Computational Experiments with Minimum-Distance Controlled Perturbation Methods

Jordi Castro*

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Pau Gargallo 5, 08028 Barcelona, Catalonia, Spain
jcastro@eio.upc.es
http://www-eio.upc.es/~jcastro

Abstract. Minimum-distance controlled perturbation is a recent family of methods for the protection of statistical tabular data. These methods are both efficient and versatile, since can deal with large tables of any structure and dimension, and in practice only need the solution of a linear or quadratic optimization problem. The purpose of this paper is to give insight into the behaviour of such methods through some computational experiments. In particular, the paper (1) illustrates the theoretical results about the low disclosure risk of the method; (2) analyzes the solutions provided by the method on a standard set of seven difficult and complex instances; and (3) shows the behaviour of a new approach obtained by the combination of two existing ones.

Keywords: statistical disclosure control, controlled perturbation methods, linear programming, quadratic programming.

1 Introduction

The safe dissemination of tabular data is one of the main concerns of national statistical agencies. The size and complexity of the data to be protected is continuously increasing, which results in a need for more efficient and versatile protection procedures. This work deals with minimum-distance controlled perturbation, a recent family of methods that meets the above requirements.

Currently, one of the widely used techniques in practice is cell suppression, which is known to be a NP-hard problem [15]. Although exact mixed integer linear programming procedures have been recently suggested [11], the main inconvenience of this approach is that, due to its combinatorial nature, the solution of very large instances (with possibly millions of cells) can result in impractical execution times [13]. Several heuristics have also been suggested to obtain fast approximate solutions [1, 4, 7, 10, 15]. Those approaches are based on the solution

* Supported by the EU IST-2000-25069 CASC project and the Spanish MCyT Project TIC2003-00997.

J. Domingo-Ferrer and V. Torra (Eds.): PSD 2004, LNCS 3050, pp. 73–86, 2004.

of several network optimization subproblems. Unfortunately, although fast, they can only be applied to certain classes of tables, e.g., two and three-dimensional, and two-dimensional with hierarchies in one dimension.

To avoid the above lacks of cell suppression, alternative approaches have been introduced. One of them is the minimum-distance controlled perturbation family of methods. Given a set of tables to be protected, they find the closest ones (according to some distance measure) that, guaranteeing confidentiality, minimize the information loss. Members of that family of methods were independently suggested in [9] (the controlled table adjustment method, which uses a L_1 distance) and [3] (the quadratic minimum-distance controlled perturbation, based on L_2). Specialized interior-point methods [2] were used in [5] for the solution of large-scale instances. A unified framework for those methods was presented in [6], including a proof of their low disclosure risk.

This paper is organized as follows. Section 2 outlines the minimum-distance controlled perturbation framework. Section 3 illustrates the theoretical results about the disclosure risk of the method. Section 4 shows the behaviour of three particular distances on a set of seven complex instances. Finally, Section 5 reports the results obtained with an approach that combines the L_1 and L_2 distances.

2 The Minimum-Distance Controlled Perturbation Framework

This section only outlines the general model, and the particular formulations for the L_1, L_2 and L_∞ distances. More details can be found in [9] and [6].

Any problem instance, either with one table or a number of (linked or hierarchical) tables, can be represented by the following elements:

- A set of cells $a_i, i = 1, \ldots, n$, that satisfy some linear relations $Ma = b$ (a being the vector of a_i's). The method will look for the closest safe values $x_i, i = 1, \ldots, n$, according to some particular distance measure L, that satisfy the above constraints. The distance can be affected by any positive semidefinite diagonal metric matrix $W = \text{diag}(w_1, \ldots, w_n)$.
- A lower and upper bound for each cell $i = 1, \ldots, n$, respectively $\underline{a}_i$ and $\overline{a}_i$, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i, $\underline{a}_i = 0$ ($\underline{a}_i = -\infty$ if $a \geq 0$ is not required) and $\overline{a}_i = +\infty$ can be used.
- A set $\mathcal{P} = \{i_1, i_2, \ldots, i_p\}$ of indices of confidential cells.
- A lower and upper protection level for each confidential cell $i \in \mathcal{P}$, respectively lpl_i and upl_i, such that the released values satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$. To add the above "or" constraint to a mathematical model we need a binary variable y_i and two extra constraints for each confidential cell:

$$\begin{array}{ll} x_i \geq -S(1-y_i) + (a_i + upl_i)y_i & i \in \mathcal{P}, \\ x_i \leq Sy_i + (a_i - lpl_i)(1-y_i) & i \in \mathcal{P}, \\ y_i \in \{0,1\} & i \in \mathcal{P}, \end{array} \tag{1}$$

S in (1) being a large value. That results in a large combinatorial optimization problem which would constrain the effectiveness of the approach to small and medium sized problems. Therefore, in practice, we will assume the sense of the protection for each confidential cell (i.e., the values of the y_i variables) is a priori fixed. This simplifying assumption permits to protect the table by the solution of a single continuous optimization problem. If the particular choice of protection senses (i.e., y_i values) results in an infeasible problem, we can solve an alternative one by relaxing the constraints $Ma = b$ with a large penalization for possible perturbations in the right-hand-side (see Section 5 for details).

