HHS Public Access

Author manuscript

Priv Stat Databases. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

Priv Stat Databases. 2020 September ; 12276: 166–179. doi:10.1007/978-3-030-57521-2_12.

On Different Formulations of a Continuous CTA Model

Goran Lesaja[1,2], **Ionut Iacob**[1], **Anna Oganian**[3]

[1]Department of Mathematical Sciences, Georgia Southern University, 65 Georgia Ave., Statesboro, GA 30460-8093, U.S.A.

[2]Department of Mathematics, US Naval Academy, 121 Blake Road, Annapolis, MD 21402-1300

[3]National Center for Health Statistics 3311 Toledo Rd, Hyattsville, MD, 20782, U.S.A.

Abstract

In this paper, we consider a Controlled Tabular Adjustment (CTA) model for statistical disclosure limitation of tabular data. The goal of the CTA model is to find the closest safe (masked) table to the original table that contains sensitive information. The measure of closeness is usually measured using ℓ_1 or ℓ_2 norm. However, in the norm-based CTA model, there is no control of how well the statistical properties of the data in the original table are preserved in the masked table. Hence, we propose a different criterion of "closeness" between the masked and original table which attempts to minimally change certain statistics used in the analysis of the table. The Chi-square statistic is among the most utilized measures for the analysis of data in two-dimensional tables. Hence, we propose a *Chi-square* CTA model which minimizes the objective function that depends on the difference of the Chi-square statistics of the original and masked table. The model is non-linear and non-convex and therefore harder to solve which prompted us to also consider a modification of this model which can be transformed into a linear programming model that can be solved more efficiently. We present numerical results for the two-dimensional table illustrating our novel approach and providing a comparison with norm-based CTA models.

Keywords

Statistical disclosure limitation; controlled tabular adjustment models; linear and non-linear optimization; interior-point methods; chi-square statistic

1 Introduction

Minimum-distance controlled tabular adjustment (CTA) methodology for tabular data was first introduced in [7, 14]. It is one of the effective statistical disclosure limitation (SDL) methods for the protection of sensitive information in tabular data. An overview of SDL theory and methods can be found in the monograph [16], and for tabular data only, in the survey [8].

CTA problem can be formulated as follows: given a table with sensitive cells, compute the "closest" additive safe (masked) table to the original table ensuring that adjusted (masked)

goran@georgiasouthern.edu.

values of all sensitive cells are safely away from their original value and that adjusted values are within a certain range of the real values. The additivity of the masked table means in most cases the requirement that the sum of cell values in each row and column of the table remains the same in the original and masked table, [16, 18].

In the standard formulation of the CTA model, the closeness of the original and masked table is measured by the weighted distance between the tables with respect to a certain norm. Most commonly used norms are ℓ_1 and ℓ_2 norms. Thus, the problem can be formulated as a minimization problem with the objective function being a particular weighted distance function and constraints being derived from the requirements stated above.

In general, the CTA problem is a Mixed Integer Optimization Problem (MICOP) which is a difficult problem to solve especially for large dimension tables. The MICOP CTA problem involves binary variables that characterize sensitive cells and their values indicate whether the value of the sensitive cell is adjusted upward or downward. Apriori fixing the values of binary variables reduces the problem to the continuous optimization problem which is easier to solve, however, the quality of the solution may be reduced since we are no longer searching for the global optimal solution but its approximation. In addition, the values of the binary variables have to be assigned carefully otherwise the problem may become infeasible. Some strategies to fix the binary variables while preserving the feasibility of the problem were discussed in [11, 12]. In the paper we assume that the binary variables are fixed upfront according to one of the strategies presented in these papers, hence we consider continuous CTA models.

As indicated above, the objective function in the continuous CTA model is based on either the ℓ_1-norm or ℓ_2-norm. The formulation of ℓ_2-CTA model leads to the Quadratic Programming (QP) problem, while ℓ_1-CTA model can be formulated as the Linear Programming (LP) problem and, as a Second-Order Cone (SOC) problem, which has recently been proposed in [18].

However, in the standard norm-based CTA model, there is no control of how well the statistical properties of the data in the table are preserved. The numerical experiments summarized in Table 1 in Section 2.2 suggest that there is no pattern which would indicate that one CTA model consistently produces the values of the Chi-square statistic, or other statistics, of the masked and original table that are closer to each other than for any other model.

