WP.16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

AN EXAMINATION OF TWO METHODS FOR CONTROLLED TABULAR ADJUSTMENT OF TABULAR DATA THAT PRESERVE DATA QUALITY

Invited Paper

Prepared by Lawrence H. Cox, National Center for Health Statistics, United States of America

An Examination of Two Methods for Controlled Tabular Adjustment of Tabular Data That Preserve Data Quality

Lawrence H. Cox

National Center for Health Statistics, Centers for Disease Control and Prevention
Hyattsville, MD 20782 USA
LCOX@CDC.GOV

Keywords. Linear programming, iterative proportional fitting, Kullback-Leibler, QP-CTA, MDI-CTA

Abstract. Two methods for balancing data quality and data confidentiality for tabular data have recently emerged. Each limits disclosure through controlled tabular adjustment (CTA). The first, Quality-Preserving (QP-) CTA, achieves CTA through a linear programming model, and preserves quality by controlling change to selected parameters and statistics associated with the original distribution via specialized capacities and constraints incorporated into the linear program. The second, Minimum Discrimination Information (MDI-) CTA, is an iterative procedure that at each stage employs iterative proportional fitting (IPF) to achieve a current CTA solution exhibiting minimum Kullback-Leibler distance (MDI) from the original distribution (table) for fixed inputs, and at the next stage refines the inputs and repeats the IPF in an attempt to produce a next CTA solution exhibiting smaller MDI. Based on limited testing, both methods appear to perform well on a practical basis. The strengths and limitations of the two methods are often opposite to each other. This paper explores these properties with an eye towards an enhanced CTA methodology.

1 Introduction

Tabular data are ubiquitous. Standard forms include count data as in population and health statistics, concentration or percentage data as in financial or energy statistics, and magnitude data such as retail sales in business statistics or average daily air pollution in environmental statistics. Tabular data remain a staple of official statistics. Data confidentiality was first investigated for tabular data [1, 2]. Tabular data are additive and expressible as specialized systems of linear equations: $\mathbf{Tx} = \mathbf{v}$, where $\mathbf{x}$ represents the *tabular cells*, $\mathbf{T}$ the *tabular equations*, and $\mathbf{v}$ fixed values. Entries of $\mathbf{T}$ are in the set {-1, 0, +1}, and each row of $\mathbf{T}$ contains at most one -1.

For decades, the prevailing *statistical disclosure limitation* (SDL) [3] method for tabular magnitude data and, to a lesser but considerable extent, tabular count data, was *complementary cell suppression* (CCS) [1, 4, 5]. Recently, an alternative SDL method named *controlled tabular adjustment* (CTA) has emerged [6, 7]. This development was motivated by computational complexity, analytical obstacles, and user dissatisfaction with CCS [8].

Disclosure in tabular data is typically defined by a *threshold rule* (count data) or a *dominance rule* (magnitude data), and more generally by a *linear sensitivity measure* [9]. Counts of 1 or 2 are likely to identify individuals, or sales data may be dominated by contributions of n = 1 or 2 contributors, and thereby provide a close estimate of the larger contributor's confidential

business information to the other contributor (n = 2) or to the public (n = 1). The sensitivity measure also determines minimal distances from the sensitive cell value to *safe values* above and below it. The *protection interval* for the sensitive cell value is defined to be the set of all unsafe values—values between its lower and upper safe values. See [9] for details.

Complementary cell suppression removes from publication the values of all sensitive cells and in addition removes sufficiently many nonsensitive cell values to ensure that the linear system $\mathbf{Tx} = \mathbf{v}$ does not reveal a sensitive cell value or locate it within an interval finer than its protection interval ([1, 9]). Drawbacks of cell suppression for statistical analysis include removal of useful and otherwise harmless information and consequent difficulties analyzing tabular systems with cell values missing not-at-random.

