

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

BALANCING DATA QUALITY AND CONFIDENTIALITY FOR TABULAR DATA

Invited Paper

Submitted by National Center for Health Statistics, Centers for Disease Control and Prevention (USA)¹

1. INTRODUCTION

1. Tabular data are the earliest form and remain a staple of official statistics data products. Familiar examples of tabular data products in official statistics include count data such as age-race-sex and other demographic data, concentration (or percentage) data such in financial or energy utilization statistics, and magnitude data such as total retail sales or air pollution data. Confidentiality problems in official statistics were first investigated for tabular data (Fellegi 1972). For magnitude data, the most studied and used disclosure limitation method has been complementary cell suppression (Cox 1980).

2. Tabular systems in official statistics tend to be large, comprising, for example, four or five levels of (not necessarily nested) geography (city, county, metropolitan area, state, country) cross-classified with (nested) industry (such as six-digit NAICS), cross-classified with several levels of employment class size. The tabulation system must be additive from each level to next higher levels. Additivity is represented by a system of linear tabular equations $\mathbf{TX} = \mathbf{0}$ wherein \mathbf{X} represents the individual cells and \mathbf{T} the tabular equations, viz., entries of \mathbf{T} restricted to the set $\{-1, 0, +1\}$, with precisely one -1 in each row.

3. Dandekar and Cox (2002) propose a disclosure limitation method for tabular data, following earlier ideas and software of Dandekar. Originally called *synthetic tabular data* this methodology has come to be called *controlled tabular adjustment* (CTA). This methodology is motivated by user dissatisfaction with complementary cell suppression, most particularly removal of otherwise useful information and consequent difficulties in analyzing data sets for which data are missing not at random. In essence, the CTA methodology replaces values of certain tabulation cells, called *sensitive cells*, that cannot be published due to confidentiality concerns, with *safe values*, viz., values sufficiently far away from the true sensitive value. Because these adjustments almost certainly will throw the additive tabular system out of kilter, controlled tabular adjustment adjusts some or all of the *nonsensitive cells* by small amounts to restore additivity. The CTA methodology is described in the next section.

4. From the standpoint of ease-of-use, controlled tabular adjustment is a definite improvement over complementary cell suppression. Because CTA changes cell values and because, for the sensitive cells, these changes cannot always be regarded as Δ_{small} , the question arises as to whether the effects of the CTA methodology on data analytical outcomes are acceptable or not. Cox and Dandekar (2003) describe how the basic methodology can be implemented with an eye towards preserving analytical utility. Herein, we probe these issues further, focusing on preserving mean, variance,

¹Prepared by Lawrence H. Cox (lcox@cdc.gov) and James P. Kelly (kelly@opttek.com).

correlation and regression between original and adjusted data. We offer two methodologies for preserving these statistics. The first is an approximate method based on linear programming. Benefits of this method include easy implementation by a wide class of users and simple but useful estimates of limitations on the extent to which these quantities can be improved. This analysis also reveals inherent conflicts between optimizing one statistic versus optimizing another, and a simple compromise to achieve acceptable results for all. The second method is a direct search strategy based on Tabu Search. Benefits of this method include the ability to produce optimal results in a wide range of applications and to optimize other nonlinear statistical measures, e.g., Chi-square statistics for categorical data.

5. Section 2 provides a brief summary of the Dandekar-Cox CTA methodology. In Section 3, we provide linear programming formulations for preserving mean, variance, correlation and regression between original and adjusted data, and introduce a compromise that can work well in all of these situations and is easy to implement. The methods are applied to a two-dimensional table of magnitude data drawn from actual data and a hypothetical three-dimensional table with complex structure. In Section 4, we provide a method based on Tabu Search and apply the method to the same two examples. Section 5 provides discussion of Issues and concluding comments.

2. CONTROLLED TABULAR ADJUSTMENT

6. The methods described below are applicable to tabular data in any form. However, it will be convenient to confine the presentation to magnitude data, which is the area most likely to benefit from the new methodology. 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 particular county) and their data (e.g., retail sales and employment data). The NSO assumes that any respondent is aware of the identity of the other respondents. The cell value, intended for publication, is the total value of a statistic of interest (e.g., total retail sales) over the values of this statistic for each respondent in the cell. The individual respondent values are their *contributions* to the cell value. Contributions are assumed to be nonnegative. Denote the cell value $v^{(i)}$ and the respondent contributions $v_j^{(i)}$, ordered from largest to smallest contribution. Then, it is possible for any respondent J to compute $v^{(i)} - v_J^{(i)}$ and obtain with certainty an upper estimate of the contribution of any other respondent. This estimate is closest, in percentage terms, when $J = 2$ and $j = 1$, viz., when the second largest contributor estimates the contribution of the largest. A standard disclosure rule, the *p*-percent rule, declares that disclosure occurs through publication of a particular cell value if this estimate is closer than p -percent of the largest contribution. The parameter value p is selected by the NSO, e.g., $p = 20\%$, and may itself be kept confidential. A cell containing disclosure is called a *sensitive cell*.

