NATIONS UNIES **E**

Conseil économique et social

Distr.
GÉNÉRALE

CES/2003/26
20 mai 2003

FRANÇAIS
Original: ANGLAIS

COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE POUR L'EUROPE

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Cinquante et unième réunion plénière
(Genève, 10-12 juin 2003)

RAPPORT DE LA RÉUNION DE TRAVAIL CEE/EUROSTAT D'AVRIL 2003 SUR LA CONFIDENTIALITÉ DES DONNÉES STATISTIQUES

1. La réunion de travail CEE/EUROSTAT sur la confidentialité des données statistiques s'est tenue à Luxembourg du 7 au 9 avril 2003. Y ont participé les représentants des pays suivants: Allemagne, Autriche, Bulgarie, Canada, Danemark, État-Unis d'Amérique, Estonie, ex-République yougoslave de Macédoine, Fédération de Russie, Finlande, France, Grèce, Hongrie, Irlande, Islande, Israël, Italie, Kirghizistan, Lettonie, Lituanie, Norvège, Nouvelle-Zélande, Pays-Bas, Pologne, Portugal, République tchèque, Roumanie, Royaume-Uni, Serbie-et-Monténégro, Slovaquie, Slovénie, Suède, Suisse et Turquie. Des représentants de l'Organisation internationale du Travail (OIT) et de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO) étaient également présents. À l'invitation d'Eurostat, des participants de plusieurs universités et instituts de recherche ont assisté à la réunion de travail.

ORGANISATION DE LA RÉUNION

2. Les représentants d'Eurostat et de la CEE se sont adressés aux participants lors de la séance d'ouverture. Soulignant la nécessité de protéger la confidentialité des données et évoquant l'impact des progrès de la technologie, ils ont rappelé que la coopération internationale était de règle en la matière. Ils ont également attiré l'attention sur l'importance que revêt cette question à la lumière des Principes fondamentaux de la statistique officielle adoptés par la Commission économique pour l'Europe à sa session annuelle de 1991. Il est d'autant plus nécessaire de protéger la confidentialité des données statistiques que des utilisateurs de plus en plus nombreux ont de plus en plus recours à l'Internet pour consulter des données statistiques et que d'autres ensembles de données sont de plus en plus aisément disponibles.

GE.03-31159 (F) 230503 280503

3. Les questions de fond suivantes ont été examinées:

i) Théories et méthodes récentes;

ii) Nouvelles techniques de publication des données;

iii) Questions qui se posent dans le domaine juridique/réglementaire;

iv) Confidentialité des statistiques se rapportant à des petites zones géographiques;

v) Évaluation des risques;

vi) Logiciels pour la limitation des données statistiques.

4. La réunion de travail a été coprésidé par M. Pedro Diaz Muñoz (Eurostat), M. Anco Hundepool (Pays-Bas), M. Lawrence Cox (États-Unis d'Amérique) et M. Jean-Louis Mercy (Eurostat). Les participants suivants ont joué leur rôle d'animateur pour les thèmes i) à vi): M. Lawrence Cox (États-Unis d'Amérique), M. Josep Domingo Ferrer (Espagne), M. John King (Eurostat), M. David Brown (Royaume-Uni), M^{me} Sarah Giessing (Allemagne), M^{me} Luisa Franconi (Italie), M. Julian Stander et M. Ramesh A. Dankedar (États-Unis d'Amérique).

TRAVAUX RECOMMANDÉS POUR LE FUTUR

5. Eurostat a proposé de publier le compte rendu de la réunion de travail dans le cadre de sa série de publications. Cette publication comportera les versions définitives des documents de travail ainsi que des résumés préparés par les intervenants.

