UNITED NATIONS **E**

Economic and Social Council

Distr.
GENERAL

CES/1999/22
6 April 1999

Original: ENGLISH

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Forty-seventh plenary session
(Neuchâtel, Switzerland, 14-16 June 1999)

REPORT OF THE JOINT ECE/EUROSTAT WORK SESSION ON STATISTICAL DATA CONFIDENTIALITY

1. The Joint ECE/Eurostat Work Session on Statistical Data Confidentiality was held in Thessaloniki, Greece from 8 to 10 March 1999. It was attended by participants from: Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Israel, Italy, Latvia, Lithuania, Netherlands, Norway, Romania, Russian Federation, Spain, Sweden, The former Yugoslav Republic of Macedonia, United Kingdom, and the United States. The European Commission was represented by Eurostat, the Institute for Systems, Informatics and Safety (ISIS), and the European Centre for Statistics and Development (CESD - Communautaire). A representative of the Food and Agricultural Organization (FAO) was also present. At the invitation of Eurostat, fourteen research and academic institutes participated as observers.

2. The provisional agenda was adopted.

3. Mr. Anco Hundepool (Netherlands) was elected the Chairperson. Ms. Luisa Franconi (Italy) and Ms. Sarah Giessing (Germany) were elected Vice-Chairs.

4. An opening statement was made by Mr. Karavitis, Secretary General of the National Statistical Service of Greece.

5. The following substantive topics were discussed at the meeting:
(i) New applications of disclosure control methods
(ii) Software and computing developments
(iii) Administration and policy of statistical data confidentiality.

6. The following participants acted as Discussants: Mr. Lawrence Cox (United States) for topic (i); Mr. Josep Domingo-Ferrer (Spain) for topic (ii); and Mr. Gordon Sande for topic (iii).

7. The Meeting recommended that a further Joint UN ECE/Eurostat Work Session on Statistical Data Confidentiality be convened in 2000/2001. It recommended, therefore, that the following text be included in the 2001/2002 Integrated Presentation of the Programme of Work of the Conference of European Statisticians:

2.1 Management of information technology infrastructure.

Activities of the ECE

The Joint UN ECE/Eurostat Work Session on Statistical Data Confidentiality (SDC) will be convened in 2000/2001 to consider:

(i) Applications of SDC methodology and software in economic statistics and social and demographic statistics;
(ii) Impact of new technological developments in software, communications and computing on SDC;
(iii) Progress in the implementation of SDC methods and techniques in transition countries;
(iv) Attitude of users and respondents towards SDC.

8. The delegation of The former Yugoslav Republic of Macedonia offered to host the 2000/2001 Work Session on SDC.

9. The participants expressed their high appreciation and gratitude to the National Statistical Service of Greece for the excellent conditions it provided for the meeting.

10. The main conclusions the participants reached in their discussions are presented in the Annex.

ANNEX

SUMMARY OF THE MAIN CONCLUSIONS REACHED AT THE MEETING

I. New applications of disclosure control methods

1. The discussion under this agenda item focused on the application of disclosure control methods in different statistical areas, such as employment and earnings, population census, and for business statistics. Disclosure control gains special importance when combining data from different sources, e.g. employment and earnings surveys, labour force survey, social security files, and health statistics.

2. Choosing the correct strategy of disclosure control often requires finding the right balance between the usefulness of the data and the information loss due to the application of disclosure control methods. Ideally, the released data must meet the appropriate balance between keeping the information content as high as possible while diminishing the risk of disclosure to an acceptable minimum. Practically, this can be difficult to accomplish and assess. While the aim to release safe microdata sets is the same in different national statistical institutes (NSIs), different disclosure control methods and different definitions of confidentiality could be used. The meeting discussed these different definitions of disclosure risks and estimation models.

3. Different approaches have to be considered for different types of data. Large populations with an inherent dependent structure and characterised by key variables of a categorical nature, such as for social data, present completely different problems compared to small populations with a skewed distribution and mainly continuous key variables, such as for business data. When dealing with social data, categorical key variables allow an approach based on finding unique cases in the data set (i.e. an individual that presents a unique combination of values of the key variables in the sample/population). In business data, continuous key variables make the same concept inappropriate, as practically all the units considered would be unique cases. Such differences influence the disclosure limitation techniques that can be applied and the definition of safety of the microdata set.

