Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data *

Juan-José Salazar-González
DEIOC, Facultad de Matemáticas, Universidad de La Laguna,
38271 La Laguna, Tenerife, Spain.
Fax: + 34 922 318170; e-mail: jjsalaza@ull.es

April 2004

Abstract

Rounding methods are common techniques in many statistical offices to protect sensitive information when publishing data in tabular form. The classical versions of these methods do not consider protection levels while searching patterns with minimum information loss, and therefore typically the so-called auditing phase is required to check the protection of the proposed patterns. This paper presents a mathematical model for the whole problem of finding a protected pattern with minimum loss of information, and proposes a branch-and-cut algorithm to solve it. It also describes a new methodology closely related with the classical controlled rounding methods but with several advantages. The new methodology is named Cell Perturbation and it leads to a different optimization problem which is simpler to solve than the previous problem. This paper presents a cutting-plane algorithm for finding an exact solution of the new problem, which is a pattern guaranteeing the same protection level requirements but with smaller loss of information when compared with the classical controlled rounding optimal patterns. The auditing phase is unnecessary on the solutions generated by the two algorithms. The paper concludes with computational results on real-world instances and discusses a modification in the objective function to guarantee statistical properties in the solutions.

Keywords: Statistical Disclosure Control; Controlled Rounding; Integer Linear Programming

1 Introduction

Statistical agencies are often required by law or policy to protect the confidentiality of the information that they collect from persons, businesses, or other units. The *microdata* is the collection of all the individual responses, and a *statistical table* is the aggregation of one variable in accordance

*This work was partially supported by the Spanish "Ministerio de Ciencia y Tecnología" (TIC2002-10231-E), by the British "Office of National Statistics" (IT-03/0763) and by the European Research project IST-2000-25069 entitled "Computational Aspects of Statistical Confidentiality" (CASC)

with other variables and including marginal sums. Before releasing statistical tables (or microdata files), these agencies use a variety of statistical methods to protect their sensitive data and to ensure that the risk of disclosure is controlled and very small. In essence, statistical agencies protect the confidentiality of the data that they collect (i.e., microdata) by restricting the amount of information in tabular data products that they release. Therefore, a common characteristic of all the methodologies is that they reduce the information to limit the disclosure risk, but with the aim of minimizing the loss of information. There are methodologies to protect microdata and others to protect statistical tables. This paper concerns only methodologies to protect statistical tables directly, i.e. modifying the table itself and not the original microdata. See, e.g., Duncan, Fienberg, Krishnan, Padman and Roehrig [7] for other perturbation techniques where some respondent contributions (i.e., values in the original microdata) are modified, such as the addition of random noise by Evans, Zayatz and Slanta [10], data swapping by Fienberg, Steele and Makov [11], and Markov perturbation by Duncan and Fienberg [8]. We refer the reader to Willenborg and de Waal [23] for a wider introduction to the Statistical Data Protection.

The importance of protecting tabular data has been clearly stated by governments awarding contracts to conduct research and issue reports on Disclosure Limitation Methods for Tabular Data Protection. For example, the *National Institute of Statistical Sciences* (`www.niss.org`) is supporting the U.S. project entitled "Digital Government", and *EUROSTAT* is coordinating the E.U. project entitled "Computational Aspects on Statistical Confidentiality", both addressing the protection of tabular data (among other topics). An important observation is that, in practice, many statistical offices consider the published statistical tables as the "end product" of the data production process, presenting the final result of the data analysis that is contained in such a table, and thus it is not assumed that there will be a secondary statistical analysis on the released data. This hypothesis is based on the fact that most of the users are interested only in the value of a very specific cell, and this opinion is supported by experts of many statistical agencies (Statistics Netherlands, German Federal Statistical Office, etc.). For the purpose of a statistical analysis, these statistical agencies do nowadays offer other ways of accessing to data, like the direct or remote access to the part of the microdata file at secure sites (see, e.g., Dunne [9]).

In the area of Statistical Disclosure Limitation experts typically distinguish two different problems. The *primary problem* concerns the problem of identifying the sensitive data, i.e., the cell values corresponding to private information that cannot be released within a prescribed exactitude. The *secondary problem* (also named the *complementary problem*) consists in applying methods to guarantee some "protection requirements" while minimizing the "loss of information". Even if

Section 2 illustrates some standard rules to address the primary problem, this paper concerns only the secondary problem, and in particular it proposes a precise definition of the two main concepts: protection requirements and loss of information. The most popular methodologies for solving the secondary problem are variants of the well-known Cell Suppression and Controlled Rounding methods. These two fundamental methodologies will be next described. Nevertheless, the different methodologies are usually applied without sharing common hypothesis by practitioners, thus making a comparison very difficult even on the same data. Even more, in practice, some implementations cannot inherently guarantee the protection requirements and a great computational effort must be applied to check the proposed output before publication. This checking is called the *Disclosure Auditing* phase and basically it consists in computing lower and upper bounds on the original value for each sensitive cell; in the literature there are several techniques to perform this third phase, including linear and integer programming, the Frechet and Bonferroni bounds, and the Buzzigoli and Giusti's shuttle algorithm (see, e.g., Duncan, Fienberg, Krishnan, Padman and Roehrig [7] for references and generalizations of these techniques).

Cell Suppression is a methodology that allows the practitioner to unpublish the values in some cells while publishing the original values of the others. In particular, once the primary problem was solved, the cells containing sensitive information must be clearly unpublished and they are the *primary suppressions*. Due to the existence of the total marginals in a table, other cells must be also unpublished to guarantee protection of the values under the primary cells, leading to the *secondary suppressions*. They must be identified by solving the so-called *Cell Suppression Problem*, which is a very interesting combinatorial problem widely addressed in the literature. Apart from satisfying the protection requirements, the output of the problem must have a minimum loss of information, which for this methodology could be considered as the sum of the unpublished cell values. See, e.g., [23] for more details on this methodology.

Controlled Rounding is an alternative classical methodology that has not been extensively analysed in the literature, and the aim of this paper is to add new results to fill this gap. When applying a rounding procedure the experts are given a base number and they are allowed to modify the original value of each cell by rounding it up or down to a near multiple of the base number. An output pattern must be associated with the minimum loss of information, which for this methodology can be considered as the distance between the original and the modified tables. In the *Random Rounding* version the experts decide to round up or down each cell by considering a probability that depends on its original cell value, without taking care of the marginal cell values. Therefore, the Random Rounding produces output tables where the marginal values are not the

sum of their internal cells, which is a disadvantage of this rounding version. Another version is the so-called *Controlled Rounding*, where probabilities are not considered and the expert should round up or down all cell values such that all the equations in the table hold in the published table. In the so-called *zero-restricted Controlled Rounding* the original values which are already multiple of the base number cannot be modified. Even not considering protection level requirements, a Controlled Rounding solution may not exist for a given table (e.g., Causey, Cox and Ernst [2] showed a simple infeasible 3-dimensional instance). Kelly, Golden and Assad [17] proposed a branch-and-bound procedure for the case of 3-dimensional tables, and Fischetti and Salazar [12] extended this procedure to 4-dimensional tables. Heuristic methods for finding solutions of this problem on multi-dimensional tables have been proposed by several authors, including Kelly, Golden and Assad [17, 19]. The problem was first introduced by Bacharach [1] in the context of replacing nonintegers by integers in 2-dimensional tabular arrays, and actually it arises in several other applications. It was introduced in a statistical context by Cox and Ernst [3]. In all these articles, the protection requirements are not considered and, therefore, they do not address the real problem arising from the Statistical Disclosure Limitation application. Indeed, a statistical office applying these incomplete approaches must also solve the auditing problem to check the protection of the output (and repeating the procedure when a protection level requirement is violated). We will address in this work the complete problem of finding a solution of the controlled rounding methodology.

In the literature there are several methodologies to protect tables by data perturbation (see, for example, the addition of random noise by Evans, Zayatz and Slanta [10], data swapping by Fienberg, Steele and Makov [11], and Markov perturbation by Duncan and Fienberg [8]) but, as far as we know, they all concern the direct modification of the microdata and, therefore, there is less control on the final protection interval of each cell in the published pattern, and less control on the error added to some aggregated cells. Moreover, as pointed out in Willenborg and De Waal [23], "adding noise to cell values of a table does not guarantee that the additivity of the table is preserved, and if preservation of additivity is important then one should look for other methods". This paper shows how Operations Research can help in Statistical Data Protection by presenting a methodology to address the problem of perturbing the cell values while preserving additivity. The paper presents the complete optimization problem of finding an output of the Controlled Rounding methodology on any type of tables including k-dimensional, linked and hierarchical tables. Solutions of this problem implicitly guarantee the required protection for different sensitive cells and against different attackers, thus saving the effort of solving the Disclosure Auditing

problem. Section 2 introduces the main concepts of the Statistical Disclosure Control problem in a general context. Section 3 considers the well-known *Controlled Rounding Methodology*, presents an integer linear mathematical model, and describes a branch-and-cut algorithm for the exact solution. Since the problem is $\mathcal{NP}$-hard (and very difficult to solve in practice) Section 4 proposes an alternative methodology called *Cell Perturbation* with the advantage that the optimization problem can be solved in polynomial time through a cutting-plane approach which is also detailed. Results from computational experiments using the proposed methods are analyzed in Section 5. Section 6 presents a variant of the whole procedures to guarantee that the optimal solutions are also unbiased. The paper ends with some conclusions in Section 7.

