

THE MATHEMATICAL BASIS FOR STATISTICS CANADA CELL SUPPRESSION SOFTWARE: CONFID

Dale Robertson¹ and Ioana Şchiopu-Kratina²

ABSTRACT

CONFID is a suite of software in operation at Statistics Canada. It is based on a mathematically sound method of cell suppression and has been successfully used to protect confidentiality in tabular data produced and published by large economic surveys. The presentation gives a short introduction to the notion of sensitivity of a cell, followed by a description of the mathematics of cell suppression and the stepwise procedure that implements it in the module SUPPRESS. In an attempt to reduce the number of complementary suppressions (which are nonsensitive cells), we propose a cell suppression method as an alternative to the stepwise procedure presently used in CONFID.

KEY WORDS: Sensitivity; pattern of suppressions; complementary suppression; objective function.

RÉSUMÉ

CONFID est un ensemble de logiciels utilisés à Statistique Canada. Les logiciels sont basés sur une méthode mathématique solide de suppression locale et ont été utilisés avec succès pour protéger le secret dans les tableaux de données d'enquêtes économiques de grande envergure. L'article présente une courte introduction à la notion de sensibilité des cellules suivi des mathématiques de la suppression locale et sa programmation dans SUPPRESS. Afin de réduire le nombre d'éliminations complémentaires (qui sont les cellules non-sensibles) nous proposons une méthode de suppression qui pourrait remplacer la procédure par étapes présentement utilisée par CONFID.

MOTS-CLÉS: Sensibilité; tendance de suppressions; élimination complémentaire; fonction objective.

1. INTRODUCTION

1.1 General description of CONFID

CONFID is a suite of software in operation at Statistics Canada which has been successfully used for protecting confidentiality for tabular data. It was designed by G. Sande, a former Statistics Canada employee and it is based on sound mathematical methods. CONFID can be applied to data collected for real-valued, positive variables (production, employment, financial activity etc) and presented in the form of additive tables at various levels of aggregation. CONFID is presently used at Statistics Canada to protect confidentiality of the data published by large economic surveys like the Census of Agriculture and the Survey of Employment, Payroll and Hours.

1.2 Organization of the article

Section 2 gives some basic concepts and discusses the definition of sensitivity of a cell. Section 3 discusses complementary suppression and the link between the ambiguity required by a sensitive cell and the amount of complementary suppression needed. In section 4 we present a different formalization of the problem of complementary suppression when more sensitive cells are present. This formalization may lead to patterns of suppression that are closer to an optimum than the patterns currently produced by CONFID. The presentation concludes with general remarks on the mathematics of CONFID (section 5).

1

Dale Robertson, SDD, Statistics Canada, R.H. Coats Building, 14th Floor, Ottawa, K1A 0T6

2

Ioana Schiopu-Kratina, HSMD, Statistics Canada, R.H. Coats Building, 16th Floor, Ottawa, K1A 0T6.
Internet:schiopu@ststcan.can

2. SENSITIVITY OF A CELL

2.1 Definition and example

Strictly speaking, a cell is an entry at the lowest level of aggregation. The term is used loosely and it also covers aggregation of elementary cells. A cell could be defined by a geographic identifier, economic activity, a measure of size, all or some of the above. In what follows, we consider only cell totals for characteristics that can take only positive, real values. The number of identifiers of a cell constitute the dimension of the cell and of the tables that contain it. The mathematics used in CONFID applies to any number of dimensions, although the software can handle at most 3 dimensions.

A cell is sensitive if the publication of its total leads to an accurate estimate of the value of the commodity associated with one respondent. Generally, sensitive cells are cells for which the distribution of the commodity of interest is skew and the published total is a good estimate of the total of a few large units. A coalition of large contributors (excluding the largest) may subtract their total from the published total and obtain a good estimate of the total of the largest respondent.

Example 1 (Sande 1984). The $(n,k) = (3, 75\%)$ sensitivity rule states that a cell is sensitive if the total of the three largest contributors constitutes more than 75% of the cell total or that the cell total is within 1/3 of the contribution of the 3 largest units.

