

# Balancing Quality and Confidentiality for Multivariate Tabular Data

Lawrence H. Cox<sup>1</sup>, James P. Kelly<sup>2</sup>, and Rahul Patil<sup>2</sup>

<sup>1</sup> National Center for Health Statistics, Centers for Disease Control and Prevention  
Hyattsville, MD 20782 USA

<sup>2</sup> OptTek Systems, Inc., Boulder, CO 80302 USA

**Abstract.** Absolute cell deviation has been used as a proxy for preserving data quality in statistical disclosure limitation for tabular data. However, users' primary interest is that analytical properties of the data are for the most part preserved, meaning that the values of key statistics are nearly unchanged. Moreover, important relationships within (additivity) and between (correlation) the published tables should also be unaffected. Previous work demonstrated how to preserve additivity, mean and variance in for univariate tabular data. In this paper, we bridge the gap between statistics and mathematical programming to propose nonlinear and linear models based on constraint satisfaction to preserve additivity and covariance, correlation, and regression coefficient between data tables. Linear models are superior than nonlinear models owing to simplicity, flexibility and computational speed. Simulations demonstrate the models perform well in terms of preserving key statistics with reasonable accuracy.

**Keywords:** Controlled tabular adjustment, linear programming, covariance

## 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 thus naturally related to specialized systems of linear equations:  $\mathbf{TX} = \mathbf{0}$ , where  $\mathbf{X}$  represents the *tabular cells* and  $\mathbf{T}$  the *tabular equations*, the entries of  $\mathbf{T}$  are in the set  $\{-1, 0, +1\}$ , and each row of  $\mathbf{T}$  contains precisely one -1.

In [3], Dandekar and Cox introduced a methodology for *statistical disclosure limitation* [4] in tabular data known as *controlled tabular adjustment* (CTA).

This development was motivated by computational complexity, analytical obstacles, and general user dissatisfaction with the prevailing methodology, *complementary cell suppression* [1, 5]. Complementary suppression removes from publication all *sensitive cells* – cells that cannot be published due to confidentiality concerns – and in addition removes other, nonsensitive cells to ensure that values of sensitive cells cannot be reconstructed or closely estimated by manipulating linear tabular relationships. Drawbacks of cell suppression for statistical analysis include removal of otherwise useful information and consequent difficulties analyzing tabular systems with cell

values missing not-at-random. By contrast, the CTA methodology replaces sensitive cell values with *safe values*, viz., values sufficiently far from the true value. Because the adjustments almost certainly throw the additive tabular system out of kilter, CTA adjusts some or all of the nonsensitive cells by small amounts to restore additivity. CTA is implemented using mathematical programming methods for which commercial and free software are widely available.

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 issue is then: Can CTA be accomplished while preserving important data analytical properties of original data?

In [6], Cox and Dandekar describe how the original CTA methodology can be implemented with an eye towards preserving data analytic outcomes. In [7], Cox and Kelly demonstrate how to extend the mathematical programming model for CTA to preserve univariate properties of original data important to linear statistical models. In this paper, we extend the Cox-Kelly paradigm to the multivariate case, ensuring that covariances, correlations and regression coefficients between original variables are preserved in adjusted data. Specialized search procedures, including Tabu Search [8], can also be employed in formulations proposed in [7]. While it is easy to formulate nonlinear programming (NLP) models to do this, such models present difficulties in understanding and use in general statistical settings and exhibit computational limitations. The NLP models are computationally more expensive than linear programming (LP) models because the LP algorithms and re-optimization processes are much more efficient compared to NLP algorithms. Moreover, LP solvers guarantee global optimality, whereas to NLP models cannot guarantee global optimality for non-convex problems. Models presented here are based on linear programming and consequently are easy to develop and use and are applicable to a very wide range of problems types and sizes.

Section 2 provides a summary of the original CTA methodology of [3] and linear methods of [7] in the univariate case for preserving means and variances of original data and for ensuring high correlation between original and adjusted data. Section 3 provides new results addressing the multivariate case, providing linear programming formulations that ensure covariance, correlation and regression coefficient between two original variables exhibited in original data are preserved in adjusted data. Section 4 reports computational results. Section 5 provides concluding comments.

