

Quality issues of minimum distance controlled tabular adjustment

Jordi Castro^{*1}, Sarah Giessing^{**}

* Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona, Catalonia, Spain
(jordi.castro@upc.edu)

** Federal Statistical Office of Germany,
65180 Wiesbaden, Germany
(sarah.giessing@destatis.de)

Abstract. Controlled tabular adjustment (CTA), and its minimum distance variants, is a recent methodology for the protection of tabular data. Given a table to be protected, the purpose of the method is to find the closest one that guarantees the confidentiality of the sensitive cells. This is achieved by adding slight adjustments to the remaining cells, preferably excluding total ones, whose values are preserved. Unlike other approaches, this methodology can efficiently protect large tables of any number of dimensions and structure. In this work, we test some minimum distance variants of CTA on a close-to-real data set, and analyze the quality of the solutions provided, defining some quality indicators. We empirically show that, for the instances tested, the quality of solutions provided by CTA is better than that of alternative procedures, as cell suppression.

Key words: Statistical Disclosure Control, Tabular Data Protection, Synthetic Tabular Data, Linear Programming, Quadratic Programming.

1 Introduction

Data collected within government statistical systems must be provided as to fulfill requirements of many users differing widely in the particular interest they take in the data. For data in tabular form, this implies that most tables made publicly available belong to a system of multiple, hierarchically structured, overlapping tables which are all publicly available. Usually, some cells of these tables contain information on single, or very few respondents. Especially in the case of establishment data, given the meta information provided along with the cell values (typically: industry, geography, size classes), those respondents could be easily identifiable. Therefore, measures for protection of those data have to be put in place. Traditionally, agencies suppress part of the information (cell suppression). Efficient algorithms for cell suppression are offered f.i. by the software package τ -ARGUS (Hundepool et al., 2004). Cell suppressions, however, must be coordinated between tables. This implies certain restrictions on the release of tabular data which is in some contrast to the flexibility and

¹Work supported by the Spanish MCyT project TIC2003-00997.

capacity of modern (OnLine) Data Base systems. Cell perturbation, as alternative to, or in combination with cell suppression may offer a way out of the dilemma.

Minimum distance controlled tabular adjustment (or CTA for short) (Dandekar and Cox, 2002; Castro, 2006) is a recent technique to generate synthetic, i.e. perturbed values that may be used to replace original entries of tables provided for a publication. Although CTA is very efficient from a computational point of view, NSAs are still reluctant to use it, because offering synthetic data might be in conflict to their responsibility to produce data that are 'as accurate as possible'. NSAs would also have to decide in which way to publish CTA protected tables. Should adjustments be flagged in a publication, and in which way? We believe that in order to introduce CTA into practice, it is essential to prove that data sets protected by CTA can provide a *sufficient* amount of *accurate* information, compared to the standards set by cell suppression. Instead of considering how to preserve second order statistics, like variance and covariance, proposed in Cox et al. (2004), in this paper we focus on the following simple criteria for a robust CTA that allow comparison to, or combination with cell suppression to some extent:

- We suggest criteria that could be used to decide, if a deviation is so small that the adjusted cell value is still sufficiently reliable for any sensible kind of analysis a user might be interested in —an adjustment fulfilling the requirements would not have to be flagged. The number of cells where the adjustment fails to meet the requirements should be as low as possible. Such large deviations are in some sense equivalent to the suppression of the cell. This is particularly important for the subset of cells that provide aggregated information on a high level (for geography, for instance, state, or whole country level).
- CTA should be able to provide a feasible solution if deviations are only allowed in a reduced subset of cells. For instance, this enables to filter through CTA data previously protected by other techniques like cell suppression: in this case the suppressed cells would be the subset of cells allowed for deviations, as suggested in Giessing (2004).

The structure of the paper is as follows. Section 2 sketches the minimum distance CTA family of methods. Section 3 analyzes the quality of results obtained with some close-to-real instances, and compares them with those obtained with cell suppression. In Section 4 we discuss a restricted CTA procedure for improving the quality of the protected tables.

