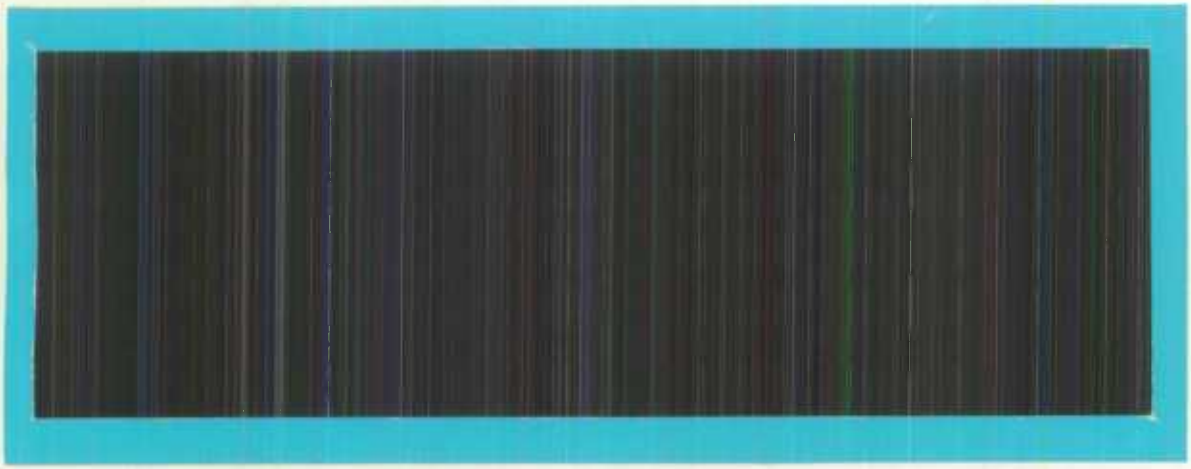
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER NO. SSMD-91-008 E

METHODOLOGY BRANCH

A CONFIDENTIALITY FRAMEWORK FOR MAGNITUDE DATA

SSMD-91-008 E

D. ROBERGE, R. KAUSHAL, A. DEMNATI

Social Survey Methods Division

Statistics Canada

May 1991


STATISTICS CANADA
STATISTIQUE CANADA
JUIL 9 1991
JULY
LIBRARY
BIBLIOTHÈQUE

A Confidentiality Framework for Magnitude Data

(D. Roberge, R. Kaushal, A. Demnati)

ABSTRACT

Data collected by Statistical Agencies are almost always collected under the seal of confidentiality. To ensure confidentiality, Statistical Agencies use various methods to avoid or limit the risk of statistical disclosure. Most of these methods have been developed without a rigorous definition of statistical disclosure. This explains the difference that can observed between the methods used by a Statistical Agency. We propose a confidentiality framework for magnitude data that allow for the development of a consistent methodology accross all products of a Statistical Agency. The cornerstone of the framework is a definition of statistical disclosure.

Résumé

Les données recueillies par les agences statistiques le sont presque toujours sous le sceau de la confidentialité. Pour assurer cette confidentialité, les agences statistiques utilisent différentes méthodes pour éviter ou limiter le risque de divulgation statistique. La plupart de ces méthodes ont été développées sans une définition rigoureuse de ce qu'est la divulgation statistique. Ceci explique en grande partie les différences substantielles qui peuvent être observées dans les méthodes utilisées par une agence statistique pour ces différents produits statistiques. Nous proposons un cadre de travail pour les données de magnitude qui permet l'élaboration d'une méthodologie cohérente pour l'ensemble des produits d'une agence statistique. La bien angulaire de ce cadre de travail est une définition de la divulgation statistique.

1. INTRODUCTION

Statistical Agencies are under increasing pressure from users to provide more statistics and in greater detail. Because confidentiality rules limit users' accessibility to statistics, they question the confidentiality rules used. It no longer suffices to proclaim that certain statistics are confidential, an explanation is warranted. This explanation must clearly state what statistical disclosure is and how

confidentiality rules operate. Users want this information to verify for themselves that confidentiality rules are formulated to maximize the value of the statistics released.

Within Statistical Agencies, on the other hand, confidentiality rules have evolved by ad hoc application without a stated definition of statistical disclosure. This manifests itself in different confidentiality rules for apparently similar applications. Also, confidentiality rules themselves are taken to be definitions of statistical disclosure. In such situations, it is difficult to justify these confidentiality rules to users.

