Statistical Disclosure Control and Data Access for Research Purposes: Critical Issues and Possible Solutions

Lucia Buzzigoli*
Luigi Biggeri
Department of Statistics "G. Parenti", University of Florence
Viale Morgagni 59, 50134 Firenze, Italy
**e-mail: lucia@ds.unifi.it*

1. Introduction

Today the huge opportunities provided by instruments available for the capture, elaboration, memorization, and transmission of data have modified and continue to modify both the supply and the demand for statistical information: on the supply side, a substantial reorganization of work within the National Statistical Institutes (NSI's) and other statistical organizations is being realized, while from the demand side, users are much more prepared and exacting, and much more autonomous in terms of management and elaboration of data. The technological advancements have brought the NSI's to a profound revision of their dissemination procedures, which now must aim to satisfy more flexible requirements than in the past, and to look for alternative methods to increase the possibility of producting results on request at a minimum cost.

On the other side, it is well known that the NSI' s must also guarantee that their data are sufficiently protected from the risk of disclosure. The transition from a rigid and prearranged dissemination regime, in which the NSI's decides what will be released and in what form, to a more flexible regime that is more geared to the specific needs of its more sophisticated users, provokes a corresponding evolution in the necessary procedures to test the confidentiality of data. This is due to the impossibility in many cases of carrying out *ex-ante* all the verifications necessary to guarantee that the disseminated data are sufficiently secure.

This is particularly true when NSI's must face the demand of data brought by the research world, which has a need of high quality service of personalized supply and is definitively oriented toward modern forms and instruments of dissemination. For instance, as far as the social science area is concerned, the possibility to access increasingly sophisticated data opens up a wide range of research options and behavioral modeling that are of great help in facing complex policy issues.

The paper tries to evidentiate the new challenges for the protection of privacy and confidentiality issued by a more and more exacting research world. After a brief review of the methodological and organizational solutions that are applied by NSI's to protect statistical data (§2), we present some critical issues that still remain to be faced (§3) and give particular attention to the role of intermediaries in the organization of data access and in the education of data users (§4). Some remarks on juridical and ethical concerns will end the paper (§5).

2. Safe data vs safe settings

NSI's use many different procedures to protect the confidentiality of data. If different strategies of data storage and dissemination are possible, the available alternatives must be evaluated in the light of a complex framework which involves statistical, technical, technological, organizational and juridical issues. In particular, the last developments in these various fields of study and in research practice have recently led to consider new implications of confidentiality, privacy and ethical issues.

The traditional method to protect data is that of limiting the informative content of the disseminated information with appropriate statistical procedures (Statistical Disclosure Control or SDC methods): the data are modified or some variables partially suppressed to avoid the spread of too highly detailed information (*safe data*). Dissemination of tables and public use microdata files heavily rely on these protection procedures. A vast literature is available on this argument including some relevant surveys (Willenborg and Waal, 1996).

But, in order to satisfy the needs of the more sophisticated users - who often do not trust on the analysis not based on original data - an increasing number of NSI's give access to more detailed data, protecting them with very rigid protocols of use, involving technical, organizational and contractual aspects (*safe setting*). In some cases specific locations are arranged (at the NSI's or at specific institutions) where bona fide researchers are provided with the necessary hardware and software and access to confidential data (Statistics Netherlands, 2001), sometimes under Fellowship programs. In other cases statistical agencies issue licenses to researchers, who can analyse and elaborate restricted data in their own premises (Committee on National Statistics, 2000). Of course, in both cases, severe penalties are provided for confidentiality violations.

Safe settings seem more adequate to research needs, but not all these kind of solution are equally satisfactory. In particular, the availability of new technologies, which offer almost unlimited possibilities of remote and decentralized access, obviously lead to strong user expectations towards more and more flexible dissemination architectures.

The role of Internet in this respect is crucial. It is doubtless that - as an instrument of data dissemination - the Internet presents clear advantages. First of all, it sets in motion unlimited possibilities for dissemination, because it eliminates distances, and it facilitates a round-the-clock undifferentiated access. Secondly, it makes an enormous amount of very detailed information available in a timely manner and at reduced costs. Finally, it allows users an impressive degree of independence in managing their own access to data, even providing them, should it be necessary, with data processing services. However, the advantages of this tool are effective only if as many steps as possible are automatized in the chain of production of the statistical data and if the technological architecture allows for an accurate disclosure protection of the disseminated data, both in term of limiting disclosure risk and in terms of access control.

