WP. 48
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
(Tarragona, Spain, 26-28 October 2011)

Topic (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

Present and future research on controlled tabular adjustment,

Prepared by Jordi Castro and José A. González, Universitat Politècnica de Catalunya, Spain

Present and future research on controlled tabular adjustment [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. Controlled tabular adjustment (CTA) can be classified within the group of approaches that perturb output data (i.e., tabular data), unlike other methods that focus on the original microdata. Being a post-tabular data perturbation technique it becomes easier to guarantee consistency and quality of the released information (e.g., table additivity, preservation of subtotal or total cells of the original table, etc.). On the other hand, it may be computationally more costly than pre-tabular strategies. The purpose of this work is twofold. First, we will review a recently used heuristic to suboptimally solve CTA (which is a mixed integer linear optimization problem). For some tables this heuristic provided decent solutions much faster than other state-of-the-art optimization methods. This approach can be useful when CTA is applied to a pre-defined "static" set of tables. The second goal of the paper is to provide and discuss variants of CTA when applied in an on-line table generation system. In this case, tables can be dynamically generated over time, and CTA has to face two new challenges: (i) it has to deliver an on-line fast solution; (ii) protection senses of sensitive cells have to be consistent (i.e, the same) when the same sensitive cell appears in two (or more) tables which are generated and protected at different moments. Some of these ideas will be implemented in the recently started Data without Boundaries FP7 EU project.

1 Introduction

Tabular data protection methods can be classified as *pre-tabular* or *post-tabular*. Pre-tabular methods perturb the microdata such that the tables produced from this modified microdata are considered safe enough. The approach of Giessing and Höhne (2011) belongs to this family. Post-tabular methods deal directly with the resulting tables, either modifying (*perturbative* methods) or hiding (*nonperturbative* methods) information. The most well known nonperturbative post-tabular tool is cell

[1]This work has been supported by grants MTM2009-08747 of the Spanish Ministry of Science and Innovation, SGR-2009-1122 of the Government of Catalonia, and the project DwB INFRA-2010-262608 of the EU FP7

suppression. Among the perturbative post-tabular approaches we may distinguish between those that preserve additivity (and thus they rely on linear optimization), like *controlled tabular adjustment* (CTA) (Castro, 2006; Dandekar and Cox, 2002) and those that perturb cells but can not guarantee preservation of subtotals or additivity (e.g., Fraser and Wooton (2006); Shlomo and Young (2008)). A recent survey on tabular data protection can be found in Castro (2011).

When table additivity or preservation of marginals is compulsory, CTA may be considered a good choice. Pre-tabular and other post-tabular methods can not easily guarantee such strong constraints. Since CTA relies on optimization (linear/quadratic programming (LP/QP) and mixed integer linear/quadratic programming (MILP/MIQP)): (i) it offers great flexibility in the control of the amount of additivity, preservation or subtotals, or cell perturbations required by the user; (ii) it can be applied to any type of table; (iii) it finds the safe closest table to the original one either using either L_1 (Manhattan distance) or L_2 (Euclidean distance) norms. The price to be paid by all these features is the solution of a (sometimes difficult) MILP/MIQP optimization problem. The RCTA (Restricted CTA) package (Castro and González, 2011a; Castro et al., 2009) is an implementation of CTA, which include many features as the choice of solver (Cplex or Xpress), extensions for nonadditive tables and negative protection levels, and a feasibility tool that deals with infeasible instances. CTA is one of the methods discussed in the handbook Hundepool et al. (2010), 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 Politècnica de Catalunya) (Giessing et al., 2009).

Given a table to be protected, CTA achieves disclosure limitation by either increasing or decreasing by at least a certain amount (*protection level*) the cell values of a subset of sensitives cells, and then adjusting the rest of cells to preserve subtotals and additivity. This problem involves two types of decisions: binary decisions to decide either the increase or decrease of sensitive cells; and continuous perturbations for the remaining cells. This can be formulated as an optimization problem. Up to now, CTA has always been applied to a predefined set of tables. In this setting, it makes sense to compute together the binary and continuous variables, obtaining a MILP model. The first part of this paper will discuss a recent heuristic used to obtain good suboptimal solutions to the MILP CTA problem.