The general minimum-distance controlled perturbation method, using some L distance, can be formulated as the following optimization problem:

$$\begin{array}{rl} \min\limits_{x} & ||x-a||_L \\ \text{subject to} & Mx = b \\ & \underline{a}_i \le x_i \le \overline{a}_i \quad i = 1,\ldots,n \\ & x_i \le a_i - lpl_i \text{ or } x_i \ge a_i + upl_i \quad i \in \mathcal{P}. \end{array} \tag{2}$$

The general problem (2) can also be formulated in terms of deviations or perturbations from the current cell values. Defining $z_i = x_i - a_i, \quad i = 1,\ldots,n$, (2) can be transformed to

$$\begin{array}{rl} \min\limits_{z} & ||z||_L \\ \text{subject to} & Mz = 0 \\ & \underline{z}_i \le z_i \le \overline{z}_i \quad i = 1,\ldots,n \\ & z_i \le -lpl_i \text{ or } z_i \ge upl_i \quad i \in \mathcal{P}. \end{array} \tag{3}$$

where $z \in \mathbb{R}^n$ is the vector of deviations, $\underline{z}_i = \underline{a}_i - a_i \le 0$ and $\overline{z}_i = \overline{a}_i - a_i \ge 0$. A benefit of (3) is that it can be solved without releasing the confidential data vector a.

Using the L_1 distance, and after some manipulation, (3) can be written as

$$\begin{array}{rl} \min\limits_{z^+,z^-} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{subject to} & M(z^+ - z^-) = 0 \\ & 0 \le z_i^+ \le \quad \overline{z}_i \quad i = 1,\ldots,n \\ & 0 \le z_i^- \le -\underline{z}_i \quad i = 1,\ldots,n \\ & \left\{ \begin{array}{l} z_i^+ \ge upl_i \\ z_i^- = 0 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} z_i^- \ge lpl_i \\ z_i^+ = 0 \end{array} \right\} \quad i \in \mathcal{P}, \end{array} \tag{4}$$

z^+ and z^- being the vector of positive and negative deviations in absolute value.

For L_2, (3) is

$$\begin{array}{rl} \min\limits_{z} & \sum\limits_{i=1}^{n} w_i z_i^2 \\ \text{subject to} & Mz = 0 \\ & \underline{z}_i \le z_i \le \overline{z}_i \quad i = 1,\ldots,n \\ & z_i \le -lpl_i \text{ or } z_i \ge upl_i \quad i \in \mathcal{P}. \end{array} \tag{5}$$

Finally, for L_∞, the general model (3) can be formulated as

$$\begin{array}{rl} \min\limits_{z^+,z^-,z_{\in\mathcal{P}},z_{\notin\mathcal{P}}} & z_{\in\mathcal{P}} + z_{\notin\mathcal{P}} \\ \text{subject to} & M(z^+ - z^-) = 0 \\ & 0 \le z_i^+ \le \quad \overline{z}_i \quad i = 1,\ldots,n \\ & 0 \le z_i^- \le -\underline{z}_i \quad i = 1,\ldots,n \\ & \left\{ \begin{array}{l} z_i^+ \ge upl_i \\ z_i^- = 0 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} z_i^- \ge lpl_i \\ z_i^+ = 0 \end{array} \right\} \quad i \in \mathcal{P} \\ & z_{\in\mathcal{P}} \ge w_i(z_i^+ + z_i^-) \quad i \in \mathcal{P} \\ & z_{\notin\mathcal{P}} \ge w_i(z_i^+ + z_i^-) \quad i \notin \mathcal{P}, \end{array} \tag{6}$$

$z_{\in\mathcal{P}}$ and $z_{\notin\mathcal{P}}$ being extra variables that store the maximum deviation for, respectively, the sensitive and nonsensitive cells.

An appropriate choice for the weights in (4–6) is $w_i = 1/a_i$, making the deviations relative to the cell value. These weights will be used in the computational results of the paper. (4) is a fixed version of the controlled tabular adjustment suggested in [9]. L_2 provides the smallest optimization problem, although it is quadratic. L_1 and L_∞ provide linear problems, with a larger number of variables and constraints. Effective approaches for the solution of (4–6) were discussed in [6].

3 Illustrating the Disclosure Risk of the Method

The theoretical results about the disclosure risk of the method were presented in [6]. This section summarizes them, and illustrates the low disclosure risk of the method through an example.

To retrieve the original table, the attacker should compute the deviations applied by solving the optimization problem (3). In practice the only term known by the attacker is the M matrix provided by the table structure. However, assume the attacker has partial information, $upl_i, i \in \mathcal{P}$, being the only unknown terms (without loss of generality we consider all the protection senses were "upper"). The problem to be solved to disclose the deviations is then

$$\begin{array}{rl} \min\limits_{z'} & ||z'||_L \\ \text{subject to} & Mz' = 0 \\ & z'_i \ge upl_i + e_i, \quad i \in \mathcal{P}, \end{array} \tag{7}$$

$upl_i + e_i$ being the approximate values used by the attacker to obtain the approximate deviations z'. The protection of the table thus depends on how sensitive the solution z'^* is to possible small e_i values. This relation is explained by the next proposition [6]:

Proposition 1. *If $z'^*(e) \in \mathbb{R}^n$ is the solution of (7) for a particular vector of $e = (e_1,\ldots,e_{|\mathcal{P}|})$ values, and $\mu \in \mathbb{R}^{|\mathcal{P}|}$ is the Lagrange multipliers vector of the bounds of z' in (7) for $e = 0$ (i.e., the multipliers obtained when protecting the table), then*

$$\nabla_e ||z'^*(e)||_L \big|_{e=0} = \mu. \tag{8}$$

a

10$_{(3)}$	15	11	9	45
8	10	**12**$_{(4)}$	15	45
10	12	**11**$_{(2)}$	**13**$_{(5)}$	46
28	37	34	37	136

(a)

z

7	0	-6	-1	0
0	0	4	-4	0
-7	0	2	5	0
0	0	0	0	0

(b)

z'

9	0	-8	-1	0
0	0	5	-5	0
-9	0	3	6	0
0	0	0	0	0

(c)

z''