As significant examples of modern solutions in web based data dissemination see, for instance, the American Factfinder of the Census Bureau or the system adopted for the First National Agricultural Census of the People's Republic of China. In any case it is essential to evaluate the disclosure risk connected to the various dissemination protocols, in order to guarantee that the re-identification probability is sufficiently low.

3. Open problems

The international debate that had risen up around all these issues is still alive and shows that several problems in very different fields still deserve proper attention.

The most important open questions regard the data access organization and the typology of data.

First of all, the technological advances, which turned out in new means of data release and data management, have introduced new opportunities in organizing data access but also new problems for assuring the safety of data, asking for a major revision of disclosure control protocols.

As a consequence, there is a strong connection between the 'classic' statistical disclosure control literature and that on computer science data access control: this is the reason why the concept of statistical disclosure control is sometimes substituted with that of statistical data protection, that gauges the security of data from within a complex process of production of the data itself and considers not only the logical protection of the data (the impossibility of retracing the disclosed datum to the respondent), but also physical protection (the impossibility of violating the hardware and software protections of the data safeguard itself).

In particular, in database literature the problem of confidentiality becomes a problem of database security and means preventing illegal data access while maintaining the maximum data availability. In this field of analysis inferences on restricted information are prevented checking the size of the database used to build the statistics and taking memory of the user's past and current queries. For a brief review on the subject, refer to Brodsky et al. (2000).

Secondly, the wealth of possibilities produced from syntheses inferable from survey data depends also on the richness of the information that can integrate the original core obtained from the starting data. In some cases, for example in the censuses, the array of census products includes also data-sets that are not directly deducible from the single survey, but that by now are an integral part of the informational patrimony necessary for a correct and complete evaluation of census results. One thinks, for example, of archives with digitized geographic references to areas of the census (digital boundary data), or to individual longitudinal archives (such as The Longitudinal Study for UK).

Large interest given rise to by geographic informational systems points out the attention for those forms of output that allow for suitable integration of diverse informational sources, in view of a general organization of the whole statistical system (and not only the census system) within a real data warehouse. GIS are becoming more and more popular as a mean to convey statistical information. On one side, geographical dimension is an essential information for analysing phenomena of different kinds (socio-economic, epidemiological, etc.), but, on the other side, it is well known that the geographical detail is maybe the most dangerous identification key for the variables in a survey, and therefore represents a threat for confidentiality of data. Moreover, for this category of data the Internet has an enormous dissemination potential, mostly thanks to the possibility of creating interactive georeferential applications that allow users to access GIS applications at very low cost and with a minimum of computer expertise. The work of Karr and Sanil (2001) in this session is an example of such a problem.

Another typology of data, which feed a wide range of research and behavioral modeling, are longitudinal data deriving from panel surveys, which usually imply a higher disclosure risk, especially when they have been linked to administrative records. An interesting debate on this subject has recently begun (see also Committee on National Statistics, 2000), which shows that longitudinal (as well as hierarchical) structures have not been adequately considered in the computing of disclosure risk: therefore the decision on what can be released and under what conditions cannot be based on objective basis. The definition of a per-record risk where the dependencies between variables or units are explicitly analysed could be a first step towards a more comprehensive approach. Some first proposals have been made by Benedetti and Franconi (1998) (who suppose that the population is constituted by a set of groups and build a model for estimating individual risk where an individual disclosure influences the risk of disclosure of the other members of the group) and Fienberg (1997) also underlined the need for methodological and empirical investigation in this field of study.

Finally, another element to be considered is the use of administrative files as a statistical source: this could have important consequences for the data dissemination policy of NSI's. Sometimes the availability of administrative records can be of invaluable importance for some research (for the linking of longitudinal files with administrative records see, for instance, §2 in Committee on National Statistics, 2000), but can seriously compromise data safety. This could lead NSI's to foresee data dissemination strategies that are more differentiated than before (see the example of the HRS longitudinal file, which is a public use file, while the versions with linkages to administrative information or with geocoding are available under differentiated restricted conditions).

