UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2003/7
7 April 2003

ORIGINAL: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-first plenary session
(Geneva, 10-12 June 2003)

STATISTICAL CONFIDENTIALITY AND MICRODATA – ISSUE PAPER

Paper submitted by Statistics Sweden[1]

I. INTRODUCTION

1. The main challenge to a National Statistical Institute (NSI) regarding statistical confidentiality and microdata is to strike a balance between the confidentiality protection and increased use of microdata. As increased use of microdata implies improved possibility of providing better data to meet the needs of users, this balance lies at the heart of official statistics which should "*…provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data…*"[2]. Simultaneously "*Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes*"[3]. In seeking this balance it is inevitable to combine different measures and actions; both legal, technical, administrative and methodological dimensions should be covered.

2. This paper aims to present some of the main issues faced when these different dimensions are addressed. Firstly, some ideas regarding the prospects of using microdata are presented. The next section deals with confidentiality and microdata. The legal issues are addressed thereafter and finally,

[1] Prepared by Matti Niva, with Bo Sundgren and Ingrid Lyberg.
[2] UN Fundamental Principles of Official Statistics, Article 1, UN Statistical Commission 1994.
[3] Ibid, Article 6.

different organisational approaches regarding access to microdata are briefly discussed. This also roughly corresponds to the outline of the seminar.

II. USE OF MICRODATA

3. Many simultaneous developments have increased the possibility to use microdata for research purposes. These include technological advances in hardware, software, data documentation and the Web. Modern PCs now have the processing capacity for advanced and large microdata sets. This implies that the NSIs can quite easily make their large data sets available to the researchers. This should be seen as an important part of the mission of an NSI: to assure that the wealth of microdata stored can be fully utilized by researchers and other legitimate users.

4. Traditionally, aggregate statistics were published according to what the NSI deemed important, although the users of course had an influence upon such decisions. The provision of tabular aggregate statistics also meant a clear limitation regarding how official statistics could be used in social and economic research.

5. The next step in the development of providing value added of the data stored at NSIs to users was to introduce statistical databases consisting of aggregated data matrixes and allowing the user to a large extent compile their own statistics.

6. The access to microdata implies a major step further as researchers and other users themselves can choose the data suitable for their research. This has also had implications to theory developments in social and economic research. Many researchers can witness the importance of the use of microdata in analysing what the consequences of policy measures may be (e.g. Erikson, p.2). Theoretical explanations of aggregate conditions can thus be supplemented with analyses of mechanisms at the individual level with help of statistical data of the NSIs.

7. The availability of large amounts of longitudinal microdata implies new analytical possibilities. For example a matching of different microdata files for several years opens new possibilities for dynamic analysis. This type of research based on microdata has been increasingly common during the past decades. This development is also obvious in economic research. A typical example from labour market economics is to link employee to employer data for analysis of both supply and demand side of the labour market (Westergaard-Nielsen, p.2).

8. NSIs can also integrate several microdata registers and create new databases. Normally, however, a lot of statistical work must be carried out to make the quality of data acceptable. Statistics Sweden has compiled some databases of this kind. The longitudinal database “Louise” with anonymised microdata on individuals and families regarding their education, income and employment might serve as an example. It should be added that this database includes annual data on all adults in Sweden from 1990 and is updated each year. Such an integrated database offers rich possibilities to carry out different analyses. An alternative to an integrated database is to link several microdata registers to each other on ad hoc bases for specific purposes.

9. The increased availability of microdata combined with IT developments has also led to a new type of approach: data mining or more broadly speaking knowledge discovery in databases. This possibility is especially interesting given the possibility of multi-database data mining (Torra et al).
10. For the NSIs, the increased use of microdata implies value added in form of better use of the data stored at the NSIs, and should also improve their legitimacy vis-à-vis respondents and the larger public. It also implies that the investments made in official statistics give higher return.

III. CONFIDENTIALITY

11. One major issue entwined in all use of all microdata is confidentiality. To put it bluntly, all use of microdata, even when anonymised, might imply a threat to confidentiality. Although a violation of confidentiality regarding microdata use has in fact hardly ever occurred in the NSIs data based research projects, the confidentiality protection is still and should be a major issue and concern. The very positive track record so far is partly due to the efforts of the NSIs. Another probable reason could be that researchers dealing with microdata have their own human capital at stake. It is also customary that the microdata issue facilitates contacts between the NSIs and the research community.

12. The need for privacy and integrity regarding statistical data is an old issue. One of the early perceptions of the need to strike a balance between the right to privacy and the increased need for information was put forward by Vincent P. Barabba in 1974 when he from the point of view of the US Census Bureau stated that "*…there is an inherent conflict in gathering data from individuals. The conflict is between the individual's right of privacy on the one hand, and on the other, government's use of mandatory processes to obtain the information it needs for valid purposes*" (Barabba, p.34).

