

11-617

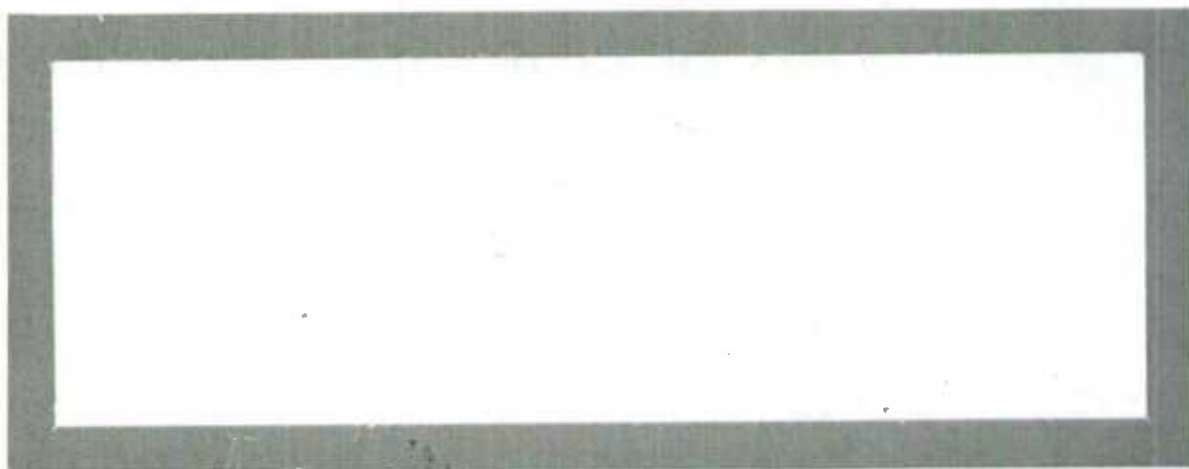
no. 96-05

c. 2



Statistics
Canada

Statistique
Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes
entreprises

Canada

WORKING PAPER NO. BSMD-96-005E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-96-005E

DIRECTION DE LA MÉTHODOLOGIE

**DISCLOSURE CONTROL ISSUES AT
STATISTICS CANADA**

by

**Douglas Yeo and Dale Robertson
January 1995**



Disclosure Control Issues at Statistics Canada

DOUGLAS YEO and DALE ROBERTSON¹

Jan. 5, 1995

Résumé

Le contrôle de la divulgation est une composante importante dans la production de produits de données à Statistique Canada. Les risques de divulgation sont en hausse continue, principalement parce que les requêtes des utilisateurs pour l'information plus détaillée augmentent, mais aussi parce qu'il y a une plus grande demande d'accès à des données disponibles électroniquement. En réponse à ce risque, l'Agence revoit présentement les procédures de contrôle de la divulgation qui sont actuellement en place.

Un des premiers aspects abordé lors de cette revue fut la conduite d'un sondage informel auprès des divisions qui diffusent des données. Des discussions ont été tenues avec ces divisions à propos de la gamme de méthodes employées actuellement. Les discussions portaient sur des points où les divisions se sentaient vulnérables ou désireraient un plus grand soutien soit au niveau des normes, des méthodes ou des logiciels. Le besoin d'utilisation de paramètres standardisés avec chacune de ces méthodes a été également abordé.

Les méthodes de contrôle de la divulgation actuellement en vigueur pour les macrodonnées incluent la suppression de cellules ainsi que plusieurs techniques d'arrondissement. Parmi les problèmes à envisager, on retrouve les difficultés inhérentes au contrôle de la divulgation dans le cas de requêtes ad hoc, la sur-suppression due au fait que les règles de divulgation sont inutilement strictes, sans oublier les problèmes de diffusion des microdonnées, surtout pour les enquêtes longitudinales. Plusieurs suggestions ont été formulées. Ainsi, on a suggéré que Statistique Canada consacre plus de ressources dans des outils généralisés. On a par exemple mentionné des extensions au logiciel de suppression de cellules connu sous l'acronyme de CONFID, la coopération avec d'autres agences statistiques, une meilleure communication à l'intérieur de l'Agence, une plus grande utilisation des techniques d'arrondissement aléatoire, des normes établies pour les méthodes et paramètres de contrôle de la divulgation, des règles différentes pour des données de sensibilité différente, ainsi que bon nombre d'autres recommandations.

