DIW Materialien
Research Notes

Research Note No. 3

Disclosure limitation methods in use: a survey

by
Flora Felsö*, Jules Theeuwes* and Gert. G. Wagner**

Berlin, August 2001

* SEO/UvA, University of Amsterdam
** DIW Berlin and European-University Viadrina, Frankfurt (Oder)

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14195 Berlin
Phone: +49-30-89789- 0
Fax: +49-30-89789- 200
Internet: http://www.diw.de

Disclosure limitation methods in use: a survey

Flora Felsö (SEO/UvA), Jules Theeuwes (SEO/UvA) and Gert. G. Wagner (DIW Berlin and EUV)

Introduction

We intend to review how statistical agencies have kept pace with new advances and whether disclosure control – combined with user friendly access to the data – is recognized as a key issue at the heart of the mission of national statistical agencies.

We collected information on practical approaches protecting confidentiality with a survey that has been sent to several national statistical agencies. We also discussed our survey results with two experts in the field, one in Europe and one in the US[1]. We were mainly interested in the methods in use and in the evolution of the applied policy, the changes in approaches in the recent past and relevant discussion of expected future changes. The geographical coverage of the survey was North America, the European Union, some aspirant members of the EU and New Zealand. One of the interviewees was a European and the other one was an American disclosure limitation expert.

We start with a literature review on disclosure control methods in use. These papers have something in common with what we are intending to do, but these papers had a different scope or purpose. Either they aimed at a comprehensive description of disclosure limitation techniques and the summary of the practice of different statistical agencies was only a side issue or only a small group of countries were taken under the loop. After the literature review, we turn to our research and we discuss our approach. First we give the explanation of the rationale for the questions in the survey. Subsequently there will be a description of the sample 'design' and the actual respondents. After that, we summarize the results of the survey. To begin with, we will sum up disclosure control methods for tabular data. Here, we will make a distinction between disclosure methods for tables of counts or frequencies and methods used for tables of magnitude data. Finally, we discuss disclosure limitation methods for microdata. A distinction will be made between demographic data and economic (establishment) data. For demographic data a distinction will be made between census data (including all members of the population) and non-census or sample databases.

[1] We would like to thank Eric Schulte Nordholt (CBS, Netherlands) and Alvan O. Zarate (NCHS) for their comments, their time and for sharing their expertise in the area of disclosure limitation methods.

Literature review

The first comprehensive work on disclosure avoidance was the Statistical Policy Working Paper #2 published by the Subcommittee on Disclosure-Avoidance of the Federal Committee on Statistical Methodology (FCSM) in the US, as early as 1978. The intention of this paper was to help managerial and technical staff of US Federal agencies to achieve "appropriate disclosure avoidance".

One of the conclusions of this paper was that comparatively little was known about disclosure. To begin with, there was no widely accepted definition of "disclosure" (p.1). The authors had to work out a framework in which disclosure practices could be reviewed and evaluated.

Nevertheless, they found that several major Federal statistical agencies had developed a variety of disclosure avoidance techniques for tabulations and microdata, but it appeared that little attention had been given to determine what constitutes disclosure and how it was decided which disclosure occurrences were or were not acceptable (p.32).

The authors recommended that all statistical agencies should formulate and apply policies and procedures designed to avoid unacceptable disclosures. Because there were wide variations in the information released, the authors did not feel that it was feasible to develop a uniform set of rules, applicable to all agencies for distinguishing acceptable and unacceptable disclosures. The only advice they gave was that agencies should apply a test of reasonableness, i.e. that no information about a specific individual or an entity would be disclosed in the manner that could harm a respondent (an individual or an entity) (p.39-42). The subcommittee also argued that special care should be taken when releases are based on a complete file - as opposed to a sample, and when data is presented for small areas (p.42).

Another interesting recommendation was that there should be a clear assignment of individual responsibilities for compliance with the chosen disclosure avoidance policies (p.42). As we will see, several (non-US) statistical agencies have not yet followed this advice 22 years later.

The Statistical Policy Working Paper contained an Appendix with the description of disclosure avoidance practices of seven Federal statistical agencies, prepared by the agencies.

The follow-up on the Subcommittee on Disclosure-Avoidance came when Hermann Habermann, Statistical Policy Office of the Office of Management and Budget, organized an ad hoc interagency committee on disclosure risks in early 1992. A new subcommittee was formed to look at methodological issues, to analyze results of an informal survey of agency practices and to develop recommendations for improvement. This subcommittee was the predecessor of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology, established in 1993.

Note, that the name has changed: disclosure limitation (or control) instead of disclosure avoidance. The new terminology reflects that it has been realized that zero-risk condition for disclosure is an impossibly high standard (website FCSM).

One of the goals of the Subcommittee on Disclosure Limitation Methodology was to update the Statistical Policy Working Paper #2. These efforts have resulted in Statistical Policy Working Paper #22 (FCSM, 1994). The aim of this work was to describe and evaluate existing disclosure limitation methods, to provide recommendations and guidelines for the selection and use of disclosure limitation techniques, to encourage the development, sharing and use of new methods and specialized software.

Chapter II of this paper gave a simple description of disclosure limitation methods for tabulations and microdata. It provided a "guideline for good practice for all agencies (p.3)". This guideline has also served as the basis for *our* survey.

Chapter III of the Statistical Policy Working Paper #22 summarized the disclosure control methods used by twelve major federal agencies and programs. This review was based on a number of resources. First, information had been used from the contribution of statistical agencies in response to a request of the Panel on Confidentiality and Data Access, Committee on National Statistics (results published in Jabine, 1993). Additional information had been taken from the reports of seven Federal statistical agencies that had been published in Statistical Policy Working Paper #2. However, the main resource of this review were the 12 responses that arrived on the request of Herman Habermann, whereby each statistical agency had been asked again to provide a description of its current practices, standards and research plans for tabular and microdata.

The main conclusion was that most of the agencies had standards, guidelines or formal review mechanisms to ensure that adequate disclosure analyses are performed and appropriate statistical disclosure methods are applied prior to the release of tables or microdata (p.37). These practices exhibit a large range of specificity: some applied only a couple of simple rules while others' rules were much more detailed.

Referring to magnitude and frequency data, they concluded that most agencies applied a minimum cell size and some type of concentration rule. Minimum cell sizes of three were almost invariably used, because each member of a cell of size two could have derived the value of the other member (p.37).

