WP. 12
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
(Tarragona, Spain, 26-28 October 2011)

Topic (ii): Software research and development

New business survey confidentiality software G-Confid

Prepared by Jean-Louis Tambay and Jean-Marc Fillion, Statistics Canada

New business survey confidentiality software G-Confid

Jean-Louis Tambay[1] and Jean-Marc Fillion[2]

[1] Statistics Canada, Ottawa, ON, Canada, jean-louis.tambay@statcan.gc.ca
[2] Statistics Canada, Ottawa, ON, Canada, jean-marc.fillion@statcan.gc.ca

1 Introduction

G-Confid is a generalized tool developed at Statistics Canada for the protection of tabular data confidentiality. The software can be used to create or validate cell suppression patterns for tabular economic data at various levels of aggregation. It is an improved version of the agency's old CONFID system and uses the same methodology to develop suppression patterns. A key feature of G-Confid is that it is made to work in a SAS environment. 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 (G-Confid requires SAS 9.2). Subject to SAS and hardware limitations, G-Confid can process large tables with many dimensions.

In section 2 we give an overview of G-Confid and of its underlying methodology. Section 3 provides detailed information on its three components and their features. Section 4 suggests ways to adapt G-Confid to handle special situations. Finally, performance results from some executions of G-Confid are given in section 5.

2 Overview

2.1 General description of G-Confid

G-Confid is a suite of three SAS components, PROC SENSITIVITY, macro SUPPRESS and macro AUDIT, that can create or validate a suppression pattern for tabular economic data at various levels of aggregation. The cell suppression technique, used to protect the confidentiality of respondents' data, consists of a primary suppression step and a residual suppression step.

The main objective of primary suppression is to identify and suppress sensitive (confidential) cells. Primary suppression alone does not provide sufficient protection to confidential data since an intruder can use remaining cells to derive the exact value of a suppressed sensitive cell or to obtain a tight range of values for that cell. That is why residual suppression is needed. The main objective of residual suppression is to identify complementary cells (complements) that will also be suppressed to protect sensitive cells, and to do so in a way that minimizes the resulting loss of information.

If users modify a suppression pattern produced by G-Confid they should run macro AUDIT to verify the validity of the modified suppression pattern. Macro AUDIT also can verify the validity of a suppression pattern produced outside of G-Confid.

G-Confid can be used with SAS Enterprise Guide 4.3. The G-Confid functions can appear in the form of personalized tasks. These will generate the SAS code necessary for G-Confid via a graphical interface.

2.2 Basic methodology underlying G-Confid

G-Confid uses the cell suppression methodology developed in the 1970s (see Cox and Sande (1979), Robertson and Şchiopu-Kratina (1997)). The three components of G-Confid correspond to the three main activities in tabular data protection: identifying sensitive cells; carrying out primary and residual suppression; and auditing suppression patterns.

A cell is sensitive if its total value allows the close estimation of some of its respondents' contribution. Such cells are identified using a sensitivity measure. Different measures are available, which are particular forms of the following formula:

$$S = \Sigma_i \, a_i x_i \,\text{, where } a_1 \geq a_2 \geq \ldots \geq a_r \geq -1 \qquad (1)$$

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 ($x_1 \geq x_2 \geq \ldots \geq x_r \geq 0$).

Two common sensitivity rules are the p-percent and the (n,k) rule. With the p-percent rule a cell is sensitive if the value of the smallest contributors, starting from the third-largest, is less than p% of the largest contributor's value. That is, $a_1 = p\%$, $a_2 = 0$ and $a_i = -1$ for $i > 2$. With the (n,k) rule the cell is sensitive if the largest n contributors account for more than k percent of the cell total. This gives $a_i = (100–k)/k$ for $i \leq n$ and $a_i = -1$ for $i > n$.

The second step, residual suppression, is done by solving a Linear Programming (LP) problem. In G-Confid sensitive cells are protected in turn, usually starting with the cell with the highest sensitivity value S. Each sensitive cell is protected by suppressing enough other cells (preferably other sensitive cells and cells suppressed to protect previously treated sensitive cells) so that its true value cannot be estimated from the table within a margin that is less than one half of its sensitivity.

