STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 16
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

IMPROVING CELL SUPPRESSION IN STATISTICAL DISCLOSURE CONTROL

Contributed paper

Submitted by the University of La Laguna, Tenerife, Spain [1]

Abstract: This paper presents the well-known Cell Suppression Methodology, herein called *complete cell suppression*. A new methodology, called *partial cell suppression*, is then described and favourably compared with complete cell suppression. Finally, a new *combined method* merging the best of both methodologies is introduced leading to a more powerful cell suppression technique to protect sensitive cells in all kinds of tables.

I. CLASSICAL CELL SUPPRESSION METHODOLOGY

1. Given a data set in tabular form (as the two-dimensional table presented in Figure 1) with sensitive cells (as the cell with nominal value 22 corresponding to Activity II and Region C in Figure 1), the *Cell Suppression Methodology* is a widely used technique to maintain confidentiality while publishing the table. It consists of omitting (or replacing by an asterisk) the necessary cells to guarantee that an external attacker cannot deduce the sensitive information. Therefore, at least the sensitive cells must be omitted, and they are called *primary suppressions*. But typically, because the existence of mathematical links between the cells in the tabular data (for example, the existence of marginal sums, etc.), the primary suppressions are not enough and other cells must also be suppressed. Those additional (non-sensitive) cells are called *secondary suppressions*.

	Region A	**Region B**	**Region C**	**TOTAL**
Activity I	20	50	10	80
Activity II	8	19	*22*	49
Activity III	17	32	12	61
TOTAL	45	101	44	190

Figure 1: investments of enterprises in millions of Euros.

2. Of course, the task of choosing the "best" secondary suppressions to:
   i) guarantee that the information under the primary suppressions is protected, and
   ii) minimise the loss of information due to the non-published data cells,

is not an easy problem, generally known as the Complementary or Secondary Cell Suppression Problem. Hereafter, we refer to this problem as complete CSP. A potential solution of the complete CSP for the instance in Figure 1 is given by the suppression pattern in Figure 2.

[1] Prepared by Juan José Salazar González.

	Region A	Region B	Region C	TOTAL
Activity I	*	50	*	80
Activity II	*	19	*	49
Activity III	17	32	12	61
TOTAL	45	101	44	190

Figure 2: potential solution to the complete CSP.

3. In practice it is not enough to protect only the exact value under the asterisk of each primary suppression, but the statistical office is also interested on guarantee certain uncertain range values. To be more precise in the definition of the complete CSP, let us introduce some basic concepts:

- External bounds:: For each cell k, we call *lower external bound* lb_k the minimum value that an attacker knows in advance on such â cell if it is suppressed. Typically it is 0, but in particular cases it could be another number. Also for each cell k, we call *upper external bound* ub_k the maximum value that an attacker knows in advance on such cell if it is suppressed. Typically it is a very big value, but in particular cases it could be another number. Both external bounds are *a priori* information from an external attacker on a suppression value, and therefore they must be input parameters for the complete CSP.
- Congruent table: Given a suppression pattern (as the one presented in Figure 2) and given external bounds for each suppressed cell, a *congruent table* is a set of values that could be the original table for an external attacker. Therefore, a congruent table coincides with the published values when the cells are not suppressed, satisfies all the mathematical equations of the tables, and contains congruent values with the external bounds when they correspond to suppressed cells.
- Protection Level Requirements: Given a sensitive cell k and three input parameters LPL_k, UPL_k and SPL_k, a suppression pattern protects cell k if and only if, an attacker (with the only knowledge of the suppression pattern and the external bounds on its suppressions) cannot classify as impossible:
  i) a congruent table with value on cell k smaller or equal than the nominal value minus LPL_k;
  ii) a congruent table with value on cell k bigger or equal than the nominal value plus UPL_k;
  iii) two congruent tables with the difference between their values on cell k bigger or equal than SPL_k.

  Parameters LPL_k, UPL_k and SPL_k are called *lower*, *upper* and *sliding protection level*, respectively. Typically the statistical office fixes them as percentages of the nominal value (for example, 20%, 30% and 60%, respectively), but they can be chosen as another numbers. Obviously, the bigger those parameters are, the more information will be lost to guarantee confidentiality.
- Loss of information: Whenever a cell is suppressed in a suppression pattern there is a loss of information. Each cell k has an associated parameter w_k known as the *cost of suppression* if cell k is suppressed, and the overall *loss of information* of a pattern is the sum of the cost of suppression for all its suppressions.

