WP. 40
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vi): Software for statistical disclosure control

TESTING VARIANTS OF MINIMUM DISTANCE CONTROLLED TABULAR ADJUSTMENT

Invited Paper

Submitted by the Federal Statistical Office, Germany and University of Catalunya, Spain[1]

[1] Prepared by Sarah Giessing (sarah.giessing@destatis.de) and Jordi Castro (jordi.castro@upc.edu).

Testing variants 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. As another alternative, we suggest a restricted CTA (RCTA) approach, where adjustments are only allowed in a subset of cells. This subset is a priori computed, for instance by a fast heuristic for the cell suppression problem. We discuss benefits of RCTA, and suggest several approaches for its solution.

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

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

certain restrictions on the release of tabular data which is in some contrast to the flexibility and 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, 2005b) 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'. In order to introduce CTA into practice, it is therefore 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:

- The number of cells with a large relative deviation (i.e., over 5%, 10%, or any other predefined threshold value) should be as low as possible (hopefully, zero). Such large deviations are in some sense equivalent to the suppression of the cell, which is exactly the technique we plan to replace by using CTA.

- Cells that provide aggregated information on a high level (for geography, for instance, state, or whole country level), should remain unchanged, or only slightly modified.

- 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 reports and analyzes the results obtained with some close-to-real instances. In Section 4 we discuss a restricted CTA procedure, which improves the quality of the protected tables, although it significantly increases the solution time. Some strategies are discussed for the efficient solution of the restricted CTA procedure.

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 $\overline{a}_i$, 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 $\mathcal{P} = \{i_1, i_2, \dots, i_p\} \subseteq \{1, \dots, n\}$ 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:

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

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

$$\begin{array}{rl} \min\limits_{z} & ||z||_L \\ \text{subject to} & Az = 0 \\ & \underline{z}_i \leq z_i \leq \overline{z}_i \quad i = 1, \dots, n \\ & z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathcal{P}, \end{array} \tag{2}$$

$z \in \mathbb{R}^n$ 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, 2005b). Using the L_1 distance, and after some manipulation,

(2) can be written as

$$\begin{array}{lllr}
\min\limits_{z^+,z^-} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) & & \\
\text{subject to} & A(z^+ - z^-) = 0 & & \\
& 0 \leq z_i^+ \leq \ \ \overline{z}_i \quad i = 1,\ldots,n & & (3)\\
& 0 \leq z_i^- \leq -\underline{z}_i \quad i = 1,\ldots,n & & \\
& \left\{ \begin{array}{lll} z_i^+ & \geq & upl_i \\ z_i^- & = & 0 \end{array} \right\} \text{ or } \left\{ \begin{array}{lll} z_i^- & \geq & lpl_i \\ z_i^+ & = & 0 \end{array} \right\} & i \in \mathcal{P}, &
\end{array}$$

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

$$\begin{array}{llr}
\min\limits_{z} & \sum\limits_{i=1}^{n} w_i z_i^2 & \\
\text{subject to} & Az = 0 & (4)\\
& \underline{z}_i \leq z_i \leq \overline{z}_i \quad i = 1,\ldots,n & \\
& z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathcal{P}. &
\end{array}$$

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:

$$\begin{array}{rcll}
x_i & \geq & -M(1-y_i) + (a_i + upl_i)y_i & i \in \mathcal{P}, \\
x_i & \leq & My_i + (a_i - lpl_i)(1-y_i) & i \in \mathcal{P}, \\
y_i & \in & \{0,1\} & i \in \mathcal{P},
\end{array} \qquad (5)$$

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

$$\begin{array}{rcccll}
upl_i y_i & \leq & z_i^+ & \leq & My_i & i \in \mathcal{P}, \\
lpl_i(1-y_i) & \leq & z_i^- & \leq & M(1-y_i) & i \in \mathcal{P}, \\
y_i & \in & \{0,1\} & & & i \in \mathcal{P};
\end{array} \qquad (6)$$

and for the L_2 model (4) we should add

$$\begin{array}{rcll}
z_i & \geq & -M(1-y_i) + upl_i y_i & i \in \mathcal{P}, \\
z_i & \leq & My_i - lpl_i(1-y_i) & i \in \mathcal{P}, \\
y_i & \in & \{0,1\} & i \in \mathcal{P}.
\end{array} \qquad (7)$$

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

Name	n	$\lvert\mathcal{P}\rvert$	m	N.coef
bts4	36570	2260	36310	136912
destatis	5940	621	1464	18180
five20b	34552	3662	52983	208335
five20c	34501	4022	58825	231345
hier13	2020	112	3313	11929
hier16	3564	224	5484	19996
nine12	10399	1178	11362	52624
nine5d	10733	1661	17295	58135
ninenew	6546	858	7340	32920
two5in6	5681	720	9629	34310

Table 1: Dimensions of the complex instances

3 Computational testing

From the perspective of a data provider, it is essential to avoid that in the released table there are large deviations in cells that provide aggregated information on a high level, and at the same time we want to keep the number of cells with large relative deviations (e.g., over 5% or 10%) low. These are contradictory objectives. Large absolute deviations in total 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.