Residual suppression is basically determined as follows. The table is represented by a set of equations establishing the relationship between table 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 LP problem is:

Minimize:	$\boldsymbol{w'y} + \boldsymbol{w'z}$	(objective function)
Subject to:	$\boldsymbol{Cy} - \boldsymbol{Cz} = \mathbf{0}$	(equations defining relationships in the table)
	$\mathbf{0} \leq \boldsymbol{y}, \boldsymbol{z} \leq \boldsymbol{t}/2$	(bounds on the movements of cells)
	$y_{sen} \geq S_{sen}/2$	(sensitive cell *sen* moved by $S_{sen}/2$)
	$z_{sen} = 0$	

Vectors ***y*** and ***z*** represent positive and negative changes in cell values, respectively, ***w*** is a fixed cost vector (w_i=0 if cell *i* is sensitive or was already suppressed), ***t*** is the vector of cell total values t_i and **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 complement and gets suppressed. One can influence the suppression pattern using different w_i's in the objective function. Because G-Confid processes one sensitive cell at a time the result can be suboptimal. To reduce the number of suppressions, once an initial suppression pattern is established for all sensitive cells the non-suppressed cells *j* are omitted ($y_j = z_j = 0$) and the process is run a second time and with a different objective function.

Auditing a suppression pattern involves finding maximum and minimum values of each sensitive 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 Components of G-Confid

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 rule(s) used to identify sensitive (confidential) cells

Following SAS standard notation PROC SENSITIVITY can have six statements:

PROC SENSITIVITY <*option(s)*>;
ID *variable*;
VAR *variable*;

SHADOW *variable*;
DIMENSION *variable(s)*;
BY *variable(s)*;

An example is:

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

Different options are specified under the **PROC** statement. One gives the names of the input microdata file and of SAS output datasets including a cell-level data file that contains the sensitivity of the cells, another file that describes the constraints (the table equations) and an optional one that contains information about the largest contributors to each cell or sensitive aggregate. The input microdata file must be at the unit (e.g., enterprise) level and can be in any format importable to SAS.

The hierarchy used for each table dimension is specified using the following format:

```
hierarchy="/*DIM1*/1 11 12 13: 11 111 112: 111 1111 -1 119;
  /*DIM2*/2 21 22: 21 211 -2 217: 22 23 -1 33;"
```

The dimensions are separated by a semi-colon (;). Each level is separated by a colon (:). A level is a parent code and the remaining values help to identify the codes for its children. SAS comments (e.g., /*DIM1*/) can be inserted for readability. One can list all the children corresponding to a parent ("1 11 12 13:" means parent 1 has children 11 12 and 13) or one can use negative numbers to denote increments ("21 211 -2 217:" is equivalent to "21 211 213 215 217:"). Multiple decompositions of a parent are allowed. So a "roulette wheel" triple decomposition could be:

```
hierarchy="/*Roulette*/ALL 0 00 EVEN ODD:ALL 0 00 1ST12 2ND12 3RD12:
  ALL 0 00 1TO18 19TO36:EVEN 2 -2 36:ODD 1 -2 35:1ST12 1 -1 12:2ND12
  13 -1 24:3RD12 25 -1 36:1TO18 1 -1 18:19TO36 19 -1 36;"
```

The range of codes associated with the lowest level of a hierarchy can be specified. As with hierarchy codes, the dimensions are separated by a semi-colon (;) each range is separated by a colon (:) and range codes can be listed individually or by increment. In the example above, the range for dimension 1 is not given, which means that code values for the dimension are reported as "1" and "2" on the microdata file.

Sensitivity rules are used to calculate cell sensitivity values and identify sensitive cells. One can specify a *p*-percent rule, up to three (*n,k*) rules applied jointly, an arbitrary rule (by giving values for a_1, a_2, a_3 and a_4 in (1)), or Statistics Canada internal rules (*duffett* and *c2*). In the example above, a (1,70) and a (2,80) rule are jointly used. One can also specify a minimum number of respondents with a nonzero value in a cell (MINRESP). If a nonsensitive cell does not meet this minimum the cell is made sensitive with sensitivity value 1.