7. The NSO may also assume that any respondent can estimate the contribution of any other respondent to within q -percent ($q > p$, of course), e.g., $q = 50\%$. This is referred to as *public knowledge*. With this additional information, the second largest contributor can estimate the quantity $v^{(i)} - v_1^{(i)} - v_2^{(i)}$, viz., the sum of all contributions excluding both the largest and second largest, to within q -percent of its true value. The resulting upper estimate provides the second largest contributor directly with a lower estimate of $v_1^{(i)}$. If this estimate is within p -percent, then disclosure occurs. 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 $v_1^{(i)}$ of the largest contribution. Numeric values below the lower protection limit or above the upper protection limit are referred to as *safe values* for the cell. A common NSO practice, adopted here for convenience, is to assume that these protection limits are equal and are denoted p_i . The corresponding limit on estimating the cell value from public knowledge is denote $q_i > p_i$. Complementary cell suppression first suppresses all sensitive cells from publication. This in effect replaces their values by variables in the system of linear tabulation equations $\mathbf{TX} = \mathbf{0}$. Because, almost surely, one or more of the suppressed sensitive cell values can be estimated via linear programming to within p -percent of its true value, it is therefore necessary to suppress some of the nonsensitive cells (viz., replace additional values by variables in the tabular system) until such estimates are no longer closer to p -percent. This amounts to a mixed integer linear programming (MILP) problem (Fischetti and Salazar 2000).

8. The controlled tabular adjustment methodology (Dandekar and Cox 2002) first replaces each sensitive value with a safe value. This on the one hand is an improvement over complementary cell suppression as it replaces a suppression symbol by an actual value. On the other hand, safe values are not unbiased estimates of true values. To minimize bias, Dandekar and Cox (2002) suggest replacing the true value with either of its protection limits, viz., with either $v^{(i)} - p_i$ or $v^{(i)} + p_i$. Because these assignments almost surely throw the additive tabular system out of kilter, CTA next adjusts nonsensitive values to restore additivity. Because the individual choices to adjust each sensitive cell value down or up are binary and not continuous, these steps combined define a mixed integer linear program, presented in Cox (2000). Dandekar and Cox present several heuristics for making the binary choices, thereby reducing the problem to a linear program.

9. Mathematical programming and statistics are very different disciplines. A linear, or mixed integer linear, program in itself will not assure that analytical properties of original and adjusted data are comparable. Cox and Dandekar (2003) address these issues in three ways. First, as described above, sensitive values are replaced by the closest possible safe values. Second, capacities are imposed on changes to nonsensitive values to ensure that adjustments to individual datum are acceptable. A statistically sensible way to select these capacities is to base them on estimated measurement error for each cell e_i . Third, the linear program is optimized with respect to an overall measure of distortion to the data. Standard choices are minimum sum of absolute adjustments, minimum sum of percent absolute adjustments, and related measures involving logarithms of adjustments.

10. Assume that there are n tabulation cells of which s are sensitive cells and that these appear first in the ordering, viz., cells $i = 1, \dots, s$ are sensitive. The MILP of Cox (2000) for the case of minimizing sum of absolute adjustments follows. The original data are represented by $n \times 1$ vector \mathbf{a} , and the adjusted data are $\mathbf{a} + \mathbf{y}^+ - \mathbf{y}^-$. Henceforth denote $\mathbf{y} = \mathbf{y}^+ - \mathbf{y}^-$. We also use this notation for a subset of cells when the context is clear.

$$\min \sum_{i=1}^s (y_i^- + y_i^+)$$

Subject to: $\mathbf{T}(\mathbf{y}) = \mathbf{0}$

$$\begin{aligned} y_i^- &= p_i(1 - I_i) \\ y_i^+ &= p_i I_i \end{aligned} \quad i = 1, \dots, s$$

$$0 \leq y_i^-, y_i^+ \leq e_i \quad i = s+1, \dots, n$$

$$I_i \text{ binary, } i = 1, \dots, s$$

There is no guarantee that this problem is feasible, necessitating a strategy for increasing capacities on nonsensitive cells until a suitable feasible region is attained. A companion strategy, albeit controversial, is to allow 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 Cox and Dandekar (2003).

