Comparing L_1 and L_2 Distances for CTA*

Jordi Castro

Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
jordi.castro@upc.edu
http://www-eio.upc.edu/~jcastro

Abstract. Minimum distance controlled tabular adjustment (CTA) is a recent perturbative technique of statistical disclosure control for tabular data. Given a table to be protected, CTA looks for the closest safe table, using some particular distance. We focus on the continuous formulation of CTA, without binary variables, which results in a convex optimization problem for distances L_1, L_2 and L_∞. We also introduce the L_0-CTA problem, which results in a combinatorial optimization problem. The two more practical approaches, L_1-CTA (linear optimization problem) and L_2-CTA (quadratic optimization problem) are empirically compared on a set of public domain instances. The results show that, depending on the criteria considered, each of them is a better option.

Keywords: statistical disclosure control, controlled tabular adjustment, linear optimization, quadratic optimization.

1 Introduction

Controlled tabular adjustment methods (CTA) [1,7] are considered an emerging technology for tabular data [10]. In terms of efficiency and quality of the solution, they usually perform well compared to other techniques [2,3].

CTA was initially [7] only formulated for L_1 norms and binary variables for deciding the sense of protection for the sensitive cells, i.e., whether to perturb up or down the original cell value. In [1] L_2 and L_∞ were also considered in continuous formulations, i.e., the protection sense was a priori fixed without paying attention to infeasibility issues [4]. Results for the two most practical distances, L_1 and L_2, were presented in [1,3], but without a detailed comparison of the reported solutions. In addition, the same cell weights were used for L_1 and L_2 in the empirical results of [1,3]; as it will be stated later in this work, the comparison was unfair, since the weights used favored L_1. This work tries to fill this void by performing a more exhaustive empirical evaluation of L_1-CTA versus L_2-CTA. A new variant L_0-CTA, closer to L_1-CTA than to L_2-CTA, will also be formulated.

* Supported by grants MTM2009-08747 of the Spanish Ministry of Science and Innovation, SGR-2009-1122 of the Government of Catalonia, and INFRA-2010-262608 of the European Union.

J. Domingo-Ferrer and I. Tinnirello (Eds.): PSD 2012, LNCS 7556, pp. 35–46, 2012.

The paper is organized as follows. Section 2 formulates the continuous CTA problem (i.e., a priory fixing the value of binary variables) for L_0, L_1, L_2 and L_∞. Section 3 introduces the criteria considered in the comparison of L_1-CTA and L_2-CTA. Finally, Section 4 reports the computational comparison.

2 Formulations of CTA for Several Distances

Any CTA instance can be represented by the following parameters:

- A set of cells $a_i, i \in \mathcal{N} = \{1, \ldots, n\}$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's). The particular structure of the table is defined by equations $Aa = b$. Each tabular constraint imposes that the inner cells have to be equal to the total or marginal cell. Any type of table can be modeled by these equations.
- A lower and upper bound for each cell $i \in \mathcal{N}$, respectively l_{a_i} and u_{a_i}, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i $l_{a_i} = 0$ ($l_{a_i} = -\infty$ if $a \geq 0$ is not required) and $u_{a_i} = +\infty$ can be used.
- A set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N}$ of indices of confidential cells.
- Nonnegative lower and upper protection levels for each confidential cell $i \in \mathcal{S}$, 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 $x_i, i = 1, \ldots, n$, according to some distance L, that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{x} & ||x - a||_L \\ \text{s. to} & Ax = b \\ & l_{a_i} \leq x_i \leq u_{a_i} \quad i \in \mathcal{N} \\ & (x_i \leq a_i - lpl_i) \text{ or } (x_i \geq a_i + upl_i) \quad i \in \mathcal{S}. \end{array} \tag{1}$$

Introducing a vector of binary variables $y \in \mathbb{R}^s$ to model the disjunctive constraints (either "upper protection sense" $x_i \geq a_i + upl_i$ when $y_i = 1$ or "lower protection sense" $x_i \leq a_i - lpl_i$ when $y_i = 0$), the above problem can be formulated as a mixed integer linear optimization problem (MILP), which can be time consuming for medium-large instances.

