Assessing the Information Loss of Controlled Adjustment Methods in Two-Way Tables*

Jordi Castro** and José A. González

Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
{jordi.castro,jose.a.gonzalez}@upc.edu

Abstract. Minimum distance controlled tabular adjustment (CTA) is a perturbative technique of statistical disclosure control for tabular data. Given a table to be protected, CTA looks for the closest safe table by solving an optimization problem using some particular distance in the objective function. CTA has shown to exhibit a low disclosure risk. The purpose of this work is to show that CTA also provides a low information loss, focusing on two-way tables. Computational results on a set of midsize tables validate this statement.

Keywords: Statistical disclosure control, controlled tabular adjustment, information loss, data utility, mixed integer linear programming.

1 Introduction

Minimum-distance controlled tabular adjustment (CTA in short) was suggested in [2,12] as a post-tabular perturbation approach for statistical disclosure control. A description of the state-of-the-art in the statistical disclosure field can be found in the monograph [20] and the survey [3]. Briefly, given a table with sensitive information, the goal of CTA is to compute the closest safe table through the solution of an optimization problem using some particular distance in its objective function. CTA is being considered an emerging technology for tabular data protection [20]. CTA can be applied to both frequency and magnitude tables (i.e., tables providing, respectively, either cell counts or aggregated information for another variable). This work only considers frequency tables, i.e., cell values are integer. For two-way tables CTA will always provide integral values, such that integrality constraints are not needed, and the two information loss measures used in this paper (one of them requiring integrality of cell values) can be applied.

* Supported by grants MTM2012-31440 of the Spanish Ministry of Economy and Competitiveness, SGR-2014-542 of the Government of Catalonia, and DwB INFRA-2010-262608 of the FP7 European Union Program.

** Corresponding author.

J. Domingo-Ferrer (Ed.): PSD 2014, LNCS 8744, pp. 11–23, 2014.

Several recent papers have been devoted to CTA. Some of them focused on the solution of the optimization problem formulated [5,7,16], whereas others dealt with quality and confidentiality issues of the computed solution [6,10].

A tabular data protection method can be seen as a map F such that $F(T) = T'$, i.e., table T is transformed to another table T'. Two are the main requirements for F: (1) the output table T' should be "safe", and (2) the information loss should be small, i.e., T' should be a good replacement for T. The disclosure risk can be analyzed through the inverse map $T = F^{-1}(T')$: if not available or difficult to compute by any attacker, then we may guarantee that F is safe. It was empirically observed in [4] that estimates $\hat{T} = \hat{F}^{-1}(T')$, $\hat{F}^{-1}$ being an estimate of F^{-1} for CTA, were not close to T for some real tables, concluding that CTA was a safe method for these tables. However, a similar analysis regarding the utility of T' has not been performed for CTA (though it was for some microdata methods, as reported in [13]). Other methods (random record swapping and semi-controlled random rounding) have been compared using a table from the 2001 UK Census in [24]. The purpose of this work is then to fill this gap by performing a computational analysis on the data utility of two-way tables protected with CTA. The same procedure may be extended to multidimensional, hierarchical or linked tables but, due to its higher complexity, is out of the scope of this work and part of the further research to be done in this field.

The paper is organized as follows. Section 2 reviews the CTA formulation used in this work. Section 3 shows the methodology developed for analyzing the information loss. Finally, Section 4 reports computational results with some midsize two-way tables.

2 The CTA Formulation

Given (i) a set of cells $a_i, i = 1, \ldots, n$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's); (ii) a lower and upper bound for each cell $i = 1, \ldots, n$, respectively l_{a_i} and u_{a_i}, which are considered to be known by any attacker; (iii) positive cell weights $w_i, i = 1, \ldots, n$, associated to the cost of perturbing cell values; (iv) a set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \{1, \ldots, n\}$ of indices of sensitive cells; (v) and a lower and upper protection level for each sensitive cell $i \in \mathcal{S}$, respectively lpl_i and upl_i, such that the released values musty satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$; the goal of CTA is to find the closest safe values $x_i, i = 1, \ldots, n$, according to some distance ℓ, that makes the released table safe. This is achieved by the solution of the following optimization problem:

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

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z_i = x_i - a_i, \quad i = 1, \ldots, n$ —and similarly $l_{z_i} = l_{x_i} - a_i$ and $u_{z_i} = u_{x_i} - a_i$—, (1) can be recast as

$$\begin{aligned} \min_{z} \ & ||z||_\ell \\ \text{s. to } & Az = 0 \\ & l_{z_i} \le z_i \le u_{z_i} \quad i = 1, \ldots, n \\ & z_i \le -lpl_i \text{ or } z_i \ge upl_i \quad i \in \mathcal{S}, \end{aligned} \tag{2}$$

$z \in \mathbb{R}^n$ being the vector of deviations. Using the ℓ_1 distance, considering the splitting $z = z^+ - z^-$, and after some manipulation, (2) can be written as

$$\begin{aligned} \min_{z^+, z^-, y} \ & \sum_{i=1}^{n} w_i (z_i^+ + z_i^-) \\ \text{s. to } & A(z^+ - z^-) = 0 \\ & 0 \le z_i^+ \le \quad u_{z_i} \quad i \notin \mathcal{S} \\ & 0 \le z_i^- \le -l_{z_i} \quad i \notin \mathcal{S} \\ & upl_i \ y_i \le z_i^+ \le u_{zi} \ y_i \quad i \in \mathcal{S} \\ & lpl_i (1 - y_i) \le z_i^- \le -l_{zi} (1 - y_i) \quad i \in \mathcal{S} \\ & y_i \in \{0, 1\} \quad i \in \mathcal{S}, \end{aligned} \tag{3}$$

$w \in \mathbb{R}^n$ being the vector of positive cell weights, $z^+ \in \mathbb{R}^n$ and $z^- \in \mathbb{R}^n$ the vector of positive and negative deviations in absolute value, and $y \in \mathbb{R}^s$ being the vector of binary variables associated to protections directions. When $y_i = 1$ the constraints mean $upl_i \le z_i^+ \le u_{zi}$ and $z_i^- = 0$, thus the protection direction is "upper"; when $y_i = 0$ we get $z_i^+ = 0$ and $lpl_i \le z_i^- \le -l_{z_i}$, thus protection direction is "lower". Model (3) is a (in general difficult) mixed integer linear optimization problem, but it may provide better quality solutions than other CTA variants without binary variables (e.g., [8,9]). In this work tables have been protected by solving (3) by the CTA package [17] recently improved within the Data without Boundaries INFRA-2010-262608 FP7 project.

3 Assessment of Information Loss

In [13] the information loss was measured by comparing several statistics on the original and protected microdata. We followed a similar approach, but restricting the analysis to a few available statistics for two-way tables to measure the association between the row and column variables. A simple statistic as the correlation between the values of the cells of the original and perturbed table a_i and $a_i + z_i$, $i = 1, \ldots, n$, is avoided, since it is meaningless: in practice it is almost 1 and it does not capture the relationship between the row and column categories. The assessment methodology is outlined in next subsections.

3.1 Generation of Tables

The analysis was restricted to two-way tables, which were randomly generated by the following algorithm:

- Input: r, number of categories for row variable (rows of the table); c, number of categories for column variable (columns of the table); N: total number of

observations or respondents; ρ: correlation between both variables (a number in $[-1, 1]$).
- Output: a contingency table of dimensions $r \times c$; table margins may also be provided.
- Step 1. We obtain a binormal random sample of N points, say $(x_i, y_i), i = 1 \dots N$, with zero mean and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

- Step 2. The variables are discretized into r and c categories, respectively. The cutpoints are randomly chosen so that very small frequencies are not possible; to be precise, at least 10 observations are required in the marginal cell of each row and column (though internal cells may be below 10).
- Step 3. A two-way table is created by cross-tabulation of both discretized variables. If required, a margin row and a margin column are created, as well as a grand-total cell (equal to N).

The software package used to produce the tables, obtain the measures described below and analyze the results was R, release 2.15 [23]. In order to get two samples with the given correlation and normal distribution we used the function `rmvnorm` from the R package `'mvtnorm'` [14,15].

3.2 Measures

Contingency tables summarize the information coming from cross-tabulation of two or more categorical variables, and there are several analytical ways to represent them through numerical estimators. Although single measures are usually too simple to catch the dependence structure underlying the variables—especially in high dimensional tables—we have chosen a few of them to allow the comparison between the original and protected tables.

Some of the most used measures of association are based on the well-known Pearson's coefficient

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i},$$

where n is the number of cells in the table, o_i means an observed frequency, and e_i an expected frequency, normally under independence of the variables. The Pearson's chi-squared test is based on the assumption that it follows a χ^2 probability distribution, with known number of degrees of freedom ($(r-1)(c-1)$ in two-way tables), depending on some conditions and whenever the variables are independent.