2 Outline of minimum distance controlled tabular adjustment

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

- A set of cells $a_i, i = 1, \dots, n$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i 's).
- A lower and upper bound for each cell $i = 1, \dots, n$, respectively \underline{a}_i and \bar{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 $\bar{a}_i = +\infty$ can be used.

- A set $\mathcal{P} = \{i_1, i_2, \dots, i_p\} \subseteq \{1, \dots, n\}$ 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 $x_i \leq a_i - lpl_i$.

CTA attempts to find the closest safe values $x_i, i = 1, \dots, n$, according to some distance L , that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{aligned} \min_x \quad & \|x - a\|_L \\ \text{subject to} \quad & Ax = b \\ & \underline{a}_i \leq x_i \leq \bar{a}_i \quad i = 1, \dots, n \\ & x_i \leq a_i - lpl_i \text{ or } x_i \geq a_i + upl_i \quad i \in \mathcal{P}. \end{aligned} \quad (1)$$

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z_i = x_i - a_i$, $i = 1, \dots, n$ —and similarly $\underline{z}_i = \underline{x}_i - a_i$ and $\bar{z}_i = \bar{x}_i - a_i$ —, (1) can be recast as:

$$\begin{aligned} \min_z \quad & \|z\|_L \\ \text{subject to} \quad & Az = 0 \\ & \underline{z}_i \leq z_i \leq \bar{z}_i \quad i = 1, \dots, n \\ & z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathcal{P}, \end{aligned} \quad (2)$$

$z \in \mathbb{R}^n$ being the vector of deviations.

It has been observed that the best quality solutions are obtained with the L_1 and L_2 distances (Castro, 2006). Using the L_1 distance, and after some manipulation, (2) can be written as

$$\begin{aligned} \min_{z^+, z^-} \quad & \sum_{i=1}^n w_i (z_i^+ + z_i^-) \\ \text{subject to} \quad & A(z^+ - z^-) = 0 \\ & 0 \leq z_i^+ \leq \bar{z}_i \quad i = 1, \dots, n \\ & 0 \leq z_i^- \leq -\underline{z}_i \quad i = 1, \dots, n \\ & \left\{ \begin{array}{l} z_i^+ \geq upl_i \\ z_i^- = 0 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} z_i^- \geq lpl_i \\ z_i^+ = 0 \end{array} \right\} \quad i \in \mathcal{P}, \end{aligned} \quad (3)$$

z^+ and z^- being the vector of positive and negative deviations in absolute value. For L_2 , we have

$$\begin{aligned} \min_z \quad & \sum_{i=1}^n w_i z_i^2 \\ \text{subject to} \quad & Az = 0 \\ & \underline{z}_i \leq z_i \leq \bar{z}_i \quad i = 1, \dots, n \\ & z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathcal{P}. \end{aligned} \quad (4)$$

Combinations of L_1 and L_2 were tested in Castro (2004).

In practice the sense for the “or” constraint is heuristically fixed a priori (Dandekar and Cox, 2002). In the computational results of Section 3 we set the “upper level protection” for all the sensitive cells. This can lead to infeasible problems, as it will be discussed in Section 4. An alternative that overcomes the infeasibility at the expense of increasing the

computational complexity, is to include the “or” decision within the mathematical model (1), adding a binary variable y_i and two extra constraints for each confidential cell:

$$\begin{aligned} x_i &\geq -M(1 - y_i) + (a_i + \text{upl}_i)y_i & i \in \mathcal{P}, \\ x_i &\leq My_i + (a_i - \text{lpl}_i)(1 - y_i) & i \in \mathcal{P}, \\ y_i &\in \{0, 1\} & i \in \mathcal{P}, \end{aligned} \tag{5}$$