A more efficient alternative for the real-time protection in on-line tabular data servers —or in other situations where processing time matters (like when protecting very large sets of linked tables)— would be to a priori fix the binary variables, thus obtaining a CTA formulation with only continuous variables [5]. Possible infeasibilities in the resulting problem could be dealt with the approaches exposed in [6], some of them already used in the context of CTA [4]. Formulating problem (1) in terms of cell deviations $z = x - a$, and fixing the binary

variables, the resulting continuous CTA problem can be, in general, formulated as the following convex optimization problem:

$$\begin{aligned} \min_{z} \ & ||z||_L \\ \text{s. to } & Az = 0 \\ & l \leq z \leq u, \end{aligned} \tag{2}$$

where

$$\begin{aligned} l_i &= \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ l_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0) \end{cases} \\ u_i &= \begin{cases} -lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1). \end{cases} \end{aligned} \tag{3}$$

Problem (2) can be specialized for several norms. For L_1, defining $z = z^+ - z^-$, we obtain the following linear optimization problem (LP):

$$\begin{aligned} \min_{z^+, z^-} \ & \sum_{i=1}^{n} w_i (z_i^+ + z_i^-) \\ \text{s. to } & A(z^+ - z^-) = 0 \\ & l^+ \leq z^+ \leq u^+ \\ & l^- \leq z^- \leq u^-, \end{aligned} \tag{4}$$

$w \in \mathbb{R}^n$ being a vector of nonnegative cell weights, $z^+ \in \mathbb{R}^n$ and $z^- \in \mathbb{R}^n$ the vector of positive and negative deviations in absolute value, and $l^+, l^-, u^+, u^- \in \mathbb{R}^n$ lower and upper bounds for the positive and negative deviations defined as

$$\begin{aligned} l_i^+ &= \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ 0 & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0) \end{cases} \\ u_i^+ &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1) \end{cases} \\ l_i^- &= \begin{cases} lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ 0 & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1) \end{cases} \\ u_i^- &= \begin{cases} 0 & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ a_i - l_{a_i} & \text{if } (i \in \mathcal{N} \setminus \mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0). \end{cases} \end{aligned} \tag{5}$$

For L_2, problem (2) can be recast as the following quadratic optimization problem (QP):

$$\begin{aligned} \min_{z} \ & \sum_{i=1}^{n} w_i z_i^2 \\ \text{s. to } & Az = 0 \\ & l \leq z \leq u. \end{aligned} \tag{6}$$

Although L_1 and L_2 are the most practical norms, we provide formulations for two additional ones, L_∞ and L_0. For L_∞, adding an extra auxiliary variable $t \in \mathbb{R}$ and considering, as for the L_1 distance, $z = z^+ - z^-$, we have:

$$\begin{array}{ll} \min\limits_{z} & t \\ \text{s. to} & Az = 0 \\ & t \geq z_i^+ + z_i^- \quad i \in \mathcal{N} \\ & l^+ \leq z^+ \leq u^+ \\ & l^- \leq z^- \leq u^-, \end{array} \tag{7}$$

$l^+, l^-, u^+, u^- \in \mathbb{R}^n$ defined as above.

The L_0 norm is a measure of the sparsity of a vector, and it is defined as the number of nonzero elements in the vector. In the context of CTA the minimization of the L_0 norm would provide the vector of deviations that modifies the smallest number of cells. The main inconvenience of this norm is that, even if the binary variables y of CTA have been fixed, it results in a combinatorial optimization problem. L_0-CTA is formulated as

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n} d_i \\ \text{s. to} & Az = 0 \\ & l \leq z \leq u \\ & l_i d_i \leq z_i \leq u_i d_i \quad i \in \mathcal{N} \\ & d \in \{0,1\}^n, \end{array} \tag{8}$$

such that $d_i = 1$ if cell i changes, and 0 otherwise.

L_0-CTA is a difficult combinatorial problem. L_1 norms are known to provide sparse enough solutions in practice [8]. Indeed, since (4) is a LP, its solutions are basic (i.e., in vertices of the polyhedron), which are sparse. This does not hold for the quadratic optimization problem (6) of L_2-CTA. Therefore, L_1-CTA is a better approximation to L_0-CTA than L_2-CTA. This assertion will be empirically observed in Section 4.