Only half of the responding agencies had established disclosure limitation procedures for microdata. The authors cited Jabine (1993) and claimed that procedures for microdata were not parameter driven like those for tabulations. The protection of microdata requires judgements that take into account whether resources are available "that might be needed by an 'attacker' to identify individual units", the expected number of unique records in file, etc. Furthermore, geography is an important factor for microdata release. Also, the number of variables is important, for example, with the use of the characteristics of the local area, the location and thereby an individual could be identified. Another conclusion was that top-coding was a commonly used method to prevent the disclosure of

individuals (or entities) with extreme values. Blurring, noise and rounding were also applied to prevent disclosure (p.38).

Like FCSM, Eurostat has also released a guide on disclosure limitation (Eurostat, 1996). The purpose of this manual was not to recommend the choice of a certain method, but rather a list of possibilities that can be applied given the type of data release and the intended level of data protection.

This release was followed by a survey on disclosure practices among the Member States of the EU in 1997 carried out by Holvast & Partner (Holvast, 1999). The main conclusion was that all national statistical institutes were aware of the importance of confidentiality and implemented the necessary safeguards. Mathematical and computing aspects were considered important by 10 out of 14 European statistical institutes (Holvast, 1999, p.201). This paper found that there is little difference in the official treatment of individual and business data, what counts is the nature of the data (e.g. frequencies versus magnitude data). An interesting point that Holvast is stressing is the specific problem of small countries. In a small country simple sector statistics could cause problems more easily because the limited number of firms in certain sectors.

A follow up of the Holvast report was the survey done by the UN Economic Commission for Europe, carried out among fourteen (Central) Eastern European countries (EEC) and five countries from the Commonwealth of Independent States (CIS) in 1998. Most EEC statistical agencies considered the protection of data to be very important, however no country reported a systematic approach to statistical data protection. The required general legal basis for statistical disclosure control already existed in most of the (Central) Eastern European countries. The main difficulties were not in the legal framework, but in the organizational and technical implementation of the legally defined confidentiality principles. Little attention had been given to mathematical and computing aspects of data confidentiality. In other countries (mainly CIS states) the legal base did not exist.

About half of the countries reported that they disseminated public use files of their labor force survey. Other forms of data release, such as detailed tabular data (twelve countries out of the sixteen) and microdata for research (nine out of the sixteen) were used more frequently. The most frequently used disclosure limitation technique for demographic data was categorization of variables. Rounding, sub-sampling, micro-aggregation, imputation of missing values and top-coding were also reported. The (Central) Eastern European countries used several techniques parallel to each other. In the CIS countries the listed techniques were practically not used.

Only three countries released public use files on business statistics. Other countries used forms of release similar to demographic data, such as tabular data and - less frequently - microdata for research and synthetic data files. On-site access was provided sometimes, but less frequently than for demographic data. The disclosure limitation methods were somewhat different for business data than for labor force data. Categorization of variables, imputation of missing values and micro-aggregation were employed in about half of the countries. Sub-sampling, top-coding, rounding and the dominance rule were

also reported. Data swapping and the use of special software were applied in one country only. Adding noise was not used at all. Access to business data was much more restricted than access to demographic data.

The 1998 survey has been updated in 2000/2001 (UN/ECE, 2001). Now all statistical offices expressed that they recognize the importance of the protection of statistical data. Increased attention is partly caused by the Population Census in 2000 and agricultural censuses conducted in some transition countries. Also, the increasing gap between rich and poor and growing criminality in some countries made public attitude toward confidentiality very sensitive, especially concerning the privacy of the individual.

According to the survey in 2000, most attention was still paid to the legislative and administrative aspects. Mathematical aspects received less attention, just like it had been the case in the previous survey. Data protection was still done on an ad hoc basis in different statistical areas. Strictly speaking, the most popular methods in 2000 were: minimum cell count rules (usually three), geographical or population thresholds (releasing data only for areas above a particular spatial or population threshold), re-coding data into broad categories and dominance rules (if fewer than a certain number of units -usually two or three- account for a certain percentage of more of the cell total). The methods reported by the transition countries are summarized in Table 1. and 2.

The situation has not changed regarding the willingness of statistical agencies to allow access to original data to other institutions. An easy solution to avoid disclosure is simply not releasing microdata at all. Out of the 18 responding countries, 11 did not release microdata on natural persons and 9 did not release microdata on enterprises in 2000. True, the pressure on statistical offices to release microdata was not so intense as in Western countries. When releasing microdata, it was often assumed that omitting explicit identification variables (namely addresses) was sufficient to avoid disclosure. The forms of release of microdata that are more often used were ‘microdata for research’ and synthetic files. There were some public use files on businesses.

The general conclusion was that the transition countries are in need of technical assistance, software and training in particular.

Table 1. The disclosure limitation techniques used for the release of data on natural persons

	Tabular data						Microdata							
	Geographic or population thresholds	Minimum cell count rules	Dominance rules	Rounding	Adding noise, blurring	Re-coding variables	Geographic or population thresholds	Sampling	Top- and bottom-coding	Re-coding data into broad categories	Data swapping	Deletion of sensitive records	Deletion of data items	Micro-aggregation
Bulgaria		+	+			+			+	+			+	
Czech Republic		+	+											
Estonia		+	+	+										+
Hungary	+	+		+		+							+	
Latvia	+												+	
Lithuania	+	+		+				+	+			+		+
Poland	+					+	+			+				
Romania	+	+				+	+	+		+			+	+
Slovakia	+	+	+			+	+	+	+	+			+	+
Slovenia	+	+	+	+				+		+			+	
Macedonia		+												
Yugoslavia														
Armenia														+
Azerbajan														
Belarus														
Kyrgyzstan										+				+
Russia														
Turkmenistan														
Total														

Source: UN/ECE Secretariat (2001) Statistical Data Confidentiality in the Transition Countries: 2000/2001 Winter Survey.

Table 2. The disclosure limitation techniques used for the release of data on enterprises

	Tabular data						Microdata							
	Geographic or population thresholds	Minimum cell count rules	Dominance rules	Rounding	Adding noise, blurring	Re-coding variables	Geographic or population thresholds	Sampling	Top- and bottom-coding	Re-coding data into broad categories	Data swapping	Deletion of sensitive records	Deletion of data items	Micro-aggregation
Bulgaria		+	+			+			+	+			+	
Czech Republic		+	+											
Estonia														
Hungary		+											+	
Latvia	+	+	+			+	+						+	
Lithuania	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Poland	+	+					+			+				+
Romania								+					+	+
Slovakia	+	+				+	+	+	+	+				+
Slovenia	+	+	+							+			+	
Macedonia														
Yugoslavia														
Armenia														+
Azerbajan														
Belarus														
Kyrgyzstan										+	+			+
Russia														
Turkmenistan														
Total														

Source: UN/ECE Secretariat (2001) Statistical Data Confidentiality in the Transition Countries: 2000/2001 Winter Survey.

