Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

A fast CTA method without the complicating binary decisions[1]

Prepared by Jordi Castro, José A. González, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Catalonia, Spain (jordi.castro@upc.edu, jose.a.gonzalez@upc.edu)

[1] This work has been supported by grants MTM2012-31440 of the Spanish research program, SGR-2009-1122 of the Government of Catalonia, and project DwB INFRA-2010-262608 of the EU FP7.

A fast CTA method without the complicating binary decisions [1]

Jordi Castro, José A. González

Department of Statistics and Operations Research, Universitat Politècnica de Catalunya Jordi Girona 1–3, 08034 Barcelona, Catalonia

(jordi.castro@upc.edu, jose.a.gonzalez@upc.edu)

Abstract. CTA is a recent approach for statistical disclosure control in tabular data. Its purpose is to compute the closest safe table to the original data, using some distance. Sensitive cells are adjusted either upwards or downwards (binary decision), and the resulting cells have to be accordingly (and minimally) modified to preserve marginals. The binary decisions are modeled as disjunctive constraints, CTA resulting in a difficult mixed integer linear problem. In this talk a variant of CTA without binary variables is discussed. Binary variables are pre-fixed, thus obtaining a linear problem (LP). Since this LP-CTA may be infeasible, changes may be needed to (1) the cell bounds; (2) the tables relations; (3) or the protection levels. Together with the original objective function, this results in a four-objective optimization problem. We discuss how it can be formulated and solved as a multiobjective optimization problem. A software package implementing this idea is presented.

1 Introduction

CTA is a post-tabular approach which looks for the closest safe table to the original unsafe table. CTA achieves disclosure limitation by either increasing or decreasing by at least a certain amount (*protection level*) the cell values of a subset of sensitive cells, and then adjusting the rest of cells to preserve some desired constraints. CTA relies on optimization methods, mainly mixed integer linear programming (MILP), and linear programming (LP). This offers great flexibility when some table properties want to be preserved in the released table (e.g., total or subtotals cells, or proximity to certain relevant cells). CTA is one of the methods discussed in the recent monograph Hundepool et al. (2012), and it has been applied within a wider scheme for the protection of structural business statistics released by Eurostat (project coordinated by Statistics Netherlands, with the participation of Destatis and Universitat

[1]This work has been supported by grants MTM2012-31440 of the Spanish research program, SGR-2009-1122 of the Government of Catalonia, and project DwB INFRA-2010-262608 of the EU FP7.

Politècnica de Catalunya) (Giessing et al., 2009). CTA was implemented in a software package (Castro et al., 2009), which is being improved within the Data without Boundaries (DwB) EU FP7 project.

The standard CTA method (Dandekar and Cox, 2002; Castro, 2006, 2011) considers as decisions of the optimization problem the direction of protection for sensitive cells (either upward or downward). These disjunctive constraints need the solution of a difficult MILP. In this talk we will discuss a more efficient procedure, where these binary decisions are a priori fixed, thus obtaining a continuous LP problem. Technical details will be skipped, since they can be found in Castro and González (2013). Although the distance between the original and released tables may be larger with this LP-CTA variant than with the MILP one, it will be in general orders of magnitude faster.

The structure of this short document is as follows. The classical MILP CTA will be outlined in Section 2. The LP-CTA variant will be briefly discussed in Section 3. Finally, details about the CTA package will be provided in Section 4

2 Outline of minimum distance MILP-CTA

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

- 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), and a vector $w \in \mathbb{R}^n$ of positive weights for the deviations of cell values.
- A lower and upper bound for each cell $i = 1, \ldots, n$, respectively l_{x_i} and u_{x_i}, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i $l_{x_i} = 0$ ($l_{x_i} = -\infty$ if $a \geq 0$ is not required) and $u_{x_i} = +\infty$ can be used.
- A set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \{1, \ldots, n\}$ of indices of s confidential cells.
- A lower and upper protection level 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 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:

$$\begin{array}{rl} \min\limits_{x} & ||x - a||_L \\ \text{subject to} & Ax = b \\ & l_x \leq x \leq u_x \\ & x_i \leq a_i - lpl_i \text{ or } x_i \geq a_i + upl_i \quad i \in \mathcal{S}. \end{array} \tag{1}$$