This work has been presented in several seminars on Disclosure Limitation Methods (Plymouth, April 2002; Ottawa, May 2002; Washington, June 2002; London, August 2003), and the concepts and algorithms are being used to develop the τ-ARGUS software package for tabular data protection, an output of the CASC research project (see, e.g., Hundepool [15]).

2 General Situation

A statistical agency is typically provided with a set of n values a_i for $i \in I := \{1, \ldots, n\}$. Vector $a = [a_i : i \in I]$ is known as "nominal table" and satisfies a set of m equations $\sum_{i \in I} m_{ij} y_i = b_j$ for $j \in J := \{1, \ldots, m\}$. For convenience of notation the linear system will be denoted by $My = b$, thus $Ma = b$ holds. Each solution y of $My = b$ is called *congruent table*. Matrix M (with n columns representing the cells and m rows representing the equations) has elements m_{ij} typically in $\{-1, 0, +1\}$ with one -1 per row associated to the marginal-cell variable, while vector b is typically the zero vector. The table in Figure 1 is a 2-dimensional table consisting in $n := 16$ cells and $m := 8$ equations (one from each row and from each column in the table), hence M has two nonzero elements per column. When the table is a 2-dimensional table, then M is the edge-node matrix of a bipartite graph, thus a congruent table can be represented as a flow circulation in a network and some tools from Graph Theory can be applied (see Cox [5]). This is not the case when M is associated to a more complex table (e.g., a 3-dimensional table). Observe that also multi-dimensional tables can be described by a general system $My = b$, where y represents the cell values and the equations define the marginal totals in the table. Since all the here-proposed ideas are based on the general system $My = b$, without any assumption on the structure of the matrix M, then these ideas apply to any type of multi-dimensional table (including hierarchical, linked and other structured tabular data).

Statistical tables typically contain sensitive data, i.e., information that cannot be disclosed since

	A	B	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: Investment of enterprises by activity and region.

they show confidential information on particular respondents. The sensitive cells in a tabular data are typically determined by common-sense *rules*. In the literature and in the practice of statistical offices there are many different rules; see, for example, [23]. As an illustration we next point out the so-called *dominance rule*. We are given the microdata from which the table is computed, and two input numbers α and β (for example, $\alpha := 80$ and $\beta := 3$). Whenever the biggest β respondents from the microdata contributing to value a_p in cell p of the table produce more than α percentage of the total value a_p, then cell p is classified as *sensitive*. This rule is widely known and used by statistical agencies, even if there are some recent critiques on this class of linear sensitivity rules (see, e.g., Robertson and Ethier [20], and Domingo-Ferrer, Torra, Mateo-Sanz and Oganian [6]). Another widely accepted sensitivity measure is what is called the *prior-posterior rule*, based on two integer parameters γ and δ with $\gamma < \delta$: assuming that, prior to the publication of the table, an intruder can estimate the contribution of every other respondent to within δ percent, a cell is considered sensitive if the intruder can estimate the contribution of an individual respondent to that cell to within γ percent posterior if the cell value is published. No matter how the primary problem is solved, we denote the subset of sensitive cells by P. In the example represented in Figure 1, the cell in Activity II and Region C is assumed to be a sensitive cell to be protected because (say) it is publicly known that there is only one respondent in Region C dedicated to Activity II.

In a general situation, all the sensitive cells in a table must be protected against a set of *attackers*. The attackers are the intruders or data snoopers that will analyze the final product data and will try to disclose confidential information. They can also be coalitions of respondents who collude and behave as single intruders. The aim of the Disclosure Limitation Methods is to reduce the risk of them succeeding. The set of attackers will be denoted by K. Each attacker knows the set of linear system $My = b$ plus extra information that bound each cell value. For example, the simplest attacker is the so-called *external intruder* knowing only that unknown cell values are, say, nonnegative. Other more accurate attackers know tighter bounds on the cell values, and they are called *internal attackers*. For example, an internal attacker could be a respondent that had contributed to cell i with, say, 10 units; then he/she knows that $y_i \geq 10$, while the

external attacker only knows $y_i \geq 0$. If the internal attacker also knows that he/she is the only contributor to cell i with value 10, then $10 \leq y_i \leq 10$ when attacking the output data. In general, attacker k is associated with two bounds lb_i^k and ub_i^k such that $a_i \in [lb_i^k \dots ub_i^k]$ for each cell $i \in I$. The literature on statistical disclosure control (see, for example, Willenborg and de Waal [23]) typically addresses the situation where $|K| = 1$, thus protecting the table against the external intruder with only the knowledge of the linear system and some external bounds; nevertheless this is a simplification of the real problem in Disclosure Limitation and statistical offices are interested in protecting tables against several intruders (see, for example, Jewett [16]).

To protect the sensitive cell p containing value a_p in the input table, the statistical office is interested in publishing an output containing several congruent tables, including not only the original nominal table but also others so that no attacker can disclose the private information a_p (neither a narrow approximation). The output of a Disclosure Limitation Method is generally called a *pattern*, and it can assume a particular structure depending on the methodology considered. The sections of this paper deal with two methodologies, and hence illustrate different patterns. In all cases they share the common definition of "protection" defined as follows.

The congruent tables associated to a pattern must differ so that each attacker analysing the pattern will not compute the original value of a sensitive cell within a narrow approximation. For each potential intruder, the idea is to define a protection range for p and to demand that the a-posteriori protection be such that any value in the range is potentially the correct cell value. To be more precise, by observing the published pattern, attacker k will compute an interval $[\underline{y}_p^k \dots \overline{y}_p^k]$ of possible values for each sensitive cell p. The pattern will be considered *valid* to protect cell p against attacker k if the computed interval is "wide enough". To set up the definition of "wide enough" in a precise way, the statistical office gives three input parameters for each attacker k and each sensitive cell p with nominal value a_p:

- Upper Protection Level: it is a number UPL_p^k representing a desired lower bound for $\overline{y}_p^k - a_p$;
- Lower Protection Level: it is a number LPL_p^k representing a desired lower bound for $a_p - \underline{y}_p^k$;
- Sliding Protection Level: it is a number SPL_p^k representing a desired lower bound for $\overline{y}_p^k - \underline{y}_p^k$.

The values of these parameters can also be defined by using common-sense rules. For example, simple values for the protection levels are percentages of the nominal value of the cell (for example, 20%, 15% and 40%, respectively). In more sophisticated situations where intruder k is an original respondent (i.e., an internal attacker), the protection levels could be chosen to be proportional to

his/her contributions s_p^k to the nominal value of the cell a_p and/or to the complement $a_p - s_p^k$ (see, e.g, Sande [21]). Of course, an elementary assumption is that

$$lb_p^k \leq a_p - LPL_p^k \leq a_p \leq a_p + UPL_p^k \leq ub_p^k$$

and

$$ub_p^k - lb_p^k \geq SPL_p^k,$$

for each attacker k and each sensitive cell p. For notational convenience, let us also define absolute protection levels and relative nominal bounds:

$$lpl_p^k := a_p - LPL_p^k,$$

$$upl_p^k := a_p + UPL_p^k,$$

$$LB_i^k := a_i - lb_i^k,$$

$$UB_i^k := ub_i^k - a_i.$$

In the example represented in Figure 1 the statistical office could be interested in protecting the sensitive cell (Activity II, Region C) against one attacker with a lower protection level of 10 units, an upper protection level of 12 units, and a sliding protection level of 0 units. Figure 2 gives four different patterns, each one coming out from a different methodology. Pattern (a) corresponds to the classical Cell Suppression Methodology (see, e.g., [13]) and Pattern (b) is an output of the recent Interval Publication Methodology (see, e.g., [14]). Pattern (c) is a solution from the classical Controlled Rounding Methodology described in Section 3 and Pattern (d) is a solution from the new Cell Perturbation Methodology introduced in Section 4.