4. Another important characteristic of **social microdata** is the hierarchical structure of the data allowing the recognition of *groups* within the population (e.g. households). The user who attempts to breach confidentiality might use the dependence structure of the data and the power of technological tools to match records with equal characteristics. Therefore, the risk evaluation has to take into account the dependence structure to reach a higher level of safety than when considering all cases individually. For social data, most of

the confidentiality protection technologies are focused on avoiding re-identification. Disclosure limitation is often performed by re-categorisation of key variables or local suppression of particularly rare categories of such variables.

5. Concerning social data, methodologies are available to estimate the individual risk of disclosure. It is feasible to evaluate the safety level for each unit in the microdata file to be released. A primary need is to develop ways to implement such methodologies and make them available to the NSI. Such software could be shared among NSIs with each one adopting its own threshold according to the final use of the file (public use file or microdata for research). A potential effect of common methodologies and common software would be to harmonise the different practices, while enhancing each NSI's application of these methods.

6. Statistics Netherlands presented the application of the disclosure control methods on employees' earning data. The use of EDI for the collection of employment and earnings data allows the collection not only of samples of data on employees, but also all the employee records of some firms. In the coming years, it is planned to collect detailed earnings information for all employees in the Netherlands. A practical strategy has been developed for the protection of employee tables using Mu-Argus and Tau-Argus software.

7. The United Kingdom demonstrated the use of the record swapping technique to release census data. Record swapping allows to maximise the degree of flexibility of outputs and to produce consistent output for different geographical areas. It can be a good trade-off between the need to protect data by altering individual records and the need to preserve the data structure. It was mentioned, however, that further studies are needed to check the level of protection reached in the released file.

8. **Business microdata** have several characteristics that require a different approach towards its release and disclosure limitation. For most of the firms, additional information is available and easily accessible, large- and medium-size enterprises can be easily identified by a few key variables, and the probability of being included in a sample can be very high for some firms. Furthermore, the information available on the enterprise level is very precise. Also, the motivation of intruders, the potential benefits as well as the costs of methods to breach confidentiality are much higher in the case of business statistics. The problem of dependencies is present also in the field of business data due to the rising importance of longitudinal and panel data in several areas of economics. NSIs often have a more cautious attitude towards releasing business microdata since the possibility of identification is perceived as too high.

9. Research, testing of new methodologies and software development are essential to reach a solution for business statistics. General software covering different methods would be welcome.

10. Canada demonstrated a method to release microdata files for small businesses. Confidentiality of the data of small businesses can be easier to guarantee since the data do not contain extreme outliers. The most effective part of the presented disclosure control methodology is the variety of methods used which makes it extremely difficult for an intruder to untangle.

11. An important aspect of the disclosure limitation techniques is their impact on **data quality**. Data quality is maintained if the distance between the unmodified data and the modified data (after disclosure control) is small. Statistical measures to describe and summarise this distance were discussed. A data quality analysis may address the original raw data (microdata), the statistics generated from the raw data (macrodata), and the impact on the above at each stage of the modification. The same disclosure limitation method may, however, have a different impact on diverse variables and diverse categories. This makes it quite difficult to predict the influence of a concrete method on data quality.

12. The discussion showed that the SDC applications have reached a well-integrated level with methodology and data use. Moreover, the applications already give feedback to further methodological development.

13. Concerns were expressed, however, regarding the implementation of SDC methodology and the maintenance of the documentation. Uniform use of SDC software can be dangerous. It was pointed out that it is important to examine carefully respondents' needs, and the political viewpoints that can impact the implementation of statistical data confidentiality (SDC) methodology. The discussion also showed the need to further develop risk assessment models, and to take into account users' needs for data analysis. NSIs should assess the "vulnerability" of their programmes to potential technological, political, social or malicious attacks against confidentiality.

II. Software and computing developments

14. The Work Session discussed research topics in the area of confidentiality software. The need to develop evaluation methods for disclosure control techniques and software tools was mentioned as one of the priorities. Both theoretical (i.e. probabilistic) work and simulation work on the security of methods implemented by software packages should be promoted.

15. When evaluating software, a **standardisation** of both the **test data** and the **evaluation criteria** used is essential. This is important in order to ensure the comparability of results. Standard testing procedures should also

be developed. Without common evaluation criteria, it is not possible to compare different techniques and/or microdata and macrodata protection.

