WP. 49
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vii): General statistical confidentiality issues

THE INSTITUT DE LA STATISTIQUE DU QUÉBEC'S APPROACH TO THE CONFIDENTIALITY OF MICRODATA FILES AND TABULAR DATA

Supporting Paper

Submitted by the Institut de la statistique du Québec, Canada[1]

[1] Prepared by Jimmy Baulne, Éric Gagnon and Lyne Des Groseilliers.

The Institut de la statistique du Québec's Approach to the Confidentiality of Microdata Files and Tabular Data

Jimmy Baulne,* Éric Gagnon* and Lyne Des Groseilliers*

* Institut de la statistique du Québec, Direction de la méthodologie, de la démographie et des enquêtes spéciales, 200 chemin Sainte-Foy, 3rd floor, Québec, Quebec, Canada G1R 5T4

Abstract: Under its act of incorporation, the Institut de la statistique du Québec (ISQ) is required to protect the confidentiality of the information it collects. Accordingly, the Institute has adopted policies enabling it to fulfil its confidentiality obligations. Two of these policies involve the confidentiality of statistical products disseminated by the Institute, i.e. microdata files and tabular data. The first part of this article deals with microdata files more specifically, as well as the statistical disclosure control (SDC) rules applied to them. The stringency of SDC rules can vary according to the environment in which the microdata files will be used. The second part of the article deals with tabular data, and outlines different aspects of the Institute's dissemination of tables. The ISQ's integrated approach is then discussed, highlighting the different components examined: SDC rules applying to demographic statistics, social surveys and business surveys.

1 Introduction

The Institut de la statistique du Québec (ISQ) is the official statistical agency of the Quebec government. Its mission is to provide reliable and objective statistical information on all aspects of Quebec society. To this end, it conducts several social surveys and business surveys every year. The Institute also prepares and updates a demographic report on Quebec, and is responsible for the register of demographic events (Registre des événements démographiques – RÉD). In keeping with its mission, the Institute must utilize the entire statistical potential of the information gathered through its different activities. The Institute lacks the internal resources to do this, however, and so it made a strategic decision to maximize the use of its statistical products by third parties. It must ensure that these data are used in accordance with its act of incorporation, which requires that the Institute preserve the confidentiality of the information it collects. It therefore adopted an approach for sharing its microdata files and tabular data, aimed at ensuring flexibility in making these products accessible while protecting their confidentiality.

Section 2 of this article describes the approach adopted for disseminating microdata files resulting from social surveys. Then section 3 describes the approach used for disseminating tabular data from social surveys and business surveys, and for disseminating tables produced from the RÉD.

2 Microdata files

2.1 Dissemination of microdata files

To ensure maximum productivity of the information gathered through its surveys, the Institute offers outside researchers access to different types of microdata files with variable analytical potential, while protecting the confidentiality of the data provided by respondents. In this section we will look at different types of microdata files produced from social surveys, as well as disclosure control measures applied to these files. Access to files produced from business surveys will not be addressed in this article. The nature of the information in such files calls for more stringent disclosure control measures than those set out here.

There are two methods of rendering microdata files available to researchers. The first is to request that respondents give their consent ahead of time for these data to be shared with researchers. The second method gives researchers access to microdata files when there is no prior consent. In this article we will consider only files shared using this latter method.

2.2 Access to microdata files without prior consent

The Institute can give researchers access to microdata files without prior consent for various reasons. Consent may not have been sought, for instance, because a researcher wants access to all respondents in the file, so that his or her results will be consistent with those of the Institute. Consent may have already been sought for researchers of one agency, but then researchers of another organization also request access to the file during a project. Prior consent clearly does not apply in this case, and another type of access must be suggested.

The Institute suggests different approaches when there is no prior consent, so as to make it possible to share microdata while protecting respondents' privacy. Different disclosure risk control measures are suggested for this purpose:

- statistical measures: statistical disclosure control (SDC);
- legal and administrative requirements;
- physical and computer security measures.

