Important Issues on Statistical Confidentiality Methods

Ph. Nanopoulos and John King

Eurostat, European Commission, L-2920 Luxembourg
photis.nanopoulos@cec.eu.int; john.king@cec.eu.int

Abstract. This paper sets out, in the context of official statistics, some of the key issues of confidentiality and the methods developed to maintain confidentiality. The relevance of the issues and methods to data mining of official data are discussed. Recent developments that will increase the availability of microdata for scientific research are outlined.

1 Introduction

The title of this workshop is "Mining Official Data". So what is the connection between data mining and confidentiality of official statistics? One link is that there may be lessons for the data-mining practitioner to be learned from the practices developed by statistical organisations in order to protect the confidential nature of their data. But to mine data, there must be access to data. So perhaps the presumption of the Workshop is that data used for the creation of official statistics could be made available for mining. This paper looks at the proposition from the viewpoint of an official statistical organisation. It therefore looks at the constraints and issues that statistical offices live with in their regular work on official economic and social data. Researchers, including data miners, who wish to have access to the datasets underlying official statistics will need to understand and respect these issues and principles.

These issues apply in particular to data obtained through surveys and censuses, either voluntary or compulsory, from individuals and households, and from companies or enterprises. Some of these issues may not arise where data have been obtained by different methods, for example from administrative records, but even then other constraints may exist in the form of specific laws protecting those records or from general Data Protection laws. An example of the latter constraint is the European Union (EU) Directive 95/46 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, and also the consequential laws implementing the directive in the EU Member States. But that is presumably a topic well known to data miners, and well understood by them, and so this paper will concentrate on confidentiality issues related to official economic and social statistics and the data on which they are based.

Data mining may seem to be the antithesis of protecting the confidentiality of official statistics. The former seems to have a carefree, heroic and exploratory image that is the opposite of the conservative approach taken in official statistics. So in any public discussion it would be important to emphasise that data mining is looking for *patterns* and *statistical* relationships in data—not for individuals and their details and their relationships. So data mining is *pattern* recognition, not *people* recognition, or company recognition.

The principle of informed consent is the basis of much of official statistics. A guarantee of confidentiality for the information provided is often the basis of obtaining the data. It also imposes a constraint on what can be done with the data and by whom. In this context, confidentiality issues and statistical disclosure methods have been developed to maximise the use of the data while keeping the original agreement with the data source.

This paper sets out some of the issues, principles and methods that have been developed, particularly in relation to traditional outputs—tabulations—of statistical offices. Then, some of the issues relating to providing access to confidential data are indicated. Finally, some recent developments, at the European level, that will provide better access to anonymised micro-datasets for scientific research are described.

2 Tabular Data Disclosure Control

Tables (including cross-tabulations) are the simplest and most common form of official statistical output. It was recognised early on that tables could be disclosive, particularly in the case of economic data. The important criterion for tables is that dissemination is only possible if data subjects (those providing the data) cannot be identified directly or indirectly. The problem is then one of checking whether a table has sensitive cells; and then of dealing with the sensitive cells.

For example, a table might show production levels of a particular commodity. If only one company makes this commodity then the table would disclose its output. A more complex table, perhaps of production by region and by size category of the production unit, might make it easy to recognise a company (from several making that commodity) and show its output of the particular commodity.

Early methods for protection of tabular information included suppressing cell values in tables and forming broader categories of the classifying variables. These methods are still in use, sometimes implemented by hand. Other methods include rounding values in cells and adding noise (e.g. adding +1, 0, or -1 to the cells in a random pattern. Some of these methods are not very satisfactory in that marginal or grand totals may not be the exact sum of the constituent cells as shown.

Over time, and following interesting research in the subject, a more sophisticated understanding of the problem has developed. And the development of more sophisticated methods for statistical disclosure control for tables has followed this understanding. For example, it was recognised that, even if a table appeared non-

disclosive to a casual reader, it might be disclosive if the reader had some additional relevant information. If a table cell had an aggregate from two observations, then one of the members of the cell could deduce, by subtraction, the contribution of the other member. Hence, the requirement, quite common in official statistics, that cells should contain at least three members before any information can be released about the cell. Various rules have developed over time, often based on the number of units in the cell, including the "dominance" rules, now commonplace.