Other parameters can be specified including some that will be covered in section 3.4.

The **ID** statement is mandatory and specifies the unit (e.g., enterprise) variable of the input data set. The variable must be a character variable. A missing value represents an anonymous respondent, i.e., a respondent whose value can be disclosed without risk. Anonymous respondents typically occur in sample data where a respondent can represent several units in the population. When calculating the cell sensitivity anonymous respondents are given a coefficient $a_i = -1$ in (1). A cell with anonymous nonzero respondents is assumed to pass the MINRESP criterion, if there is one.

The **VAR** statement is mandatory and specifies the variable for which the procedure calculates the sensitivity. The variable must be a numeric nonnegative variable.

The **SHADOW** statement is optional. It specifies an auxiliary variable that is processed alongside the main variable mostly for diagnostic purposes. If processing a variable that can be negative, like Profits, the processed variable can be absolute profits and the shadow variable, actual profits. The variable must be a numeric.

The **DIMENSION** statement is mandatory and specifies the variables that the procedure uses as the dimension variables. The variables must be character variables.

The optional **BY** statement specifies variables the procedure uses to form BY groups. A BY statement could serve to process unrelated questionaire items. The BY variables can be numeric or character. They will appear in the output data sets.

3.2 Macro SUPPRESS

This macro carries out the residual suppression. Aggregated data are processed through a LP solver and using the constraints generated in PROC SENSITIVITY. The syntax is:

```
%Suppress(InCell=, Constraint=, CFunction1=, CFunction2=, CVar1=,
   CVar2=, OutCell=, OutComplement=, By, ScaleCost=, DebugInfo= );
```

InCell and Constraint identify the datasets containing the actual table cells (and sensitive aggregates, see 3.4) and the linear constraints coefficients, respectively.

As noted in section 2.2, to reduce the number of suppressions the LP process can be run twice, with different objective function values (w_i). CFunction1 and CFunction2 identify the cost function used each phase. CVar1 and CVar2 identify the numeric variables to which these apply (their default is TOTAL, which is the total of the processed variable, i.e., t_i). With t_i , cost function choices are SIZE ($w_i = t_i$), DIGITS ($w_i = \log_{10}(t_i+1)$), CONSTANT ($w_i =1$) and INFORMATION ($w_i =\log_{10}(t_i+1)/(t_i+1)$). For CFunction1 the default value is DIGITS. INFORMATION is often used in the second run to “free up” suppressed small cells. If CFunction2 is not specified the process is run once only.

The ability to use other variables than the processed variable is useful if one wants to influence a suppression pattern. One could wish to favour the suppression of cells suppressed historically or those with a high coefficient of variation, or to avoid suppressing cells for industries that represent a particular importance in a province.

OutCell gives the name of the SAS output data file. In addition to the fields contained in the InCell data set, this file contains OUTSTATUS, which indicates the status of the cell (‘P’ for published or ‘X’ for suppressed) and NETVARIATION, which is the largest amount by which the cell value was moved in protecting sensitive cells (see section 2.2). Published cells have zero net variation.

OutComplement gives the name of the output complement data file, which identifies all the complements used in protecting each sensitive cell. It can be useful to analyse the outcome of a suppression. For example, sensitive cells responsible for most suppressions could be targeted for *waivers* – i.e., to obtain respondents’ permission to publish.

By specifies the variable name(s) used to create BY groups, if any. It is optional.

ScaleCost rescales the cost function coefficients, which can help the LP to run more smoothly. Possible values are NONE (the default), MEAN and SCALE. When SCALE is used the cost coefficients (w_i) are replaced by $B(w_i - w_{min})/(w_{max} - w_{min})$ for some constant B. With MEAN they are replaced by w_i / w□, where w□ is the average of the w_i’s.

If DebugInfo=YES additional information is printed in the log for debugging purposes. This parameter is optional and the default value is NO.

An example of an output report from Macro SUPPRESS is given in Figure 1.

