Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (vii): Software developments and demonstrations

G-Confid: Turning the tables on disclosure risk

Prepared by Peter Wright, Statistics Canada

G-Confid: Turning the tables on disclosure risk

Peter Wright

Statistics Canada, Ottawa, ON, Canada, peter.wright@statcan.gc.ca

Abstract: Under the *Statistics Act*, Statistics Canada must protect respondents' confidential data. Cell suppression is a technique used to protect tabular data. The automated disclosure control software G-Confid, developed at Statistics Canada, is used to implement this technique. This software is user-friendly, has the same look and feel as Statistics Canada's other generalized systems and can incorporate new approaches. It can also be used to deal with potentially voluminous multi-dimensional tables. The main objective of G-Confid is to provide the appropriate level of protection for confidential cells while minimizing the loss of information. To achieve this objective, G-Confid uses an automated linear programming procedure to optimize complementary cell suppression. This article provides an overview of the functionality and characteristics of G-Confid. The emphasis is on newer features as well as techniques to improve the analysis of disclosure risk using G-Confid.

Key Words: Confidentiality; Complementary cell suppression.

1 Introduction

G-Confid is a generalized system developed at Statistics Canada that protects the confidentiality of tabular economic data. Users of G-Confid can specify the levels of aggregation at which G-Confid creates or validates cell suppression patterns. G-Confid identifies the cells (usually estimates of totals) of tables requiring primary suppression, and then optimizes the selection of cells for complementary cell suppression. G-Confid is capable of processing large, multi-dimensional tables as well as hierarchies that involve multiple decompositions.

G-Confid functions within a SAS environment. The SAS macros that implement and assess complementary cell suppression use the SAS/OR® LP solver, and are limited only by SAS and hardware restrictions. Current versions of G-Confid are available for SAS 9.2 and 9.3. Separate versions are available for use with 32-bit and 64-bit processors. G-Confid is supported on Windows XP, Windows Vista and Windows 7. Also available are both a Unix (AIX) version and a Linux version. G-Confid also has a graphical user interface for users to run the system with SAS Enterprise Guide, version 4.3 or 5.1.

2 Overview

2.1 General description of G-Confid

G-Confid features a suite of three SAS components for use with tabular economic data at various levels of aggregation. PROC SENSITIVITY identifies cells requiring primary suppression. The macro SUPPRESS protects the cells identified by PROC SENSITIVITY by selecting an optimal set of cells for complementary suppression. The macro AUDIT validates a suppression pattern

that was not provided by the macro SUPPRESS or which the G-Confid user altered after running the macro SUPPRESS. Two additional macros provide further information to G-Confid users, although not needed in the cell suppression process. The auxiliary macro AGGREGATE provides further information about sensitive unions of cells. The auxiliary macro REPORTCELLS provides a visual snapshot of the suppression pattern to facilitate the creation of output tables of the economic data under study.

2 Methodology underlying G-Confid

The three main components of G-Confid, PROC SENSITIVITY, the macro SUPPRESS and the macro AUDIT, correspond to the three activities that protect tabular economic data: identify sensitive cells, identify cells for complementary cell suppression, and audit the suppressions patterns to check for exact or partial disclosure. G-Confid uses the cell suppression methodology developed in the 1970s (see Cox and Sande (1979), and also Robertson and Şchiopu-Kratina (1997)).

2.1 Cell sensitivity

The main objective of primary suppression is to identify and suppress sensitive cells. A cell is sensitive if its total value allows the close estimation of the contribution of some of its respondents. Such cells are identified using a sensitivity measure. Different measures are available, and which are particular forms of the following subadditive formula:

$S = \Sigma_i \, a_i \, x_i$, where $a_1 \geq a_2 \geq \ldots \geq a_r \geq -1$

where S is the cell sensitivity (a cell is sensitive if $S > 0$),

a_i are fixed coefficients (usually $a_i = -1$ for $i > f$ where f is 1, 2 or 3),

x_i are the ordered values of the r contributors to the cell ($x_1 \geq x_2 \geq \ldots \geq x_r \geq 0$).

Using PROC SENSITIVITY, users can select the p-percent rule or the (n,k) rule, or create a customized rule by setting the values of the a_i coefficients.