Controlled tabular adjustment replaces sensitive cell values with safe values and, because these adjustments almost certainly throw the tabular system $\mathbf{Tx} = \mathbf{v}$ out of kilter, CTA adjusts some or all of the nonsensitive cells by small amounts to rebalance the additive system. In terms of ease-of-use, controlled tabular adjustment is unquestionably an improvement over cell suppression. As CTA changes sensitive and other cell values, the data quality issue is then: Can CTA be accomplished while preserving important data analytical properties of original data?

A preliminary discussion of quality aspects of CTA was introduced in [8]. A methodology for preserving distributional parameters of linear models for univariate distributions was introduced in [6], and named *quality-preserving controlled tabular adjustment* (QP-CTA). In [10], QP-CTA was extended to multivariate distributions. QP-CTA is based on mathematical (mostly, linear) programming. A statistical approach based on iterative proportional fitting to preserve the original data distribution, as measured by Kullback-Leibler distance ([11]) between the original and adjusted tables, was introduced in [12], and named *minimum discrimination information controlled tabular adjustment* (MDI-CTA).

QP-CTA and MDI-CTA each appear to perform well and efficiently based on limited testing. The strengths and limitations of the two methods are in many cases opposite to each other. The purpose of this paper is to examine and compare these properties, with an eye towards an enhanced methodology that exploits their combined strengths. In Section 2, we specify the two methods mathematically. In Section 3, we examine their quality characteristics and, in Section 4, compare them. Section 5 provides concluding comments.

2 Controlled Tabular Adjustment and Data Quality

2.1 The Basic CTA Methodology

CTA is applicable to all tabular data, but for convenience we focus on magnitude data. A simple paradigm for statistical disclosure in magnitude data follows. Tabulation cell i comprises k respondents (e.g., retail clothing stores in a county) and a statistic of interest (e.g., retails sales). The NSO assumes that any respondent is aware of the identity of the other respondents. The cell value is the total value of the statistic of interest, viz., the sum of nonnegative values of this statistic (called *contributions*) by each respondent in the cell. Denote the cell value $v^{(i)}$ and the contributions $v_j^{(i)}$, ordered from largest to smallest. It is possible for any respondent j to

compute $v^{(i)} - v_j^{(i)}$ which is an upper estimate of the contribution of any other respondent. This estimate is closest, in percentage terms, when the target is the largest respondent and j = 2. The *p-percent rule* declares that the cell value represents disclosure whenever this estimate is less than (100 + p)-percent of the largest contribution. The sensitive cells are those failing this condition. This is a standard rule and has a corresponding linear sensitivity measure ([9]).

The NSO may also assume that any respondent can use public knowledge to estimate the contribution of any other respondent to within q-percent (q > p, e.g., q = 50%). This information allows the second largest to estimate $v^{(i)} - v_1^{(i)} - v_2^{(i)}$ to within q-percent. This upper estimate provides the second largest a lower estimate of $v_1^{(i)}$. This is referred to as the *p/q-ambiguity rule*, also associated with a linear sensitivity measure ([9])

The *lower* and *upper protection limits* for the cell value equal, respectively, the minimum amount that must be subtracted from (added to) the cell value so that these lower (upper) estimates are at least p-percent away from the true value $v_1^{(i)}$. Numeric values outside the protection limit range of the true value are its safe values. A common NSO practice is to assume that both protection limits are equal to a common value p_i. Complementary cell suppression suppresses all sensitive cells from publication, replacing sensitive values by variables in the tabular system **Tx = v**. Because, almost surely, one or more suppressed sensitive cell values can be estimated via linear programming to within its unsafe range, it is necessary to suppress additional nonsensitive cell values until no sensitive estimates can be obtained. This yields a mixed integer linear programming (MILP) problem over binary suppression variables ([4, 5]).