This observation motivated us to consider different criteria of “closeness” between masked and original table which attempts to minimally change certain statistical properties of the table. For example, the Chi-square statistic is an important statistical measure often used to analyze tabular data [10]. Hence, we propose, what we call *Chi-square* CTA model, which minimizes the objective function that depends on the difference of Chi-square statistics of the original and masked table. The Chi-square CTA model is smooth, non-linear, and non-convex which makes it harder to solve the problem. This motivated us to also consider a modification of this model, called *Chi-linear* CTA which can be transformed into a LP problem that can be solved more efficiently.

The Chi-square and Chi-linear CTA models are applied to the two-dimensional table used previously in the literature [18, 5] as a two-dimensional test table to compare solutions of different CTA models. Chi-square and Chi-linear CTA models for this table are solved using interior-point methods (IPMs) and compared with results obtained in [18] when norm-based CTAs models were applied to the same table. The Chi-square statistic, Cramer V, and Chi-linear measures were calculated for the original table and for masked tables and compared to illustrate the validity of our approach.

The paper is organized as follows. In Section 2 the norm-based CTA models are outlined. In Section 3 a novel continuous Chi-square CTA model is presented, as well as its modification, Chi-linear CTA model, and the transformation to LP problem is derived. Section 4 contains numerical results of applying Chi-square, Chi-linear, and norm-based CTA models to the two-dimensional table. The concluding remarks and possible directions for future research are given in Section 5.

2 Preliminaries

2.1 Norm-based CTA models

In this section, we review the standard norm-based CTA model as it is presented in [18]. Given the following set of parameters:

- **i.** A set of cells a_i, $i \in \mathcal{N} = \{1, \ldots, n\}$. The vector $a = (a_1, \ldots, a_n)^T$ satisfies certain linear system $Aa = b$ where $A \in \mathbb{R}^{m \times n}$ is an $m \times n$ matrix and and $b \in \mathbb{R}^m$ is m-vector.
- **ii.** A lower, and upper bound for each cell, $l_{a_i} \leq a_i \leq u_{a_i}$ for $i \in \mathcal{N}$, which are considered known by any attacker.
- **iii.** A set of indices of sensitive cells, $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N}$
- **iv.** 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 must be outside of the interval $(a_i - lpl_i, a_i + upl_i)$.
- **v.** A set of weights, w_i, $i \in \mathcal{N}$ used in measuring the deviation of the released data values from the original data values.

A standard CTA problem is a problem of finding values z_i, $i \in \mathcal{N}$, to be released, such that z_i, $i \in \mathcal{S}$ are safe values and the weighted distance between released values z_i and original values a_i, denoted as $\|z - a\|_{l(w)}$ is minimized, which leads to solving the following optimization problem

$$\begin{aligned} &\min_{z} \|z - a\|_{l(w)} \\ &s.t.\ \ Az = b, \\ &l_{a_i} \leq z_i \leq u_{a_i},\ \ i \in \mathcal{N}, \\ &z_i, i \in \mathcal{S} \text{ are safe values}. \end{aligned} \tag{1}$$

As indicated in the assumption (iv) above, safe values are the values that satisfy

$$z_i \leq a_i - lpl_i \text{ or } z_i \geq a_i + upl_i,\ i \in \mathcal{S}. \tag{2}$$

By introducing a vector of binary variables $y \in \{0, 1\}^s$ the constraint (2) can be written as

$$\begin{aligned} z_i &\geq -M(1 - y_i) + (a_i + upl_i)y_i,\ i \in \mathcal{S}, \\ z_i &\leq My_i + (a_i - lpl_i)(1 - y_i), \quad i \in \mathcal{S}, \end{aligned} \tag{3}$$

where $M \gg 0$ is a large positive number. Constraints (3) enforce the upper safe value if $y_i = 1$ or the lower safe value if $y_i = 0$.

Replacing the last constraint in the CTA model (1) with (3) leads to a mixed-integer convex optimization problem (MICOP) which is, in general, a difficult problem to solve; however, it provides a globally optimal solution [6]. The alternative approach is to fix binary variables upfront which leads to a CTA model which is a continuous convex optimization problem that is easier to solve. It is worth noting that the obtained solution is optimal for the fixed combination of binary variables which is different from the global optimum obtained when solving MICOP CTA problem, however, in most cases, it is quite a good approximation that serves the purpose of protecting sensitive values in the table quite well. It is also important to mention that a wrong assignment of binary variables may result in the problem being infeasible. Strategies on how to avoid this difficulty are discussed in [11, 12].

In this paper, we consider a continuous CTA model where binary variables are fixed according to one of the strategies suggested in these papers. Furthermore, vector z is replaced by the vector of *cell deviations*, $x = z - a$.