## 2 Controlled Tabular Adjustment and Data Quality for Univariate Data

### 2.1 The Original CTA Methodology

CTA is applicable to tabular data in any form but for convenience we focus on magnitude data, where the greatest benefits reside. A simple paradigm for statistical disclosure in magnitude data is as follows. A tabulation cell, denoted  $i$ , comprises  $k$  respondents (e.g., retail clothing stores in a county) and their data (e.g., retail sales and employment data). It is assumed that each respondent knows the identity of the other respondents. The *cell value* is the total value of a statistic of interest (e.g., total retail

sales), summed over nonnegative *contributions* of each respondent in the cell to this statistic. Denote the cell value  $v^{(i)}$  and the respondent contributions  $v_j^{(i)}$ , ordered from largest to smallest. It is possible for any respondent  $j$  to compute  $v^{(i)} - v_j^{(i)}$  which yields 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  $j = 2$ . A standard disclosure rule, 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.

We also may 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 additional information allows the second largest to estimate  $v^{(i)} - v_1^{(i)} - v_2^{(i)}$ , the sum of all contributions excluding itself and the largest, to within  $q$ -percent. This upper estimate provides the second largest a lower estimate of  $v_1^{(i)}$ . 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 response  $v_1^{(i)}$ . Numeric values outside the protection limit range of the true value are *safe values* for the cell. A common practice assumes that these protection limits are equal, to  $p_i$ . Complementary cell suppression suppresses all sensitive cells from publication, replacing sensitive values by variables in the tabular system  $\mathbf{TX} = \mathbf{0}$ . 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 some nonsensitive cells until no sensitive estimates are obtainable. This yields a mixed integer linear programming (MILP) problem, as in [9].

The original controlled tabular adjustment methodology [3] 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, [3] replaces the true value by either of its nearest safe values,  $v^{(i)} - p_i$  or  $v^{(i)} + p_i$ . Because this will 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 [10]. Dandekar and Cox [3] present heuristics for the binary choices. The resulting *linear programming relaxation* is easily solved.

A (mixed integer) linear program in itself will not assure that analytical properties of original and adjusted data are comparable. Cox and Dandekar [6] address these issues in three ways. First, sensitive values are replaced by nearest safe values to reduce statistical bias. Second, lower and upper bounds (*capacities*) are imposed on changes to nonsensitive values to ensure adjustments to individual datum are acceptably small. Statistically sensible capacities would, e.g., be based on estimated measurement error for each cell  $e_i$ . Third, the linear program optimizes an overall measure of data distortion such as minimum sum of absolute adjustments or minimum sum of percent absolute adjustments. The MILP model is as follows.

Assume there are  $n$  tabulation cells of which the first  $s$  are sensitive, original data are represented by the  $n \times 1$  vector  $\mathbf{a}$ , adjusted data by  $\mathbf{a} + \mathbf{y}^+ - \mathbf{y}^-$ ; and  $\mathbf{y} = \mathbf{y}^+ - \mathbf{y}^-$ . The MILP of [10] corresponding to the methodology of [3] for minimizing sum of absolute adjustments is:

$$\begin{aligned} & \min \sum_{i=1}^n (y_i^- + y_i^+) \\ \text{subject to:} & \quad \mathbf{T}(\mathbf{y}) = 0 \tag{1} \\ & y_i^- = p_i(I - I_i), \quad y_i^+ = p_i I_i, \quad I_i \text{ binary, } i = 1, \dots, s \\ & 0 \leq y_i^-, y_i^+ \leq e_i \quad i = s+1, \dots, n \end{aligned}$$

If the capacities are too tight, this problem may be *infeasible* (lack solutions). In such cases, capacities on nonsensitive cells may be increased. A companion strategy, albeit controversial, 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; see [6] for details.

The constraints used in [6] are useful. Unfortunately, choices for the optimizing measure are limited to linear functions of cell deviations. In the next two sections, we extend this paradigm in two separate directions, focusing on approaches to preserving mean, variance, correlation and regression coefficient between original and adjusted data.