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z = x - a$, $l_z = l_x - a$, $u_z = u_x - a$, using the L_1 distance weighted by w, and introducing variables $z^+, z^- \in \mathbb{R}^n$ so that $z = z^+ - z^-$ and $|z| = z^+ + z^-$, the final MILP model for CTA is:

$$\begin{aligned}
\min_{z^+,z^-,y} \quad & \sum_{i=1}^{n} w_i(z_i^+ + z_i^-) && \text{(2a)}\\
\text{subject to} \quad & A(z^+ - z^-) = 0 && \text{(2b)}\\
& 0 \le z^+ \le u_z, \quad 0 \le z^- \le -l_z && \text{(2c)}\\
& y \in \{0,1\}^s && \text{(2d)}\\
& \left.\begin{array}{ll} upl_i\, y_i & \le z_i^+ \le u_{z_i} y_i \\ lpl_i(1-y_i) & \le z_i^- \le -lz_i(1-y_i) \end{array}\right\} i \in \mathcal{S} && \text{(2e)}
\end{aligned}$$

Constraints (2b) impose feasibility of the published perturbed table. Constraints (2c) guarantee perturbations are within allowed bounds. Constraints (2d)–(2e) force the new table is safe. When $y_i = 1$ the constraints mean $upl_i \le z_i^+ \le u_{z_i}$ and $z_i^- = 0$, thus the protection sense is "upper"; when $y_i = 0$ we get $z_i^+ = 0$ and $lpl_i \le z_i^- \le -l_{z_i}$, thus the protection sense is "lower".

3 The LP-CTA variant

Problem (2) is a difficult MILP, whose solution may take a long time for large tables. The LP-CTA variant is obtained by fixing in (2) the binary variables y. However, this may result in an infeasible LP problem. Several procedures can be devised to get some good y values (e.g., SAT approaches (Castro and González, 2013)), but none of them can guarantee feasibility of the LP. Therefore we may need to modify the original problem. Changes may be needed to (1) the cell bounds; (2) the table relations; (3) or the protection levels. Together with the original objective function, this results in a four-objective optimization problem. This multicriteria optimization problem may be solved using the lexicographic-minimization (*lexmin*) approach, assigning priorities to the four different objectives. The lexmin method is known to guarantee Pareto optimal solutions (details in Castro and González (2013)). This LP-CTA variant only needs the solution of four LPs, thus being orders of magnitude faster than the standard MILP-CTA formulation. Therefore it may be used for either the protection in online servers systems, or the protection of large fine-grained tables obtained by crossing all the available categorical variables in some microdate file (from which it could be possible to quickly reproduce any other table) (Giessing and Höhne, 2011). The continuous LP-CTA variant has been implemented in the CTA package discussed in next Section.

4 The CTA package

The package implements both the MILP-CTA and the continuous LP-CTA variants. The continuous LP-CTA implements the lexmin multiobjective method.

The package can be used in three different ways:

- As a standalone application through the command line.
- As a standalone application through a Graphical User Interface (GUI), especially useful for non-expert users. Figures 1–3 show three screenshots for some particular states of the GUI. Figure 1 corresponds to the screen for the solution of a MILP-CTA problem; Figure 2 shows the screen for a LP-CTA; and Figure 3 shows a solver log file, and the output file with the input and protected cell table values.
- As a callable library which allows creating an own program or to be used in other ad-hoc applications..

The CTA package is linked with six state-of-the-art solvers: CPLEX, XPRESS, GLPK, CBC, CLP and SYMPHONY. CLP is only valid for LPs; the other solvers can deal with both LPs and MILPs. CBC uses CLP as the LP solver. This multi-solver platform was developed using Osi (Open Solver Interface), which provides an abstract interface to communicate with solvers. CPLEX and XPRESS are commercial solvers and a license is needed, but GLPK, CBC, CLP, and SYMPHONY are license free solvers.

The current version of the CTA package has about 15000 lines of C/C++ code. Approximately 9000 of them have been developed within the DwB project. Some of its relevant features are:

- It implements the MILP method, which allows to find optimal directions to the sensitive cells in order to provide a table as close as possible to the original one.
- It implements a Block Coordinate Descent (BCD) heuristic. This heuristic suboptimally solves the MILP CTA problem, by decomposing it into simpler subproblems. More details about this approach can be found in González and Castro (2011).
- It implements the fast LP-CTA version, where binary decisions are pre-fixed. In this case CTA reduces to the solution of four continuous LPs. The multi-objective problem is solved by a lexmin optimization by assigning priorities to the four objectives. The user may choose, among many other parameters, this priority order.