Cell suppression itself has developed into a sophisticated art. We now consider the suppression of "secondary" cells to protect the "primary" cells, and also the degree (or interval) of protection afforded by different suppression patterns.

Recent developments include methods for the simultaneous protective treatment of tables with several dimensions that have one or more overlapping (i.e. shared) dimensions. These methods have been implemented in software by statistical offices and continue to be the subject of research. An example is the *GHMITER* engine developed at the Statistisches Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen. A user-friendly graphical interface for this, *CIF*, has been developed by Eurostat; and the engine is also being incorporated into the software suite *t-Argus.*

Another issue that is becoming recognised as very important is the constraint imposed by published tables on the to-be-published tables. There is active research in this area. Karr and Sanil indicate that the whole tabulation plan for a particular dataset should be considered at the outset. Otherwise publication of a particularly interesting table (or at least the interesting parts of it) may be prevented because of the information already published. In the Eurostat context, this could mean that an aggregate for the 15 Member States (MSs) might not be publishable because of the aggregates already published (or suppressed) at MS level.

3 Dissemination of Anonymised Records

The dissemination of anonymised records is an issue that has become increasingly important and will undoubtedly continue to be so. It is probably the issue discussed in this paper that will be of greatest interest because the records could be the raw material for data mining.

3.1 Research Interest

Anonymised records, or anonymised micro-datasets, are becoming important because of increasing interest from researchers in access to them. This interest has two related drivers. An aspect of modern life is the increasing interest in and demand for evidence-based policy, policy analysis, and monitoring policies and their impact. This kind of activity requires timely, detailed information and frequently requires more detailed analyses than are presently published by statistical organisations. Sometimes these analyses are seen as outside the remit of national statistical organisations (NSIs)

or even as activities that could compromise the perceived independence of NSIs. Indeed, these analyses are performed often by academic institutions or independent research institutions.

Detailed data are needed for these types of analyses. The obvious and most relevant source is often identified as the data collected and held by NSIs. Hence there is an increasing pressure on NSIs and other statistical organisations to provide detailed data on a wide range of topics. In particular, for the European Union (EU), pan-EU analyses and research are becoming more and more important. The same could also be said for the Euro-zone. So the need is for access to pan-EU datasets for this research. Eurostat holds many such datasets, and so it is seen, by analogy with the national situation, as the natural, simple and direct potential source for these datasets.

The second driver here is the changing nature of research itself. Much modern research cannot be satisfied with aggregate data—micro-data are needed for fine analysis and model building. Hand-in-hand with this there has been an evolution (perhaps revolution would be a more appropriate description) of research computing capacity—both hardware and software tools. This has considerably increased the demand for access to micro-data records for computing correlation matrices, estimating models and other analyses, depending on the context of the research topic.

3.2 Providing Data while Maintaining Confidentiality

At the same time, statistical organisations, both NSIs and supra-national and international institutions, are increasingly seeing making more use of the data held by them as an important contribution to society and as part of an obligation to make better use of their resources (data). But there are constraints on what statistical organisations, particularly NSIs, can do and on how they can do it. The role of researchers and research organisations is thus an important one, and it is an increasing one too. NSIs can assist in this development by providing some form of access to some form of data to these researchers and research organisations.

But there are problems in providing micro-datasets to researchers—the main one being the confidential nature of the data themselves. The question is, how can the confidentiality of the data be maintained while at the same time providing researchers with the information they would like.

There are two key issues here. The first is the basis on which the data has been obtained; and the second, related to the first, is the perception of the data supplier. Most information used for official statistics has been obtained against a pledge of confidentiality and a guarantee that it will be used for statistical purposes only. The ideas of informed consent and confidentiality underlie official statistical data collection. The principle is that the data subject has a right to know what the information will be used for and who will see their information. The argument here is that if there is a new kind of use of the information, then the data subject should be made aware of it.

The perception of the data supplier is very important. Many statistical offices feel that this is a major factor affecting response rates to surveys and data collections. Damage to perceptions, or a loss of confidence through even inadvertent disclosure of confidential data, would result in worse data in the future.