16. Performance comparisons of the various software packages would be greatly appreciated by the SDC practitioners. The packages could be compared in terms of speed, applicability, information loss, disclosure risk and software utility. Reducing information loss can be aimed at minimising the number of cells suppressed, the total value suppressed, or total value suppressed of a weight variable. Disclosure risk analysis requires that the protection of linked tables (avoiding the situation where cells suppressed in one table are published in another) and pre-published cells (already published and therefore ineligible for suppression) are taken into account.

17. Developing criteria for measuring the degree of data protection against disclosure requires the analysis of the opportunities and motives of potential intruders, and the modelling of their behaviour. This allows the development of safety criteria reflecting the potential users of a particular data set (e.g. researchers, the general public, etc.) as well as the legal and organisational measures that accompany the release of these data.

18. It is also important to have a relevant knowledge of respondent perceptions about confidentiality. There is a need to know the extent to which respondents understand the confidentiality promises made by statistical institutes. Which variables are seen as sensitive and require the greatest protection, and which data become less confidential with ageing could be examples.

19. An important research task would be improving the ability of software for dealing with complex microdata, tables with hierarchical structure and linked tables. Some concern was expressed about the insufficient ability of currently used software packages to deal with such complex data structures.

20. Automated tools are needed to optimise the application of SDC techniques in order to minimise information loss. One possible solution was proposed using the fingerprint technique. The method aims at finding records with many short unique combinations of key variables (fingerprints), so that appropriate data protection techniques can be applied to make the file safer by reducing the number of fingerprints.

21. An automated cell suppression system ensuring the confidentiality of business statistics was demonstrated. The system takes into account the requirements for historical continuity of suppression patterns and the need to accommodate subject-matter concerns into the design. This method is independent of other processing steps in the survey. It leaves cell suppression independent of the software systems that support the processing of a typical business survey.

22. Attention should be drawn to the design and development of safe access and processing of confidential data in distributed environments. Also, distributing and handling sensitive statistical data over the Internet is becoming a key issue of concern for NSIs. Often greater security is achieved through more difficult and cumbersome access, but that puts an increased burden on the legitimate users of data. Cryptographic techniques can be used to implement new ways of access and communication with the same or higher levels of security than with traditional measures.

23. Cryptography can enhance all three main stages of statistical production (data collection, data processing and data dissemination). In **data collection**, it would allow the burden placed on the data collector to be reduced, or even data to be collected remotely over Internet in a secure way. In **data processing**, it would enable the secure storage, distribution and handling of data. In **data dissemination** it could be used to ensure copyright protection and electronic (micro-) sale of statistical information. Cryptography also allows some shortcomings in the area of statistical data protection to be overcome that are not covered by statistical disclosure control, namely the need to provide the general public with data which are disclosure-protected but still usable to obtain exact statistics.

III. Administration and policy of statistical data confidentiality

24. The discussion revealed that confidentiality problems have recently become very important for several reasons. The most important are the following: identification is easier due to more sophisticated and widely available technology; more microdata exist in government and business circles; fears of disclosure are greatly diminishing the public's trust and cooperation; and the motivation for identification has increased. The concerns are heightened by privacy threats due to the use of new technologies such as Internet, and intruders who might have new tools to break disclosure protection.

25. NSIs must continuously address the public's concerns that the level of data protection is not sufficient. Once public trust is lost, it becomes a huge obstacle to gain future cooperation in surveys. The perception that the agency cannot be trusted can be as damaging to future response rates as an actual breach. An important factor in the perception of confidentiality is the degree of knowledge about procedures used to process and protect information. NSIs make a considerable effort to protect against disclosure but often do not provide enough resources to prepare and deliver messages that can effectively deal with perceptions.

26. An important part of the NSI's general policy for data protection and confidentiality could be a **data security plan**. The plan should specify the layers of confidentiality and security, and the data classes belonging to each

layer. There could also be a difference in securing information for storage and for communication. Research and development in this area is hindered by the lack of consensus on the security requirements of NSIs. Even within an NSI, security requirements are often not well defined. Consensus for computer security requirements of NSI premises would be needed to undertake joint research in the area of NSI computer security.

27. An efficient confidentiality policy should address all the different aspects of guaranteeing data confidentiality aimed both at external users and the NSI's own staff. These include: evaluating the security and disclosure limitation activities to ensure that they are adequate; creating a culture of confidentiality awareness and building up an institutional climate that values security; providing security training for staff; and developing and testing messages that reassure both users and the general public that confidentiality is being maintained.

