Catalogue no. 11-522-XIE

Statistics Canada International
Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada’s Symposium 2005
Methodological Challenges for Future Information Needs

THE CONFIDENTIALITY OF VITAL STATISTICS TABLES: THE ISQ’S APPROACH

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

ABSTRACT

One of the missions of the Institut de la statistique du Québec is to collect and compile data on vital events, particularly births, marriages, deaths and stillbirths. Under its establishing legislation, the ISQ is required to safeguard the confidentiality of the information it collects. To that end, it has a policy on the confidentiality of data tables intended for release. One section of that policy deals with the confidentiality of vital statistics tables. Following a brief review of the various steps in the table dissemination process, we will describe the statistical disclosure control methodology used in the case of vital statistics tables.

KEYWORDS: Statistical disclosure control, vital event, data table

1. INTRODUCTION

The Institut de la statistique du Québec is the Quebec government’s official statistics agency. Its mission is to provide reliable, objective statistical information about all aspects of Quebec society. To that end, the ISQ conducts a number of social, household and business surveys each year. It also creates and maintains Quebec’s record of vital events. In particular, the ISQ is legally required to have a system that collects demographic data such as births, marriages and deaths. Those statistics provide a clearer understanding of the composition, trend and rate of population growth. When this information is broken down by age and gender, it can be used as a guide by people who plan, carry out and evaluate a wide variety of public health, medical research and economic and social development activities.

The ISQ’s mandate is to produce and disseminate provincial statistics on vital events. It is also required to respond to special requests from various quarters, such the public sector, the government and the research community.

The ISQ must comply with its establishing legislation, which requires it to safeguard the confidentiality of the information it collects. Hence, in disseminating its data tables, the ISQ has taken an approach aimed at maximizing the use of its statistical products while scrupulously maintaining confidentiality.

Section 2 of this article covers the general confidentiality protection concepts that the ISQ follows in disseminating data tables from social, household and business surveys. It also presents the concepts used by the ISQ to ensure the confidentiality of data tables from the *Registre des événements démographiques* [registry of vital events]; that approach is described in section 3. That section also contains a description of the characteristics of the vital events data and two brief comparative studies.

2. DISSEMINATION OF SURVEY DATA TABLES

To ensure the widest possible access to its products while safeguarding confidentiality, the ISQ has adopted a policy that sets out a procedure for ensuring the confidentiality of data tables intended for release. The procedure is different depending on whether the data are vital statistics, social or household survey data, or business survey data.

[1] Lyne Des Groseilliers, Jimmy Baulne and Éric Gagnon, Institut de la statistique du Québec, Direction de la méthodologie, de la démographie et des enquêtes spéciales, 200 Sainte-Foy, 3rd Floor, Québec City, Quebec, Canada, G1R 5T4.

It involves the application of statistical disclosure control (SDC) rules in two steps. The first step involves identifying the risk of disclosure, and the second consists in masking the data to reduce the risk.

This section provides an overview of the SDC rules that the ISQ uses for tables based on a non-masked microdata file, in the case of business surveys and social or household surveys. These concepts will serve as a lead-in to the SDC rules developed for vital events.

2.1 SDC rules for social or household surveys

In the case of tables of social or household survey data, the risk of disclosure can be associated with the small number of respondents in each cell and the presence of an ethnicity variable or a "sensitive" variable in the table (Beaulne *et al*., 2005). A sensitive variable is a variable that contains information about an individual's private life, which is not usually known and which the individual does not wish to disclose to just anyone. One's sexual conduct, the reason for a disability, one's country of birth and one's religion are examples of sensitive information. Variables are designated as sensitive by the survey's project leader, with the survey manager's approval. Tables containing cells with a small number of respondents, sensitive variables or ethnicity variables may be subjected to more stringent SDC rules than other tables.

Among the masking techniques used to reduce the risk of disclosure are the following:

- aggregation of categories,
- local data suppression (including secondary cell suppression).