Statistical Agencies have to maintain the difficult balance between the demands of protection for the respondents and open access for the users. Coherent and justifiable confidentiality rules, which are to the satisfaction of both the respondents and user communities, have to be adopted by Statistical Agencies. This should be based on an understandable definition of statistical disclosure. Confidentiality rules based on this definition, should be employed for all applications within the Agency. Of foremost importance in confidentiality rules is that it is clear what is to be protected and how the rules are doing this. Statistical agencies must have a clear concept of confidentiality and be able to communicate this to both the users and respondents.

While it is true that "without safeguarding the anonymity of respondents we would find that our data sources would dry up rapidly" [Fellegi 1972], it is also true that the perception of respondents of the confidentiality process is equally important. To secure their long-term cooperation, crucial to Statistical Agencies, public perception of disclosure must be based on an understanding of the disclosure limitation process.

This article provides a framework to build efficient confidentiality rules for magnitude data. Magnitude data are data that can be measured on a continuous scale, and for which interest centers on population characteristics such as the mean, total, or the ratio of two totals or means. Examples are revenue, expenditure and age. Efficient confidentiality rules control the risk of disclosure to an acceptable level while usefulness of the statistics to the users is maximized. The proposed framework contains three elements: a definition of statistical disclosure, the identification of threats, and the usefulness of the statistics to the users.

The cornerstone of this framework is a definition of statistical disclosure, a part too often overlooked as most researches have been on protection techniques. We choose a definition that does not include any reference to an intruder or to the usefulness of statistics. These two considerations are left as separate elements of the framework, thus allowing for a single definition across a Statistical Agency. The consistency establishes by the use of a single definition will reassure users and respondents of both the concern for and control over the confidentiality issue that the Statistical Agency has. This in turn will reinforce the Statistical Agency image as a competent and professional organization.

The emphasis is on protection methodologies based on a definition of statistical disclosure. In section 2, we discuss the definition of statistical disclosure and provide some examples. Next, disclosure is seen from the point of view of the intruder and threats posed by intruders are discussed (section 3). To complete the framework, in section 4, confidentiality is seen from the perspective of the user, who is interested in maximizing the usefulness of the data. Finally, in section 5, we illustrate the development of confidentiality rules using the three elements of the framework.

2. DEFINITION OF STATISTICAL DISCLOSURE

Several definitions of disclosure are present in literature. We consider the following three:

> ... if the population is sufficiently narrowly defined, it will contain only one identifiable respondent or at least, information can be deduced from the estimates that can be related to a particular identifiable respondent. [Fellegi 1972]

> If the release of the statistic S makes it possible to determine the value (of a characteristic) more accurately than is possible without access to S, a disclosure has taken place ... [Dalenius 1977]

> If sufficient accurate data are presented for correct identification of a respondent and a good approximation of confidential data, and if it is possible to correctly associate that data with the respondent, then statistical disclosure has occurred. [Cox and Sande 1979]

Protection of privacy of the respondent is the cornerstone of various Statistics Acts [Fellegi 1972] and all three definitions of confidentiality aim to protect the respondent. In his definition, Fellegi implies to protect respondents by hiding their identity. Hence, this definition suggests a methodology that focuses on the respondent. Dalenius presents another view by suggesting a methodology based on controlling the

level of access gained about the value of a characteristic. Duncan and Lambert [1986] discuss Dalenius' definition and suggest that a measure of disclosure would incorporate measures of knowledge, knowledge gain and relative knowledge gain by the intruder. Cox and Sande [1979] extend Fellegi's concept of respondent identity to include a good approximation of respondent's characteristics. Cox [1980] refers to disclosure in categorical and magnitude data arising from an "unacceptably narrow estimate" of a respondent's characteristics.

We adopt the Cox and Sande definition and focus on the concept of "unacceptably narrow estimates" or "good approximation" of magnitude data to define statistical disclosure. The primary reason for embracing this definition is that it is important that the intention of the definition be clear. The main consideration is that the respondent be protected. Limiting access is a consequence but not the primary objective of confidentiality. The purpose of Statistical Agencies is to provide as much statistical information as possible and the definition should not be limiting this basic function. While the ultimate intention of all definitions is the same, the Cox and Sande definition is appealing for its intuitive simplicity.

Another important reason for adopting the Cox and Sande definition is that it does not include any information about the intruder. While a good approximation is equivalent to knowledge gain under certain conditions, prior information that an intruder may have, must be modelled as required by the particular situation. By not incorporating the model for prior knowledge and statistics released, the definition becomes independent of the source of data and nature of statistics released. Thus, the definition is applicable to all possible situations and statistical products, and can be implemented throughout an Agency. Prior information and released statistics are discussed in more detail under the section on identification of threats.