MOTS CLÉS : Suppression de cellules ; Contrôle de la divulgation ; Publication des microdonnées ; Arrondissement.

Abstract

Disclosure control is an important component in the production of data products at Statistics Canada. The risk to the data is on the increase, due to the rising level and detail of user requests, and the demand for more data to be made available electronically. In response to this risk, the Agency is currently reviewing the disclosure control procedures currently in place.

One of the first aspects of this review to be undertaken was an informal survey of the data-disseminating divisions. Discussions were held with these divisions about the range of methods currently being used. The talks focused on areas where the divisions felt vulnerable or would welcome greater support through standards, methods, or software. The need for standard parameters to be used in these methods was also discussed.

Disclosure control methods now in use for macro-data include cell suppression and several rounding techniques. Issues of concern include the difficulty in controlling disclosure in ad hoc requests, oversuppression due to unnecessarily strict rules, and problems with micro-data dissemination, especially for longitudinal surveys. Several suggestions were made: that Statistics Canada put more resources into generalized software, such as extensions to the cell suppression software currently in use (CONFID); cooperation with other statistical agencies; better communication within the Agency; greater use of rounding; standards for disclosure control methods and parameters; different rules for data with varying degrees of sensitivity; and other recommendations.

KEY WORDS: Cell suppression; Disclosure control; Micro-data release; Rounding.

¹ D. Yeo, Business Survey Methods Division, D. Robertson, System Development Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Introduction

Due to the increasing concern over the risk of disclosure of confidential data, coupled with the constant increase in user demand for more detailed data, Statistics Canada has undertaken a study of disclosure control issues. Informal discussions were held with survey managers from most of the data-disseminating divisions, covering the range of methods currently being used. The talks focused on areas where the divisions felt vulnerable. Information was also gathered about areas of standards, methods, or software where survey managers would welcome greater support. The need for standards in the choice of methods and in the parameters used in these methods was also discussed.

Typically, each discussion, lasting from an hour to an hour and a half, was held with a small group of managers of a particular survey area. It was stressed from the outset that information collected was not to be used as an audit of that *specific* survey area, but to gather data over a *range* of Statistics Canada surveys. These data were meant as an aid to evaluating the current state of disclosure control at the agency, and to pinpoint areas of concern and potential future activity. In general, survey managers felt that this was an important topic, and welcomed the opportunity to discuss some of their concerns.

Methods

Magnitude Data

The use of cell suppression (please see the Glossary) is very common, especially for the magnitude data published by economic surveys. The methods in use typically test for dominance and place a lower bound on cell size. The generalized system most often used is CONFID. (See "Software", below). Processes in place range in size and automation from small and manual to large and fully automated. In the past, suppression rules were determined by an intuitive feel for the dangers of disclosure, and the appropriate level of protection needed. More recent research has led to a class of cell sensitivity measures. A key feature of these measures is that they are quantitative, giving a measure of the amount of sensitivity of particular cells. The rules are adjustable in strictness and may be designed to give a desired degree of protection. The desired amount of protection or relative uncertainty given to individuals is the starting point for setting up a rule.

Random rounding is also used for magnitude data in some cases, especially for multi-dimensional tables. (One survey area noted its successful use for eight-dimensional tables.) Rounding with perturbation of estimated totals is also done. Microaggregation is used by at least one area for micro-data.