Formulation (1) is a mixed integer linear program, the integer part can be solved by exact methods in 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 [3,11]. Cox and Kelly [11] propose different heuristics to fix the binary variables and to reduce the number of the binary variables in order to improve the computational efficiency, and report good solutions in reasonable time. 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. For that reason, for convenience we occasionally abuse terminology and refer to (1) and subsequent formulations as “linear programs”.

## 2.2 Using CTA to Preserve Univariate Relationships

We present linear programming formulations of [7] for preserving approximately mean, variance, correlation and regression slope between original and adjusted data while preserving additivity.

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 capitate changes to the corresponding totals to zero. To fix the mean of any set of variables for which a corre-

sponding variable has not been defined, incorporate a new constraint into the linear system:  $\sum_i (y_i^+ - y_i^-) = 0$ , where the sum is taken over the set of variables of interest.

To ensure feasibility and allow a richer set of potential solutions, it is useful to allow adjustments to sensitive cells to vary a bit beyond nominal protection limits, using capacities. The corresponding MILP is:

$$\begin{aligned}
 & \min c(\mathbf{y}) \\
 \text{subject to: } & \mathbf{T}(\mathbf{y}) = 0 \tag{2} \\
 & \sum_{ii} (y_i^+ - y_i^-) = 0 \\
 & q_i(1 - I_i) \leq y_i^- \leq p_i(1 - I_i), \quad q_i I_i \leq y_i^+ \leq p_i I_i ; \quad I_i \text{ binary } \quad i=1, \dots, s \\
 & 0 \leq y_i^-, y_i^+ \leq e_i \quad \quad \quad i = s+1, \dots, n
 \end{aligned}$$

$c(\mathbf{y})$  is used to keep adjustments close to their lower limit, e.g.,  $c(\mathbf{y}) = \sum(\mathbf{y}^+ + \mathbf{y}^-)$ .

Cox and Kelly [7] demonstrate that the data quality objectives---variance, correlation and regression slope---can be realized by forcing the covariance between original data  $\mathbf{a}$  and adjustments to original data  $\mathbf{y}$  to be close to zero while preserving the corresponding mean value(s). For variance, any subset of cells of size  $t$  with  $\bar{y} = 0$ ,

$$\begin{aligned}
 \text{Var}(\mathbf{a} + \mathbf{y}) &= (1/t)(\sum((a_i + y_i - (\bar{a} + \bar{y})))^2) \\
 &= \text{Var}(\mathbf{a}) + (2/t)\sum(a_i - \bar{a})y_i + \text{Var}(\mathbf{y})
 \end{aligned}$$

Define  $L(\mathbf{y}) = \text{Cov}(\mathbf{a}, \mathbf{y})/\text{Var}(\mathbf{a})$ . As  $\bar{y} = 0$ ,

then  $L(\mathbf{y}) = (1/(t\text{Var}(\mathbf{a}))) \sum_{i=1}^t (a_i - \bar{a})y_i$ , so

$$\begin{aligned}
 \text{Var}(\mathbf{a} + \mathbf{y})/\text{Var}(\mathbf{a}) &= 2L(\mathbf{y}) + (1 + \text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})) \quad \text{and} \\
 |\text{Var}(\mathbf{a} + \mathbf{y})/\text{Var}(\mathbf{a}) - 1| &= |2L(\mathbf{y}) + (\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a}))|
 \end{aligned}$$

Thus, relative change in variance can be minimized by minimizing the right-hand side. As  $\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})$  is typically small, it suffices to minimize  $|L(\mathbf{y})|$  or at least to reduce it to an acceptable level. This can be accomplished as follows:

- a) Incorporate two new linear constraints into the system (2):
 
$$w \geq L(\mathbf{y}), \quad w \geq -L(\mathbf{y}) \tag{3}$$
- b) Minimize  $w$ .

Using standard methods (e.g., constraining  $\mathbf{w}$  and  $-\mathbf{w}$  to be less than a small quantity) we treat (3) as a set of constraints leaving us free to specify a convenient linear objective function, such as sum of absolute adjustments,  $\sum y_i$ . In such cases, we say that we are setting the corresponding linear functional (e.g.,  $L(y)$ ) to a numeric value (typically, zero) “exactly or approximately”.