M in (5) being a large value. In terms of deviations, the equivalent constraints for the L_1 model (3) are

$$\begin{aligned} \text{upl}_i y_i &\leq z_i^+ \leq My_i & i \in \mathcal{P}, \\ \text{lpl}_i(1 - y_i) &\leq z_i^- \leq M(1 - y_i) & i \in \mathcal{P}, \\ y_i &\in \{0, 1\} & i \in \mathcal{P}; \end{aligned} \tag{6}$$

and for the L_2 model (4) we should add

$$\begin{aligned} z_i &\geq -M(1 - y_i) + \text{upl}_i y_i & i \in \mathcal{P}, \\ z_i &\leq My_i - \text{lpl}_i(1 - y_i) & i \in \mathcal{P}, \\ y_i &\in \{0, 1\} & i \in \mathcal{P}. \end{aligned} \tag{7}$$

The above constraints result in a combinatorial optimization problem, which is discussed in Section 4.

3 Quality issues and computational experience

From the perspective of a data provider, one question with CTA is how to publish the protected data set. If any adjusted cell is flagged as such, how should users know which cells are still sufficiently reliable for the kind of analysis they are interested in? It might even seem to users that information loss caused by CTA is still worse than information loss caused by cell suppression, because CTA tends to affect more cells, although many cells are adjusted only very slightly. In the following we therefore suggest criteria that could be used to decide whether an adjustment is so small that there is no need for the cell to be flagged as “adjusted”. Needless to say, that it is of course absolutely crucial for a data provider to be very strict in the choice of such a criterion.

For those cells which still have to be flagged as “adjusted”, data users should be provided with a quality indicator, like for instance the distribution of adjusted cells by ranges of their relative deviation.

In that sense, we could say the quality of the result of a CTA procedure is acceptable, if, on one hand, the number of cells that have to be flagged as adjusted because the adjustment is not “sufficiently small”, does not exceed the number of cells that are suppressed, when cell suppression is applied instead of CTA. This is especially important for the subset of cells on a high level of a table. On the other hand, also among those cells that indeed have to be flagged as adjusted the number of cells with very large relative deviations should be small.

While for smaller cells with a “limited” number of respondents it will probably be enough to request that the adjustment should be below a given percentage in order to consider it as “sufficiently small”, for large cells we do not think that such a criterion is strict enough. For those cells, relative deviations should tend to zero with increasing cell size, in order to be considered “sufficiently small”.

For the following evaluation of alternative CTA variants with different weight functions, we consider three criteria: The adjustment of cell i will be considered “sufficiently small”, when the relative deviation is less than

- $1/r_i$ (condition 1),
- or $1/\sqrt{a_i}$ (condition 2),
- or $1/\sqrt[3]{a_i}$ (condition 3),

where r_i denotes the number of respondents, and a_i the cell value. These conditions can also be expressed in terms of absolute deviations, which should be limited by either

- a_i/r_i (condition 1),
- or by $\sqrt{a_i}$ (condition 2),
- or by $\sqrt[3]{a_i^2}$ (condition 3).

So, the objective is on one hand, to avoid that in the protected table there are large deviations in cells that provide aggregated information on a high level and tend to be large, and on the other the number of cells with large relative deviations (e.g., over 5% or 10%) should be low. These are contradictory objectives. Large absolute deviations in large cells are avoided if we choose cell weights $w_i = 1$ in (3) or (4). On the other hand, relative deviations are kept small for $w_i = 1/a_i$ (if $a_i = 0$ the cell can not be perturbed, and we set w_i to any value, e.g., 1). Both weights belong to the family $w_i = 1/a_i^\gamma$, for $\gamma = 0$ and $\gamma = 1$. Weights with $\gamma = 0.5$ are also a reasonable choice, since in theory they should balance relative and absolute deviations.

3.1 Computational results

We tested the three weights for $\gamma = 0, 1/2, 1$ and the L_1 and L_2 distances with the set of seven complex instances used in Dandekar (2003) and Castro (2006). For results on those instances see Castro and Giessing (2006).