2.2 Complementary suppression

The main objective of complementary suppression is to identify complementary cells that will also be suppressed to protect sensitive cells, and to do so in a way that minimizes the resulting loss of information. Complementary suppression involves representing a table by using a set of equations that establish the relationship between cells, e.g., between cells in a row (or column) and their corresponding total cell. Each cell i is actually represented by two variables y_i and z_i corresponding to a positive and a negative change in its value, respectively. The sensitive cell

sen is moved by one half of its sensitivity S_{sen} and values y_i or z_i of other cells are moved to restore table additivity.

The matrix formulation of the linear programming (LP) problem is:

Minimize: $\boldsymbol{w^T y + w^T z}$ (objective function)

Subject to: $\boldsymbol{Cy - Cz = 0}$ (equations defining relationships in the table)

$\mathbf{0} \leq \boldsymbol{y, z} \leq \boldsymbol{t}/2$ (bounds on the movements of cells)

$y_{sen} \geq S_{sen}/2$ (sensitive cell *sen* moved by $S_{sen}/2$)

$z_{sen} = 0$

Vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ represent positive and negative changes in cell values, respectively, $\boldsymbol{w}$ is a fixed cost vector (w_i =0 if cell i is sensitive or was already suppressed), $\boldsymbol{t}$ is the vector of cell total values t_i and $\mathbf{C}$ is a matrix of coefficients (0, 1 or –1) that represents the relations between the cells in the table. In solving the LP problem any cell i that was moved (has y_i >0 or z_i >0) is a complementary cell and gets suppressed.

G-Confid protects sensitive cells one at a time, usually starting with the cell i with the highest sensitivity value S_i. G-Confid identifies complementary suppressions for each sensitive cell and chooses cells for complementary suppression, typically by using a cost function that favours cells with small values. Using a sequential approach in this manner may lead to a suboptimal solution in terms of the overall LP objective. For that reason G-Confid undertakes a second step, selecting complements from among the set that were already selected in the first step, and using a different cost function that favours the largest cells as complements. In doing so the second step identifies cells that were labelled as complements during the first step but are not sensitive or needed after all to protect other sensitive cells. Compared to using the first step alone, the two-step approach leads to publishing more cells and reducing information loss while maintaining the required level of protection.

2.3 Auditing a suppression pattern

If users modify the suppression pattern produced by the macro SUPPRESS, e.g., in order to force the publication of certain key results, they should run macro AUDIT to verify the validity of the modified suppression pattern. The macro AUDIT can also verify the level of protection offered by a suppression pattern that was produced not using the SUPPRESS macro of G-Confid. Auditing a suppression pattern involves finding maximum and minimum values of each suppressed cell *sen* subject to the values of other suppressed cells i being between predetermined bounds (e.g., between $0.5t_i$ and $1.5t_i$).

3 Main components of G-Confid

This section presents an example of each of the three main components of G-Confid. Readers may consult Tambay and Fillion (2011) for a detailed description of the syntax and options of each component.

3.1 PROC SENSITIVITY

To run G-Confid a user usually supplies four inputs to PROC SENSITIVITY:
• A microdata file
• A definition of the hierarchy(-ies) for each dimension of the table
• The ranges of codes associated with the lowest level of each hierarchy (*optional*)
• The rules used to identify sensitive cells

An example of PROC SENSITIVITY code featuring the dominance rules (n=1, k=70) and (n=2, k=80) is:

```
proc sensitivity data=microfile outconstraint=consfile
  outcell=cellfile outlargest=largestfile
  hierarchy="0 East West; 0 1 2 3;"
  srule="nk 1 70 2 80"
  range="East A B: West C D;
         1 101 201 301: 2 102 202 302: 3 103 203 303;"
minresp=5;
id Enterpriseid;
var Income;
dimension EastWest Industry;
by QuestionNumber;
run;
```

PROC SENSITIVITY processes the microdata to ensure that *false complements* are detected and assessed along one-dimensional lines (rows, columns, etc.). A false complement is a potential complement cell that seems to offer more protection to a sensitive cell that it actually offers. A common false complement situation is when two one-respondent cells are used to mutually protect each other. Their union or *aggregate*, having two respondents only, is still sensitive. Other false complement situations may occur when a complementary cell has respondents in common with the sensitive cell, or has respondents that are larger than the second respondent in the sensitive cell. PROC SENSITIVITY identifies the unions of sensitive cells, and the unions of sensitive and nonsensitive cells, that comprise *sensitive aggregates*. Once sensitive aggregates are identified using PROC SENSITIVITY they can be protected using the macro SUPPRESS, a measure that will prevent the occurrence of false complements.