10	4	-11	-3	0
0	0	6	-6	0
-10	-4	5	9	0
0	0	0	0	0

(d)

Fig. 1. Example of sensitivity of the method to changes in the protection levels. (a) Original data a to be protected. Sensitive cells are in boldface, and upper protection levels are given in brackets. (b) Optimal deviations z computed with the L_1 distance, weights $w_i = 1$, and inactive bounds $\underline{a_i} = 0$ and $\overline{a_i} = \infty$ for all the internal cells. Marginal cells were fixed. The Lagrange multipliers of the bounds $z_i \geq upl_i$ for the sensitive cells are $\mu_{11} = 0$, $\mu_{23} = 2$, $\mu_{33} = 4$ and $\mu_{34} = 4$. The objective function – the sum of deviations in absolute value – is 36. (c) and (d) Deviations z' and z'' computed by the attacker using approximate protection levels with errors $e_{11} = e_{23} = e_{33} = e_{34} = 1$, and $e_{11} = 1$, $e_{23} = 2$, $e_{33} = 3$, $e_{34} = 4$, respectively. The objective functions are respectively 46 and 68 which satisfy (9).

Moreover, for, respectively, the L_1 and L_∞ distances, problem (7) is linear, and, for small enough vectors $e = (e_1, \ldots, e_{|\mathcal{P}|})$, (8) can be recast as

$$||z'^*(e)||_L - ||z^*||_L = \sum_{i \in \mathcal{P}} \mu_i e_i, \tag{9}$$

z^* being the deviations used to protect the table.

To illustrate the above result, consider the example of Figure 1. Table (a) shows the original data to be protected. Sensitive cells appear in boldface, and their upper protection levels upl_i are given in brackets. Using the L_1 distance, weights $w_i = 1$, and bounds $\underline{a_i} = 0$ and $\overline{a_i} = \infty$ for all the internal cells, the optimal deviations computed are shown in Table (b). The objective function value is $||z||_{L_1} = \sum_{i=1}^{n} |z_i| = 36$. The Lagrange multipliers of the constraints $z_i \geq upl_i$ for the sensitive cells are $\mu_{11} = 0$, $\mu_{23} = 2$, $\mu_{33} = 4$ and $\mu_{34} = 4$. Since bounds $\underline{a_i} = 0$ are inactive in the solution, the attacker can use (7) to disclose the deviations of Table (b). If, for instance, the attacker can adjust all the original upl_i protection levels, but for cell a_{11}, (in this case, if $e_{11} \leq 4$, $e_{23} = e_{33} = e_{34} = 0$), from (9) and since $\mu_{11} = 0$, a solution with the same objective function (and possibly with the same deviations) that for Table (b) (i.e., 36) will be obtained. However, if all the protection levels are adjusted with errors, a different solution will be computed. For instance, if problem (7) is solved with $e_{11} = e_{23} = e_{33} = e_{34} = 1$, the deviations z' obtained are those of Table (c). The objective function (i.e., sum of deviations) is 46, which satisfies (9): $46 - 36 = 1\mu_{11} + 1\mu_{23} + 1\mu_{33} + 1\mu_{34}$. If problem (7) is solved with slightly larger values $e_{11} = 1$, $e_{23} = 2$, $e_{33} = 3$, $e_{34} = 4$, the deviations z'' obtained are shown in Table (c). Again, the objective function, 68, satisfies: $68 - 36 = 1\mu_{11} + 2\mu_{23} + 3\mu_{33} + 4\mu_{34}$.

Note that $||(\mu_{i:i \in \mathcal{P}})||$ (the norm of the Lagrange multipliers of constraints $z_i \geq upl_i$ for the sensitive cells) can be used as an indicator of the protection of

z'_{L_1}

7	0	-6	-1	0
0	0	4	-4	0
-7	0	2	5	0
0	0	0	0	0

(a)

z''_{L_1}

6	0	-6	0	0
1	0	4	-5	0
-7	0	2	5	0
0	0	0	0	0

(b)

z_{L_2}

$3.41\hat{6}$	$3.41\hat{6}$	-6	$-8.\hat{3}$	0
$0.08\hat{3}$	$0.08\hat{3}$	4.0	$-4.1\hat{6}$	0
-3.5	-3.5	2	5	0
0	0	0	0	0

(c)

Fig. 2. Example of alternative solutions with complete information by the attacker. The original data a to be protected are those of Table (a) of Figure 1. Sensitive cells are in boldface, and upper protection levels are given in brackets. (a) and (b) Alternative solutions z'_{L_1} and z''_{L_1}, computed with two different linear programming solvers, using the L_1 distance, weights $w_i = 1$, and bounds $\underline{a_i} = 0$ and $\overline{a_i} = \infty$ for all the internal cells. Marginal cells were fixed. The objective function – the sum of deviations in absolute value – of both solutions is 36. (c) Unique solution z_{L_2} for the L_2 distance, again with weights $w_i = 1$, and bounds $\underline{a_i} = 0$ and $\overline{a_i} = \infty$ for all the internal cells. The 2-norm of the deviations vector is 12.12.

the table. In theory, the larger this value, the more difficult is for an attacker to retrieve the original data. Real tables, with a large number of sensitive cells, often will have a high $||(\mu_{i:i\in\mathcal{P}})||$ value, and thus confidential.

