Strategies for Processing Tabular Data Using the G-Confid Cell Suppression Software

Jean-Louis Tambay and Jean-Marc Fillion
Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6

Abstract

Statistics Canada's G-Confid system (Statistics Canada, 2011) uses cell suppression to protect the values of sensitive (confidential) cells in tables of magnitude. It uses a heuristic approach to generate a suppression pattern that minimizes the resulting loss of data. While G-Confid achieves its primary purpose users would like it to handle additional situations such as the treatment of weighted survey data, of negative values and of waivers. Waivers are used when, in an attempt to disseminate more data, a statistical agency obtains from certain large respondents the permission to release information that may disclose their value. Some users would also like to influence the generated suppression patterns, for example to decrease the likelihood of suppressing cells that are of greater interest, or to orient the suppression towards cells of poor quality. After giving an overview of G-Confid, the paper will describe approaches that can be used to address these and other needs. Although the approaches operate within the confines of the G-Confid system they may be implemented within other cell suppression programs.

Key Words: Confidentiality, Cell Suppression

1. Introduction

Complementary cell suppression has been used for the protection of data confidentiality in tables of magnitude for several decades. Several cell suppression programs are now available. Giessing (1999, 2013) has reviewed and compared some of them including G-Confid (*aka* Confid2), a generalized system developed at Statistics Canada. G-Confid can be used to create or validate cell suppression patterns for tabular economic data at various levels of aggregation. The system consists of one SAS procedure and two SAS macros. The macros use the SAS/OR® LP solver to create and audit suppression patterns.

Cell suppression programs typically work with microdata, but they are usually not designed to work with data that can be negative or weighted, as with survey data. Some users would like to have those features, as well as other ones such as the possibility to deal with waivers, i.e., enterprises that have waived their right to confidentiality protection, and the ability to influence cell suppression patterns, say to decrease the likelihood of suppressing cells of greater interest, or to orient the suppression towards cells of poor quality. While some of these features may be available, or in the process of being implemented, in cell suppression programs this paper proposes alternate ways to address those needs, at least partially. Although the solutions offered exploit present features of G-Confid, some of them may be implementable in the other cell suppression programs.

The following section describes the methodology used for the protection of tables by cell suppression. Section 3 gives an overview of the G-Confid system, focussing more on aspects that are relevant to the problems presented and their proposed solutions. Strategies for handling special situations are given in section 4. These include the processing of weighted survey data, the treatment of waivers, the treatment of negative values, the influencing of suppression patterns and the protection of statistically small populations. Concluding remarks are given in section 5.

2. Cell Suppression

The methodology used by many cell suppression systems dates back to the 1970's (Cox and Sande, 1979). Its two main components are the identification of sensitive (i.e., confidential) cells and the generation of a suppression pattern to protect them. Sensitive cells are identified using a sensitivity measure. Different measures are available, which are particular forms of the following formula:

$$S = \Sigma_i \, a_i x_i \,,$$

where S is the cell's sensitivity measure,
a_i are fixed coefficients (with $a_1 \geq a_2 \geq \ldots \geq a_r$, and usually $a_i = -1$ for $i > 3$), and
x_i are the ordered values of the r contributors to the cell ($x_1 \geq x_2 \geq \ldots \geq x_r \geq 0$).

A cell is sensitive if $S>0$. A good sensitivity rule is the p-percent rule, which deals with the worst case scenario of an attempt by the second-largest contributor to estimate the largest contributor's value (i.e., x_1). The rule makes a cell sensitive if the value of the smallest contributors, starting from the third-largest, is less than p% of x_1. Letting T denote the cell total value, the sensitivity measure corresponding to the p-percent rule is $S = p\% \, x_1 - x_{3+}$, where $x_{3+} = \Sigma_{i>2} \, x_i = T - x_1 - x_2$. With this rule $p\% \, x_1$ represents the amount of protection sought for the largest contributor; x_{3+} represents the amount of protection (from the second largest contributor) that is provided to x_1 from within the cell; and S, if positive, represents the amount of protection that x_1 still needs to get from other cells.

Sometimes the contribution of smaller units in the cell is represented by an undifferentiated anonymous mass. This could happen, for example, when sampling or modeling is used to estimate for those units and a single aggregate is used to represent the contribution from the sample or model. We represent this aggregate value as x_{anon} and the sensitivity measure becomes a function of the identifiable and anonymous unit values, i.e., $S = (\Sigma_i \, a_i x_i) - x_{anon}$.