4. The complete CSP looks for the set of suppressions that guarantee the three protection level requirements for all the sensitive data of the table, minimising the total loss of information. Therefore, it is an *optimisation* problem.

5. Complete CSP has been extensively studied in literature (see, for example, Willenborg and De Waal (1996) for references. Fischetti and Salazar (1999) propose a new algorithm based on modern mathematical programming techniques, and solve 2-dimensional instances with up to 500 rows and 500 columns in few minutes of a very standard PC computer. Fischetti and Salazar (2000) extend their proposal to other tabular data including k-dimensional tables with k>2, hierarchical tables, linked tables, etc., providing also very interesting computational results.

II. PARTIAL CELL SUPPRESSION METHODOLOGY

6. In the classical Complete Cell Suppression methodology, even if a suppression pattern consists on publishing or suppressing each cell of the table, from the attacker point of view, the final output is a set of interval. Indeed, from the suppression pattern and from the external bounds, the attacker will compute a *feasibility interval* for each missing value. For example, an attacker watching the suppression pattern in Figure 2, and knowing that the external bounds for the missing values are 0 and infinity (i.e.,

the values under asterisk are nonnegative) will know that the values in congruent tables are in the intervals of Figure 3.

	Region A	**Region B**	**Region C**	**TOTAL**
Activity I	[0...28]	50	[2...30]	80
Activity II	[0...28]	19	*[2...30]*	49
Activity III	17	32	12	61
TOTAL	45	101	44	190

Figure 3 : feasibility interval for pattern in Figure 2

7. In fact, from Figure 2 and the non-negativity of the missing values, an attacker will know that (say) 1 is not a possible value under the asterisk in Activity II and Region C, even if apparently the complete cell was suppressed. To compute the feasibility intervals even for very complex tabular data like a linked table, the attacker needs only a Linear Programming optimiser, which is very standard and easy to have. Indeed, today almost all data-processor software (e.g. *MS Excel*, *Maple*, *Mathematica*, etc.) offer a Linear Programming optimiser, and also in *Internet* there are several public-domain codes to perform the numerical task of computing feasibility intervals. Therefore, the classical complete cell suppression methodology could be understood as publishing intervals instead of yes-or-not numbers. Of course, this observation is well-known and typically referred to as the output of the *auditing* phase.

8. The motivation for developing a new methodology arises from the following question. If the optimisation problem in classical (complete) cell suppression looks for cells to publish intervals instead of the exact original values, why not look for the extreme values of the intervals too?

9. If the objective function in Cell Suppression is to minimise the loss of information, then extra freedom of computing also the extreme values of the feasibility intervals gives the optimiser the option of finding patterns with smaller loss of information while guarantee the same protection level requirements. For example, if the statistical office requires protection levels of *LPL*=2, *UPL*=4 and *SPL*=0 on the sensitive cell in Figure 1, then the suppression pattern in Figure 2 could be one with minimum information loss using classical cell suppression with certain external bounds and costs of suppression. But with the new methodology (*partial cell suppression*) other suppression patterns satisfying the required protection levels and with smaller loss of information are also possible. In the example, Figure 4 exhibits a possible pattern satisfying the required protection levels, losing less information than the one presented in Figure 3. Therefore a first advantage of using the new methodology instead of the classical one is that partial suppression saves loss of information.

	Region A	**Region B**	**Region C**	**TOTAL**
Activity I	[18...24]	50	[6...12]	80
Activity II	[4...10]	19	*[20...26]*	49
Activity III	17	32	12	61
TOTAL	45	101	44	190

Figure 4 : potential solution to the partial CSP.

10. A second clear advantage of using the new methodology is that the “auditing phase” is not necessary. Indeed, after solving the complete CSP one has to proceed solving also the auditing phase to compute the feasibility intervals for the suppressions. In the new methodology this extra work is not necessary since the feasibility intervals are automatically computed by the partial CSP itself.

11. Apparently the optimisation problem of the new methodology (partial CSP) is more complicated than the old one (complete CSP), but not! Comparing both optimisation problems with standard tools in Complexity Theory, the partial CSP is much easier than the complete CSP. To be more precise, the complete CSP is *NP*-hard in the strong sense, while the partial CSP is polynomially solvable. Therefore, a third advantage of using the new methodology is that the optimisation problem is (in theory) much easier.