11. The Cox-Dandekar constraints are worthwhile. Unfortunately, choices for the optimizing measure are limited to linear functions. In the next two sections, we extend this paradigm in two separate directions, focusing on approaches to preserving mean, variance, correlation and regression between original and adjusted data.

3. LINEAR PROGRAMMING APPROACHES TO PRESERVING DISTRIBUTIONAL PROPERTIES

12. In this section, we present linear programming formulations for preserving mean, variance, correlation and regression between original and adjusted data. Some of these formulations are approximate as the statistics they address are nonlinear. However, they offer the significant advantage of being implementable by any organization with linear programming software. In addition, close examination reveals that they offer a reasonable approach to balancing competing statistical objectives. Numerical results are presented for a two-dimensional table based on actual magnitude data and a hypothetical three-dimensional table of size 13x13x13 with nonhierarchical structure among the 13 variables along each dimension, developed by Dandekar.

13. Two observations are important at the outset. First, a putative weakness of CTA is that, while it may restrict changes to individual nonsensitive cells to within measurement error, it may change sensitive values by a significant amount. This is to be expected and is no worse than the effects of complementary cell suppression because CTA estimates of sensitive cell values, based on minimum protection limits, are at least as accurate as those available to a sophisticated user under cell suppression, and provide the less sophisticated user reliable estimates not otherwise available. Nevertheless, the issue of change to sensitive cells deserves examination, and consequently we focus on the sensitive cells in the remainder of this paper. From an analytical perspective, this is artificial as no user would care to analyze only the sensitive cells, assuming the user even knew which cells were sensitive. And, a conservative data protector might argue that preserving properties of sensitive cells potentially undermines disclosure limitation. Nevertheless, to address the issue of distortion to sensitive cells and its effects on statistical analysis, we demonstrate several ways in which this distortion can be controlled so as to preserve statistics of interest. Moreover, our formulations extend easily to the case of all cells or to any subset(s) of cells for which the issue of preserving analytical outcomes is of importance. The second observation is that the methods described in the remainder of this paper can be applied either in conjunction with the mixed integer linear program, or after feasible choices for the direction of change (down/up) for each sensitive cell have been determined via an appropriate heuristic. Thus, we do not distinguish between these cases further.

3.1 Preserving Mean Values

14. Adjusted cell values are different than original values and can change analytical outcomes. Restricting adjustments to nonsensitive cell values to within, e.g., measurement error, is a step in the right direction. However, adjustments to sensitive cells must be safe and, as formulated in Dandekar and Cox (2002), are likely to be larger. Effects of adjustments on data analysis, particularly analysis using linear models, are mitigated by preserving mean values. If the grand total is fixed (viz., adjustment capacitated to zero), then the grand mean is preserved. Similarly, any mean whether corresponding to a set of cells explicitly aggregated within the tabular system or to any preselected aggregation of cells can be preserved simply by capacitating its adjustment to zero.

15. We focus on the largest changes, viz., to the sensitive cells. The mean of the adjusted sensitive values will equal the mean of the original sensitive values if and only if the sum of the adjustments equals zero, viz., if $\sum (y_i^+ - y_i^-) = 0$.

The corresponding mathematical problem, namely, selecting two subsets of a set of (positive) values whose difference is minimal, is known as a *partitioning problem*. With fixed values for the downward and upward adjustments, achieving minimum difference at or near zero may be impossible. Consequently, we allow the sensitive adjustments to vary sufficiently to achieve a zero-partition. Assume for convenience that the adjustments may vary as much as public knowledge. The corresponding MILP is as follows.

$$\min c(y)$$

Subject to: $T(y) = 0$

$$\sum_{i=1}^s (y_i^+ - y_i^-) = 0$$

$$\begin{aligned}
p_i(1 - I_i) \leq y_i^- \leq q_i(1 - I_i) & & i = 1, \dots, s \\
p_i I_i \leq y_i^+ \leq q_i I_i & & \\
0 \leq y_i^-, y_i^+ \leq e_i & & i = s+1, \dots, n \\
I_i \text{ binary, } i = 1, \dots, s & &
\end{aligned}$$

$c(y)$ is constructed to keep the adjustments close to their lower limit, e.g., $c(y) = \sum y^+ + y^-$, and to achieve other optima as presented below. Recall that adjustment directions and adjustment amounts can be selected simultaneously by the MILP or, alternatively, adjustment directions can be selected heuristically and adjustment amounts determined by the linear program relaxation.