Another initiative was the study on behalf of the (German) Federal Ministry of Education and Research (BMBF) by an expert committee (*Kommission zur Verbesserung der Statistischen Infrastruktur in Zusammenarbeit zwischen Statistik und Wissenschaft [KVI]*)[2]. In this report the results of an international survey of users of official microdata are summarized, undertaken by a panel of experts to analyze Germany's statistical infrastructure. As opposed to the earlier mentioned studies, this survey approached the users of microdata to ask their opinion on the user-friendliness of different dissemination and disclosure limitation approaches and about the general judgment on the possibilities of microdata analysis in the respondent's home country.

The KVI survey was sent to 16 experts in eight countries (one sociologist and one economist each). The countries selected were similar to our survey: USA, Canada, France, Germany, Netherlands, Norway, Sweden and the United Kingdom. This overlap is not a lucky accident but rather an indication for the fact that those countries are the most interesting ones with respect to microdata release. Fifteen experts returned the questionnaire.

According to the KVI survey (cf. KVI 2001a) the major complaint of users of official microdata were not about limitations due to disclosure control but concerned the high costs, bad documentation and non-transparent policies of access.

A major barrier to access was the high cost of disclosure control measures charged to individual users. Whereas statistical microdata is treated as a 'public good' in the USA and Canada, in some European countries substantial fees are charged which was a major reason for complaints. The cases of the USA and Canada showed that minimal use charges (fees, which cover just the marginal costs of dissemination,) could be realized by different means. In the US the statistical agencies covered the cost of producing public use and/or scientific use files out of their own budgets, whereas in Canada the universities paid to Statistics Canada a flat rate which covered the cost of producing public use and scientific use files. Members of those universities could obtain these data without paying any fees. In the Netherlands the "National Science Foundation" (WSA) paid a flat rate to Statistics Netherlands (CBS) and researchers pay a marginal fee per data source.

The KVI survey brought up an even more important observation. Even in the case of low costs for access to microdata there can be limitations of access if the rules are unclear and the documentation of the data is weak (or even non-existent). All respondents in all countries judged the documentation on microdata which was provided by the statistical agencies themselves to be inferior (in need of improvement) or nearly non-existent. Special service agencies like those established in Norway and the Netherlands were evaluated as very positive. These agencies, which were funded by the funding bodies for research (in other words: which are not funded by statistical agencies), provided transparent service and reasonable documentation.

[2] For an English written summary of the work of KVI see KVI (2001b).

Survey questionnaire: a motivation

We now turn to our own survey. The main objective of our paper and survey was, to detect which methods (if any) are adopted by national statistical agencies in the sample.

The questions of the survey[3] were mainly based on the paper Statistical Policy Working Paper # 22 (FCSM, 1994). We have used the framework of chapter II to ask simple questions (mostly yes/no) about specific methods being used. The respondents were invited to comment anywhere where they felt like sharing personal thoughts or comments. Many of them enclosed papers written by themselves or colleagues, thereby supplying us with a lot of extra information.

In the survey we are looking at tabular and microdata separately. For tabular data, we make a further distinction for disclosure methods for tables of counts or frequencies and for tables of magnitude data. While looking at microdata-practices, we consider demographic microdata and economic microdata separately. Within demographic data, we differentiate between census- and non-census data.

At the end of the questionnaire we asked about the evolution of disclosure limitation methods in the recent past, whether the agency has ever experienced disclosure related problems and whether the institute has publications on this topic. We also asked about the use of specific software packages. With these questions, we wanted to trace a number of relevant issues, such as the evolution of practices and the implicitly or explicitly formulated disclosure control methods. We also wanted to explore whether disclosure control is a matter of concern within the statistical agencies. The results were discussed two disclosure limitation experts, one from Europe and one from the US.

Sample “design” and distribution

The survey was distributed by e-mail. We used e-mail for several reasons. A major factor was our concern about the speed of the survey procedure. We did not have an address list of staff members within the statistical agencies currently dealing with these issues. As a start, we used participants’ lists of conferences dealing with confidentiality issues. Obviously, a lot of the addresses were not valid any more. Some persons asked were no longer in charge of disclosure control and we had to ask them to forward the questionnaire to colleagues who would be able to help us. The first mailing was sent to 25 national statistical agencies.

The general objective was to get as many responses as possible, but in any case, we wanted to cover the following countries: USA, Canada, France, Germany, Netherlands, Norway, Sweden and the United Kingdom. This turned out to be quite an ambitious aim, but thanks to the keen interest of statistical agencies in disclosure issues and the co-operative disposition of many staff members, we largely succeeded in our objective. We

[3] See Appendix

wish to thank the many people involved in the survey for their kindness and their help, sometimes above and beyond the call of duty[4].

For reasons mentioned before, and also due to simple non-response, we did not always find the best point of entry at the statistical institute. We had to ask help from co-authors of this book and their many connections to bring us in contact with the right people at the right places.

All these efforts have resulted in the responses of the (national) statistical agencies of Canada, the Czech Republic, Denmark, Estonia, Germany, Hungary, Italy, Lithuania, Netherlands, New Zealand, Norway, Sweden and the USA (in the US even from 4 different sources).

Our impression was that most of the time, (chief) statisticians in charge of many statistical procedures, including policies related to disclosure completed the questionnaire. Only in some cases, we noticed that someone has been appointed specifically to co-ordinate disclosure related problems. Some even note that these concerns are rather new, and often are not yet fully centralized in one hand.

Of course, we realize that our survey suffers from severe selectivity in response. Those agencies that have an interest in the topic and have already taken measures have probably responded more often. Hence we cannot claim that our results are representative for national statistical agencies in general.

The 16 questionnaires that were returned are not equally complete. Some left out the section about microdata for the census or microdata on economic data, simply because they do not release that type of data or it was not the responsibility of the one who filled out the questionnaire. There were however some questionnaires sent back that were the collective response of different departments of an agency, where all sections had been answered by the staff member in charge of the specific issue.

The results

Tabular data: tables of counts or frequencies

Tables of counts or frequencies are usually referring to individuals or households. Those tables display numbers of individuals by category or fraction of individuals by category.