The situation drastically changes when the set of tables is a priori not predefined, and they are rather produced by on-line tabular data servers. In the second part of this work we will discuss the pros and cons of the classical MILP CTA formulation for on-line systems, and how CTA can be adjusted for this new paradigm, following the recommendations of Giessing (2011).

We will first start by outlining the classical MILP CTA formulation in next section.

2 Outline of minimum distance 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:

$$\min_{z^+, z^-, y} \quad \sum_{i=1}^{n} w_i (z_i^+ + z_i^-) \tag{2a}$$

$$\text{subject to} \quad A(z^+ - z^-) = 0 \tag{2b}$$

$$0 \leq z^+ \leq u_z, \quad 0 \leq z^- \leq -l_z \tag{2c}$$

$$y \in \{0, 1\}^s \tag{2d}$$

$$\left.\begin{array}{ll} upl_i\, y_i & \leq z_i^+ \leq u_{z_i} y_i \\ lpl_i (1 - y_i) & \leq z_i^- \leq -l z_i (1 - y_i) \end{array}\right\} i \in \mathcal{S} \tag{2e}$$

Constraints (2b) impose feasibility of the published perturbed table. Constraints (2c) guarantee perturbations are within allowed bounds. Constraints (2d)–(2e) force

Table 1: Dimensions of instances

instance	n	s	m	N. coef.	cont.	bin.	constr.
case 35	499298	55527	20747	1007124	998596	55527	242855
case 36	1200439	107743	45638	2417196	2400878	107743	476610
case 37	296004	42652	10904	597057	592008	42652	181512
case 38	572373	81359	18873	1152345	1144746	81359	344309

the new table is safe. When $y_i = 1$ the constraints mean $upl_i \leq z_i^+ \leq u_{z_i}$ and $z_i^- = 0$, thus the protection sense is "upper"; when $y_i = 0$ we get $z_i^+ = 0$ and $lpl_i \leq z_i^- \leq -l_{z_i}$, thus the protection sense is "lower".

3 A heuristic approach for CTA

In the recent work González and Castro (2011) a block coordinate descent (BCD) approach was applied to CTA. Briefly, given an optimization problem, BCD first decomposes it in subproblems that contain a subset of the variables while the remaining ones kept fixed; next, it optimally solves a sequence of those subproblems, fixing the solution of the previously solved subproblem in the current one. For CTA, the set of binary variables y (protection senses) is assigned to each subproblem, while the continuous variables z^+ and z^- (perturbations) are not fixed, so they can be modified in any subproblem.

BCD does not guarantee convergence to an optimal solution (except for strictly convex optimization problems) but in practice it has a good behaviour. Since BCD needs a feasible starting point, we considered a SAT-based strategy for warm-starting the algorithm (SAT comes from Boolean satisfiability, a well-known problem in complexity theory). It works as follows. First, from the linear constraints of the table, a set of infeasible combinations of protection senses (i.e., values of binary variables) were detected. Second, an assignment that avoided all the infeasible combinations was found for the binary variables. This assignment is obtained formulating the problem as a SAT problem. Avoiding these infeasible combinations is a necessary, but not sufficient, condition for feasibility. However in practice the solution obtained by SAT was generally feasible. We note that the infeasible combinations detected can be added as feasibility cuts to strengthen the CTA model; this is work in progress.

Figure 1 shows illustrative results obtained with four large 1H2D tables (two dimensional tables with one hierarchical variable). The dimensions of these tables are shown in Table 1 (number of cells n, number of sensitive cells s, number of table constrains m, number of coefficients in tabular constraints "N. coef", number of continuous and binary variables "cont." and "bin", and overall number of constraints "constr."). Figure 1 shows the evolution of the best objective function for four variants, named BCD, Tree BCD, SAT B&C and B&C. BCD and Tree BCD are