12. Of course, one could think the last advantage is only an analysis from a theoretical point of view, but in practice it could be hard to implement an efficient algorithm to solve the partial CSP. This is not the case, and the mathematical programming details are in Fischetti and Salazar (1998). In fact, in this report the computational results in the following table where obtained using a PC Pentium 133 Mhz. with 32 RAM mbytes running under *MS-Windows 95*.

type	cells	link	sensi	levels	sup	loss	Time	sup'	loss'	time'	time''
41x31	1271	72	3	6	9	153	4.78	7	486	367.62	0.07
183x61	11163	244	2467	4934	6	140297	17.28	3	281398	11.42	667.06
359x46	16514	405	4923	9846	85	9357	77.67	38	19099	116.93	2486.07

13. The first five columns describe the features of the three real-world instances: "type" gives the internal structure of each instance, "cells" gives the number of potential cells to be unpublished, "links" gives the number of mathematical equations between cells in the table, "sensi" gives the number of sensitive cells, and "levels" the number of non-zero protection level requirements. The next three columns regard to partial CSP features: "sup" is the number of cells with intervals in an optimal solution, "loss" the loss of information, and "time" the computational effort in seconds of a Pentium 133. The last four columns regard to complete CSP features: "sup' " is the number of missing values in an optimal suppression pattern, "loss' " is the loss of information in an optimal suppression pattern, "time' " the effort in seconds to solve the complete CSP, and "time'' '' the effort to solve the auditing phase.

14. A fundamental hypothesis for preferring partial cell suppression instead of classical cell suppression is that pattern in Figure 4 publishes more information than pattern in Figure 3 (which is equivalent to pattern in Figure 2 from the attacker point of view). This is based on the idea that intervals with smaller width release more detailed information.

15. An important observation is that the loss of information has a different meaning in both methodologies. While in complete CSP there is a yes-or-no cost w_k depending if the cell k is suppressed or published, in partial CSP the loss of information is proportional to the width of the published interval. The cost of losing a unit of information under or over the nominal value could be even different, if convenient for the statistical office. In other words, in complete CSP the statistical office must consider a fixed cost of suppression w_k to each cell k if it is unpublished, independently of the feasibility interval that its suppression will have in the final pattern. For example, a cell with cost of suppression 100 will add exactly 100 to the loss of information if it is unpublished, no matter the width of the feasibility interval in the final suppression pattern. In partial CSP the statistical office can consider a loss of information proportional to the width of the published interval. For example, the statistical office can set a loss of information of (say) 5 units for each unit in the feasibility interval under the nominal value, and (say) 2 units for each unit in the feasibility interval over the nominal value. Then, a cell with original value of 300 but replaced by the interval [290...320] implies a loss of information equals to 90, while if replaced by the interval [295...310] implies a loss of information equals to 45. In general, in partial CSP each cell k must have two input parameters w_k^- and w_k^+.

III. IMPROVING THE CLASSICAL CELL SUPPRESSION METHODOLOGY

16. From the above computational experiments one observes that usually the partial cell suppression tends to substitute more values by intervals than the ones that classical cell suppression substitutes by asterisk. The substitution of a value is done even if the interval width is very narrow but it helps to decrease the information loss incurred. To solve this difficulty it is very simple to combine both methodologies and have a new technique that also provide the statistical office with additional control on the number of the replacements, the minimum length of the intervals, etc.

17. Indeed, a mathematical model for solving the complete CSP needs a set of binary variables, one variable x_k for each potential suppressed cell k:

$x_k = 1$ if cell k must be suppressed
$= 0$ otherwise.

In other words, the pattern to be published will contain the original value of cell k when x_k=0, and will substitute the original value by an asterisk when x_k=1. Then an integer-programming model is:

minimise	$\Sigma_k \; w_k \, x_k$	
subject to:	$x_k \in \{0,1\}$	for all cells k

plus a set of linear constraints to impose the protection level requirements over all the sensitive cells. The idea for writing these requirements as linear constraints is based on *Duality Theory* in Linear Programming, and we refer the reader to Fischetti and Salazar (2000) for mathematical details.