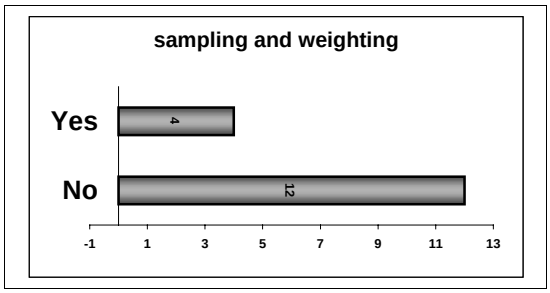
The first question considering tables of counts and frequencies was whether sampling and weighting, without publishing the weight was used as a statistical disclosure limitation method. Only four agencies report that they use this technique, see Figure 1.

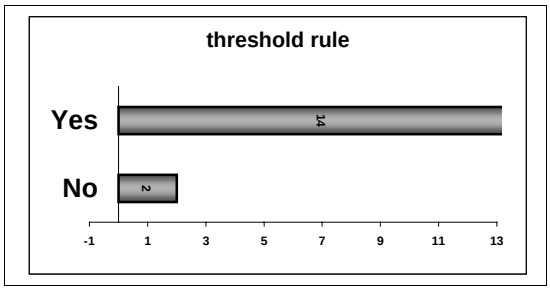
[4] We are very grateful for the help we got from Juergen Chlumsky, Mark Elliot, Virág Erdei, Francis Kramarz, Julia Lane, Mark Schipper, Rainer Winkelmann and Virgina de Wolf .

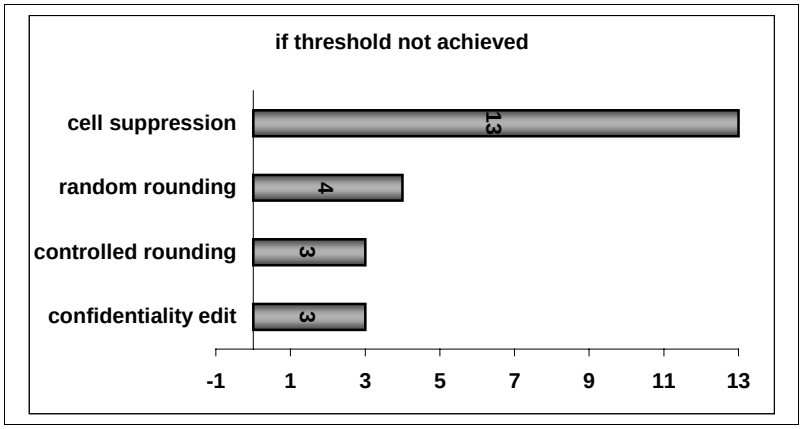
Most agencies report that they apply a threshold rule where a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number. The most frequently reported threshold was 3 (9 agencies). Others say that the threshold varies by application and gave 4, 5, 10, 30 and 75 as thresholds. Some comment that the minimum cell size is chosen largely for quality purposes (e. g. avoidance of big sampling errors) rather than disclosure avoidance.

We find that all agencies that apply the threshold rule (and answered the question) apply cell suppression; other techniques are much less widespread. Of course, many respondents note that different techniques are used for different applications. An interesting comment is also that random rounding is used occasionally but never as a sole protection method.

Figure 1. Disclosure methods for tables of counts or frequencies







Other special rules that respondents report are special restrictions on the level of detail that can be provided in a table, such as:

- other population thresholds;
- rules against the publication of certain levels of geographic detail;
- collapsing categories;
- releasing only total frequency counts and percentages;
- there should be no cross-tabulations with levels of detail that would allow spontaneous recognition of a population unique[5];
- combining categories that may include top and bottom coding as well as collapsing intermediate categories;
- applying the dominance rule[6] ;
- only concerning highly confidential frequency data: apply the threshold rule to the difference between number of respondents to the (sub-)marginal in each dimension of the table.

Tabular data: Magnitude data

One of the agencies reports that it does not release magnitude data, one did not comment and another respondent reported that all items are still being studied. A total of thirteen agencies' responses are used in this section.

Magnitude data are generally economic data reporting nonnegative quantities about certain business establishments or institutions. The distribution of these values is likely to be skewed, with a few entities having very large values. Disclosure limitation for these types of data concentrates on making sure that the values reported by the largest, most visible respondents cannot be estimated too closely. To this end primary suppression rules (see Box 1.) have been developed to determine whether a cell could reveal sensitive information, especially to make it difficult for one of the respondents of a survey to estimate the value reported by another respondent too closely. We have asked the statistical agencies how they identify these sensitive cells.

- [5] E.g. the 'occupation' of someone being 'member of parliament' and 'location' 'small town' would enable recognition of a population unique, even if the weighted up frequency was around 100.

- [6] E.g. if two units account for more than 80% in terms of turnover.

Primary suppression rules

(n,k) rule: "Regardless of the number of respondents in a cell, if a small number (n or fewer) of these respondents contribute a large percentage (k percent or more) of the total cell value, then the so-called n respondent, k percent rule of cell dominance defines this cell as sensitive" (page 48 in the Statistical Policy Working Paper # 22).

P-percentage rule: "Approximate disclosure of magnitude data occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a prespecified percentage, p." (page 46)

pq rule: "In the derivation for the p-percent rule, one assumes that there was a limited prior knowledge about respondent's values. Some believe that agencies should not make this assumption. In the pq rule, agencies can specify how much prior knowledge there is by assigning a value q which represents how accurately respondents can estimate another respondent's value before any data are published ($p<q<100$)." (page 47)

The most widespread technique to identify sensitive cells is the (n,k) rule (eight agencies). One respondent commented, however, that in the future they would shift to the p-percent rule (currently applied by three agencies). Another agency noted that they have already made this change. Only one of the respondents reported that they apply the pq rule. Two agencies reported that they use another method. One claimed that they use the threshold rule. The other one reported a comprehensive set of special guidelines:

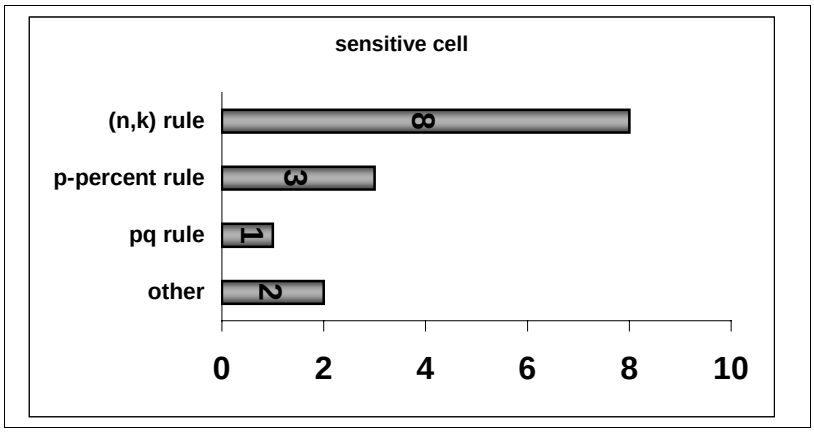
- In no table should all cases of any line or column be found in a single cell.
- In no case should the total figure for a line or column of a cross-tabulation be less than three.
- In no case should the quantity figure based upon fewer than three cases.
- In no case should a quantity figure be published if one case contributes more than 60% of the amount.
- In no case should the data on an identifiable case, nor any of the kinds of data listed in the preceding items, be derivable through subtraction or other calculation from the combination of tables published on a given study
- Data published by the agency should never permit disclosure when used in combination with other known data.