Given a pattern, the mathematical problems of computing values $\underline{y}_p^k$ and $\overline{y}_p^k$ are known as *attacker problems* for cell p and attacker k. The overall problem of solving the attacker problems for all cells is called *Disclosure Auditing Problem*, which should not be confused with the Disclosure Auditing Phase mentioned in Section 1 and which is an unnecessary phase for the methodologies proposed in this paper since they will implicitly guarantee the protection requirements on the output pattern. The attacker problems associated with cell p and attacker k can be formulated as two Linear Programming (LP) models on an array of variables $y = [y_i : i \in I]$ representing a table. Indeed, an attacker problem is

$$\underline{y}_p^k := \min y_i$$

subject to

$$\begin{aligned} My &= b \\ lb_i^k \leq y_i &\leq ub_i^k \qquad \text{for all } i \in I, \end{aligned}$$

	A	B	C	Total
Activity I	20	50	10	80
Activity II	*	19	*	49
Activity III	*	32	*	61
Total	45	101	44	190

(a) Cell Suppression pattern.

	A	B	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

(b) Interval Publication pattern.

	A	B	C	Total
Activity I	20	50	10	80
Activity II	10	20	20	50
Activity III	15	30	15	60
Total	45	100	45	190

(c) Controlled Rounding pattern (base 5).

	A	B	C	Total
Activity I	20	50	10	80
Activity II	7	16	26	49
Activity III	18	35	8	61
Total	45	101	44	190

(d) Cell Perturbation pattern (base 5).

Figure 2: Four different patterns.

plus a set of additional constraints that make y feasible in accordance with the published pattern. The precise additional constraints depend on the structure of the pattern, and therefore on the considered methodology. The other attacker problem is obtained by replacing the objective function with $\overline{y}_p^k := \max y_i$. Section 3 of this paper shows the precise attacker problems for the Controlled Rounding methodology and Section 4 for the Cell Perturbation methodology.

Finally, among all possible valid patterns, the statistical office is interested in finding one with minimum information loss. The *information loss* of a pattern is intended to be a measure of the number of congruent tables in the pattern. Indeed, a valid pattern must always allow the nominal table to be a feasible congruent table, but it must also contain other different congruent tables so as to keep the risk of disclosure controlled. For example, when the pattern contains only the original table (because there is no sensitive data to be protected) then the loss of information is clearly zero. The precise definition of loss of information depends on the structure of the pattern, and hence on the methodology to be considered. In practice, since it is not always easy to count the number of congruent tables in a pattern from the point of view of an intruder k, the loss of information of a pattern is replaced by the sum of the loss of information of its cells. In this case, the individual cost for cell p is generally proportional to the difference between the worse-case situations (i.e., to $\overline{y}_p^k - \underline{y}_p^k$), it is proportional to the number of respondents contributing to the cell value a_p, or it is simply a positive fixed cost when a_p is not published (i.e., when $\overline{y}_p^k - \underline{y}_p^k > 0$). It could be interesting to use a definition of loss of information for a pattern given by a distance between the original table and, for example, the most-probable table for the intruder k among the congruent ones with the final pattern, but this is not an easy task without knowing the probability distribution of y_p in $[\underline{y}_p^k \dots \overline{y}_p^k]$.

In practice most of the available software is based on techniques for finding "good" patterns with no inherent guarantee on the protection level requirements, i.e. not necessarily valid (see, for example, [7]). Therefore, it is necessary to check the proposed pattern before it is made public by solving the Disclosure Auditing Problem, and to try a different technique when the result is negative. It is well-known (see, for example, Duncan, Fienberg, Krishnan, Padman and Roehrig [7]) that auditing a pattern could consume many computing resources. In the next sections we introduce precise methodologies to find a valid pattern (if any exists) with minimum (or near-minimum) information loss, hence the Disclosure Auditing Phase is not required.

3 Controlled Rounding Methodology

In *Controlled Rounding Methodology* we are provided with an input base number r_i for each cell i. In practice, the statistical office uses a common base number r_i for all cells, but the method can also be applied when there are different base numbers, as required by some practitioners (e.g., when protecting some hierarchical tables, bigger base numbers are preferred on top levels than on low levels).

Let us denote by $\lfloor a_i \rfloor$ the multiple of r_i obtained by rounding down a_i, and by $\lceil a_i \rceil$ the multiple of r_i obtained by rounding up a_i. To follow the well-accepted *zero-restricted* version of the Controlled Rounding Methodology, if r_i is such that $\lfloor a_i \rfloor = \lceil a_i \rceil$ then we redefine $r_i := 0$, thus $r_i = \lceil a_i \rceil - \lfloor a_i \rfloor$ for all $i \in I$.

A pattern in the Controlled Rounding Methodology is a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in \{\lfloor a_i \rfloor, \lceil a_i \rceil\}. \tag{1}$$

Figure 2(c) gives an example of pattern when $r_i := 5$ ($i \in I$) for the instance in Figure 1. The values r_i are published with the output pattern by the statistical office, thus they are assumed to be known by the attackers. The feasible region for the attacker problems associated with attacker k is defined by

$$\begin{aligned} My &= b \\ v_i - r_i \le y_i \le v_i + r_i &\qquad \text{for all } i \in I \\ lb_i^k \le y_i \le ub_i^k &\qquad \text{for all } i \in I. \end{aligned}$$

The natural concept of "loss of information" of a cell is defined as the difference between the nominal value and the published value, and then the loss of information of a pattern is the sum of all the individual loss of information:

$$\delta(v, a) = \sum_{i \in I} |v_i - a_i| \tag{2}$$

We now present a mathematical model for the combinatorial problem of finding a protected controlled rounding pattern with minimum loss of information, and then we describe an algorithm for solving this model. The optimization problem is refereed as *Controlled Rounding Problem* (CRP).

3.1 Mathematical model

Let us consider a binary variable x_i for each cell i, representing

$$x_i = \begin{cases} 0 & \text{if } v_i = \lfloor a_i \rfloor, \\ 1 & \text{if } v_i = \lceil a_i \rceil, \end{cases}$$

i.e., $x_i = 1$ if and only if the published value v_i is obtained by rounding up a_i. Note that when a solution x_i is given, then the published table is determined by $v_i := \lfloor a_i \rfloor + r_i x_i$ for all $i \in I$, and therefore the attacker problems for a given pattern $[x_i : i \in I]$ have a feasible region defined by

$$\left.\begin{array}{rl} My = b & \\ \lfloor a_i \rfloor + r_i x_i - r_i \leq y_i \leq \lfloor a_i \rfloor + r_i x_i + r_i & \text{for all } i \in I \\ lb_i^k \leq y_i \leq ub_i^k & \text{for all } i \in I. \end{array}\right\} \tag{3}$$

The loss of information of a cell i can now be written as a constant when $x_i = 0$, plus a (positive or negative) parameter w_i if $x_i = 1$. For example, if $w_i = \lceil a_i \rceil + \lfloor a_i \rfloor - 2a_i$, then w_i represents the cost of rounding up instead of rounding down from value a_i. Then, the loss of information (2) of the pattern v defined by $[x_i : i \in I]$ is a constant plus $\sum_{i \in I} w_i x_i$.

The CRP is to find a value for each x_i such that the total loss of the information in the released pattern is minimized, i.e.:

$$\min \sum_{i \in I} w_i x_i \tag{4}$$

subject to, for each sensitive cell $p \in P$ and for each attacker $k \in K$,

- the upper protection requirement must be satisfied, i.e.:

$$\max \{ y_p : (3) \text{ holds } \} \geq upl_p^k \tag{5}$$

- the lower protection requirement must be satisfied, i.e.:

$$\min \{ y_p : (3) \text{ holds } \} \leq lpl_p^k \tag{6}$$

- the sliding protection requirement must be satisfied, i.e.:

$$\max \{ y_p : (3) \text{ holds } \} - \min \{ y_p : (3) \text{ holds } \} \geq SPL_p^k \tag{7}$$

Finally, each variable must assume value 0 or 1, i.e.:

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I. \tag{8}$$

Mathematical model (4)–(8) contains all the requirements of the statistical office (in accordance with the definition given in Section 2), and therefore a solution $[x_i^* : i \in I]$ defines an optimal protected controlled rounding pattern. The inconvenience is that it is not an easy model to be solved, since it does not belong to the standard (Mixed) Integer Linear Programming (ILP). In fact, the existence of optimization problems as part of the constraints of a main optimization problem

classifies the model in the so-called "Bilevel Mathematical Programming", which today is not provided with effective solution algorithms. Observe that a drawback of model (4)–(8) is not the number of variables, which is at most the number of cells, both for the master optimization problem (first optimization level) and for each subproblem (second optimization level). The inconvenience of model (4)–(8) is the fact that there are optimization problems nested in the two levels. A way to avoid this inconvenience is to look for a transformation into a classical ILP model, as it is next done.

