QUALITY-PRESERVING AND MINIMUM DISCRIMINATION INFORMATION CONTROLLED TABULAR ADJUSTMENT:

ALTERNATIVES TO COMPLEMENTARY CELL SUPPRESSION FOR DISCLOSURE LIMITATION OF TABULAR DATA

Lawrence H. Cox, Ph.D.
Associate Director, Research & Methodology
National Center for Health Statistics
LCOX@CDC.GOV

ICES III
Montreal, Canada
June 18-21, 2007

WHERE WE ARE HEADED

(Nearly) Actual Example of Magnitude Table with Disclosures

167	317	1284	587	4490	3981	2442	1150	70 (21)	**14488**
57(1)	1487	172	667	1006	327	1683	1138	46 (7)	**6583**
616	202	1899	1098	2172	3825	4372	300(40)	787	**15271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1: 4x9 Table of Magnitude Data & Protection Limits for the 7 Disclosure Cells (red)

D	317	1284	D	4490	3981	2442	1150	D	**14488**
D	1487	172	667	1006	327	1679	D	D	**6583**
616	D	1899	1098	2172	3825	4371	D	787	**15271**
0	D	0	D	0	0	70	0	D	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1a: After Optimal Suppression: 11 Cells (*30%*) & 2759 Units (*7.5%*) Suppressed

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1683	1138	39	**6571**
617	196	1899	1095	2172	3825	4372	260	797	**15232**
0	26	0	12	0	0	65	0	180	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1b: After Controlled Tabular Adjustment

OUTLINE

1. Describe statistical disclosure limitation in tables

2. Describe complementary cell suppression

3. Describe controlled tabular adjustment

4. Describe one approach to preserving data quality and utility subject to controlled tabular adjustment:
   Quality-Preserving Controlled Tabular Adjustment

5. Describe a second approach to preserving data quality and utility subject to controlled tabular adjustment:
   Minimum Discrimination Information
   Controlled Tabular Adjustment

Statistical Disclosure Limitation (SDL) for Tabular Data

Tabular data
- * frequency (*count*) data organized in *contingency tables*
- * *magnitude* data (income, sales, tonnage, # employees, ..) organized in sets of tables

Tables
- * there can be *many*, many, many tables (national censuses)
- * tables can be 1-, 2-, 3-, ...up to many *dimensions*
- * tables can be *linked*
- * table entries: *cells* (industry = retail shoe stores & location = Washington DC)
- * data to be published: *cell values* (first quarter sales for shoe stores in Washington DC = $17M)

What is disclosure?

Count data: disclosure = small counts (1, 2, ...)
Magnitude data: disclosure = dominated cell value

Example: Shoe company # 1: $10M
Shoe company # 2: $ 6M
<u>Other companies (total): $ 1M</u>
Cell value: $17M

2 can subtract its contribution from cell value and infer contribution of #1 to within 10% of its true value = *DISCLOSURE*

Cells containing disclosure are called *sensitive cells*

How is disclosure in tabular data *limited* by statistical agencies?
- identify cell values representing disclosure
- determine *safe values* for these cells

Example: If estimation of any contribution to within 20% is deemed safe (policy decision), then a safe value is \$18M viz., \$18M - 6M = 12M $\geq$ (120%) \$10M

- traditional methods for statistical disclosure limitation
  - Count data:
    - rounding
    - data perturbation
    - swapping/switching
    - cell suppression
  - Magnitude data:
    - cell suppression

What is *complementary cell suppression* (CCS)?

- replace each senstive cell value by a symbol (*variable*)
- replace selected other cell values by a symbol (*variable*) to prevent narrow estimates of sensitive cell values
- process is complete when resulting system of equations divulges no *unsafe estimates* of sensitive cell values

Some properties of CCS:

* based on mathematical programming
* very complex theoretically, computationally, practically viz., NP-hard even for 1-dimensional tables
* destroys useful information
* thwarts many analyses; favors sophisticated users

How does CCS address *data quality?*

CCS uses a linear objective function to control *oversuppression*
Namely, the mathematical program minimizes either:

* total value suppressed
* total percent value suppressed
* number of cells suppressed
* logarithmic function related to cell values (*Berg entropy*)
* etc.

