INTERNATIONAL SEMINAR ON STATISTICAL CONFIDENTIALITY PROCEEDINGS

28 to 30 November 1994
Luxembourg

Organized by

Eurostat
Statistical Office of the
European Communities

Statec
National Statistical Institute
of Luxembourg

Theme
Miscellaneous
Series
Methods

0
E

STATISTICAL DOCUMENT

THE INFLUENCE OF THE TECHNOLOGICAL DEVELOPMENT ON DATA COLLECTION METHODS IN SURVEYS. HOW IS THE PROTECTION OF PERSONAL DATA AFFECTED

Lars Thygesen

Danmarks Statistik

1. Introduction

This paper discusses the impact on confidentiality of new methods of data capture. I shall not limit the exposition to discussion of methods on the absolute technological forefront. Many of the methods have been maturing for years and are now being used in practical statistical work in a number of countries.

2. The new collection methods

The classical way of collecting survey data is of course based on questionnaires which are filled in either by an interviewer or by the person or company who is to answer the questions. In recent years a number of techniques have been developed, aiming at reducing the costs of data collection or increasing timeliness of the results. Some of these techniques resemble the classical interviewing method while others are more revolutionary. Electronic transmission of data is often involved.

A number of new methods are listed in the following paragraphs and their data protection implications are discussed.

2.1 Computer assisted data capture

The whole family of methods for computer assisted data capture has become very popular and widespread in central statistical offices in recent years: The basic idea of these techniques is not at all a new one (Weeks, 1992) but it has taken many years to penetrate into the practical day-to-day operations of central statistical bureaus in many countries. In Denmark, for instance, it has only been used since 1991. The idea is that a PC with an appropriate program (e.g. Blaise) is used instead of a paper questionnaire. The answers are entered immediately and checked for errors, resulting in cheaper and faster statistics. The methods applied differ in the sense that the person operating the PC may have different positions.

CATI

Probably the most widely used method is Computer Assisted Telephone Interviewing (CATI) where the interviewer using the PC is connected with the respondent via phone.

CAPI

Another application is Computer Assisted Personal Interviewing (CAPI) where the interviewer carries out a face-to-face interview carrying a notebook PC.

Electronic questionnaire (CADI)

The person feeding the PC may also be the respondent himself. The term Computer Assisted Direct Input (CADI) has been used for this situation. In Denmark we usually call it the Electronic Questionnaire because the questionnaire is sent to the respondent in an electronic form. While CATI and CAPI have been employed for some time, the electronic questionnaire method has only quite recently become used in Danmarks Statistik but is spreading rapidly.

Data protection aspects

CATI and CAPI seem to pose very little security problems that were not already there in conventional survey taking. Of course data have to be transmitted as quickly as possible from the relatively unprotected PC environment to a safe setting, e.g. a mainframe environment. This can easily be managed in CATI if the PC's are connected in a local area network in the central statistical office. In CAPI, interview data should be transmitted daily

from the interviewer over the tele network to the safe environment so that only data from one day's work is kept locally.

In relation to CADI we have encountered some special problems in Denmark. CADI is used in two surveys:

- The Survey on Labour Market Measures (SLMM), aiming at statistics on the use of activating measures for unemployed people, such as subsidised employment, special training courses; the respondents are the municipalities
- The Wages and Salaries Survey (WSS); the respondents are the employers (business companies)

The respondents receive a diskette with the questionnaire, fill it in, and return it: No problem. But it turns out that some of the respondents, especially in the SLMM, will keep the program and currently enter new cases (including identifications) during the following quarter so that the answers to the next questionnaire are already filled in at the end of the quarter. In this way the statistical questionnaire becomes a register that the municipality may use also for other purposes. Although it is a register meant for statistics there is a risk, that it is used for administrative purposes, such as to manage the labour market measures. Is this the problem of the central statistical office?

2.2 EDI

Electronic Data Interchange (EDI) is becoming increasingly used by private enterprise for the exchange of business information, e.g., invoices. It has also been proposed as the future data collection method of business statistics. Some statisticians have gone as far as to argue that unless we employ EDI techniques, companies will simply refuse giving information to NSIs. EDI depends on an agreement on a standard format for the exchange, and exchange can take place in different electronic media, on-line or off-line. I believe e-mail will be the typical medium in the future.

In Denmark, a technique similar to EDI is being employed in the new Wages and Salaries Survey (WSS) where a common format for very detailed basic information on each employee has been agreed upon. This format is used for reporting by the employers, either to Danmarks Statistik, or to the Danish Employers Union who will in their turn report to us. In many cases, the reporting is done by a data center running wage systems for a large number of employers.

2.3 Telephone based techniques

TDE, VR

In some countries, notably the US (Werking, 1992), telephones are used for collection of statistical data from private enterprise. These seem to be cheap and fast techniques well suited for very small questionnaires that are used repeatedly, e.g. monthly. The simplest form is Touch-tone data entry (TDE) where the numerical keyboard of the telephone is used to give code-numbers for the answers. Voice recognition self response (VR) is easier for the respondent as his answers just have to be spoken.

We have no experience in any of these techniques in Denmark.

2.4 Register based statistics

Finally, mention should be made of statistics based on (public) registers, a technique that is widely used in Denmark (Danmarks Statistik, 1995). This is potentially an extremely rich source for statistics as public authorities hold large amounts of data on persons, companies, and land use. The Danish experience is that a statistical system may be built on this foundation, especially if it is possible to use general identifiers (Person Number, etc.) to combine data from different sources.

This puts the NSI in the position where it has to process and hold large amounts of identified information, and to protect this information. The system is vulnerable to public distrust, so protection measures must be given extremely high priority. In Denmark, this has lead to very strict procedures of privacy protection, involving among other things that no data on an individual basis, with or without identifiers, are given to outside users.

It should be mentioned that register based statistics in many ways resemble the above-mentioned EDI techniques.

3. Conclusion

When I agreed to make this presentation some months ago we were discussing with the Danish Data Surveillance Authority some data protection problems in connection with especially CADI. I had the idea that in a few months time I would be much wiser - but unfortunately this is not the case! I have now searched the literature available to

me and found, in Weeks (1992), an excellent exposition of the properties of many of the newer techniques. But to my disappointment, I have found no discussion of security.

New methods new problems

My conclusions from Danish experience is that the problems of data protection associated with the new methods differ from the problems we already know. As many of the techniques imply telecommunication of confidential data, we must devote some effort to securing this communication more than at present time. We must be absolutely sure that data arrive to the correct receiver (the NSI) and nowhere else. This will involve encryption techniques.

If the new methods give the statisticians access to very large volumes of identified data, as in the case of EDI and register based statistics, the role of confidentiality protection is stressed.

References

Danmarks Statistik (1995): Social and demographic statistics in Denmark - A register based statistical system. Eurostat (in press)

Thygesen, L. (1993): *Technological aspects of confidentiality: New technology - threat or greater protection?* In "International seminar on statistical confidentiality, Proceedings, 8-10 September, Dublin, Ireland". Eurostat

Weeks, M. F. (1992): Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. Journal of Official Statistics, 1992-4. Stockholm

Werking, G. S., & R. L. Clayton (1992): *Enhancing data quality through the use of a mixed mode collection.* In "New technologies and techniques for statistics, Proceedings of the conference, Bonn, 24 to 25 February 1992". Eurostat