In the following we discuss results on an instance given by a tabulation of a strongly skewed variable (like “turnover”, f.i.), typical for business statistics. It is a 3 dimensional table where one of the 3 variables is hierarchical with 3 levels. The table has 5940 cells, 621 sensitive cells, 1464 linear relations, and 18180 coefficients in constraints matrix. Plots a), and b) of Figure 1 show the deviations obtained for the cell values (in log scale), for $\gamma = 0$ and $\gamma = 1/2$. As expected the pattern for $\gamma = 0$ provides lower variability, and most deviations concentrate around 0. The number of cells by ranges of relative deviations is shown in Table 1. From that table it is clear that $\gamma = 0$ gives the greatest number of cells with large relative deviations. The opposite behaviour is observed for $\gamma = 1$. For $\gamma = 1/2$ we get a small number of cells with large relative deviations, although, from Figure 1, deviations are still fairly large for the highest-valued cells, mainly for L_1 . As a compromise closer to $\gamma = 0$ we considered weights $w_i = 1/\log a_i$, both for L_1 and L_2 ; corresponding results are shown in last two columns of Table 1.

However, the patterns of Figure 1 only give a first impression of the performance with respect to the quality issue we are actually interested in, e.g. that cells on a high level

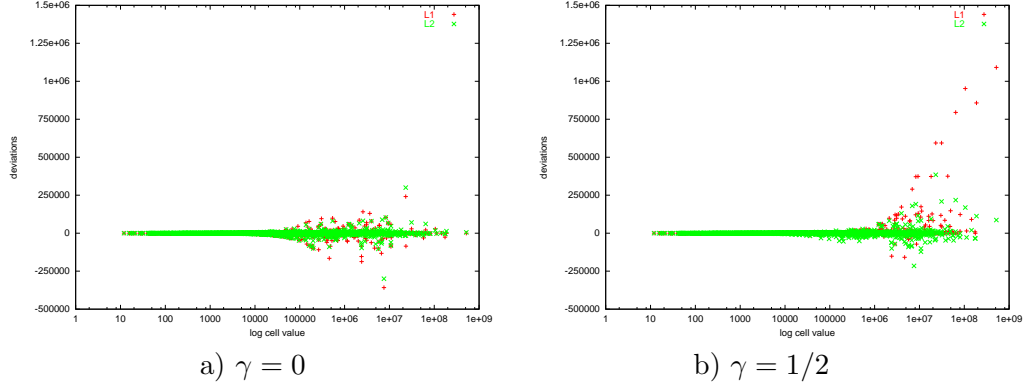


Figure 1: Deviations for a) $\gamma = 0$, b) $\gamma = 1/2$ for L_1 and L_2

| Range | $\gamma = 0$ | | $\gamma = 1/2$ | | $\gamma = 1$ | | $w_i = 1/\log a_i$ | |
|------------|--------------|-------|----------------|-------|--------------|-------|--------------------|-------|
| | L_1 | L_2 | L_1 | L_2 | L_1 | L_2 | L_1 | L_2 |
| 0% | 2164 | 0 | 2407 | 0 | 2439 | 0 | 2300 | 0 |
| (0%,2%] | 540 | 1812 | 677 | 2280 | 655 | 2841 | 644 | 1857 |
| (2%,5%] | 164 | 396 | 125 | 416 | 119 | 309 | 136 | 402 |
| (5%,10%] | 78 | 233 | 7 | 195 | 4 | 61 | 42 | 252 |
| (10%,100%] | 274 | 779 | 4 | 329 | 3 | 9 | 98 | 709 |

Table 1: Number of cells by ranges of relative deviation for $\gamma = 0, 1, 1/2$ and $w_i = 1/\log a_i$

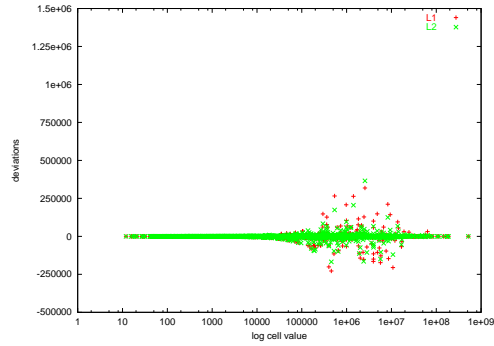


Figure 2: Deviations for adaptive γ according to cell hierarchies for L_1 and L_2

| Range | L_1 | L_2 |
|------------|-------|-------|
| 0% | 2320 | 0 |
| (0%,2%] | 577 | 2233 |
| (2%,5%] | 124 | 423 |
| (5%,10%] | 63 | 223 |
| (10%,100%] | 136 | 341 |