The second component of the cell suppression methodology, complementary suppression, involves the generation of a suppression pattern to ensure that the value of each sensitive cell c, T_c, cannot be estimated within certain bounds (e.g., within $\pm S_c/2$). In programs like G-Confid this complementary suppression is done by moving the sensitive cell by the protection value and identifying a set of other cells to move to restore table additivity. This is done under the constraint that no cell's value can be moved beyond a certain point, which is usually 50% of its total value T. All moved cells get suppressed. G-Confid uses linear optimization with a cost function to minimize the total amount of suppression in the table. The suppression cost for each non sensitive cell is usually expressed as a

function of that cell's total value. The tables below present two possible suppression patterns obtained when sensitive cell 22 is moved by one-half of its sensitivity value ($S_{22}/2$). Cells in bold are suppressed. The table on the left may have been produced by an attempt to minimize the number of suppressions while the table on the right may have been produced by an attempt to minimize the total value suppressed.

Tables 1(a) & (b): Two possible suppression patterns for a sensitive cell (22)

T_{11}	T_{12}	T_{13}	T_{14}	T_{1+}
T_{21}	$\boldsymbol{T_{22}+.5S_{22}}$	$\boldsymbol{T_{23}-.5S_{22}}$	T_{24}	T_{2+}
T_{31}	T_{32}	T_{33}	T_{34}	T_{3+}
T_{41}	$\boldsymbol{T_{42}-.5S_{22}}$	$\boldsymbol{T_{43}+.5S_{22}}$	T_{44}	T_{4+}
T_{51}	T_{52}	T_{53}	T_{54}	T_{5+}
T_{+1}	T_{+2}	T_{+3}	T_{+4}	T_{++}

T_{11}	T_{12}	T_{13}	T_{14}	T_{1+}
T_{21}	$\boldsymbol{T_{22}+.5S_{22}}$	T_{23}	$\boldsymbol{T_{24}-.5S_{22}}$	T_{2+}
T_{31}	T_{32}	T_{33}	T_{34}	T_{3+}
T_{41}	$\boldsymbol{T_{42}-.5S_{22}}$	$\boldsymbol{T_{43}+.3S_{22}}$	$\boldsymbol{T_{43}+.2S_{22}}$	T_{4+}
T_{51}	T_{52}	$\boldsymbol{T_{53}-.3S_{22}}$	$\boldsymbol{T_{54}+.3S_{22}}$	T_{5+}
T_{+1}	T_{+2}	T_{+3}	T_{+4}	T_{++}

Note that, with a the *p*-percent rule, just as x_{3+} can represent an amount of internal "noise" or protection provided to the largest contributor within the cell, the values T for other suppressed cells can represent the maximum amount of external "noise" or protection that they provide to that contributor.

3. Overview of G-Confid

To protect a table of magnitude G-Confid needs to work with micro level data. The program consists of three SAS modules. Proc SENSITIVITY reads the table's microdata, defines the structure of the table, and calculates cell total values and their sensitivities. Macro %SUPPRESS carries out complementary suppression using SAS/OR® linear optimization. An optional third module, Macro %AUDIT, serves to validate a cell suppression pattern.

The main inputs to Proc SENSITIVITY are a microdata file, the definitions of hierarchies for each classification variable in the table (e.g., row and column variables), and the chosen sensitivity measure (e.g., *p*-percent rule with p=20). Each microdata record represents one unit (e.g., enterprise) and contains the unit's identifier, its values for each table classification variable (e.g., province, industry) and its value for the table's magnitude variable (e.g., revenues). Optionally, a shadow variable can be added to each microdata record. While G-Confid does not process shadow variables, it provides cell totals for both the magnitude variable and the shadow variable. A shadow variable can be useful when the variable of interest cannot be processed as is (e.g., it needs to be transformed, or the suppression pattern for the variable of interest is generated using a related "representative" variable instead) but users still want to have cell totals for the variable of interest along with those for the processed variable. Other uses for a shadow variable will be given later.

G-Confid does not process survey weights (yet) but setting a unit identifier to blank means that the record's value for the magnitude (and shadow) variable represents an anonymous contribution. That is, records with a blank unit identifier do not require protection, but they contribute to x_{anon} for their cell and can serve to protect other units.

Along with summary results Proc SENSITIVITY produces a cell level file and an equations file. Each record in the cell level file contains the cell identifier (cellid), its total value for the magnitude variable (and the shadow variable, if requested), its sensitivity and the cell status (sensitive or not). There is also a flag to indicate if the record represents a table cell or an aggregate – but this flag is outside the scope of this paper and will be ignored. The equations file defines the structure of the table and is also outside the scope of this paper.