Frequency Data

Frequency data, most commonly produced by household surveys, is generally well protected by rounding techniques. Random rounding is the most common method, as opposed to controlled rounding, because of its relative simplicity. In some instances, rounding is used with suppression or perturbation. For micro-data, data swapping and perturbation are typical methods, although they can have a serious impact on the usefulness of the data.

Other Methods and Issues

Some surveys use processes that are more ad hoc in nature. Some use data analysis packages to try to track down unique records. Other areas rely on "latent suppression", where data are only suppressed if respondents express a desire to reduce the level of publication detail in their cells. This method is only used in areas where the confidentiality of the data is perceived to be of lesser (or no) concern.

It was noted during several interviews that more research is needed into useful methods of disclosure control, for both new and ongoing surveys. It was generally felt that Methodology had a role to play in developing disclosure control techniques at Statistics Canada.

The Process

Pre-planned Versus Ad Hoc Data Releases

The biggest disclosure control issue facing Statistics Canada today is the need to employ methods to handle ad hoc requests. Most survey areas have stable processes in place to handle pre-planned data releases. However, there is a rapidly increasing demand for detailed ad hoc requests from data users. These requests place a heavy burden on the disclosure control and dissemination processes in place. As well, users naturally want their data demands to be met with as little suppression as possible.

The risk of residual disclosure is a particular concern. As more ad hoc requests are received, it becomes more difficult to provide data that are not heavily suppressed. Multiple requests may put the data at risk, whether or not deliberate attempts are being made to discover confidential data. The problem is exacerbated in repeated surveys.

If the data in question can be divided in several ways using different variables, the problem of multiple boundaries appears. The overlap between these different sets of boundaries can lead to disclosure. If a manual process is being used, survey managers are not confident that proper disclosure control is being achieved. As well, a manual process can be very expensive; often, most of the cost charged to the user is for the disclosure control.

Macro- Versus Micro-Data

Entirely different processes are needed for these two types of data. Although there are several well-grounded methods for macro-data, such as cell suppression, the situation for micro-data releases is completely different. Currently, the Micro-Data Release Committee is mandated to examine micro-data releases on a case-by-case basis to decide what can be published. Any manager proposing a micro-data release applies the general criteria to the data and makes a formal proposal to the Committee, which may order changes such as the deletion of variables. The Committee's decisions are binding and final.

One general problem to be faced is that certain types of micro-data have extremely high proportions of unique records. It was stated by some survey managers that they feel the confidentiality of micro-data cannot be guaranteed, especially in the face of demanding users who may be misusing the data.

In particular, it is clearly difficult to release detailed micro-data for hierarchical or longitudinal datasets in any form. It is felt by some survey areas that the combination of unique records, high levels of detail, and repeated publications may make disclosure control impossible.

In terms of methods used, some questions were raised about random rounding. For macro-data, the possibility arises that random-rounded data can be deciphered, especially in small cells. Another problem may occur when other organizations sign data-sharing agreements with Statistics Canada. Because of the confidential nature of the Agency's disclosure control *rules*, the sharing agency may have better access to the micro-data themselves than to the rules meant to protect the data. Finally, weights on micro-data files also may lead to disclosure of the identity of respondents.

Cell Suppression Rules

The rules used to determine cell sensitivity at Statistics Canada have been employed for quite a number of years. Since the time of their formulation, much has been learned, and computers have entered into widespread use. Statistics Canada's traditional cell suppression rules are a combination of various N-k rules, to be applied in different situations. The parameters are considered confidential, but the strength of the rules is similar to that of the rules used by other agencies. These rules have a few technical defects in their formulation, leading to behaviour that violates common sense in some situations. Several survey areas found that these rules yield too much suppression, and cause troublesome results due to discontinuities in their strength. They can also cause sporadic difficulties in automated systems. Originally two strengths of rules were proposed, for more and less sensitive statistics, but in practice only the more stringent rules are used. The adaptation of the strength of the protection mechanism to the sensitivity of the data is, however, a sound idea.