Table 2: Number of cells by ranges of relative deviation for adaptive γ

| condition | $\gamma = 1$ | | $\gamma = 0$ | | $w_i = 1/\log a_i$ | adaptive γ |
|-----------|--------------|-------|--------------|-------|--------------------|-------------------|
| | L_1 | L_2 | L_1 | L_2 | L_1 | L_1 |
| 1 | 85 | 90 | 37 | 90 | 51 | 8 |
| 2 | 83 | 82 | 33 | 65 | 48 | 11 |
| 3 | 38 | 17 | 15 | 19 | 18 | 7 |

Table 3: Number of high level cell values that changed too much according to conditions 1 to 3 formulated above

of aggregation should remain unchanged, or be only slightly modified. Most of these cells are among the cells with the largest values, but some are not. A more direct approach to achieve the goal of small deviations for high-level cells is to choose the parameter γ adaptively according to the cell hierarchy, such that cells with large hierarchies (i.e., national cells) have γ close to 0 (i.e., absolute deviations minimized), and low hierarchy cells have γ close to 1 (i.e., relative deviations minimized). Assuming that $h_i, i = 1, \dots, n$ gives the hierarchy of cell i , and that $\bar{h} = \max\{h_i, i = 1, \dots, n\}$ the rule considered was

$$\gamma_i = \frac{(\bar{h} - h_i)}{\bar{h}}.$$

Figure 2 shows the deviations by cell value for these adaptive γ values. We observe that the adaptive γ provides deviations closer to those obtained with $\gamma = 0$. As for the relative deviations, Table 2 reports the number of cells by ranges of relative deviations. The adaptive γ provides better results than $\gamma = 0$, but the number of cells with large relative deviations is still greater than for $\gamma = 1$.

We imagine now that data providers request that on the top levels of a hierarchical table, CTA should present as many *reliable* results as cell suppression. For our instance we consider as top levels the 2 top levels of the hierarchical variable which are *inner* cells with respect to at most one of the non-hierarchical variables. Within this set of 111 cells, the modular method of τ -ARGUS selects 20 secondary suppressions. If, for instance, we consider a high-level cell value as changed too much a for publication, when, according to condition 2 above for relative deviations the amount of change exceeds $1/\sqrt{a_i}$, we find that only adaptive γ for L_1 leads to an acceptable result: 11 cells change too much, while all other CTA variants lead to more than 20 cells lost because they lack precision (see Table 3).

| Hierarchy | w_i | | | | | | |
|-----------|--------------------|-------|----------------------|--------------------|------------------------|---------|------------------------|
| | $1/a_i^{\gamma_i}$ | r_i | $r_i/a_i^{\gamma_i}$ | $1/r_i^{\gamma_i}$ | $1/a_i^{\gamma_i/r_i}$ | r_i^2 | $r_i^2/a_i^{\gamma_i}$ |
| 0 | 311 | 1048 | 438 | 376 | 302 | 2172 | 753 |
| 1 | 34 | 76 | 48 | 55 | 59 | 163 | 58 |
| 2 | 62 | 128 | 68 | 105 | 82 | 768 | 85 |
| 3 | 5 | 18 | 7 | 11 | 19 | 45 | 7 |
| 4 | 2 | 2 | 2 | 4 | 2 | 53 | 2 |
| 5 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

a)

| Hierarchy | w_i | | | | | | |
|-----------|--------------------|-------|----------------------|--------------------|------------------------|---------|------------------------|
| | $1/a_i^{\gamma_i}$ | r_i | $r_i/a_i^{\gamma_i}$ | $1/r_i^{\gamma_i}$ | $1/a_i^{\gamma_i/r_i}$ | r_i^2 | $r_i^2/a_i^{\gamma_i}$ |
| 0 | 358 | 1125 | 540 | 352 | 297 | 2179 | 856 |
| 1 | 50 | 94 | 68 | 65 | 70 | 165 | 77 |
| 2 | 81 | 171 | 100 | 125 | 96 | 775 | 122 |
| 3 | 7 | 20 | 8 | 13 | 17 | 45 | 8 |
| 4 | 2 | 2 | 2 | 5 | 2 | 53 | 2 |
| 5 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

b)