Nevertheless, several statistical offices have created anonymised micro-datasets for access by researchers. These are intended to provide researchers with a dataset in which the information content has been reduced sufficiently for the risk of identification of a record or of disclosure to be acceptably small. Some datasets are available, with little bureaucratic procedures and at minimum cost, to academic researchers. Other statistical offices have a more conservative approach. Differences reflect conditions, attitudes, legal issues and past practices in different countries. There are differences, too, in the meaning of "anonymisation" and in the degree of risk that would be acceptable in releasing anonymised micro-datasets.

3.3 Creating Anonymised Records

For the creation of anonymised micro-datasets, various techniques have been proposed—broadly the same as those used to protect tabular information but also micro-aggregation techniques, suppressing variables and ensuring there is a delay between collection and reference period of the data and its release as microdata.

Some methods used, often in combination, at present include:

- reducing the geographic coverage;
- rounding;
- grouping or combining categories;
- adding noise or perturbations;
- micro-aggregation;
- data swapping;
- top- (or bottom-) coding;
- imputing values from a model;
- suppressing fields or cells;
- suppressing variables;
- time delay.

An innovative approach suggested recently by Abowd and Woodcock is the creation of synthetic datasets—retaining the internal structure of the variables but containing no real records. Although these synthetic datasets might satisfy some research purposes, it is not yet clear whether they would also be of use in situations where the relationship being searched for may exist in the original dataset but not have been specified in the creation rules for the synthetic dataset.

3.4 Micro-aggregation

Micro-aggregation may need a little explanation as it is relatively unknown and unused. It is one method (or, rather, a set of methods) for anonymising potentially disclosive data and thus for creating anonymised micro-datasets. In essence, a

variable that is sensitive or potentially disclosive is perturbed in a particular way with the intention of retaining as much information and pattern in the data as possible but at the same time reducing the risk of the information about that variable being disclosive. Records are clustered into small groups on one or more variables. Then the aggregate or average value of the variable over the group is assigned to each member of the group. In an early presentation of the approach, Defays and Nanopoulos proposed fixed same sized groups, but more recent research has considered variable sized groups.

For example, a variable may be of the amount of local taxes paid by a household. Depending on the way this amount is determined or calculated, this amount may enable a researcher to identify the local authority in which the household resides. This information, particularly in conjunction with other information about the household in the record, may raise the risk of identification of the household to an unacceptable level. So a simple anonymisation process would be to suppress this variable as it provides information enabling the local authority of the household to be identified. But suppose that the amount of local taxes paid could be averaged (for similar households) over several similar local authorities. This average (the micro-aggregate) could replace the actual amount paid in the records for the households. This would mean that the particular local authority of a household could not be identified unambiguously, thus reducing the risk of identification. But the patterns in the data would, to a large extent, be unchanged.

Different approaches have considered different notions of "similarity" for forming the groups, and these have lead to different ways of creating the groups of records to be aggregated. What they have in common is first the ranking of a variable according to a defined criterion and then the grouping of successive records. Some approaches have used values of a variable; others, the first principal component of a set of variables; yet others, the sum of z-scores.

4 Confidential Data at Eurostat—the Legal Context

The principle of statistical confidentiality is effectively the contract connecting the statistician with all those providing their individual data, either voluntarily, as is frequently the case, or by legal obligation, with a view to producing the statistical data essential for the society as a whole. From the formal legal point of view most of the European countries have established legal provisions for statistical confidentiality a long time ago. At the European level, the principle has been enshrined in Article 285 of the Treaty establishing the European Community as a fundamental principle for Community statistics. Article 285 provides that the production of Community statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality. The confidentiality principle is therefore part of the European basic constitutional charter and has thus acquired the highest status in legal terms.

The principle has been further specified and data received, held, used and disseminated by Eurostat are controlled by a set of laws that have developed since the Treaty founding the European Communities. In 1990, Council Regulation 1588/90 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Community set out basic rules and safeguards for the handling of confidential data. Subsequently, in 1997, the "Statistical Law"—EU regulation 322/1997 on Community Statistics—expanded on these basic rules. In particular, a legal definition of statistical disclosure was introduced. Article 13 states:

"1. Data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.

To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit."