We considered the coefficient known as Cramér's V [11], computed as

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}.$$

Cramér's V ranges from 0 (in case of no association between the variables) to 1 (maximum association), being only 1 when the variables are identical. Cramér's V is invariable to changes in the order of the categories of the variables. This measure was computed using the function `assocstats` from the R package `'vcd'` [21]. Cramér's V was one of the measures employed in [24].

The second technique considered in this work to explore the relationships between the two variables of the table is correspondence analysis (CA). CA is frequently employed as an exploratory tool, with the aim to identify more detailed ways of association between the variables, instead of a single measure of the strength of such a relationship. For our purposes, we used the variant for two-way contingency tables named Simple Correspondence Analysis (SCA).

SCA reduces the high dimensionality of the original data (given by the number of categories of our variables) to a low-dimensional space which retains as much information as possible. Briefly, SCA involves the generalized singular value decomposition [18] of a matrix M computed as follows. Denoting by T the matrix containing the $r \times c$ entries of the two-way contingency table, by e_t the column vector of 1's of dimension t, and by $\text{diag}(v)$ a diagonal matrix containing the elements of vector v in its diagonal positions, M is computed as

$$M = R - e_r c^\top \quad \text{where} \quad R = \text{diag}(Te_c)^{-1}T \quad \text{and} \quad c = (e_r^\top Te_c)^{-1}(e_r^\top T).$$

Denoting

$$W_r = \text{diag}(e_r^\top Te_c)^{-1}(Te_c) \quad \text{and} \quad W_c = \text{diag}(c)^{-1}$$

then M is decomposed by the generalized singular value decomposition as

$$M = U\Sigma V \quad \text{where} \quad U^\top W_r U = I_r \quad \text{and} \quad V^\top W_c V = I_c,$$

where I_t is the $t \times t$ identity matrix, U and V contain the row and column singular vectors, and $\Sigma \in \mathbb{R}^{r \times c}$ contains l nonzero singular values (where $l \leq \min(r, c)$) in its diagonal entries (see, for instance, [19] for a comprehensive description). The rows of the two-way table can be projected onto the singular vectors U, obtaining the factor scores. The variance of the factor scores for a given dimension is equal to the squared singular value of this dimension. The squared singular values of M are equal to the eigenvalues of $MM^\top$ [18]. It is worth to remind that the concept of *inertia* is equal to the χ^2 statistic divided by N, that the sum of all the eigenvalues of $MM^\top$, $\sum_{i=1}^{l} \lambda_i$, is equal to the inertia, and that a few dimensions (or directions, or eigenvectors) related to the largest eigenvalues may explain most of the information in the table.

In this work we focus on the larger eigenvalue (λ_1) from the SCA, and the contribution of λ_1 among all the eigenvalues, i.e., the ratio $\lambda_1 / \sum_{i=1}^{l} \lambda_i$ between λ_1 and the inertia as a percentage, denoted as π_1. The relation between V and the contribution of λ_1 is not straightforward, and much less between V and π_1. The singular values were computed with the function `ca`, from the R package of the same name [22].

3.3 Description of the Experiments

Two experiments have been designed. They are independent since different tables have been considered for them. Alternatively, the same tables could have been used in both experiments, but we decided to consider two different sets. The procedure is similar in both cases:

- Set the parameters of the instances: Percentage of sensitive cells in the tables, and correlation ρ; other factors have been fixed; 15 instances will be generated for each combination of percentage of sensitive cells and ρ.
- For each instance:
  - generate a table with random r, c and N;
  - compute the measures from the original table;
  - write the table in a format allowed by the CTA package;
  - run the CTA package, and write the protected table;
  - computed the measures from the protected table;
  - save the results;
- Read the results file, and compare the outcomes.