Let us order the units in a cell according to the magnitude of the reported commodity, e.g. $x(1)$ is the largest contributor, $x(2)$ is the second largest, $x(n)$ is the smallest etc. The $(3,75\%)$ rule is a linear sensitivity rule and the sensitivity $s(x)$ of the cell x is given by: $s(x) = 1/3[x(1)+x(2)+x(3)] - [x(4)+ \dots x(n)]$. A cell x is sensitive if $s(x) > 0$ and nonsensitive otherwise. If a cell is sensitive according to this rule, the coalition of the second and the third largest contributors could arrive at an estimate of the largest contributor total within a margin of error that is considered unacceptably low. Notice that this formula actually attaches a numerical value of sensitivity to each cell.

Good sensitivity rules satisfy (1) (G. Sande, L. Cox (1981)):

$$(1) \quad s(x) - t(y) \leq s(x \cup y) \leq s(x) + s(y),$$

where $x \cup y$ is the aggregation of x and y and $t(y)$ is the total of cell y . Formula (1) justifies 'collapsing categories' of appropriate sizes in an attempt to avoid disclosure of confidential information. The right hand side of (1) shows that, aggregating a sensitive cell x with

a nonsensitive cell y leads to a new cell with a sensitivity that is lower than that of x . The left side of the inequality places a lower bound on the sensitivity of the aggregated cell: one cannot create a nonsensitive cell by combining a sensitive cell with a very small nonsensitive cell. Notice that, by symmetry, we also have $s(y) - t(x)$ as a lower bound in (1). However, this bound is redundant when x is sensitive, y is not and $t(y) > t(x)$.

2.2 Relationship with the required ambiguity

Assume that a cell x is sensitive with positive sensitivity $s(x)$. Publishing its total $t(x)$ will disclose information about its largest contributor. It is easy to see that, publishing $t'(x) = t(x) + s(x)$ instead offers the minimum amount of protection required according to the specified sensitivity rule. Thus, $a(x) = s(x)$ is the amount of ambiguity required for the total $t(x)$ to offer a 'safe' estimate of the largest respondent in the cell. The larger the sensitivity of the cell, the larger the amount of ambiguity required and, as we shall see in the next section, the larger the amount of complementary suppression needed.

3. COMPLEMENTARY CELL SUPPRESSION

3.1 The mechanism of complementary cell suppression

The sensitive cells would have to be suppressed in order to avoid disclosure of information related to the value associated with the largest contributor to the cell total. One must also suppress cells that are algebraically related to a sensitive cell (called complementary suppressions), or else an unacceptable sensitive range of the suppressed cell would be recovered. This process should be repeated for each sensitive cell. Overall, we would like to suppress as little information as possible from the tables. The following example illustrates the procedure.

Example 2 Internal cells have been suppressed from the following table:

	100	3
100	x_{11}	x_{12}
3	x_{21}	x_{22}

Using the information provided by the marginals, we recover the ranges of the suppressed entries: $[97, 100]$ for x_{11} and $[0, 3]$ for the others. Assume that $t_{11} = 99$ and that the ambiguity required is $a_{11} = 2.5$. Under these conditions, the recovered range is

too narrow. Indeed, the least sensitive value is $99 - 2.5 = 96.5$, $97 [96.5, 99]$, in other words, the lower bound of the recovered interval is well within the sensitive range. In this example, we must suppress the marginals to avoid disclosure.

Definition: A pattern of suppressions offers upper protection to a sensitive cell x if one can fill out the suppressed entries with values that are within 50% of the reported values and so that the total $t'(x)$ assigned to x , $t'(x) [u(x), t(x)/2]$ and that the additivity of the resulting tables, which include publishable values, is preserved. Here $u(x) = t(x) + a(x)$ is the upper limit of the sensitive interval (upper tolerance).

It follows from the above that, in order to create a safe pattern, one must create sufficient ambiguity in the tables to ensure adequate protection for each sensitive cell. It is easy to see that, due to symmetry considerations, a one sided pattern is sufficient to ensure both upper and lower protection.

G. Sande assumed that a potential intruder is capable of estimating any total within 50% of the true value. It is now commonly accepted that, in order to preserve the meaningfulness of the data, the modification of totals should be kept within 50% of their true values.

3.2 Construction of a safe pattern for a sensitive cell

The creation of a safe pattern for a sensitive cell is carried out along the following steps. First, the total of the sensitive cell is modified to the upper (or lower) limit of its sensitive range. Other cell totals must be modified so that the additivity in tables is preserved. There are usually more possibilities (solutions) that render the tables additive. For each solution, the cells that require a nonzero adjustment are the complementary cells. Of all possible solutions, the module SUPRESS of CONFID selects a solution that minimizes the information lost in the published tables as a result of complementary suppression. Finally, the sensitive cells along with the complementary cells are suppressed and a safe pattern is created.