Controlled tabular adjustment replaces each sensitive value with a safe value. This is an improvement over complementary cell suppression as it replaces a suppression symbol by an actual value. However, safe values are not necessarily unbiased estimates of true values. To minimize bias, it is often desirable to replace the true value by either of its nearest safe values, $v^{(i)} - p_i$ or $v^{(i)} + p_i$. Because these assignments almost surely throw the tabular system out of kilter, CTA adjusts nonsensitive values to restore additivity. Because choices to adjust each sensitive value down or up are binary, combined these steps define a MILP. Typically, its *linear programming relaxation* is solved. Even as a MILP, CTA is superior to CCS because, for CCS, the MILP assigns one binary variable to each nonsensitive cell while, for CTA, binary variables are assigned to the sensitive cells, which almost certainly are far fewer in number.

A MILP by itself will not assure that analytical properties of original and adjusted data are comparable. Three simple procedures aimed at preserving quality were introduced in [8]. First, sensitive values are replaced by nearest safe values to reduce statistical bias. Second, lower and upper bounds (*capacities*) are imposed on adjustments to nonsensitive values to keep individual adjustments acceptably small, e.g., capacities might be based on estimated cell value measurement error e_i. Third, the objective function for the linear program is an overall measure of data distortion such as minimum sum of absolute, or percent, adjustments. An (arbitrarily large) upper bound on adjustment of sensitive cell value v_i is denoted m_i. The MILP model for CTA subject to the second and third criteria--while relaxing the first--is as follows ([10]).

Assume there are *n* tabulation cells of which the first *s* are sensitive, original data are represented by the nx1 vector **a**, adjusted data by $a + y^+ - y^-$; and $y = y^+ - y^-$. The MILP is:

$$\min \sum_{i=1}^{n} (y_i^+ + y_i^-) \quad \text{subject to:}$$

$$T\,y = 0 \tag{1}$$

$$p_i(1 - I_i) \le y_i^- \le m_i(1 - I_i), \quad p_i I_i \le y_i^+ \le m_i I_i \;;\; I_i \;\; \textit{binary} \qquad i = 1,\ldots s$$

$$0 \le y_i^-,\; y_i^+ \le e_i \qquad i = s+1, \ldots, n$$

If the capacities on adjustments to nonsensitive cells are too tight, it is possible that problem (1) be *infeasible* (lack solutions), requiring that some capacities be increased. A companion strategy, allows sensitive cell adjustments smaller than p_i in well-defined situations. This is justified mathematically because the intruder does not know if the adjusted value lies above or below the original value, but nevertheless is perceived by some as controversial.

Problem (1) is a MILP. The integer part can be solved by exact methods for small to medium-sized problems or via heuristics which first fix the integer variables and subsequently use linear programming to solve the linear programming relaxation. The remainder of this paper focuses on the problem of preserving data quality under CTA, and is not concerned with how the integer portion is being or has been solved.

2.2 QP-CTA: Using CTA to Preserve Parameters of Linear Models

For univariate data, we seek to preserve approximately mean and variance of original data and also correlation and regression slope between original and adjusted data, while maintaining additivity. For multivariate data, in addition adjusted data should preserve approximately covariance, correlation and regression coefficients between variables in original data.

Preserving mean values is straightforward. Any cell value a_i can be held fixed by forcing its corresponding adjustment variables y_i^+, y_i^- to zero, viz., set each variable's upper capacity to zero. Means are averages over sums. So, for example, to fix the grand mean, simply fix the grand total. Or, to fix means over all or a selected set of rows, columns, etc., in the tabular system, simply capacitate to zero adjustments to the corresponding totals.

In [6], it is shown how data quality objectives---variance, correlation and regression slope--- can be achieved by forcing covariance between original data **a** and adjustments **y** to them to be close to zero, while preserving the corresponding means, viz., over the indices comprising the mean, sum of upper adjustments equals sum of lower adjustments. Define

$$L(y) = \mathrm{Cov}(a, y)/\mathrm{Var}(a) \tag{2}$$

For any (renumbered) subset of *t* cells, as $\bar{y} = 0$: $L(y) = (1/(t\mathrm{Var}(a)))\sum_{i=1}^{t}(a_i - \bar{a})\,y_i$.