The CTA (1) model with constraints (3) reduces to the following convex optimization problem:

$$\begin{aligned} &\min_x \|x\|_{l(w)} \\ &s.t.\ Ax = 0, \\ &l \leq x \leq u, \end{aligned} \tag{4}$$

where upper and lover bounds for x_i, $i \in \mathcal{N}$ are defined as follows:

$$l_i = \begin{cases} upl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 1 \\ l_{a_i} - a_i & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 0) \end{cases} \tag{5}$$

$$u_i = \begin{cases} -lpl_i & \text{if } i \in \mathcal{S} \text{ and } y_i = 0 \\ u_{a_i} - a_i & \text{if } (i \in \mathcal{N}\backslash\mathcal{S}) \text{ or } (i \in \mathcal{S} \text{ and } y_i = 1) \end{cases} \tag{6}$$

The two most commonly used norms in problem (4) are the ℓ_1 and ℓ_2 norms. For the ℓ_2-norm the problem, (4) reduces to the following ℓ_2-CTA model which is a QP problem:

$$\begin{aligned} &\min_{x} \sum_{i=1}^{n} w_i x_i^2 \\ &s.t.\ \ Ax = 0 \\ &l \le x \le u. \end{aligned} \tag{7}$$

For the ℓ_1-norm the problem, (4) reduces to the following ℓ_1-CTA model:

$$\begin{aligned} &\min_{x} \sum_{i=1}^{n} w_i |x_i| \\ &s.t.\ \ Ax = 0, \\ &l \le x \le u. \end{aligned} \tag{8}$$

The above ℓ_1-CTA model (8) is a convex optimization problem; however, the objective function is not differentiable at $x = 0$. Since most of the algorithms require differentiability of the objective function, problem (8) needs to be reformulated. The standard reformulation is the transformation of the model (8) to the following LP model:

$$\begin{aligned} &\min_{x^-, x^+} \sum_{i=1}^{n} w_i \left(x_i^+ + x_i^-\right) \\ &s.t.\ \ A\left(x_i^+ - x_i^-\right) = 0, \\ &l \le x^+ - x^- \le u, \end{aligned} \tag{9}$$

where

$$x^+ = \begin{cases} x \text{ if } x \ge 0 \\ 0 \text{ if } x < 0, \end{cases} \quad x^- = \begin{cases} 0 \quad \text{if } x > 0 \\ -x \text{ if } x \le 0, \end{cases} \tag{10}$$

The inequality constraints can further be split into lower and upper bounds constraints for x^+ and x^- separately (see [18]).

Recently, another reformulation of ℓ_1-CTA has been proposed. In [18] it was observed that the absolute value has an obvious second-order cone (SOC) representation

$$t_i = |x_i| \ \rightarrow \ \mathcal{K}_i = \left\{(x_i, t_i) \in \mathbb{R}^2 : t_i \ge \sqrt{x_i^2}\right\}$$

which leads to the following SOC formulation of the ℓ_1-CTA (8)

$$\begin{aligned} &\min_{x} \sum_{i=1}^{n} w_i t_i \\ &s.t.\ \ Ax = 0, \\ &(x_i, t_i) \in \mathcal{K}_i; \quad i = 1, \ldots, n, \\ &l \le x \le u. \end{aligned} \tag{11}$$

The three CTA models outlined above can be solved using interior-point methods (IPMs). IPMs have been developed in the past three decades and have proven to be very efficient in

solving large linear and non-linear optimization problems that were previously hard to solve. Nowadays almost every relevant optimization software, whether commercial or open-source, contains an IPM solver. For more information on IPMs see [19–21, 23, 22] and references therein. Specifically, for conic optimization problems and methods see [3, 4, 15].

We conclude this section by listing several references where numerical experiments and comparisons of different methods for norm-based CTA models were presented [5, 9, 13, 18].

2.2 Motivation to consider different CTA models

In traditional, norm-based continuous CTA models we are finding the closest safe table to the original table with respect to a certain norm, usually l_2 or l_1 norm. However, in norm-based CTA models, there is no control of how well the statistical properties of the data in the table are preserved. The analysis of the data in the masked table with respect to the original table is usually done after the masked table is produced using a CTA model. One of the most utilized measures of analysis is the Chi-square statistic. For example, in [10] Chi-square and Cramer V statistical measures were used in assessing information loss of the masked table produced by the LP CTA model (9). See also references therein.

The definitions of Chi-square statistic and Cramer V measure are well known, however, we list them below for the sake of completeness.

Chi-square statistic of a table is

$$\chi^2 = \sum_{i=1}^{n} \frac{(0_i - e_i)^2}{e_i}, \tag{12}$$

where o_i is an observed cell value and e_i is an expected cell value.