Macro %SUPPRESS generates the cell suppression pattern for the table. Its inputs are the cell level file and the equations file from Proc SENSITIVITY (in their original or modified form), the chosen cost function(s) and, optionally, the names of variables to which the cost functions are applied. Unlike other programs, G-Confid can carry out complementary cell suppression using a cost variable other than the table's magnitude variable. This could be the shadow variable or any other variable that the user has added to the cell level file. Along with summary results Macro %SUPPRESS produces another cell level file with variables added (e.g., the final cell suppression status) and, if requested, a complements file. The complements file identifies, for each sensitive cell, which cells were used to protect it (i.e., which cells were moved when establishing a pattern as in tables 1(a) or 1(b)).

Users can influence the cell suppression pattern in %Macro SUPPRESS globally, by changing the cost variables and/or functions, or locally, by setting the status of nonsensitive cells in the input cell level file to *published* or *suppressed*. Published cells cannot be chosen as complementary suppressions. Suppressed cells carry a zero suppression cost, which makes them more likely to be used as complementary suppressions.

4. Strategies for Handling Special Situations

4.1 Processing Weighted Survey Data

When dealing with survey data survey weights are typically used to represent nonsampled and/or nonresponding units in the population. For a cell with r respondents, the estimated population total becomes $\hat{T} = w_1 x_1 + w_2 x_2 + \ldots + w_r x_r$ where the w_i (≥ 1) are the respondent survey weights. With $N–r$ other population units in a cell the sensitivity measure under a p-percent rule with p=20 should be

$$S = 0.2\, x_1 - x_3 - \ldots - x_r - (x_{r+1} + x_{r+2} + \ldots + x_N),$$

where the portion in parentheses represents the unknown but estimated contribution from those $N–r$ units. Because we do not know or use these units' values they do not need protection, but their estimated contribution can protect the values of responding units.

Instead of using weighted data, G-Confid uses anonymous respondents and sensitivity measure

$$S = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_r x_r - (x_{anon})$$

e.g.,

$$S = 0.2\, x_1 - x_3 - \ldots - x_r - (x_{anon}).$$

So when dealing with weighted data we can let the respondents represent themselves and put the residual contribution from their weights in x_{anon}, that is, set $x_{anon} = (w_1\text{-}1)x_1+(w_2\text{-}1)x_2+\ldots+(w_r\text{-}1)x_r$. We note that with a *p*-percent rule if the largest respondent has weight $w_1 \geq (100+p)\%$ we get $S \leq 0$, so with p=20 as long as $w_1 \geq 1.2$ the cell will not be sensitive.

Sometimes this is not considered to be a sufficient level of protection. Usually weights that are slightly above 1 are more this way due to nonresponse adjustment or calibration than to sampling, so the identity of respondents may be known to some other individuals. One practice used at Statistics Canada consists of protecting the unweighted contribution of respondents whose weights are below 3, while puttting the weighted contribution of respondents with weights of at least 3 in the anonymous portion. So on the microdata file we use x_i and keep the unit identifier when $w_i<3$, and replace x_i by $w_i x_i$ and blank out the identifier otherwise (which puts $w_i x_i$ in x_{anon}). A positive feature of this practice is that a cell with only one or two respondents will still be sensitive if their weights are lower than 3. A less desirable feature is the fact that when weights are below 3 the cell total values are not preserved in G-Confid. Another undesirable feature is that there is a jump in the protection requirement for x_1 when w_1 reaches the threshold value of 3 (e.g., with a *p*-percent rule with p=20 the needed protection goes from $0.2x_1$ to 0 when w_1 reaches 3).

One alternative, when a *p*-percent rule is used, is to replace values for the largest respondent in each inside cell, x_1, by $x_1^* = \delta_1 x_1$ if $1<w_1<3$, where $\delta_1 = (w_1 + \alpha_1)/(1+\alpha_1)$ and to add the difference $(w_1 - \delta_1) x_1$ to the anonymous portion (x_{anon}). This is mathematically equivalent, when w_1 is small, to using sensitivity measure

$$S = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_r x_r - \{(w_2 - 1)x_2 + \ldots + (w_r - 1)x_r\}$$

instead of

$$S = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_r x_r - \{(w_1 - 1)x_1 + (w_2 - 1)x_2 + \ldots + (w_r - 1)x_r\}.$$

In other words, the weight of the largest respondent, if below 3, is not used to protect it. When $w_1 \geq 3$ the cell is not sensitive and we may put w_1x_1 in the anonymous portion or, for reasons explained below, keep x_1 as is and put $(w_1-1) x_1$ in the anonymous portion. Cell total values will be preserved, but there will still be a jump in the protection requirement for x_1 when w_1 reaches value 3.