Example 3 Consider the 2x2 table of cells with marginals:

2	3	5
2	10	12
4	13	17

Assume that the cell with total 10 is sensitive, and the upper tolerance of its sensitive interval is 11.

We replace 10 by 11 and modify other cells to render the table additive:

2+1	3-1	5
2-1	10+1	12
4	13	17

Notice that all internal entries have changed. They are complementary cells and they must be suppressed.

The following example shows that the pattern of complementary suppression and the amount of information lost in tables depends on the amount of ambiguity required and ultimately on the value of the sensitivity of the cell.

Example 3' In Example 3, we assume that the ambiguity required is now 2 rather than 1. Because of the constraints imposed on the variation of cell totals, the internal cells may no longer serve as complements and marginals must be used for balancing the table:

2	3	5
2	10+2	12+2
4	13+2	17+2

Three marginal cells will have to be suppressed and so the amount of information lost is larger as a result of an increased amount of ambiguity required (the sensitivity of the cell is larger).

We now present the general situation on the simplest possible example: a single linear equation with three totals.

Example 4 Assume that the totals of three cells satisfy the equation below and that cell x_{11} is sensitive with ambiguity a_{11} :

$$t_{11} + t_{12} = t_1$$

We substitute $t_{11} + a_{11}$ for t_{11} , $t_2 + y_{12} - z_2$ for t_{12} and $t_1 + y_1 - z_1$ for t_1 in the equation above. We require that the y 's and the z 's be positive: they represent positive and negative changes to the totals. It is necessary to break the changes into two positive changes because linear programming techniques require that all variables be positive. We also require that the variables do not exceed the limits for their cells, i.e. $t(x)/2$.

The equation now becomes:

$$(2) \quad a_{11} + y_{12} - z_{12} - y_1 + z_1 = 0,$$

and the constraints on the variables are: $0 \leq y_{12} \leq t_{12}/2$, $0 \leq z_{12} \leq t_{12}/2$, $0 \leq y_1 \leq t_1/2$, $0 \leq z_1 \leq t_1/2$. Some of these variables will be nonzero in order to offset a_{11} .

3.3 The objective function

Among all safe patterns of suppression, we would like to select one for which a combination of the number of suppressed cells and their total, which will no longer be available to users after suppression, is a minimum.

Linear functions of cell variations allow for a full use of the linear programming technology, which is a great advantage. CONFID offers several such objective functions. They differ by the coefficients that multiply these variations (the $y+z$'s). The constant cost function, or CONCST, assigns equal coefficients to all nonsensitive cells. Because larger variations are absorbed by cells with larger totals, minimization of this objective function may lead to suppression of fewer cells with larger totals. The size cost SIZCST has the cell total $t(x)$ as coefficient for cell x . Its use leads to the suppression of (perhaps many) small cells. The information cost INFCST has as coefficient $\log[1+t(x)]/t(x)$. Its minimization leads to the suppression of a small number of large cells, as it is a decreasing function of $t(x)$. The digit cost DIGCST assigns the coefficient $\log[1+t(x)]$ to cell x and strikes a balance between SIZCST and INFCST. G. Sande suggests that the objective functions be used in conjunction with one another. For example, SIZCST can be used in a first run, at the end of which a subset of the original system containing all sensitive cells and their complements will have been found. Because this cost function may select many small cells and as optimization is done in a stepwise manner, the subsystem may be still too large and may contain redundant complements. The procedure can then be repeated for the subsystem with INFCST as objective function. As this cost function tends to retain larger cells, some of the small complements picked in the first pass would not be reselected. The net result is a safe pattern with a smaller number of complements. The reason why applying the procedure twice may reduce the set of complements is that the stepwise manner in which suppression is performed is not optimal (see Example 5).

3.4 The sequential approach to finding complements

So far we discussed the creation of a safe pattern for one sensitive cell. In most applications, there are hundreds of sensitive cells which must be protected. CONFID 'treats' sensitive cells one at a time, starting

with the sensitive cell requiring the largest amount of ambiguity. All sensitive cells are given a 0 coefficient in every objective function so that they are extensively used for rebalancing tables as they will be suppressed anyway. As a particular cell is protected, other sensitive cells may be fully or partially protected. This is exploited in the sequential algorithm. The complements found in one run will be assigned a 0 cost in all subsequent runs so that they are re-used for balancing tables. Although this optimizes the procedure at every step, globally we are not obtaining an optimum solution. Example 5 below illustrates this statement.