Cramer's V statistical measure is derived from Chi-square statistic

$$V = \sqrt{\frac{\chi^2}{n(r-1)(c-1)}} \tag{13}$$

where r is a number of rows and c is a number of columns and n is a number of cells in the table, i.e. $n = rc$.

An absolute value of the differences instead of a square of the differences as in (12) can also be considered. We call this measure a *Chi-linear* measures.

$$\chi_{abs} = \sum_{i=1}^{n} \frac{|0_i - e_i|}{\sqrt{e_i}}, \tag{14}$$

We performed numerical experiments on a set of randomly generated tables of different dimensions and different numbers of sensitive cells and applied the QP, the LP, and SOC (Conic) CTA models listed above to obtain the masked tables. We used different weights, $w_i = 1/a_i$ and $w_i = 1/e_i$ for QP and the square root of these weights for LP-CTA models. We

calculated the Chi-square statistic, and Cramer V and Chi-linear measures for each masked table and for the original table. The summary of the results is presented in Table 1 below

The review of the results in Table 1 leads to the following observations:

- There is no pattern that would indicate that the masked table produced by one of the CTA models consistently exhibits values of the Chi-square, or other statistics, closer to the values of these statistics for the original table than for other models.
- This is consistent with the findings in [10]. The authors compared the original and masked tables generated using only LP CTA model. They observed that Chi-square and Cramer's V measures are affected by the size of the table, the way the cell values are generated, the number of sensitive cells, upper and lower safe values for sensitive cells, etc.
- The conclusion is that standard norm-based CTA models do not guarantee that Chi-square value, or values of other statistics, computed on the masked table will be as close as possible to the corresponding values of the original table given the constraints.

Given these observations, we propose to consider a different measure of "closeness" between masked and original table which attempts to minimize an objective function that depends on the difference between values of statistics of the original and masked table. In the sequel, we specifically focus on designing a CTA model for Chi-square statistic.

3 Chi-square CTA model and a modification

3.1 Chi-square CTA model

In this section, we propose a CTA model that we call *Chi-Square* CTA where the minimization of the norm-based objective function in (4) is replaced with the minimization of the absolute value of the differences of values of Chi-square statistic of the masked and original table.

The model is as follows:

$$\begin{aligned} &\min \left| \sum_{i=1}^{n} \frac{(z_i - e_i)^2}{e_i} - \sum_{i=1}^{n} \frac{(a_i - e_i)^2}{e_i} \right| \\ &s.t.\ \ Ax = 0, \\ &l \le x \le u. \end{aligned} \tag{15}$$

Below, the objective function is transformed in terms of cell deviations, $x = z - a$ rather than the original masked values z.

$$
\begin{aligned}
f(x) &= \left| \sum_{i=1}^{n} \frac{(z_i - e_i)^2}{e_i} - \sum_{i=1}^{n} \frac{(a_i - e_i)^2}{e_i} \right| \\
&= \left| \sum_{i=1}^{n} \frac{(x_i + a_i - e_i)^2}{e_i} - \sum_{i=1}^{n} \frac{(a_i - e_i)^2}{e_i} \right| \rightarrow d_i := a_i - e_i \\
&= \left| \sum_{i=1}^{n} \frac{(x_i + d_i)^2}{e_i} - \sum_{i=1}^{n} \frac{d_i^2}{e_i} \right| \\
&= \left| \sum_{i=1}^{n} \frac{(x_i + 2d_i)x_i}{e_i} \right| \\
&= \left| \sum_{i=1}^{n} \frac{x_i^2 + 2d_i x_i}{e_i} \right|
\end{aligned}
\tag{16}
$$

The difficulty with the Chi-square CTA model (15) is that it is non-linear, non-convex, and non-smooth. The non-smoothness that is caused by absolute value can be removed by replacing the absolute value with a square of the differences.

$$
\begin{aligned}
&\min \left[\sum_{i=1}^{n} \frac{(z_i - e_i)^2}{e_i} - \sum_{i=1}^{n} \frac{(a_i - e_i)^2}{e_i} \right]^2 \\
&s.t. \ Ax = 0, \\
&l \leq x \leq u.
\end{aligned}
\tag{17}
$$

Using the same substitutions as in (16) we obtain the following nonlinear and non-convex but smooth problem with linear constraints that we call *Chi-square* CTA model.

$$
\begin{aligned}
&\min \left[\sum_{i=1}^{n} \frac{(x_i + 2d_i)x_i}{e_i} \right]^2 \\
&s.t. \ Ax = 0, \\
&l \leq x \leq u.
\end{aligned}
\tag{18}
$$