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. The syntax is:

```
%Audit(InCell=, Constraint=, OutCell=, LBFactor=, UBFactor=, By=,
   SasConnect=, DebugInfo=, ReportLevel= );
```

Parameters InCell, Constraint, By and DebugInfo were defined above. OutCell, the output datafile, has one record for each suppressed cell or sensitive aggregate. It contains many of the same variables as the OutCell in Macro SUPPRESS plus the derived minimum, maximum and midpoint values of the cell or sensitive aggregate and a problem indicator. For sensitive cells or aggregates the problem indicator is 0 for a good protection, 1 for an unachieved protection and 2 for an exact disclosure. For complements it is 0 for a good protection and 2 for an exact disclosure.

LBFactor and UBFactor set bounds for suppressed cells in the LP solver. The value for each suppressed cell i is set to be between LBFactor*t_i and UBFactor*t_i. The parameters are optional with $0 \leq$LBFactor≤ 1 and $1 \leq$UBFactor≤ 10 (default values are 0.5 and 1.5).

SasConnect specifies to use SAS/Connect to perform parallel processing on the same machine to decrease execution time. If SasConnect=YES the cells to be processed in AUDIT are distributed to the processors available to the machine. The process is transparent to the user. This parameter is optional and the default value is NO.

ReportLevel specifies the level of reports to be generated when AUDIT is completed. Possible values are 1 and 2. When 1, a report giving statistics for “Problem Indicator” is generated. When 2, two supplementary reports are generated, giving statistics of the “Midpoint Problem” for suppressed cells and sensitive aggregates. This parameter is optional and the default value is 1. An example of an output report from Macro AUDIT is given in Figure 2.

3.4 Creation of sensitive aggregates in PROC SENSITIVITY

For computational efficieny reasons Macro SUPPRESS works with cell-level data rather than respondent microdata. But microdata are needed to avoid what is called the *false complement* problem, which occurs when a complement seems to offer more protection to a sensitive cell than it actually does. A common false complement situation is when two one-respondent cells are used to mutually protect each other. Their union, having two respondents only, is still sensitive. Other false complement situations may occur when a complement 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 processes the microdata to ensure that false complement situations are treated along one-dimensional lines (rows, columns, etc.). It does this by calculating the true sensitivity of unions of sensitive cells, and of unions of sensitive and nonsensitive cells, along each line to identify which unions remain sensitive. Pseudo-cells called *sensitive aggregates* are generated for the sensitive unions and added to the cell-level file that is passed to Macro SUPPRESS. The relations between sensitive aggregates and their component cells are added to the constraints file.

A large number of sensitive aggregates may be examined and generated by this process, which will hinder the execution of G-Confid. For example, if one province out of 10 is sensitive, $2^9-1=511$ unions of that province with other provinces could be checked for false complements. For this reason, the PROC SENSITIVITY statement includes parameters to control the number of sensitive aggregates that are examined or generated. One parameter, M, limits the number of nonsensitive cells that can be considered for a sensitive aggregate. By setting M=3, 129 unions would be checked instead of 511. A percentage X prevents the generation of sensitive aggregates for milder cases of false complements. Two other percentages, Y and Z, allow the screening out of uninteresting cells such as very small ones from consideration for sensitive aggregates.

4 Adapting G-Confid to handle special situations

4.1 Changing the status of cells

Two important variables on the cell-level file that is created in PROC SENSITIVITY are SENSITIVITY, the sensitivity of the cell, and STATUS, its status. Status values are ‘S’ for sensitive cells and ‘V’ (variable) for others. Users can change the STATUS values to ‘P’ for published and ‘X’ for suppressed. This is useful if running G-Confid on linked tables, for example, a table at the 5-digit industry level that is processed after a table at the 4-digit level. Status ‘P’ indicates that the cell was already published and hence cannot be used as a complement. The status can also be used to prevent a key cell from being used as a complement. Status ‘X’ indicates that a cell is suppressed. Such cells will have zero cost associated with them in the objective function, which will make them more likely to be used as complements.

Changing statuses must be done carefully. The use of ‘P’ may lead to LP problems with infeasible solutions. And Macro SUPPRESS will not find complements for cells with status ‘X’. To do that, one should assign a sensitivity value to such cells.