2.1 SDC rules for business surveys

For tables of business survey data, the risk of disclosure can be due to the fact that there is a small number of respondents in each cell. In addition, in the case of tables of magnitude data, the risk is associated with the fact that a small number of enterprises makes up a large proportion of the estimate. Sensitivity measures such as the (n, k) dominance rule and the p-percent rule are used to compute these monopoly cases (Baulne *et al*., 2005; Willenborg, 2001). An adaptation of the "sensitive variable" concept is also used. The variables in a table are identified as strategic or non-strategic. In an effort to improve their performance, businesses are always looking for information that might give them an edge over their competitors. Any information that provides such an advantage is regarded as strategic, like any information from a company's strategic plan. Variables are designated as strategic by the survey's project leader, with the survey manager's approval. However, a list of variables that are automatically strategic has been developed to simplify the task and encourage consistency within the ISQ.

Tables containing cells with a small number of respondents, strategic variables or, in the case of tables of magnitude data, a small number of companies making up a large proportion of the estimate, may be subjected to more stringent SDC rules than other tables. Among the masking techniques used to reduce the risk of disclosure are the following:

- local data suppression (including secondary cell suppression),
- aggregation of categories,
- addition of random noise;
- controlled or random rounding.

To summarize, SDC rules for social, household or business surveys can vary with the number of respondents per cell, the type of data (proportions or magnitudes) and the nature of the data (sensitive, strategic or ethnicity variable). In addition, a thorough understanding of the survey is needed to choose the right sensitive or strategic variables.

3. DISSEMINATION OF VITAL STATISTICS TABLES

As noted above, the ISQ is responsible for collecting and compiling data on vital events such as births, marriages, deaths and stillbirths. In addition, the ISQ and the MSSS are co-owners of the *Registre des événements* *démographiques*. This registry contains the annual files of births, deaths, stillbirths, marriages and civil unions and hence can be used to compute the actual number of events in a given period.

3.1 SDC rules for vital events

To fulfil its mandate to produce and disseminate statistics on vital events, the ISQ had to develop a special confidentiality procedure for this type of data.

Unlike survey files, which can cover a wide range of subjects and large numbers of variables, the *Registre des événements démographiques* focuses on a fixed number of indicators used to produce a recurring set of tables each year. The data for births, marriages and deaths are quite similar: name, residence address, date and place of the event, age, marital status, language and sex. For births, the child's weight, the length of the pregnancy and the type of birth are recorded as well. For deaths, the medical cause of death is included. Though not comprehensive, this list shows the repetitive nature of the data for each type of event and the small number of variables for each type of event (about 30).

Taking advantage of these characteristics of the registry, we analyzed each variable in terms of its risk of disclosure. Three attributes were used:

- degree of identification,
- sensitivity,
- level of detail.

The degree of identification relates to how rarely an event occurs; the more unusual a characteristic is, the greater the chance of identification. For example, a birth rank of 15 can be very revealing. The concept of sensitivity is similar to that of the sensitive variable. It refers to information about an individual's private life which is not usually known and which the individual does not wish to disclose to just anyone, such as country of birth or cause of death. The level of detail affects almost all of the variables in the registry, since it is possible to provide a less detailed breakdown for a variable to reduce the risk of disclosure. For example, age groups are less revealing than single years of age. This attribute is particularly important for data produced for various geographies.

On the basis of these criteria, a list of all the variables in the *Registre des événements démographiques* was prepared in descending order of disclosure risk. This ranking is maintained by assigning a weight to each variable, with the highest weight going to the variable with the greatest risk. The weight assigned to a variable that appears in the births, deaths and marriages files is the same and remains constant.

For a number of variables in the registry, preset aggregations of categories were proposed, and a weight was assigned to them. The weight was related to the level of detail selected: the more detailed the breakdown, the higher the weight.

Since small area data are in high demand, various regional breakdowns were considered. A weighting system was adopted to accommodate tables produced for health regions, regional county municipalities and various municipality sizes. As a result, the regional breakdown used in a table is considered a variable which is assigned a weight based on the level of detail.

Note that no weights were assigned to individuals' names, addresses and complete birthdates,[2] since that information can never be published. However, the location of an event and the date on which it occurred were kept on the list of variables analyzed.

The disclosure risk associated with data tables from the *Registre des événements démographiques* is identified in one step:

- add up the weights associated with all the variables in the table;

[2] Not to be confused with the date of the event in the births file.