3.2 Chi-linear CTA Model

The Chi-square CTA model (18) which is a smooth non-linear and non-convex problem can be solved using an appropriate IPM for non-linear problems. However, the non-linearity and non-convexity of the problem make it harder to solve the problem. In other words, the IPM will be able to handle problems of the smaller size and will perform slower than if it is applied to the LP or QP CTA model. At this point, it is still an open question ofwhether model (18) can be transformed into a more tractable problem that can be efficiently solved by IPMs. One option is to consider the modification of the Chi-square CTA formulation (18) by minimizing the absolute value of the sum of absolute values of differences (errors) rather than squares of differences as in (18).

The model is as follows:

$$\begin{aligned} &\min_z \left| \sum_{i=1}^{n} \frac{|z_i - e_i|}{\sqrt{e_i}} - \sum_{i=1}^{n} \frac{|a_i - e_i|}{\sqrt{e_i}} \right| \\ &s.t.\ \ Ax = 0 \\ &l \le x \le u, \end{aligned} \tag{19}$$

We call this model *Chi-linear* CTA model. In what follows we will show that this model can be transformed into the LP problem which then can be solved efficiently using IPMs or simplex based algorithms.

The objective function in (19) can be transformed in a similar way as in (16) for Chi-square CTA model (18).

$$\begin{aligned} f(x) &= \left| \sum_{i=1}^{n} \frac{|z_i - e_i|}{\sqrt{e_i}} - \sum_{i=1}^{n} \frac{|a_i - e_i|}{\sqrt{e_i}} \right| \quad \rightarrow \quad x_i := z_i - a_i \\ &= \left| \sum_{i=1}^{n} \frac{|x_i + a_i - e_i|}{\sqrt{e_i}} - \sum_{i=1}^{n} \frac{|a_i - e_i|}{\sqrt{e_i}} \right| \quad \rightarrow \quad d_i := a_i - e_i \\ &= \left| \sum_{i=1}^{n} \frac{|x_i + d_i|}{\sqrt{e_i}} - \sum_{i=1}^{n} \frac{|d_i|}{\sqrt{e_i}} \right| \quad \rightarrow \quad G := \sum_{i=1}^{n} \frac{|d_i|}{\sqrt{e_i}} \\ &= \left| \sum_{i=1}^{n} \frac{|x_i + d_i|}{\sqrt{e_i}} - G \right| \quad \rightarrow \quad g(x) = \sum_{i=1}^{n} \frac{|x_i + d_i|}{\sqrt{e_i}} \\ &= |g(x) - G| \end{aligned} \tag{20}$$

The Chi-linear CTA model (19) can now be written in the form

$$\begin{aligned} &\min_x |g(x) - G| \\ &s.t.\ \ Ax = 0, \\ &l \le x \le u, \end{aligned} \tag{21}$$

The transformation of (21) to the LP model is derived below. The objective function transformation:

$$\begin{aligned} f(x) &= \left| \sum_{i=1}^{n} \frac{|x_i + d_i|}{\sqrt{e_i}} - G \right| \ \rightarrow \ y_i = x_i + d_i; i = 1, \cdots, n \\ &= \left| \sum_{i=1}^{n} \frac{|y_i|}{\sqrt{e_i}} - G \right| \ \rightarrow \ |y_i| = y_i^+ + y_i^-; y_i^+, y_i^- \ge 0; i = 1, \cdots, n \\ &= \left| \sum_{i=1}^{n} \frac{y_i^+ + y_i^-}{\sqrt{e_i}} - G \right| \ \rightarrow \ t = \sum_{i=1}^{n} \frac{y_i^+ + y_i^-}{\sqrt{e_i}} - G \\ &= |t| \ \rightarrow \ t = t^+ - t^-, \ t^+, t^- \ge 0 \\ &= t^+ + t^- \end{aligned} \tag{22}$$

Equality constraints transformations:

$$\begin{aligned} &Ax = 0 \rightarrow\ A(y-d) = 0 \\ &\rightarrow\ A(y^+ - y^-) = Ad \\ &\rightarrow\ Ay^+ - Ay^- = Ad \end{aligned} \tag{23}$$

It is not hard to show that

$$Ad = 0, \tag{24}$$

hence, we have

$$Ax = 0 \quad \rightarrow \quad Ay^+ - Ay^- = 0. \tag{25}$$

Inequality constraints transformations:

$$\begin{aligned} &l \le x \le u \rightarrow\ l \le y - d \le u \\ &\rightarrow\ l + d \le y \le u + d \\ &\rightarrow\ l + d \le y^+ - y^- \le u + d \end{aligned} \tag{26}$$