6. Les participants ont jugé utile que les services nationaux et internationaux de statistique continuent de procéder à des échanges de données d'expérience dans le domaine de la confidentialité des statistiques. Aussi, ont-ils recommandé que les travaux engagés soient poursuivis sous les auspices communs de la Conférence des statisticiens européens et d'Eurostat. Ils ont proposé que soit créé un groupe directeur chargé de soumettre au bureau de la Conférence une proposition sur les activités futures dans ce domaine. L'examen des questions suivantes a été recommandé:

- Accès à distance en ligne/par l'Internet (techniques et questions d'organisation);
- Sécurité des résultats des analyses, assurances contre la divulgation des produits analytiques;
- Glossaire de termes relatifs à la confidentialité des données statistiques;
- Coordination des liens avec le «Comité du secret»;
- Risques de divulgation, pertes d'informations et possibilités d'utilisation des données;

- Confidentialité des données statistiques recueillies aux fins des recensements de la population et des logements effectués selon des méthodes classiques, sur la base de registres et de façon continue;

- Confidentialité des données mises en tableaux (tableaux de fréquence, etc.).

7. Les participants ont suggéré que soient également examinées les questions suivantes:

- Les intrus; jusqu'où peut-on aller? Contre qui?

- Logiciels de contrôle de la divulgation des données statistiques;

- Moyens de réduire le coût du contrôle de la divulgation des données statistiques;

- La divulgation de données limitées comme substitut acceptable de la divulgation de données brutes à des fins d'analyse;

- Évaluation des méthodes de contrôle de la divulgation des données statistiques;

- Mise en place d'un ensemble type de séries de données d'essai;

- Fichiers synthétiques de données à des fins de formation à la recherche/l'analyse;

- Méthodes et approches nouvelles (mixtes), analyse au centre de recherche et accès à distance pour l'analyse finale;

- Accès à des microdonnées de gestion à des fins d'analyse;

- Protection des données d'entrée/de sortie;

- Scénarios de divulgation/d'intrusion?

8. Les participants ont recommandé que les organisations internationales sous les auspices desquels se tiendront les futures réunions de travail sur la confidentialité des données statistiques choisissent des lieux faciles d'accès pour les participants des pays en transition de manière à encourager une plus large participation de ces pays à ces réunions.

CONCLUSIONS DE LA RÉUNION DE TRAVAIL

9. Les conclusions auxquelles sont parvenus les participants à l'issue de l'examen des questions de fond inscrites à l'ordre du jour sont présentées dans l'annexe (en anglais seulement).

ANNEX

Summary of the main conclusions reached at the April 2003 Joint UNECE/Eurostat work session on statistical data confidentiality

Topic (i): New theories and emerging methods
Discussant: Lawrence Cox (United States)
Documentation: Invited papers by Italy, Spain and United States (2 papers)
Supporting papers by Germany (2 papers), Norway, United States and New Zealand.

1. Papers on this topic drawn upon a diversity of ideas and methods in the mathematical sciences including mathematical statistics, graph theory, optimization theory and mathematical programming, focused on an important area of official statistics. Participants stressed that it is important that the disclosure protection methods work "with a memory", because by combining several independent queries, the intruder may disclose the sensitive information.

2. The participants discussed the statistical versus mathematical approaches to statistical disclosure limitation. In this connection the question was raised: If a protection method relies on deterministic techniques, then how effective is it (how effective can it be) against a probabilistic attack, and conversely? Another question was whether it is preferable to ensure that totals and other statistics are preserved exactly or approximately, e.g., by using deterministic methods, or to preserve expected values of important quantities, e.g., by using randomization.

3. Data quality, utility and analyzability issues were also considered under this topic. The participants considered whether it is preferable to control changes to individual data (e.g. cell values) or to control overall measures of data quality (e.g. mean values). The participants also considered what are the appropriate, usable measures or criteria for judging the effects of statistical disclosure limitation on data quality. The discussion also raised the question whether it is preferable to represent data quality and analyzability criteria as constraints rather than mathematical objectives, which is to say that the two solutions meeting the quality constraints are for most purposes interchangeable.