Another alternative, which eliminates the jump in the protection requirement for x_1 when w_1 reaches value 3, consists of replacing x_1 by $x_1^* = (w_1 - \delta_1) x_1$ when $1<w_1<3$ and putting $\delta_1 x_1$ in the anonymous portion, where $\delta_1 = 3\alpha_1(w_1 - 1)/\{2(1+\alpha_1)\}$. When $w_1 = 1$, $\delta_1 = 0$ and $x_1^* = x_1$. When $w_1 = 3$, $\delta_1 = 3\alpha_1/(1+\alpha_1)$ and $x_1^* = 3x_1/(1+\alpha_1)$, which has the effect of eliminating the contribution of x_1 from the sensitivity measure since we then have

$$S = \alpha_2 x_2 + \ldots + \alpha_r x_r - \{(w_2 - 1)x_2 + \ldots + (w_r - 1)x_r\}.$$

With a *p*-percent rule $\alpha_2 = 0$ and $\alpha_i = -1$ for $i>2$, which makes $S \leq 0$, so the cell sensitivity disappears as w_1 reaches 3.

The residual contribution from weights for other units, i.e., $(w_i - 1) x_i$, can be put in the anonymous portion. Usually we only need to do this for x_2, and can put all of $w_i x_i$ for other units in x_{anon}. An exception may be when the same enterprises can contribute to more than one cell in the table – in which case we may not want to anonymize the reported x_i values, just the residual contribution from their weights. This will allow G-

Confid to merge unweighted reported values for the same enterprise when processing aggregate cells (there will be an impact due to the use of δ_1).

Weights can lie below 1 due to calibration or some other processing step. For example, weights can be used to allocate an enterprise's values for survey variables among its provincial establishments according to their share of the enterprise's total revenue. When weights are below 1 it may be sufficient to use $w_i x_i$ instead of the reported values in the input microdata file.

4.2 Treatment of Waivers

In running a cell suppression program for a table some sensitive cells can generate a disproportionately large number of complementary suppressions. This can be due to exceptionally high sensitivity values for those cells and/or to the sparseness of the table around them. If the sensitive cells are dominated by 1 or 2 enterprises, say, we may try to obtain waivers from those enterprises in an attempt to reduce the number of complementary suppressions in the table. A waiver is an agreement where the respondent (enterprise) gives consent to a statistical agency to release their individual information. Since waivers are often difficult to obtain from respondents, and they cause operational burden on the part of the statistical agency, waiver candidates must be carefully selected. We can look for waivers in cells that generate many complementary suppressions (as identified by G-Confid) or we can use methods that assign scores to candidate cells, as in Provençal *et al*. (2004) and seek waivers from the highest scoring cells.

Assuming that we have obtained waivers from respondents, we may wish to process those respondents so that: (a) they do not require protection in their cell and (b) they are not used to protect non-waivers in their cell or in other cells. The former means that the cell sensitivity measure should not target the protection of those respondents' values. The latter means that we should treat the values of waivers as public knowledge (this can often be the case, although the point may be subject to debate). We can meet (a) by setting those respondents to *anonymous*, i.e., blank out their identifiers in the microdata file, but this would violate (b). Instead, a proposed solution is to replace the respondent values by zero in the microdata file. Their cell's sensitivity will be calculated based on the largest non-waiver, if any. If the cell is still sensitive then more waivers may be needed. This would be the case with a two-respondent cell. For cells with more respondents there may be a domino effect on the need for waivers, which could indicate that the cell was not such a good candidate for waivers. Setting waiver values to zero is equivalent to ignoring them in calculating cell sensitivities, and those respondents will not be used to protect other units or cells. In G-Confid the original respondent values can be stored in the shadow variable, so that users can still see the true cell totals.

It should be cautioned that setting the value for waivers to zero may cause some aggregate cells to become sensitive. For example, suppose that we have 5 records for the 3 cells in a given industry: {Unit 1, East, $500}, {Unit 2, Central, $500}, {Unit 3, Central, $50}, {Unit 4, Central, $35} and {Anonymous, West, $20}. According to a *p*-percent rule with p=20 the cells for East and Central are both sensitive with sensitivity values 100 and 65, respectively. The all-region marginal total cell is not sensitive, however. Now if we obtained a waiver for Unit 1 in the East, say to avoid suppressing other industries in the East, and changed its value to 0 then the marginal total cell would become sensitive with a sensitivity value of 45. In general it is preferable not to have sensitive marginal total cells.