The combined application of these measures makes it possible to control the risk adequately. It is possible to vary the stringency of each measure, while still ensuring adequate risk control. For instance, if a decision is made to apply less stringent statistical measures, then legal and administrative requirements and physical and computer security measures should be more strict.

2.2.1 Public-use microdata files

The first type of access suggested when no prior consent exists is to give researchers access to a public-use microdata file (PUMF) at their place of work. Before

producing this kind of file, the variables in the file must be classified into three categories: direct identifiers, indirect identifiers and non-identifying variables. Direct identifiers are variables that can be used to identify a person directly, e.g. a name, address or telephone number. Indirect identifiers can be used to identify a person when they are cross-referenced, e.g. sex, age and profession. Finally, all other variables in the file are non-identifiers and consequently are excluded from SDC analysis. To create a PUMF, SDC rules are applied to direct and indirect identifiers in the file. As a first step, direct identifiers are removed, and then very strict SDC rules are applied to indirect identifiers. SDC rules are applied in two stages: risk identification followed by masking to minimize the risk. The following criteria are used to identify the risk:

- A region that can be distinguished in a file must have at least 80,000 inhabitants.
- Each cell obtained by combining the categories of three indirect identifiers must comprise at least 800 individuals in the population.
- One of the indirect identifiers in the combination must be the distinguishable region mentioned above.

The minimal criterion of 80,000 inhabitants in a region has been used in the past by Statistics Canada for creating PUMFs (Béland, 1999). Under certain circumstances, this criterion and the minimal number of individuals in a cell may be relaxed or tightened. In using varied thresholds it is important to take into account the sensitivity of information in a file and the type of population covered by the survey. Moreover, the use of lower thresholds must not substantially increase the risk.

For the second stage in applying SDC rules, the following masking techniques may be applied:

- global recoding of regional variables and indirect identifiers at risk;
- removal of an indirect identifier at risk from the file;
- removal of the indirect identifier at risk for certain respondents;
- top-coding and bottom-coding;
- rounding off or adding random noise.

Applying these SDC rules permits minimization of the disclosure risk and, consequently, relaxation of the other risk control measures. For instance, it is not necessary for researchers to apply SDC rules to tabular data produced from this type of file. When using such files, however, researchers must agree:

- to use the file for analysis and research purposes;
- not to combine the file with another file or attempt re-identification;
- not to make back-up copies of the file.

Researchers who do not comply with these requirements may be denied access to the PUMF.

2.2.2 Scientific-use microdata files (SUMFs) available at researchers' place of work

A second type of proposed access is to provide researchers with a SUMF at their workplace. This allows them to work on a file with greater analysis potential than that offered by a PUMF. To obtain access, however, researchers must sign an agreement with the Institute, agreeing to protect the confidentiality of the data provided. A SUMF is created by classifying the file variables into the same categories as in a PUMF; direct identifiers are removed and SDC rules applied to indirect identifiers are less stringent than those used in creating a PUMF. As with PUMFs, SDC rules are applied in two stages: risk identification and masking. Risk identification for SUMFs uses the following criteria:

- A region that can be distinguished in the file must have at least 10,000 inhabitants.
- Each cell produced by combining the categories of three indirect identifiers must comprise at least 100 individuals in the population.
- One of the indirect identifiers in the combination must be the distinguishable region mentioned above.

These risk identification criteria, inspired by the methods developed by Statistics Netherlands (Schulte Nordholt, 2001), are less stringent than those described for PUMFs. Moreover, under certain circumstances these criteria can be relaxed or tightened using the same conditions that apply to PUMFs.

In the second stage of applying SDC rules, masking is used for SUMFs just as it is used for PUMFs. However, the masking is less stringent than that applied to PUMFs, given the lower risk identified at the first stage.