Condition (5) can be replaced by the existence of a congruent table $[f_i^{kp} : i \in I]$ such that it is feasible (i.e., it satisfies (3)) and it guarantees the upper protection level requirement, i.e.:

$$f_p^{kp} \geq upl_p^k.$$

In the same way, the optimization problem in condition (6) can be replaced by the existence of a congruent table $[g_i^{kp} : i \in I]$ such that it is also feasible (i.e., it satisfies (3)) and it guarantees the lower protection level requirement, i.e.:

$$g_p^{kp} \leq lpl_p^k.$$

Finally, the two optimization problems in condition (7) can be replaced by the above congruent tables if they guarantee the sliding protection level, i.e.:

$$f_p^{kp} - g_p^{kp} \geq SPL_p^k.$$

Figure 3 shows a first attempt to have an ILP model, where x_i, f_i^{kp}, g_i^{kp} are the variables.

As mentioned by Fischetti and Salazar [13] on a similar model for the Cell Suppression Methodology, an important disadvantage of model in Figure 3 is the large number of variables even for small instances. A way to skip this disadvantage is by projecting away the continuous variables f_i^{kp} and g^{kp_i} by using the Benders' Decomposition technique for mixed integer programming models as it is next described.

Imposing the upper protection level requirements

Based on the Farkas' Lemma, it is possible to replace the second level subproblems of model (4)–(8) by linear constraints on the x_i variables. Indeed, assuming that values y_i in a congruent table are continuous numbers, the two LP models in conditions (5)–(7) can be rewritten in their dual format. More precisely, by Duality Theory in Linear Programming (see, for example, Wolsey [24]):

$$\max \{y_p : (3) \text{ holds } \}$$

$$\min \sum_{i \in I} w_i x_i$$

subject to:

$$\begin{array}{ll} \sum_{i \in I} m_{ij}(\lfloor a_i \rfloor + r_i x_i) = b_j & \text{for all } j \in J \\ x_i \in \{0,1\} & \text{for all } i \in I \end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\begin{array}{ll} \sum_{i \in I} m_{ij} f_i^{kp} = b_j & \text{for all } j \in J \\ lb_i^k \leq f_i^{kp} \leq ub_i^k & \text{for all } i \in I \\ \lfloor a_i \rfloor + r_i x_i - r_i \leq f_i^{kp} \leq \lfloor a_i \rfloor + r_i x_i + r_i & \text{for all } i \in I \\ \sum_{i \in I} m_{ij} g_i^{kp} = b_j & \text{for all } j \in J \\ lb_i^k \leq g_i^{kp} \leq ub_i^k & \text{for all } i \in I \\ \lfloor a_i \rfloor + r_i x_i - r_i \leq g_i^{kp} \leq \lfloor a_i \rfloor + r_i x_i + r_i & \text{for all } i \in I \\ f_p^{kp} \geq upl_p^k & \\ g_p^{kp} \leq lpl_p^k & \\ f_p^{kp} - g_p^{kp} \geq SPL_p^k. & \end{array}$$

Figure 3: Basic ILP model for Controlled Rounding.

is equivalent to

$$\min \sum_{j \in J} \gamma_j b_j + \sum_{i \in I} [\alpha_i^1 ub_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i) - \beta_i^1 lb_i^k - \beta_i^2(\lfloor a_i \rfloor + r_i x_i - r_i)]$$

subject to

$$\left.\begin{array}{rl} \alpha_p^1 + \alpha_p^2 - \beta_p^1 - \beta_p^2 + \sum_{j \in J} m_{pj}\gamma_j = 1 & \\ \alpha_i^1 + \alpha_i^2 - \beta_i^1 - \beta_i^2 + \sum_{j \in J} m_{ij}\gamma_j = 0 & \text{for all } i \in I \setminus \{p\} \\ \alpha_i^1 \geq 0 & \text{for all } i \in I \\ \alpha_i^2 \geq 0 & \text{for all } i \in I \\ \beta_i^1 \geq 0 & \text{for all } i \in I \\ \beta_i^2 \geq 0 & \text{for all } i \in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j \in J, \end{array}\right\} \quad (9)$$

Because of (9) and $[a_i : i \in I]$ is a consistent table, we have

$$\sum_{j \in J} \gamma_j b_j + \sum_{i \in I}(\alpha_i^1 a_i + \alpha_i^2 a_i - \beta_i^1 a_i - \beta_i^2 a_i) = \sum_{i \in I}\sum_{j \in J} \gamma_j m_{ij} a_i + \sum_{i \in I}(\alpha_i^1 + \alpha_i^2 - \beta_i^1 - \beta_i^2)a_i = a_p.$$

Hence the above LP model can be rewritten as

$$a_p + \min \sum_{i \in I} (\alpha_i^1 UB_i^k + \alpha_i^2 (\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2 (a_i - \lfloor a_i \rfloor - r_i x_i + r_i))$$

subject to $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \gamma_j$ satisfying (9).

From this observation, condition (5) can be now written as:

$$\textstyle\sum_{i \in I} (\alpha_i^1 UB_i^k + \alpha_i^2 (\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2 (a_i - \lfloor a_i \rfloor - r_i x_i + r_i)) \geq UPL_p^k$$

for all $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \gamma_j$ satisfying (9).

In other words, the last system defines a family of linear constraints, in the x-variables only, representing condition (5) which concerns the upper protection level requirement for sensitive cell p and attacker k.

Notice that this family contains in principle an infinite number of constraints, each associated with a different point $[\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2 : i \in I; \gamma_j : j \in J]$ of the polyhedron defined by (9). However, it is well known that only the extreme points (and rays) of such a polyhedron can lead to undominated constraints, i.e., a finite number of such constraints is sufficient to impose the upper protection level requirement for a given sensitive cell p and a given attacker k.

Imposing the lower protection level requirements

In a similar way, the optimization problem in (6) is:

$$-\max \{-y_p : (3) \text{ holds } \},$$

which, by the Duality Theory, is equivalent to

$$-\min \sum_{j \in J} \gamma_j b_j + \sum_{i \in I} [\alpha_i^1 ub_i^k + \alpha_i^2 (\lfloor a_i \rfloor + r_i x_i + r_i) - \beta_i^1 lb_i^k - \beta_i^2 (\lfloor a_i \rfloor + r_i x_i - r_i)]$$

subject to

$$\left.\begin{array}{rl} \alpha_p^1 + \alpha_p^2 - \beta_p^1 - \beta_p^2 + \sum_{j \in J} m_{pj} \gamma_j = -1 & \\ \alpha_i^1 + \alpha_i^2 - \beta_i^1 - \beta_i^2 + \sum_{j \in J} m_{ij} \gamma_j = 0 & \text{for all } i \in I \setminus \{p\} \\ \alpha_i^1 \geq 0 & \text{for all } i \in I \\ \alpha_i^2 \geq 0 & \text{for all } i \in I \\ \beta_i^1 \geq 0 & \text{for all } i \in I \\ \beta_i^2 \geq 0 & \text{for all } i \in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j \in J. \end{array}\right\} \quad (10)$$

As it was done before, the above linear program can be rewritten as

$$-a_p - \min \sum_{i \in I} (\alpha_i^1 UB_i^k + \alpha_i^2 (\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2 (a_i - \lfloor a_i \rfloor - r_i x_i + r_i))$$

subject to $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \gamma_j$ satisfying (10).

From this observation, condition (6) can be now written as:

$$\sum_{i\in I}(\alpha_i^1 UB_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2(a_i - \lfloor a_i \rfloor - r_i x_i + r_i)) \geq LPL_p^k$$

for all $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \gamma_j$ satisfying (10).

In other words, the last system defines a family of linear constraints, in the x-variables only, representing condition (6) which concerns the lower protection level requirement for sensitive cell p and attacker k.