4.2 Treatment of weighted data

Many business surveys use weights (wt_i) to expand sample results to the population level, adjust for nonresponse and/or calibrate results to population totals. The input microdata file does not handle survey weights but ways to handle weighted data have been suggested. One option is to use unweighted values y_i for respondents with wt_i <3 and weighted values $wt_i y_i$ for all others (and make them anonymous respondents). Weighted values $wt_i\, y_i$ for all units could be carried in the SHADOW variable in order to reproduce actual survey totals. Another option, which preserves totals, is to duplicate records as follows: if $1<wt_i \leq 2$ create two respondent records with values y_i and $(wt_i -1)y_i$ and if $wt_i >2$ create two records, each having value y_i , plus an anonymous respondent record with value $(wt_i-2)y_i$.

4.3 Treatment of negative values

PROC SENSITIVITY does not treat negative values. Observations with a negative or missing value for the VAR variable are skipped. Several ways have been suggested to process variables y_i than can be negative including:

- use number of respondents (option MINRESP),
- replacing y_i by their absolute values $|y_i|$,
- replacing negative values of y_i by 0,
- adding a large enough constant $K \geq \max\{-y_i | y_i < 0\}$ to make all values nonnegative,
- combining y_i with a nonnegative size variable x_i (e.g., $\max\{|y_i|, \alpha x_i\}$ for some α).

The solution adopted will depend on the nature of the variable y_i. For methods that modify the respondent data it is suggested to keep the original values as the SHADOW variable.

5 Performance of G-Confid

PROC SENSITIVITY is very fast. Results from executions of Macro SUPPRESS are presented below. They were obtained with a computer that had an AMD Athlon(tm) 64x2 Dual core Processor 4800+ 2.5GHz with 4 GB Ram and a 32 Bit operating system. SAS version 9.2 was used. The LP process was run twice, first using the SIZE cost option and then using the INFORMATION cost option. As results show, this can make the number of suppressions decrease significantly.

Run times	Number of dimensions	Number of cells	Num. of sensitive cells	Number of sensitive aggregates	Number of complements after Phase1	Number of complements after Phase 2
9 sec.	2	3046	333	68	357	312
32 sec.	2	5245	856	118	712	506
6 sec.	3	1329	147	42	592	442
4 sec.	3	2149	69	15	230	172
10 sec.	3	2825	306	55	709	593
53m.14s.	3	8074	608	146	2116	1183
2h. 45m.	4	16992	2527	582	6007	4481

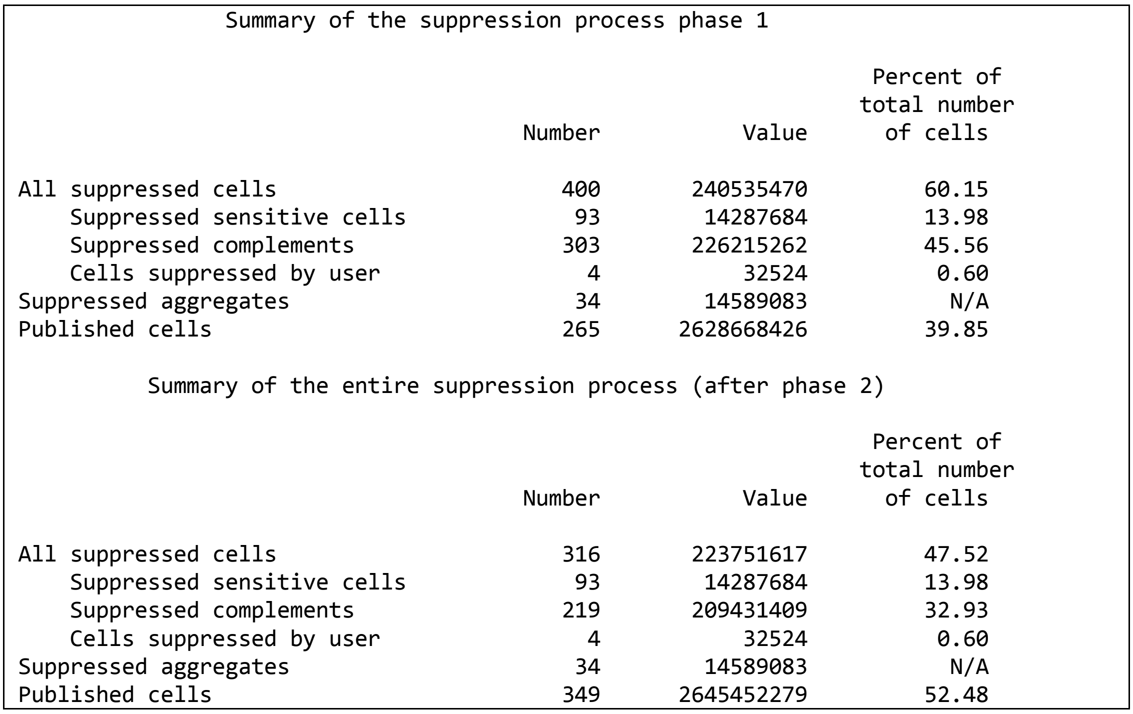
Table 1. Sample output report from Macro SUPPRESS