3 Criteria for the Comparison of L_1-CTA and L_2-CTA

L_1 and L_2 are widely used in regression analysis. L_2 regression (ordinary least squares) is simpler to compute, but L_1 regression is known to be more robust in the presence of outliers. However, there is one main difference between linear regression and CTA: in linear regression we have observations (possibly with outliers) and look for the theoretical linear model; on the other hand, in CTA we already have the "theoretical model" (the original table) and look for the "observations" (a perturbed safe table). Therefore, by the same reason that L_2 regression is more sensitive to outliers than L_1 regression, L_2-CTA will hardly provide a table with large cell deviations ("outliers"), compared to L_1-CTA. This is also clear from the different objectives of L_2-CTA (quadratic function) and L_1-CTA (linear function).

However, if we look at the comparison made in [3] between the absolute relative deviations of L_1-CTA and L_2-CTA, we note that, unexpectedly, L_2-CTA provided larger absolute relative deviations than L_1-CTA. However, this was due to the use of the weights of (4) in (6). If our purpose is to minimize the overall absolute relative deviations, then the cell weights of (4) have to be defined as $w_i = 1/a_i$, and the objective function becomes $\sum_{i=1}^{n} \frac{z_i^+ + z_i^-}{a_i}$. To achieve the same goal in (6) these weights are no longer valid, and instead we should use $w'_i = 1/a_i^2$, such that the objective function of (6) is $\sum_{i=1}^{n} \left(\frac{z_i}{a_i}\right)^2$. The computational results of Section 4 for L_1-CTA and L_2-CTA were obtained using these two different sets of weights for each problem. As it will be shown, using these correct weights, L_2-CTA provides less large relative deviations, as expected.

The particular criteria selected for the comparison of problems (4) and (6) are the following:

- CPU time, for the efficiency of each model.
- Mean and standard deviation of the absolute relative deviations. They are included as basic statistics.
- The maximum absolute relative deviation, which according to the previous discussion, should be in general smaller for L_2-CTA than for L_1-CTA (as it will be validated by the computational results of next section). From now on, "relative deviations" will be used for "absolute relative deviations".
- The number of cells with "large" relative deviations. This criteria was introduced in [3] as a measure of "data utility": the smaller the number of cells with large relative deviations, the higher the data utility. In this work we considered "large" relative deviations greater than a certain threshold value: one fourth of the maximum relative deviation provided by L_1-CTA (which was always greater than the maximum relative deviation of L_2-CTA in the tests performed). This particular threshold value was obtained by observing in the distribution of relative deviations some few very large values, whereas most of them were concentrated around 0.
- The number of nonzero relative deviations, i.e., the L_0 norm of the vector of relative deviations, or equivalently, the number of cells that change its value in the protected table. According to the discussion in Section 2, in theory this number should be smaller for L_1-CTA than for L_2-CTA. The computational results of next section validate this assertion.

4 Computational Results

We have considered a set of 36 public instances which can be found in the literature (e.g., in [1,2]). Table 1 shows the main dimensions of these instances and the solution time needed to solve the optimization problems (4) and (6) with the interior-point algorithm of the state-of-the-art optimization package Cplex [11]. Interior-point algorithms [12] have shown to be, in general, the most efficient approach for CTA formulations involving only continuous variables [1].