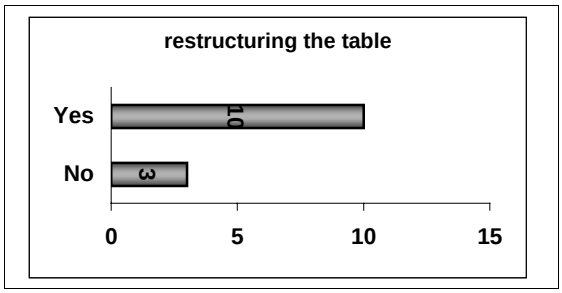
However there are some exceptional cases where no suppression is applied. There are some cases 'accepted traditionally', these cases being special cases of some special statistics (NCHS p.19).

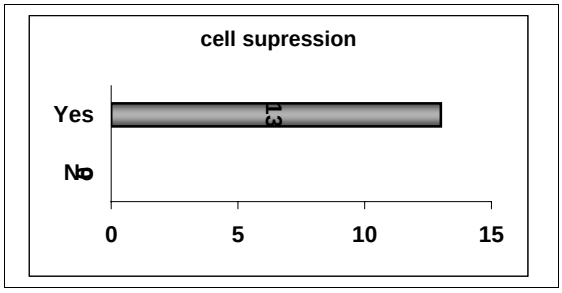
After having the sensitive cell identified, an agency could choose to restructure the table by collapsing cells until no sensitive cell remains, or one could choose to use cell suppression. Ten agencies have reported that they restructure the table, some claim that they usually choose this option and some occasionally, e.g. if the number of primary cell suppressions would be large. All thirteen respondents practice cell suppression.

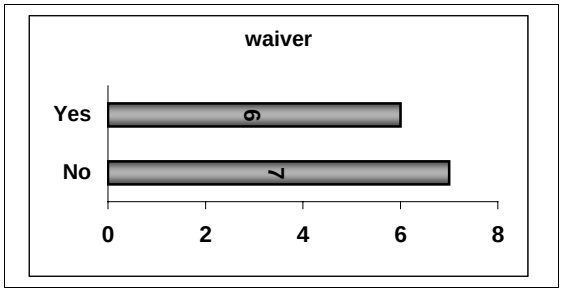
There is also an administrative way to avoid cell suppression: obtaining a permission from a respondent in the sensitive cell, to publish the cell. This is occasionally used by six of the thirteen agencies. One agency comments that, only in rare cases, they request a "waiver" for instance for a large public company where similar information is already in the public domain.

Figure 2 Tables of magnitude data









Demographic microdata

Most agencies report that they do not release microdata to the public, but they provide access to researchers under certain circumstances, mostly in controlled environments, in some exceptional cases even outside the premises of the statistical agency.

Nine agencies reported that they release microdata from a census either for public use or for restricted research purposes. One respondent wrote that they are not yet releasing microdata from their census, but they are planning to do so in the future, with all identifiers removed. Another respondent is currently in de process of creating an on-site working facility for researchers.

Respondents also report that their main tools for protecting census microdata are excluding obvious identifiers, limiting geographic detail, and limiting the number of variables on the file. Other methods that have been reported (but are not included in Figure 3) are:

- issuing multiple files, one with more detailed geography and less detailed characteristics and the other with less detailed geography and more detailed characteristics;
- grouping, by splitting continuous variables into ranges to reduce detail, as opposed to top or bottom-coding
- deletion of sensitive records.

Non-census demographic micro-data is being released by thirteen agencies. All files are protected by excluding obvious identifiers and limiting geographic detail. Limiting the number of variables (the file is often customized to provide only the variables needed for the research project) and top- and bottom-coding is also applied quite frequently.

Other methods (not listed in Figure 3) have also been reported:
- deletion of sensitive records, deletion of sensitive items, recoding into broad categories, sampling, micro aggregation;
- blanks/missing values are imputed in unsafe records when other methods fail to protect these records;
- local suppression;
- grouping;
- eliminating any variables that can be used to link to external sources that contain individual identifiers or more geographic detail than can be released on the microdata files.

One agency reported that for release to the public the file is reviewed in a systematic manner by a disclosure review team composed of subject matter specialists, sampling experts and experts in disclosure analyses. Procedures are codified in a disclosure potential ‘checklist’ which subject matter specialists are required to fill out.

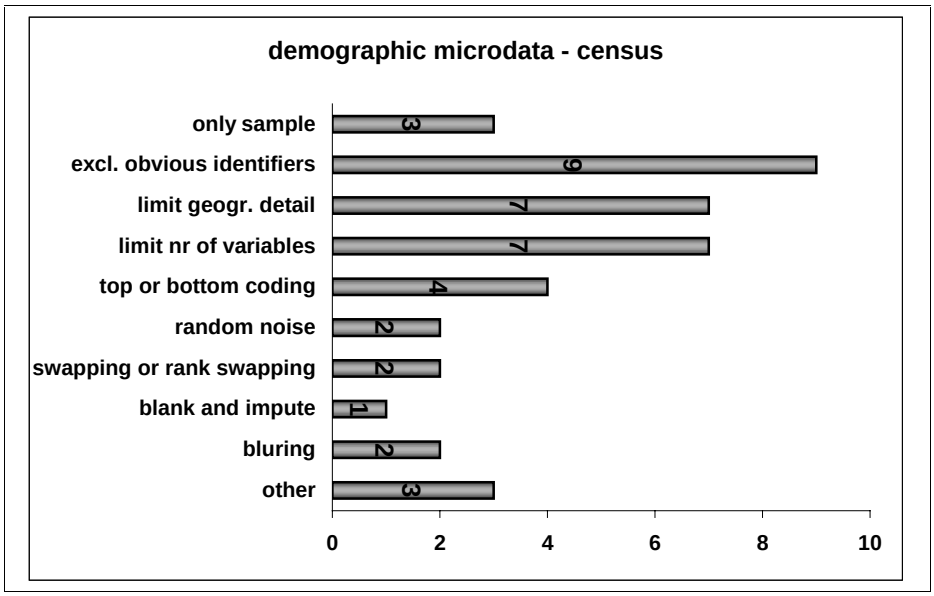
An interesting comment on the question whether they release data from only a sample of the respondents was that in case of a panel they always provide all respondents in a given year, but the years selected are limited.