For variance, $Var(a + y) = (1/t)(\sum((a_i + y_i - (\bar{a} + \bar{y}))^2) = \mathrm{Var}(a) + (2/t)\sum(a_i - \bar{a})y_i + \mathrm{Var}(y)$.

So, $|Var(\text{a + y})/\text{Var(a)} - 1| = |2\text{L(y)} + (\text{Var(y)}/\text{Var(a)})|$ and relative change in variance can be minimized by minimizing the right-hand side. As $Var(\text{y})/\text{Var(a)}$ is typically small, it suffices to minimize $|L(\text{y})|$. This is accomplished by:

a) adjoin two additional linear constraints to (1):
$w \geq L(\text{y})$, $w \geq -L(\text{y})$, and (3)
b) minimize w

For univariate correlation, we seek Corr (a, a + y) = 1 approximately. As $\bar{y} = 0$,

$Corr(a,a+y) = Cov(a,a+y) / \sqrt{Var(a)Var(a+y)} = (1 + L(y)) / \sqrt{\text{Var(a + y)/ Var(a)}}$

As Var(y)/Var(a) is typically small, $\min |L(\text{y})|$ will typically suffice.

Finally, to preserve ordinary least squares regression $Y = \beta_1 X + \beta_0$ of adjusted data Y = a + y on original data X = a, we want β_1 near one and β_0 near zero:

$$\beta_1 = Cov(\text{a + y, a}) / \text{Var(a)} = 1 + \text{L(y)},\ \beta_0 = (\bar{\text{a}} + \bar{\text{y}}) - \beta_1 \bar{\text{a}}$$

As $\bar{y} = 0$, then $\beta_0 = 0$, $\beta_1 = 1$ if L(y) = 0 is feasible. Again, min $|L(\text{y})|$ suffices.

For multivariate data, in place of a single data set organized in tabular form, viz., **Ta = v**, to which adjustments **y** are to be made for confidentiality purposes, we have multiple data sets, each organized within a common tabular structure **T**. This is typical in official statistics where, e.g., tabulations would be shown at various levels of geography and industry classification for a range of variables such as total retail sales, cost of goods, number of employees, etc.

For concreteness, we focus on the bivariate case. Original data are denoted **a**, **b** and corresponding adjustments to original values are denoted by variables **y** and **z**. In the univariate case, the key to preserving variance, correlation and regression slope was to force Cov (**a, y**) = 0. It is easy to overlook in the univariate case that as Var (**a**) = Cov (**a, a**), then preserving variance via Cov (**a, y**) = 0 is equivalent to requiring Cov (**a, a + y**) = Cov (**a, a**). In the multivariate situation, however, preserving covariance (and variance) is of key importance and not to be overlooked. Namely, if we can preserve mean values and the variance-covariance matrix of original data, then we have preserved essential properties of the original data, particularly in the case of linear statistical models. We also would like to preserve simple linear regression of original data **b** on original data **a** in the adjusted data. And, of course, we wish to preserve the univariate properties of each variable.

To preserve Cov (**a, b**), we require: Cov (a, b) = Cov (a + y, b + z) = Cov (a, b) + Cov (a, z) + Cov (b, y) + Cov (y, z). Consequently, we seek:

min {|Cov (**a, z**) + Cov (**b, y**) + Cov (**y, z**)|} , subject to (1, 2, 3) (4)

The last term in the objective function is quadratic. For some problems, use of quadratic programming would be acceptable computationally. A linear approach to solving (4) heuristically is: perform successive alternating linear optimizations, viz., solve (2) for $\mathbf{y} = \mathbf{y_0}$, substitute $\mathbf{y}_0$ into (4) and solve for $\mathbf{z} = \mathbf{z_0}$, and continue in this fashion until an acceptable solution is reached.