These are overall (*global*) measures of data distortion
Further, individual cell *costs* or *capacities* can be set to control individual cell (*local*) distortion

These are all sensible criteria and worth doing

However, they do not preserve statistical properties (*moments*)

Moreover, *suppression destroys data and thwarts analysis*

Controlled Tabular Adjustment (CTA)

* recent method for SDL in tabular data
* perturbative method–changes, does not eliminate, data
* alternative to complementary cell suppression
* attractive for *magnitude data* & applicable to count data

<u>Original CTA Method</u>

* identify sensitive tabulation cells
* replace each sensitive value by a *safe value*–namely, move the cell value *down* or *up* until safety is reached
* use linear programming to adjust nonsensitive values in order to restore additivity (*rebalancing*)
* if second and third steps are performed simultaneously, a *mixed integer linear program* (MILP) results. MILP is extremely computationally demanding
* otherwise (most often), the down/up decision is made heuristically, followed by rebalancing via linear programming (LP) which computes efficiently even for large problems

(Nearly) Actual Example of Magnitude Table with Disclosures

167	317	1284	587	4490	3981	2442	1150	70 (21)	**14488**
57(1)	1487	172	667	1006	327	1683	1138	46 (7)	**6583**
616	202	1899	1098	2172	3825	4372	300(40)	787	**15271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1: 4x9 Table of Magnitude Data & Protection Limits for the 7 Disclosure Cells (red)

D	317	1284	D	4490	3981	2442	1150	D	**14488**
D	1487	172	667	1006	327	1679	D	D	**6583**
616	D	1899	1098	2172	3825	4371	D	787	**15271**
0	D	0	D	0	0	70	0	D	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1a: After Optimal Suppression: 11 Cells (*30%*) & 2759 Units (*7.5%*) Suppressed

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1683	1138	39	**6571**
617	196	1899	1095	2172	3825	4372	260	797	**15232**
0	26	0	12	0	0	65	0	180	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1b: After Controlled Tabular Adjustment

MILP for Controlled Tabular Adjustment

Original data: nx1 vector **a**
Adjusted data: nx1 vector $\boldsymbol{a} + \boldsymbol{y}^{+} - \boldsymbol{y}^{-}$
T denotes the coefficient matrix for the tabulation equations
Denote $\boldsymbol{y} = \boldsymbol{y}^{+} - \boldsymbol{y}^{-}$
Cells i = 1, ..., s are the *sensitive cells*
Upper (lower) *protection* for sensitive cell i denoted p_i $(-p_i)$

<u>MILP for case of minimizing sum of absolute adjustments</u>

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

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

$$q_i(1 - I_i) \geq y_i^{-} \geq p_i(1 - I_i) \qquad i = 1, \ldots, s$$
$$q_i I_i \geq y_i^{+} \geq p_i I_i$$

(sensitive cells)

$$0 \leq y_i^{-},\ y_i^{+} \leq e_i \qquad i = s+1, \ldots, n$$

(nonsensitive cells)

$I_i = 0, 1$ (binary)

$q_i \geq p_i$: bounds on adjustments to sensitive cells

Capacities e_i on adjustments to nonsensitive cells are typically small, e.g., within measurement error

PRESERVING DISTRIBUTIONAL PARAMETERS SUBJECT TO CONTROLED TABULAR ADJUSTMENT:

QUALITY-PRESERVING CONTROLLED TABULAR ADJUSTMENT (QP-CTA)

Joint work with:

James P. Kelly Rahul J. Patil
OptTek Systems, Inc.

Data Quality Issues

Based on mathematical programming, in like manner to cell suppression, CTA can minimize any of:

* total (or max) of absolute values of adjustments
* total (or max) percent absolute adjustment
* number of cells changed
* logarithmic functions of absolute adjustments
* etc.

In addition, adjustments to nonsensitive cells can be restricted to lie within *measurement error*

Still, this may not ensure good statistical outcomes, namely,

<u>Objective</u>

analyses on original vs adjusted data yield comparable results

Towards Ensuring Comparable Statistical Analyses

Verification of "comparable results" is mostly empirical
Many, many analyses are possible: Which analysis to choose?

We focus on preserving key statistics and linear models

In the univariate case, we seek to preserve:

* mean values
* variance
* correlation
* regression slope

between original and adjusted data

Preserve means that adjusted data approximate reasonably well values for these quantities from original data

Can do this using direct (*Tabu*) search

I will describe ***how to do so well in most cases using LP***

For simplicity, assume that the down/up decisions for sensitive cells have already been made (by *heuristic*)

Preserving Mean Values

When the LP holds a total fixed, it *preserves the mean* of the cell values contributing to the total

e.g., fixing the grand total preserves the overall mean

In general, to preserve a mean, introduce (new) constraint:

$\sum$ (adjustments to cells contributing to the mean) = 0

Most of these are already expressed by the tabular constraints

Example: Preserving the mean of the sensitive cell values

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

The MILP is:

$$\min \ c(\boldsymbol{y})$$

Subject to:

$$\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{0}$$

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

$$\left.\begin{array}{l} p_i(1 - I_i) \le y_i^- \le q_i(1 - I_i) \\ p_i I_i \le y_i^+ \le q_i I_i \end{array}\right\} \quad i = 1, \ldots, s$$

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

$I_i = 0, 1$(binary)

$q_i \ge p_i$: bounds on adjustments to sensitive cells

$c(\boldsymbol{y})$ = linear cost fcn., e.g., sum of absolute adjust.