4. Most of the papers on this topic involved perturbative methods (e.g. cell suppression, controlled rounding, controlled tabular adjustment, etc.). For tabular data, perturbation can be applied either to the tabulations or to the underlying microdata (if available). If applied in a randomized manner to the microdata, then any and all tabulations created by summing from these microdata can be released. Conversely, if tabulations are first perturbed and disclosure-limited, then using certain techniques one can work backwards to produce microdata consistent with tabular cell values.

5. The participants, therefore, discussed whether it is preferable to perturb tabular data or to perturb its underlying microdata. Some participants pointed out that in making the choice of the method, the purpose of disclosure protection has to be taken into account, that is, what data are to be protected, where are the data coming from, etc. This purpose has to be taken in its complexity, because it is not known whether the intruder is focused on microdata or sensitive aggregates. The key importance of microdata protection was stressed in this respect, while it was underlined that protecting microdata is more difficult than to protect tabulations. One of the particular problems with microdata is that they may have a large number of dimensions, which is less often the case with the tabular data.

6. Re-sampling was suggested as one of the alternative methods. The procedure creates datasets, the re-sample, which have the same characteristics as the original survey data. Re-sampling was compared with a traditional method of disclosure control, disturbance with multiplicative error, concerning confidentiality on the one hand and the use of the disturbed data for different kinds of analyses on the other hand.

Topic (ii): New data release techniques
Discussant: Josep Domingo-Ferrer (Spain)
Documentation: Invited papers by Israel, Luxembourg, Spain and United Kingdom
Supporting papers by Denmark and Italy.

7. Some of the presentations touched on the question of technology for remote data access and its fast development. A specific viewpoint was the release of data needed by academics and researchers. While they require a relatively high level of flexibility, trust and few restrictions on data manipulation, this may sometimes be difficult for the statistical offices to align with data protection requirements. While on the one hand it was pointed out that academics and researchers accessing data remotely do not have the time nor intention to identify individuals, on the other hand this assumption may not be valid for other users having access to the same data over the Internet.

8. Another specific issue considered at the meeting was the shift, in some countries, from survey-based to register-based statistics. This is usually created through linking various administrative data sources. While the new techniques allowing linking records from various data sets may seem to be cost-efficient and useful, they represent a new threat from the viewpoint of statistical data confidentiality. In order to achieve a desirable level of data protection, realistic estimates of risks are indispensable. Various methods and techniques for preventing data disclosure through remote access were presented including "blocking at source" (using user authentication, looking for suspicious search strings, etc.).

9. The presentations stressed that the choice of the release policy depends on the data to be released. The choice between "pre-" and "post-" approaches was discussed, and it was suggested that general recommendations could be useful. It was emphasized that post-statistical disclosure control methods are wise under the assumption that users/intruders do not cooperate between themselves. There was a discussion on whether this is a reasonable assumption, and whether it does not call for user pre-SDC methods.

10. Widespread online remote access by a growing number of users, along with the ability of users to link to several data sources with a large amount of data, raise the following basic questions:

- What are the disclosure scenarios that should be considered when empirically computing the disclosure risk using records linkage?
- What is the input on disclosure scenarios to be considered that can be reasonably requested from a standard user?
- How can empirical disclosure risk computation be included in a generic SDC package such as μ-Argus?

11. Participants agreed that the development of remote access systems is inevitable, and it implies the need for automated prevention of disclosure of confidential data. Non-automated ("manual") methods are labour-intensive and difficult to be managed along with on-line remote access. As analytical disclosure risk estimation has been developed for non-perturbative methods (sampling, collapsing, etc.), it would be interesting to be able to assess the disclosure risk for perturbative methods without resorting to record linkage. The suggestion was made to consider the user perspective and perception when choosing between perturbative and non-perturbative methods.

12. There was also a discussion on the perception of respondents. In this connection some participants suggested that the term "data mining" may be perceived quite negatively on the reporting units side, as it suggests that someone is trying to look for their private information. On the data users side, many researchers are taking data mining seriously and they focus on purely scientific purposes.

13. Participants also recommended pursuing the work by identifying the best practices on statistical disclosure control for remote access.