The next objective is to preserve the estimated regression coefficient under simple linear regression of **b** on **a**. We seek approximately:

Cov (a, b) / Var (a) = Cov (a + y, b + z) / Var (a + y)

Var (a + y) / Var (a) = Cov (a + y, b + z) / Cov (a, b)

= 1 + Cov (a, z) / Cov (a, b) + Cov (b, y) / Cov (a, b) + Cov (y, z) / Cov (a, b)

Observe: $\mathrm{Var}(a + y)/\mathrm{Var}(a) = 2L(y) + 1 + \mathrm{Var}(y)/\mathrm{Var}(a)$

$2L(y) + \mathrm{Var}(y)/\mathrm{Var}(a) = \mathrm{Cov}(a, z) / \mathrm{Cov}(a, b) + \mathrm{Cov}(b, y) / \mathrm{Cov}(a, b) + \mathrm{Cov}(y, z) / \mathrm{Cov}(a, b)$

To preserve the regression coefficient, then we solve the linear program:

$$\min |(\mathrm{Cov}(\mathbf{a}, \mathbf{z}) + \mathrm{Cov}(\mathbf{b}, \mathbf{y})) / \mathrm{Cov}(\mathbf{a}, \mathbf{b})|, \text{ subject to (4)} \qquad (5)$$

To preserve correlation, we seek: Corr (**a, b**) = Corr (**a + y, b + z**). Equivalently:

$$\sqrt{\frac{Var(a+y)}{Var(a)}}\sqrt{\frac{Var(b+z)}{Var(b)}} = \frac{Cov(a+y,b+z)}{Cov(a,b)}$$

2.3 MDI-CTA: Achieving CTA While Preserving the Distributions Subject to Kullback-Leibler Divergence

Kullback-Leibler minimum discrimination information (MDI) ([11]) is a measure of distance between two statistical distributions. Often, the first distribution is known and the unknown second distribution is the closest distribution to the first within a predefined class of distributions, where "close" is measured by MDI. In our setting, the first distribution is the original distribution (table) and the class is the set of tables satisfying prespecified fixed marginal totals (*minimal sufficient statistics* = MSS). It is well-known that the iterative proportional fitting (IPF) procedure can be used to compute a unique solution that minimizes MDI. IPF permits fixing a subset of the cell values. In our setting, this subset includes sensitive cells set at selected safe values and structural zeroes. In [12], it is shown how to apply IPF to preserve distributions under CTA, as follows.

CTA fixes the values of the sensitive cells to safe values. Typically, these values are set equal to either the maximum lower or minimum upper safe value, viz., $v^{(i)} - p_i$ or $v^{(i)} + p_i$. This results in a binary choice for each sensitive cell, resulting in 2^s possible choices. Conditional on any one of these choices and on fixed MSS, it is possible to compute the IPF solution. The IPF solution maintains additivity to the MSS and, conditional on choices for the safe values, minimizes MDI relative to the original table. Based on a heuristic algorithm for updating choices of safe values to improve the MDI solution, the procedure of [12] iterates the IPF until the difference in MDI between the original and adjusted tables is statistically insignificant.

3 Data Quality Characteristics of the Two CTA Methods

In [13], two broad classes of data quality indicators for tabular data were introduced—local quality and global quality. *Local quality* refers to preserving or remaining close to original data values and relationships between them. *Global quality* refers to preserving the original distribution, its properties, and characteristics. A third category, arguably subsumed under both, is *structural quality*, e.g., preserving additivity.

3.1 Characteristics of QP-CTA

Operational characteristics of QP-CTA are as follows.

