Disclosure control that accounts for survey realities: assessing the risk using G-Confid

Peter Wright

Methodology Branch, Statistics Canada. peter.wright2@canada.ca

Abstract: Under the *Statistics Act*, Statistics Canada must protect respondents' confidential data. Assessing the disclosure risk of business survey data traditionally presupposes a census of units, non-negative values, and the protection of every contributor as a legal right and ethical obligation. In practice, it can be challenging to incorporate the aspects of survey sampling that impact the disclosure risk including waivers, negative values and weights. The automated disclosure control software G-Confid, developed at Statistics Canada, is used to measure the sensitivity of tabular business data. This paper identifies ways to assess the impact of some of the aspects of survey sampling on the disclosure control, as recently implemented in G-Confid.

Key Words: Disclosure Control; Sensitivity; Confidentiality; Waiver.

1 Introduction

In the spirit of modernization and being more user-centric, Statistics Canada is reviewing its processes. We have had increasing pressure from internal and external data users to publish more data and provide greater automation of the confidentiality process. In order to meet these demands while ensuring rigorous scientific methods are used, we have improved our approach to the confidentiality of tabular economic data. The current set of improvements stem from incorporating three aspects of survey sampling into the assessment of the sensitivity: the use of waivers, the assessment of variables that contain mixed-sign (positive and negative) values, and the calculation of survey weights.

1.1 Adopting a modern perspective at Statistics Canada

Traditionally, Statistics Canada assessed the disclosure risk of tabular economic data by representing sensitivity rules using the well-known linear sensitivity measure formulation (Cox and Sande, 1979) where the contributions in descending order (x_1, x_2,…) are

weighted by the coefficients ($\alpha_r : -1 \leq \alpha_r \leq 1$) that represent a particular sensitivity rule:

$$S = \sum_r \alpha_r x_r \quad (1)$$

Typically the value (x_r) is the value using during the estimation process to calculate a domain (cell) total. Whereas the estimation process involves the survey weight, the confidentiality process does not use the weight (or at best, the weight is used indirectly a part of a workaround solution).

Statistics Canada has reassessed the methodology of the disclosure risk assessment of its business data. To this end three main objectives have been sought in the modernization exercise: to quantify the impact of waivers, mixed-sign data and survey weights; to assess cell sensitivity without requiring manual adjustment after calculating its value; and to automate the process within G-Confid.

1.2 A new framework

In order to include the survey weights in the assessment of confidentiality, but without distorting the unweighted value (x_r) of each contributor, Gray (2016) proposed the Precision Threshold and Noise (*PTN*) framework. *PTN* allows more than one input variable at the microdata level to be used with the assessment of the sensitivity of the cell. Statistics Canada has adopted the flexible approach offered by the *PTN* framework in order to assess the confidentiality in the presence of survey weights. Additionally, *PTN* offers the advantage of assessing the confidentiality in the presence of two or more aspects, and without having to worry whether the workaround solution for each aspect works well in tandem with the other workaround solutions.

1.3 Applicability to commonly used sensitivity rules

This paper focuses on the assessment of sensitivity using either the PQ rule or the NK rule. The PQ rule (or prior-posterior rule) assumes that a contributor that is most at-risk, or the *target*, is most vulnerable to disclosure by an internal attacker, or *intruder*. The contributor with the largest value takes the role of the target, while the contributor with the second-largest value is considered the intruder. The NK rule (or dominance rule)

indicates that disclosure occurs if the largest N contributors provide more than K% of the total contributions to the cell.

2 Waivers

In order to increase the granularity of its published results, particularly at increasingly detailed levels of geography, Statistics Canada relies upon the goodwill of business enterprises to provide waivers in order not to protect their data. As defined in Elliot et al. (2005), a waiver is a signed record of the respondent granting permission to publish its data.

In the context of the PQ rule, the presence of a waiver affects the roles of the business enterprises. The largest contributor, having supplied a waiver, changes its role from the target to the intruder. The next largest contributor, not having supplied a waiver, becomes the target. In assessing the worst-case disclosure scenario, the automated use of waiver information requires that the appropriate roles be assigned to each business enterprise that contributes to a domain. By extension, if the top two (or more) contributors provide waivers, the sensitivity cannot be positive and the cell is safe to publish.