The optimality gap is a bound for the maximum relative difference allowed between the computed and the optimal solutions. The value considered for all the executions, 2.5%, was chosen after some exploration with different values. It became apparent that the CTA procedure was robust (i.e., there were no large deviations between the original and the protected tables) even with large gaps such as 50%. However, the number of sensitive cells protected upwards was significantly higher with those larger gaps, while smaller gaps produced tables with a good balance among the protection directions of their sensitive cells (i.e., the number of sensitive cells upper and lower protected was similar, which reduces the disclosure risk against an attacker). On the other hand, very small gaps may result in large CTA executions for the solution of (3). We set a limit time of 300 seconds for all the executions, which was enough for most of the cases. In particular, CTA took more than one minute in 78 tables (3.42% of the overall 2280 tables protected—720 tables for the first experiment with Cramér's V, and 1560 tables for the second experiment with SCA), and 25 (1.1% of tables) reached the maximum limit of five minutes. Median time to solution was 0.22 seconds.

Sensitive cells were chosen at random, and protection levels were 10% of the cell value, rounded to the nearest integer. The dimensions of the table were taken at random between 10 and 40. The table margins were included as cells for convenience, but we don't allow them to differ from the original value. The total number of observations N is dependent of r and c, so larger tables usually have more observations. The percentage of zero cells in the generated tables is approximately 5%; the percentage of cells with one respondent is also 5%. By construction a complete row or column cannot be empty. Zero cells are preserved in the protected table.

Table 1. Summary of dimensions and V for generated tables

	median	min	max
cells	600	121	1640
sensitive cells	57	2	309
N	29640	7050	94831
original V	0.0666	0.0181	0.1825

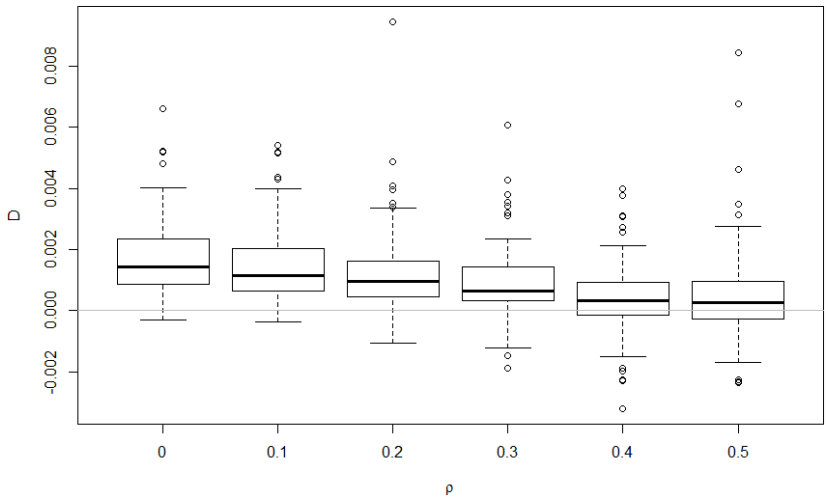

Fig. 1. Boxplots of D for different ρ values

4 Computational Results

4.1 Cramér's V

We generated 720 tables, with a percentage of sensitive cells between 5% and 19%, and values of $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. A summary of the dimensions of these tables and their V values is reported in Table 1.

The median of V was 0.0680 in the protected tables, ranging between 0.0197 and 0.1839. Since V is highly correlated with ρ and moderately related to N, we studied the difference $D = V_{prot} - V_{orig}$. Relative differences were discarded because the original quantities can be close to zero, especially for uncorrelated variables, and V ranges from 0 to 1, thus absolute differences can be easily interpreted. Figure 1 reports boxplots of D for different ρ values, showing that D increases when ρ is close to 0. The change is small in magnitude, compared with its variability, as shown in Table 2.

The tables with larger deviations in the Cramér's V measure are small tables (300 cells in average) with a high percentage of sensitive cells (16%). The most significant factors by a general linear model for D are: the correlation ρ (coefficient $-3 \cdot 10^{-3}$), the percentage of sensitive cells (coefficient $8.3 \cdot 10^{-5}$), and the

number of cells (coefficient $8 \cdot 10^{-7}$). However, these factors explain only 25.8% of the total variability observed in D.

4.2 Simple Correspondence Analysis

For this second experiment we generated 1560 tables, using values of $\rho \in \{-0.6, -0.5, \ldots, 0, \ldots, 0.5, 0.6\}$. Unlike for the Cramér's V, we considered negative correlations for if they might influence the results. For each original and protected table the measures λ_1 and π_1 were computed. Table 3 shows a summary of collected values.