4.3 Treatment of Negative Values

The cell suppression methodology works on the assumption that all data are nonnegative. This does not fit all economic data, so heuristic approaches have been proposed for the treatment of tables containing negative data (see, for example, Giessing (2008)). The approaches proposed are easy to use with standard cell suppression programs, or are incorporated in such programs. We review some of them and introduce a new one that is based on the application of cell suppression to a nonnegative derived variable that is related to the real-valued variable of interest.

For some variables, like net income or net change, negative values can occur regularly while for others, like shipments, they occur exceptionally. In both situations the presence of negative values raises a number of issues for cell suppression. First of all, how do we determine which units need the most protection? Sensitivity measures protect the most vulnerable contributor in a cell, which is typically the respondent with the largest value (x_1). But with negative data is the most vulnerable contributor the respondent with the largest value or the largest absolute value ($\max_i \{|x_i|\}$); or is it the "largest" respondent based on a nonnegative size variable like total assets or revenues?

Also, if respondent i is identified as needing the most protection, how do we determine the amount of protection that it needs? Under the p-percent rule for nonnegative values it is $p\%\ x_1$. Would $p\%|x_i|$ be sufficient if x_i happened to be very close to zero (e.g., a huge enterprise with near-zero profits)? And how do we determine the amount of protection ("noise") offered to it? The noise from small respondents in its cell can be much larger than $|x_{3+}|$. Likewise, with negative data, the protection from other suppressed cells can be much greater than T. Finally, what would be a suitable sensitivity measure in the presence of negative values? Sensitivity measures used in cell suppression algorithms have two properties. The subadditivity property states that the sensitivity of a union of cells cannot be greater than the sum of the sensitivities of its components (e.g., $S_{a\cup b} \leq S_a + S_b$). The other property puts a limit to how much a cell b can protect a sensitive cell a ($S_{a\cup b} \geq S_a - T_b$). A \$500 cell alone cannot protect a cell with sensitivity 1000. These properties may not hold with negative data.

The problem gets worse when the variable represents the difference between two nonnegative variables that are also being published. We could suppress the variable's cell when either of the nonnegative variable cell gets suppressed. But what if both those cells get suppressed? Or what if they are both nonsensitive but only one respondent has a nonzero difference? Proper protection would require the three variables to be processed jointly. More on this in subsection 4.3.6.

A list of approaches is given, ending with a proposed new one. Many process a modified version of the variable of interest, or a related variable. In G-Confid the original variable of interest can become the shadow variable (it is allowed to have negative values).

4.3.1 Use another suppression criterion

If values can be both positive and negative it may be argued that, since they are essentially unrestricted, it suffices to adopt a rule based on the number of respondents in a cell. The CENEX SDC-handbook (Hundepool, *et al*., 2007) suggests relaxing the parameters of the disclosure rule when variables can be negative but no cell total is negative. This is motivated by the idea that, when data can be negative, uncertainty about respondents' values can extend beyond 100%. With the *pq*-rule, a variant of the *p*-percent

rule, as the uncertainty goes to infinity we get closer to a 3-respondent minimum rule, which the handbook also proposes. An issue would be how to carry out residual suppression with such a rule. One option, implemented in G-Confid, is to put a default sensitivity value of 1 for such cells.

4.3.2 Use the suppression pattern for a related nonegative variable

This is a common approach when a size variable like total revenue is available. The approach may be suitable in certain situations. Surveys sometimes apply the tabular suppression pattern for a "representative" variable to tables for related characteristics. The approach is recognized by G-Confid and τ-Argus (Hundepool *et al*., 2009), for example.

4.3.3 Replace negative values by zeroes

This may be acceptable if negative values are rare and not particularly large. Examples could be the occasional negative shipments or revenues.

4.3.4 Replace negative values by their absolute value

This option is also common. It is discussed in some detail in Daalmans and de Waal (2010). It may be acceptable if one thinks of the absolute value for a respondent as indicative of the level of protection that it needs as well as of the level of protective noise that it can offer to others (but then what do values $T = \Sigma |x_i|$ represent: cumulated noise?) That option is less attractive when the absolute value is not too indicative of the information that is of interest. For example, if the variable of interest is profits then the fact that a respondent with 6 millions in revenues has generated profits of only 32,000 makes the latter figure inadequate as an indicator of the amount of protection required or provided. Under a *p*-percent rule with p=20 we would try to protect its profits by only 6,400 whereas the amount of protection sought for its revenues would be 1.2 millions.