The Chi-linear CTA model (21) transforms now to the following LP problem.

$$\begin{aligned} &\min\left(t^+ + t^-\right) \\ &s.t.\ \ Ay^+ - Ay^- = 0, \\ &t^- - t^- = \sum_{i=1}^{n} \frac{y_i^+ + y_i^-}{\sqrt{e_i}} - G \\ &l + d \le y^+ - y^- \le u + d \\ &t^+, t^- \ge 0 \\ &y^+, y^- \ge 0, \end{aligned} \tag{27}$$

4 Numerical results

In this section a two-dimensional table stated in Figure 1 in [18] is considered. The table is listed in Figure 1 below as Table (a).

In order to make the paper more self-contained we give the description of the parameters used to formulate constraints of continuous CTA models that are based on table (a), as it was given in [18]:

- The linear constraints are obtained from the requirement that the sum of the elements in each row (or column) remains constant and is equal to the corresponding component in the last column (or row) of the table (a).
- The sensitive cells are cells a_1 and a_{12}. For both of them the upper safe values are enforced, which are listed in the parentheses in the lower right corners of the cells, $upl_1 = 3$ and $upl_{12} = 5$ respectively. Hence, in the transformed tables the upper safe value of the cell a_1 should be 13 or above and for a_{12} the upper safe value should be 18 or above.

- For the nonsensitive cells the lower and upper bounds are set to be zero and positive infinity respectively, that is, $l_{a_i} = 0$ and $u_{a_i} = \text{inf}$ for $i = 2, \ldots, 11$.
- We take the weights in the objective function to be $w_i = 1$; $i = 1, \ldots, 12$.

This table is used to build five different CTA models, ℓ_1-CTA LP formulation (9), ℓ_1-CTA SOC formulation (11), ℓ_2-CTA QP formulation (7), Chi-linear CTA LP formulation (27), and Chi-square CTA formulation (18). These CTA models are solved using appropriate interior-point methods (IPMs). The first four models were solved using MOSEK solver [1] while the last one is solved using IPOPT solver [2]. The results are listed in Figure 1.

In the next Table 2 the values of Chi-square, Cramer V, and Chi-linear statistical measures are listed for the original table and related masked tables produced by five CTA models.

From Table 2 we observe that the value of the Chi-square statistic of the masked table produced by the Chi-square CTA model indeed differs the least from the value of the Chi-square statistic of the original table. Similarly, the Chi-linear measure of the masked table produced by the Chi-linear CTA model is the closest to the Chi-linear measure of the original table.

The second observation is about p-values of the Chi-square statistic for the tables listed in Table 2. The p-values for the tables are as follows: original: 0.82, ℓ_2-QP: 0.15, ℓ_1-LP: 0.11, ℓ_1-SOC: 0.08, Chi-square: 0.34, Chi-linear: 0.29. As expected, the p-value of the Chi-square table is the closest to the p-value of the original table. However, the discrepancy between the p-values of the original and masked tables is significant and deserves comment.

The Chi-square statistic is very sensitive to the number of sensitive cells in the table and the level of perturbation needed to get the safe values for these cells [6, 10]. The larger the number of sensitive cells and the level of perturbation, in general, the larger the discrepancy between Chi-square statistic values and, consequently, p-values. Therefore, the discrepancy is more due to the requirements for the protection of tabular data and less due to the CTA model used to obtain the masked table which satisfies these requirements. Nevertheless, the new Chi-square CTA model proposed in this paper achieves the p-value of the masked table that is the closest to the p-value of the original table among all other CTA models. On the more general note, this is an illustration of the interplay between maximizing the utility of the masked data while keeping disclosure risk under control which is at the heart of the theory and methods of SDL.

5 Concluding remarks and future work

In this paper, a novel approach to building Continuous CTA models for statistical disclosure limitation of tabular data is discussed. The standard norm-based CTA model finds the closest safe (masked) table to the original table while satisfying additivity equations and safe value inequality constraints, as described in Section 2.1. The measure of closeness is usually measured using an ℓ_1 or ℓ_2 norm.

The numerical experiments summarized in Table 1 in Section 2.2 suggest that there is no pattern which would indicate that one CTA model consistently produces the values of the

Chi-square statistic, or other statistics, of the masked and original table that are closer to each other than for any other model. Hence, we propose a CTA model, which we call *Chi-square* CTA model (18), that produces a masked table with Chi-square statistic closest to Chi-square statistic of the original table.