In some cases, even if the attacker has complete information, the right perturbations can not be disclosed [6]:

Proposition 2. *Assume the attacker knows all the terms of problem (7). If the L_2 distance is used, the solution of that problem will provide the deviations used to protect the table. However, for L_1 or L_∞, the attacker can obtain alternative deviations.*

For instance, Tables (a) and (b) of Figure 2 show two alternative solutions with the L_1 distance for the data of Table (a) of Figure 1. They were obtained with two different implementations of the simplex algorithm, using weights $w_i = 1$, and bounds $\underline{a_i} = 0$ and $\overline{a_i} = \infty$ for all the internal cells. Marginal cells were fixed. The sum of deviations is 36 in both solutions. Table (c) of Figure 2 shows, for the same data, the unique solution for the L_2 distance. Since L_2 involves a quadratic function, the solution attempts to distribute the deviations among all the cells, obtaining a non-integer solution (valid for magnitude tables). Proposition 2 means that L_1 and L_∞ are a bit safer when the attacker knows all the terms of (7), which in practice is equivalent to that the attacker knows the original data (thus, very unlikely). Therefore, in practice, it can be concluded that the three distances have the same low disclosure risk.

4 Computational Comparison

For the computational comparison of models (4–6) (i.e., L_1, L_2 and $L\infty$) we used the seven most complex instances of CSPLIB. CSPLIB is the unique currently available set of instances for tabular data protection [11]. It can be freely

Table 1. Properties of the seven complex instances.

Name	Dimensions	Size	n	$\|\mathcal{P}\|$	m	N.coef
bts4	4D, hierarchical	54,54,4,4	36570	2260	36310	136912
hier13	3D, hierarchical	13,13,13	2020	112	3313	11929
hier16	3D, hierarchical	16,16,16	3564	224	5484	19996
nine12	9D, linked	10,6,6,6,6,6,6,6,6	10399	1178	11362	52624
nine5d	9D, linked	4,29,3,4,5,6,5,4,5	10733	1661	17295	58135
ninenew	9D, linked	10,6,6,6,6,6,6,6,6	6546	858	7340	32920
two5in6	6D, linked	6,4,16,4,4,4	5681	720	9629	34310

obtained from `http://webpages.ull.es/users/casc/#CSPlib:`. These seven instances were also the choice in [8] and are challenging for other approaches, as cell suppression. As shown below, they can be solved in few seconds with the minimum-distance approach. Table 1 provides their main features: identifier (column "Name"), number of dimensions and structure – linked or hierarchical – (column "Dimensions"), size for each dimension (column "Size"), number of total cells and sensitive cells (columns "n" and "$|\mathcal{P}|$", respectively), number of constraints (column "m"), and number of coefficient of the M matrix (column "N.coef"). The structure and size information was obtained from [8].

Problems (4–6) were implemented using the AMPL modelling language [12] and CPLEX 8.0 [14]. All runs were carried on a notebook with a 1.8 GHz processor and 512 Mb of RAM. For L_2 we used the primal-dual interior-point algorithm [16], which can be considered the most efficient choice. L_1 and L_∞ were solved with the two best linear programming algorithms: the simplex method and the primal-dual interior-point method. Although the optimal objective function is the same, both algorithms can provide different solutions. In this work we used those of the simplex method, which, in practice, provided better deviations.

For each of the three distances, Table 3 of Appendix A show the following information. Row "CPU" gives the CPU time in seconds for each algorithm. Rows "Abs. dev." provide the mean (columns "mean"), standard deviation (columns "std") and maximum (columns "max.") of the absolute deviations (i.e., $|z_i|$), for all the cells (row "all"), for the sensitive cells (row "$\in \mathcal{P}$"), and for the nonsensitive cells (row "$\notin \mathcal{P}$"). A similar information is provided for the percentage absolute deviations (i.e., $100|z_i|/a_i$) in rows "Perc. dev.". Finally, rows "2-norm" report the 2-norm of the deviations (i.e., $||z||_2$), again for sensitive, nonsensitive, and all the cells.

Looking at Table 3 we see that most of the optimization problems were solved until optimality in few seconds on a standard personal computer. L_∞ provides the slowest executions, due to the large number of constraints considered in (6). L_2, solved through a quadratic interior-point solver, was always the most efficient choice (except for the smallest instance hier13). In most instances the solution time of the L_2 was about half the time of the second fastest option. This is because, first, the complexity of solving a quadratic separable optimization problem is the same that for a linear one, if we use an interior-point algorithm; and second, problem (4) involves the double of variables that (5). The solu-

tion times obtained with the interior-point algorithm, for the three objectives, can even be improved using specialized solvers that exploit the tables structure [2, 5].

For the absolute deviations, L_2 provides the lowest means and, mainly, the lowest standard deviations. Such lowest standard deviations are not surprising, since L_2, due to its quadratic nature, attempts to evenly distribute the required deviations among all the cells. As for the other two distances, L_∞ provided better absolute deviations than L_1, but for instances hier13 and hier16. That was, a priori, an unexpected result, since only two cells appear in the objective function of (6), whereas all the perturbations are considered in (4). The distribution of the absolute deviations (not reported in the tables) showed that L_1 provided the greater number of cells with small deviations.

For the percentage deviations, L_1 must clearly provide the best mean values, since its objective function is exactly the sum of percentage absolute deviations. However, L_2 provides similar mean percentage deviations, and, for most instances, with slightly better standard deviations. L_∞ provided worser means and standard deviations, but, as a consequence of its objective function, the lowest maximum values.

Finally, the lowest 2-norms of the deviations vector are provided in all the instances by L_2 . This is a consequence of L_2 being the only quadratic objective of the three tested. Except for instance hier13, L_∞ always provides deviations with better 2-norms than L_1.

From the above comments, we can conclude that L_1 provides the best results when a first-order comparison measure, as the mean percentage deviation, is considered. However, when a second-order measure is used, as the 2-norm of the deviations or the standard deviation of the percentage deviations, L_2 seems to be the best choice. The above is an immediate result of the objective functions (linear or quadratic) of the respective optimization problems. That suggests that a method combining L_1 and L_2 could provide fairly good values for the first and second-order comparison measures. This alternative is exploited in next section.