As  $\bar{y} = 0$ , then  $Var(y) = \sum y_i^2$ , minima of which are minima of sum of absolute adjustments,  $\sum |y_i|$ . Thus, an improved alternative to (3) for preserving variance would be to solve (2) for minimum sum absolute adjustments, compute minimum variance of adjustments, and set  $L(y) = -\min Var(y)/Var(a)$  exactly or approximately. This works well if we are interested in preserving mean and variance only. Formulation (3) is useful below for preserving correlation and regression slope, and consequently may be favored in most applications.

Regarding correlation, the objective is to achieve high positive correlation between original and adjusted values. We seek  $Corr(a, a + y) = 1$ , exactly or approximately. As  $\bar{y} = 0$ ,

$$\begin{aligned} Corr(a, a + y) &= Cov(a, a + y) / \sqrt{Var(a)Var(a + y)} \\ &= (1 + L(y)) / \sqrt{Var(a + y) / Var(a)} \end{aligned}$$

With  $Var(y)/Var(a)$  typically small, the denominator should be close to one, and  $\min |L(y)|$  subject to (2) should do well in preserving correlation. Note that denominator equal to one is equivalent to preserving variance, which as we have seen also is accomplished via  $\min |L(y)|$ .

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

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

As  $\bar{y} = 0$ , then  $\beta_0 = 0, \beta_1 = 1$  whenever  $L(y) = 0$  is feasible. This again corresponds to  $\min |L(y)|$  subject to the constraints of (2), viz., to (3).

Thus, data quality as regards means, variances, correlation and regression slope can be preserved under CTA in the univariate case by the mathematical program (3). Cox and Kelly [7] report computational results for a 2-dimensional table of actual data and a hypothetical 3-dimensional table that are nearly perfect in preserving all quantities of interest.

### 3 Using CTA to Preserve Multivariate Relationships

In place of a single data set organized in tabular form, viz.,  $\mathbf{T}\mathbf{a} = \mathbf{0}$ , to which adjustments  $\mathbf{y}$  are to be made for confidentiality purposes, henceforth we consider multiple

data sets, each organized within a common tabular structure  $\mathbf{T}$ . This is the typical situation in official statistics where, for example, 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  $\mathbf{a}$ ,  $\mathbf{b}$  and corresponding adjustments to original values are denoted by variables  $\mathbf{y}$ ,  $\mathbf{z}$ . In the univariate case, the key to preserving variance, correlation and regression slope was to force  $\text{Cov}(\mathbf{a}, \mathbf{y}) = 0$ . Though trivial, it is easy to overlook in the univariate case that as  $\text{Var}(\mathbf{a}) = \text{Cov}(\mathbf{a}, \mathbf{a})$ , then preserving variance via  $\text{Cov}(\mathbf{a}, \mathbf{y}) = 0$  is equivalent to requiring  $\text{Cov}(\mathbf{a}, \mathbf{a} + \mathbf{y}) = \text{Cov}(\mathbf{a}, \mathbf{a})$ . 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  $\mathbf{b}$  on original data  $\mathbf{a}$  in the adjusted data. These are the objectives of Section 3.

### 3.1 Preserving Means and Univariate Properties

This was the subject of the preceding section. Here we need only establish notation. Continuing our focus on the bivariate case, we begin with two copies of the mathematical program (3), one expressed in  $\mathbf{a}$  and  $\mathbf{y}$  and the other expressed in  $\mathbf{b}$  and  $\mathbf{z}$ . By virtue of the preceding section, this is sufficient to preserve both univariate means and variances.

### 3.2 Preserving the Variance-Covariance Matrix

The separate copies of model (3) of the preceding section preserve the univariate variances  $\text{Var}(\mathbf{a})$  and  $\text{Var}(\mathbf{b})$ . To preserve  $\text{Cov}(\mathbf{a}, \mathbf{b})$ , we require:

$$\text{Cov}(\mathbf{a}, \mathbf{b}) = \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) = \text{Cov}(\mathbf{a}, \mathbf{b}) + \text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y}) + \text{Cov}(\mathbf{y}, \mathbf{z})$$

Consequently, we seek a precise or approximate solution to:

$$\begin{aligned} \min & \quad | \text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y}) + \text{Cov}(\mathbf{y}, \mathbf{z}) |, \\ \text{subject to} & \quad (3) \end{aligned} \tag{4}$$

The first two terms in this objective function are linear and not pose a problem, but the last term is quadratic. We could apply quadratic programming to (4), and for some problems this will be acceptable computationally. For general problems and interest, we continue to pursue an approximate linear formulation. For practical purposes this formulation will express the objective in the constraint system. Our 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.