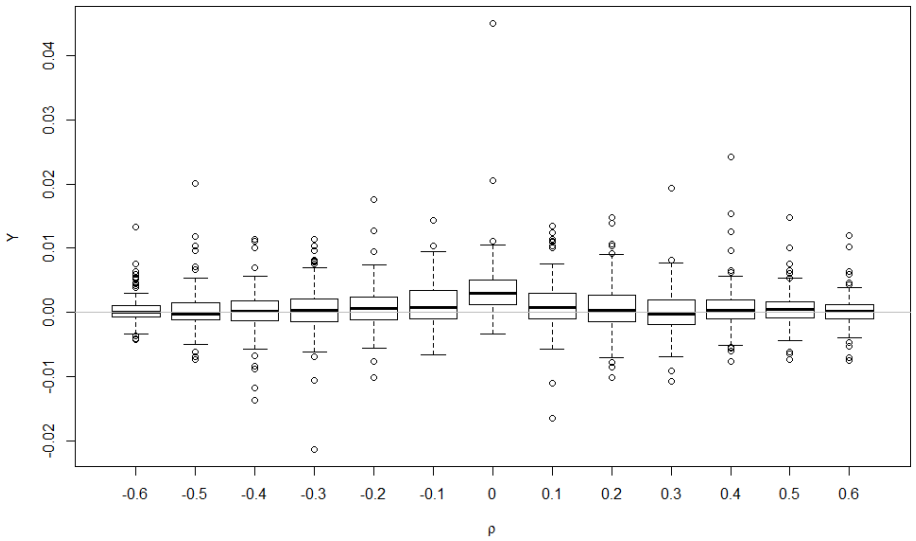

Fig. 2. Boxplots of Y for different ρ values

We studied the singular value $\sqrt{\lambda_1}$ instead of the eigenvalue because it appeared to be proportional to $|\rho|$ and it showed a greater stability in variance. As before, the effect observed after the protection performed by CTA is studied through the change $Y = \sqrt{\lambda_{1,prot}} - \sqrt{\lambda_{1,orig}}$. As for D, Y is defined as an absolute difference since the original eigenvalues $\lambda_{1,orig}$ are close to zero, especially with null ρ (relative differences are used below in Table 4). Figure 2 shows boxplots of Y for the different ρ values. The outlier at $\rho = 0$ appearing on top of

Table 2. Mean and standard deviation of D with respect to ρ

ρ	0	0.1	0.2	0.3	0.4	0.5
mean	0.0017	0.0015	0.0012	0.0009	0.0004	0.0004
std. dev.	0.0012	0.0012	0.0013	0.0011	0.0012	0.0014

Table 3. Summary of λ_1 and π_1 for original and protected tables

	median	min	max
orig. λ_1	0.0875	0.0011	0.3586
prot. λ_1	0.0881	0.0014	0.3587
orig. π_1	76.33%	9.5%	92.6%
prot. π_1	75.45%	9.6%	90.9%

Table 4. Bounds for the intervals containing 90% of relative changes in the singular value, expressed as $Y/\sqrt{\lambda_{1,orig}} \cdot 100$

$\lvert\rho\rvert$	0	0.1	0.2	0.3	0.4	0.5	0.6
Lower (%)	0.27	-4.30	-2.47	-1.33	-1.27	-0.86	-0.56
Upper (%)	24.23	7.64	3.86	2.20	1.43	1.27	0.80

the figure was produced by a table of 392 cells, whose eigenvalues λ_1 before and after protection were 0.001074 and 0.006043, respectively; it was the table with the smallest λ_1. Aside from this outlier, it can be seen that in general changes due to the protection were small.

Table 4 shows the intervals which include 90% of the relative changes, expressed as $Y/\sqrt{\lambda_{1,orig}} \cdot 100$, in the singular values, depending on ρ. The sign of ρ is not important for the analysis, so we considered only its absolute value. It is shown that relevant changes only appear for $\rho = 0$ (as large as, e.g. 25%). Indeed, the 95% confidence interval for the mean of Y when $\rho = 0$ was $(0.0032, 0.0046)$. For nonzero correlations there is no evidence of change. Moreover, from Table 4 it is clear that relative changes in $\sqrt{\lambda_1}$ are a decreasing function of $|\rho|$.