Although the identification of sensitive cells is handled well, the characterization of an ideal suppression pattern is difficult to formulate, much less achieve at acceptable cost. Choosing complementary suppressions is not an exact science. In the current version of CONFID, a refinement run is needed after the suppression pattern is constructed, in order to help minimize the number of complementary cells suppressed.

Currently, automated cell suppression is done for two-dimensional and, occasionally, three-dimensional tables. It is felt that more support is needed for higher-dimensional problems, especially when large amounts of data are involved. As well, a process is needed to handle a hierarchy of suppression, such as the processing of major groups first, followed by more detailed groups. Without an automated system to handle such problems, survey teams are forced to break up large problems, and examine multiple tables for residual disclosure, manually. To reduce the risk of residual disclosure, an automated cell suppression system should be able to keep track of patterns for multiple tables, including those produced because of ad hoc requests.

The Impact of Disclosure Control

Many survey managers feel that the disclosure control process in place has a significant impact on published data. Many believe that there is too much suppression in their data tables, especially for cost-recovery data, where a user may pay a large sum for heavily suppressed data. In particular, fine detail often suffers extreme suppression. It is felt that the fear of heavy suppression often limits requests for custom tabulations. In general, survey managers feel that disclosure control methods are too restrictive, and would like to see more data published.

Performance Statistics

The disclosure control process should be managed with the aid of performance and management statistics. For example, when one or more steps in an automated disclosure control system fail, diagnostics should be produced by the system to help pinpoint the problem. However, in practice, few statistics—if any—are produced, and the cost of the process is rarely recorded. CONFID has some built-in performance statistics available to users, but these appear to be rarely used. There is often little documentation, especially for the more manual processes. Disclosure control information in publications is usually minimal, although users rarely demand more details.

Software

There is currently a range of software being used at Statistics Canada, much of it custom-built. The only generalized software in use is CONFID.

Generalized Software

CONFID is a generalized cell suppression software package. It is used both for actual disclosure control and as an auditing tool, to test the outputs of other disclosure control processes. The current version of CONFID began as a research prototype and test bed. It has been found adequate for many practical problems, and as the only software available, it has become almost indispensable. However, because it originally was built for use by someone with a detailed knowledge of its internal workings, its usage is not easy for less-knowledgeable individuals. Since its algorithms involve linear programming and heuristics, its detailed operations are complex and difficult to explain. In addition, CONFID is now being used to handle very large problems, which sometimes exceed certain (artificial) limitations present on the Agency's mainframe computer, for which CONFID was originally developed. Today's more powerful (multiprocessor) workstations may be a more suitable environment for these types of problems.

Because CONFID is a prototype, decomposition or partitioning of large problems was not implemented. If these methods were to be used, very large problems could be treated consistently, at least in principle. This implementation would be enormously preferable to ad hoc manual decomposition, which often happens in practice. There has been much work in the general areas of mathematical programming and operations research on problems that have some similarity to those of cell suppression. It is time to reexamine the core algorithms and heuristics in the package in the light of developments.

The technique of suppression is powerful, but subtle. More documentation on the subject suitable for subject matter users needs to be prepared. The present CONFID program, together with support and documentation, remains usable for many surveys, but it is time for a rewrite with additional capability.

User Needs

Most survey managers believe that some form of generalized software is needed. It is too costly to develop custom-built software survey by survey; resources spent on CONFID will be recovered in the long run. Survey managers have several demands. First of all, fast turnaround is essential; users do not want to be burdened by a slow, batch-oriented approach. The software should be modular, with a simpler user interface and simple setup. Another common request was for the ability to keep a history of requests, perhaps as a

database of previous tables produced. This history should be kept for all requests, and by individual requester, allowing residual checks between tables. The software should also be extended to handle larger problems, and to make hierarchical suppression more automatic. As well, survey managers want better output control to tabulation, to avoid painstaking manual transfers. Finally, performance statistics should be extended.