Topic (iii): Emerging legal/regulatory issues
Discussant: John King (Eurostat)
Documentation: Invited papers by Germany, United Kingdom and Eurostat
Supporting papers by Armenia, Germany, Kyrgyzstan and Eurostat.

14. A specific issue discussed at the work session was the European Commission Regulation 831/2002 concerning access to confidential data for scientific purposes. This has a significant implication on Eurostat, which can build on past experiences, like the European Community Household Panel, when a controlled access to the confidential (anonymized) microdata sets was made available. The implementation of the Regulation at Eurostat takes place in close cooperation with the research community as well as the national statistical institutes of the Member States. Concerning the impact on the Member States the Regulation encourages the national statistical institutes and Eurostat to work closely together on developing a system for access to anonymized confidential microdata. Some national experiences were referred to in this connection. Finally the Regulation has an impact on the research community, for which it opens new possibilities and partnerships, but on which it imposes tight discipline and limitations such as the price of the opportunities.

15. Data research centres were one of the issues discussed under this programme element. Experiences with dissemination of microdata stripped of identification were presented and discussed at the work session. With respect to the sensitivity of this issue, very often the access is given only to a restricted category of users who use the workstations within the premises of the statistical offices. Activities aiming at controlled remote access are also under way, while due care is taken to the sensitivity of the remote access to anonymized microdata.

16. It was pointed out that a proper definition and/or agreement on the meaning of the term “scientific purposes” would be needed. Very often it is felt that “scientific purposes” means the use of data by academics, but some opinions were expressed about extending this understanding (e.g. access to students, non-academic researchers, etc.). However, some participants stressed that this issue is very sensitive and it is not always easy to widen this access. There were two aspects discussed in relation to licensing of the institutions/researchers, which, on the one hand may provide optimal data protection, while on the part of researchers, may be perceived as a discriminatory measure.

17. The participants also discussed the technical and legal aspects on controlling what the researchers are doing with the data and whether they respect the rules (rules related to data manipulation as well as with respect to the obligation to safeguard the data against third party intrusion). It was suggested that trust is a very important element in the present practice of providing confidential microdata. This trust has two parts – the trust between statisticians and researchers and the trust between data providers (respondents) and statistical agencies. On the other hand, some participants stressed that there were practical occurrences when the intruders tried to use the access to microdata in order to obtain a commercial advantage.

18. The legislative, administrative and regulatory aspects of statistical data confidentiality were considered in relation to the professional framework. These aspects may have a particular importance in countries with decentralized statistical systems. Some countries reported that they have their statistical acts which protect the position of official statistics including the confidentiality issues, while other countries do not have specific statistics acts and the confidentiality issues are dealt with on the basis of various regulations related to the administrative registers and records. Examples were presented of developing a code of practice and protocols to given activities.

19. Some countries referred to the Fundamental Principles of Official Statistics (adopted by the UNECE in 1992), and these were taken on board when creating their national statistical legislation. The principle of data confidentiality and protection of respondents' privacy is incorporated into these principles.

Topic (iv): Confidentiality issues for small areas
Discussant: Sarah Giessing (Destatis, Germany); Organizer: David Brown, (United Kingdom)
Documentation: Invited papers by United Kingdom (2 papers) and United States
Supporting paper by United States

20. The presentations concentrated mostly on social and demographic data rather than on business-related statistics. Various geo-coding methods and related disclosure risks were discussed.

21. Population and housing census data are typically collected with their geographical references, and the resulting tabulations may be produced for very small areas. This represents a specific challenge for the statistical disclosure protection. Some concrete examples were presented, mostly using the perturbative methods (swapping and controlled rounding), but also non-perturbative techniques (collapsing categories and applying thresholds) on tabulations with totals and some other statistical characteristics calculated before perturbing the data.