### 3.3 Preserving the Simple Linear Regression Coefficient

Our objective is to preserve the estimated regression coefficient under simple linear regression of  $\mathbf{b}$  on  $\mathbf{a}$ . We do not address here related issues of preserving the standard error of the estimate and goodness-of-fit. We seek exactly or approximately:

$$\text{Cov}(\mathbf{a}, \mathbf{b}) / \text{Var}(\mathbf{a}) = \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) / \text{Var}(\mathbf{a} + \mathbf{y})$$

$$\begin{aligned} \text{Var}(\mathbf{a} + \mathbf{y}) / \text{Var}(\mathbf{a}) &= \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) / \text{Cov}(\mathbf{a}, \mathbf{b}) \\ &= 1 + \text{Cov}(\mathbf{a}, \mathbf{z}) / \text{Cov}(\mathbf{a}, \mathbf{b}) + \text{Cov}(\mathbf{b}, \mathbf{y}) / \text{Cov}(\mathbf{a}, \mathbf{b}) + \text{Cov}(\mathbf{y}, \mathbf{z}) / \text{Cov}(\mathbf{a}, \mathbf{b}) \end{aligned}$$

Recall:

$$\begin{aligned} \text{Var}(\mathbf{a} + \mathbf{y}) / \text{Var}(\mathbf{a}) &= 2L(\mathbf{y}) + 1 + \text{Var}(\mathbf{y}) / \text{Var}(\mathbf{a}) \\ 2L(\mathbf{y}) + \text{Var}(\mathbf{y}) / \text{Var}(\mathbf{a}) &= \text{Cov}(\mathbf{a}, \mathbf{z}) / \text{Cov}(\mathbf{a}, \mathbf{b}) + \text{Cov}(\mathbf{b}, \mathbf{y}) / \text{Cov}(\mathbf{a}, \mathbf{b}) \\ &\quad + \text{Cov}(\mathbf{y}, \mathbf{z}) / \text{Cov}(\mathbf{a}, \mathbf{b}) \end{aligned}$$

To preserve univariate properties, impose the constraint  $L(\mathbf{y}) = 0$  exactly or approximately. To preserve bivariate covariance, impose  $\text{Cov}(\mathbf{y}, \mathbf{z}) = 0$  exactly or approximately. So, if in addition we seek to preserve the regression coefficient, then we must satisfy the linear program:

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

In implementation, the objective is represented as a near-zero constraint on the absolute value.

### 3.4 Preserving Correlations

The objective here is to ensure that correlations between variables computed on adjusted data are close in value to correlations based on original data, viz., that, exactly or approximately  $\text{Corr}(\mathbf{a}, \mathbf{b}) = \text{Corr}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z})$ . After some algebra, preserving correlation is equivalent to satisfying, exactly or approximately:

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

Methods included in (5) for preserving univariate variances and covariance in many cases will preserve correlation. Otherwise, iteration aimed at controlling the right hand product may help.

## 4 Results of Computational Simulations

We tested the performance of the proposed linear formulations on three 2-dimensional tables for both univariate and bivariate statistical measures. Three tables were taken from a 4x9x9 3-dimensional table. This table contained actual magnitude data and disclosure was defined by a (1 contributor, 70%) dominance rule, viz, a cell is sensitive if the largest contribution exceeds 70% of the cell value. This results in protection

limits  $p_i = (v_1^i)/0.7 - v^i$ . Tables A, B, C contain 6, 5, 4 sensitive cells respectively. Upper bounds (capacities) for adjustments to nonsensitive cells were set at 20-percent of cell value.

First, we used the MILP formulation to compute exact solutions for the three instances AB, AC, BC. The MILP formulation used the mean and variance preserving constraints and a covariance change minimization objective, aimed at preserving both univariate and bivariate statistics. This kind of modeling has the advantage of building flexibility for controlling information loss. For example, if preserving the covariance is not important, then covariance preserving constraints would be relaxed to get better performance with respect to other statistics.