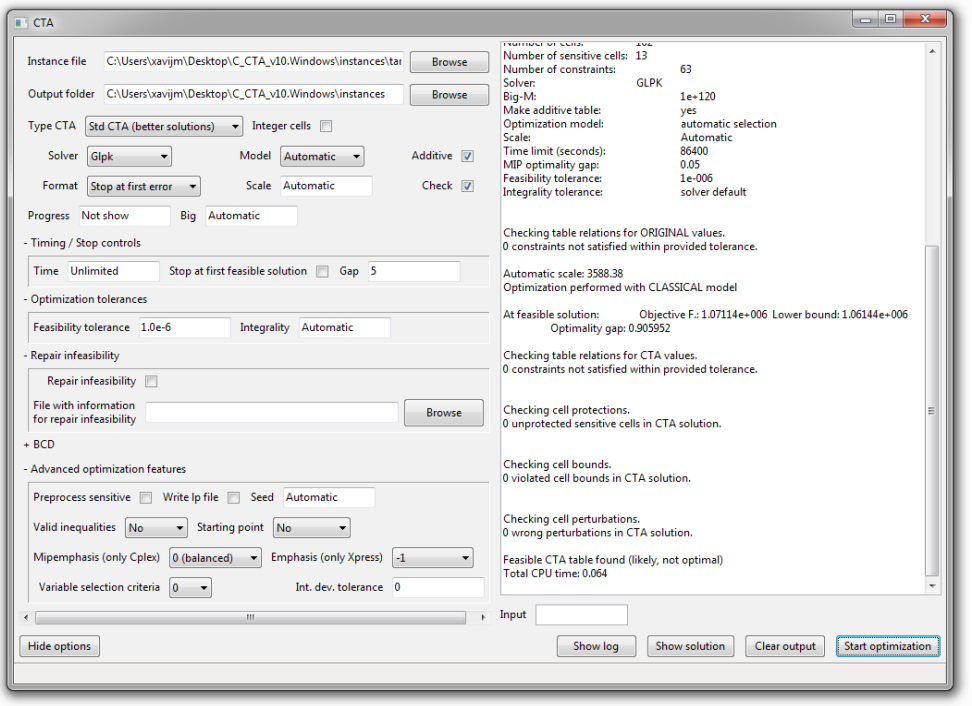

Figure 1: Screenshot of the GUI for the solution of a MILP-CTA.

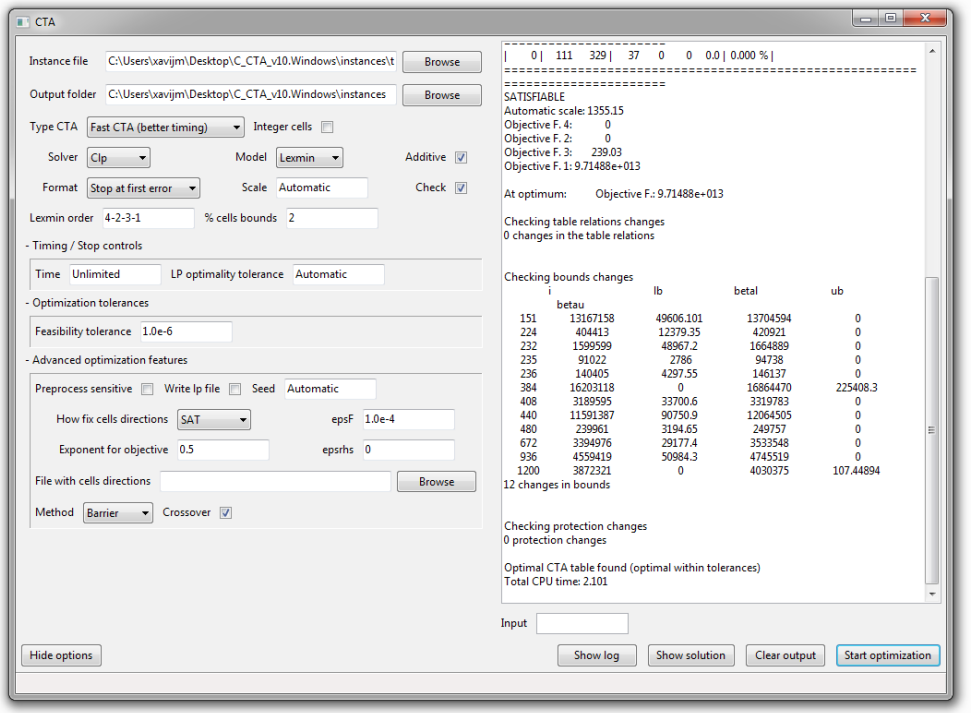

Figure 2: Screenshot of the GUI for the solution of a LP-CTA.

