturbation but it might be more acceptable for business data than social survey data.
- There are many perception issues with data perturbation. Users don't like the idea of many (or all) of the values not being the true values. A dialogue is required with the user. Do we need to educate the user or educate the data supplier or both? There is a need for transparency here.
- Queries were raised concerning criteria and scores for particular methods. There is a danger of comparing methods without putting them into context. In order to compare methods common, realistic datasets are needed.
- Panellists gave no consideration to licensing, safe centres or remote access and assumed complete reliance on public use files. Are there cases where data quality (for specific users/uses) cannot be maintained using SDC techniques therefore requiring a reliance on access options for managing risk?
- Public use files should come with appropriate metadata and health warnings. For example, what analysis is appropriate (with respect to the sampling methods and the sample size).

* * * * *