- if the sum of the weights is less than or equal to a preset threshold, the table is deemed not to involve any risk of disclosure and can be published; otherwise, the table presents a risk of disclosure and cannot be published as is.

The threshold was set so as to make the data as useful as possible while limiting the risk of disclosure. The desire to produce tables that would be comparable to the vital statistics tables published in an annual report by the other provinces and the central government was also a factor in choosing the threshold. The SDC rules applied to survey data were also considered.

When the sum of a table's weights exceeds the threshold, there are two options for reducing the risk of disclosure:

- either collapse some categories of one or more variables used in the table, which will lower the individual weights, and recalculate the sum of the table's weights;
- or round each of the table's values to the nearest multiple of 5 on a probabilistic basis (this option should be used with caution).

It is important to note that the risk identification procedure is not subject to the minimum cell size rule, as is the case for survey data. This difference is due to the fact that the registry contains a small number of variables and that the number of allowable cross-tabulations is limited by the weight threshold. Hence, it is possible to find unique cases in published tables, but it should be noted that the risk that those cases will be identified is lower because of the higher weights associated with rare categories, sensitive variables and regional breakdowns.

The ISQ adopted this approach in 2003. Since then, it has been used to gauge the disclosure risk for all statutory tables based on data from the *Registre des événements démographiques*. About 76% of the 160 statutory tables published each year were considered to involve no risk of disclosure. Tables with an excessively high risk of disclosure are no longer published.

3.2 Two comparative studies

A study was conducted to compare the SDC rules for vital statistics tables with the SDC rules for the ISQ's social and household surveys. The baseline for comparison was the statutory tables produced by the ISQ before 2003, *i.e.*, before the current approach was adopted. In this particular context, the study showed that the SDC rules for vital statistics tables were slightly more stringent than the ones used for social and household surveys. Specifically, about half of the statutory tables that the SDC rules for vital events identified as involving a risk of disclosure (24% of the tables) were publishable under the survey SDC rules. The results reveal the difference between the two approaches: the first approach applies to data from the entire population, while the second applies to data collected from a probabilistic sample, for which the risk of disclosure is lower, from the standpoint of a statistics agency, because the sample was selected at random.

A second comparison was carried out to determine whether the disclosure control applied to tables based on data from Quebec's registry was similar to the disclosure control used in the annual vital statistics reports published by Statistics Canada and several Canadian provinces. Of the published tables that were studied, a number of the cause-of-death tables involved a risk of disclosure under the ISQ's approach. Hence, the ISQ's SDC rules for vital statistics tables differ slightly from those of other agencies.

3.3 Computerized approach

One advantage of the approach proposed by the ISQ is simplicity. An Excel spreadsheet was created, and the variable weights and the weight threshold were entered in it. To determine whether a table involves a risk of disclosure, the user simply selects the variables that make up the table.

This tool is particularly useful for responding quickly to special requests from various quarters, such as the public sector, the government or the research community.

4. CONCLUSION

The approach developed by the ISQ provides disclosure control for tables of vital events data. The MSSS, co-owner of the *Registre des événements démographiques*, also adopted the approach in 2003 for the publication of its data tables. The approach makes it possible to release vital events data for use in planning and research in the areas of public health and social development, while minimizing the risk of information disclosure.

ACKNOWLEDGEMENTS

The authors wish to thank Paul Berthiaume, manager of the *Registre des événements démographiques*, and Normand Thibault for their invaluable assistance and Line Beauchesne for her constructive comments.

REFERENCES

Baulne, J., É. Gagnon and L. Des Groseilliers (2005), "L'approche de l'ISQ pour assurer la confidentialité des fichiers de microdonnées et des tableaux de résultats", Proceeding of the *Quatrième Colloque francophone sur les sondages,* to appear.

British Columbia (2003), "Annual Report – Selected Vital Statistics and Health Status Indicators", British Columbia Vital Statistics Agency.

Nova Scotia (2003), "Annual Report – Vital Statistics", Service Nova Scotia and Municipal Relations.

Statistics Canada (1999), "Causes of Death – Shelf Tables", Catalogue No. 84F0208XPB.

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