Sometimes an enterprise may be contributing to several cells, say for different provinces. If positive and negative contributions are present then as its results get aggregated to the national level the sum of its absolute provincial values would be higher than its absolute national value. But this is not necessarily a problem because the presence of positive and negative values means that, nationally, the protection needed and/or offered by its profits are higher than what is indicated by the magnitude of profits.

4.3.5 Add a constant to all values that is large enough to make them all positive

The constant should be greater than the absolute of the lowest negative value. Standard disclosure protection can be applied to the modified values. If the added constant is very large we may end up with the equivalent of a minimum-respondent rule in most cells; cells with more respondents may rarely become sensitive. Sometimes adding a constant can make safe cells become sensitive (e.g., after adding 100, responses 300, 100 and -100 become 400, 200 and 0, which are sensitive). For data series the value of the constant should be large enough to allow it to be stable over time. If the constant changes significantly every period we will have an undesirable situation that is similar to changing the disclosure rule every year.

A constant can be added to the cell values instead of to individual values. The τ-Argus system's modular approach breaks down tables with a certain structure into subtables that are processed separately. Its user manual notes "If in a subtable during the process negative values are found, all cell values are increased such that the lowest value

becomes positive. Of course the margins have to be recalculated, but a safe protection pattern can be found." (Hundepool, *et al*., 2009).

4.3.6 Processing the variable as the difference of two nonnegative variables

Most variables that can be negative are differences of two nonnegative variables that are also published. So one possibility would be to base the cell suppression pattern for the variable of interest on the suppression patterns for the two related nonnegative variables. In appendix A of WP22 (Federal Committee on Statistical Methodology, 2005) it is suggested, for variables representing a net change over two periods, to suppress the change if either period's value is suppressed. For variables that are differences of two nonnegative variables (e.g., revenues and expenses), it is suggested to use the suppression pattern for the variable that is generally higher, if such is the case. If neither variable dominates the other two, suggestions offered are to use a minimum-respondent rule or to process the absolute value of the variable of interest.

Suppressing the variable of interest whenever either of its contributing nonnegative variables is suppressed may lead to oversuppression – especially if residual suppressions for the two variables are determined independently. One way to reduce oversuppression is to better align the suppression patterns for the two variables. In G-Confid and τ-Argus a cost variable can be used to influence cell suppression patterns. After applying cell suppression to one of the variables, the suppression pattern for the second variable can be generated with a lower suppression cost given to cells suppressed for the first variable.

Note that it may not always be necessary to suppress the variable of interest when both nonnegative variables are suppressed – but it is not always harmless to publish it either. And when a direct relationship exists between the three variables, such as $a–b=c$, publishing any two of the variables for a cell is equivalent to publishing all three.

4.3.7 Integrate the variable of interest with a nonnegative size variable

We revisit the cell suppression problem for variables that can take on negative values by asking a basic question – what do we want? The answer is, three things: to be able to determine a protection level for respondents, to be able to determine units' contribution to "noise" and to have suitable sensitivity measure. Let's start with the first point.

Under a *p*-percent rule, the level of protection sought when the values x_i are nonnegative is $p\%\, x_i$. When x_i can be negative $p\%\, |x_i|$ could work in many cases, but we need something for when a huge unit has very small $|x_i|$ and $p\%\, |x_i|$ is too small as an indicator of the protection needed. We also need something that works with one or two respondents only. In one survey of manufacturers, 70% of the sensitive cells had only one or two respondents. A reasonable approach would be to use a nonnegative proxy variable for x_i, such as $z_i = \max\{|x_i|\, , \alpha\, y_i\}$ or $z_i = |x_i| + \alpha\, y_i$, where y_i is a nonnegative size variable like revenues and constant α serves to cover cases where $|x_i|$ is very small. The protection level sought would be $p\%\, z_i$. As a safety threshold we can use a rough, small, value for α. The use of a proxy variable z_i addresses the related question – how to determine who needs the most protection? It would be the unit with the highest z_i.

Turning to the second point, other respondents and cells contribute "noise" to protect the value of a unit at risk of disclosure. The noise is related to the respondents' values, but is usually not more than ±100% of the values when these are nonnegative. When values x_i can be negative the relative span of possible values for units (x_i) and cells (T) is wider

than ±100%. Deviations like $|x_i - x_{median}|$ could be used as measures of the noise associated with values x_i. But it would be difficult to process such data since the median value of x_i would be cell-specific. We would also prefer to have some consistency between the way internal (x_{3+}) and external (T) noises are calculated. And we would prefer to use the same proxy for protection sought for x_1 and protection offered by x_{3+}. It turns out that the proxy variable z_i suggested above can meet those needs reasonably well. Requirements for α for noise could be different from those for protection (e.g., for noise one could use the first quartile or decile of $|x_i|/y_i$ across the entire population).