Table 1. Dimensions and solution time of the test instances

					L_1		L_2	
instance	n	s	m	nz	vars	CPU	vars	CPU
australia_ABS	24420	918	274	13224	48840	0.2	24420	0.08
bts4	36570	2260	36310	136912	73140	9.54	36570	4.87
cbs	11163	2467	244	22326	22326	0.04	11163	0.01
dale	16514	4923	405	33028	33028	0.32	16514	0.07
destatis	5940	621	1464	18180	11880	0.31	5940	0.5
five20b	34552	3662	52983	208335	69104	54.31	34552	29.28
five20c	34501	4022	58825	231345	69002	152.88	34501	41.69
hier13d4	18969	2188	47675	143953	37938	809.37	18969	45.23
hier13	2020	112	3313	11929	4040	0.78	2020	0.6
hier13x13x13a	2197	108	3549	11661	4394	0.7	2197	0.67
hier13x13x13b	2197	108	3549	11661	4394	0.88	2197	0.65
hier13x13x13c	2197	108	3549	11661	4394	0.77	2197	0.54
hier13x13x13d	2197	108	3549	11661	4394	0.75	2197	0.58
hier13x13x13e	2197	112	3549	11661	4394	0.68	2197	0.44
hier13x13x7d	1183	75	1443	5369	2366	0.19	1183	0.07
hier13x7x7d	637	50	525	2401	1274	0.06	637	0.03
hier16	3564	224	5484	19996	7128	2.59	3564	2.68
hier16x16x16a	4096	224	5376	21504	8192	2.21	4096	2.3
hier16x16x16b	4096	224	5376	21504	8192	2.12	4096	3.26
hier16x16x16c	4096	224	5376	21504	8192	2.08	4096	2.49
hier16x16x16d	4096	224	5376	21504	8192	2.02	4096	2.52
hier16x16x16e	4096	224	5376	21504	8192	2.07	4096	2.41
nine12	10399	1178	11362	52624	20798	5.99	10399	9.47
nine5d	10733	1661	17295	58135	21466	3.48	10733	4.86
ninenew	6546	858	7340	32920	13092	4.11	6546	6.61
osorio	10201	7	202	20402	20402	0.17	10201	0.07
table1	1584	146	510	4752	3168	0.1	1584	0.03
table3	4992	517	2464	19968	9984	0.39	4992	0.43
table4	4992	517	2464	19968	9984	0.38	4992	0.41
table5	4992	517	2464	19968	9984	0.35	4992	0.79
table6	1584	146	510	4752	3168	0.07	1584	0.02
table7	624	17	230	1872	1248	0.04	624	0.01
table8	1271	3	72	2542	2542	0.02	1271	0
targus	162	13	63	360	324	0	162	0
toy3dsarah	2890	376	1649	9690	5780	0.06	2890	0.04
two5in6	5681	720	9629	34310	11362	1.58	5681	1.71

The AMPL modeling language was used to implement the L_1-CTA and L_2-CTA models (but the CPU time shown in Table 1 only corresponds to the time spent by Cplex in the optimization process). Columns n, s, m and "nz" report, respectively, the number of cells, number of sensitive cells, number of constraints and number of nonzeros of the constraints matrix A. Columns "vars" and "CPU" provide the number of variables and solution time of, respectively, the optimization problems (4) and (6) for L_1 and L_2 distances. All runs were carried out on a Fujitsu Primergy RX300 server with 3.33GHz Intel Xeon X5680 CPUs and 144 GB of RAM, under a GNU/Linux operating system (Suse 11.4), without exploitation of parallelism capabilities (these continuous LP problems can also be solved in a much smaller laptop or desktop PC).

Tables 2 and 3 report some statistics about the vector of absolute relative deviations (i.e., $|z_i|/a_i, i \in \mathcal{N}$) provided by problems (4) and (6), respectively for all the cells $i \in \mathcal{N}$ and for nonsensitive cells $i \in \mathcal{N} \setminus \mathcal{S}$. This distinction is made to avoid the possible bias introduced by sensitive cells, which are by its nature always perturbed. Separate results for sensitive cells are not provided to avoid an excessive length of the document. Columns "mean" show the mean relative deviation. Columns "stdev" give the standard deviation of the vector of relative deviations. Columns "max" show the maximum relative deviation. Columns "#large" report the number of large relative deviations, computed as the number of cells with a relative deviation greater than a certain threshold value; the threshold value considered was one fourth of the maximum relative deviation obtained with L_1. Columns L_0 show the L_0 norm of the vector of relative deviations (i.e., the number of nonzero deviations, or, equivalently, the number of cells that changed their value).

From columns "mean" and "stdev" of tables 2–3 we clearly see that L_1 provides smaller means while L_2 provides smaller standard deviations of the relative adjustments. This is consistent with the behaviour of the linear and quadratic objectives of the optimization problems (4) and (6): L_2 usually adds small changes to a larger number of cells but the values of deviations are more concentrated (large values are avoided, as seen below). The rest of information in tables 1–3 is partly summarized in figures 1–4. Figure 1 shows the difference of the CPU time needed by L_1-CTA and L_2-CTA. We observe that problem (4) always required more CPU time than (6), and in the most difficult instance "hier13d4" about 764 more CPU seconds (809 vs 45 seconds). If efficiency was instrumental, e.g., for the (real-time) protection of large tables in on-line servers, L_2 may be more appropriate than L_1.

