Waivers in business surveys: A systematic approach to increase the amount of publishable information

Jean-Sébastien Provençal, Hélène Bérard

Statistics Canada
11-L and 11-K, R.H. Coats Building, Ottawa, Canada K1A0T6 / provjea@statcan.ca and berahel@statcan.ca

1. Introduction

The suppression in information (cells) of published tabulated survey data is a technique commonly used to preserve the confidentiality of a respondent's data. In business surveys, the suppression patterns are usually driven by two criteria: a minimum number of respondents and dominance rules where, for example, a cell could be suppressed if a business accounts for more than 80% of the cell total.
Due to the highly skewed business populations with few large businesses and many small ones, the suppression of cells frequently stems from the presence of one or several businesses that dominate a sector (dominance rule). The suppression of one cell will likely trigger the suppression of other cells in order to reduce the risk of residual disclosure. In an effort to publish more cells, waivers are sometimes obtained from large contributors that allow the publication of the cell's data despite the risk of disclosure. These businesses are often selected in an *ad hoc* manner. The decision to obtain a waiver from a respondent for a given cell and not another may lead to different suppression patterns, some of them can be more optimal than others according to survey and user requirements.
In this paper, we present a method that attempts to identify the eligible potential waivers that will lead to an optimal suppression pattern. First, we describe some of the notions related to confidentiality and disclosure. Second, we give some details regarding CONFID, a software used in Statistics Canada to identify the information that should be suppressed. We summarize the methodology used in this software. Third, we discuss how waivers could be used to minimize the amount of suppressed information and we explain the method developed to identify who should be targeted for a waiver in order to be more optimal in terms of survey and user requirements. Finally, we discuss results obtained with practical example using the Canadian Annual Survey of Manufacturing.

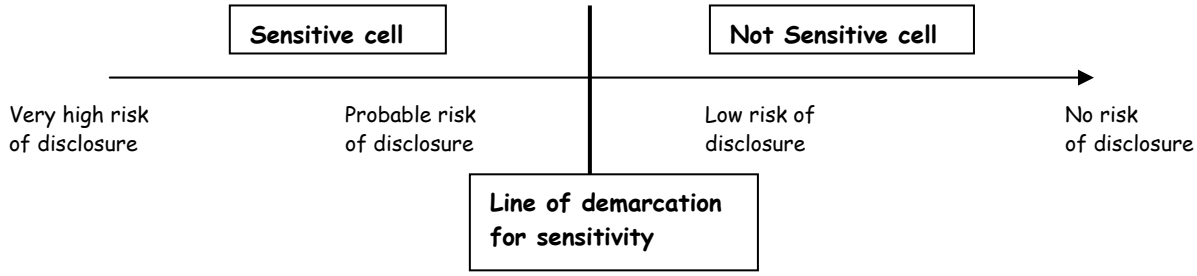
2. Concepts related to confidentiality

2.1 Waiver

A waiver is a written agreement between a statistical agency, in this case Statistics Canada, and a respondent. Under this agreement, the respondent releases the statistical agency from any liability regarding its obligation to preserve the confidentiality of the respondent.

2.2 Sensitive Cell

In the area of confidentiality we use the concept of sensitivity to describe and quantify the risks of potential disclosure within a cell. A cell is deemed sensitive if it is felt that there is a probable risk of disclosure. The line of demarcation is established according to confidentiality rules that reflect how secure we want to be.



We use confidentiality rules to quantify this risk,. One type of rule we use is the (n,k) rule. With this rule, the sensitivity of a cell is established according to the main contributors of the cell. If the n top

contributors represent more than k percents of the cell's total, then the cell is sensitive. We can define this rule as a function.
For example, with a (n,k) rule where n = 3 and k = yy, we can say there is a risk of disclosure if the three most dominant contributors: $X_{(1)}, X_{(2)}$ and $X_{(3)}$, account for more than *yy%* of the cell's total.
The sensitivity function, $s(x)$, could be formulated as follows:

$$s(x) = \frac{100}{yy}\left[X_{(1)} + X_{(2)} + X_{(3)}\right] - t(x)$$ where $t(x)$ is the cell's total.