If the down/up directions are pre-selected, this is an LP

Preserving Univariate Statistics

Preserving variances

Seek: $Var(\boldsymbol{a} + \boldsymbol{y}) \doteq Var(\boldsymbol{a})$, assuming $\bar{y} = 0$

$$Var(\boldsymbol{a} + \boldsymbol{y}) = Var(\boldsymbol{a}) + 2Cov(\boldsymbol{a}, \boldsymbol{y}) + Var(\boldsymbol{y})$$

Define: $\boldsymbol{L}(\boldsymbol{y}) = (1/(sVar(\boldsymbol{a})))\sum_{i=1}^{s}(a_i - \bar{a})y_i = Cov(\boldsymbol{a}, \boldsymbol{y})/Var(\boldsymbol{a})$

L(**y**) is a *linear function* of the adjustments **y**

$$Var(\boldsymbol{a} + \boldsymbol{y})/Var(\boldsymbol{a}) = 2L(\boldsymbol{y}) + (1 + Var(\boldsymbol{y})/Var(\boldsymbol{a}))$$

$$|\ Var(\boldsymbol{a} + \boldsymbol{y})/Var(\boldsymbol{a}) - 1\ | = |\ 2L(\boldsymbol{y}) + (Var(\boldsymbol{y})/Var(\boldsymbol{a}))\ |$$

Typically, *Var(**y**)/Var(**a**) is small*
Thus, variance is approximately preserved by minimizing
$|L(\boldsymbol{y})|$

The absolute value is minimized as follows:

* incorporate two new linear constraints in the system:

$$w \geq L(\boldsymbol{y})$$
$$w \geq -L(\boldsymbol{y})$$

* minimize w

Assuring high positive correlation

Seek: $Corr(\boldsymbol{a}, \ \boldsymbol{a} + \boldsymbol{y}) \doteq 1$

Corr (**a**, **a + y**) = $Cov(\boldsymbol{a}, \ \boldsymbol{a} + \boldsymbol{y}) \div \sqrt{Var(\boldsymbol{a}) \ Var(\boldsymbol{a} + \boldsymbol{y})}$

$$= (1 + L(\boldsymbol{y})) \div \sqrt{Var(\boldsymbol{a} + \boldsymbol{y}) \ / \ Var(\boldsymbol{a})}$$

Note:

1. Denominator near one
2. min $|L(\boldsymbol{y})|$ drives numerator to one

Preserving regression coefficients

Seek: under ordinary least squares regression

$$Y = \beta_1 X + \beta_0$$

of adjusted data Y = **a + y** on original data X = **a**,

we want (approximately): $\beta_1 = 1$ and $\beta_0 = 0$

$$\beta_1 = Cov(\boldsymbol{a} + \boldsymbol{y}, \boldsymbol{a}) / Var(\boldsymbol{a}) = 1 + L(\boldsymbol{y}),$$

$$\beta_0 = (\bar{a} + \bar{y}) - \beta_1 \bar{a}$$

As $\bar{y} = 0$, then $\beta_0 = 0$ if $\beta_1 = 1$

This corresponds (approximately) to $L(\boldsymbol{y}) = 0$ (if feasible)

Note again: best result achieved for min $|L(\boldsymbol{y})|$

Comment: $L(\boldsymbol{y}) = 0$ is motivated statistically because,

as solutions **y** and **-y** are equally good,

data **a** and adjustments **y** must be uncorrelated

Examples

4x9 Table									
Original	*Table*								
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**
Protection	*(+/-)*								
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 Its Seven Sensitive Cells (in red)

min $\sum \lvert y_i \rvert$									
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 To Preserve Statistical Properties of Sensitive Cells Only

167	317	1284	587	4490	3981	2442	1150	70 (21)	**14488**
57(1)	1487	172	667	1006	327	1683	1138	46 (7)	**6583**
616	202	1899	1098	2172	3825	4372	300(40)	787	**15271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1: 4x9 Table of Magnitude Data & Protection Limits for the 7 Disclosure Cells (red)

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1679	1138	39	**6571**
617	196	1899	1095	2172	3825	4371	260	797	**15232**
0	26	0	12	0	0	70	0	180	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1b: Table After Controlled Tabular Adjustment

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1683	1138	35	**6571**
617	202	1899	1098	2172	3825	4372	260	787	**15232**
0	20	0	9	0	0	65	0	194	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1c: Table After Optimal Controlled Tabular Adjustment (Regression)

Results for 4x9 table

Summary: 4x9 Table		**Linear**	**Programmin g**
Sensitive Cells	Corr.	Regress. Slope	New Var. / Original Var.
min $\|y_i\|$	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			
All 4 Functions	**1.00**	**1.00**	**1.00**