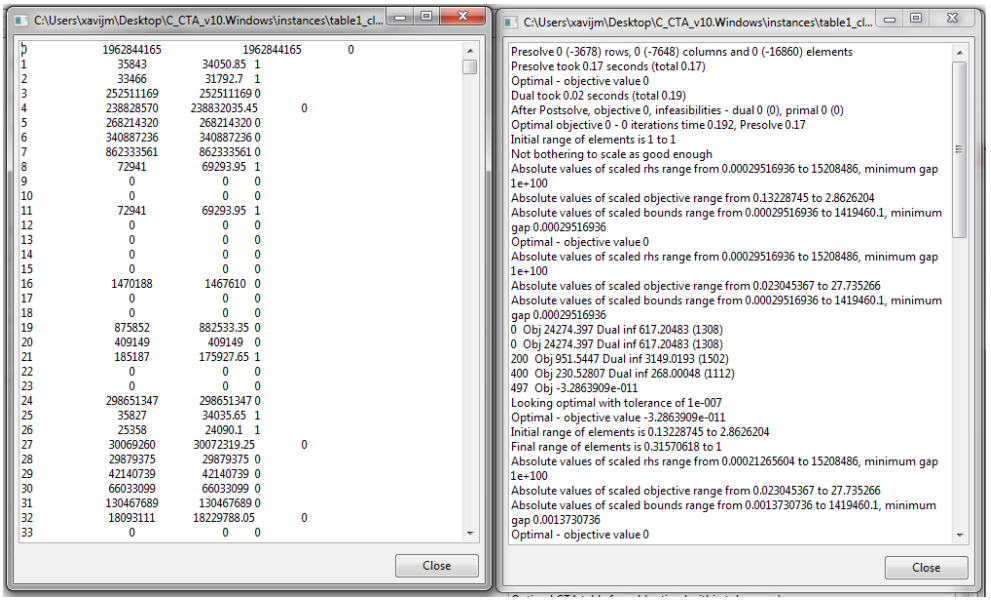

Figure 3: Screenshot of the solver log and output file with original and protected values.

- Graphical user interface. Some screenshots are reported in Figures 1–3. All the capabilities of the command-line version are available with the GUI version.
- Both Linux and Windows versions.
- Extension for integrality in cell values. In general integrality is guaranteed, but in some tables is necessary and it may be lost; this option allows to force integrality. The resulting model is harder and more time consuming, and thus it is not recommended for large tables.
- Auto-scaling of input data. This avoids several numerical problems related to the optimization solvers.
- Usage of Osi for communication with several LP and MILP solvers. The package is linked with two commercial (CPLEX and XPRESS) and four free solvers (CBC, SYMPHONY, GLPK, CLP). The user sees a unique front-end to control the many parameters of the different solvers (e.g., optimality tolerance). Internally this is translated to the particular solver functions, either using Osi or directly interacting with the solver interface for advanced parameters.
- The CTA package is free. Binaries can be obtained directly from the authors or from the DwB project. A license of CPLEX and XPRESS is needed if these solvers are to be used. Additionally, National Statistical Agencies requiring the source code may contact the authors.

References

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection, *European Journal of Operational Research*, 171, 39–52.

Castro, J. (2011). Recent advances in optimization techniques for statistical tabular data protection, *European Journal of Operational Research*, 216, 257–269.

Castro, J., and González, J.A. (2013). A multiobjective LP approach for controlled tabular adjustement in statistical disclosure control. Working paper, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya.

Castro, J., González, J.A., and Baena, D. (2009). User's and programmer's manual of the RCTA package, Technical Report DR 2009-01, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya.

Dandekar, R.A., and Cox, L.H. (2002). Synthetic tabular data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. Department of Energy.

Giessing, S., and Höhne J. (2011). Eliminating small cells from census counts tables: some considerations on transition probabilities, *Lecture Notes in Computer Science* 6344, 52–65.

Giessing, S., Hundepool, A., and Castro, J. (2009). Rounding methods for protecting EU-aggregates, in *Worksession on statistical data confidentiality. Eurostat methodologies and working papers*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 255–264.

González, J.A., and Castro, J. (2011) A heuristic block coordinate descent approach for controlled tabular adjustment, *Computers & Operations Research* 38, 1826-1835

Hundepool, A, Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P.-P. (2012), *Statistical Disclosure Control*, Chichester, Wiley.