18. In a similar way, a mathematical model for solving the partial CSP needs a set of non-negative continuos variables, two variables z_k^- and z_k^+ for each potential replaced cell k:

z_k^- represents the decrement respect to the nominal value
z_k^+ represents the increment respect to the nominal value.

In other words, the pattern to be published will contain [$a_k - z_k \;\; ... \;\; a_k + z_k$] instead of the nominal value a_k. Of course, when $z_k^- = z_k^+$ then the original value of cell k is published. Then an integer-programming model is:

minimise	$\Sigma_k \; (w_k^- z_k^- + w_k^+ z_k^+)$	
subject to:	$z_k^- \geq 0$	for all cells k,
	$z_k^+ \geq 0$	for all cells k,

plus a set of linear constraints to impose the protection level requirements over all the sensitive cells. The idea for writing these requirements as linear constraints is based on *Duality Theory* in Linear Programming, and we refer the reader to Fischetti and Salazar (1998) for mathematical details.

19. In order to describe the optimisation problem of the combined cell suppression methodology by mixing both methodologies, we only need to add the binary variables to the partial CSP with the following meaning:

x_k= 1 if cell k must be replaced by an interval
= 0 otherwise,

plus the following constraints:

$z_k^- \leq (a_k - lb_k) \, x_k$	for all cells k
$z_k^+ \leq (ub_k - a_k) \, x_k$	for all cells k

where a_k is the nominal value, lb_k the lower external bound and ub_k the upper external bound of cell k. These constraints ensure that x_k=1 when z_k^- or z_k^+ is positive, i.e., when the exact value a_k of cell k is selected to be unpublished.

20. Of course, the new mathematical model contains both continuous and integer variables, so it belongs to Mixed Integer Mathematical Programming, and hence in general the optimisation problem is classified as *NP*-hard problem in Complexity Theory. Therefore, the new *combined methodology* has not the advantage of having a polynomially-solvable optimisation problem, as the partial cell suppression has. Nevertheless, a branch-and-cut algorithm can be implemented for finding heuristic and optimal

solutions on medium-sized instances, and therefore the proposed methodology can be applied. The basic idea of this algorithm follows the same line of the algorithm proposed in Fischetti and Salazar (2000) for solving the complete CSP.

21. Regarding the other two advantages presented in the partial cell suppression and not in the classical complete cell suppression, the combined methodology has both. In fact, the loss of information can be smaller and the auditing phase is unnecessary because the new methodology computes also the extreme values of the feasibility intervals.

22. Moreover, as said before, the new combined methodology allows also controlling the number of unpublished cells, as the mathematical model can consider the inequality

$$\Sigma_k \; x_k \;\; \leq \;\; \text{MAX\_INTERVALS}$$

for a given input parameter MAX_INTERVALS representing the maximum number of proper intervals that we allow in the final pattern. Also, the minimum width of a proper interval can be controlled in advance by considering constraints like:

$$z_k^- + z_k^+ \geq \text{MIN\_WIDTH} \cdot x_k \qquad \text{for all cells } k$$

for a given input parameter MIN_WIDTH representing the minimum width that the statistical office wants to have on a proper interval in the published pattern. Finally, also other controls on the features of the final pattern can be provided whenever they can be mathematically written as linear constraints on the above variables.

IV. CONCLUSIONS

23. We have presented three methodologies for protecting sensitive data before publishing a statistical table. The first methodology is the classical Cell Suppression, called “complete cell suppression”, consisting of choosing cells that can be replaced by missing values (or asterisk). The second is a new methodology called “partial cell suppression” where the nominal value of some cells can be replaced by an interval. The third methodology joins the advantages of the previous ones leading to a combined technique. The underlying optimisation problems of the three methodologies are treated.

References

M. Fischetti, J.J. Salazar, “Partial Cell Suppression: a New Methodology for Statistical Disclosure Control”, *working paper*, University of La Laguna, 1998.

M. Fischetti, J.J. Salazar, “Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control”, *Mathematical Programming* 84 (1999) 283-312.

M. Fischetti, J.J. Salazar, “Models and Algorithms for Optimising Cell Suppression in Tabular Data with Linear Constraints”, *Journal of American Statistical Association*, 451 (2000).

L. Willenborg, T. De Waal, “Statistical Disclosure Control in Practice”, Lecture Notes in Statistics, Vol. 111, *Springer Verlag*, New York, 1996.