Imposing the sliding protection level requirements

As to the sliding protection level for sensitive cell p and attacker k, the requirement is that

$$SPL_p^k \leq \max\{y_p : (3) \text{ hold } \} + \max\{-y_p : (3) \text{ hold } \}.$$

Again, by LP Duality, this condition is equivalent to

$$SPL_p^k \leq$$

$$\min\{\textstyle\sum_{j\in J}\gamma_j b_j + \sum_{i\in I}[\alpha_i^1 ub_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i) - \beta_i^1 lb_i^k - \beta_i^2(\lfloor a_i \rfloor + r_i x_i - r_i)] : (9) \text{ holds } \} +$$

$$\min\{\textstyle\sum_{j\in J}\gamma_j b_j + \sum_{i\in I}[\alpha_i^1 ub_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i) - \beta_i^1 lb_i^k - \beta_i^2(\lfloor a_i \rfloor + r_i x_i - r_i)] : (10) \text{ holds } \}.$$

Therefore, the feasibility condition can now be formulated by requiring

$$SPL_p^k \leq \textstyle\sum_{j\in J}(\gamma_j + \gamma'_j) b_j +$$

$$\textstyle\sum_{i\in I}[(\alpha_i^1 + \alpha_i'^1) ub_i^k + (\alpha_i^2 + \alpha_i'^2)(\lfloor a_i \rfloor + r_i x_i + r_i) - (\beta_i^1 + \beta_i'^1) lb_i^k - (\beta_i^2 + \beta_i'^2)(\lfloor a_i \rfloor + r_i x_i - r_i)]$$

for all $\alpha^1, \alpha^2, \beta^1, \beta^2, \gamma$ satisfying (9) and for all $\alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma'$ satisfying (10),

or, equivalently,

$$\textstyle\sum_{i\in I}(\alpha_i^1 + \alpha_i'^1) UB_i^k + (\alpha_i^2 + \alpha_i'^2)(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) +$$

$$(\beta_i^1 + \beta_i'^1) LB_i^k + (\beta_i^2 + \beta_i'^2)(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq SPL_p^k$$

for all $\alpha^1, \alpha^2, \beta^1, \beta^2, \gamma$ satisfying (9) and for all $\alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma'$ satisfying (10).

Overall model

Figure 4 summarizes an alternative model to (4)–(8) with only the 0-1 variables. The inequalities in the model are called *capacity constraints* in analogy with similar constraints introduced in Fischetti and Salazar [13] for the Cell Suppression Methodology to enforce a sufficient "capacity" of certain

$$\min \sum_{i \in I} w_i x_i$$

subject to:

$$\begin{array}{ll} \sum_{i \in I} m_{ij}(\lfloor a_i \rfloor + r_i x_i) = b_j & \text{for all } j \in J \\ x_i \in \{0,1\} & \text{for all } i \in I \end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\sum_{i \in I} \alpha_i^1 UB_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq UPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (9)}$$

$$\sum_{i \in I} \alpha_i'^1 UB_i^k + \alpha_i'^2(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i'^1 LB_i^k + \beta_i'^2(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq LPL_p^k$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (10)}$$

$$\sum_{i \in I} (\alpha_i^1 + \alpha_i'^1) UB_i^k + (\alpha_i^2 + \alpha_i'^2)(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) +$$
$$(\beta_i^1 + \beta_i'^1) LB_i^k + (\beta_i^2 + \beta_i'^2)(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq SPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (9) and}$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (10).}$$

Figure 4: Second ILP model for Controlled Rounding.

cuts in the network representation of problem on 2-dimensional tables with marginals. Intuitively, the capacity constraints force to modify a sufficient number of cell values whose positions within the table and contributions to the overall protection are specified by the dual variables $(\alpha, \alpha', \beta, \beta', \gamma)$ of the attacker subproblems.

The overall model is appropriate for being solved within a branch-and-cut framework. Indeed, the advantage of model in Figure 4 when compared to model in Figure 3 it that the first one contains only the 0-1 variables, which are order of the number of cells. At a disadvantage one can observe that it has a huge number of constraints, one for each (extreme) point of (9) and (10). Nevertheless, not all these constraints are necessary from the beginning of the resolution and, given a pattern $[x_i^* : i \in I]$, one can generate a most violated inequality from these families by solving a linear program. More precisely, given a (possibly fractional) pattern $[x_i^* : i \in I]$, to check if there is a violated capacity constraint not yet generated, one has to solve the two linear programs which

maximize and minimize y_p, respectively, subject to (3) for all sensitive cell p and all attacker k. If a protection level requirement does not hold, then the dual variables of the corresponding linear programs define a violated capacity constraint. Hence, it is not necessary to work in practice with the feasible regions (9) and (10), which would imply a large number of unnecessary variables. On the contrary, the implementation of the described separation procedure can be done directly with the primal version of the attacker problems, working only with the dual variables $[\gamma_j : j \in J]$. Indeed, each extreme point of (9) (resp. (10)) has at most one of the four components $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2$ at a non-zero value for each cell i, and this non-zero value can be computed using $[\gamma_j : j \in J]$ and the equation in (9) (resp. (10)).

In practice an important observation is that we can fix some variables in a preprocessing: $x_i = 1$ if $\lfloor a_i \rfloor < lb_i^k$ and $x_i = 0$ if $\lceil a_i \rceil > ub_i^k$. Moreover, we can also strengthen a generated capacity constraint $\sum_{i \in I} \delta_i x_i \geq \delta_0$ with $\delta_i \geq 0$ for all $i \in I$ (thus $\delta_0 > 0$) by redefining the left-hand side coefficients $\delta_i := \min\{\delta_i, \delta_0\}$ because $x_i \in \{0, 1\}$. Special cases of these stronger inequalities are the trivial fixings: $x_i = 1$ if $UPL_i^k > \lceil a_i \rceil - a_i$ and $x_i = 0$ if $LPL_i^k > a_i - \lfloor a_i \rfloor$, arising when there is one $i \in I$ with $\delta_i \neq 0$.

Another important observation to make the algorithm works on large instances is the following. Not all the attacker subproblems should be solved to check the protection levels requirements. Indeed, one can arrange the attacker problems to be solved in a list $\mathcal{L}$ sorted by decreasing protection levels. The first subproblem is taken from $\mathcal{L}$ and solved with an LP-solver with a limit in the objective function equal to the protection level. If the limit is achieved then the protection level is attained even if we have not computed the exact optimal value. Otherwise the dual values define a violated capacity constraints. In all cases, the primal solution of the solved attacker subproblem (which is a feasible pattern) is used to check other protection levels for this attacker, and the correspondent subproblems are removed from $\mathcal{L}$ when they are satisfied. The next subproblem to be taken from $\mathcal{L}$ is the one associated to the protection level close to be satisfied by the last primal solution (ties are broken by considering the order in the list), and it is solved with the primal-simplex LP-solver. In this way, many attacker subproblems do not need to be solved, and when a subproblem must be solved then the LP-solver will benefit from a previous good primal solution and from the fact that the objective value is limited by the protection level. This elaborated procedure is very important to reduce the computational effort to generate the violated cuts, which is similar in a sense to solve the *auditing phase* and therefore it could be a time-consuming task (see, e.g., Duncan, Fienberg, Krishnan, Padman and Roehrig [7]).

3.2 Dealing with infeasibility

The version of the Controlled Rounding methodology modeled in the previous section is known as *zero-restricted*, and can lead to an infeasible optimization problem as observed by Causey, Cox and Ernst [2]. This is due to the strong constraints (1), and this section presents a way of relaxing such conditions to possibly find a congruent table to be released.

When the zero-restricted problem is infeasible, the statistical office is still interested in rounding the cell values and producing a congruent table protected according to the given protection level requirements. To this end, we look for congruent tables where a cell value is allowed to be rounded to a multiple of the base number different than the closest ones. To keep controlled the distance of a rounded value v_i from the original value a_i, we solve a sequence of problems where $|v_i - a_i| \leq sr_i$ for all $i \in I$ and for a give parameter s. Starting with $s = 2$, the parameter s is increased by one unit through the sequence. The problem solved at each iteration can be modeled in a similar way as done for the zero-restricted version. Indeed, instead of one 0-1 variable x_i, we now need two integer variables associated to each cell i: a variable x_i^+ giving the number of roundings below $\lfloor a_i \rfloor$, and a second integer variable x_i^+ giving the number of roundings over $\lfloor a_i \rfloor$. The variable x_i^- can assume values in $\{0, 1, \ldots, s\}$ and the variable x_i^+ can assume values in $\{0, 1, \ldots, s-1\}$. Since the value of s will be published by the statistical office when publishing the output to maximize the utility of the released data, the attacker problems are defined by:

$$\left.\begin{array}{rll} My = b & & \\ \lfloor a_i \rfloor - r_i x_i^- + r_i x_i^+ - sr_i \leq y_i \leq \lfloor a_i \rfloor - r_i x_i^- + r_i x_i^+ - sr_i & \text{for all } i \in I & \\ lb_i^k \leq y_i \leq ub_i^k & \text{for all } i \in I. & \end{array}\right\} \qquad (11)$$

Then the problem of finding a protected pattern (if any exists) can be modeled in a similar way as done in Figures 4 and 3, which correspond to the problem when $s = 1$. For simplicity, we will not go into more details.