Pro

- preserves additivity
- relies on standard linear programming software
- capacities, constraints and objective are easily modified
- typically computationally efficient
- applicable to arbitrary tabular structure, dimension, and size
- can be performed in a multivariate setting
- does not require that marginal totals be fixed

Con

- can be solved exactly for certain structures, dimension, and size, but typically relies on heuristics to assign safe values to the sensitive cells
- objective function(s) and heuristics not tied to statistical criteria

Data quality characteristics of QP-CTA are as follows.

Local Quality

- adjustments to individual sensitive cells can be minimized
- adjustments to individual nonsensitive cells can be limited in size
- structural zeroes and other selected cell values can be exempt from change
- nonstructural zero cells can be adjusted away from zero

Global Quality

- can minimize global distance or average distance
- preserves univariate properties: mean, variance, correlation, regression
- preserves multivariate properties: covariance, regression
- can preserve these quantities for arbitrarily defined subsets of cells

3.2 Characteristics of MDI-CTA

Operational characteristics of MDI-CTA are as follows.

Pro

- preserves additivity
- relies on standard statistical algorithms available as software
- typically computationally efficient
- objective function(s) and heuristics tied to statistical criteria

Con

- relies on a heuristic for assigning safe values to sensitive cells
- not easily applied to arbitrary tabular structure
- no clear generalization to a multivariate setting
- requires (some) marginal totals to be fixed

Data quality characteristics of MDI-CTA are as follows.

Local Quality

- structural zeroes and other selected cell values can be exempt from change

Global Quality

- preserves original distribution, not just selected parameters or statistics
- nonstructural zero cells remain fixed at zero

4 Comparison of QP-CTA and MDI-CTA for Preserving Data Quality

Global data quality is concerned with preserving the original distribution and its properties. MDI-CTA preserves the original distribution conditional on the safe values. That is, if the number of sensitive cells is small relative to the total number of cells, and if safe values are not exceptionally large relative to nonsensitive values and estimated measurement errors, then it is reasonable to expect that MDI-CTA will preserve the original distribution. Whether this is the case or not can be verified by computing MDI or another test statistic to detect a statistically significant distance between original and adjusted data. If MDI-CTA has preserved the original distribution, then it is reasonable to expect that it also preserved important distributional parameters and statistics. This is not guaranteed but also can be verified. When releasing the adjusted data, it would be useful for the NSO also to release estimates of these quantities computed from original data.

QP-CTA preserves means, variances, covariances and regressions, so it is reasonable to expect that original and adjusted distributions are not too far apart, conditional on the safe values. Furthermore, if it is possible to limit adjustments to nonsensitive cells to within estimated measurement error, and if sensitive cells are adjusted to values at or near minimal safe values, then, conditional on the safe values, it is reasonable to expect that original and QP-CTA adjusted distributions are similar. This is not guaranteed, but can be verified by testing for a statistically significant MDI between original and adjusted tables.

Local data quality is concerned with changes to individual cell values and relationships between them. QP-CTA preserves local data quality directly via capacity constraints on adjustments to individual cell values, and in addition preserves covariances, correlations and regressions. Both QP-CTA and MDI-CTA can exempt selected values, including structural zeroes, from change. However, MDI-CTA does not control local changes to nonexempt cells, and there does not appear to be a way to modify IPF to incorporate capacity constraints. QP-CTA is able to adjust nonstructural zeroes away from zero. Replacing nonstructural zeroes with small “epsilon” values would enable MDI-CTA to do likewise.

Both methods preserve additivity to marginal totals. In some applications, marginal totals can be sensitive, and therefore adjusted, and in addition likely to require adjustment of additional marginal totals. In other cases, it may be desirable to permit adjustment of all or some nonsensitive marginal totals, e.g., to improve global quality. Conversely, if no marginal totals are sensitive, it may be desirable not to adjust any marginal total, e.g., when totals have been published previously. QP-CTA enables all of these choices. Currently, MDI-CTA does not enable any of them. This could be overcome if a set of MSS involving only nonsensitive marginals is identified. IPF then is performed based on the MSS, and each marginal total set equal to the sum of its constituent internal entries. As designed, MDI-CTA applies easily to a single, standard multi-dimensional table but not to arbitrary tabular structures.