For the third point it is suggested that using values z_i in the sensitivity measure S also would be reasonable.

If we were to use some proxy variable z_i instead of x_i we would need to determine what form of z_i we wished to use, what to use for y_i and how to determine α. Other forms of z_i could be considered, such as $|x_i|+\alpha y_i^{½}$. And we may decide not to generate z_i's if there is no x_i (e.g., if x_i = net exports and y_i = total sales, do not generate z_i's for enterprises without exports). When x_i is the difference of two nonnegative variables, say $x_i = u_i - v_i$, y_i could be u_i , v_i or a combination such as $(u_i + v_i)/2$ or $\max\{u_i, v_i\}$. The value of α could be determined by examining the relationship between x_i and y_i at some global level. Inasmuch as proxy variable z_i essentially serves to come up with a suppression pattern, choices can be influenced by practical considerations.

Figure 1 shows scatterplots of the ratio |profits|/sales against sqrt(sales) from a dataset of 834 companies (Brand, Domingo-Ferrer and Mateo-Sanz, 2002). The square root was used to compress the image. In searching for a value for α we note that 45% of the observations have a ratio below 0.015. This number is very low and may be suitable.

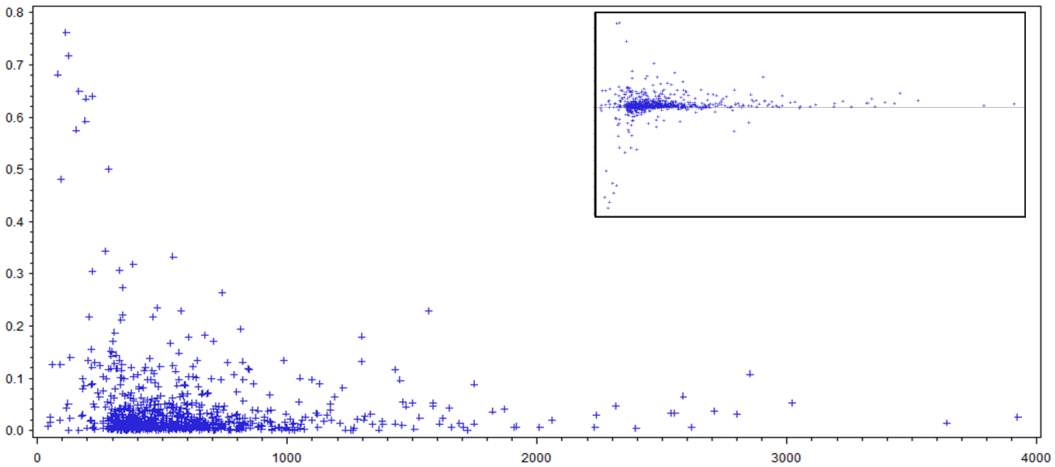

Figure 1: Scatterplot of ratios |profits|/sales *vs* sqrt(sales) – truncated at 0.8 (10 obs.). Inset, ratios profits/sales *vs* sqrt(sales). Source: 1995 Tarragona Dataset.

4.4 Influencing the Suppression Pattern

For a number of reasons users may wish to influence or modify the suppression pattern produced by a program like G-Confid. In Macro %SUPPRESS the publication of nonsensitive cells of importance to their region or industry can be assured by giving these cells a status of "Published" in the input cell level file. Published cells are ineligible for complementary suppression. But imposing a status of Published for some cells may result in an infeasible solution if the cells are absolutely necessary to protect a sensitive cell.

To avoid infeasible solutions, we could give such cells a very high suppression cost, which would still make them eligible for suppression if there were no alternatives. Macro %SUPPRESS allows the use of variables other than the magnitude variable in the cost function. A modified version of the magnitude variable, with much higher values for cells of interest to users, could be added to the input cell level file and used by the cost function. Thus higher suppression costs can be given to cells of importance to their region or industry. Alternatively, values for the cost variable could be reduced for cells that are less important, such as those of poor quality (high coefficient of variation). The cost variable could also serve to influence the suppression pattern for greater consistency over time. The cost could be decreased for cells that were suppressed in previous periods.

4.5 Protecting Statistically Small Populations

Some statistically small populations (e.g., small specialized industries or remote regions) have very few units in them. Their cells tend to be sensitive, often because they have very few respondents. The problem is that these small populations, although statistically insignificant overall, may cause complementary suppressions in other industries or regions. While their cells are often sensitive and suppressed, we would like to limit the impact that they have on the release of cell data for other populations (the next larger region may get an undue number of suppressions because of them). It may be desirable to have the equivalent of waivers for them.