Clearly, a disadvantage of this way of escaping from the infeasibility of the zero-restricted version is the number of iterations that this method requires and the complexity of the integer programming problem of each iteration. The next section presents a simpler alternative leading to a new methodology in Statistical Data Protection.

4 Cell Perturbation Methodology

The main disadvantage of the Controlled Rounding methodology is that a protected pattern does not always exist due to the tight constraints (1). Therefore, a different way of ensuring the existence of protected patterns is to relax conditions (1) in the controlled rounding model (e.g., consider the

linear programming relaxations of models in Figures 3 and 4) and to look for a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in [\lfloor a_i \rfloor \dots \lceil a_i \rceil]. \tag{12}$$

where $\lfloor a_i \rfloor$ and $\lceil a_i \rceil$ are given in advance from the statistical office such that $\lfloor a_i \rfloor \leq a_i \leq \lceil a_i \rceil$. These extreme values can be defined as the nearest numbers to a_i which are multiples of a given number (i.e., defined as in the standard Controlled Rounding methodology from a given base number), but they can also be the two values within a given difference with respect to a_i (i.e., $\lfloor a_i \rfloor := a_i - t_i$ and $\lceil a_i \rceil := a_i + t_i$ for a given base number $t_i > 0$). Figure 2 (d) shows a possible Cell Perturbation pattern for the nominal table in Figure 1. Table v is then a pattern in the *Cell Perturbation Methodology* and the novelty with respect to the controlled rounding is that now v_i can be any value between the two extremes of the interval $[\lfloor a_i \rfloor \dots \lceil a_i \rceil]$. As in the Controlled Rounding methodology, the loss of information of a cell i could be defined to be proportional to $|v_i - a_i|$, and the "loss of information" of a pattern is the sum of the loss of information of all the cells.

Obviously, if all constraints (1) are removed and no new one is added to the continuous relaxation of a model minimizing the non-linear function $\sum_{i \in I} |v_i - a_i|$ over the feasible region defined by (5)–(8), then the valid pattern with minimum loss of information is the nominal table a. A way to avoid this disappointing solution is to keep some constraints from (1) (for example, the one concerning the sensitive cells) or simply require that the published values in each sensitive cell must be equal to some given values (for example, $v_p = \lceil a_i \rceil$ for all $p \in P$). Practitioners in statistical offices prefer another way of avoiding the nominal table as published table: it consists in defining a different objective function. Indeed, by considering the objective as the distance between each published value v_i and the value in $\{\lfloor a_i \rfloor, \lceil a_i \rceil\}$ closest to a_i we get the same criteria used in the classical Controlled Rounding methodology, and allow the objective function to be linear on the variables x_i.

Let $r_i := \lceil a_i \rceil - \lfloor a_i \rfloor$ a (possibly) known information for attackers. Then the attacker problems associated with attacker k are now exactly the same as in the Controlled Rounding Methodology, i.e.

$$\begin{aligned} My &= b \\ v_i - r_i \leq y_i &\leq v_i + r_i \qquad \text{for all } i \in I \\ lb_i^k \leq y_i &\leq ub_i^k \qquad \text{for all } i \in I. \end{aligned}$$

As in the Controlled Rounding methodology, a necessary (but not sufficient) condition for feasibility is that $\max_{k \in K}\{SPL_i^k, UPL_i^k + LPL_i^k\} \leq 2r_i$ for all $i \in I$.

Mathematical models for the underlying optimization problem in this Cell Perturbation Method-

ology are simply given by the continuous relaxations of the (Mixed) Integer Linear Programming models given in the previous section. Indeed, the published value can be modeled by $v_i := \lfloor a_i \rfloor + r_i x_i$ where x_i is now a continuous variable in $[0, 1]$. This output must be an additive table, which is guaranteed by constraints

$$\sum_{i \in I} m_{ij}(\lfloor a_i \rfloor + r_i x_i) = b_j \qquad \text{for all } j \in J$$

Other constraints in Figure 3 (or in Figure 4), except the integrality on the x_i variables, impose the existence of additive tables to guarantee the protection level requirements.

As described for solving the classical Controlled Rounding Problem, a branch-and-cut technique is appropriated for solving this model with an exponential number of constraints. Hence, we do not need to solve the full master model, but a sequence of relaxed problems. Then violated constraints (if any) can be easily generated by solving the attacker subproblems. This phase of finding potential violated constraints for a given fractional solution of the relaxed master problem is known as *separation problem*.

Clearly, the capacity constraints cannot be strengthened as mentioned in the classical Controlled Rounding Methodology. Still, relevant constraints in practice are the following trivial inequalities:

$$\frac{UPL_p^k - \lfloor a_p \rfloor - r_p + a_p}{r_p} \leq x_p \leq \frac{r_p - LPL_p^k - \lfloor a_p \rfloor + a_p}{r_p}$$

for all sensitive cell p, which are capacity constraints $\sum_{i \in I} \delta_i x_i \geq \delta_0$ when $\delta_i = 0$ for all $i \in I \setminus \{p\}$. In particular, the left-hand side inequality arises when there is one positive dual variable which is $\alpha_p^2 = 1$ and the right-hand side inequality arises when there is one positive dual variable which is $\beta_p^2 = 1$. Note that both are solutions of (9) and (10), respectively. Other trivial valid constraints are

$$\frac{lb_i^k - \lfloor a_i \rfloor}{r_i} \leq x_i \leq \frac{ub_i^k - \lfloor a_i \rfloor}{r_i}$$

In our implementation we have generated all these bound constraints in the initial step of our algorithm, hence these trivial constraints do not appear as violated constraints while solving the separation problems.

5 Computational Results

We have implemented the branch-and-cut algorithm described in Section 3 for solving the classical Controlled Rounding Problem, and the cutting-plane algorithm described in Section 4 for solving the new Cell Perturbation Problem. The implementation has been done in ANSI C using the

Microsoft Visual C 6.0 compiler and the branch-and-cut framework of CPLEX 8.0. The experiments have been executed on a personal computer with a PC Pentium IV 2.5 Ghz under Microsoft Windows XP.

Both codes succeeded in finding optimal patterns for a collection of benchmark instances received from the "Department for Work and Pensions", United Kingdom. This collection of instances contains real-world data containing confidential information, and were provided for this research under a confidentiality agreement. Therefore the data are not available for other researchers. Still, we can mention that they concern with neighbourhood statistics considering a hierarchical subdivision of Great Britain: 10524 wards, 408 local authorities, 55 counties, 11 government office regions, 3 countries and 1 kingdom. The data contains different information for each of these 11002 groups. In particular, the larger table corresponds to the Income Support at August 2000 and consists of 20 values for each group, some being partial marginal values of others. This large table is modelled through 220040 cells and 75572 links, and both algorithms found optimal pattern in less than one minute. A relevant feature of the tables in our collection is that the pattern found by both algorithms was the same in all cases. In other words, the continuous relaxation of the integer models produced an integer solutions in all the instances. To explain this result we illustrate the structure of our real-world table with the dummy data in Figure 5, where an original and a rounded table with a similar structure to our real-world tables are given. The column named "total" is the sum of "male" and "female", but also the sum of "young" and "adult", and the sum of "thin" and "fat". Moreover the row "England" is the sum of "North East",...,"South West", and the row "Great Britain" is the sum of "Wales", "Scotland" and "England". Then the rounding problems can be modelled as finding min-cost flow circulations on three capacitated networks (sex, age and weight) plus the additional constraints that the flow of some arcs (corresponding to the marginal cells) must coincide. An optimally rounded table of the original table with 220040 cells and 75572 equations was found by our branch-and-cut algorithm in 21 seconds of a personal computer Pentium 2533 Mhz.