Figure 2 shows the difference of the maximum relative deviations provided by L_1-CTA and L_2-CTA. Although this is not guaranteed in theory, in all these executions the difference was positive (i.e., the maximum deviations reported by L_1-CTA were greater than those of L_2-CTA). Looking at the plot of Figure 2 we also observe that the difference increases with the number of cells of the table. It is worth to note that for instance "australia_ABS" the maximum deviations are significantly larger than for the other instances. This can be explained because this (likely frequency) table has some few large sensitive cells (and accordingly,

Table 2. Results for the relative deviations of all the cells

	L_1					L_2				
instance	mean	stdev	max	#large	L_0	mean	stdev	max	#large	L_0
australia_ABS	1.64	17.55	1305.77	12	1134	2.37	12	364.02	1	6608
bts4	0.74	1.97	11.11	3000	35834	0.78	1.93	11.11	2795	31955
cbs	10.45	19.22	100	2690	2762	10.5	19.19	100	2681	2875
dale	16.87	30.19	100	4021	5284	16.94	30.15	100	4019	14931
destatis	1	4.25	50	95	2380	1.02	4.24	50	95	3841
five20b	1.38	2.28	17.38	4428	34478	1.44	2.18	10	4283	34551
five20c	1.53	2.36	14.77	5247	34473	1.59	2.27	10.85	4996	34500
hier13d4	1.38	2.24	9.98	3541	13018	1.45	2.12	9.98	3155	18968
hier13	0.81	1.72	9.97	194	1523	0.84	1.67	9.97	176	2020
hier13x13x13a	0.72	1.67	9.97	190	1454	0.75	1.63	9.97	176	2020
hier13x13x13b	0.72	1.67	9.97	190	1454	0.75	1.63	9.97	176	2020
hier13x13x13c	0.72	1.67	9.97	190	1454	0.75	1.63	9.97	176	2020
hier13x13x13d	1.44	3.33	19.94	190	1376	1.5	3.25	19.94	176	2020
hier13x13x13e	1.44	3.33	19.94	190	1289	1.5	3.25	19.94	176	2020
hier13x13x7d	0.72	1.78	9.97	109	538	0.75	1.75	9.97	102	1040
hier13x7x7d	0.73	1.88	9.97	63	236	0.77	1.86	9.97	58	519
hier16	0.83	1.84	10	309	2715	0.87	1.8	10	289	3564
hier16x16x16a	0.7	1.74	10	309	3009	0.74	1.71	10	289	3564
hier16x16x16b	0.7	1.74	10	309	3009	0.74	1.71	10	289	3564
hier16x16x16c	0.7	1.74	10	309	3009	0.74	1.71	10	289	3564
hier16x16x16d	1.4	3.48	20	309	2675	1.47	3.42	20	289	3564
hier16x16x16e	1.4	3.48	20	309	2675	1.47	3.42	20	289	3564
nine12	1.35	2.34	12.55	1561	8013	1.43	2.23	10	1452	10398
nine5d	1.67	2.69	10	2497	6336	1.78	2.53	10	2061	10732
ninenew	1.56	2.47	16.16	1056	4406	1.64	2.35	10.3	1020	6545
osorio	0.03	1.15	100	2	16	0.05	1.09	100	2	9997
table1	0.57	1.53	5.02	192	849	0.59	1.52	5	184	962
table3	1.43	3.49	66.19	1	1860	1.45	3.32	11.93	0	2397
table4	1.43	3.49	66.19	1	1860	1.45	3.32	11.93	0	2397
table5	1.43	3.49	66.19	1	1860	1.45	3.32	11.93	0	2397
table6	0.57	1.53	8.39	177	891	0.59	1.52	5	173	962
table7	4.56	25.2	160	20	113	4.63	25.18	160	19	484
table8	0.03	0.52	11.11	3	10	0.03	0.52	11.11	3	1200
targus	2.88	9.32	33.4	14	61	2.89	9.32	33.4	14	115
toy3dsarah	0.62	1.38	4	484	727	0.62	1.37	4	482	727
two5in6	1.46	2.49	10	1071	4451	1.56	2.37	10	957	5680