It was suggested that Statistics Canada work with other agencies when developing these generalized tools, to take advantage of each agency's strengths.

Other Issues

Some survey managers pointed out that not all data can be handled well by generalized systems. One example given was commodity data. There may always be a need for some custom-built software.

Another issue is the choice of software platform. Workstations are powerful and are not constrained by proprietary software, but they often have hidden costs. Mainframe computing is often more expensive on paper, but there is often excess capacity, and this excess could be used more efficiently. In many survey areas there is also demand for a microcomputer platform, given their widespread use. Microcomputers could be used for small surveys, enabling a simple, cheap process with quick turnaround.

Guidelines

A Global Strategy

Many survey managers stated that a global strategy must be implemented. Statistics Canada needs to determine what *disclosure* means, and work needs to be done to achieve state-of-the-art knowledge of the subject. Better communication within the Agency should be fostered, perhaps through the use of a software toolkit, a document summarising techniques, or a database of methods. Any global strategy should allow for the occasionally contradictory aims of internal flexibility and alignment with other statistical agencies.

Many managers feel that this strategy should include the widespread use of random rounding. This technique could be extended to business surveys, alleviating the problems arising when cell suppression is used to handle multiple ad hoc user requests.

What is a Safe Level of Risk?

It must be stressed that there is no such thing as disclosure avoidance; there is only disclosure control. Any guidelines must recognize this fact and ensure consistency across the Agency.

A point often stressed was that any disclosure control guidelines should recognize the varying degrees of risk inherent in any dissemination process. For example, risk varies by *age* of data: the older the data, the less sensitive it is. However, in current practice, old data—even from bankrupt companies—cannot be disclosed. As well, sensitivity varies depending on the type of *variable* in question. One possibility is to declare some variables less sensitive than others. Perhaps some variables, such as import data, should not be treated as confidential. Cell counts are a special case; is a zero cell confidential? Can cell counts be released if they

are cross-classified against a more sensitive variable? Another complication arises when true cell counts are unclear because of cross-ownership.

The *source* of data also affects sensitivity. For example, data from administrative sources may be less sensitive than survey data. On the other hand, if the survey data are collected from a public company and are public knowledge, there is some conviction that they be declared non-confidential. One last factor involved in the degree of risk is the *sampling rate* of the data. The smaller the sampling fraction, the greater the degree of protection that should be afforded the data.

A Standard Suppression Rule

A replacement for the current cell suppression rules is under consideration: it is known as a C-times rule (or the p/q rule, or with a further assumption the p% rule). This rule is the simplest and most general of the linear sensitivity measures. It is designed to guard against a universal danger, namely one cell respondent estimating another, using only his own total plus generally available knowledge. This general knowledge is simply the identity of the cell respondents, together with rough estimates of their contributions (to within $\pm q\%$ say). The most vulnerable individual is the largest contributor, and the most dangerous estimate can be made by the second largest. Any uncertainty in the leading respondent's contribution comes from the inexactly known contributions of the third largest, fourth largest, etc. contributors. The second largest contributor simply subtracts his own value from the published total to get an upper bound for the amount due to the largest entity. If the tail end (from the third largest onwards) is not big enough then an unacceptably accurate estimate for the largest contribution is available. Thus the tail end must supply a contribution that is not negligible compared to the largest one. The C-times rule says that the tail should be at least $1/C$ times the value of the leading contributor for safety.

By making a few reasonable assumptions, one may relate the value of C to the relative ambiguity guaranteed to an individual response. If the value of the relative uncertainty after the cell total is known is $\pm p\%$ then $C = q/p$. C can be interpreted as the ratio describing the decrease in the relative uncertainty of the estimate of the largest contributor. This ratio should not be allowed to become too large.

If a certain ambiguity—a desired value of p%—is to be ensured, one chooses a C given by the above, using a reasonable value for q ($\pm 50\%$ say). Values of C in the range 5 to 10 are representative of the typical strengths used in statistical agencies. For example, $C = 5$ corresponds to $p=10\%$ (with $q=50\%$). Note that when $q=100\%$ is assumed, the rule is known simply as the p% rule.