References

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

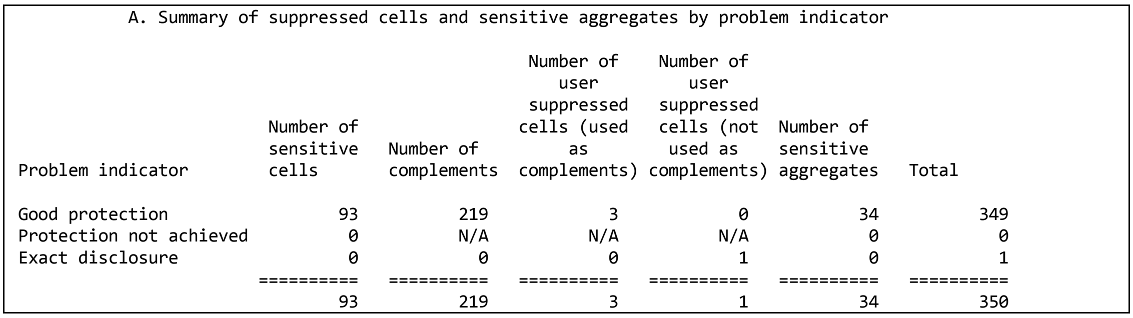
Robertson, D. and Şchiopu-Kratina, I. (1997). The mathematical basis for Statistics Canada cell suppression software: CONFID. *SSC Annual Meeting – Proceedings of the Survey Methods Section*.

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



```
                 Summary of the suppression process phase 1

                                                                  Percent of
                                                                 total number
                                       Number            Value     of cells

All suppressed cells                      400        240535470        60.15
    Suppressed sensitive cells             93         14287684        13.98
    Suppressed complements                303        226215262        45.56
    Cells suppressed by user                4            32524         0.60
Suppressed aggregates                      34         14589083          N/A
Published cells                           265       2628668426        39.85

           Summary of the entire suppression process (after phase 2)

                                                                  Percent of
                                                                 total number
                                       Number            Value     of cells

All suppressed cells                      316        223751617        47.52
    Suppressed sensitive cells             93         14287684        13.98
    Suppressed complements                219        209431409        32.93
    Cells suppressed by user                4            32524         0.60
Suppressed aggregates                      34         14589083          N/A
Published cells                           349       2645452279        52.48
```

Fig 1. Sample output report from Macro SUPPRESS.



```
            A. Summary of suppressed cells and sensitive aggregates by problem indicator

                                                  Number of    Number of
                                                    user         user
                                                  suppressed   suppressed
                        Number of                 cells (used  cells (not  Number of
                        sensitive   Number of         as        used as    sensitive
Problem indicator       cells       complements  complements) complements) aggregates   Total

Good protection                 93         219            3            0          34         349
Protection not achieved          0         N/A          N/A          N/A           0           0
Exact disclosure                 0           0            0            1           0           1
                       ==========  ==========   ==========   ==========  ==========  ==========
                                93         219            3            1          34         350
```

Fig 2. Sample output report from Macro AUDIT.