4. The role of intermediaries

The effective accessibility and usability of data for research purposes is boosted also by the existence of auxiliary structures that integrate the dissemination activity of the producer institute. In fact, researchers usually have very specific informational needs: they do not need only data, but they also need accessible meta-data, information on the quality of data, on their format, on the software available for analysing them. All these exigencies require an adequate support that only specific specialized structures can guarantee.

We can distinguish the local branches of the NSI' s from those other units which, instead, assume the role of actual intermediaries between statistical data producers and users. The first are territorial support units both for operations of collection of surveys and for the dissemination of data produced. The second can be public or private entities that help the process of data dissemination or simply make them more visible, or add to their informative value by providing counseling activities, planning and general research.

Many of these structures turn to the academic world, or, more in general, to research institutions, for which they constitute preferential channels of access to data. The example of social science data archives is typical.

The involvement of these intermediary figures has significant legal, financial and practical aspects. From a legal point of view it is necessary that the ownership of the data and the responsibility for its management be very clear; moreover, the set of legal penalties for misuse must be clearly defined. The financial aspect involves substantial decisions on rates policy (which is an essential part of the establishment of a marketing strategy with outside users) and on the problem of finding financial support. The practical aspect calls for the predisposition of modern and efficient units often completely dedicated to the management of dissemination for the predisposition of adequate access control systems, that seem the most critical components of these architectures.

The role of these intermediaries seems to acquire particular significance in society today, where, to complement an indiscriminate increase in the creation and exchange of information flows there often cannot be found an adequate culture of information in general, and particularly, that expressed as statistical data. The presence of a gap in information is addressed by the policy of dissemination to universities which should bridge it with their educational mission (a significant example is that of the Data Liberation Initiative in Canada, which has already significantly influenced research and, especially, teaching) and by the diffusion of data sharing principles.

5. Concluding remarks: juridical and ethical issues

Technical (i.e., organizational and methodological) solutions to data access for research purposes are not enough, if they cannot rely on sound juridical basis, which should guarantee both the right of research and the right to privacy of the individuals who provide the data. Therefore, the body of laws regulating data access obviously interacts with the laws concerning privacy and the management and diffusion of statistical information. An articulated and complex legal architecture is consequently defined that operates at various levels (sovranational, national, etc.) and that reflects the globalization process which is modifying the socio-economic development cycles in our society.

In this scenario the rights and the duties of researchers must be clearly pointed out: in particular, impersonal and generic freedom of scientific research must translate in a specific, personal responsibility of researchers who use confidential data and this necessity opens new important areas of development for sectorial recommendations (in form of ethical codes and research policy handbooks) and for the institution of overseeing agencies.

REFERENCES

Benedetti R. and L. Franconi (1998), "Applied Issues on Disclosure Avoidance Complex Microdata Files", Servizio Studi Metodologici, Istat, Roma.

Brodsky A., C. Farkas, D. Wijesekera and X.S. Wang (2000), "Contraints, Inference Channels and Secure Databases", *Tecnical Report*, George Mason University.

Committee on National Statistics (2000), *Improving Access to and Confidentiality of Research Data. Report of a Workshop*, National Academy Press, Washington, D.C..

Fienberg S.E. (1997), "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research", *Tech. Rep. 10/97*, Carnegie Mellon University.

Karr A.F. and A.P. Sanil (2001), "Web Systems that Disseminate Information but Protect Confidential Data", *Invited Paper*, 53rd ISI Session, Seoul.

Statistics Netherlands (2001), "SDC in Practice: Some Examples in Official Statistics of Statistics Netherlands", *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Skopje.

Willenborg L., T. de Waal (1996) *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, 111, Springer-Verlag, New York.

RESUME

Ce papier examine les problèmes d'organisation et techniques qu'il est nécessaire de résoudre pour garantir l'accès à les données individuelles de part de rechercheurs, en assurant la protection de les données confidentiel. Quelque solution est suggérée pour les aspects soit techniques que juridiques, en utilisant aussi spécifiques codes éthiques et de conduite pour les rechercheurs.