22. Some participants pointed out that small area statistics usually concerns information on individuals and small businesses. In the case of the neighbourhood statistics, rounding (controlled rounding, random rounding and related auditing methods) was applied. Aggregation to a higher territorial unit level was one of the suggested solutions to improve reliability of tabulations computed on the basis of perturbed data. It was stressed that in many cases, confidentiality protection has to take place "at source", that is, the government departments and agencies who are the original keepers of the data have the obligation to protect the data vis-a-vis the reporting units. Often, they are not empowered to transfer the data even for statistical purposes. Therefore, it would be the original keepers of the data who will apply the data protection procedures on the data before passing them on to statistical offices.

23. Uncertainty of geographical location is one of the possible means of disclosure protection. This can be achieved, for example, through the classical untargeted swapping of records. The efficiency of local records swapping depends on the frequency of zeros and specific low-frequency values. Some examples were presented on how to assess the impact of the swapping on information loss and how the statistical characteristics and estimates are perturbed. Another issue mentioned was that in disseminating microdata stripped of identification details, the small geographic area information should not be provided – otherwise depending on the size of the area the re-identification risk may be significantly higher.

24. The suggestion was made that when swapping the records, it may be useful to swap them between the blocks with the same statistical characteristics, rather than between the immediate neighbours. The advantage would be the expected lower perturbation on statistical estimates. With respect to confidentiality protection, it was emphasized that only the blocks are to be matched, while the record matching would create a threat to confidentiality protection.

25. ZIP codes are often used to provide geographical references to data for small areas, although the ZIP codes often do not follow any geopolitical areas and were created purely for postal delivery purposes. The use of refined ZCTA codes (refined ZIP codes) allows following the logical boundaries, but may increase the risk of disclosure. It was stressed that the risk grows with a smaller number of households sharing the same code (e.g. less than 500 households).

26. In addition to the above-mentioned perturbative methods, some of the non-perturbative methods were also mentioned (filtering, recoding and top coding, etc.). These also lead to the loss of information, in particular with respect to the restriction of geographic detail and aggregation of the data.

Topic (v): Risk assessment
Discussant: Julian Stander; Organizer: Luisa Franconi (Italy)
Documentation: Invited papers by Israel, Italy (2 papers) and Spain. Supporting paper by Germany.

27. The basic question addressed under this topic was how to decide whether the data file can be released, and if it cannot what should be done so that it can be released. Different aspects of risk estimation were discussed such as choice of appropriate model; optimizing the measures for disclosure risk and information loss and combining them; individual versus global estimates of disclosure risk; measures of disclosure risk for perturbed data; and estimating the risk related to record linkage.

28. Risk estimation in the context of statistical disclosure is usually based on a probabilistic model. Different models were compared at the meeting, some of which were based on Poisson and negative binomial distributions. It was pointed out that it might be beneficial not to focus on a single model, but to consider several models, and to re-evaluate the adequacy of the models on the basis of the data obtained. Some participants suggested considering the inclusion of additional models in the risk estimation part of Argus software.

29. A method aiming at estimating a risk of disclosure for each unit in the sample to be released was presented. The individual risk is a record level probability of re-identification. The assessment of the risk then allows selecting the units (units with individual risk higher than a given threshold), which have to undergo protection. The individual risk model (under the European project CASC) is currently under implementation in μ-Argus. The methodology was presented from a theoretical point of view giving advice also on ways to select the threshold. A second paper on this topic tackled the problem of assessing the behaviour of the individual risk methodology. The study took, as an example, the labour force survey and compared the estimates of the individual risks obtained through the model with real risks obtained through the data from the 1991 population census. There was general agreement that both individual and global risk estimates are needed.

30. A study, which tried to address two problems simultaneously, - achieving a given level of safety and a given level of validity (data utility) was discussed at the meeting. Examples were presented based on decomposing the task into elementary problems along with the sample measures for safety and validity. Future plans include defining future suitable measures for safety and validity and formalizing the elicitation process.

31. Another study presented at the meeting used the graph theory to estimate risks related to records linkage. A 3-phase (according to the hierarchy) algorithm with polynomial run time was also presented. An illustrative application of record linkage for structure of costs survey was demonstrated.