A cell would be qualified as *sensitive* if $s(x) < 0$. The value taken by $s(x)$ allows for quantifying a cell's degree of sensitivity.
In determining if a cell is sensitive or not, we can also add other criteria related to the number of units. For example, if a cell contains fewer than three units, we can say this cell is sensitive.

2.3 Complementary cell
If a cell has been identified as sensitive, it is not published because there is a risk of disclosure and it is masked. In a published table containing several totals and subtotals, masking a cell may require blocking one or more other cells in order to protect the sensitive cell. Concealing these make it impossible to deduce the masked value from the other table entries. These complementary cells are hidden for reasons of *residual confidentiality*. For example, consider the following table:

	1	2	3	TOT
A	12	8	9	29
B	4	4	6	14
C	5	3	3	11
TOT	21	15	18	54

Suppose that we need to mask cell C2 because it is a sensitive cell. We must also select complementary cells to protect C2, therefore we obtain:

	1	2	3	TOT
A	12	8	9	29
B	C	C[1]	6	14
C	C	S[2]	3	11
TOT	21	15	18	54

In protecting cell C2, we need to mask B1, B2 and C1. If any one of these three cells is not masked, the value of C2 can be deducted. For example, if B1 is not blocked, we can calculated C1=21-12-B1=5 and C2=11-3-C1=3, rendering C2 not confidential.
We could have blocked another set of cells to ensure confidentiality, such as B2-B3-C3. The choice of the complementary cells depends on various parameters and considerations. For example, one consideration is the size of the sensitive cell in terms of the variable of interest. We need complementary cells large enough to protect the confidentiality of cell C2. The size of the complementary cells required depends on the confidentiality rules described earlier.

3. Methodology with CONFID

CONFID is a software that identifies sensitive cells according to your confidentiality rules and identifies all the complementary cells that should be masked in order to produce tabulations where the confidentiality of each contributing unit is preserved. For a given problem, i.e. a set of tabulations to produce, there are many possible solutions. Those solutions are called confidentiality patterns. CONFID finds an optimal solution according to some pre-specified criteria.

[1]Complementary Cell.
[2]Sensitive Cell.

The goal is to mask as few cells as possible and withhold as little information as possible. Sometimes, you must have to choose between publishing two small cells or one large cell. The strategy will depend on your objectives for the survey and the publications. CONFID allows you to prioritize the amount of information in dollars value or the number of cells. It is also possible to set up CONFID to find an optimal confidentiality pattern that is a trade-off between the two criteria. CONFID will find an optimal solution by resolving a system of numerical equations that take into account the tabulations to be published, the data set itself, your confidentiality rules and the criteria you want to prioritize.

4. Use of Waivers

With the confidentiality pattern obtained using CONFID, you can suppress a considerable amount of information. This happens when you have a survey for which you produce very detailed tabulations composed of small cells with few units. The use of waivers can help to reduce the amount of suppressed information. There are two ways waivers can lead to an increase of publishable information:

1) the sensivity of a cell may be caused by one or a few large contributors. Obtaining waivers for these companies will allow you to publish the whole cell.
2) when you eliminate the sensitivity of a cell, the complementary cells, that were chosen initially, may no longer be required to be masked to protect the sensitive cell. Publishing these cells may lead to substantial gains in terms of dollars and cells.

5. New approach : Score function

Because of cost and operational constraints, it is not always possible or feasible to obtain waivers for each sensitive units. You may need to be selective and choose only a subset of units for which you will try to obtain waivers. Our approach attempts to identify systematically which units should be prioritized. Our goal is to find which waiver provides the best return in terms of the amount of publishable information. In other words, we identify which sensitive units are responsible for the largest amount of suppressed information originating from sensitive and complementary cells.
In order to be able to rank potential waiver requests, we sought to establish some criteria. We based these criteria on the characteristics of the cell and of the block containing the sensitive unit.
The cell and the block are the two first levels of a tabulation. A cell corresponds to the most detailed element that is being published. In business survey, a cell is usually made of two dimensions. Typically, it will include a detailed industrial dimension like a subset of four-digit Standard Industrial Classification (SIC) codes and a geographical dimension e.g. a province. A block is an aggregation of cells (e.g. four-digits SIC by province) that includes the totals for the first dimension categories, the totals for the second dimension categories and the grand total.
Fig1. Block