With the NK rule, one approach is to require a waiver from each enterprise among the top N contributors. When N>1 Statistics Canada proposes an alternative approach whereby the *relative protection* offered by the NK rule is preserved for the contributor that is most at-risk of disclosure *as if using the PQ rule*. Consider an example in which the business enterprise with the largest contribution provides a waiver, where N=2 and K=80, and all survey weights equal one. The same value of the sensitivity could have been found using a PQ rule:

$$\text{Sensitivity ignoring waivers} = \frac{100-80}{80} \times (600 + 300) - 100 = 125$$

$$\text{Equivalent result using PQ rule} = 0.375 \times 600 - 100 = 125$$

In this example, we obtain the same sensitivity using the PQ rule as we obtained using the NK rule if we set $(p/q) = 0.375$. If the largest contributor were to provide a waiver, and the second largest were not, they exchange roles. We then recalculate the sensitivity as

Sensitivity after applying waivers = $0.375 \times 300 - 100 = 12.5$

This approach does not necessarily ensure sufficient protection if the largest contributor were part of an attacking coalition. Hundepool et al. (2012) note the similarity between the PQ rule and the NK rule when N=2. They suggest deriving the value (p/q) from the coefficient α_1 (or equivalently α_2) of the NK rule. Their approach does not require pre-calculating the sensitivity prior to accounting for waivers.

3 Options for treating negative values in a mixed-sign variable

Statistics Canada is adopting two new approaches to assessing the sensitivity of cells in the presence of mixed-sign data: (1) calculating the absolute values of contributions at the level of the union of cells, and (2) using a proxy variable.

3.1 The use of absolute values

Various workaround solutions have been proposed to treat mixed-sign data. Both the Federal Committee on Statistical Methodology (2005) as well as Daalmans and de Waal (2010) discuss the use of absolute values. Tambay and Fillion (2013) support the use of an absolute value for a contributor if it is indicative of the level of both the protection required as well as the protective noise that the contributor can offer to other contributors. If the attacker cannot determine of the sign of the contribution, Giessing (2008) proposed to permit a wider prior interval to estimate the target contribution.

In the presence of mixed-sign data, Statistics Canada has traditionally calculated the absolute values of the contributions ($|x_r|$) in the cells at the most detailed level of the output table. However, the contribution of a particular business enterprise to a union of cells may become inflated by having added its positive contributions to some cells with the absolute values of its negative contributions to other cells.

For example, suppose three business enterprises provide contributions to two cells, I1 and I2. The absolute values are calculated within each of I1 and I2, equivalent to pre-treating the input microdata. The contribution to their union M12 are defined as the sum of

contributions to I1 and I2. The contribution of Enterprise 3 is calculated as 40 (see Table 1). This approach preserves the subadditivity of cells, whereby the sensitivity of a union of cells cannot be greater than the sum of the cells' individual sensitivities (Elliot et al., 2005). Statistics Canada proposes the option to recalculate the absolute values of the contributions at the level of the union of cells. To complete the example, the contribution of Enterprise 3 to marginal cell M12-alt is -20, or 20 in absolute value, and the sensitivity becomes 16. Care must be taken when specifying whether Enterprise 3 contributes noise of either 20 or 40, depending on our assumption of attacker knowledge.

Table 1. Contributions to a marginal cell preserving subadditivity

Contributor	**I1**	**I2**	**M12**	**M12-alt**
Enterprise 1	80	100	180	180
Enterprise 2	60	70	130	130
Enterprise 3	10	|-30|	40	|-20|
Sensitivity[1]	6	-10	-4	16

1. The PQ rule with $(p/q) = 0.2$ is applied.

3.2 The use of a non-negative proxy variable

To make use of a proxy variable, let X denote the original variable and let Y denote a non-negatively valued variable that is representative of the relative magnitude of each business. The variable Y should be judiciously chosen from among the variables that correlate strongly with the amount of protection that X would ordinarily require. For a given business enterprise r, let $Z_r = \max\{|X_r|, \delta Y_r\}$ denote the variable that will be analyzed for sensitivity, with parameter $0 \leq \delta \leq 1$. For example, if X represented the profit (or loss) of a business enterprise, the total gross revenue might serve as the auxiliary variable Y. The definition of Z was proposed by Tambay and Fillion (2013) who recommended using a small value of δ suitable for data scenarios in which $|x_r|$ was very much smaller than $|y_r|$. Applying the PQ rule to proxy variable Z would lead to protecting the ratio $|X/Y|$ to within $\delta p \times 100\%$.

A suitable choice of the value of δ may not be obvious. With that in mind, a data-driven approach is proposed to determine its value: among cells at the most detailed level of the table, the ratio $|x_r|/y_r$ for the r^{th} contributor to a cell is calculated. The resulting values are ranked in ascending order. The percentile of the ranked distribution (π) at which δ is to be selected may then be specified. As a consequence, the term δy_r is used in place of contributions with the smallest π% of ratio values, and $|x_r|$ otherwise.