Definition of Good Approximation:

We defined nmathematically the Cox and Sande concept of a good approximation as an interval around a respondent datum. For a datum, x, from a respondent, let this interval be defined by a lower bound L(x) and an upper bound U(x) as follows:

$$L(x) < x < U(x) \quad (1)$$

If it is possible to estimate from the released statistics that the value of x is within an interval L'(x) and U'(x) such that

$$L'(x) < L(x) < x < U(x) < U'(x) \quad (2)$$

then a good approximation is not obtained. An upper or lower estimate of x within the interval defined by (1) is a "good approximation" of x, resulting in statistical disclosure. The distance of x from each bound (upper and lower) gives the protection. In effect, this does not permit an accurate estimate of either the minimum or the maximum value of the data collected about a respondent. Confidentiality rules based on this definition must ensure that such an interval exists for all characteristics collected about a respondent and must do so for all respondents.

As stressed in Dalenius' definition the released statistics must contributes to the process of obtaining a good approximation for statistical disclosure to take place. Otherwise, if an intruder is already in possession of a good approximation through other means, then the released statistics are irrelevant to this process of approximation and confidentiality is not violated.

Good Approximation Intervals: Setting the Level of Protection

The bounds for the good approximation interval, U(x) and L(x), can be expressed as functions of the respondent datum. We present three examples on how to set the protection: relative, absolute, and a combination of the two.

Relative protection is a percentage of the datum (x) and can be written as follows:

$$(x - px) = L(x) < x < U(x) = (x + px) \quad (3)$$

where,

$$p > 0$$

This function provides protection relative to the datum. Therefore, the larger the value of the datum, the larger the good approximation interval. The drawback of relative protection is that respondents with small values of x are vulnerable because they are protected by small intervals. Based on this reason, absolute

protection should be offered. To provide a constant equivocation for all respondents, irrespective of the value of their data, the following function can be used:

$$(x - c) = L(x) < x < U(x) = (x + c) \tag{4}$$

where,

$$c > 0.$$

Intervals in (3) and (4) can be used in conjunction to define another good approximation interval as follows:

$$x - \max(|px|,c) = L(x) < x < U(x) = (x + \max(|px|,c)) \tag{5}$$

where,

$$c > 0 \text{ and } p > 0.$$

This combination of relative and constant protection ensures that small values are protected with a minimum constant equivocation while large values are protected with relative protection. This function is simple, justifiable and understandable. Therefore, it can be easily communicated to respondents and users. Functions other than the ones suggested can be used for the upper and lower bounds. The interval does not need to be symmetric around the datum. Yet, it is desirable that the above mentioned qualities be maintained.

Once the good approximation definition is accepted with the defining function, a complete definition is established. This definition binds all products of Statistical Agencies. The value of the various parameters of the defining function must be selected for Agency-wide application. The values of the parameters do not need to be the same for all the variables. Indeed, they should be chosen independently for each variable considering its nature and sensitivity. Respondents should be active participants of the protection process and consulted on protection parameters. Still, ultimately the decision becomes a value judgement. For an Agency to have a single definition, the parameter values must be independent of the source of the data. Differences caused by differing sources can be considered while determining the confidentiality rules. For example, a respondent providing income to the Census and to a sample survey should be given equal level of protection. Still, sampling does provide some ambiguity

that census data is not privileged to. The confidentiality rules, enforcing the definition should take into account the ambiguity bestowed on sampled data.

Once the definition and the level of protection are established, one can devise techniques and rules that will protect this interval. These rules would disallow an intruder from obtaining an estimate of a respondent's data that is within the above mentioned interval. To put a complete confidentiality methodology together, the various processes used to obtain estimates along with prior or public knowledge and released statistics have to be considered. In the next section, we discuss processes used by intruders, to obtain an approximation of a respondent's data.

3. IDENTIFICATION OF THREATS

The discussion so far has reflected on the definition of protection of respondent data. We now consider what happens when the data becomes a contribution toward a set of statistics. The question then becomes, is it possible to obtain a good approximation of any respondent datum from this set of statistics? If any set (or subset) of statistics allows a good approximation of a respondent datum, then they are sensitive and procedures to protect the respondent datum need to be carried out. In this section we shall consider threats posed by intruders and the resulting risk involved in releasing a set of statistics.