	GEO A	GEO B	GEO A+B
Industry1	CELL1A	CELL1B	TOT1,A+B
Industry2	CELL2A	CELL2B	TOT2,A+B
Ind. 1+2	TOT1+2,A	TOT1+2,B	TOT1+2,A+B

5.1 Criteria for a cell
We identified two characteristics that could be used to score a cell. Each criterion is represented as a ratio and is read as a percentage.

4.1.1) Number of waivers required before a cell becomes fully publishable
4.1.2) Size of the cell based on the main variable of interest

The percentages are combined to get a unique score. Both criteria are weighted according to their deemed importance. Here are a few details on each criterion.

5.1.1 Number of waivers required. We are better off working on a cell if it requires a smaller number of waivers. Our criterion is based the ratio of the number of waivers required to make a cell publishable over the maximum number of potential waivers required among all the sensitive cells. This criterion scores best if the cell presents a smaller number of waivers. It is defined by:

$$ce_1 = \frac{(\max(x) - x)}{\max(x)} \times 100$$

where x is the number of waivers that have to be requested to remove the cell's sensitivity and $max(x)$ is the largest possible value of x across all the sensitive cells. ce_1 will be a high percentage if the number of required waivers is small.

5.1.2 Size of cell in terms of the variable of interest. We consider here the relative size of the cell in terms of the main variable of interest. The reason we use the relative size is to grant more priority to the largest cells. The formula for this criterion, in terms of the variable of interest, is:

$$ce_2 = \frac{y}{\max(y)} \times 100 .$$

where y is the cell's total for the variable of interest and $\max(y)$ is the largest possible value of y across all the sensitive cells. ce_2 will be a higher percentage if y value is larger.

5.1.3 Function for scoring cells. Using the two criteria for the cells, we can create a function to score and rank each cell. This function is formulated as follows:

$$f(ce) = w_1 ce_1 + w_2 ce_2$$

where the ce_i variables represent each criterion's value. The w_i variables are the weights applied in light of the criterion's deemed importance. The sum of the w_i variables is 1. The $f(ce)$ score for a cell is a percentage.

5.2 Criteria for a block

We came up with four criteria for the block containing sensitive units. Each criterion is represented as a ratio and is read as a percentage.

- 5.2.1 Number of waivers required before a cell becomes fully publishable
- 5.2.2 Size of the block based on the main variable of interest
- 5.2.3 Quantity of masked confidential and residual information
- 5.2.4 Number of cells in the block containing a sensitive unit. The form of these criteria resembles that used for cells, as follows.

The percentages are combined to get a unique score. Both criteria are weighted according to their deemed importance. Here are a few details on each criterion.

5.2.1 Number of waivers required. We are better off working on a block if it requires a smaller number of waivers. We calculate the ratio of the number of waivers required to make the block completely publishable over the maximum number of potential waivers required among all the blocks where we observe sensitivity. This criterion therefore scores best if the cell presents a smaller number of waivers. It is defined by:

$$bl_1 = \frac{(\max(x) - x)}{\max(x)} \times 100$$

where x is the number of waivers that have to be requested to remove the sensitivity observed in the block and $max(x)$ is the largest possible value of x across all the blocks. bl_1 will be a high percentage if the number of required waivers is small.

5.2.2 Size of block. We want to grant more priority to the largest blocks. Hence, we defined this criterion considering the relative size of the block in terms of the main variable of interest. The formula for this criterion is:

$$bl_2 = \frac{y}{\max(y)} \times 100$$

where y is the block's total for the variable of interest and $\max(y)$ is the largest possible value of y across all the blocks presenting sensitivity. bl_2 will be a high percentage if the number of deliveries is large.