5 Combining the L_1 and L_2 Distances

The optimization problem that results from the combination of the L_1 and L_2 distances can be written in a general form as

$$\begin{aligned}
\min_{z^+,z^-} \quad & \omega_1 \sum_{i=1}^{n} w_{1,i}(z_i^+ + z_i^-) + \omega_2 \sum_{i\in\mathcal{S}} w_{2,i}(z_i^+ + z_i^-)^2 \\
\text{subject to} \quad & M(z^+ - z^-) = 0 \\
& 0 \le z_i^+ \le \quad \overline{z}_i \quad i = 1,\ldots,n \\
& 0 \le z_i^- \le -\underline{z}_i \quad i = 1,\ldots,n \\
& \left\{ \begin{array}{l} z_i^+ \ge upl_i \\ z_i^- = 0 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} z_i^- \ge lpl_i \\ z_i^+ = 0 \end{array} \right\} \quad i \in \mathcal{P},
\end{aligned} \tag{10}$$

z^+ and z^- being the vector of positive and negative deviations in absolute value, ω_1 and ω_2 weights for the overall contribution to the objective function of re-

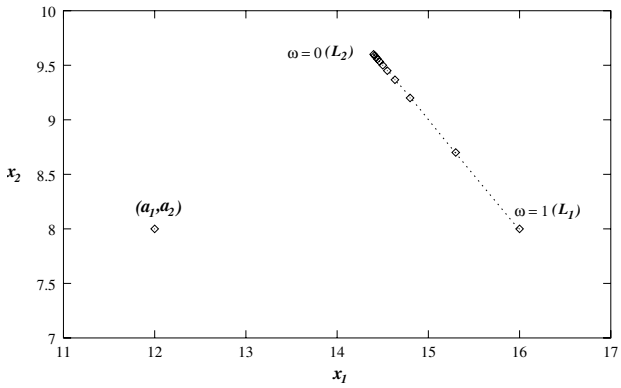

Fig. 3. Solutions of the L_1 and L_2 distances for the one dimensional table $a_1 + a_2 = a_3$, imposing a perturbation $z_3 \geq 4$ for the marginal cell. Point $(a_1, a_2) = (12, 8)$ corresponds to the original internal cell values. The other eleven points are the solutions obtained with the objective function of (4) using $\omega_1 = \omega$ and $\omega_2 = 1 - \omega$, for $\omega = 0, 0.1, 0.2, \ldots, 0.9, 1$, which combines the L_1 and L_2 distances through the weight factor ω. The L_2 solution (computed with $\omega = 0$) is closer to (a_1, a_2), but the L_1 point ($\omega = 1$) preserves the value of a_2.

spectively the linear and quadratic terms, $\mathcal{S}$ a subset of cells affected by L_2, and $w_{1,i}$ and $w_{2,i}$ cell weights for respectively L_1 and L_2. This formulation is general enough to accommodate to several situations. For instance, it provides an always feasible problem if we apply the L_1 and L_2 terms to respectively the internal and marginal cells (i.e., $\mathcal{S}$ is the set of marginal cells), with a large penalization for changes in marginal values (i.e., $w_{2,i} \gg 0$).

Before presenting results for the seven instances of Section 4, we first illustrate the behaviour of (10) on the small example of Figure 3. The table considered is $a_1 + a_2 = a_3$, with $a_1 = 12$ and $a_2 = 8$. We imposed $z_1 + z_2 = z_3$ and $z_3 \geq 4$, i.e., an upper protection level of 4 is forced for the marginal sensitive cell. We set $\mathcal{S} = \{1, 2, 3\}$ (i.e., the three cells appear in the quadratic term of the objective function), and $\omega_1 = \omega$, $\omega_2 = 1 - \omega$, $\omega \in [0, 1]$ being a predefined parameter. For $\omega = 1$ and $\omega = 0$ the combined objective of (10) corresponds to the L_1 and L_2 distances, respectively. Using $w_{1,i} = 1/a_i$ the optimal solution obtained with L_1 is $z_1 = 4$, $z_2 = 0$ and $z_3 = 4$. With the same weights $w_{2,i} = 1/a_i$, the optimal solution provided by L_2 is $z_1 = 2.4$, $z_2 = 1.6$ and $z_3 = 4$. If integer values were required, the z_1 and z_2 values could be rounded through some heuristic postprocess (in that case the most reasonable choice would be $z_1 = 2$ and $z_2 = 2$). Figure 3 shows the perturbed internal cell values obtained for $\omega = 0, 0.1, \ldots, 0.9, 1$, and the original ones (a_1, a_2). Clearly, the L_2 point is closer to $(12, 8)$, but the L_1 solution preserves the value of cell a_2. This is consistent with the results of Section 4. The combined L_{1-2} objective provides solutions on a curve joining the L_1 and L_2 points. Because of the larger costs of the quadratic term, the optimal solution was only far enough from the L_2 point for $\omega = 0.8$ and $\omega = 0.9$.

Table 2. Results for the seven complex instances, for L_1, L_2 and L_{1-2}.