As for the percentage π_1 explained by the first dimension, we observed: a) a symmetrical pattern with respect to $\rho = 0$, b) small values of π_1 for $\rho = 0$ (about 16%), quickly increasing with $|\rho|$ until approximately $|\rho| = 0.4$ (about 83%), and decreasing slowly beyond that point, both before and after the table protection. Figure 3 shows the ratio $Z = \pi_{1,prot}/\pi_{1,orig}$. The outlier at $\rho = 0$ appearing on top of the Figure 2 was not drawn in Figure 3, since it modified π_1 from 18.6% to 53.4% ($Z \approx 3$ is out of the range of the vertical axis of Figure 3).

Changes in π_1 can be analyzed through Figure 3 and Table 5, which report the intervals with 90% of observed Z for different ρ. In general, the π_1 of the protected table tends to decrease for small $|\rho|$ values, though the trend in uncorrelated factors points to an increase; for large $|\rho|$ the change in π_1 can be negligible.

Table 5. Lower and upper bounds for the intervals containing 90% of the ratios Z

$\lvert\rho\rvert$	0	0.1	0.2	0.3	0.4	0.5	0.6
Lower	0.936	0.885	0.945	0.967	0.972	0.970	0.969
Upper	1.180	1.031	1.001	0.998	1.004	1.013	1.018

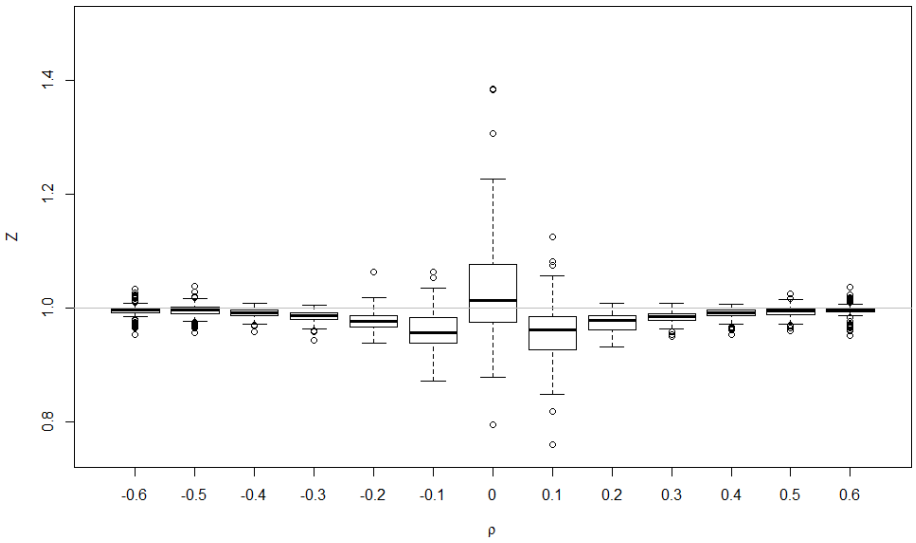

Fig. 3. Boxplots of Z for different ρ values

CTA provided solutions with well balanced sensitive cells with respect to the direction of the deviation. The percentage of sensitive cells protected upwards lied between 41.7% and 58.8% in 90% of the tables, which makes the procedure unpredictable, thus safer.

5 Conclusions

Through the measures considered in this work, we may conclude that a two-way table protected with CTA experiment a slight information loss. It was observed that only the tables from independent factors could suffer significant alteration in Cramér's V or in indicators related to SCA. For V, we have found that the chance of change is higher in small tables or tables with a high percentage of sensitive cells. Anyway, in absolute numbers V barely changed: an average increase of 0.0017 if uncorrelated factors were present.

With respect to SCA, relative changes in λ_1 were significant only when $\rho = 0$. However, we have found that the absolute change in $\sqrt{\lambda_1}$ is usually insignificant: while the average first singular value is 0.052, it increases at most (in 95% of cases) by 0.0093. Differences tend to increase for π_1: for $\rho = 0$ the variation of π_1 can be large, normally above the original value; for $\rho \neq 0$ the variation of π_1 is lesser though usually below the original value. It should be kept in mind that, even when $\rho = 0$, the absolute changes in π_1 were small: only two tables —1 out of 1000— modified by more than 10% the original π_1 value.

There is not a conclusive explanation of why the greatest information loss occurred for $\rho = 0$. One possible reason could be that, since for $\rho = 0$ cells

values are evenly scattered through the table, the number of additional cells with deviations (aside from the sensitive ones) increases; whereas in two-way tables from correlated variables it might be easier to compensate deviations due to protection levels just using sensitive cells. However, a deeper analysis is part of the additional work to be done.