3.2 Macro SUPPRESS

This macro carries out the complementary cell suppression. It requires the use of two output files from the PROC SENSITIVITY step: the cell-level data file that contains the sensitivity of cells (and includes the sensitive aggregates, if any), and the file that describes the linear constraints that relate two or more cells to a subtotal or aggregate.

Aggregated data are processed through a LP solver and using the linear constraints generated by PROC SENSITIVITY. An example of the macro SUPPRESS is:

%SUPPRESS(**InCell**=cellfile, **Constraint**=consfile,
CFunction1=size, **CFunction2**=information,
CVar1=costvar, **CVar2**=costvar2, **OutCell**=suppcell,
OutComplement=suppcomp, **ByVars**=QuestionNumber, **ScaleCost**=None);

As noted in section 2.2, to reduce the number of suppressions the LP process is run using two steps, with different objective function values (w_i). Users may specify the cost functions (CFunction1 and CFunction2) to use in each step. The cost function choices are SIZE ($w_i=t_i$), DIGITS ($w_i=\log_{10}(t_i+1)$), CONSTANT ($w_i=1$) and INFORMATION ($w_i=\log_{10}(t_i+1)/(t_i+1)$). G-Confid users are recommended to use either SIZE or DIGITS in step 1, while INFORMATION is often used in step 2 to "free up" small cells that were identified for suppression in step 1.

Users may also assign customized cost values to each cell (CVar1 and CVar2), or accept the default values which are the cell totals. Assigning customized cost values allows the G-Confid user to influence the suppression pattern, e.g., to favour the selection of complementary cells with high coefficients of variation. In step 1 of the SUPPRESS example shown above, the variable costvar will be used instead of the variable of interest (cell total) in calculating the suppression cost. The lower a cell's cost, the greater the likelihood of its being suppressed.

3.3 Macro AUDIT

This macro verifies the validity of a suppression pattern by calculating minimum and maximum values for each suppressed cell or sensitive aggregate using the LP solver. An example of the code is:

%**AUDIT(InCell**=suppcell, **Constraint**=consfile, **OutCell**=audcell, **LBFactor**=0.5,
UBFactor=1.5);

The purpose of the AUDIT macro is to validate the suppression pattern and ensure that each sensitive cell is protected. If not, the AUDIT macro indicates whether someone who uses the resulting table can identify the exact value of the sensitive cell (exact disclosure) or can calculate too precisely the interval of possible values (unachieved protection).

4 Improving the analysis of disclosure risk using G-Confid

In this section we identify some current challenges faced by analysts of disclosure risk, and explain how to use G-Confid to overcome them. This section also features recent and upcoming improvements to help users examine and apply the results.

4.1 Favouring cells for suppression or publication

Often G-Confid users want to publish key cell results and are disappointed to find that they serve as complementary suppressions. To be sure to publish a key result, users can set the status of a cell to 'P' (publish) prior to running the macro SUPPRESS. This will prevent the cell from being selected as a complement. However, the use of 'P' may lead to LP problems with infeasible solutions if G-Confid cannot identify other cells for complementary suppression.

Users can nevertheless favour the cell for publication, and yet allow G-Confid to suppress it if no other cells sufficiently protect the sensitive cells. For example when CFunction1=SIZE, the cost value of the cell equals the cell value by default. Users may instead choose to supply a custom set of cost values (CVar1) for use with step 1, as described in section 3.2, and assign a higher value to the cell representing a key result. One possible value is to assign the maximum value of all cells (e.g., the grand total at the top of the hierarchy of dimensions) to a cell representing a key result, while setting the cost value equal to the cell total for all other cells. Users may implement a more complex strategy, too. Rondeau and Fillion (2011) present an example using the relative contribution to employment by geographic area as the measure of importance to favour the publication of cells.