Example 5 (presented to the ASA by C. Sullivan and L. Zayatz of the U.S.B.C.) The two tables below are linked and the first column of the first table represents the row totals in the second table. We use P for primary suppressions (sensitive cells) and C for complementary suppressions. There are 3 primary suppressions and the final suppressions pattern depends on the order in which sensitive cells are protected in the sequential approach. If we start with the cell with total 42 in the second table and attempt to minimize the total value suppressed, we obtain the following final pattern:

95C	2259	6730P	23758	32842
554	4325	9449	22766	37094
1067	11308	16902	25462	54739
1716C	17892	33081P	71986	124675
53C	42P	95C		
306C	248C	554		
357	710	1067		
716C	1000	1716C		

Notice that a better solution would have required the suppression of the marginal cell with total 1000 in the second table, in lieu of 4 other complements, thus saving $53 + 306 + 716 + 248 - 1000 = 1323 - 1000 = 323$ in units and 3 in the number of cells published in the final pattern.

If the cell 42P above had the largest sensitivity (which is possible because cells at the lowest level of aggregation tend to have larger sensitivity, cf. formula (1)), CONFID would have created the same pattern as above. Once the corner cell 53C is selected, it will be given a 0 cost in the next run. Sensitive cells 6730P and 33081P could 'protect' each other in the second run, as they are on the same row and thus belong

to the same 'balancing' equation. Once the cell 95C is selected as complement, the non-optimal pattern will be found: 53C will be re-used in the second table and consequently 716C.

4. A DIFFERENT MATHEMATICAL FORMULATION

4.1 Motivation for a new formulation

We propose a different mathematical formulation of the cell suppression problem when several sensitive cells must be protected. The idea is to 'protect' each sensitive cell independently and then minimize the loss of information in a common, final pattern. We feel that this formulation leads, at least in theory, to a global optimum to the problem of complementary suppression. In other words, we hope to minimize the amount of suppressions in tables so that situations like the one described in example 5 above do not occur. On the other hand, this amounts to considering a much larger system, which consists of as many copies of the original system as there are sensitive cells. As objective function to minimize, we can consider various possibilities, all of which give a bonus to complements that are used to protect more than one sensitive cell. It is conceivable that this formulation had been considered at the time CONFID was built but was discarded as unfeasible. What was unfeasible 20 years ago may be feasible now and so we feel that any solution that may require less suppression in the published tables would be worth pursuing now. We restricted our investigation of objective functions to linear functions of cell variations. Again, the progress made in programming methods and techniques may grant considering a broader class of objective functions, e.g. integer valued objective functions in order to minimize suppression in tables.

4.2 Examples of the new formulation

The optimal solution in the example that follows is obvious. The example was given for the purpose of illustration only, as it is the simplest.

Example 6 (Example 4 revisited) We consider one table containing two internal sensitive cells (ambiguities a_{11} and a_{12}) and their aggregation:

$$t_{11} + t_{12} = t_1$$

We attempt to protect each sensitive cell using two different systems with variables $y_{11}, z_{11}, y_{12}, z_{12}, y_1, z_1, y_1', z_1'$. The equation above generates, as in (2), the system:

$$\begin{aligned} a_{11} + y_{12} - z_{12} - y_1 + z_1 &= 0 \\ a_{12} + y_{11} - z_{11} - y_1' + z_1' &= 0 \end{aligned}$$

The constant cost objective function CONCSST (recall that the sensitive cells are assigned 0 cost) is:

$$y_1 + z_1 + y_1' + z_1'$$

and there are other constraints on the size of the ranges of the variables. If:

$$(3) \quad a_{11} \leq t_{12}/2, a_{12} \leq t_{11}/2$$

we may take $z_{12} = a_{11}, z_{11} = a_{12}$, all other changes equal to 0 and obtain the optimal solution.

Notice that, even though the optimum solution was recovered through the use of CONCSST, this function does not identify the two variations as belonging to the same cell x_1 (the aggregate cell). We next define functions that actually decrease the cost of a nonsensitive cell if it is being used twice. In the example above, consider $Y \geq 0, Z \geq 0$ with:

$$(4) \quad Y - Z = y_1 + y_1' - z_1 - z_1'$$

We may wish to find:

$$(5) \quad \min\{Y + Z\} = \min |y_1 + y_1' - z_1 - z_1'|, \text{ or } \min t_1(Y+Z) = \min t_1 |y_1 + y_1' - z_1 - z_1'|,$$

where t_1 is the cell total and $|a|$ is the absolute value of a . These functions allow subtraction of opposite changes that affect the same cell that is used repeatedly. Subtraction is encouraged in (5), as it diminishes the total variation of the cell.