A more exhaustive study considering also real-world tables is needed, and part of the further work to be done. Some preliminary results with two standard two-way tables used in the literature (named "table8" and "dale") confirm that changes in measures increase with the size of the table and the percentage of sensitive cells. For instance, for the 40×30 "table8" instance with only 3 sensitive cells, the V statistic was almost the same before and after protection (0.09270563 vs 0.09280493). On the other hand, for the 358×45 "dale" instance with a 30% of sensitive cells the change in V was significant: from 0.0692391 to 0.1093475. However, for "dale", the information loss was small according to the other measure: $\lambda_{1,orig} = 0.09809$ and $\lambda_{1,prot} = 0.10264$.

Alternative measures could have been applied. One of them would be hypothesis testing on the independence of the two variables using Pearson's χ^2 test. However, even for original independent tables, it is likely that the null hypothesis is rejected for CTA-protected tables, since sensitive cells are forced to be "significantly" perturbed, and this perturbation affects quadratically to the Pearson's χ^2 statistic. This effect may increase with the percentage of sensitive cells. Some preliminary tests with synthetic independent tables confirmed this assertion. Anyway, hypothesis testing might not be a suitable measure in this context: data may not come from random sampling and, furthermore, there is considerable debate around the hypothesis testing nature and the use of p-values [1].

In summary, it can be concluded that the data utility of the CTA-protected tables used in this work is in general acceptable/high and comparable to that of the original tables. Among the further lines of work we find:

- Extension of the above measures to higher-dimensional, hierarchical and linked real-world tables.
- Extension to magnitude tables, using other information loss measures (e.g., generalized linear models).
- Joint analysis of the data utility and disclosure risk of CTA-protected tables, likely in the form of risk-utility plots.

References

1. Batanero, C.: Controversies around the role of statistical tests in experimental research. Mathematical Thinking and Learning 2, 75–97 (2000)
2. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
3. Castro, J.: Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research 216, 257–269 (2012)

4. Castro, J.: On assessing the disclosure risk of controlled adjustment methods for statistical tabular data. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20, 921–941 (2012)
5. Castro, J., Frangioni, A., Gentile, C.: Perspective reformulations of the CTA problem with L_2 distances. Operations Research (in press, 2014)
6. Castro, J., Giessing, S.: Testing variants of minimum distance controlled tabular adjustment. Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 333–343 (2006)
7. 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)
8. Castro, J., González, J.A.: A fast CTA method without the complicating binary decisions. In: Documents of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Statistics Canada, Ottawa, pp. 1–7 (2013)
9. Castro, J., González, J.A.: A multiobjective LP approach for controlled tabular adjustment in statistical disclosure control. Working paper, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2014)
10. Cox, L.H., Kelly, J.P., Patil, R.: Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 87–98. Springer, Heidelberg (2004)
11. Cramér, H.: Mathematical Methods of Statistics. Princeton University Press, Princeton (1946)
12. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. (2002)
13. Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Proceedings of ETK-NTTS 2001, pp. 807–826. Eurostat, Luxemburg (2001)
14. Genz, A., Bretz, F.: Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, vol. 195. Springer, Heidelberg (2009)
15. Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T.: mvtnorm: Multivariate Normal and t Distributions, R package version 0.9-9999 (2014), `http://CRAN.R-project.org/package=mvtnorm`
16. González, J.A., Castro, J.: A heuristic block coordinate descent approach for controlled tabular adjustment. Computers & Operations Research 38, 1826–1835 (2011)
17. Giessing, S., Hundepool, A., Castro, J.: Rounding methods for protecting EU-aggregates. In: Eurostat Methodologies and Working Papers. Worksession on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 255–264 (2009) ISBN 978-92-79-12055-8.
18. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins Univ. Press, Baltimore (1996)
19. Greenacre, M. J.: Theory and applications of correspondence analysis. Academic Press, New York (1984)
20. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K., de Wolf, P.P.: Statistical Disclosure Control. Wiley, Chichester (2012)

21. Meyer, D., Zeileis, A., Hornik, K.: vcd: Visualizing Categorical Data. R package version 1.3-1 (2013), http://CRAN.R-project.org/package=vcd
22. Nenadic, O., Greenacre, M.: Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20, 1–13 (2007)
23. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2012), http://www.R-project.org/
24. Shlomo, N., Young, C.: Statistical disclosure control methods through a risk-utility framework. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 68–81. Springer, Heidelberg (2006)