Users may also prefer that certain cells be considered first for complementary suppression, for example, cells representing survey estimates with high coefficients of variation. To favour a cell to be selected for complementary suppression, users may set its cost to a value lower than its cell total, or assign an arbitrarily low value such as 1.

4.2 Using a cost function that is scaled to the range of cells being treated

The SCALECOST parameter, for use with the macro SUPPRESS, allows users to choose from a set of functions by which to adjust the cell costs used in the LP objective function. The parameter values are MEAN, NONE and SCALE. The default is SCALECOST=NONE, which leaves the coefficients as they are. Prior to 2011 G-Confid only used SCALECOST=MEAN, defined as:

$$w_i^* = w_i \Big/ \frac{\sum_1^n w_i}{n}$$

where w_i is the value of the current coefficient, w_i^* is the value of the rescaled coefficient and n is the number of coefficients of the cost function. Users may instead choose SCALECOST=SCALE, defined as:

$$w_i^* = \frac{(b-a)(w_i - w_{min})}{(w_{max} - w_{min})}$$

where b and a are the upper and lower bounds of the scaling interval, w_{max} and w_{min} are the maximum and minimum values of the cost function coefficients, w_i and w_i^* are as previously defined.

Each of the three SCALECOST options leads to a complementary suppression pattern that is optimal for the LP objective function that was used. The suppression pattern using one option may not be the same as the suppression pattern using another option.

Decreasing the run time of the macro SUPPRESS

G-Confid is capable of processing large tables with several dimensions of data. Even so, the limitations of processing speed and random access memory may lead to excessively long run times. When the problem is large, the SCALECOST=SCALE option takes the least processing time because there are fewer extreme values to process after rescaling. Preliminary testing have supported this assertion. One test, using SCALECOST=SCALE and involving a very large data set, completed execution in 45 hours on each of a personal computer and on a server. Using SCALECOST=NONE, neither the server nor the personal computer had sufficient resources to complete the execution within one week. Further testing involving smaller data sets showed that each of the three SCALECOST options results in a similar count of suppressed cells.

4.3 Reducing the sensitivity by referring to the sampling weights

Business surveys at Statistics Canada use sampling weights to expand the sample results to the population level. Business enterprises that contribute large, dominant values to cell totals usually have weights near one as they have been sampled with certainty ("take-all" units) or with high probability of selection. Smaller enterprises ("take-some" units) tend to have higher sampling weights, having been selected from a larger population. A reduced sampling fraction among smaller enterprises leads to greater uncertainty whether a particular smaller enterprise has contributed to a cell.

If the sampling weights are ignored while analyzing the risk of disclosure, users might suppress results to a greater degree than needed. For that reason the current practice at Statistics Canada is to render *anonymous* any enterprise with a sampling weight of more than three. An anonymous enterprise is an enterprise for which the G-Confid user has set the value of the identification variable to missing. By rendering an enterprise anonymous, its contribution reduces the sensitivity of the cell when using G-Confid. This practice reflects the view that the dominance of

a smaller enterprise within a domain is far from certain given that at least two similar, unsampled enterprises exist in the population.

4.4 Setting the minimum number of respondents

The MINRESP option was introduced so that G-Confid users may specify the minimum number of observations that contribute to a cell total. If any cell has fewer than the minimum number of observations, G-Gonfid assigns that cell a sensitivity of $\max\{1,S\}$ where S is the cell's sensitivity calculated according to the sensitivity rule (see section 2.1). Using the SUPPRESS macro, a cell with a sensitivity of one needs protection of 0.5. Note that G-Confid ignores observations with negative or zero values; they are excluded when counting the number of observations.

The MINRESP criterion is primarily intended to provide additional protection to small domains in the population with high sampling fractions. If any positive value that contributes to a cell belongs to an anonymous enterprise, G-Confid does not apply the MINRESP rule to that cell.

4.5 Decreasing the run time of the macro AUDIT

The macro AUDIT calculates the minimum and maximum values of each suppressed cell and sensitive aggregate. To do this, AUDIT begins with an initial set of bounds and attempts to narrow the interval of possible values. By default the initial lower bound is set at $0.5\times t_i$ and the initial upper bound is $1.5\times t_i$, consistent with the assumption that a potential intruder may estimate a cell total within ± 50% of its true value.