Table 4: Number of nonsensitive cells with a relative deviation a) greater than $1/r_i$, and b) greater than $1/\sqrt{a_i}$, for L_1

3.2 Exploiting cell respondents information

Number of respondents for the cell (i.e., number of individuals contributing to the cell value) is a second alternative for adjusting weights. Several weights can be devised using this information, but, in general, the common rationale for them is to allow more deviations in cells with few respondents. The reasons are, first, that statisticians consider cells with few respondents less reliable; and second, that cells with many respondents are usually associated to top hierarchy cells, which should remain unchanged.

Table 4 shows the results obtained with L_1 and the following seven different weights (r_i being the number of respondents of cell i):

1. $w_i = 1/a_i^{\gamma_i}$ is the adaptive weight according to cell hierarchy of previous subsection.
2. $w_i = r_i$: cells with few respondents will be more likely changed.
3. $w_i = r_i/a_i^{\gamma_i}$: it combines the hierarchy and respondents cell information.
4. $w_i = 1/r_i^{\gamma_i}$: the same that option 1 above replacing a_i by r_i since cells with large value usually involve more respondents.

5. $w_i = 1/a_i^{\gamma_i/r_i}$: dividing γ_i by r_i increases the weights of cells with large respondents, such that cells with few respondents will more likely be changed.
6. $w_i = r_i^2$: the same that option 2 above using the square of r_i .
7. $w_i = r_i^2/a_i^{\gamma_i}$: the same that option 3 above using the square of r_i .

The table shows, for each weight, the number of cells of each of the seven hierarchies that changed too much. Subtable 4.a reports results observed on basis of condition 1 above, i.e. relative deviations exceeding $1/r_i$ are considered too large. On basis of condition 2, considering relative deviations exceeding $1/\sqrt{a_i}$ as too large, we obtain the results reported in subtable 7.b. .

Looking at the higher hierarchies (e.g., from 3 to 6) we see that $w_i = 1/a_i^{\gamma_i}$ provides the minimum number of cells with large deviations for both criteria. Although in other instances some of the alternative weights involving the respondents information could be better, $w_i = 1/a_i^{\gamma_i}$ should be regarded in general as a good choice.

4 The restricted CTA method

Large relative deviations, independently of the weights used, can be avoided by imposing constraints

$$(1 - \alpha_i)a_i \leq x_i \leq (1 + \beta_i)a_i \quad i = 1, \dots, n, \quad (8)$$

for some $\alpha_i, \beta_i \geq 0$, to the general model (1), or, equivalently,

$$\begin{aligned} 0 &\leq z_i^+ \leq \beta_i a_i & i = 1, \dots, n \\ 0 &\leq z_i^- \leq \alpha_i a_i & i = 1, \dots, n \end{aligned} \quad (9)$$

for the L_1 model (3), and

$$-\alpha_i a_i \leq z_i \leq \beta_i a_i \quad i = 1, \dots, n \geq 0, \quad (10)$$

for the L_2 model (4). The parameters α_i and β_i bound the relative deviations on cell values. Imposing, e.g., $\alpha_i = \beta_i = 0.05$ for all $i = 1, \dots, n$ we avoid relative deviations larger than 5%. Imposing $\alpha_i = \beta_i = 0.0, i \in \mathcal{F}$ for some subset of cells \mathcal{F} , we guarantee that cells of \mathcal{F} will remain unchanged in the protected table. The resulting procedure is more restrictive than the original CTA method, since deviations are only allowed in some cells, and such deviations are confined within some bounds. We call the new procedure the Restricted Controlled Tabular Adjustment (RCTA for short).