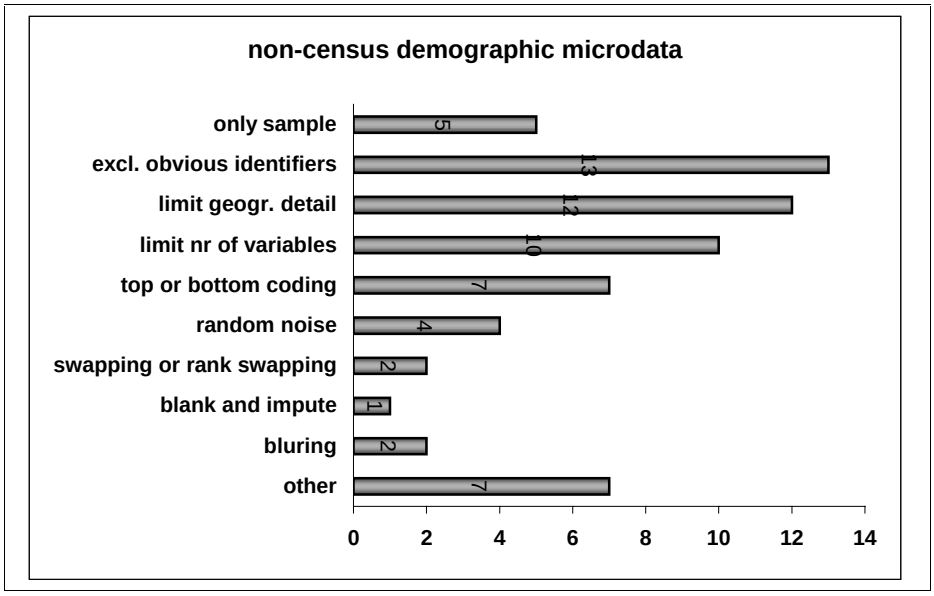
One respondent reports that it is sometimes possible to release a very limited number of variables. A single request, or perhaps even several can be considered, so long disclosure risk analyses does not indicate confidentiality problems. However they assume that any release to any member of the public is a release to the entire public, so that successive releases of information on a few variables are regarded as cumulative, even though the data were released to different parties. Hence at some point the marginal release is seen as an unacceptable disclosure risk.

One agency reports that they are offering Analytical Programming Services, meaning that users submit computer programs electronically and receive output that is subject to disclosure review before delivery to the user. There is also an automated version of this service, provided through a modem, called Remote Access.

Short-term appointments of outside researchers in the statistical office that implies that they sign confidentiality statements are also applied by several statistical agencies.

Figure 3 Microdata of demographic data





Economic microdata

There are virtually no public use microdata files released for establishment data. For research purposes however, statistical agencies do provide access to business microdata. The problem with this type of data is that the informations collected about business establishments are primarily magnitude data. Their distribution is skewed and large establishments would be easily identifiable with the use of other publicly available information.

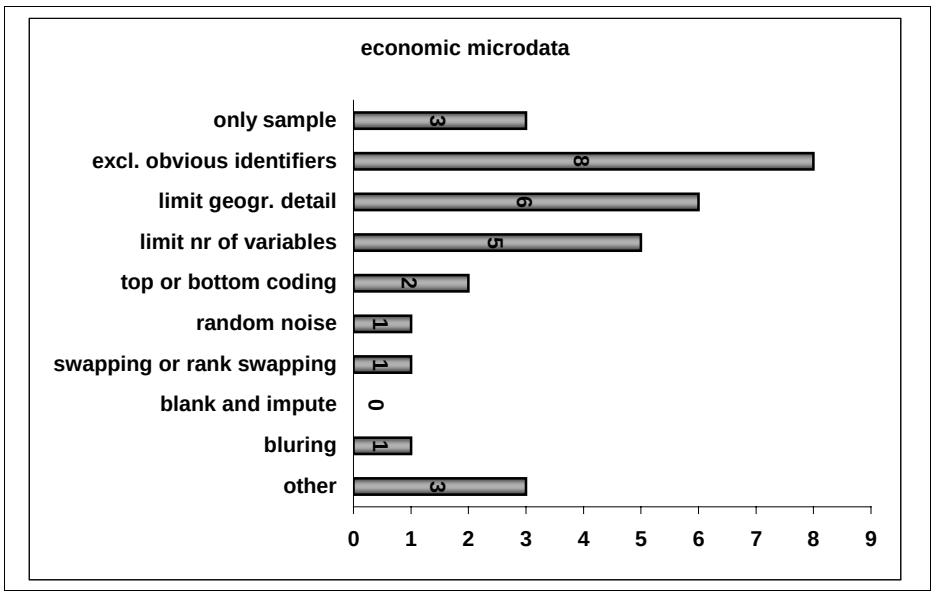
In our sample we have eight agencies that provide researchers access to establishment microdata. All of them exclude obvious identifiers while providing access to researchers. Most of them limit geographical detail and the number of variables on file. There were some answers that were not on our survey list, such as:

- grouping;
- deletion of sensitive records, deletion of sensitive items, recoding into broad categories, sampling, micro aggregation;
- imputing blanks/missing values in the unsafe records when other methods fail to protect these records.

One agency reports that it does not provide microdata for public use or on a scientific use basis, but it does provide another service, similar to the Analytical Programming Service reported in the section of demographic microdata. In certain (exceptional) cases the statistical institute carries out user-requested specialized data analysis. In these cases the user (researcher) is expected to develop a computer program to do the analysis, which will be run by staff members of the statistical agency on the original microdata file. Staff will do a confidentiality check of the output, before it is sent to the researcher. To

facilitate these checks, only programs developed on the basis of standard software (such as SPSS) are accepted.