We used the ILOG-CPLEX-Concert Technology optimization software to solve the resulting MIP and LP problems. The program code was written using C++ and Concert Technology libraries. Table 1 reports performance on covariance, correlation, regression coefficient, ‘i’ vector variance, ‘j’ vector variance (e.g.,  $i = B, j = C$ ). Table values are percent change in the original statistics. Means were preserved for all three instances. The results are encouraging: information loss (change) to the key statistics was low using the linear formulation. This is desirable because using the linear formulation to preserve univariate and bivariate statistics simultaneously offers advantages on scalability, flexibility, and computational efficiency. The loss on the bivariate statistical measures (regression coefficient, correlation) was considerably lower because the bivariate constraints offer more flexibility for adjusting cell values.

**Table 1.** Performance of the linear formulations on key statistics (in percent change).

Cases	Covariance change	Correlation change	Regression Coeff. change	Variance i change	Variance j change	Original Correlation. Coefficient
AB	3.15	1.09	5.94	-3.22	6.2	0.77
AC	1.13	2.63	1.14	-2.43	0.1	0.40
BC	3.6	6.12	6.7	-3.6	-1.89	0.49
Average	2.62	3.28	4.59	-3.08	1.47	0.55

We formulated the problem as a binary integer program, which in general has many feasible solutions. It would be interesting to study the behavior of these feasible solution vectors on the various statistics to determine the distribution of information loss over particular statistics. As an illustration, we evaluated performance of all the feasible solution vectors for the instance AB. We plotted the performance of these solutions on covariance and absolute cell deviation as shown in Figures 1- 2. Covariance varied considerably across different feasible solutions. However, the absolute cell deviation did not. Absolute cell deviation is one of the widely used performance measure in the statistical disclosure control literature. This indicates that it is possible to improve performance on the key statistics without hampering performance on cell deviation, and demonstrates the importance of incorporating “statistical properties preserving” optimization models in the traditional “minimizing cell deviation” framework.

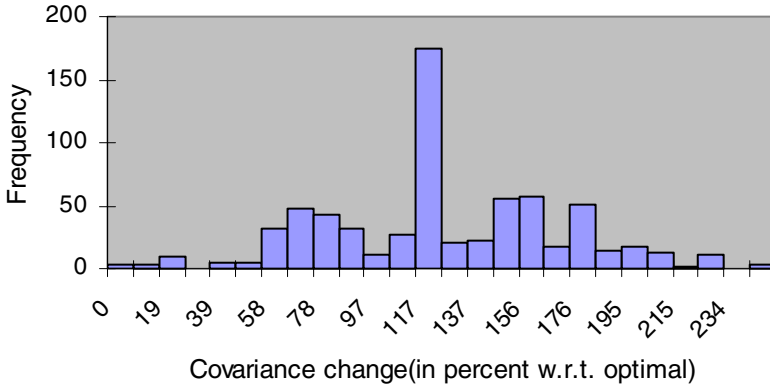


Fig. 1. Distribution of solutions w.r.t. covariance change.

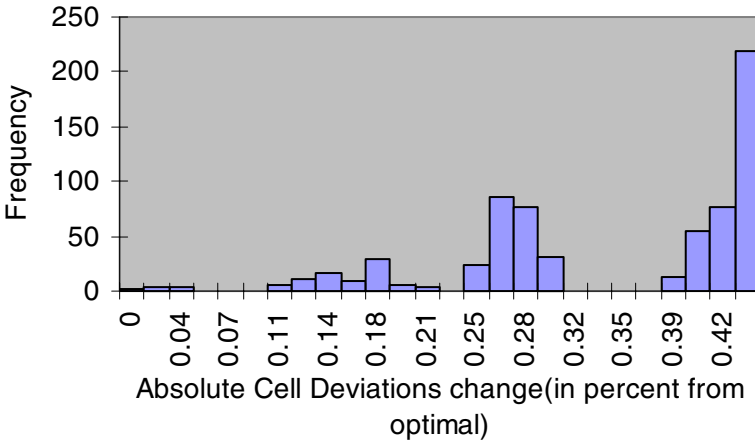


Fig. 2. Distribution of solutions w.r.t absolute cell deviation.