4. The use of survey weights

Previously, Statistics Canada has only used the survey weights indirectly to determine that a contributor's identity is protected. In a survey sampling context, an elevated survey weight ($w_r: w_r \geq 1$) suggests that ($w_r - 1$) other contributors could have been selected into sample. As described in Tambay and Fillion (2013), Statistics Canada refers to these contributors as *anonymous*, thereby dissociating the contribution from its contributor. Suppose *u* contributors to a particular cell have elevated survey weights, and define

$$x_{anon} = \sum_u x_u \tag{2}$$

When applying the traditional linear sensitivity measure, the anonymous contribution is ranked last ($x_1, x_2, \ldots, x_{R-u}, x_{anon}$) and always has coefficient $\alpha_{anon} = -1$.

Statistics Canada seeks henceforth to use the survey weights directly, depending on the role of the contributors. An intruder that does not know its own weight may be considered to provide protection of $(w_{intruder} - 1)x_{intruder}$. Contributors that are neither target the intruder provide *noise*; using the weights we can use their weighted contribution ($w_r x_r, r \neq target, intruder$). In the presence of survey weights, we are challenged to assign contributors to the roles of target, intruder and noise as the roles do not necessary follow the same order as ($x_1, x_2, \ldots$). The next section describes how this challenge has been met.

5. The *PTN* framework

Each of the survey aspects discussed in this paper – waivers, mixed-sign data, survey weights – necessitate auxiliary information beyond merely the unweighted contribution (x_r) to a survey estimate. The proposed solutions for waivers and mixed-sign data rely on

additional processing steps prior to using the unweighted contributions. However, the use of survey weights directly in the assessment of sensitivity requires more than one input variable. Born of necessity, the *PTN* framework permits users to encapsulate a vector of inputs in order to assess the sensitivity.

5.1 Description and input variables

To assess cell sensitivity, the *PTN* framework assigns three values to represent the impact on the confidentiality assessment of the contribution of each enterprise to a cell:

- Precision threshold (*PT*): the degree of protection that must be accorded to the enterprise contribution to ensure that sufficient protection
- Self-noise (*SN*): the amount of protection provided by an intruder's own contribution; typically zero unless the intruder's contribution has undergone some transformation unknown to the intruder, such as weighting, calibration, or other forms of adjustment
- Noise (*N*): the amount of protection offered by the contribution in scenarios where it is neither the target nor the intruder

Once determined, the sensitivity of any target-intruder pair (t, s) can be calculated as the difference between the target's precision threshold, and the sum of the intruder's self-noise and the noise offered by all other contributors:

$$S(t,s) = PT(t) - SN(s) - \sum_{r \neq t,s} N(r) \tag{3}$$

In the context of a PQ rule, of paramount importance within the *PTN* framework is the identification of the target-intruder pair that is *most sensitive*; this not only determines whether or not the cell as a whole is safe or sensitive, but also which other cells may be suitable for residual disclosure protection. In his paper Gray provides a general algorithm for finding these pairs, labelled *maximal pairs*, among all possible pairs of contributors to a cell. Cell sensitivity is then defined as:

$$S_{cell} = \max\{S(t,s) |\, t \neq s\} \tag{4}$$

Among contributions indexed in descending order of magnitude (i.e., $x_1 \geq x_2 \geq \cdots \geq x_N$) the pair (x_1, x_2) is maximal and the cell sensitivity is equal to

$$S = S(x_1, x_2) = px_1 - q \sum_{r \geq 3} x_r \tag{5}$$

which is equivalent to the PQ rule expressed in the form of the traditional linear sensitivity measure (scaled with respect to *q*.) For information about using the *PTN* framework to express other sensitivity rules such as the NK rule, see the paper by Gray.

Although the *PTN* framework was initially designed to permit the use of survey weights, *PTN* also allows two or more survey aspects to be be represented in the assessment of sensitivity. The *PTN* framework even permits the use of weights below one (due to calibration, indexing or time-series adjustments), a feature that workaround solutions do not generally provide.

5.2 Representing survey realities using *PTN*

Within the *PTN* framework, the r^{th} contributing business enterprise is assigned values to the *PT*(*r*), *N*(*r*) and *SN*(*r*) variables. Table 2 presents three examples: (1) no waivers, no negative values and all weights equal to one, (2) survey weights not necessarily equal to one, and (3) survey weights and waivers.

Table 2. Examples of *PTN* variables[1]

	Example 1	**Example 2**	**Example 3**
$\boldsymbol{PT(r)}$	px_r	$px_r - f(x_r, w_r \mid w_r \neq 1)$	0 if waiver
$\boldsymbol{SN(r)}$	0	$q(w_r - 1)x_r$	$q(w_r - 1)x_r$
$\boldsymbol{N(r)}$	qx_r	$qw_r x_r$	$qw_r x_r$

1. The PQ rule is applied.

Example 1 provides an equivalent result to the traditional linear sensitivity measure (see equation 5). Regarding the function *f*, Statistics Canada looks to use the approach described in section 4 of this paper, or a linear decreasing function as the weight deviates from one.