Figure 4. Microdata of economic data



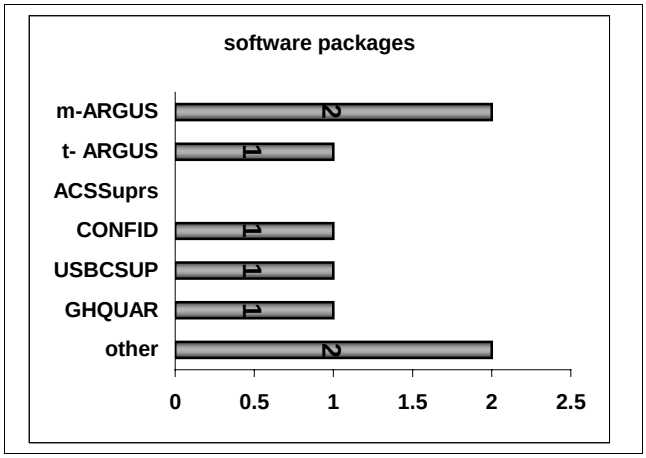
Software packages in use

There is no clear consensus of which software packages should be used. In fact, there are no widespread software packages. The only program that is being used by more than one agency in our sample is μ-ARGUS. Two statistical agencies report that they use "home-made" programs.

Some agencies comment on these software. One of our respondents reports the following: "CONFID has been evaluated in 1996 and decided that it would be difficult to integrate into the system of the agency. ACSSuprs has also been evaluated in 1997-98. It is promising, but some questions as to availability of on-going support. τ-ARGUS has been evaluated in 1997-98. It is also promising, but cell suppression methodology is not complete. Will keep up-to-date with developments. μ-ARGUS is not a major issue for the agency, but they would be interested if they provided confidentialised microdata."

Another agency does not use any specific software packages currently, but is intending to implement τ-ARGUS and μ-ARGUS in the future. τ-ARGUS and μ-ARGUS are also being tested for potential use by three other respondents; however, one of them reckons that packages that are more easily integrated into their general production system (based on SAS) would be convenient.

Figure 5. Software packages in use



The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project of the Fourth Framework Programme of the European Union. New versions of the ARGUS packages are being developed by the methods and informatics department of Statistics Netherlands, as part of the Computational Aspects of Statistical Confidentiality project in the Fifth Framework Programme of the European Union. For τ-ARGUS, for example research is being done in the field of disclosure limitation for linked tables, more options for secondary suppressions and consecutive years of the same survey. In the case of μ-ARGUS, the difference between protecting microdata for research and protecting public use files is of interest (Schulte Nordholt, 2001 p.11-12).

We have also been notified of the collective effort of the Confidentiality and Data Access Committee (CDAC) and the FCSM to develop an audit software for large tabular presentations. To this end, the Subcommittee on Disclosure Auditing has been formed to develop a framework, after which the Energy Information Administration (EIA) has taken the lead in this project. The software is designed to assist in releasing the most detailed data possible with acceptably small risk of the identification of confidential data. The software is currently referred to as Disclosure Audit System (DAS).

Evolution of disclosure limitation methods used

It seems that the rather complicated disclosure probability models of the literature are not directly applied by statistical agencies. Instead, some more or less ad hoc rules are set and applied for a whole family of statistical data. Some institutes are exceptions. It seems that there are a number of statistical institutes (mostly in the US) where special teams are formed consisting of experts from different fields (such as mathematicians, economists,

sampling experts) to examine data files before it is released to the public. These teams are referred to as Disclosure Review Boards. They are responsible for disclosure limitation policy regarding all publicly available data products of a given federal statistical agency. They review cases (or sometimes all publications) where some concern arises, they are responsible for communication inside and outside the institute and coordinate research. A major concern of these Disclosure Review Boards is whether there are external databases available that can potentially be matched to databases to be published.

Concerning the technical part of disclosure limitation efforts, our impression from our and other surveys is that most of the aspirant member countries of the European Union have experienced a dramatic change in their disclosure control practices: they have greatly developed new methods in the past ten years.

For the other agencies, it was not so much the 'rules' that changed, but the tools that are available. The appearance of different specialized software packages triggered a lot of interest in automated procedures. Most agencies are experimenting or gathering knowledge of these software but so far are cautious to install them, mainly because of difficulties in integrating them in their current systems. The most frequently mentioned software packages are the ARGUS packages, but so far it is often in the phase of studying and testing of it. We got the same impression from the UN/ECE studies. The CDAC/FCSM audit software is not yet available, and might also be subject to restrictive access in the future. Nevertheless, some agencies attempt to develop 'home-made' programs.
We also find evidence that there is a shift from interest in blank (cell suppression) and imputation to data swapping and from the (n,k) rule to the p-percent rule. However, our evidence is suggestive only.

Apart from the challenge of automated systems there was an external influence that surely will have an effect on future research and development. There has been a great increase in the demand for access to microdata from researchers and others. This in itself is the consequence of the development of statistical software packages. A lot of agencies provide access to microdata now. Some have established data laboratories and access sites on the premises of the statistical agency, where researchers are provided access in a controlled environment. These developments have challenged and will continue to challenge traditional confidentiality issues. In this context it is interesting to report that three of our respondents have experienced disclosure-related problems in the recent past.

Literature

Energy Information Administration (2001) Example of a Table Suppression Audit, Confidentiality and Data Access Committee (CDAC): Demonstration of Disclosure Limiting Software (DLS).

Erdei, Virág (2001) *A statisztikai adatok vedelme a magyar Központi Statisztikai Hivatalban* Paper presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, 14-16 March 2001.

Eurostat (1996) *Manual on disclosure control methods*. Luxemburg: Office for Publications of the European Communities.

Federal Committee on Statistical Methodology (1978) *Report on Statistical Disclosure and Disclosure Avoidance Techniques* (Statistical Policy Working Paper #2) Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Federal Committee on Statistical Methodology (1994*) Report on Statistical Disclosure Limitation*, Washington DC (Policy Working Paper #22) US Office of Management and Budget, Statistical Policy Office.

Holvast, Jan (1999) *Statistical dissemination, confidentiality and disclosure* In: Eurostat: Statistical Data Confidentiality – Proceedings of the Joint Eurostat/UN-ECE Work session on Statistical Data Confidentiality held in Thessaloniki in March 1999.

Hoy, E., M.M. McMillen, F. Scheuren, G.D. Stamas, G.T. Therriault and A.O. Zarate (2000) *Panel on Disclosure Review Boards of Federal Agencies: Characteristics, Defining Qualities and Generalizability*, Paper presented at the Joint Statistical Meetings, August 17, Indianapolis, Indiana.

Jabine, T.B. (1993) Procedures for Restricted Data Access, Journal of Official Statistics, vol.9. no.2 537-589.