Table 3. Results for the relative deviations of nonsensitive cells

	L_1					L_2				
instance	mean	stdev	max	#large	L_0	mean	stdev	max	#large	L_0
australia_ABS	0.8	12.59	809.94	24	218	1.42	9.2	267.5	4	5692
bts4	0.31	0.83	11.03	758	33574	0.35	0.75	11.03	546	29695
cbs	0.98	6.01	54.55	231	295	1.04	5.97	54.55	231	408
dale	0.43	4.16	83.33	90	361	0.53	4.13	83.33	88	10008
destatis	0.13	0.85	38.89	3	1759	0.15	0.84	38.89	3	3220
five20b	0.74	1.15	13.71	1228	30816	0.81	0.99	9.91	936	30889
five20c	0.84	1.23	10.5	2103	30451	0.91	1.06	9.89	1638	30478
hier13d4	0.74	1.23	8.68	1667	10830	0.82	1	8.68	1125	16780
hier13	0.49	1.02	8.28	103	1411	0.52	0.95	8.28	80	1908
hier13x13x13a	0.44	1.03	9.36	89	1346	0.47	0.97	9.36	71	1912
hier13x13x13b	0.44	1.03	9.36	89	1346	0.47	0.97	9.36	71	1912
hier13x13x13c	0.44	1.03	9.36	89	1346	0.47	0.97	9.36	71	1912
hier13x13x13d	0.87	2.06	18.72	89	1268	0.94	1.95	18.72	71	1912
hier13x13x13e	0.85	1.97	17.27	92	1177	0.91	1.85	16.56	73	1908
hier13x13x7d	0.35	0.93	8.28	43	463	0.38	0.88	8.28	40	965
hier13x7x7d	0.26	0.74	6.78	21	186	0.29	0.69	6.78	17	469
hier16	0.43	0.78	7.59	138	2491	0.47	0.7	7.59	101	3340
hier16x16x16a	0.34	0.74	7.59	141	2785	0.38	0.67	7.59	105	3340
hier16x16x16b	0.34	0.74	7.59	141	2785	0.38	0.67	7.59	105	3340
hier16x16x16c	0.34	0.74	7.59	141	2785	0.38	0.67	7.59	105	3340
hier16x16x16d	0.69	1.47	15.18	141	2451	0.76	1.34	15.18	105	3340
hier16x16x16e	0.69	1.47	15.18	141	2451	0.76	1.34	15.18	105	3340
nine12	0.67	1.15	12.55	389	6835	0.76	0.94	8.95	280	9220
nine5d	0.73	1.31	9.78	873	4675	0.85	0.99	8.79	423	9071
ninenew	0.79	1.29	16.16	210	3548	0.88	1.06	10.3	172	5687
osorio	0.01	0.36	20	7	9	0.03	0.05	0.57	0	9990
table1	0.12	0.62	5.02	46	703	0.14	0.6	5	38	816
table3	0.44	2.01	66.19	1	1343	0.46	1.68	10.86	0	1880
table4	0.44	2.01	66.19	1	1343	0.46	1.68	10.86	0	1880
table5	0.44	2.01	66.19	1	1343	0.46	1.68	10.86	0	1880
table6	0.12	0.62	8.39	31	745	0.14	0.6	5	27	816
table7	0.59	8.44	144	3	96	0.66	8.41	144	2	467
table8	0	0.03	0.81	3	7	0	0.03	0.78	3	1197
targus	0.25	2.74	33.36	1	48	0.26	2.74	33.36	1	102
toy3dsarah	0.14	0.62	4	108	351	0.15	0.61	4	106	351
two5in6	0.69	1.23	9.69	376	3731	0.79	0.99	8.53	248	4960