28. The **concept of "informed consent"** was discussed. The notion becomes especially important with the increasing use of data from administrative registers, and the linking of register data with data collected by statistical surveys. Since administrative data are collected by administrative agencies, access arrangements have to be worked out between these agencies and NSIs.

29. Examining respondents' attitudes is very important. Results of case studies show quite often that respondents had objections to matching register-based data with survey responses because they were afraid of losing control over their data. Asking permission to use personal data should be as concrete as possible, including the list of registers against which the data would be matched. The permission to retrieve and match administrative data depends on the sensitivity of the issue (e.g. income is considered a very delicate issue). Also, professional codes of ethics must draw attention to other domains where statisticians might be active, such as research, teaching, and consultancies.

30. The opinion was expressed that in official statistics, informed consent is not a moral imperative to be followed under all circumstances. The availability of official statistics for society is a public asset without which government and the market could not function properly. Administrative data by themselves could be used for statistical purposes without data subjects being aware of such operations.

31. The special role of **Eurostat** with regard to data confidentiality was discussed. Eurostat itself is not involved in the data collection process, it is a trusted receiver of data from NSIs, and as such must uphold the confidentiality promises made by the NSIs. The European Statistical Law increases the role of Eurostat in a number of areas relating to statistical

confidentiality. It envisages the agreement of common EU-wide definitions of confidentiality based on the identifiability of statistical units.

32. Policies and working procedures have been put in place at Eurostat as part of the implementation of European statistical legislation. These procedures cover details of transmission, storage, processing and dissemination of data. According to these principles, the protection of confidentiality is of central importance, overriding considerations of ease of access and processing. Adequate levels of protection must be provided for all stages of data handling; access to confidential data is limited and for each data set there is a named official responsible for the protection of confidentiality.

33. The results of a survey of confidentiality practices in EU member countries and in transition countries were presented. It can be concluded that, on the whole, the member states of the EU have achieved a reasonable balance between respondents' interests in privacy and confidentiality and those of data users in obtaining useful data. However, the distinction must be clear between the use of data for administrative and statistical purposes. The potential use of statistical data by the police or intelligence services can raise concern among the respondents.

34. All EU NSIs are convinced of the importance of protection of statistical data relating to natural persons and businesses. There is little difference in the treatment of personal data and company data. The special care in the treatment of company data in EU countries may be attributed to company attitudes that are in general more suspicious and negative than those of natural persons. The reason for this negative attitude is often the burden of answering all the surveys.

35. Based on the outcome of the survey conducted by the secretariat, the attention to data confidentiality issues in transition countries is somewhat lower, especially concerning mathematical and computing aspects. Often the methods that would allow the safe release of microdata are not known. This can be one of the reasons for a more cautious attitude towards allowing access to original data in these countries in general. However, in many transition countries, much attention is paid to the legislative and administrative aspects of confidentiality. The legal basis for confidentiality has often been founded during the 1990s, and is being further developed by potential EU candidate countries to bring it into accordance with the respective legislation in the EU.

36. Disclosure protection measures are vital for NSIs because of their dependency on the provision of reliable information by respondents. However, the type of measures used and strictness of their application is sometimes inconsistent with the level of disclosure risk and the degree of intruder

interest in disclosing confidential information. A more active campaign by NSIs is required to inform the public of the statistical information available and the measures that are taken to protect confidentiality and prevent disclosure.

IV. Major problems concerning statistical data confidentiality in the transition countries.

37. The discussion revealed that, in general, the application of SDC methods and techniques as well as legal protection of statistical data in transition countries is still in its infancy. The level of implementation differs from country to country. No country reported a systematic approach to statistical data protection. All participating transition countries, however, stated that this task should be increasingly important for NSIs in the near future.

38. The Work Session was informed that in many transition countries serious problems with SDC are found in business statistics, population censuses and labour statistics (especially household income). Although some experience has been gained from Eurostat and bilaterally from developed countries (e.g. Romania-France, Czech Republic-Finland cooperation), all participating countries requested the UN/ECE and Eurostat to urgently assist in the implementation of SDC methods and techniques in the above-mentioned statistical domains.

39. With regard to the application of SDC methods in practice, several transition countries highlighted the lack of software, the need for training in confidentiality methods, and for increasing awareness of confidentiality problems in society. International cooperation and learning from the experiences of more developed countries can play an important role in this respect.