Applying SDC rules to create SUMFs makes reduction of the disclosure risk possible, but does not eliminate it. Consequently, adequate control of this risk requires stricter legal, physical and computer measures than those used for PUMFs:

- Users may not transport the microdata.
- Paper copies must be kept in a secure location.
- Access to the copy of the original microdata file or its subproducts must be controlled and restricted to authorized individuals.
- The file must be kept in a secure location and encrypted.
- Once the project is finished, the copy of the original microdata file must be destroyed and a note confirming its destruction sent to the Institute.
- Researchers must apply SDC rules to the tables produced from the file. Details on these rules are given in section 3.3.1.
- Etc.

Researchers who fail to comply with these requirements may be denied access to the SUMF. The Institute may even take legal action against them.

2.2.3 Microdata files without direct identifiers but non-masked, available at CADRISQ

A third type of suggested access is to provide researchers with a microdata file, without direct identifiers but non-masked, on the premises of the Institute's research data consultation centre (Centre d'accès aux données de recherche de l'Institut de la statistique du Québec – CADRISQ). This approach may be preferable for researchers who are not satisfied with the analysis potential of SUMFs.

In these files, only direct identifiers are removed. No SDC rules are applied to indirect identifiers. Consequently, the disclosure risk for such files is considerable. To make up for the lack of SDC rules, however, legal, physical and computer measures relating to the use of such files are more stringent than those used for SUMFs:

- The microdata file remains on CADRISQ premises.
- Analyses are conducted under the supervision of the CADRISQ supervisor.
- Researchers are sworn to secrecy and subject to the same confidentiality obligations as ISQ employees.
- SDC rules must be applied by researchers to tabular data they wish to remove from the CADRISQ. Section 3.3.1 gives further details concerning these rules. The CADRISQ supervisor ensures that SDC rules have been applied adequately by the researcher, so as to safeguard the confidentiality of the tables.

2.2.4 SUMFs consultable by remote access

A fourth type of proposed access consists of giving researchers remote access to SUMFs, so that they can work on the files on their own premises, in terminal mode. Note that although researchers are conducting their analyses at work, the physical file remains at the Institute. The SDC rules applied to create such files are the same as those used to create SUMFs given to researchers. However, the other measures relating to the file's use are stricter. All operations by researchers are monitored remotely by an ISQ employee. This supervision is similar to supervision at the CADRISQ. In addition, researchers cannot download parts of the file to their computers, or print excerpts. Note that all tabular data that researchers wish to extract from this secure environment must be checked by an ISQ employee to ensure that there is no disclosure risk.

The Institute is considering relaxing SDC rules applied to remotely accessible microdata files, since the computer security measures used are stricter than those applied when researchers have access to SUMFs. In future, the Institute would like to make files with less masking than SUMFs available by remote access.

3 Tabular data

3.1 Dissemination of tabular data

The approach described in the paragraphs below concerns the dissemination of tabular data produced from a microdata file belonging to the Institute. This data may be disseminated by non-ISQ researchers using a file from the Institute, but also by ISQ employees when publishing survey findings.

Unlike the approach used when disseminating microdata, which concerns only social survey data, the approach for disseminating tabular data concerns both social and business survey data.

In the section outlining the approach for the dissemination of microdata files, we made a distinction between different types of files that can be made accessible to users. This distinction has a direct impact on the SDC rules applied to tabular data. That is why the approach concerning the dissemination of tables takes into account the type of file used to produce the table (i.e. whether it is a non-masked file or SUMF). Once again, SDC rules applied to tables also depend on the type of user who disseminates the table (i.e. an ISQ employee or a non-ISQ user).

Keeping all these distinctions in mind, we will now take a closer look at the ISQ approach concerning SDC rules applicable to the dissemination of tables.

3.2 Developing a policy

Whether research is being done for its own publications or for a publication by a researcher using one of its microdata files, the Institute must provide users of its files with a procedure that lays out rules to be followed. The purpose of these rules is to ensure the confidentiality of the information disseminated. Furthermore, if a researcher uses an Institute file, failure to comply with this procedure could result in legal action against the individual and his or her employer.