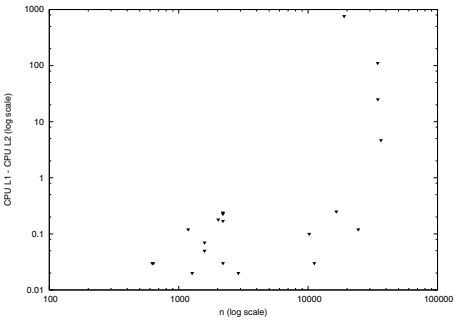

Fig. 1. Difference of the CPU time of L_1 and L_2 (in log scale) vs number of cells (in log scale)

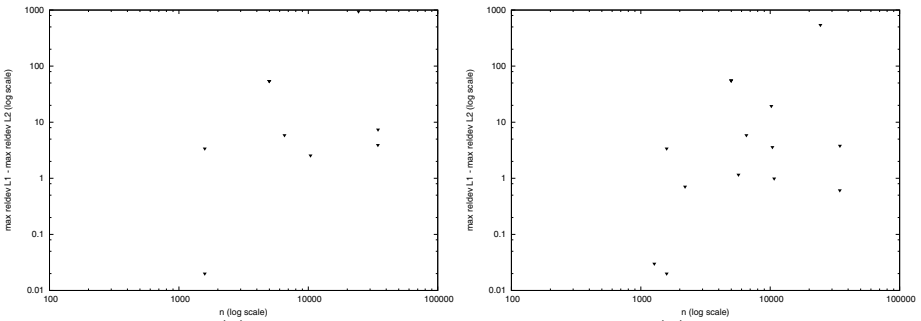

Fig. 2. Difference of the maximum relative deviations provided by L_1 and L_2 (in log scale) vs number of cells (in log scale), for (a) all the cells, and (b) nonsensitive cells

with large protection levels), much larger than the other sensitive and nonsensitive cells. Therefore, it may be stated than, specially in frequency tables, when some relatively small cells must compensate the necessary perturbations of some large sensitive cells, we can expect large maximum relative deviations.

Figure 3 plots the difference of the number of cells with large relative deviations between L_1-CTA and L_2-CTA, where "large" mean deviations greater than one fourth of the maximum relative deviation obtained with L_1. This threshold value depends on whether sensitive cells are considered (Figure 3.(a)) or not (Figure 3.(b)). From this figure we clearly see that the number of cells with large deviations was higher for L_1-CTA than for L_2-CTA (in the extreme cases, around 500 cells). If a smaller number of cells with large relative deviations can be seen as a measure of "data utility", L_2-CTA provides better results.

Finally, Figure 4 shows the difference of the L_0 norms of the vector of relative deviations provided by L_2-CTA and L_1-CTA, i.e, the difference in the number of perturbed cells. We remark that both plots (a) and (b) in this Figure are the same, since sensitive cells have always nonnegative deviations. We observe that in most instances L_2 perturbs more cells than L_1 (much more in some

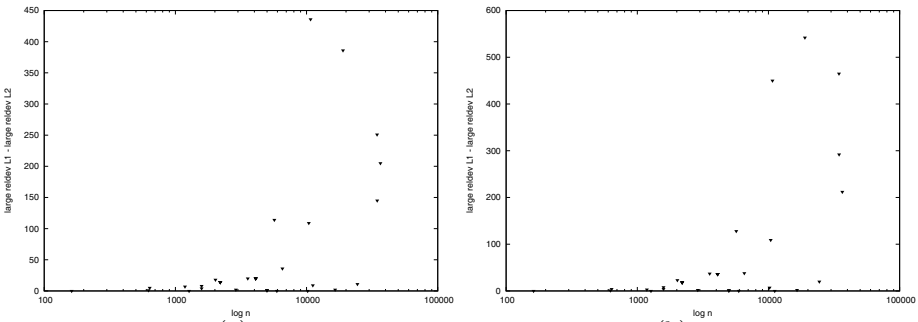

Fig. 3. Difference of the number of large relative deviations provided by L_1 and L_2 vs number of cells (in log scale), for (a) all the cells, and (b) nonsensitive cells