We tested the three weights for $\gamma = 0, 1/2, 1$ and the L_1 and L_2 distances with a set of complex instances: the seven most complex instances used in (Dandekar, 2003; Castro, 2005b) (named “bts”, “hier13”, “hier16”, “nine12”, “nine5d”, “ninenew”, and “two5in6”) which seem to present frequency counts, and a close-to-real instance provided by Destatis (named the ”destatis” instance in the following). The latter instance represents a tabulation of a strongly skewed variable (like ”turnover”, f.i.), typical for business statistics. We also attempted the recently released “five20b” and “five20c” twenty-dimensional tables (Dandekar, 2005). However, unlike the former, which are solved in seconds, these two instances are computationally challenging. For instance, the protection procedure was stopped after 10 hours of CPU time without a solution for “five20b”, using either the dual or primal simplex algorithm of Cplex 9.1 on a Pentium-4 at 1.8GHz; “five20c” was not attempted with simplex algorithms. On the other hand, interior-point algorithms seem to be a more efficient choice for large multidimensional tables. For instance, the interior-point option of

Cplex 9.1 protected the "five20b" and "five20c" instances in, respectively, 10 and 20 minutes of CPU using the L_1 distance, and 5 and 10 minutes of CPU using the L_2 distance. In principle there is room for improvement using specialized interior-point methods, as done for three-dimensional tables in Castro (2005a). Table 1 provides the dimensions of each instance: number of cells (column "n"), number of sensitive cells (column "$|\mathcal{P}|$"), number of constraints (column "m"), and number of nonzeros in constraints matrix (column "N.coef"). Table 2 shows the number of cells with relative deviations between 2% and 5% and over 5% for each value γ. It is observed that, in general, the number of cells with large relative deviations increases when γ tends to zero. Another observation is that for the business data instance the choice of the cost function seems to have a stronger effect as with the other instances.

In the following, we analyze in more detail this instance "destatis". It is a 3 dimensional table where one of the 3 variables is hierarchical with 3 levels. Plots a), b) and c) of Figure 1 show the deviations obtained for the cell values (in log scale). As expected the pattern for $\gamma = 0$ provides the lowest variability, and most deviations concentrate around 0. The number of cells by ranges of relative deviations is shown in Table 3. 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 Figure 3. Another alternative approach has been suggested in Giessing (2004), a heuristic implementation of a 'restricted CTA' (RCTA) procedure which is presented in the following section 4. Table 4 proves that this particular RCTA heuristic, referred to as SUP8 in the figures, outperforms the CTA variant with weights $w_i = 1/\log a_i$ in the sense that it reduces the number of cells with a relative deviation beyond 10% from 98 (for L_1; 709 for L_2) to 1. Comparison of Figures 2 (referring to SUP8) and 3 shows that large changes in large values are also prevented more efficiently as by the L_1 variant.

However, the patterns of Figures 1 to 3 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 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

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Instance	L_1	L_2	L_1	L_2	L_1	L_2
bts4	1402	1515	1016	1184	962	933
destatis	164	396	125	416	119	309
five20b	2841	3013	2478	2815	2426	2605
five20c	3218	3477	2769	3096	2777	2822
hier13	101	103	75	82	79	68
hier16	127	145	108	124	112	95
nine12	787	889	685	787	695	709
nine5d	875	999	947	993	978	918
ninenew	613	646	521	598	531	510
two5in6	451	529	388	499	424	384

a) relative deviation between 2% and 5%

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Instance	L_1	L_2	L_1	L_2	L_1	L_2
bts4	741	799	353	521	279	292
destatis	352	1012	11	524	7	70
five20b	1284	1434	650	1161	445	579
five20c	1352	1542	699	1202	559	706
hier13	32	32	26	27	26	24
hier16	60	69	29	46	17	112
nine12	378	427	162	310	120	149
nine5d	606	724	223	523	163	128
ninenew	298	360	154	258	107	131
two5in6	244	80	128	163	90	86

b) relative deviation greater than 5%

Table 2: Number of cells with a relative deviation between 2% and 5% (a)), and greater than 5% (b)), for $\gamma = 0, 1/2, 1$ and the complex instances

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Range	L_1	L_2	L_1	L_2	L_1	L_2
0%	2164	0	2407	0	2439	0
(0%,2%]	540	1812	677	2280	655	2841
(2%,5%]	164	396	125	416	119	309
(5%,10%]	78	233	7	195	4	61
(10%,100%]	274	779	4	329	3	9