Since the Institute is obliged to protect the confidentiality of information published, it has developed a policy setting out guidelines governing the confidentiality of tabular data of survey results for dissemination.

This policy covers different types of tables: frequency count tables, tables of magnitude – mean, total or ratio –, percentile and model analysis results (regression). In addition, tables may be produced from social or business survey files or from the RÉD, co-owned by the Institute and the Ministère de la Santé et des services sociaux du Québec (MSSS).

Thus there are a variety of circumstances that the Institute must take into account when developing its policy on the dissemination of tabular data. For example, who wants to disseminate the table: an ISQ employee or an external researcher? What kind of file (non-masked or SUMF) was used to produce the table? What kind of data

(social, business or demographic) does the table contain? All these considerations help determine the choice of SDC rules to be applied to tables. The Institute has had to come up with policies, each complemented by a separate procedure, to ensure that it can respect its commitment to confidentiality in all situations involving the dissemination of tabular data.

3.3 Organization of guidelines

Guidelines on the confidentiality of tabular data for dissemination have been split into guidelines for social surveys, business surveys, and demographic statistics. Each section has a procedure for every different situation involving the dissemination of tabular data.

3.3.1 Social surveys component

The social surveys component has three procedures. The first deals with tables produced by non-ISQ users, using non-masked files. Such files can be made available to researchers either at the CADRISQ, or on the premises of the researcher's public organization, if respondents have given their prior consent. Since no SDC rules have been applied to the indirect identifiers in microdata files of this type (see section 2.2.3), there is a very high disclosure risk and strict SDC rules are applied to the tabular data.

For this procedure, a table represents a disclosure risk if there is not a minimum number of respondents in each of the cells of the table, or if there are zero cells or full cells. A cell is full when it contains all the respondents; a zero cell contains no respondents. The masking techniques applied to tables considered at risk depend on the variables they contain. Indeed, this procedure uses two important concepts: the presence of a variable related to ethnicity and the size of the geographic classification. Distinguishing tables on the basis of ethnicity is justified by the fact that this is a very sensitive concept in Quebec, and sub-populations formed by different cultural communities are relatively small, with consequently higher risk of identification. The same observation applies to the sub-populations defined by certain geographic territories, which makes such tables highly specific.

Accordingly, SDC rules are stricter when there is a variable linked to ethnicity and when the geographic classification is small. Among the masking techniques used in this procedure are:

- table redesign;
- local suppression of data (including secondary cell suppression);
- limiting the number of cross-referenced variables used in a table;
- prohibiting the regional dissemination of tables (in certain cases).

The second procedure concerns tabular data produced from SUMFs used by non-ISQ researchers. The disclosure risk associated with these tables is less than for tables

produced from non-masked files, for SDC rules have been applied to the microdata. Thus less stringent SDC rules can be applied to tables. This procedure uses the same concepts as the first (i.e. the presence of a variable linked to ethnicity and the size of the geographic classification). Disclosure risks are also identified in the same way, and only some of the masking techniques are used to reduce this risk. This is more or less what distinguishes the first two procedures, and this distinction comes from the fact that SDC rules are applied to the microdata of a SUMF, but not to the microdata in a non-masked file. For example, a table may be disseminated regionally if it is produced from a SUMF file and complies with the rules of the second procedure, while producing such a table from a non-masked file (first procedure) may not be allowed.

The third procedure in the social surveys component concerns tables produced by ISQ employees. Non-masked microdata files are used to produce such tables, of course. The presence of a variable linked to ethnicity is once again an important concept for determining the SDC rules to be applied to the tables. However, the second concept used in this procedure is the presence or absence of a delicate variable in the table. A variable is considered delicate if it contains information relating to the respondent's private life, which is not generally known and the respondent does not wish to disclose, such as sexual behaviour or the cause of a disability. Tabular data must therefore be classified into one of the following four categories:

- table containing a delicate variable cross-referenced with a variable linked to ethnicity;
- table containing a delicate variable not cross-referenced with a variable linked to ethnicity;
- table containing a non-delicate variable cross-referenced with a variable linked to ethnicity;
- table containing a non-delicate variable not cross-referenced with a variable linked to ethnicity.