3.2 Preserving Variance

16. For any data x_k :
$$Var(x) = (1/t) \sum_{k=1}^t (x_k - \bar{x})^2 = (1/t) \sum_{k=1}^t x_k^2 - \bar{x}^2$$

viz., variance equals the mean of squared values minus the square of the mean value. In particular, when the mean is zero, variance equals the mean of squared values. We seek that: $Var(a + y) - Var(a)$ near 0 over all cells or a subset of cells of interest, such as the sensitive cells. We assume that the mean is also preserved, viz., $\bar{y} = 0$.

For any subset of cells of size t with $\bar{y} = 0$, e.g., the sensitive cells:

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

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

$$L(y) = (1/(tVar(a))) \sum_{i=1}^t (a_i - \bar{a}) y_i$$

Then

$$Var(a + y)/Var(a) = 2L(y) + (1 + Var(y)/Var(a))$$

and

$$|Var(a + y)/Var(a) - 1| = |2L(y) + (Var(y)/Var(a))|$$

Thus, relative change in variance can be minimized by minimizing the right-hand side. $L(y)$ is a linear function and $Var(a)$ is a constant. $Var(y)$ is a nonlinear function, but can be linearly approximated. We illustrate the procedure for the case of the sensitive cells.

17. As $\bar{y} = 0$, then $\text{Var}(\mathbf{y})$ equals the mean of squared y -values. Let m and M denote, respectively, lower and upper bounds on $\text{Var}(\mathbf{y})$. Obvious choices are given by:

$$m = (1/s) \sum_{i=1}^s p_i^2 \leq \text{Var}(\mathbf{y}) \leq (1/s) \sum_{i=1}^s q_i^2 = M$$

Note that m is a strong choice and M a weak one.

Dividing all terms by $\text{Var}(\mathbf{a})$ yields a linear approximation of $\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})$. Consequently,

$$2L(\mathbf{y}) + m / \text{Var}(\mathbf{a}) \leq \text{Var}(\mathbf{a} + \mathbf{y})/\text{Var}(\mathbf{a}) - 1 \leq 2L(\mathbf{y}) + M / \text{Var}(\mathbf{a})$$

and, up to the uncertainty in $\text{Var}(\mathbf{y})$, $|\text{Var}(\mathbf{a} + \mathbf{y})/\text{Var}(\mathbf{a}) - 1|$ is minimized by minimizing $|2L(\mathbf{y}) + (M + m)/(2\text{Var}(\mathbf{a}))|$. The absolute value is minimized as follows:

a) Incorporate two new linear constraints into the system:

$$w \geq 2L(\mathbf{y}) + (M + m)/(2\text{Var}(\mathbf{a}))$$

$$w \geq -2L(\mathbf{y}) - (M + m)/(2\text{Var}(\mathbf{a}))$$

b) Minimize w

If $\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})$ is small, as typically is the case, a good approximation is given by minimizing $|L(\mathbf{y})|$ directly.

Maximizing Correlation

18. We seek to assure high positive correlation between original and adjusted cell values. We illustrate the method for the case of the sensitive cells.

$$\text{Corr}(\mathbf{a}, \mathbf{a} + \mathbf{y}) = \frac{\sum_{i=1}^s (a_i - \bar{a})(a_i + y_i - \bar{a} - \bar{y})}{\sqrt{\sum_{i=1}^s (a_i - \bar{a})^2 \sum_{i=1}^s (a_i + y_i - \bar{a} - \bar{y})^2}}$$

Again, $\bar{y} = 0$ and $L(\mathbf{y}) = (1/s\text{Var}(\mathbf{a})) \sum (a_i - \bar{a}) y_i$

$$\text{Corr}(\mathbf{a}, \mathbf{a} + \mathbf{y}) = (s \text{Var}(\mathbf{a}) + L(\mathbf{y})) / \sqrt{s \text{Var}(\mathbf{a}) [s \text{Var}(\mathbf{a}) + 2s \text{Var}(\mathbf{a})L(\mathbf{y}) + \sum y_i^2]}$$

$$= (1 + L(\mathbf{y})) / \sqrt{(1 + \text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})) + 2L(\mathbf{y})} \geq (1 + L(\mathbf{y})) / \sqrt{(1 + M/\text{Var}(\mathbf{a})) + 2L(\mathbf{y})}$$

The right-hand function is maximized by maximizing $L(\mathbf{y})$ subject to the constraints. Again, $\min |L(\mathbf{y})|$ yields a good approximation to the optimum when $\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})$ is small.