Table 3: Number of cells by ranges of relative deviation for $\gamma = 0, 1, 1/2$ for "destatis" instance

	$w_i = 1/\log a_i$		SUP8
Range	L_1	L_2	
0%	2300	0	2341
(0%,2%]	644	1857	641
(2%,5%]	136	402	169
(5%,10%]	42	252	88
(10%,100%]	98	709	1

Table 4: Number of cells by ranges of relative deviation for $w_i = 1/\log a_i$ and SUP8 for “destatis” instance

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 5: Number of cells by ranges of relative deviation for adaptive γ

$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$		adaptive γ		$w_i = 1/\log a_i$		SUP8
L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2	
33	65	76	86	83	82	11	28	48	66	40

Table 6: Number of high level cell values changed too much for publication

that $\overline{h} = \max\{h_i, i = 1, \ldots, n\}$ the rule considered was

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

Figure 4 shows the deviations by cell value for these adaptive γ values. We observe that the adaptive γ outperforms $\gamma = 1$ and $\gamma = 1/2$, and provides deviations closer to those obtained with $\gamma = 0$. As for the relative deviations, Table 5 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 such highly aggregated data, even a change of 1% is usually considered far too much. For our instance "destatis" 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. For the following analysis, we consider a high-level cell value *a* as *changed too much for publication*, when the amount of change exceeds $\sqrt{a}$. With this concept, 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 6).

In the next section we present ideas to combine cell suppression and CTA methodology which may turn out to be of special interest in the context of protecting linked tables.

4 The restricted CTA method

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

$$(1 - \alpha_i)a_i \le x_i \le (1 + \beta_i)a_i \quad i = 1, \ldots, n, \tag{8}$$

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

$$\begin{array}{lll} 0 & \le z_i^+ \le \beta_i a_i & i = 1, \ldots, n \\ 0 & \le z_i^- \le \alpha_i a_i & i = 1, \ldots, n \end{array} \tag{9}$$

for the L_1 model (3), and

$$-\alpha_i a_i \le z_i \le \beta_i a_i \quad i = 1, \ldots, n \ge 0, \tag{10}$$

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 $i = 1, \ldots, n$ 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. Such a set could f.i. be the set of cells a table has in common with another table that has already been protected in the case of linked tables. In the procedure SUP8 presented in Giessing (2004) this set has been determined by a fast heuristic for the cell suppression problem, i.e. the GHMITER hypercube algorithm (Repsilber, 2002; Giessing, 2003) considering +/-8% a priori bounds on the cell values. For the CTA step deviations in this subset of cells were allowed with at most $\alpha_i = \beta_i = 0.09$. For the cost function we used weights with $\gamma = 1$ and L_1 distances. 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. For instance, imposing $\alpha_i = \beta_i = 0$ in the subset of cells previously computed by the GHMITER hypercube heuristic for cell suppression, instance "destatis" becomes infeasible, naturally, when we use the "upper protection sense" for all primary cells. Even when we allow deviations in all cells with $\alpha_i = \beta_i = 0.1$, instance "destatis" remains infeasible using the "upper protection sense" for all primary cells. For $\alpha_i = \beta_i = 0.5$ the instance becomes feasible, again with the "upper protection sense" for all primary cells. However a 50% of relative variation is impractical.

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 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.

  The SUP8 procedure of Giessing (2004) makes a heuristic choice of the protection senses (Rabenhorst, 2003), solving infeasibility problems by penalizing bound violations. In the “destatis” instance, this resulted in 3 significantly underprotected sensitive cells.

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 with a special focus 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 certain 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. (2005). Quadratic interior-point methods in statistical disclosure control, *Computational Management Science* 2, 107–121.

Castro, J. (2005). Minimum-distance controlled perturbation methods for large-scale tabular data protection, *European Journal of Operational Research*, in press.

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. (2003), Co-ordination of cell suppressions: strategies for use of GHMITER, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg. Available from `http://www.unece.org/stats/documents/2003.-04.confidentiality.htm`.