6 G-Confid: a generalized system with improved functionality

In the early 1980's, Statistics Canada developed its first automated system to identify sensitive cells and to protect data via cell suppression, as described in Sande (1984). In 2009 Statistics Canada introduced a user-friendly version that functions within a SAS® environment and makes use of the SAS/OR® LP solver to select cells for complementary suppression. Known as G-Confid, it forms part of the suite of generalized systems that Statistics Canada developed, uses and maintains for use with various steps of the survey process. The PROC SENSITIVITY module of G-Confid is used to identify sensitive cells. Readers are encouraged to consult Rondeau and Fillion (2011) for further information on G-Confid. The latest release of G-Confid (version 1.07) includes new functionality to process waivers, mixed-sign data and survey weights.

In order to make use of waivers, the G-Confid user includes a numeric variable on the input microdata file that indicates whether a waiver is associated with each record on the file (1 if yes, 0 otherwise). When using PROC SENSITIVITY, the user specifies the Waiver statement and specifies the name of the waiver indicator variable. If the NK rule is specified, G-Confid implements the approach described in section 2 of this paper.

When providing input microdata that include mixed-sign variables, G-Confid calculates their absolute values either by preserving subadditivity (by specifying the AdditNoise option) or by calculating the contributions to unions of cells independently (the NoAdditNoise option). Alternatively, the G-Confid user may include the ProxyVariable statement and name a proxy variable on the input microdata file. The G-Confid user then chooses whether to specify the ProxyRatio parameter and provide a value of δ or to specify the ProxyPercentile and provide a value of π, as described in section 3.2 of this paper.

In the presence of either waivers or mixed-sign data, G-Confid continues to assess the sensitivity using the traditional linear sensitivity measure.

To make use of the survey weight variable, the G-Confid user includes this variable on the input microdata file. The user then identifies the name of this variable within the Weight statement of PROC SENSITIVITY. If a G-Confid user specifies the Weight statement, or a combination of statements associated with treating waivers and mixed-sign data, G-Confid automatically creates and populates the variables specified by the *PTN* framework. G-Confid then identifies the maximal pair associated with each cell and assesses the sensitivity using the *PTN* variables.

7 Conclusion

In order to modernize its assessment of the sensitivity of its business data, with a view to publishing more data and providing greater automation of the confidentiality process, Statistics Canada undertook to (1) improve the disclosure risk methodology in light of survey realities, and (2) automate their application within its generalized system, G-Confid. To this end, G-Confid was upgraded to automate the treatment of waivers, mixed-sign data and survey weights. In particular to account for survey weights, the functionality of the *PTN* framework was built into G-Confid. This framework permits a more refined assessment of cell sensitivity compared to the traditional linear sensitivity measure. As a result, we have been successful in meeting user needs to publish more data, while continuing to ensure that respondent confidentiality is protected.

References

Cox, J.L. and Sande, G. (1979). Techniques for Preserving Statistical Confidentiality. *Proceedings of the 42nd Session of the International Statistical Institute*, Manila.

Daalmans, J., de Waal, T. (2010). An improved formulation of the disclosure auditing problem for secondary cell suppression. Trans. Data Priv. **3**(3), 217–251

Elliot, M., Hundepool, A., Schulte Nordholt, E., Tambay, J.-L., Wende, T. (2005). Glossary on Statistical Disclosure Control. UNECE document.

Federal Committee on Statistical Methodology. (2005). *Statistical Policy Working Paper 22 (Second version, 2005) - Report on Statistical Disclosure Limitation Methodology*. U.S. Office of Management and Budget, Washington, D.C.

Giessing, S. (2008). Protection of tables with negative values. ESSNet report, Destatis.

Gray, D. (2016). Precision Threshold and Noise: An Alternative Framework of Sensitivity Measures. In: Domingo-Ferrer, J. and Pejić-Bach, M. (Eds.) *Privacy in Statistical Databases*, LNCS 9867, Dubrovnik. 15-27

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., De Wolf, P.-P. (2012). *Statistical Disclosure Control*, Chichester. Wiley Series in Survey Methodology

Rondeau, C. and Fillion, J.-M. (2011). G-Confid: Statistics Canada's confidentiality software. *Proceedings of Statistics Canada Symposium*, Ottawa. 114-119

Sande, G. (1984). Automated cell suppression to preserve confidentiality of business statistics. *Statistical Journal of the United Nations ECE*. 2: 33-41

Tambay, J.-L., Fillion, J.-M. (2013). Strategies for processing tabular data using the G-Confid cell suppression software. *Proceedings of the Joint Statistical Meetings*, Montreal. 3652-3663