Preserving Regression Slope

19. The objective is to preserve ordinary least squares regression $Y = b_1 X + b_0$ of the adjusted data $Y = \mathbf{a} + \mathbf{y}$ on the original data $X = \mathbf{a}$, that b_1 is near one while b_0 is near zero.

$$b_1 = \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{a}) / \text{Var}(\mathbf{a}) = 1 + L(\mathbf{y}), \quad b_0 = (\bar{\mathbf{a}} + \bar{\mathbf{y}}) - b_1 \bar{\mathbf{a}}$$

We continue to assume that $\bar{\mathbf{y}} = \mathbf{0}$. Therefore, $b_0 = \mathbf{0}$ if $b_1 = \mathbf{1}$. Thus, $b_0 = \mathbf{0}$ and $b_1 = \mathbf{1}$ if the constraint $L(\mathbf{y}) = \mathbf{0}$ is imposed and is feasible. Once again, this corresponds to minimizing $|L(\mathbf{y})|$.

A Compromise Solution for Balancing Competing Objectives

20. Variance is preserved by minimizing $L(\mathbf{y})$; correlation by maximizing $L(\mathbf{y})$; and regression by setting $L(\mathbf{y}) = \mathbf{0}$ (if feasible), all subject to $\bar{\mathbf{y}} = \mathbf{0}$. Whenever $\text{Var}(\mathbf{y})/\text{Var}(\mathbf{a})$ is small, typically the case, imposing the objective $\min |L(\mathbf{y})|$ will assure good results for variance, correlation and regression simultaneously. A shortcut is to incorporate the constraint $L(\mathbf{y}) = \mathbf{0}$. If feasible, this will produce the desired solution. Choosing $L(\mathbf{y})$ near zero is motivated statistically as it implies (near) zero correlation between original cell values \mathbf{a} and cell value adjustments \mathbf{y} . Lack of correlation is evident, e.g., because solutions \mathbf{y} and $-\mathbf{y}$ are in most situations interchangeable.

Results from Numerical Simulations

21. Below we report results of numerical simulations for a two-dimensional table and for hypothetical three-dimensional table of Dandekar. The two-dimensional table of size 4x9 was constructed from actual magnitude data for which 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, resulting in protection levels $p_i = v_i^{(i)} / 0.7 - v^{(i)}$. The three-dimensional table is of size 13x13x13 for which the variables along each dimension obey nonhierarchical tabular constraints; in the k-direction: (13) = (4) + (5) + (6); (12) = (1) + (6); (11) = (3) + (4); (10) = (3) + (6); (9) = (4) + (5); (8) = (1) + (2); and (7) = (3) + (8) + (13), and contains approximately 100 sensitive cells.

22. The 4x9 two-dimensional example is presented in Table 1. The original table contains $s = 7$ sensitive cells (in red). Their respective protection limits p_i are provided below the cell data. Absolute adjustments to sensitive cell values are capacitated to lie between p_i and 50% of the original cell value; absolute nonsensitive cell adjustments are capacitated to lie between zero and 50% of the original cell value. This enforces zero adjustment to zero cells, which is common practice. The 50% upper bounds are much broader than in practice, but facilitate computation and presentation for a table of this small size (36 cells) and high sensitivity.

4x9 Table									
<i>Original</i>	<i>Table</i>								
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	70000	14490006
56250	1487000	172500	667503	1006253	327500	1683000	1138250	46000	6584256
616752	202750	1899502	1098751	2172251	3825251	4372753	300000	787500	15275510
0	35000	0	16250	0	0	65000	0	140000	256250
840502	2042251	3355753	2370005	7669255	8133752	8562754	2588250	1043500	36606022
<i>Protection</i>	<i>(+/-)</i>								
0	0	0	0	0	0	0	0	21000	
625	0	0	0	0	0	0	0	7800	
0	0	0	0	0	0	0	40000	0	
0	10500	0	4875	0	0	0	0	42000	

Table 1: 4x9 Table of Magnitude Data and Protection Limits for Seven Sensitive Cells (in red)