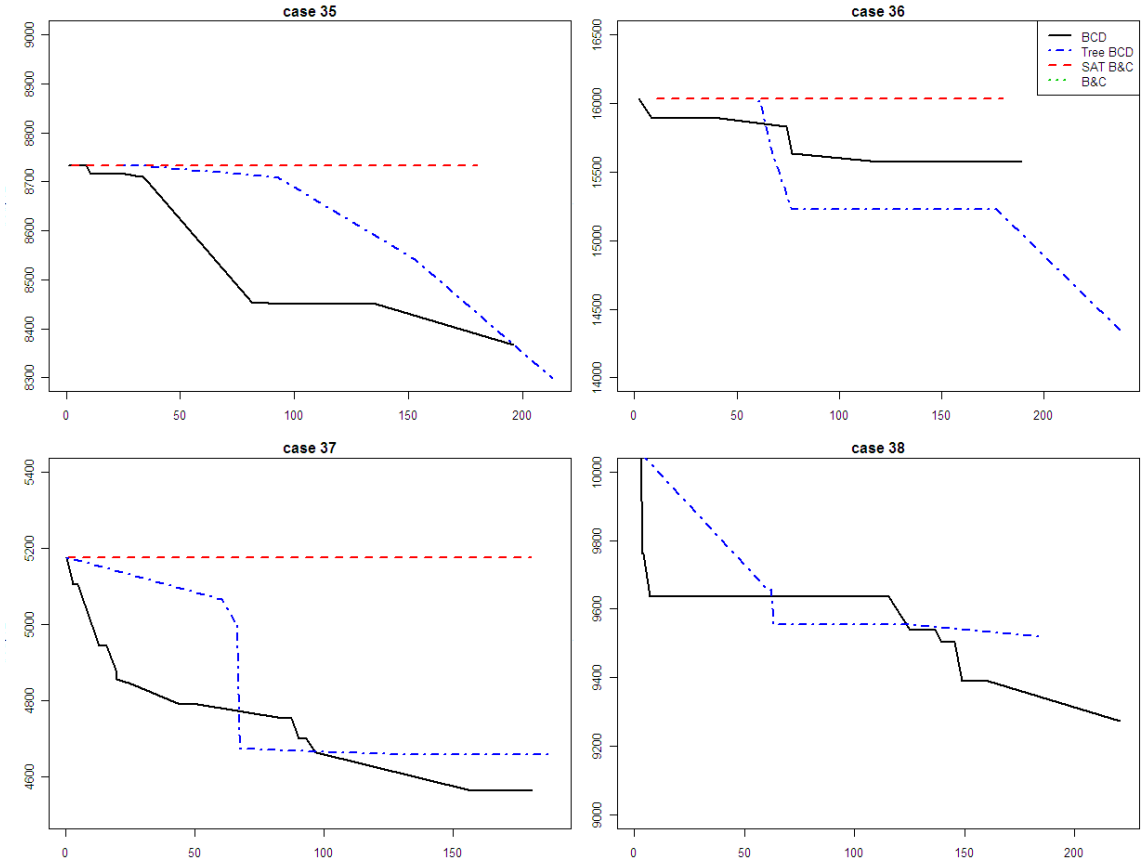

Figure 1: Evolution of incumbent. The horizontal axis shows the CPU time in minutes, and the vertical axis the best objective function value achieved. The B&C solutions are not shown since the objective functions were 1000 times those of the other approaches (exceeding the scale of the vertical axis).

two BCD variants obtained by respectively partitioning the set of binary variables randomly or according to the subtables of the 1H2D table. B&C corresponds to the standard Cplex branch-and-cut algorithm. SAT B&C was obtained by initializing the Cplex B&C with the feasible point returned by SAT. We note that BCD subproblems were solved by Cplex B&C. Clearly, BCD variants are very competitive against standard B&C, and SAT B&C was much more efficient that B&C.

4 CTA for on-line tabular data servers

Unlike the "static view" that represents the protection of a pre-defined set of tables, on-line data servers provide a "dynamic view", where tables are generated over time, and, in addition, the set of tables is a priori unknown. CTA has been used, up to now, only for the "static view". The features of a CTA-like procedure for a "dynamic view" should include (Giessing, 2011):

- *Consistency on input*: if a cell appears in several tables, it is either nonsensitive, or sensitive in all these tables (i.e., it can not be sensitive in one table, and nonsensitive in others, otherwise its true value could be released). This feature has to be guaranteed by the sensitivity rules considered, rather than for the protection method.

- *Consistency on output*: if a sensitive cell of value a and protection levels (for instance, both upper and lower) p appears in two (or more) different tables, the protection sense (either lower or upper) should be the same in all tables. Otherwise, the released values for this cell in different tables may be both $a-p$ and $a+p$; then, just taking the average value any attacker would recompute a.

- *Efficiency*: the tool should provide a quick solution in an on-line table generation system, unlike the standard CTA approach which can take a long time to process very large tables.