The second example refers to a table that appeared in the 2001 Scottish Census of Population. For this data, various pre-tabulation disclosure control techniques were considered sufficient to protect the confidentiality of the respondents. So there is no need to apply a further stage of rounding for confidentiality protection. Nevertheless, they form an interesting example of data (a) that are close in form to unprotected data, and (b) that are also in the public domain, so they can be used to test our algorithm. The table we used was of age (20 levels) by sex (2 levels) by living arrangements (7 levels), together with a number of marginal tables. Considering geography,

Unrounded data	total	male	female	young	adult	thin	fat
North East	60593	29225	31368	13856	46737	34565	26028
North West	174414	78129	96285	25673	148741	3432	170982
Yorkshire and Humberside	108769	46119	62650	2342	106427	32223	76546
East Midlands	93346	43201	50145	23443	69903	23434	69912
West Midlands	131817	61046	70771	23878	107939	432	131385
East	107060	47376	59684	24532	82528	34233	72827
London	110811	49053	61758	17635	93176	3423	107388
South East	123359	50949	72410	34223	89136	4567	118792
South West	119863	44718	75145	35980	83883	56356	63507
Wales	95388	49579	45809	34989	60399	6454	88934
Scotland	124678	61327	63351	36789	87889	5643	119035
England	1030032	449816	580216	201562	828470	192665	837367
Great Britain	1250098	560722	689376	273340	976758	204762	1045336

Rounded data ($r_i = 5$)	total	male	female	young	adult	thin	fat
North East	60595	29225	31370	13855	46740	34565	26030
North West	174415	78130	96285	25675	148740	3430	170985
Yorkshire and Humberside	108770	46120	62650	2340	106430	32225	76545
East Midlands	93345	43200	50145	23445	69900	23435	69910
West Midlands	131815	61045	70770	23875	107940	430	131385
East	107060	47375	59685	24530	82530	34235	72825
London	110810	49055	61755	17635	93175	3420	107390
South East	123360	50950	72410	34225	89135	4570	118790
South West	119860	44715	75145	35980	83880	56355	63505
Wales	95390	49580	45810	34990	60400	6455	88935
Scotland	124675	61325	63350	36790	87885	5640	119035
England	1030030	449815	580215	201560	828470	192665	837365
Great Britain	1250095	560720	689375	273340	976755	204760	1045335

Figure 5: Dummy table with a similar structure of our real-world data.

there were 32 local authorities (LADs), within which were 1176 wards. The total number of cells was 431613, and the number of equations was 105960. Each geographical area has 80 equations, and there are 1209 geographical areas (1176 wards + 32 LADs + 1 country). The full problem in its zero-restricted form was difficult to solve with $r_i = 5$: no solution was found within two days. The optimisation problems were not solved even when g was increased to 5. However, if the 9240 equations representing relationships between geographical areas were removed, separate zero-restricted solutions for all geographical areas and levels were found in 17 seconds on a computer Pentium 2533 Mhz. Once the control-rounded tables for the wards were found, these could be added to obtain control-rounded tables for LADs and for the whole of Scotland that are consistent with the control-rounded ward tables.

6 Avoiding biased solutions

The aim of many statistical offices is to provide detailed statistics for small neighbourhoods with target populations of 200–250 people. Information is provided for these small areas not only for their own interest, but also so that the neighbourhoods can be used as building blocks from which to construct an approximation to any larger area, A. In order to produce the frequency tables for the large area, A, it is then simply necessary to add the tables for the individual neighbourhoods contained within A. However, it is important when doing this that the perturbations made when rounding the frequencies for each small area do not accumulate into very large resulting perturbations for the table for area A. In order to reduce the chances of this happening, the distortions made in random rounding are often specified in such a way that they introduce no bias: the stochastic (i.e. random) process is defined so that the expectation of the rounded frequency in any cell is equal to the original frequency, $E(y_i) = a_i$, for all a_i in the original table. *Expectation* is used here in the usual statistical sense to mean the average over a large number of realizations of the stochastic process concerned. Note that this provides no guarantee about what happens in any individual table. It just that in the very long run (i.e. in any aggregate statistic consisting of a sum of several individual rounded tables), we are not likely to introduce any substantial distortions in the table frequencies.

It is possible to define bias in a systematic way for random rounding, since the protection mechanism involves a stochastic process. It is not immediately apparent how to define bias for controlled rounding, since in many implementations it does not involve any random elements. However, just as it is possible to view the deterministic mechanisms used in congruential pseudo-random number generators as stochastic processes, we can apply a similar approximation to controlled rounding. It is a complex process whose properties we can study statistically as if it were a stochastic process. It is in this sense that we specify that ideally the process is approximately unbiased.

We carried out a statistical analysis on the perturbations in the controlled rounding for each original frequency in one of our large datasets from the Scottish Census (we also got similar results for some other datasets). Table 1 shows the obtained proportions of cells rounded up to the next multiple of the rounding base for each original frequency, for the original algorithm (i.e., with the loss of information defined by (2), and with an alternative objective function (13), to be defined next. For each value $a_i \in \{0, \ldots, 20\}$ the second column in Table 1 shows the number of occurrences of this value in the table, and the third column shows the expected proportion of values rounded up for an unbiased solution.

From the fourth column, it is clear that there is a substantial bias in the rounding for all

a_i	occurrences	Expected	using (2)	using (13)
0	131,831	0.0	0.00 ± 0.000	0.00 ± 0.000
1	33,241	0.2	0.04 ± 0.002	0.14 ± 0.004
2	19,535	0.4	0.34 ± 0.007	0.41 ± 0.007
3	13,934	0.6	0.83 ± 0.006	0.66 ± 0.008
4	11,406	0.8	0.99 ± 0.002	0.85 ± 0.007
5	9,476	0.0	0.00 ± 0.000	0.00 ± 0.000
6	8,561	0.2	0.02 ± 0.003	0.13 ± 0.007
7	7,412	0.4	0.26 ± 0.010	0.41 ± 0.011
8	6,820	0.6	0.81 ± 0.010	0.66 ± 0.011
9	6,250	0.8	0.99 ± 0.002	0.85 ± 0.009
10	5,466	0.0	0.00 ± 0.000	0.00 ± 0.000
11	4,906	0.2	0.01 ± 0.003	0.12 ± 0.009
12	4,560	0.4	0.26 ± 0.010	0.39 ± 0.014
13	4,076	0.6	0.82 ± 0.012	0.67 ± 0.015
14	3,870	0.8	0.99 ± 0.003	0.86 ± 0.011
15	3,660	0.0	0.00 ± 0.000	0.00 ± 0.000
16	3,320	0.2	0.02 ± 0.004	0.11 ± 0.011
17	3,180	0.4	0.24 ± 0.015	0.40 ± 0.017
18	2,985	0.6	0.83 ± 0.014	0.66 ± 0.017
19	2,913	0.8	0.99 ± 0.003	0.87 ± 0.013
20	2,681	0.0	0.00 ± 0.000	0.00 ± 0.000

Table 1: Proportion of cells rounded up to the next multiple of the rounding base

frequencies when using objective function (2), but especially for original frequencies immediately adjacent to integer multiples of the rounding base. The figures given after ± are the half widths of 95% confidence intervals on the proportions (allowing for the sampling error as a result of carrying out a finite number of numerical experiments). These are generally very narrow.

There is tendency for the controlled rounding driven by the objective function (2) to round to the closest multiple of the rounding base much more frequently than an unbiased rounder would. In this respect, controlled rounding with objective function (2) behaves in a fashion intermediate between standard unbiased random rounding, and ordinary deterministic rounding (in which frequencies always round to the nearest multiple of the rounding base, e.g. 3, 4, 6 and 7 always round to 5, etc). The bias found in this version of controlled rounding is unlikely to be important in almost all practical applications, since it would be unlikely that an undue preponderance of frequencies would fall in the areas immediately above or below the rounding base.

However, we considered whether changes in the method could be made to reduce this type of bias. Cox [4] discussed procedures for unbiased controlled rounding for 2-way tables. A more accurate method to guarantee this statistical property in the solutions provided by the described

algorithm consists on replacing the objective function (2) by

$$\delta'(v,a) := \delta(v,\xi(a)) = \sum_{i\in I} |v_i - \xi(a_i)| \tag{13}$$

where $\xi(a) = [\xi(a_i) : i \in I]$ is a vector of random variables derived from $a = [a_i : i \in I]$. So the rounding procedure is now aiming to be as close as possible to $\xi(a)$ rather than to a itself, where $\xi(a_i)$ is a random transformation of a_i that takes the form of a map from $\{nr_i + 1, nr_i + 2, nr_i + 3, ..., nr_i + r_i - 1\}$ onto itself for all integer n. We regard the transformation from a to $\xi(a)$ (using objective function (2)) as a stochastic mechanism with probabilities of rounding up given in the 4th column of Table 1, and as we are using the same algorithm to go from $\xi(a)$ to v, they apply also to that transformation. We then obtain a matrix of transition probabilities from a to v that would result in the required probabilities of rounding up for the overall process. In effect what we are now doing is making two transformations, first $a \mapsto \xi(a)$, then $\xi(a) \mapsto v$, where the second transformation is carried out by solving the controlled rounding problem using the same form of objective function as in (2) but with $\xi(a)$ replacing a, as in (13). In order to determine what is the appropriate transformation from $a \mapsto \xi(a)$, it is necessary, of course, to first determine what the bias of the standard procedure is for the particular data set being rounded. It can be done in a preliminary stage using the standard controlled rounding procedure. This approach is only possible when the dataset involved provides many occurrences in which the same original frequencies are rounded, from which the statistical properties of the controlled-rounder in the specific context can be approximately estimated. If the dataset is too small for this, then the only guide available is general experience of controlled rounding of other datasets. It remains to be seen what consistency in statistical properties of the rounding perturbations occurs in different datasets, as further experience is accumulated. The result of this exercise is that the bias is substantially reduced (see Table 1). For almost all practical purposes, the bias properties of the rounding procedure with objective function (13) are adequate.