13. The issue of confidentiality has been developed by and reflected in the documents of the international statistical community. In the ISI declaration on professional ethics from 1985 it is underlined that "*Statistical data are unconcerned with individual identities"* which implies that …"*identities and records of cooperating (or non-cooperating) subjects should …be kept confidential, whether or not confidentiality has been explicitly pledged*". (ISI, 4.5 Maintaining confidentiality of records). Further, statisticians should prevent their data from being published "*… in a form that would allow any subject's identity to be disclosed or inferred* " (ISI, 4.6 Inhibiting disclosure of identities).

14. Also the UN Fundamental Principles of Official Statistics is very clear on this point: "*Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes*" (Article 6). In many countries a national code of practice has been developed for the NSIs. Normally, confidentiality protection is one of the corner stones of such declarations. This is also the case in the recent UK National Statistics Code of Practice, where the principle of protecting confidentiality is one of the main commitments of the National Statistics (National Statistics Code of Practice).

15. Confidentiality protection is clearly a crucial commitment of the NSIs. For ethical reasons the NSIs are of course concerned about the integrity of the citizens as well as business establishments, but the commitment to confidentiality protection also has a specific explanation. The NSIs must be fully trustworthy in this respect to be able to gather data from respondents.

16. However, bearing in mind that use of microdata always might imply that the confidentiality is at risk (there is no such thing as completely safe microdata), the real issue is to strike a balance between increased information and confidentiality. NSIs normally estimate the risk of disclosure in different areas and phases of the statistical process regarding different types of uses and try to minimise such risks. An array of different methodological solutions has been put in place by the NSIs.

17. Some of the other issues regarding confidentiality protection are the specific features regarding the provision of microdata on businesses as opposed to individuals Another is the timeframe, i.e. should confidentiality protection apply regardless of time (that is, forever), or should it last for a lifetime or even less?

18. In Central and Eastern European countries the institutional and legal situation regarding statistics has changed significantly in the transition process. Data confidentiality in these countries may have some special features (e.g. legislative situation, implementation of one-way flow principle[4], implementation of confidentiality protection throughout the statistical system). In a study carried out by the ECE secretariat it was concluded that generic guidelines for statistical confidentiality would be valuable (CES seminar …).

19. Recent developments regarding the need for improved national security may pose the problem of undermining statistical confidentiality by security concerns. If this would imply an unlimited access to microdata that has been reserved for statistical purposes, the credibility of the whole statistical community could be at stake.

20. So far confidentiality protection has to a large extent been a national issue, but in the EU-context it becomes an issue for the EU institutions, such as Eurostat as well. New possibilities along with new tensions appear.

IV. LEGAL ISSUES

21. The arrangements for confidentiality protection can be based on legal acts and/or rules and regulations applied by the NSI. The legislative situation varies across countries and regions. If there is a Statistics Act of relatively recent data in a country, it normally contains regulation of statistical information. One central principle is usually that data collected for statistical purposes, regardless whether it has been collected in accordance with prescribed obligation or is given voluntarily, may in

[4] One-way flow principle implies that the NSI has access to administrative records kept by ministries or other government agencies at the microdata level, whereas the possibility for a reverse flow of micro-data subject to statistical confidentiality, whether from statistical surveys or from any other source, is strictly excluded.

principle only be used for the production of statistics. However, the data may also be used for research purposes under certain preconditions.

22. Normally, other legislation, such as The Personal Data Act, applies to the production of statistics and the release of microdata. In the EU-context, the so-called Data Protection Directive of the Council and the European Parliament (No 95/46 of 24 October 1995) is important as it strengthens legal protection of individuals with regard to automatic procession of personal information (applied to computerised personal data and data held in structured manual files) related to the individuals in question. All the Member States should have a corresponding national Personal Data Act based on the EU Directive. This might also imply that other authorities such as the Data Inspection Agency have a final say regarding use of microdata on individuals.

23. When it comes to purely statistical confidentiality legislation in the EU, there is the EU Council regulation on Community statistics (No 322/97) according to which both national authorities and Eurostat shall protect confidential data. The main principle is that confidential data obtained exclusively for the production of Community statistics shall be used for statistical production only, unless the respondents have given their consent to the use for other purposes in an unambiguous way. The recent implementation regulation under the regulation on Community Statistics concerns access to confidential data for scientific purposes (Commission Regulation 83/2002). This regulation is on rules concerning conditions to allow access to confidential data from four different EU-surveys. According to this, scientific researchers might have this access if a contract regulating the terms of access is signed with Eurostat. However it remains to be seen how Member States will meet the implementation of this legal arrangement in practice.

24. Regarding data possessed by public authorities, one point of departure could be quite the opposite of confidentiality protection. It can be claimed that for the sake of democracy all information that is created within the public sector should be public. Accordingly all decisions taken by a public authority and background documents facilitating these decisions including correspondence of the public authority should be made public. This principle of transparent public administration applies in Sweden. However, all exceptions to this principle of publicity due to motivated secrecy etc . must be stated in specific laws such as Statistics Act.