The status assigned to variables (i.e. whether they are delicate or not) is up to the survey project leader, to be approved by his or her manager. This strategy makes it possible to relax the SDC rules applied to tabular data in certain cases. The tables in the fourth category are an example. The disclosure risk identification methods are the same as for the other two procedures. Once again, the strictness of the rules depends on the classification of the table. Tables containing delicate variables or variables linked to ethnicity will be subject to stricter SDC rules, whereas the other tables will be subject to less strict measures, in particular allowing low-frequency cells in the tables. The masking techniques used in this procedure are combining categories and local suppression of data deemed confidential, including secondary cell suppression.

3.3.2 Business surveys component

The business surveys component includes just one procedure, involving tabular data produced by ISQ employees. Like their counterparts in the social surveys component, these tables are produced using non-masked files, meaning that no SDC rules have been applied to the indirect identifiers in microdata files of this type.

The procedure for this component uses a concept equivalent to the third procedure in the social surveys component (i.e. delicate variables, but adapted to the business context). Tables in this component are categorized according to whether they contain a strategic variable or a non-strategic variable. Any information likely to give a business a competitive advantage may be considered a strategic variable.

The SDC rules applied to tables containing a strategic variable are stricter than those applied to other tables. In the only procedure in this component, disclosure risk is identified by the absence of a minimum number of respondents in each cell or the presence of zero or full cells and, for table of magnitude data, a sensitivity measure such as the dominance rule (n,k) or the p-percent rule (Willenborg, 2001). The following masking techniques are used to limit this risk:

- local suppression of data (including secondary cell suppression);
- table redesign;
- adding random noise;
- controlled or random rounding.

Just as for the social surveys component, the choice of strategic and non-strategic variables is up to the survey project leader, subject to approval by his or her manager. For the business surveys component, however, a committee consisting of ISQ employees was struck specially to draw up a list of variables, grouped into themes, that must be considered strategic. Employees wishing to disseminate tabular data are required to use this list to determine the status of variables.

3.3.3 Demographic statistics component

The Institute disseminates data based on birth, marriage, death and stillbirth records, in its own name and as an agent of the MSSS.

The demographic statistics component is particular in that the publication produced from the RÉD consists of statutory tables to which are added a limited number of individual requests. Unlike survey files that may cover a range of subjects, and hence a range of variables, the RÉD focuses on a set number of indicators that are used to produce a recurring series of tables every year.

Unlike the procedures for the other components, based more on obtaining a minimum number of units in each table cell, procedures for demographic statistics use characteristics of the variables in the table. Since the variables are the same from year to year, they are weighted and the risk is identified on the basis of this

weighting. Of course the weight of the variables and the threshold values used in decision making are included in confidential documents.

4 Conclusion

As the Quebec government's official statistical agency, the Institute has an obligation to protect the confidentiality of the information it releases. The approaches described in this article allow it to fulfil its obligations while offering researchers access to data with satisfactory analytical potential.

References

Béland, Y. (1999). "Release of Public Use Microdata Files for NPHS? Mission… Partially Accomplished!" *Proceedings of the Survey Research Methods Section.* American Statistical Association, p. 404-409.

Schulte Nordholt, E. (2001). "Statistical Disclosure Control (SDC) in Practice. Some Examples in Official Statistics of Statistics Netherlands." Paper presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, The former Yugoslav Republic of Macedonia.

Willenborg, L. and T. De Waal (2001). *Elements of Statistical Disclosure Control*. *Lecture Notes in Statistics 155*. New York: Springer-Verlag.