Given the non-linearity and non-convexity of the Chi-square CTA model, we also consider a modification of this model, a *Chi-linear* CTA model (27) that can be transformed to LP problem, hence allowing IPMs to solve high dimensional tables efficiently. The price to pay is that the closeness between Chi-square statistics of an original and masked table may be affected. Further examination of this topic is the subject of future research.

The goal of the paper is mainly theoretical, that is, to present a novel Chi-square CTA model and its modification, Chi-linear CTA model, as an illustration of a possible new approach in building CTA models which produce masked tables that are the closest to the original table in terms of a certain statistic, rather than in terms of a distance between tables.

The rationale behind the new approach is to consider *Analysis Specific* CTA models. On one hand, they may be more narrow in scope, however, they produce the optimal result for the specific analysis. On the other hand, norm-based CTA models may be wider in scope and produce tables that may have ‘‘relatively good” results for multiple different statistical measures, but ‘really good” (optimal) for none. In addition, we have no explicit control of the quality of results in the norm-based CTA approach.

Directions for future research include more extensive numerical experiments on a larger set of randomly generated two-dimensional tables of different sizes and different numbers of sensitive cells. A more theoretical direction for future research is to examine whether the Chi-square CTA model (18) can be transformed into a more tractable problem that can be efficiently solved by IPMs.

Acknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors only and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. Andersen ED, MOSEK solver. https://mosek.com/resources/doc, 2020.
2. Waechter A, Laird C IPOPT solver. https://coin-or.github.io/Ipopt/
3. Alizadeh F, and Goldfarb D, Second-order cone programming. Math. Programming, 95(1):3–51, 2003.
4. Andersen ED, Roos C and Terlaky T, On implementing a primal-dual interior-point method for conic quadratic optimization. Math. Programming, 95(2):249–277, 2003.
5. Castro J, A CTA Model Based on the Huber Function. Privacy in Statistical Databases 2014, LNCS, 8744:79–88, 2014.
6. 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.
7. Castro J, Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research, 171:39–52, 2006.

8. Castro J, Recent advances in optimization techniques for statistical tabular data protection. European Journal of Operational Research, 216:257–269, 2012.
9. Castro J and Cuesta J, Solving ℓ_1-CTA in 3D tables by an interior-point method for primal block-angular problems. TOP, 21:25–47, 2013.
10. Castro J and Gonzalez JA, Assessing the information loss of controlled adjustment methods in two-way tables. Privacy in Statistical Databases 2014, LNCS, 8744:79–88, 2014.
11. Castro J and Gonzalez JA, A fast CTA method without complicating binary decisions. Documents of the Joint UNECE / Eurostat Work Session on Statistical Data Confidentiality, Statistics Canada, Ottawa, 1–7, 2013.
12. Castro J and Gonzalez JA, A multiobjective LP approach for controlled tabular adjustment in statistical disclosure control. Working paper, Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, 2014.
13. Castro J and Giessing S, Testing variants of minimum distance controlled tabular adjustment. In Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 333–343, 2006.
14. Dandekar RA and Cox LH, Synthetic tabular Data: an alternative to complementary cell suppression. Manuscript, Energy Information Administration, U.S 2002.
15. Gu G, Interior-Point Methods for Symmetric Optimization. Ph.D. Thesis, TU Delft, 2009.
16. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Schulte Nordholt E, Spicer K and DeWolf P-P, Statistical Disclosure Control. Wiley, Chichester, United Kingdom, 2012.
17. Karr AF, Kohnen CN, Oganian A, Reiter JP, and Sanil AP, A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician, 60(3):224–232, 2006.
18. Lesaja G, Castro J, and Oganian A A Second Order Cone Formulation of Continuous CTA Model. Lecture Notes in Computer Science 9867, Springer, 41–53, 2016.
19. Lesaja G, Introducing Interior-Point Methods for Introductory Operations Research Courses and/or Linear Programming Courses. The Open Operational Research Journal, 3:1–12, 2009.
20. Lesaja G and Roos C, Kernel-based interior-point methods for monotone linear complementarity problems over symmetric cones. J. Optim. Theory Appl, 150(3):444–474, 2011.
21. Nesterov Y and Nemirovski A, Interior-Point Polynomial Algorithms in Convex Programming. SIAM Studies in Applied Mathematics, Volume 13, SIAM, Philadelphia, 1994.
22. Roos C, Terlaky T, and Vial J. Ph., Theory and Algorithms for Linear Optimization. An Interior-Point Approach. Springer Science, 2005.
23. Wright SJ, Primal-Dual Interior-Point Methods. SIAM, Philadelphia, 1996.