$\min \sum y_i $									
166875	307001	1283751	587501	4490751	3981001	2442001	1150000	91000	14499881
56875	1487000	172500	667503	1006253	327500	1683000	1141875	38200	6580706
616752	202750	1899502	1103626	2172251	3825251	4372753	260000	816300	15269185
0	45500	0	11375	0	0	65000	36375	98000	256250
840502	2042251	3355753	2370005	7669255	8133752	8562754	2588250	1043500	36606022
$\min L-Bnd $ (Variance)									
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	91003	14511009
55625	1487000	172500	667503	1006253	327500	1683000	1146675	38200	6584256
616752	202750	1899502	1098751	2172251	3825251	4372753	260000	787498	15235508
0	18791	0	8125	0	0	65000	0	191756	283672
839877	2026042	3355753	2361880	7669255	8133752	8562754	2556675	1108457	36614445
$\max L$ (Corr.)									
167500	317501	1283751	587501	4490751	3981001	2442001	1129000	91000	14490006
55313	1499637	172500	667503	1006253	327500	1683000	1138250	34300	6584256
616752	202750	1899502	1098751	2172251	3825251	4372753	359884	787500	15335394
937	19250	0	8938	0	0	65000	0	94815	188940
840502	2039138	3355753	2362693	7669255	8133752	8562754	2627134	1007615	36598596
$\min L $ (Regress.)									
167500	317501	1276439	587501	4490751	3981001	2442001	1150000	91000	14503694
55625	1487000	172500	667503	1006253	327500	1683000	1138250	34420	6572051
616752	202750	1899502	1106063	2172251	3825251	4372753	260000	787500	15242822
0	19250	0	8938	0	0	65000	0	194267	287455
839877	2026501	3348441	2370005	7669255	8133752	8562754	2548250	1107187	36606022

Table 2: Original Table After Various Controlled Tabular Adjustments Using Linear Programming

23. The first adjusted table minimizes total absolute deviation over all cells, as in Cox and Dandekar (2003). The next table preserves the mean of the sensitive cells. The next three attempt to preserve variance, correlation and regression slope for the sensitive cells. Certainly, the set of sensitive cells is of no particular interest analytically, and does not motivate preserving its statistical properties. We choose to optimize over the sensitive cells for several reasons. First, they provide a convenient means to demonstrate how important statistical properties can be preserved over subsets of the data. Second, doing so serves to refute the notion that CTA cannot mitigate the effects of (large) changes to sensitive values. Third, the sensitive cells represent a worst case scenario and provide a basis of comparison for analyses based on the full table. Subject to optimizing the statistic of interest, each table optimizes minimum total absolute deviation. The last solution corresponds to $L(y)$ near zero and represents the A compromise solution@.

24. The statistics of interest corresponding to these simulations are summarized in the Table 2. The same simulations are performed for the three-dimensional table (not shown) and summarized in Table 3.

Summary: 4x9 Table		Using	Linear Programming
Sensitive Cells	Correlation	Regress. Slope	New Var. / Original Var.
min v./	0.98	0.82	0.70
min L-Bound (Var.)	0.95	0.93	0.94
max L (Cor.)	0.97	1.20	1.52
min L (Reg.)*	0.95	0.93	0.95
All Cells	Correlation	Regress. Slope	New Var. / Original Var.
All 4 Functions	1.00	1.00	1.00

Table 3: Summary of Results of Numeric Simulations on 4x9 Table Using Linear Programming

Summary: 13x13x13 Table		Using	Linear Programming
Sensitive Cells	Correlation	Regress. Slope	New Var. / Original Var.
min v./	0.995	0.96	0.94
min L-Bound (Var.)	0.995	1.00	1.00
max L (Cor.)	0.995	1.00	1.21
min L (Reg.)*	0.995	1.00	1.01
All Cells			
All 4 Functions	1.00	1.00	1.00

Table 4: Summary of Results of Numeric Simulations on 13x13x13 Table Using Linear Programming

* = compromise solution

4. CONTROLLED TABULAR ADJUSTMENT USING TABU SEARCH

25. The linear programming-based methods discussed in the previous sections require that the objective function and constraints be linear functions. Section 3 demonstrates that, through careful analysis, one can encourage these linear models to address nonlinear functions such as variance, correlation, and regression slope. However, the linearity restriction ultimately limits the effectiveness of these approaches. So also might size and computational demands for massively large problems. Recently, a heuristic method for controlled tabular adjustment was developed by OptTek Systems, Inc. for the U.S. Bureau of Transportation Statistics to process nonlinear objective functions and constraints (OptTek 2003). The new approach cannot guarantee optimality but does provide general nonlinear capabilities. Additionally, since the new method does not rely on linear programming it can be used to process extremely large, high-dimensional or complex tables.

Heuristic Algorithm

26. The algorithm contains three phases. In the first phase, a feasible solution is obtained. In the second phase, the solution is improved relative to the quality measure such as correlation, slope, variance or any other appropriate measure of information loss. The final phase uses Tabu Search (Glover and Laguna 1997) to further improve the best solution. Tabu Search applied to controlled tabular adjustment provides the opportunity to exploit the underlying structures by using adaptive memory and responsive exploration. The use of adaptive memory contrasts with "rigid memory" designs, such as those exemplified by branch and bound and associated processes that lie at the core of exact methods. The use of responsive exploration also affords the ability to guide the solution process in ways that are not accessible to exact methods.