The *ordering heuristic* as proposed in [3] first sorts the sensitive cells in descending order and then assigns the directions for the sensitive cells in an alternating fashion. It is intended to find good solutions with respect to absolute cell deviation in a computationally efficient manner. We studied its performance on preserving statistical measures by comparing its performance to that of the exact MILP method and to an optimal (nonlinear) solution. By conducting explicit enumeration of the feasible vectors it was determined that the nonlinear formulations preserve covariance.

We used EXCEL-Solver to solve this nonlinear problem (which in general would be computationally impractical). Table 2 reports the comparison between the MILP and ordering heuristic methods. Given the variation in the quality of the solutions on the covariance and variance measures, the performance of the ordering heuristic was good and, in fact, the ordering heuristic improved performance on absolute cell deviation. This is consistent with [3], which reports superior performance of the ordering heuristic on cell deviation.

**Table 2.** Performance of ordering heuristic (in percent).

Solution Method	Covariance change	Correlation change	Variance A change	Variance B change	Absolute Cell Deviation
Exact (MIP)	3.15	1.09	5.94	-3.22	8.81e+7
Ordering Heuristic	5.34	2.19	4.49	-4.56	8.78e+7
Performance w.r.t. optimal	69	100	-24	41	-0.34

## 5 Concluding Comments

Developments over the past two decades have resulted in a variety of methods for statistical disclosure limitation in tabular data. Among these, controlled tabular adjustment yields the most useable data product, particularly for magnitude data, thereby supporting broad opportunities to analyze released data. The question is then how well adjusted data preserve analytical outcomes of original data. Cox and Kelly [7] addressed this issue in the univariate case. This paper provides effective linear effective formulations for preserving key statistics in the multivariate case.

There is an aspect of our formulations worth elaborating, namely that to the extent possible quality considerations be incorporated into the constraint system rather than the objective function. We find this approach more comprehensible and flexible than the traditional approach of defining an information loss measure and optimizing it over the tabular and nominal constraints. More importantly, our approach reflects a point of view that while mathematical optimization is an essential and flexible tool in modeling and solving quality/confidentiality problems, obtaining a mathematically optimal solution is of secondary, sometimes negligible, importance. This is because there are typically many solutions (adjusted tabulations) that by any statistical standard, such as measurement error, are indistinguishable. The issue we claim is then to develop a mathematical programming model incorporating essential relationships and quality considerations, and to accept any feasible solution. The objective function, such as sum of absolute adjustments, is used merely as one among several quality controls. Consequently, near-optimal solutions are in most cases acceptable. This can be significant in reducing computational burden, especially for national statistical offices that deal with large and multiple tabulation systems.

Related research not reported here looks at combining all constraints to produce a single set of multivariate solutions, e.g., ABC, that perform well on key statistical measures, at producing formulations for other statistical measures, e.g., for count data, the Chi-square statistic, and at investigating computational and scale issues. Space limitations preclude presenting here the original tables A, B, C and adjusted pairs AB, AC, BC, available from the authors.

## 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. Dandekar, R.A., Cox, L.H.: Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. (2002) (manuscript)
4. U.S. Department of Commerce.: Statistical Disclosure and Disclosure Limitation Methods, Statistical Policy Working Paper 22, Washington, DC: Federal. Committee on Statistical Methodology.(1994)
5. Cox, L.H.: Network Models for Complementary Cell Suppression. *Journal of the American Statistical Association.* 90(1995) 1153-1162.
6. Cox, L.H., Dandekar, R.A.: 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) (in press).*
7. Cox, L.H., Kelly, J. P.: 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 (in press)*
8. Glover, F., Laguna, M.: *Tabu Search.* Kluwer Academic, Amsterdam (1997)
9. Fischetti, M., Salazar-Gonzalez, J.J.: Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. *Journal of the American Statistical Association.* 95 (2000) 916-928.
10. 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.
11. Cox, L.H., Kelly, J. P.: *Controlled Tabular Adjustment: An Empirical Study of Heuristic and Optimal Methods.* (2004) (submitted).