A possible solution would be to apply a perturbative method to these populations' data so that the perturbed cell totals are considered safe and not in need of protection. For example, additive noise (Evans, Zayatz and Slanta, 1998) could be applied to their microdata. Since the noise is focused on one dimension (e.g., one region) of the table it can be applied in a balanced way so that, even though the noise level for some industries in the region may be higher than acceptable for quality and publication purposes, the overall noise level for the region is acceptable. Since these populations have a minor impact on the overall population total, so will their noise (even more so).

We can run Proc SENSITIVITY with respondents' microdata in those populations replaced by perturbed microdata, and enterprise identifiers set to blank (anonymous) so that their cells cannot become sensitive. If we wish, we can use the cost variable in Macro %SUPPRESS to assign a low suppression cost to the perturbed cells, or even set the status of overly perturbed cells to suppressed. In G-Confid cells with a status of suppressed do not require complementary suppression if they are not sensitive. We tried this method with monthly employment data from the Survey of Employment, Payrolls & Hours, focussing our attention on cells with small sensitivity values instead of cells for small populations (it is another worthy objective). We used a two-dimensional table crossing industry (with up to 6 hierarchical levels) with geography (2 levels), giving 5270 cells. Running G-Confid with a *p*-percent rule with p=10 gave 1349 suppressions: 850 sensitive cells and 499 complementary suppressions. Suppressions accounted for 25.6% of the total employment value for all the cells (including the marginal totals cells). We then assumed that we perturbed 182 sensitive cells. Those were cells at level 5 or 6 in the industry hierarchy and level 2 in the geography hierarchy whose sensitivity value was less than 0.2% of the employment marginal subtotals (i.e., at the next higher level) for both their industry and geography. Making the 182 cells nonsensitive made the total number of suppressions go down to 1246 (23.6% of total employment) including 659 sensitive cells, 111 user-suppressed cells and 476 complementary suppressions. The 111 user-suppressed cells were those cells among the 182 that were never used as complementary suppressions. If we chose to publish those "perturbed" cells the total

number of suppressions would have gone down from 1349 to 1135 (from 25.6% of employment to 21.5%).

5. Conclusion

In this paper we propose ways to handle five situations in tabular cell suppression that are not specifically addressed by the current version of G-Confid. The solutions exploit features of G-Confid that may or may not be present in other cell suppression programs. If not, the article may serve to demonstrate the benefits of incorporating some of those features in the other programs. The solutions proposed are heuristic and may not always be suitable for the purpose, and better ones may be available. It is ultimately the user's responsibility to determine if they will work for them. The article also promotes the idea that it is not always necessary for a generalized system, such as G-Confid, to handle all situations, but that the design should allow some flexibility so that it can be adapted for other purposes.

References

Brand, R., Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002). Reference data sets to test and compare SDC methods for protection of numerical microdata. Unscheduled deliverable, *Computational Aspects of Statistical Confidentiality Project*. (http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf)

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

Daalmans, J. and de Waal, T. (2010). A general formulation of the secondary cell suppression problem. Discussion paper (10009), Statistics Netherlands.

Evans, T., Zayatz, L. and Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics*, **14**, 537–551.

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. (1999). A Survey on Software Packages for Automated Secondary Cell Suppression. *Presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki, Greece, 8-10 March 1999.

Giessing, S. (2008). Protection of tables with negative values. ESSNet SDC Document. (http://neon.vb.cbs.nl/casc/ESSnet/PosNegReport.pdf)

Giessing, S. (2013). Software tools for assessing disclosure risk and producing lower risk tabular data. *Data Without Boundaries Deliverable 11.1 – Part B*, February 2013. (http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d11-1b_software-tools-disclosure-risk-assessment.pdf).

Hundepool, A. *et al*. (2009). *τ-ARGUS Version 3.3 (revised) – User's Manual*. Statistics Netherlands.

Hundepool, A. Domingo-Ferrer, J., Franconi, L. Giessing, S. Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. and De Wolf, P.P. (2007). *Handbook on Disclosure Control*, Version 1.01, CENEX SDC.

Provençal, J.S., Bérard, H., Fillion, J.M. and Tambay, J.L. (2004). Approaches to Identify the Amount of Publishable Information in Business Surveys through Waivers. *Privacy in Statistical Databases: CASC Project International Workshop, PSD*.

Statistics Canada. (2011). *G-Confid User Guide*. Internal Report.