Kommission zur Verbesserung der statistischen Infrastruktur (KVI*)* (2001a), *Wege zur Verbesserung der statistischen Infrastruktur (Gutachten der Kommission)*, Baden-Baden 2001: mimeo.

Kommission zur Verbesserung der statistischen Infrastruktur (KVI*)* (2001b), Ways Towards an Improved Informational Infrastructure, *Schmollers Jahrbuch – Journal of Applied Social Science Studies*, 121(3) (in print).

Luige, Tiina and Jana Meliskova (1999) *Confidentiality practices in the transition countries*, In: Eurostat: Statistical Data Confidentiality – Proceedings of the Joint Eurostat/UN-ECE Work session on Statistical Data Confidentiality held in Thessaloniki in March 1999.

NCHS (1980) Staff Manual on Confidentiality (under revision).

Schulte Nordholt, E. (2001) Progress in the Implementation of SDC Methods and Techniques in Central and Eastern Europe – List of Key issues for discussion for the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, 14-16 March 2001.

UN/ECE Secretariat (2001) *Statistical Data Confidentiality in the Transition Countries: 2000/2001 Winter Survey*, Paper presented on the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, 14-16 March 2001.

Zarate A.O. (1998) *Legal, Administrative and Statistical Aspects of Confidentiality Procedures at the National Center for Health Statistics Presentation,* Paper presented as expert testimony on Issues of "Privacy and Confidentiality" for the public Meeting on the President's Initiative on Immunization Registries, Atlanta, Georgia, July 16, 1998.

Disclosure limitation methods in use

Questionnaire

Explanatory notes

There is a fundamental tension at the heart of every statistical agency's mission. Each is charged with collecting high quality data to inform national policy and enable statistical research. This necessitates dissemination of both summary and micro-data. Each statistical agency is also charged with protecting the confidentiality of survey respondents. This often necessities blurring the data to reduce the probability of re-identification of individuals. The tradeoff dilemma, which could well be stated as protecting confidentiality but maximizing access, has become more complex as both technological advances and public perceptions have altered in the information age.
We intend to make a review of how statistical disclosure techniques have kept pace with these changes.

The **objective** of this **questionnaire** is to gather information on practical approaches protecting confidentiality of respondents but maximising access to demographic and economic data. We are interested in the methods in use and also in the evolution of the applied policy, the changes in approaches in the recent past and relevant discussion of expected future changes.

On the contrary to the topic, we will not treat your answers as confidential. We do want to make an overview of the applied techniques. The results of our survey will be published in a book *Confidentiality, Disclosure and Data Access: Theory and Practical Approaches for Statistical Agencies* edited by Pat Doyle, Julia Lane, Jules Theeuwes and Laura Zayatz and to be published by Elsevier North Holland in the beginning of 2002. The book will contain technical chapters on the techniques of disclosure limitation methods as well as two chapters on the practical approaches protecting confidentiality used by national statistical agencies. We will have one chapter on demographic and one on economic data.

Of course, if you desire to share personal thoughts or critique with us that is no part of the policy of your agency we will keep that information confidential if requested.

Answering this questionnaire will take approximately 20 minutes of your time. Your contribution will be highly appreciated.

Please fill in into this file.

If necessary please duplicate specific sections !

You are kindly requested to fill in this questionnaire and return it to floraf@seo.fee.uva.nl.

In case you have specific difficulties answering the questions, you may contact Prof. Gert G. Wagner: (00)-49-30-89789-290 (gwagner@diw.de) concerning demographic data or Prof. Jules Theeuwes or Flóra Felsö: (00)- 31-20-6242412 (theeuwes@seo.fee.uva.nl or floraf@seo.fee.uva.nl) concerning economic data.

Questionnaire

0. General

Can we have your name?:

And your e-mail address, telephone and fax number?:

Which Agency do you work for?:

Which department?:

We are interested in the methods in use and also in the evolution of the applied techniques, changes in the recent past and relevant discussion of future changes. Mostly, you can simply answer the questions with yes or no, but you are also welcome to comment wherever you feel you want to share something with us.

I. Disclosure methods for table of counts or frequencies

1) Do you use 'sampling and weighting' (not publishing the weight) as a statistical disclosure limitation method?

2) Do you apply a threshold rule where a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number?
   If yes, how many respondents are required?

3) What method do you apply if the threshold is not achieved? Do you use
   - cell suppression?

- random rounding?
- controlled rounding?
- confidentiality edit?

4) Do you use other special rules that impose restrictions on the level of detail that can be provided in a table?
If yes, what special rules?

II. Tables of magnitude data

1) How do you identify sensitive cells? Do you use the
- (n,k) rule?
- p-percent rule?
- pq rule?
- other…

2) If a sensitive cell is identified do you restructure the table?

3) If a sensitive cell is identified do you apply suppression?

4) If a sensitive cell is identified do you try to get a "waiver" of the promise to protect sensitive cells?

III. Microdata

Please distinguish between demographic data (surveys/register data of persons and households) and economic data (surveys/register data of firms)

III.1 Demographic Data

Do you release microdata of the census?
If you release both microdata of the census and non-census data, please duplicate the section hereunder and answer the questions separately for census data and non-census data.

For public use of microdata or for microdata for research do you:
- include data from only a sample of the population?
- exclude obvious identifiers?
- limit geographic detail?
- limit the number of variables on the file?
- apply top or bottom-coding?
- add random noise (adding or multiplying by random numbers)?
- apply swapping or rank swapping (also called switching)?
- apply blank and impute (electing records at random, blanking out selected variables and imputing for them)?

- blur the data (aggregating across small groups of respondents and replacing one individual's reported value with the average)?
- other…

III.2 Economic Data

For public use of microdata or for microdata for research do you:

- include data from only a sample of the universe?
- exclude obvious identifiers?
- limit geographic detail?
- limit the number of variables on the file?
- apply top or bottom-coding?
- add random noise (adding or multiplying by random numbers)?
- apply swapping or rank swapping (also called switching)?
- apply blank and impute (electing records at random, blanking out selected variables and imputing for them)?
- blur the data (aggregating across small groups of respondents and replacing one individual's reported value with the average)?
- other…

IV. The evolution of disclosure limitation methods

1) What where the major changes in your policy in the past ten years, and why? Since when do you apply the combination of methods described above?

2) Have you ever experienced disclosure related problems?

3) Do you use specific software packages, such as:

- GHQUAR?
- USBCSUP?
- CONFID?
- ACSSuprs?
- τ-ARGUS?
- μ-ARGUS?
- other…

4) Do you or your institute have publications on this topic? Please note the title and authors.

5) Finally, can we contact you for further information?

Thank you for your contribution!