Both methods employ heuristics for selecting the safe values. More work, e.g., [14], is needed on developing appropriate, effective heuristics. Regarding assessing goodness of fit, whereas MDI is convenient to perform the CTA and preserve the original distribution, it is not clear what is the most appropriate test statistic and work on that is needed also. Research into ways to combine these rather different methods into a stronger, combined method is indicated.

5 Concluding Comments

QP-CTA and MDI-CTA are two methods based on controlled tabular adjustment for producing a quality-preserving, disclosure-limited set of tabulations from an original set of tabulations that contains disclosure. We have presented mathematical/statistical models for applying these methods and examined their respective strengths and weaknesses operationally and for preserving data quality. We observed that often a strength of one method is a weakness of the other, and vice-versa, which motivated our comparison of their respective quality characteristics.

Disclaimer This paper represents the work of the author and is not intended to represent the policies or practices of the Centers for Disease Control and Prevention or any other organization.

References

1. Cox, L.H.: Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association. 75(1980) 377-385
2. Fellegi, I.P.: On the Question of Statistical Confidentiality. Journal of the American Statistical Association. 67 (1972) 7-18
3. U.S. Department of Commerce.: Statistical Disclosure and Disclosure Limitation Methods, Statistical Policy Working Paper 22, Washington, DC: Federal Committee on Statistical Methodology.(1994)
4. Cox, L.H.: Network Models for Complementary Cell Suppression. Journal of the American Statistical Association. 90(1995) 1153-1162.
5. Fischetti, M. and J.J. Salazar-Gonzalez: Models and Algorithms for Optimizing

Cell Suppression in Tabular Data with Linear Constraints. Journal of the American Statistical Association. 95 (2000) 916-928.
6. Cox, L.H. and J.P. Kelly: Balancing Data Quality and Confidentiality for Tabular Data. Proceedings of the UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003, Monographs of Official Statistics. Luxembourg: Eurostat (2003) 11-23.
7. Cox, L.H.: Discussion. ICES II: The Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions. Alexandria, VA: American Statistical Association. (2000) 905-907.
8. Cox, L.H. and R.A. Dandekar: A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use. Proceedings of the 2002 FCSM Statistical Policy Seminar, Washington, DC: U.S. Office of Management and Budget (2003) http://www.fcsm.gov/working-papers/wp35_1.pdf
9. Cox, L.H.: Linear Sensitivity Measures in Statistical Disclosure Control. Journal of Statistical Planning and Inference 5 (1981) 153-164.
10. Cox, L.H., J.P. Kelly and R. Patil: Balancing Quality and Confidentiality for Multivariate Tabular Data. in: Privacy in Statistical Databases, Lecture Notes in Computer Science 3050 (J. Domingo-Ferrer and V. Torra, eds.), Berlin: Springer-Verlag (2004) 87-98.
11. Kullback, S. and R.A. Leibler: On Information and Efficiency. Annals of Mathematical Statistics, Volume 86 (1951) 79-86.
12. Cox, L.H., J.G. Orelien and B.V. Shah: A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment. in: **P**rivacy and Statistical Data Bases 2006, Lecture Notes in Computer Science 4302 **(**J. Domingo-Ferrer and L. Franconi, eds.), Heidelberg: Springer-Verlag (2006) 1-11.
13. Cox, L.H.: Balancing Quality and Confidentiality of Statistical Data. Proceedings of the 56th Session of the International Statistical Institute: Invited Papers, Voorburg: International Statistical Institute (2007), CD-ROM.
14. Glover, F., L.H. Cox, J.P. Kelly and R. Patil: Exact, Heuristic and Metaheuristic Methods for Confidentiality Protection by Controlled Tabular Adjustment, International Journal of Operations Research (2007), under review.