If the matrix of constraints is large, the macro AUDIT may take a very long time to run. In order to reduce the execution time, G-Confid will soon incorporate the Shuttle algorithm so that the initial interval is narrower prior to running the macro AUDIT. Using one linear constraint at a time the Shuttle algorithm recalculates the bounds of each sensitive cell. For example, to update the lower bound of a sensitive cell, it takes the difference defined by the lower bound of the marginal cell minus the sum of upper bounds of the other internal cells. The lower bound remains unchanged if the difference is a smaller value than the lower bound. A similar rule exists to update the upper bound of each sensitive internal cell. Separate rules are applied to marginal cells.

The execution time to run the Shuttle algorithm to adjust the lower and upper bounds is relatively short because only one linear constraint at a time is used to recalculate the bounds (see Buzzigoli and Giusti (2006) for information on the Shuttle algorithm). Although the use of the Shuttle algorithm does not lead to the optimal (narrowest) intervals, the resulting bounds may serve as inputs to the macro AUDIT.

Table 1 presents the results of three preliminary tests. The results show that the use of the Shuttle algorithm prior to running the macro AUDIT reduces considerably the total processing time to audit the data set. The run time is strongly associated with the number of suppressions that are

required to protect the data.

	Data set 1	Data set 2	Data set 3
No. of records, approximate	200 000	1 000 000	1 000 000
No. of linear constraints	22 899	1 128	32 684
No. of suppressed cells	7 604	1 480	3 033
Run time (in h:mm:ss)			
Macro AUDIT only	06:14	:13	1:13:12
Shuttle algorithm	:14	:03	:24
Macro AUDIT	03:06	:07	43:36

Table 1. The impact on processing time of using the Shuttle algorithm prior the AUDIT macro

4.6 Producing summary tables of the suppression pattern

The macro REPORTCELLS enables the G-Confid user to produce two-dimensional tables of cell totals. If three or more dimensions are specified, a two-dimensional table is produced for each level of a third dimension, or for each intersection of levels of two or more other dimensions. Using the SAS Output Delivery System the G-Confid user can create the table in Microsoft Excel, Microsoft Word or Adobe pdf format. The table is colour coded with a red cell indicating a primary suppression and a yellow cell indicating a complementary suppression. These tables provide G-Confid users with a visual reference to identify the cells in their tables to suppress.

4.7 Examining sensitive aggregates

In the context of G-Confid, an aggregate is a union of two or more internal cells. As described in section 3.1 when a complement offers less protected than the amount implied (a false complement situation), G-Confid identifies the union of the sensitive cell and the false complement a sensitive aggregate. At least one sensitive cell contributes to a sensitive aggregate, consistent with the subadditive property of the dominance rules that G-Confid uses. Sensitive aggregates can contain several sensitive and non-sensitive cells. In the macro SUPPRESS, sensitive aggregates as well as sensitive individual cells are protected using complementary suppression.

By using the macro AGGREGATE, G-Confid users can easily identify the individual cells (sensitive and perhaps also non-sensitive) that comprise each sensitive aggregate.; G-Confid thereby allows users to refine the identification of sensitive aggregates after analysis. For each sensitive aggregate, the output file identifies its contributing cells and displays their respective totals and measures of sensitivity. This file is not needed for the macro SUPPRESS as all of the information regarding sensitive aggregates is included in the other files generated by PROC SENSITIVITY.

References

Buzzigoli, L. and Giusti, A. (2006) From Marginal to Array Structure with the Shuttle Algorithm. *JSDA Electronic Journal of Symbolic Data Analysis*, Vol. 4, No. 1, 1-14

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

Robertson, D. and Şchiopu-Kratina, I. (1997). The mathematical basis for Statistics Canada cell suppression software: CONFID. *SSC Annual Meeting – Proceedings of the Survey Methods Section*.

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

Statistics Canada. (2011). *G-Confid User Guide. Internal Statistics Canada document*, Ottawa, Canada.

Tambay, J.-L. and Fillion, J.-M. (2011) New business survey confidentiality software G-Confid. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Tarragona, Spain, 26-28 October 2011*, Working Paper 12