27. The basis for implementing Tabu Search in the controlled tabular adjustment context may be described as follows. Consider first an abstract representation of the problem as that of optimizing a function $f(x)$ over a set X . Tabu Search begins by proceeding iteratively from one solution to another until a chosen termination criterion is satisfied. Each $x \in X$ has an associated neighborhood $N(x)$, and each solution in $N(x)$ is reached from x by an operation called a *move*. Tabu Search employs a strategy of modifying $N(x)$ as the search progresses, effectively replacing it by another neighborhood $N^*(x)$, based on the use of adaptive memory structures. The solutions admitted to $N^*(x)$ by these memory structures are determined in several ways. One of these, which gives Tabu Search its name, identifies solutions encountered over a specified horizon (and implicitly, additional related solutions), and forbids them to belong to $N^*(x)$ by classifying them *tabu*. The implementation of this mechanism allows the search process to overcome local optimality in the quest for the globally optimal solution.

28. The objective functions employed for controlled tabular adjustment can be summarized in the expression:

$$\text{Minimize } \{a(\text{Sum Abs. Dev.}) + b(1 - \text{Corr.}^2) + c|1 - \text{Slope}| + d|\text{New Variance}/\text{Orig. Var.} - 1|\}$$

where coefficients a, b, c, d are selected to provide the desired results and scale the sum of absolute deviations. Note that the functions used above are actual correlation, variance, and slope as opposed to the linear surrogates of Section 3.

29. The heuristic algorithm changes sensitive and non-sensitive cells first seeking to obtain a feasible solution, and then once feasibility is obtained, it moves on to optimize the quality measure. The algorithm only changes one cell or sum at a time. Eventually, no single change can improve the solution. At this point, the algorithm utilizes Tabu Search to move beyond the locally optimal solution to seek a globally optimal solution. In this phase of the algorithm, non-improving changes are forced into the solution to allow the search to move to better solutions. These forced changes are maintained within the solution for a fixed number of iterations. The best feasible solution found during the entire search is returned to the user.

30. Both tables and all measures examined in Section 3 were processed using Tabu Search. The results for the 4x9 table are shown in Table 5 and the statistics of interest for both tables presented in Tables 6 and 7, respectively.

5. CONCLUDING COMMENTS

31. We addressed the issue of preserving analytic utility of data subjected to controlled tabular adjustment for confidentiality purposes. We provided linear programming formulations capable of preserving mean values exactly and for approximately preserving variances and correlation and regression slope between original and adjusted cell values. Results from numeric simulations on an actual two-dimensional table and a hypothetical three-dimensional table with complex structure are encouraging.

32. The statistics variance, correlation and slope are not harmonious in the sense that optimizing one typically degrades another. We provided a single linear programming formulation for a compromise solution that strikes an acceptable balance between these objectives under expected conditions, viz., when variation in the adjustments is small relative to variation in the data. This condition is very likely to be met when the number of sensitive cells is small relative to the total number of cells. We also provided a strategy based on Tabu Search for which the search can be driven by any statistic of interest—linear or nonlinear—capable of producing good to optimal results in a wide variety of settings. The compromise solution also offers the possibility of speeding the search procedure by incorporating a constraint that forces $L(y)$ to be zero or small.

Variance (Tabu Search)									
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	34900	14454906
56875	1487000	172500	667503	1006253	327500	1683000	1138250	53800	6592681
616752	202750	1899502	1098751	2172251	3825251	4372753	260000	787500	15235510
0	45500	0	8125	0	0	65000	0	204300	322925
841127	2052751	3355753	2361880	7669255	8133752	8562754	2548250	1080500	36606022

Correlation (Tabu Search)									
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	92184	14512190
58058	1487000	172500	667503	1006253	327500	1683000	1138250	38200	6578264
616752	202750	1899502	1098751	2172251	3825251	4372753	341183	787500	15316693
0	24500	0	11375	0	0	65000	0	98000	198875
842310	2031751	3355753	2365130	7669255	8133752	8562754	2629433	1015884	36606022

Slope (Tabu Search)									
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	34900	14454906
56875	1487000	172500	667503	1006253	327500	1683000	1138250	53800	6592681
616752	202750	1899502	1098751	2172251	3825251	4372753	260000	787500	15235510
0	45500	0	8125	0	0	65000	0	204300	322925
841127	2052751	3355753	2361880	7669255	8133752	8562754	2548250	1080500	36606022

Table 5: Original Table After Various Controlled Tabular Adjustments Using Tabu Search