Figure 6 gives a geometrical interpretation to replace the classical definition of lost of information in (2) by the new definition in (13). The black dots represent tables with rounded values, and the ones inside the polytope in the figure represent additive and protected solutions. Given an unrounded table a, represented by a white dot in the figure, the classical loss of information (2) leads the optimization procedure to generate the biased table v. Instead, by first generating the random table $\xi(a)$ through the transition probabilities, it is possible to generate the unbiased table w in a second run of the procedure. Note that $\xi(a)$ is not necessary additive, as it occurs

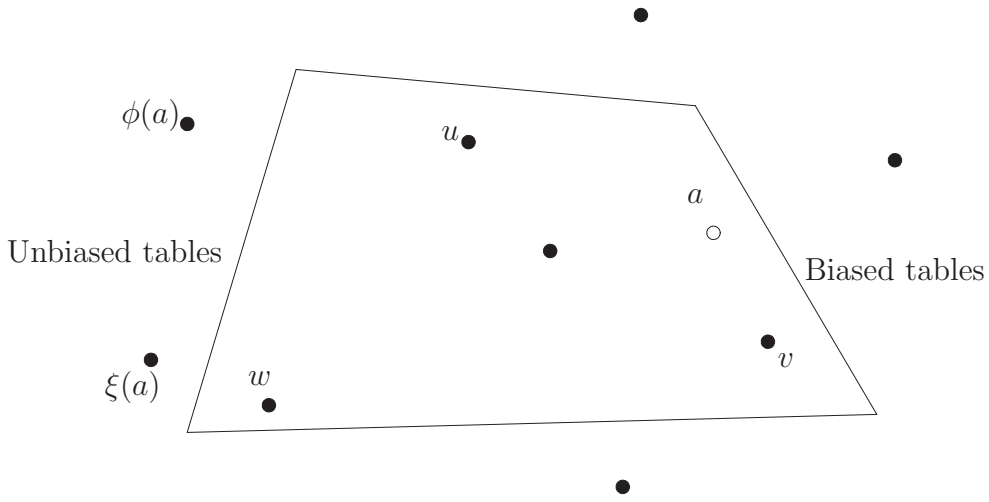

Figure 6: Interpretation of steering the mathematical model with the two objective functions

with the table $\phi(a)$ generated as follows:

$$\phi(a_i) = \begin{cases} \lfloor a_i \rfloor & \text{with probability } (\lceil a_i \rceil - a_i)/r_i \\ \lceil a_i \rceil & \text{with probability } (a_i - \lfloor a_i \rfloor)/r_i \end{cases}$$

Clearly $\phi(a)$ is unbiased and can be very easily generated, but it has the disadvantage that it is not additive neither protected. Using the mathematical model instead, the generated solution v is always additive and protected, but it is biased if (2) was used. Improving the mathematical model with the double transformation procedure, our computational results have shown that the generated table w is also almost unbiased, thus satisfying the statistical office wishes. We also conducted experiments where we first generated $\phi(a)$ and second run the algorithm to generate a solution u closest to $\phi(a)$, but u resulted to have worse statistical features when compared with w.

7 Conclusions

We have addressed the classical Controlled Rounding Methodology integrated with the protection level requirements. This paper is the first work presenting a mathematical model for the underlying optimization problem considering both the minimization of the loss of information and the protection level guarantees. It allows lower, upper and sliding protection levels for protecting a subset of sensitive cells, each one against a set of intruders or a coalition of intruders with different information. There is not assumption on the structure of the tabular data, thus the presented proposal can be applied to k-dimensional tables, hierarchical and linked tables. A first model is in Mixed Integer Programming with a large number of continuous variables, and a second model

with one binary variable for each cell is obtained through Benders' Decomposition. A branch-and-cut algorithm is presented for solving this second model. An alternative methodology, named Cell Perturbation, is proposed to find solution when the classical Controlled Rounding problem is infeasible. The paper has also addressed the critique about the bias implicit in the optimal solutions, and have introduced a two-level procedure to generate unbiased solutions. Computational experiments on real datasets shows the performance of the proposed algorithms.

References

[1] Bacharach, M. (1966) "Matrix Rounding Problem", *Management Science*, **9**, 732–742.

[2] Causey, B.D., Cox, L.H. and Ernst, L.R. (1985) "Applications of Transportation Theory to Statistical Problems", *Journal of the American Statistical Association*, **80**, 903–909.

[3] Cox, L.H. and Ernst, L.R. (1982) "Controlled Rounding", *INFOR*, **20**, 423–432.

[4] Cox, L.H. (1987) "A Constructive Procedure for Unbiased Controlled Rounding", *Journal of the American Statistical Association*, **82**, 520–524.

[5] Cox, L. H. (1995) "Network Models for Complementary Cell Suppression", *Journal of the American Statistical Association*, **90**, 1453–1462.

[6] Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J. M. and Oganian, A. (2002) "Disclosure risk assessment in statistical data protection", in *International Conference on Computational and Applied Mathematics - ICCAM'2002*, Leuven, Belgium.

[7] Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001) "Disclosure Limitation Methods and Information Loss for Tabular Data" in Doyle, P., Lane, J., Theeuwes, J. and Zayatz, L. (editors) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science.

[8] Duncan, G.T. and Fiendberg, S.E. (1998) "Obtaining information while preserving privacy: a Markov perturbation method for tabular data", Proceedings of the *Statistical Data Protection* conference, 351–362.

[9] Dunne, T. (2001) "Issues in the Establishment and Management of Secure Research Sites" in Doyle, P., Lane, J., Theeuwes, J. and Zayatz, L. (editors) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science.

[10] Evans, T., Zayatz, L. and Slanta, J. (1998) "Using Noise for Disclosure Limitation of Establishment Tabular Data", *Journal of Official Statistics*, **14/4**, 537–551.

[11] Fienberg, S. E., Steele, R. J. and Makov, U. E. (1996) "Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Log-Linear Models", Proceedings of *Bureau of the Census 1996 Annual Research Conference*, Washington D.C.

[12] Fischetti, M. and Salazar, J. J. (1998) "Computational Experience with the Controlled Rounding Problem in Statistical Disclosure Control", *Journal of Official Statistics*, **14/4**, 553–565.

[13] Fischetti, M. and Salazar, J. J. (2000) Solving the Cell Suppression Problem on Tabular Data with Linear Constraints. *Management Science*, **47**, 1008–1026.

[14] Fischetti, M. and Salazar, J.J. (2003) "Partial Cell Suppression: a New Methodology for Statistical Disclosure Control", *Statistics and Computing*, **13**, 13–21.

[15] Hundepool, A. (2002) "The CASC project", 172–180, in Domingo-Ferrer, J. (editor) *Inference Control in Statistical Databases: From Theory to Practice*, Lecture Notes in Computer Science 2316, Springer.

[16] Jewett, R. (1993) "Disclosure Analysis for the 1992 Economic Census", Working paper, U.S.B.C.

[17] Kelly, J. P., Golden, B. L. and Assad, A. A. (1990) "Using Simulated Annealing to Solve Controlled Rounding Problems", *ORSA Journal on Computing*, **2**, 174–185.

[18] Kelly, J. P., Golden, B. L., Assad, A. A. and Baker, E. K. (1990) "Controlled Rounding of Tabular Data", *Operations Research*, **38**, 760–772.

[19] Kelly, J. P., Golden, B. L. and Assad, A. A. (1993) "Large-Scale Controlled Rounding Using TABU Search with Strategic Oscillation", *Annals of Operations Research*, **41**, 69–84.

[20] Robertson, D. A. and Ethier, R. (2002) "Cell Suppression: Experience and Theory", 8–20 in Domingo-Ferrer, J. (editor) *Inference Control in Statistical Databases: From Theory to Practice*, Lecture Notes in Computer Science 2316, Springer.

[21] Sande, G. (1984) "Automated Cell Suppression to preserve confidentiality of business statistics", *Statistical Journal of the United Nations ECE*, **2**, 33–41.

[22] Sande, G. (1995) "ACS documentation", *Sande & Associates*, 600 Sanderling Ct. Secaucus NJ, 07094 U.S.A.

[23] Willenborg, L. C. R. J. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics 155, Springer.

[24] Wolsey, L.A. (1998) *Integer Programming*, Wiley-Interscience.