This definition has replaced the former definition laid down in Regulation 1588/90 where confidential data were defined as "data declared confidential by the Member States in line with national legislation or practices governing statistical confidentiality." The notion of confidential data has consequently become an objective notion with a clear Community dimension.

This definition uses explicitly five different concepts: individual information; a third party; identification (direct or indirect) of a statistical unit; disclosure of individual information to a third party; and means that can be reasonably used by a third party to identify the said unit.

The concept of "individual information" is not explicitly defined in the European statistical law and so interpretation should be as wide as possible thus making confidential any information concerning a statistical unit. Nevertheless, article 13 goes on to state:

"2. By derogation from paragraph 1, data taken from sources which are available to the public and remain available to the public at the national authorities according to national legislation, shall not be considered confidential."

So, in conclusion, we can say that "individual confidential information is any individual information which is not normally publicly available".

The Statistical Law also states that confidential data must be used exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes (article 15). The law also makes provision for access to confidential data for scientific purposes (article 17).

5 Recent developments on access to datasets for scientific research

At the European level, there have been developments recently which will help to meet the demand for access to microdata and to open up datasets for scientific research purposes. A new Commission Regulation, 831/2002, concerning access to confidential

data for scientific purposes, was adopted on 17 May and came into force on 7 June 2002. This was the culmination of a long process of discussion, negotiation and drafting with 15 Member States (MSs). The Regulation sets out procedures and conditions under which access to confidential data for scientific purposes may be granted. The regulation refers to four important sources:

- European Community Household Panel (ECHP);
- Labour Force Survey (LFS);
- Community Innovation Survey (CIS);
- Continuing Vocational Training Survey (CVTS).

In summary, researchers must belong to research institutions and organisations within the MSs. A detailed proposal must be prepared stating the purpose of the research, methods to be used and details of the data to be used. Safeguards for the secure holding of the datasets will be necessary and controls on access by individuals will be required. Agreement to conditions and safeguards will be through a contract with the researchers' institution. There is no right of access to confidential data under the Regulation. MSs can withhold the data of their country from any particular research request. Access to confidential datasets can be on the premises of Eurostat with checks on the output and results to maintain confidentiality; or access can be through anonymised micro-datasets. Work is now proceeding in Eurostat, in close collaboration with the NSIs of MSs and with the research community, in putting this into practice.

Incidentally, the new Regulation 831/2002 now provides a legal definition of anonymised micro-datasets. " "anonymised microdata" shall mean individual statistical records which have been modified in order to minimise in accordance with current best practice the risk of identification of the statistical units to which they relate."

For some of the data sources mentioned in the Regulation, the first step will be the creation of anonymised micro-datasets. Although there is a wealth of knowledge of these datasets in all the NSIs, and some experience of creating anonymised micro-datasets, the work will proceed through discussion with, and agreement from, the NSIs of the MSs, using established methods as described above. The creation of anonymised micro-datasets for household and individual data seems to be relatively simple. At present work is proceeding on investigating whether satisfactorily anonymised micro-datasets can be created for records for business and enterprise data as well.

In the longer term, it is hoped that access could be provided to other datasets on the same, or a very similar, basis.

The data miner interested in accessing these anonymised datasets will need to study the Regulation carefully. The purpose must be statistical and the activity and output must be scientific research. This would seem to exclude any activities for commercial purposes. The datasets may not be brought together with, or compared with, any other datasets. The credentials of the researcher and the university or research organisation for the research proposed will need to be established.

Some of this may seem off-putting. But for genuine scientific research, the Regulation offers new opportunities to access large, respected and comprehensive

datasets covering the MSs of the European Union. The potential is great, as is the opportunity and the challenge to researchers.

References

Abowd, J.M. and Woodcock, S.D. Disclosure limitation in longitudinal linked data. In Confidentiality, Disclosure, and Data Access, (pp.215-78), North-Holland, 2001.

Defays, D. and Nanopoulos, P. "Panels of Enterprises and Confidentiality: The Small Aggregates Method" in Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada, pp.195-204.

European Union. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995, pp.0031-0050.

Karr, Alan F. and Sanil, Anish P. Web systems that disseminate information but protect confidential data. Proceedings of 53rd Session of the International Statistical Institute, 2001, Bulletin of the ISI, 53rd Session, volume LIX, book 1 pp. 265-8.