The main benefit of RCTA is that we can precisely control through constraints, instead of through the weights, the relative deviations of the cells. The drawback is that small values for α_i and β_i result in infeasible problems, at least if the sense of protection (“upper” or “lower”) is a priori fixed.

To avoid infeasibility problems with RCTA we are forced to include in the optimization problem the binary decision for the “upper” or “lower” protection sense, either adding constraints (6) to the L_1 model (3) or adding (7) to the L_2 model (4). Unfortunately this transforms the linear and quadratic models for L_1 and L_2 to combinatorial ones, significantly increasing the solution time. For instance, for L_1 and the instance of section 3 we attempted

the optimization problem (3,6), using the mixed-integer-programming solver of Cplex 9.1 on a Pentium-4 at 1.8GHz. We stopped the procedure after 10 hours of CPU without a solution. The same model without the binary constraints (6) is solved in about 1 second.

Possible solution strategies to overcome the excessive time of RCTA with the binary variables are:

- Optimal solution through Bender’s decomposition, moving binary decisions to a master problem, and solving a sequence of the easy continuous subproblems (3) or (4).
- Use of a heuristic for a good initial choice of the protection senses (either “lower” or “upper”). Once fixed, only one solution of either (3) or (4) is needed.
- Metaheuristic, as genetic algorithms, for adjusting the binary decisions, which involves the solution of a sequence of subproblems (3) or (4).
- The last option consists of removing the binary decisions, and to allow deviations go beyond their bounds, penalizing such bound violations in the objective function by a large penalty term. This guarantees an always feasible problem, at the expense of providing a table with some unprotected sensitive cells. Only one easy linear or quadratic problem has to be solved in that case, but some kind of post processing is eventually required to fix underprotection problems.

All the previous approaches are currently being investigated by the authors.

5 Summary and final conclusions

In this paper, we have compared several variants of CTA on an instance from business statistics. Our experiments show that at least in the context of strongly skewed business data, the parameters of a CTA approach, such as the choice of a particular cost function, have considerable effect on the output data quality. Spending some effort here on fine tuning of a method seems to be worthwhile.

As CTA is discussed as an alternative to well established cell suppression, we also included a quality criterion that allows direct comparison of the performance of CTA to cell suppression, to some extent. First results are promising, indicating that it may be possible to make CTA procedures provide at least as much data meeting the high data quality standards of official statistics for data of a particular relevance as cell suppression. We also suggested restricted RCTA as an option to combine cell suppression and CTA, or to facilitate use of CTA in the context of linked tables. RCTA allows to control relative and absolute deviations more precisely than CTA. Unfortunately, RCTA is more sensible to the protection sense (“upper” or “lower”) of sensitive cells than CTA, leading to infeasibility problems. Several strategies have been discussed for a proper choice of protection sense, leading to both optimal and heuristic solutions. Heuristic solutions are likely to be the best practical option, since they will provide a reasonable quality protected table within reasonable time. All these approaches for RCTA are currently under development by the authors.

References

- Castro, J. (2004), Computational experiments with minimum-distance controlled perturbation methods, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 73–86. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.
- Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection, *European Journal of Operational Research*, 171, 39–52.
- Castro, J., Giessing S. (2006). Testing variants of minimum distance controlled tabular adjustment, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 333-343.
- Cox, L. H., Kelly, J. P., and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 87–98. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.
- Dandekar, R.A. (2003), Cost effective implementation of synthetic tabulation (a.k.a. controlled tabular adjustments) in legacy and state of the art statistical data publication systems, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg. Available from <http://www.unece.org/stats/documents/2003.-04.confidentiality.htm>.
- Dandekar, R.A. (2005), personal communication.
- Dandekar, R.A., and Cox, L.H. (2002), Synthetic tabular data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. Department of Energy. Available from the first author on request (Ramesh.Dandekar@eia.doe.gov).
- Giessing, S. (2004), Survey on methods for tabular data protection in ARGUS, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 1–13. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.
- Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P. (2004), τ -ARGUS users’s manual, version 3.0.