Summary: 4x9 Table		Using	Tabu Search
Sensitive Cells	Correlation	Regress. Slope	New Var. / Original Var.
Min Abs Dev	0.98	0.82	0.70
Variance	0.94	0.92	0.96
Correlation	0.98	1.13	1.32
Regression	0.94	0.92	0.96
All Cells	Correlation	Regress. Slope	New Var. / Original Var.
Min Abs Dev	1.00	1.00	1.00
Variance	1.00	1.00	1.00
Correlation	1.00	1.00	1.00
Regression	1.00	1.00	1.00

Table 6: Summary of Results of Numeric Simulations on 4x9 Table Using Tabu Search

Summary: 13x13x13		Using	Tabu Search
Sensitive Cells	Correlation	Regress. Slope	New Var. / Original Var.
Min Abs Dev	0.995	0.96	0.94
Variance	0.995	1.00	1.00
Correlation	0.995	1.00	1.02
Regression	0.995	1.00	1.02
All 4 Functions	Correlation	Regress. Slope	New Var./Original Var.
	1.00	1.00	1.00

Table 4: Summary of Results of Numeric Simulations on 13x13x13 Table Using Tabu Search

33. Preserving data quality and analytical utility under CTA is two-faceted. First, adjustments to individual cell values need to be as small as possible. Sensitive cells typically require larger adjustments, so differential adjustments are required for sensitive and nonsensitive cell values. This is accomplished under CTA by linear capacity constraints on adjustments: nonsensitive cell adjustments are limited to be within, e.g., a small percentage or measurement error, and sensitive cell adjustments are related to protection limits computed from the data and the disclosure rule. CTA opens the possibility of assigning protection below protection limits in most cases, thereby reducing bias to sensitive cell values. This is controversial but does have a mathematical foundation and should be considered by national statistical offices.

34. Capacities ensure that original and adjusted data are close Alocally@. The second facet is to ensure that they are close Aglobally@ in the sense that important statistical properties of the data set or relevant subsets are approximately preserved. Previous work, most notably Cox and Dandekar (2003), focused on minimizing deterministic measures of change such as total (percent) absolute change. This is worth doing to preserve analytical utility, but does not address statistical properties directly. Here we provide directly methods for preserving important statistical properties of the data: mean, variance, correlation and regression slope, further closing the gap between confidentiality protection and data quality and utility.

35. Preserving Alocal@ and Aglobal@ properties of the data are competing objectives. This can be seen directly from the formulations of Section 3. If sensitive values can be adjusted only to protection limits, then it is unlikely that the mean of sensitive values will be preserved. This necessitates broadening capacities on sensitive cells. If $L(\mathbf{y})$ cannot be forced near to zero, then the corresponding approximations to preserve variance, correlation and regression slope will be poor, again necessitated relaxation of capacities. Conversely, many users are interested primarily in individual values or sets of values. The more capacities are relaxed for these values, the less useful and reliable they become for the user. Rules of thumb are that it is easier to accommodate these competing objectives either if the data set is large or if the number of sensitive cells and the protection required do not overwhelm the nonsensitive cells: a certain amount of elbow room enables acceptable solutions. How and to what extent should capacities be relaxed? A balance can be struck by basing capacities on a small number of percentages applied to cell values, treating these percentages as variables, and optimizing statistics of interest and percentages jointly. We are investigating this and are extending these formulations to the multivariate case to ensure that relationships such as correlation and regression slope between variables exhibited by original data are preserved in adjusted data.

REFERENCES

- OptTek Systems, Inc. (2003). ADisclosure Limitation Methods.@ Report to the U.S. Bureau of Transportation Statistics..
- Cox, L.H. (1980). ASuppression Methodology and Statistical Disclosure Control.@ *Journal of the American Statistical Association* **75**, 377-385.
- _____. (2000). ADiscussion.@ **ICES II: The Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions**. Alexandria, VA: American Statistical Association, 905-907.
- _____ and R.A. Dandekar (2003). AA New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use.@ **Proceedings of the 2002 FCSM Statistical Policy Seminar, Federal Committee on Statistical Methodology**, Washington, DC: U.S. Office of Management and Budget, in press.
- Dandekar, R.A. and L.H. Cox (2002). ASynthetic Tabular DataBAN Alternative to Complementary Cell Suppression.@ submitted.
- Fellegi, I.P. (1972). AOn the Question of Statistical Confidentiality.@ *Journal of the American Statistical Association* **67**, 7-18.
- Fischetti, M. and J.J. Salazar-Gonzalez (2000). AModels and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints.@ *Journal of the American Statistical Association* **95**, 916-928.
- Glover, F., and M. Laguna (1997). **Tabu Search**. Amsterdam: Kluwer Academic Publishers.