	L_1			L_2			L_{1-2}		
name	CPU	$\overline{\%\text{Dev.}}$	2-norm	CPU	$\overline{\%\text{Dev.}}$	2-norm	CPU	$\overline{\%\text{Dev.}}$	2-norm
bts4	16.5	0.74	18243	11.5	0.83	7912	45.0	0.76	10217
hier13	3.3	0.81	2609	3.8	0.87	2149	7.1	0.82	2306
hier16	19.9	0.83	3203	17.1	0.90	2706	31.0	0.84	2845
nine12	382.1	1.35	5840	18.3	1.53	4878	43.7	1.38	5234
nine5d	126.7	1.67	8316	20.4	1.90	5468	30.9	1.72	5845
ninenew	27.0	1.55	5448	11.1	1.76	4444	26.3	1.56	4731
two5in6	13.6	1.46	4917	9	1.65	3749	16.8	1.50	4045

In the computational results of this section we used $\mathcal{S}=\{1,\ldots,n\}$ (i.e., all the cells are involved in the quadratic term) and $w_{1,i} = w_{2,i} = 1/a_i, i = 1,\ldots,n$. According to the previous small example, we also set $\omega_1 = 0.99$ and $\omega_2 = 0.01$. Table 2 shows the results obtained with L_1, L_2 and the combined L_{1-2} objective. For each distance, the execution time (columns "CPU"), average percentage deviation for all the cells (columns "$\overline{\%\text{Dev.}}$"), and 2-norm of the deviations vector (columns "2-norm") are provided. Executions were carried on the same hardware and with the same software (i.e., AMPL+CPLEX 8.0) than in Section 4. The results reported for L_1 were obtained with the simplex method, while the quadratic interior-point algorithm was used for L_2 and L_{1-2}. Looking at Table 2 we see the combined L_{1-2} distance provides average percentage deviations close to those of L_1, while the 2-norm has been significantly reduced. As expected, the combined L_{1-2} distance inherited the good properties of L_1 and L_2.

6 Conclusions

As shown by the computational experiments of this work, the minimum-distance approach is efficient, versatile and safe. The three methods tested, for L_1, L_2 and L_∞, provided different patterns of deviations, each of them with a clear behaviour. As done in the paper with L_1 and L_2, it is possible to combine them in a new approach with the good features of the original methods.

One of the fields of research to be be explored deals with the optimization solvers. In a static environment, the final goal might be the protection, in a single run, of all the tables derived from the same microdata. The resulting problem is huge. In a dynamic environment, the goal would be the online protection of particular tables (e.g., obtained from end-user queries from a data-warehouse). Speed is instrumental in that case. In both situations, we may need highly-efficient implementations of the optimization methods used in this work, which exploit the problem structure. Some steps have already been done in this direction for large (i.e., one million cells) three-dimensional tables and L_2 [5], where a specialized implementation was two orders of magnitude faster than the CPLEX 8.0 solver. Extending those achievements to general tables is part of the future work to be done.

References

1. Carvalho, F.D., Dellaert, N.P., Osório, M.D.: Statistical disclosure in two-dimensional tables: general tables. J. Am. Stat. Assoc. **89** (1994) 1547–1557
2. Castro, J.: A specialized interior-point algorithm for multicommodity network flows. SIAM J. on Opt. **10** (2000) 852–877
3. Castro, J.: Internal communication to partners of the European Union IST-2000-25069 CASC project (2002).
4. Castro, J.: Network flows heuristics for complementary cell suppression: an empirical evaluation and extensions. Lect. Notes in Comp. Sci. **2316** (2002) 59–73. Volume Inference Control in Statistical Databases, ed. J. Domingo-Ferrer. Springer.
5. Castro, J.: Quadratic interior-point methods in statistical disclosure control, Computational Management Science, in press (2004). Previously appeared as Research Report DR 2003/10, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2003). Available from the author webpage.
6. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research, accepted subject to revision (2003). An extended version previously appeared as Research Report DR 2003/14, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2003). Available from the author webpage.
7. Cox, L.H.: Network models for complementary cell suppression. J. Am. Stat. Assoc. **90** (1995) 1453–1462
8. Dandekar, R.A.: Cost effective implementation of synthetic tabulation (a.k.a. controlled tabular adjustments) in legacy and new statistical data publication systems. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, (2003). Available from `http://www.unece.org/stats/documents/2003.04.confidentiality.htm`.
9. Dandekar, R.A., Cox, L.H.: Synthetic tabular data: an alternative to complementary cell suppression. Manuscript, Energy Information Administration, U.S. Department of Energy (2002). Available from the first author on request (`Ramesh.Dandekar@eia.doe.gov`).
10. Dellaert, N.P., Luijten, W.A.: Statistical disclosure in general three-dimensional tables. Statistica Neerlandica **53** (1999) 197–221
11. Fischetti, M., Salazar, J.J.: Models and algorithms for optimizing cell suppression in tabular data with linear constraints. J. Am. Stat. Assoc. **95** (2000) 916–928
12. Fourer, R, Gay, D.M., Kernighan, B.W.: AMPL: A Modeling Language for Mathematical Programming. Duxbury Press (1993)
13. Giessing, S.: New tools for cell-suppression in τ-Argus: one piece of the CASC project work draft. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje (2001).
14. ILOG CPLEX: ILOG CPLEX 8.0 Reference Manual Library. ILOG (2002)
15. Kelly, J.P., Golden, B.L, Assad, A.A.: Cell Suppression: disclosure protection for sensitive tabular data. Networks **22** (1992) 28–55
16. Wright, S.J.: Primal-Dual Interior-Point Methods. SIAM (1997).

Appendix

A Tables with Results for the Seven Instances

Table 3. Results for the seven complex instances, for L_1, L_2 and L_∞.