Example 7 (Example 5 revisited). We will show that the method described above will retain the optimal solution over the solution provided by the stepwise procedure.

Let $a = a_1 = a_2$ be ambiguities associated with the sensitive cells in the first tables, b the ambiguity associated with the only sensitive cell in the second table. Assume that $b \geq a$ and that $a \geq b$

In the first table, we add the value of the ambiguity a to the top sensitive cell and subtract the same amount from the column total, which is a sensitive cell. The table is balanced by the variation of the complements in the first column, in the directions shown by the arrows:

First table

a	95C	2259	6730P	a	23758	32842
	554	4325	9449		22766	37094
	1067	11308	16902		25462	54739

a	1716C	17892	33081P	a	71986	124675
---	-------	-------	--------	---	-------	--------

The cost incurred by the first table is the same in both solutions so it will be ignored in the ω

The second table in the sequential approach was balanced twice in the following manner:

Second table in the sequential approach:

b	53C	a	42P	b	95C	a
b	306C		248C	b	554	
	357		710		1067	

716C	a	1000	1716C	a
------	---	------	-------	---

To calculate our cost function for the sequential approach, we look first at the complements that appear only in the sequential approach. Their contribution to our cost function, which allows subtraction before taking the absolute value, is larger than 1270a. Indeed, $53(b - a) + 306b + 248b + 716a = 53 + 554b + 716a - 1270a$

Assume now that the total of the sensitive cell was moved down by b (this will further increase the cost of the second table in the sequential solution). We now calculate the cost incurred by the optimum solution:

Second table in the optimum solution:

53	42P	b a	95C	b a
306	248		554	
357	710		1067	

716	1000C	b a	1716C	b a
-----	-------	-----	-------	-----

There is a contribution to our cost function of $1000(b-a)$ from suppression that occurs in the optimum solution only. This is smaller than 1270a. Common complements contribute: $95a + 1716a$ in the sequential approach, $95(b-a) + 1716(b-a)$, which is smaller, in the optimal solution. Therefore, our cost function eliminates the sequential solution in favour of the optimum solution

5. ON THE MATHEMATICS OF CONFID

The mathematical method which is at the basis of CONFID is sound. Among existing cell suppression methods, the method used by CONFID is the safest.

Sensitivity of unions of sensitive cells along a row or a column of a table is also taken into account. The applicability of the method is quite broad. In principle, any system of linear equations can be used and the limitation in dimensionality of the cells is basically a software limitation. CONFID may over suppress due to the stepwise manner in which sensitive cells are protected. There is therefore scope in trying to obtain a global minimum for a suitably selected objective function in order to reduce the amount of complementary suppression in published tables.

REFERENCES

- [1] Cox, L. (1981): Linear sensitivity measures and statistical disclosure control, *Journal of Statistical Planning and Inference* 5, 153-164.
- [2] Kelly, J.P., Golden, B.L., Assad, A.A.(1990): Cell suppression: Disclosure protection for sensitive tabular data, *Working Paper Series MS/S 90-001, College of Business and Management, University of Maryland, College Park, Maryland 20742.*
- [3] Lougee-Heimer, R.(1989): Guarantying confidentiality \The protection of tabular data. *M SC thesis, Department of Mathematical Science of Clemson University, April 1989.*
- [4] Meyer, S., chiopu-Kratina, I.(1997): Testing CONFID suppression pattern using the midpoint strategy, *Technical Report, Statistics Canada.*
- [5] Technical report prepared by Science Applications International Corporation: Preserving confidentiality in energy publications - Introduction to the use of CONFID (1985), Modelling and Analysis Division, Norristown, PA 19403.
- [6] Robertson., D. (1993): Cell suppression at Statistics Canada, *Proceedings of the Annual Research Conference, U.S. Bureau of the Census, Washington D.C., 107-138.*
- [7] Robertson., D. (1994): Automated disclosure control at Statistics Canada, presented at the *Second International Seminar on Statistical Confidentiality, Luxembourg, November 1994.*