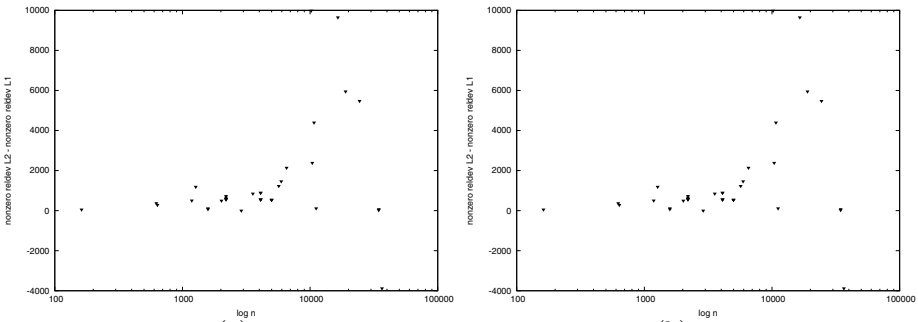

Fig. 4. Difference of L_0 norms of the relative deviations provided by L_2 and L_1 (i.e., difference of the number of nonzero relative deviations) vs number of cells (in log scale), for (a) all the cells, and (b) nonsensitive cells

instances). This is consistent with theory, since L_1-CTA solutions are basic solutions (or vertices of the feasible polyhedron, with many zero components), while the quadratic optimization problem derived from L_2 tries to evenly distribute deviations among all the components. Indeed, the solution with the minimum number of perturbed cells would be provided by formulation (8) using the L_0 norm, and L_1 is "closer" to L_0 than L_2. It is worth to mention that in some situations this is not a main drawback for L_2-CTA, since CTA can be used in practice as a second stage after the introduction of stochastic noise, such that original cell values are anyway modified [9].

5 Conclusions

From the computational results of this work comparing the continuous formulations of L_1-CTA (LP) and L_2-CTA (QP) we conclude that both approaches have their merits and drawbacks. If we focus on efficiency, L_2-CTA requires less

CPU time. If we focus on the relative adjustments provided by both models we observe that: (i) L_1-CTA provides in general smaller means but larger standard deviations of relative adjustments than L_2-CTA; (ii) L_2-CTA provided for all the instances tested smaller maximum relative deviations; (iii) L_2-CTA provided a smaller number of cells with large relative deviations (which can be associated to a measure of data utility); (iv) L_1-CTA provided a much larger number of cells without deviations, since it is a better approximation to L_0 than L_2. If preserving the original values in as many cells as possible is an objective, then L_1-CTA should be chosen. If we look for efficiency and a smaller number of cells with large deviations, then L_2-CTA could be used. The best option is likely having implementations of both models at hand, and, depending on the particular instances or goals, either use one or the other.

References

1. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
2. Castro, J.: Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research 216, 257–269 (2012)
3. Castro, J., Giessing, S.: Testing variants of minimum distance controlled tabular adjustment. In: Monographs of Official Statistics, pp. 333–343. Eurostat-Office for Official Publications of the European Communities, Luxembourg (2006)
4. Castro, J., González, J.A.: A Tool for Analyzing and Fixing Infeasible RCTA Instances. In: Domingo-Ferrer, J., Magkos, E. (eds.) PSD 2010. LNCS, vol. 6344, pp. 17–28. Springer, Heidelberg (2010)
5. Castro, J., González, J.A.: Present and future research on controlled tabular adjustment. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (2011), http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/48_Castro-Gonzalez.pdf
6. Chinneck, J.W.: Feasibility and Infeasibility in Optimization: Algorithms and Computational Methods. Springer (2008)
7. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. (2002)
8. Donoho, D.L., Tsaig, Y.: Fast solution of l_1-norm minimization problems when the solution be sparse. IEEE Transactions on Information Theory 54, 4789–4812 (2008)
9. Giessing, S.: Personal communication in the scope of the "DwB. Data without Boundaries" Project INFRA-2010-262608, VII Mark Program of the European Union (2012)
10. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte-Nordholt, E., Seri, G., de Wolf, P.P.: Handbook on Statistical Disclosure Control (v. 1.2), Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (2010), http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
11. IBM ILOG CPLEX: CPLEX 12.4 User's Manual, IBM (2012)
12. Wright, S.J.: Primal-Dual Interior-Point Methods. SIAM, Philadelphia (1997)