Topic (vi): Software tools for statistical disclosure control
Discussant: Ramesh Dandekar (United States)
Documentation: Invited papers by the Netherlands, United Kingdom, United States and Eurostat. Supporting papers by Germany (2 papers), the Netherlands and United States (2 papers).

32. The presentations focused mainly on practical applications but also discussed some methodological issues and algorithms. Several presentations were made on Argus software. Argus software has two major components. The μ-Argus component is designed to protect sensitive microdata. The τ-Argus component protects tabular data from disclosure. As a part of the CASC project, which is partially funded by the 5th framework program of the EU, both components of the Argus software are undergoing modifications to accommodate increasing complexities associated with protecting microdata and tabular data. The following presentations were made on Argus:

- A microdata protection method in combination with a tabular data protection method to create a fully protected public-use product. The author observed that the collapsing of regional variables using μ-Argus offered the most protection. Follow-up protection using τ-Argus ensured the disclosure-free final product.

- Using τ-Argus software on business statistics addressing four separate areas: (i) confidentiality rules to protect business statistics; (ii) processing of linked tables; (iii) experience with the GHMITER method and; (iv) τ-Argus software.

- Existing as well as planned capabilities for GHMITER software (part of τ-Argus). The software uses the hypercube method for tabular data protection offering the most practical alternative to protect extremely large tabular data sets from statistical disclosure. One of the planned improvements is combining GHMITER with an auditing tool in order to compensate for GHMITER's tendency towards over-suppression of cells.

- Practical application for statistical disclosure control using a microdata protection method in combination with a tabular data protection method to create a fully protected public-use product. The collapsing of regional variables using μ-Argus offered the most protection. Follow-up protection using τ-Argus ensured the disclosure-free final product.

33. A further simplification to the Dandekar/Cox method (2002) was also presented at the work session. The simplified logic allows using the existing computer code of the NSO to generate tabular data. The primary objective of the simplified implementation approach is to make the least number of changes to the existing software code. The simplified procedure could also be used to protect newly developed multi-dimensional, linked tabular data containing a hierarchical structure. The author demonstrated the simplified method on seven different complex test data sets available in the public domain for researchers in tabular data protection.

34. The participants considered a prototype software developed to protect sensitive tabular data based on the synthetic tabular data method proposed by Dandekar/Cox in 2002. The software uses a heuristic algorithm along with the TABU SEARCH method to protect multi-dimensional tabular data from statistical disclosure. The software has a user-friendly interface, which allows efficient operation. The synthetic tabular data are also referred to as the controlled tabular adjustments in the most recent technical discussions.

35. The outcomes of the projects aimed at the development of new methods were presented at the work session:

- A new method using the Data Intruder Simulation (DIS) Algorithm to provide an accurate file-level measure of disclosure risk. The measure is used to calibrate the output from the Special Uniques Detection Algorithm (SUDA) in a high-performance-computing environment. This implementation allows cross-file risk assessment of disclosure control.

- A new data masking procedure for categorical and continuous microdata to balance information loss and statistical disclosure. Maintaining the analytical properties of microdata is always crucial. To achieve that objective, Research Triangle Institute International (RTI) uses a combination of sampling, data adjustment and data imputation techniques to generate public-use microdata files. The new method also addresses the issues arising from data protection from inside intruders. Empirical examples to demonstrate the effectiveness of the proposed method were also presented.

- A new method implemented for statistical disclosure control of microdata containing multiple categorical and continuous variables. The categorical variables are first used to construct multi- dimensional tables for all continuous variables. The sensitive table cells in each table are then identified and linked directly to related records in the microdata file. The paper describes a step-by-step procedure required to create a safe microdata file by using the information contained in the multivariate table structure. Some test results were also provided to demonstrate the effectiveness of this method.

36. There was a discussion on protecting ad hoc tables, and a related issue to publish all available tables through the Internet. The approach depends on publishing policies of individual statistical agencies. The opinion was expressed to protect first the core tables aimed at publishing, and to follow with a “value added” table at a later stage.