The C-times rule is generally felt to be superior to 1-k or 2-k rules; experiments on real data support this view. This rule is in use elsewhere. At Statistics Canada, it is envisaged that a small number of strengths be proposed, for example, one strength for more sensitive statistics and a second one for less-sensitive statistics. In certain, less common, circumstances, it may be thought that there is some safety in numbers, some uncertainty about the actual cell respondents. Here, variations of the C-times rules may be blended so there is a continuous weakening of the rule as the number of respondents in the cell increases. This could be of value occasionally, but we do not suggest it as the first choice.

Other Standards

There are currently no standards for micro-data, and many survey managers displayed strong convictions that Agency standards must be set in place.

It was proposed that the disclosure control waiver process be standardised. Instead of the multitude of survey-specific processes currently in place, a single Agency-wide group should handle all waiver requests. This could be done at the profiling stage, to allow for greater efficiency.

Changes to data-sharing agreements were suggested, including the extension of these agreements to other government departments. A related suggestion was to allow users to use data files directly, under signed agreements providing for heavy fines if confidentiality is breached.

Another suggestion was to implement a standard auditing tool. This testing procedure could be performed on tables before publication. The auditing process could include a "data-crackers group", charged with searching for flaws in disclosure control processes.

Lastly, the importance of the awareness of disclosure control was stressed. Statistics Canada staff should be made more aware of all aspects of confidentiality, for example, the safe handling of questionnaires, and not just about disclosure control issues.

Conclusions

This informal survey of data-disseminating divisions proved to be a valuable exercise. The discussions held with survey managers provided a good overview of the range of disclosure control methods currently in use at Statistics Canada. Clearly these managers feel vulnerable in several areas, such as the release of ad hoc tables and of micro-data. In general, it is felt that disclosure control rules should be made more consistent, and also less conservative in many cases.

Survey managers would like to see more standards for the disclosure control process. Agency-wide communication and documentation need to be implemented. Terms like disclosure, disclosure control, and risk need to be defined. Specifically, the confidentiality rules and the parameters to be used should be standardized; improved generalized software would be an important step in this direction. Other areas in need of standardization include the disclosure control waiver process and the auditing of tables to be disseminated.

Glossary

Cell Suppression: A method of protecting sensitive cells in tables by concealing their exact values. Typically, an "X" appears in place of the value. In fact, this is usually equivalent to specifying a range of possible values for this entry. These ranges may be calculated from the table with "X"s present.

Complementary Suppression: (See cell suppression, primary suppression.) A table cell whose value is chosen to be concealed to help achieve the protection for a primary suppression.

Controlled Rounding: (See rounding.) In controlled rounding, the movements of the table values are correlated in such a manner that the resulting table still adds up correctly.

Microaggregation: Microaggregation is based on data modification, where individual data is not published; instead, small aggregated data, such as triads are published.

Perturbation: A cell value is perturbed by adding "noise" to its value. In random perturbation, this added noise is in the form of a random variable.

Primary Suppression: (See cell suppression.) A table cell whose value must not be revealed in order to protect the confidentiality of a respondent to the cell. Typically, a certain ambiguity must be given to this value. This may be expressed as a range of possible values that the cell could assume, consistent with the released table.

Random Rounding: (See rounding.) In random rounding, the value is moved up or down randomly. The probabilities of moving up (p , say) and down ($1-p$) are chosen so that the expected value of the rounded value is the same as the original value. The additive nature of the table is generally destroyed by this process. (See controlled rounding.)

Rounding: Applied to a table, the process of adjusting all table entries to be multiples of some integer called the base. In general, a value is moved up or down to the first (closest) multiple encountered, and no further. (See random rounding, controlled rounding.)

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA



1010224109

H 72354

C. 2

000