- *Reliability*: a solution (table) should always be provided, possibly non-optimal or slightly violating some of the constraints of the table (additivity, bounds on cell values, or, ultimately, protection levels).

We first note that the standard CTA approach applied to an on-line table would not satisfy consistency on output. Indeed, the protection sense (i.e., the value of the binary variable y_i) could be both lower and upper for the same cell in two different tables (unless they are protected together, which is not possible if they are obtained at different moments).

A possible CTA-like approach satisfying all the above features could be based on a two-stages approach. The proposal is:

- *First stage.* Given an on-line table to be protected, the first stage would compute: (i) all the parameters of the CTA model (2) (i.e., set of sensitive cells, protection levels, cell bounds, and additivity constraints); (ii) and also the protection senses of sensitive cells (i.e., it would set $y_i, i \in \mathcal{S}$ either to 0 or 1 in (2)). We note that if sensitive cells are computed using the standard sensitivity rules (minimum frequency or p-% rules), then consistency on input is guaranteed; indeed, whether a cell is sensitive or not only depends (according to the above rules) on the particular respondents to this cell, not on the structure of the table being generated. Consistency on output would also be guaranteed if the fixed protection sense is always the same for this cell, independently of the table where it appears. Therefore, the protection sense should be fixed using only local information to the cell (as done by sensitivity rule). This is likely not the best approach, since some particular protection sense could make the second stage problem (see below) infeasible. An alternative would be to set y_i to 0 or 1 considering the constraints of the table where this cell appears for the first time, and recording (for future tables) the selected value y_i. However, this means to store in the on-line server information about protection senses for all sensitive cells, and in no way it guarantees that second stage problems of future tables will be feasible.

- *Second stage.* The basic problem to be solved in the second stage is (2) but considering binary variables y_i as parameters (they were fixed in the first stage). Since some combinations of y_i may make the resulting problem infeasible, it may be necessary to include extra variables and constraints to permit slight "violations" in bounds and constraints ("soft constraints"), and ultimately, in protection levels. This would guarantee the reliability of the approach. In addition, since binary variables are a priori fixed, the resulting problem would be a continuous LP or QP (if Euclidean distances or quadratic penalizations to violations are considered). This would satisfy efficiency, and very large tables could be protected in few seconds or minutes. For nonsensitive cells slightly different values could be released in different tables, unless the values of previously published cells are recorded in the on-line system (which could be prohibitive or impractical).

The above two-stage scheme is similar to that used in the field of stochastic programming (SP), with the (very important) difference that in SP the binary first stage decisions are selected to minimize (in average) all the second stage scenarios. In our case, the different second stage scenarios would correspond to all the possible on-line tables that can be released, which may be extremely large. In theory, all the possibles tables (or a sample of them) could be considered to a priori compute the "best" binary decisions, but the resulting problem would be a massive MILP

problem. This problem is interesting from a theoretical point of view, but perhaps it exceeds the capacities of on-line table generation systems.

5 Conclusions and future work

In this work we discussed two practical extensions to the RCTA package. The first one consists on implementing and including the BCD heuristic in the current package. Suboptimal, hopefully good, solutions could be then obtained in a fraction of the time needed by the optimal algorithms. The second extension is to develop a version of the package for CTA instances derived from on-line servers, implementing the two-stages procedure, which should provide a quick and reliable solution. Fortunately, since the resulting second-stage problem is continuous (either LP or QP), this problem is a priori simpler than the standard CTA. These extensions are planned to be done within the scope of the Data without Boundaries project.

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. (2011). A tool for analyzing and fixing infeasible rcta instances, *Lecture Notes in Computer Science* 6344, 17–28.

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. Available from the first author on request (`Ramesh.Dandekar@eia.doe.gov`).

Fraser, B., and Wooton, J. (2006). A proposed method for confidentialising tabular output to protect against differencing, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 299–302.

Giessing, S. (2011). German Federal Statistical Office. Personal communication.

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., Lenz, R., Naylor, J., Schulte Nordholt E., Seri, G., and De Wolf, P.-P. (2010), *Handook on Statistical Disclosure Control*, Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control. Available online at `http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf`.

Shlomo, N., and Young, C. (2008). Invariant post-tabular protection of census frequency counts, *Lecture Notes in Computer Science*, 5262, 77–89.