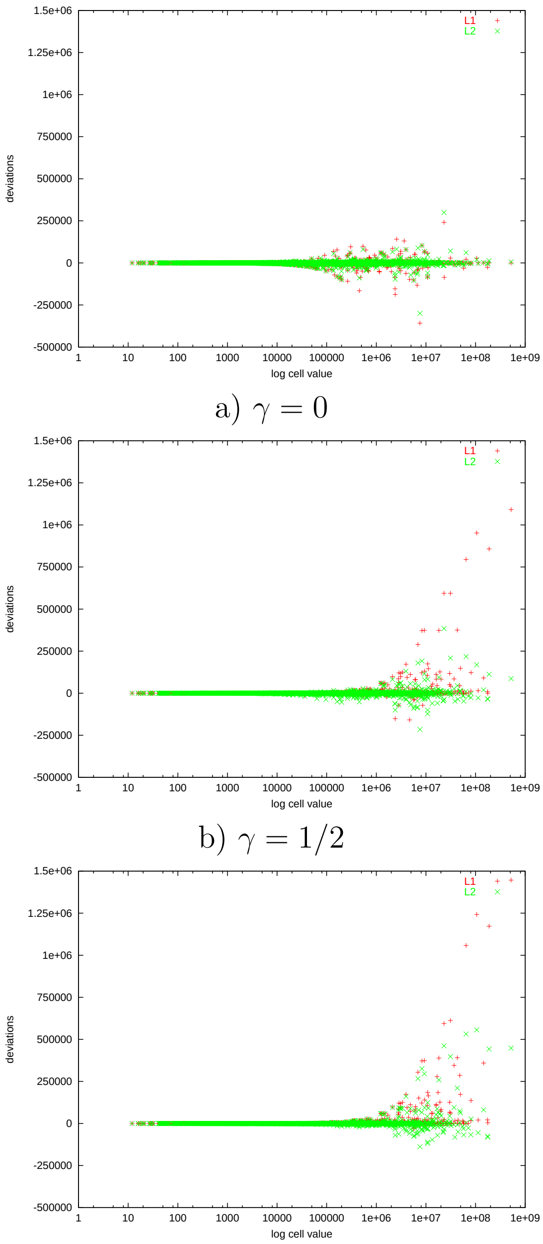
Giessing, S. (2004), Survey on methods for tabular data protection in ARGUS, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 1–13. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.

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.

Rabenhorst, A. (2003), Bestimmung von Intervallen und Ersatzwerten für gesperrte Zellen in statistischen Tabellen, Diploma Thesis, Manuscript, University Ilmenau (in German).

Repsilber, D.(2002), Sicherung persönlicher Angaben in Tabellendaten' - in Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 1/2002 (in German).

Appendix: Figures of the paper



a) $\gamma = 0$

b) $\gamma = 1/2$

c) $\gamma = 1$

Figure 1: Deviations for a) $\gamma = 0$, b) $\gamma = 1/2$ and c) $\gamma = 1$ in the “destatis” instance

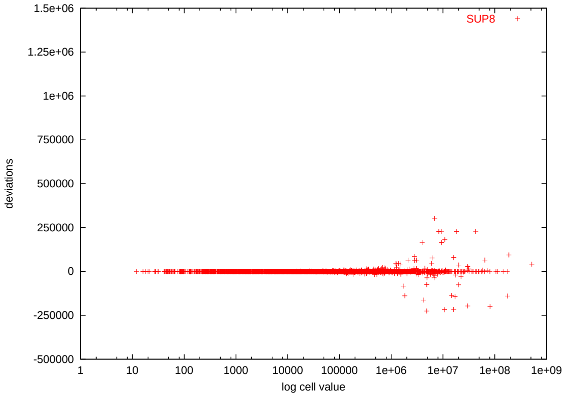


Figure 2: Deviations for SUP8 in the “destatis” instance

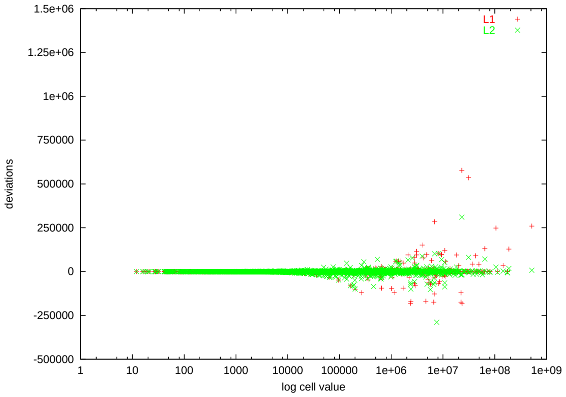


Figure 3: Deviations for weights $w_i = 1/\log a_i$ in the “destatis” instance for L_1 and L_2

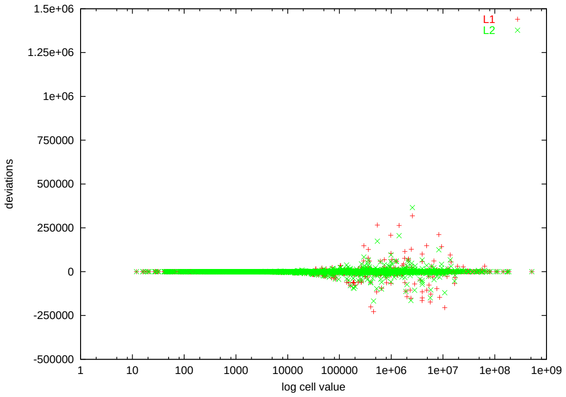


Figure 4: Deviations for adaptive γ according to cell hierarchies. in the “destatis” instance for L_1 and L_2