V. ACCESS TO MICRODATA – DIFFERENT APPROACHES

25. One of the main challenges facing NSIs regarding microdata is to provide access to users in different ways. This access can in principle be organised in a number of ways and normally the NSI itself should find a suitable and feasible solution considering the prevailing institutional and organisational circumstances. At the same time a lot of benchmarking has taken place during recent years.

26. One early way of providing controlled access to microdata has been to compile anonymised Public Use Microdata Files (PUMFs). This solution was introduced by Statistics Canada in the early 1970's (Boyko & Watkins, p. 3). A rigorous processing is carried out before release of PUMF to reduce the probability of disclosure. Since the outset of this program more than 350 PUMFs have been

created and several other countries have chosen similar solutions. The PUMFs have been valuable for researchers in universities and government departments. Some of the problems related to their use have been relatively high costs, especially under the 1980's and early 1990's (use of mainframe, pricing policy) and the fact that the anonymisation process decreases the value of the data (ibid, p. 5).

27. In many countries the delivery of de-identified microdata to researchers and other legitimate users outside the NSI is still the main way to release microdata. There were around 200 such cases in 2002 at Statistics Sweden and the number of releases is rapidly increasing. If the released microdata is detailed register-based data, it is normally in fact not anonymous. It is obvious that the NSI in such cases must base the approval of the release of microdata on prevailing legislation and other rules stipulating the confidentiality rules. In Sweden attempts to re-identify data are criminalized.

28. Because of the sensibility of microdata and the possibility of re-identification, the NSIs in many countries do not allow an off-site use of microdata. Instead the NSI creates an on-site Research laboratory for the researchers. This is the case for example in Denmark (Access to…, p. 14). This option also includes a solution where the NSI puts up and runs a Research Data Centre e.g. at a University. In both cases it is easier for the NSI to check that the confidentiality is not violated.

29. A still more cautious solution is to allow the use of microdata only by the NSI staff. In some cases the researcher becomes a staff member for a period of time to be able to carry out the microdata based research. Also, the NSI might have a policy of inclusion of research staff to be able to exploit the wealth of microdata for external clients.

30. However, it is becoming more common to authorise given research institutes to be able to have on-line access to anonymised microdata of the NSI. This solution has been chosen in Australia, in Denmark and in Portugal. The data sets available on-line might be limited due to their sensitivity and also modified according the specific needs and orientations of research institutes concerned. It is obvious that the remote access systems such as Internet based on-line access are highly appreciated by researchers. It might also be attractive to a growing number of NSIs as this solution allows a certain control of the use of microdata.

31. The question of pricing is also relevant when discussing access to microdata. It is quite common that the NSI has already been funded with appropriations for the major part of the work to compile and maintain microdata registers. If so, it would seem reasonable that the price charged corresponds to the extra costs due to the release of microdata. Such costs can of course be defined in a number of ways, pending on the calculation principles applied. However, the pricing of the release of microdata might also be based on other principles such as market pricing or even a free of charge basis under certain conditions.

VI. CONCLUDING REMARKS

32. This paper has underlined that a balance must be struck between the protection of statistical confidentiality and improved access to microdata. It has also shown that a number of legal,

administrative and technical measures must be combined to reach such a goal. This also implies that there are many different ways of reaching this balance. The statistical community could also work for a common policy and agree upon core principles regarding access to microdata. The CES seminar will hopefully contribute to a richer understanding of the options available.

REFERENCES

Access to Microdata in the Nordic Countries (2003). Statistics Sweden.

Barabba Vincent P. (1974): *The Right of Privacy and the Need to Know*. Proceedings of the Social Statistics Section, American Statistical Association 33.

Boyko Ernie & Watkins Wendy (2002) *Safe Data, Safe Places: No Either/Or Solutions*. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'

CES *Seminar on Data Confidentiality* (2003). Note by the ECE secretariat. CES/BUR.2003/27/Add.1

Declaration on professional ethics (1985). International Statistical Institute.

Erikson Robert (2002): *The right to privacy and the right to information*. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'

National Statistics Code of Practice. Statement of Principles (2002). National Statistics, UK.

Perpétuo Fernanda (2002) *Statistical Information System for Researchers*. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'

Sundgren Bo (2001): *Statistical Microdata – Confidentiality Protection vs. Freedom of Information*. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, FYROM.

Torra Vincenç & Domingo-Ferrer Josep & Torres Àngel (2003): *Data Mining Methods for Linking Data from Several Sources*. Paper to the joint ECE/Eurostat work session on statistical data confidentiality. Luxembourg 7-9 April 2003.

UN Fundamental Principles of Official Statistics (1994). Adopted by the UN Statistical Commission.

Westergaard-Nielsen Niels (2002*): Linking employer-employee data.* Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'