Results for the bts4 instance

	L_1				L_2				L_∞			
	Simplex		Int. Point		Int. Point				Simplex		Int. Point	
CPU	16.46		39.7		11.45				1594.69		207.02	
		mean	std	max.		mean	std	max.		mean	std	max.
Abs.	all	33.9	89.2	4483.0	all	24.9	33.0	795.9	all	30.1	49.0	947.8
dev.	$\in\mathcal{P}$	56.0	32.2	155.0	$\in\mathcal{P}$	56.0	32.2	155.0	$\in\mathcal{P}$	57.0	32.6	168.5
	$\notin\mathcal{P}$	32.4	91.5	4483.0	$\notin\mathcal{P}$	22.9	32.0	795.9	$\notin\mathcal{P}$	28.4	49.4	947.8
		mean	std	max.		mean	std	max.		mean	std	max.
Perc.	all	0.74	1.97	11.11	all	0.84	1.95	20.23	all	1.10	2.36	11.11
dev.	$\in\mathcal{P}$	7.27	2.60	11.11	$\in\mathcal{P}$	7.27	2.59	11.11	$\in\mathcal{P}$	7.46	2.61	11.11
	$\notin\mathcal{P}$	0.31	0.83	11.03	$\notin\mathcal{P}$	0.42	0.84	20.23	$\notin\mathcal{P}$	0.68	1.63	11.03
	all		18243.0		all		7912.0		all		10997.2	
2-norm	$\in\mathcal{P}$		3072.3		$\in\mathcal{P}$		3070.3		$\in\mathcal{P}$		3120.5	
	$\notin\mathcal{P}$		17982.4		$\notin\mathcal{P}$		7292.0		$\notin\mathcal{P}$		10545.2	

Results for the hier13 instance

	L_1				L_2				L_∞			
	Simplex		Int. Point		Int. Point				Simplex		Int. Point	
CPU	3.25		6.86		3.83				5.85		35.23	
		mean	std	max.		mean	std	max.		mean	std	max.
Abs.	all	37.8	44.1	344.0	all	33.9	33.7	313.4	all	52.0	58.2	463.7
dev.	$\in\mathcal{P}$	55.6	28.0	97.0	$\in\mathcal{P}$	55.2	27.8	97.0	$\in\mathcal{P}$	59.0	27.8	97.0
	$\notin\mathcal{P}$	36.8	44.6	344.0	$\notin\mathcal{P}$	32.7	33.6	313.4	$\notin\mathcal{P}$	51.5	59.4	463.7
		mean	std	max.		mean	std	max.		mean	std	max.
Perc.	all	0.81	1.72	9.97	all	0.87	1.95	45.84	all	1.04	1.91	9.97
dev.	$\in\mathcal{P}$	6.20	2.17	9.97	$\in\mathcal{P}$	6.18	2.19	9.97	$\in\mathcal{P}$	6.65	2.37	9.97
	$\notin\mathcal{P}$	0.49	1.02	8.28	$\notin\mathcal{P}$	0.56	1.42	45.84	$\notin\mathcal{P}$	0.71	1.25	8.28
	all		2609.6		all		2149.3		all		3504.9	
2-norm	$\in\mathcal{P}$		658.5		$\in\mathcal{P}$		654.1		$\in\mathcal{P}$		689.3	
	$\notin\mathcal{P}$		2525.1		$\notin\mathcal{P}$		2047.4		$\notin\mathcal{P}$		3436.4	

Table 3. (Continued).

Results for the hier16 instance

	L_1				L_2				L_∞			
	Simplex		Int. Point		Int. Point				Simplex		Int. Point	
CPU	19.85		28.36		17.19				66.52		136.86	
		mean	std	max.		mean	std	max.		mean	std	max.
Abs.	all	35.8	40.0	280.5	all	33.4	30.6	258.3	all	36.8	36.6	300.9
dev.	$\in\mathcal{P}$	48.3	27.4	131.0	$\in\mathcal{P}$	48.3	27.4	131.0	$\in\mathcal{P}$	48.7	27.4	131.0
	$\notin\mathcal{P}$	34.9	40.6	280.5	$\notin\mathcal{P}$	32.4	30.6	258.3	$\notin\mathcal{P}$	36.0	37.0	300.9
		mean	std	max.		mean	std	max.		mean	std	max.
Perc.	all	0.83	1.84	10.00	all	0.90	1.81	10.00	all	1.13	2.05	10.00
dev.	$\in\mathcal{P}$	6.89	2.38	10.00	$\in\mathcal{P}$	6.89	2.38	10.00	$\in\mathcal{P}$	7.04	2.41	10.00
	$\notin\mathcal{P}$	0.43	0.78	7.59	$\notin\mathcal{P}$	0.50	0.75	7.59	$\notin\mathcal{P}$	0.73	1.26	7.59
	all		3203.5		all		2706.3		all		3098.4	
2-norm	$\in\mathcal{P}$		830.2		$\in\mathcal{P}$		830.2		$\in\mathcal{P}$		836.8	
	$\notin\mathcal{P}$		3094.1		$\notin\mathcal{P}$		2575.9		$\notin\mathcal{P}$		2983.2	

Results for the nine12 instance

	Simplex		Int. Point		Int. Point				Simplex		Int. Point	
CPU	382.13		47.38		18.29				727.28		338.8	
		mean	std	max.		mean	std	max.		mean	std	max.
Abs.	all	36.3	44.3	490.9	all	34.6	33.0	377.4	all	32.6	36.5	268.0
dev.	$\in\mathcal{P}$	51.7	28.3	154.0	$\in\mathcal{P}$	51.6	28.2	154.0	$\in\mathcal{P}$	52.1	28.2	154.0
	$\notin\mathcal{P}$	34.4	45.6	490.9	$\notin\mathcal{P}$	32.4	33.0	377.4	$\notin\mathcal{P}$	30.1	36.7	268.0
		mean	std	max.		mean	std	max.		mean	std	max.
Perc.	all	1.35	2.34	12.55	all	1.53	2.32	25.43	all	1.74	2.64	10.00
dev.	$\in\mathcal{P}$	6.71	2.38	10.00	$\in\mathcal{P}$	6.70	2.39	11.97	$\in\mathcal{P}$	6.82	2.40	10.00
	$\notin\mathcal{P}$	0.67	1.15	12.55	$\notin\mathcal{P}$	0.87	1.23	25.43	$\notin\mathcal{P}$	1.09	1.85	8.95
	all		5840.1		all		4878.1		all		4988.1	
2-norm	$\in\mathcal{P}$		2022.2		$\in\mathcal{P}$		2017.2		$\in\mathcal{P}$		2034.2	
	$\notin\mathcal{P}$		5478.8		$\notin\mathcal{P}$		4441.5		$\notin\mathcal{P}$		4554.5	