The threats in releasing a set of statistics come from intruders, who are attempting to obtain estimates of respondents' datum. Intruders have many processes at their disposal. To block all possible attempts by intruders, to obtain a good approximation, confidentiality rules would have to address all such processes. This, however, would result in the complete suppression of released statistics. This is not an acceptable solution to the problem of confidentiality, since the raison d'être of Statistical Agencies is to publish statistics in the interest of society. To respect this mandate and provide protection to respondents, only reasonable threats should be entertained. Therefore, we need to assess the risk of omitting a particular intruder process. The assessment of the threat is a subjective decision and related to the level of risk that a Statistical Agency is willing to take.

Processes available to intruders depend on the set of statistics released, prior knowledge possessed by the intruder and technological facility at their disposal. Intruders have different levels and kinds of prior information. It is not possible to know with certainty the prior knowledge and the technological

facility an intruder may have. Therefore, assumptions must be made about them. The risk taken by Statistical Agencies can be lowered by assuming that intruders have a large amount of prior information and extensive computing facilities. Making such a strong assumption is unreasonable for most situations and unduly limits the statistics available for release.

Let us illustrate this with an example. Is it reasonable to assume that an intruder has access to the data belonging to 25 respondents? The answer to this question depends on the situation. For population census data, only employees handling the forms and processing will have access to data from many respondents. Often Statistics Acts put legal constraints on its employees, which reduces such risk. Therefore, in such a case this assumption is unreasonable. A more reasonable assumption is that an intruder has access to data contributed by him and his spouse. In such a situation, the agency runs a high risk if from the released statistics, an intruder can use his and his spouse's data to obtain easily a good approximation of an identifiable respondent. Similarly, reasonable assumptions must take into account technological facilities, cost, feasibility and other options to obtain the same information without using the released statistics. For example, we should not consider threats requiring an intruder to have unlimited computer resources.

Once the assumptions about the knowledge and capabilities of the intruder are established, the various processes available to intruders need to be identified. Processes are the mechanisms by which the released statistics and the asumed prior knowledge are combined to obtain approximations of respondents' data. Now let us consider a more specific example, where the released statistics are the respondent count and the total investment income of a certain population. Assume that an intruder has access to information on his and his spouse's total investment income and can identify individuals in the population. Let us also assume the worst possible circumstance, that is, the intruder and his spouse are in the same population as the identified respondent and are contributing their income toward the total. Given this prior knowledge, what is the process that the intruder would go about trying to get an estimate of a respondent's income? A simple process that he can follow is that he can subtract his and his spouse's total investment income from the population total to obtain an upper bound estimate of the respondent.

The processes identified as having reasonable risk are used to build a protection methodology that relies on presenting obstacles to those processes. On the other hand, the remaining threats are the risk of disclosure that a Statistical Agency is taking. Therefore, the risk of disclosure is the risk that an

intruder will have greater prior knowledge than assumed and use these processes that the Statistical Agency did not specifically protect the data against, and that the use of these processes will lead to the disclosure of confidential data. This risk depends on the two factors. The first factor is difficult to measure. It is the probability that an intruder will have a certain knowledge and use a given process. This factor can be estimated either by modelling or by a survey measuring such a likelihood. The second factor is the number of good approximations of respondent data obtained by the use of a non-protected threat. An empirical study using the data to be protected, can measure this factor. However, given the difficulty in evaluating the first factor, judging the risk often becomes a subjective decision.

4. USEFULNESS OF STATISTICS

Equipped with a definition of statistical disclosure and having identified reasonable threats, we now are able to construct an effective protection methodology. Protection methodologies are based on setting constraints on the release of statistics to avoid a breach of respondent confidentiality. However, Statistical Agencies have a responsibility toward both respondents and users. Therefore, a protection methodology needs to be not only effective but also efficient. An efficient protection methodology maximizes the usefulness of the released statistics, while providing the required protection. In this section, we are going to discuss usefulness of the released statistics.

While the definition of statistical disclosure and identification of threats pertain to respondent protection, the usefulness of released statistics concerns the user. Therefore, the usefulness is best established through user consultations. The consultation should evolve from general discussions to a more focused form, where specific criteria about the usefulness of the released statistics can be addressed. The initial consultation should include a review of the major protection techniques used for magnitude data like suppression and perturbation. In the suppression technique all the protection is concentrated on a few statistics, whereas in perturbation, the uncertainty is dispersed throughout all statistics. Therefore, the utilization of the statistics influences the selection of protection techniques. More specific discussions with the users involve further decisions that need to be made depending on the technique selected. For example, in suppression, the choice of which set of statistics to suppress can be made from any of the following criteria: the set that minimizes the number of statistics suppressed, the set of statistics that represent the least number of respondents or the set with the smallest aggregate numeric value of the

statistic. Such criteria allow the protection process to be optimised from the user's point of view.