5.2.3 Masked Information. We also want account for the quantity of concealed confidential and residual information in a block where we observe sensitivity in some of its cells. To consider this we defined:

$$bl_3 = \frac{y_1 + y_2}{\max(y_1 + y_2)} \times 100$$

where, within the block, y_1 is the variable of interest total concealed because of sensitivity, and y_2 is the amount of the variable of interest masked for complementary purposes (residual disclosure). The variable $\max(y_1 + y_2)$ is the largest possible value of these two variables across all the blocks where information is hidden for confidentiality purposes. bl_3 will have a high value if a large quantity of information is blocked.

5.2.4 Number of blocked cells. We want to prioritize the block where a large number of cells is suppressed. The goal is to rank blocks according to the number of cells masked. The formula for this criterion is:

$$bl_4 = \frac{x}{\max(x)} \times 100$$

where x is the number of sensitive cells and $max(x)$ is the largest possible value of x across all the blocks presenting sensitivity. bl_4 will be a high percentage if many cells are masked.

5.2.5 Function for scoring blocks. Using the four criteria for the blocks, we can create a function to score and rank each block. This function is formulated as follows:

$$f(bl) = w_1 bl_1 + w_2 bl_2 + w_3 bl_3 + w_4 bl_4$$

where the bl_i variables represent each criterion's value. The w_i variables are the weights applied according to the criterion's deemed importance. The sum of the w_i variables is 1. The $f(bl)$ score for a block is therefore a percentage.
By taking the two score functions for the cell ($f(ce)$) and the block ($f(bl)$), we can establish the relative priority of each cell. In the following section, we analyze the impact of this kind of classification method using a practical example.

6. Practical example

In this section, we quantified the gain obtained by using waivers. We used the food industries data from 1997 Annual Survey of Mining (ASM). That year, ASM was census of units above a certain threshold. In this sector, roughly 3100 units contribute to a total manufacturing shipments of 50,468,609K $. Data are tabulated to cover a total of 319 cells at various industry by geographical area levels (including totals and subtotals). Without using waivers we would publish roughly 55% of these cells.
We compared our score function method to identify waivers versus other type of ranking. All options in table 5.1 are expressed in terms number of cells and percentage of manufacturing shipments published. In this table, we are comparing 4 options : (i) without using waivers, (ii) using top-10 waivers according to our score-function, (iii) using top-10 waivers according to manufacturing shipments by unit and (iv) using top-10 waivers according to total cell manufacturing shipments. We

gave results for each level of industry i.e. SIC2 (10 : Food industries), SIC3 (ex.: 101 : Meat and poultry products industry) and SIC4 (Ex.: 1012 Poultry products industry).

Table 6.1 Efficiency of each option in terms of number of cells published (N) and percentage of manufacturing shipments (% of Ship.)

Industry Level	N Total	Option (i)		Option (ii)		Option (iii)		Option (iv)	
		N	% of Ship.	N	% of Ship.	N	% of Ship.	N	% of Ship.
SIC2	**13**	13	100	13	100	13	100	13	100
SIC3	**88**	61	97.1	65	97.5	61	97.4	62	97.1
SIC4	**218**	100	87.6	116	91.4	105	90.2	103	89.5

As we can see, by comparing option (i) with (ii), (iii) and (iv), using waivers will increase the amount of published information. The gains are greater at the most detailed level (SIC4). Also, if you compare the three options used to prioritized waivers, you can notice that our score function (option (i)) allows us to publish a non-negligible extra amount of information, especially again for the most detailed levels. At the most detailed level (SIC4), we can see that using 10 waivers with our score function allow us to publish an extra 5% of manufacturing shipments and, roughly, an extra 15% of cells.

7. Conclusion

To summarize, we can say that using waivers and prioritizing them can really lead to a significant increase of published information when you have to deal with tabulated survey data. Our score function based on cell and block characteristics of a sensitive cell help to optimise gains obtained by using waivers.
Further work related to this score function could be to define a more optimal way to combine cell and block characteristics. It would be also possible to look at the weighting used to quantify each criteria in each function (cell and block). Certain aspects of the weighting could be optimised, but there will be always a part of the decision that will depend on users priorities.