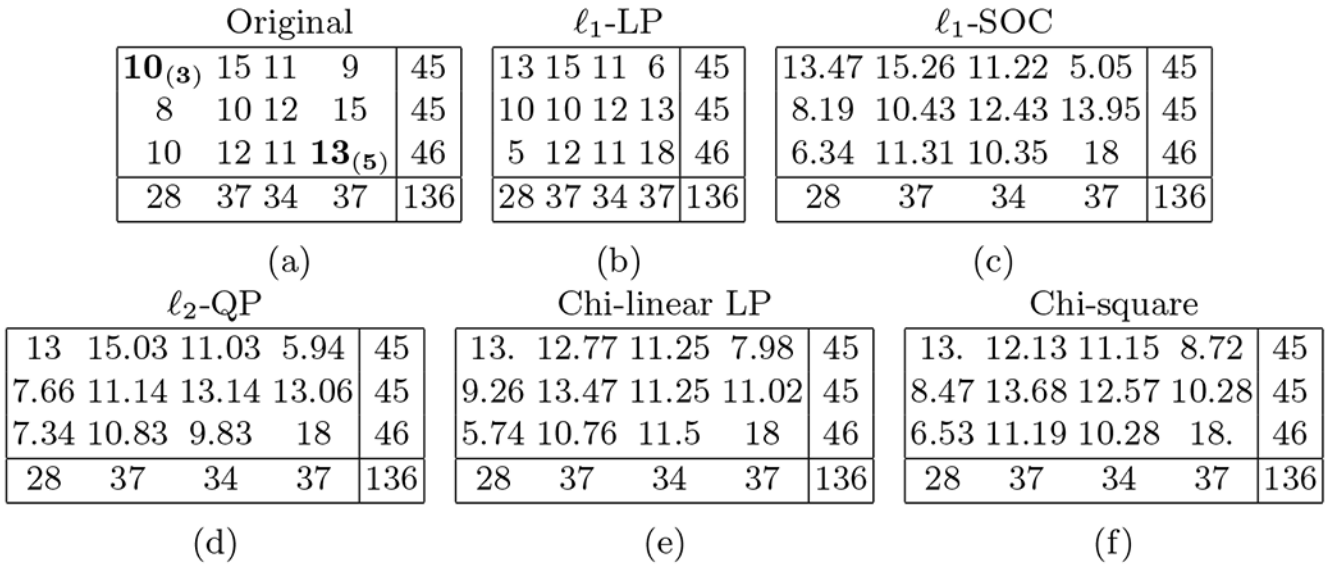

Fig. 1.
Masked tables produced by different CTA models for table (a)

Table 1.

Values of statistical measures

Percentage of sensitive cells	Tables	Statistical measures		
		Chi-square	Chi-linear	Cramer V
	Original	1790.69	506.40	0.228821
10				
	LP	2039.68	536.77	0.244212
	$W(a_i) - LP$	2034.97	518.96	0.243930
	$W(e_i) - LP$	2111.99	540.48	0.248504
	QP	1971.56	519.52	0.240100
	$W(a_i) - QP$	1940.21	512.82	0.238183
	$W(e_i) - QP$	1976.04	519.89	0.240372
	Conic	1954.00	520.86	0.239028
15				
	LP	2008.59	532.59	0.242344
	$W(a_i) - LP$	1960.27	520.67	0.239411
	$W(e_i) - LP$	2046.51	539.46	0.244621
	QP	2012.38	534.54	0.242573
	$W(a_i) - QP$	1952.36	526.04	0.238928
	$W(e_i) - QP$	2019.59	535.99	0.243007
	Conic	1968.98	525.72	0.239942
20				
	LP	1950.37	513.28	0.238806
	$W(a_i) - LP$	1922.12	511.77	0.237070
	$W(e_i) - LP$	1993.11	522.35	0.241408
	QP	1949.98	516.88	0.238782
	$W(a_i) - QP$	1881.74	505.13	0.234566
	$W(e_i) - QP$	1947.82	516.88	0.238650
	Conic	1957.35	520.91	0.239233

Table 2.

Values of the three statistical measures

	Statistical measures		
Tables	**Chi-square**	**Chi-linear**	**Cramer V**
original	2.89	4.70	0.20
ℓ_2-QP	9.49	8.74	0.36
ℓ_1-LP	10.44	8.49	0.32
ℓ_1-SOC	11.30	9.26	0.39
Chi-Square	**6.81**	7.17	0.31
Chi-Linear	7.38	**6.55**	0.32