Results for the nine5d instance

	Simplex		Int. Point		Int. Point				Simplex		Int. Point	
CPU	126.67		43.03		20.36				784.52		137.33	
		mean	std	max.		mean	std	max.		mean	std	max.
Abs.	all	41.4	68.8	1010.0	all	37.2	37.5	499.4	all	34.4	38.4	306.5
dev.	$\in\mathcal{P}$	50.6	29.3	156.0	$\in\mathcal{P}$	50.6	29.3	156.0	$\in\mathcal{P}$	50.8	29.3	156.0
	$\notin\mathcal{P}$	39.7	73.6	1010.0	$\notin\mathcal{P}$	34.7	38.3	499.4	$\notin\mathcal{P}$	31.4	39.1	306.5
		mean	std	max.		mean	std	max.		mean	std	max.
Perc.	all	1.67	2.69	10.00	all	1.90	2.53	10.00	all	2.23	3.02	10.00
dev.	$\in\mathcal{P}$	6.83	2.42	10.00	$\in\mathcal{P}$	6.83	2.42	10.00	$\in\mathcal{P}$	6.87	2.42	10.00
	$\notin\mathcal{P}$	0.73	1.31	9.78	$\notin\mathcal{P}$	1.00	1.11	9.31	$\notin\mathcal{P}$	1.38	2.25	8.79
	all		8316.4		all		5468.3		all		5343.4	
2-norm	$\in\mathcal{P}$		2383.6		$\in\mathcal{P}$		2383.2		$\in\mathcal{P}$		2389.6	
	$\notin\mathcal{P}$		7967.5		$\notin\mathcal{P}$		4921.7		$\notin\mathcal{P}$		4779.4	

Table 3. (Continued).

Results for the ninenew instance

	Simplex		Int. Point		Int. Point			Simplex		Int. Point	
CPU	27.08		24.02		11.15			199.39		120.52	
		mean	std	max.		mean	std	max.	mean	std	max.
Abs.	all	41.6	53.0	602.7	all	38.6	39.0	522.8	all 39.0	43.2	439.1
dev.	$\in\mathcal{P}$	52.4	28.6	192.0	$\in\mathcal{P}$	52.4	28.3	192.0	$\in\mathcal{P}$ 53.0	28.3	192.0
	$\notin\mathcal{P}$	39.9	55.5	602.7	$\notin\mathcal{P}$	36.6	40.0	522.8	$\notin\mathcal{P}$ 36.8	44.7	439.1
		mean	std	max.		mean	std	max.	mean	std	max.
Perc.	all	1.56	2.47	16.16	all	1.76	2.44	22.86	all 2.19	2.93	10.00
dev.	$\in\mathcal{P}$	6.66	2.38	10.00	$\in\mathcal{P}$	6.66	2.36	10.00	$\in\mathcal{P}$ 6.79	2.39	10.00
	$\notin\mathcal{P}$	0.79	1.29	16.16	$\notin\mathcal{P}$	1.02	1.35	22.86	$\notin\mathcal{P}$ 1.50	2.32	9.93
	all		5447.5		all		4444.3		all	4708.1	
2-norm	$\in\mathcal{P}$		1749.6		$\in\mathcal{P}$		1744.5		$\in\mathcal{P}$	1759.5	
	$\notin\mathcal{P}$		5158.9		$\notin\mathcal{P}$		4087.6		$\notin\mathcal{P}$	4366.9	

Results for the two5in6 instance

	Simplex		Int. Point		Int. Point			Simplex		Int. Point	
CPU	13.58		16.88		9			83.48		86.47	
		mean	std	max.		mean	std	max.	mean	std	max.
Abs.	all	38.3	52.8	530.0	all	35.4	34.9	340.1	all 38.3	39.3	281.8
dev.	$\in\mathcal{P}$	49.1	32.0	169.0	$\in\mathcal{P}$	49.1	32.0	169.0	$\in\mathcal{P}$ 49.7	31.8	169.0
	$\notin\mathcal{P}$	36.7	55.0	530.0	$\notin\mathcal{P}$	33.5	34.9	340.1	$\notin\mathcal{P}$ 36.7	40.0	281.8
		mean	std	max.		mean	std	max.	mean	std	max.
Perc.	all	1.46	2.49	10.00	all	1.65	2.40	17.88	all 2.08	2.81	10.00
dev.	$\in\mathcal{P}$	6.80	2.42	10.00	$\in\mathcal{P}$	6.80	2.42	10.00	$\in\mathcal{P}$ 6.99	2.42	10.00
	$\notin\mathcal{P}$	0.69	1.23	9.69	$\notin\mathcal{P}$	0.90	1.17	17.88	$\notin\mathcal{P}$ 1.37	2.04	8.44
	all		4917.2		all		3749.3		all	4137.1	
2-norm	$\in\mathcal{P}$		1573.0		$\in\mathcal{P}$		1572.0		$\in\mathcal{P}$	1582.4	
	$\notin\mathcal{P}$		4658.8		$\notin\mathcal{P}$		3403.8		$\notin\mathcal{P}$	3822.5	