5. USING THE FRAMEWORK - AN ILLUSTRATION

The proposed framework has three elements: a definition of statistical disclosure, identification of threats and usefulness criteria. In this section we describe how these three elements are used in the development of a protection methodology.

We illustrate the framework outlined in this paper with a simple example. Consider aggregate by sources of income for small areas in Canada. Let us use the function given by equation (5) to define a good approximation. The first step is to decide the desired relative and absolute protection. Here consultations with both the respondents and users are recommended. Communication between respondents and users can demonstrate to the respondents that they also gain from the statistical exercise. Better statistical knowledge of the population allows the users to directly and indirectly help and serve the respondents better. Similarly, users will gain an appreciation of the respondents' concern of privacy, and that respondents' privacy must be protected if their cooperation is to be secured and maintained. Then and only then will it be possible for Statistical Agencies to strike a balance between the respondents' right to privacy and the increasing demand for more and more detailed statistical information.

For our illustration let us assume that for each income source a relative protection of 10% and an absolute protection equal to the median of the distribution for Canada is chosen. The purpose of the median, or any other function of variable distribution, is to have a definition that can be extended to several variables. To be succint, let us restrict ourselves to income from salaries, and assume a relative protection of 10% and that the median is $10,000. Therefore, statistical disclosure occurs if an approximation of a salary is within 10% or $10,000 of the respondent's data.

Having defined disclosure, let us identify threats. Though in a real application several threats may be identified, we in this illustration are going to consider only one threat, and assume that this threat is reasonable. We illustrate a process for an intruder who is interested in obtaining the salary of a particular respondent, the target. We assume that the intruder is also a respondent and that he belongs to the same small area as the target. For each small area, the statistic to be released is the total aggregated salary. The

process considered in this illustration is for the intruder to subtract his salary from the total to obtain an upper bound on the salary of the target. In equation form this becomes:

$$U'(x) = T(x) - C(x) \qquad (6)$$

where, T(x) is the total aggregated salary for a small area

U'(x) is the intruder's estimate of the upper bound of the target's contribution to T(x)

C(x) is the contribution of the intruder to T(x)

This process becomes a problem if it leads to a good approximation of the target respondent salary. Cox [1990] has shown theoretically that the closest approximation is obtained when the target respondent has the largest value and the intruder has the second largest. Therefore, the total aggregated salary is deemed sensitive if the second highest respondent can obtain a good approximation of the highest respondent. We shall refer to this statement as a sensitivity rule, and write it as:

$$U(x) + \text{Max}\,[10\%\ U(x),\ 10{,}000] < T(x)\text{-}C(x) \qquad (7)$$

This rule can then be used for all small areas to determine whether it is possible from the identified threat to obtain a good approximation. If at least one good approximation is obtainable, a protection methodology must be built to protect against the threat. The developer should be careful, when considering the various sensitivity rules found in the literature, as some rules may be based on a different definition of disclosure or addressed threats that may not be reasonable in his particular situation. Also, it may be impossible for some sensitivity rules to figure out the underlying definition or threat, as they are based on other rationale.

The next step is for the developer in collaboration with the users to identify the various protection methodologies worthy of consideration, and criteria to measure the usefulness of the statistics. For our illustration let us say that suppression is the preferred protection technique and that the usefulness criterium is to minimize the number of small area statistics suppressed. In our cases the small areas to be suppressed are directly identified by the sensitivity rule (7) and suppressing these statistics will protect the data against the identified threats. However, the developer needs to consider secondary threats that aim to get past the protection provided. When considering secondary threats the developer must assess

their reasonableness as was done for the primary threats. In our illustration, let assume that the small areas statistics are aggregated to a higher level, and that residual disclosure is possible through marginal totals and deemed a reasonable threat. Then complementary suppression should be carried out to protect against an intruder obtaining a good approximation of a respondent through the extended process of primary and secondary threats. Hence, the development of a protection methodology is often an iterative process that includes the identification of threats, of protection methodology and of usefulness criteria.

REFERENCES

Cox, Lawrence (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, 75, 377-385.

Cox, Lawrence and Sande Gordon (1979). Linear Sensitivity Measures in Statistical Disclosure Control", Proceedings of the 42nd session of the International Statistical Institute.

Dalenius T. (1977). Towards a Methodology for Disclosure Control. Statistics Tidsskrift, 3, 213-225.

Duncan George T. and Lambert Diane (1986). Disclosure Limited Data Dissemination. Journal of the American Statistical Association, 81, 10-28

Fellegi, Ivan P. (1972). On the Question of Confidentiality. Journal of the American Statistical Association, 67, 7-18