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

Results for 13x13x13 table

Summary: 13x13x13 Table		**Linear**	**Programming**
Sensitive Cells	Corr.	Regress. Slope	New Var. / Original Var.
min $\lvert y_i \rvert$	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

PRESERVING MULTIVARIATE STATISTICS

Preserving the variance-covariance matrix

Data: **a**, **b**
Adjustments: **y**, **z**

Variances approximately preserved by preserving means and adjoining
L (**y**) = 0 to CTA constraints; together = *univariate constraints*

Cov (**a + y**, **b+ z**) = Cov (**a**, **b**) + Cov (**y**, **b**) + Cov (**a**, **z**) + Cov (**y**, **z**)

Thus, Cov **(a + y**, **b + z**) = Cov (**a**, **b**) iff

Cov (**a**, **z**) + Cov (**y**, **b**) + Cov (**y**, **z**) = 0

Last term is nonlinear
Could use quadratic programming
We prefer to solve

min |Cov (**a**, **z**) + Cov (**y**, **b**) + Cov (**y**, **z**)|
subject to univariate constraints

as a sequence of LPs: for $\mathbf{y} = \mathbf{y_0}$ (constant), solve optimal $\mathbf{z} = \mathbf{z_0}$
fix $\mathbf{z} = \mathbf{z_0}$ (constant), solve optimal $\mathbf{y} = \mathbf{y_1}$
Continue
STOP when sufficiently close to 0

Call this system the *variance-covariance constraints*

Preserving the simple linear regression coefficient

Simple linear regression of **b** on **a**
Simple linear regression coefficient $\beta_1 = Cov\ (\boldsymbol{a}, \boldsymbol{b}) \ / \ Var(\boldsymbol{a})$

So, we seek:

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

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

Variance-covariance constraints assure left-hand side near 1
Univariate constraints assure right-hand side near 1

Preserving correlations

$$\text{Corr}\,(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) = \text{Corr}\,(\mathbf{a}, \mathbf{b}) \quad \text{iff}$$

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

and again this is assured by the *variance-covariance constraints*

COMPUTATIONAL RESULTS (multivariate)

Data

Three 4x9 tables (**A**, **B**, **C**) selected from a 4x9x9 table of actual data
Disclosure rule: (1, 70)-dominance rule
Sensitive cells: **A** (6) **B** (5) **C** (4)

Effect of CTA on univariate and bivariate statistics

Case	Covariance	Correlation	Reg.Coef.	Var 1	Var 2
AB	3.15	1.09	5.94	-3.22	6.20
AC	1.13	2.63	1.14	-2.43	0.10
BC	3.60	6.12	6.70	-3.60	-1.89
Avg.	2.62	3.28	4.59	-3.08	1.47

(in percent change)

QP-CTA: SUMMARY

Controlled tabular adjustment (**CTA**) can

- provide disclosure-protected tabular data
- preserve additive tabular structure
- be implemented using linear programming (**LP**)

Univariate case

CTA can be extended using LP to preserve

- means and variances
- correlation and regression between original and adjusted data

Multivariate case

Univariate CTA can be extended using LP to preserve

- multivariate variance-covariance matrix
- bivariate correlations
- bivariate simple linear regression coefficient

We call this method quality-preserving controlled tabular adjustment (**QP-CTA**)

REFERENCE

Cox, L.H., Kelly, J.P., and Patil, R.J. Balancing quality and confidentiality for multi-variate tabular data. In: **Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050**,(J. Domingo-Ferrer, V. Torra, eds), New York: Springer Verlag, 2004, 87-98.

PRESERVING STATISTICAL DISTRIBUTIONS SUBJECT TO CONTROLLED TABULAR ADJUSTMENT:

MINIMUM DISCRIMINATION INFORMATION CTA (MDI-CTA)

Joint work with:

Jean G. Orelien Babubhai Shah

SciMetrika, LLC

KULLBACK-LEIBLER DISTANCE

Kullback-Leibler is a probability-based distance function between two distributions. For tables, K-L is defined:

1. Given a probability distribution $\pi(w)$ over the set of cells for a table or space Ω such that $\sum_{\Omega} \pi(w) = 1$, and a family of distributions $P\{p(w)\}$ which satisfies certain constraints (e.g., $\sum_{\Omega} p(w) = 1$), *K-L distance* is given by

$$I(p:\pi) = \sum_{\Omega} p(w) \log\left(\frac{p(w)}{\pi(w)}\right)$$

2. The distribution $p^{*}(w)$ of P that is closest to $\pi(w)$ in terms of $I(p:\pi)$ is the *minimum discrimination information* or MDI

Properties of MDI

1. $I(p:\pi)$ is a convex function, hence the procedure yields a unique MDI solution
2. If $p^*(w)$ is the MDI, it can be shown that for any member $p(w)$ of P
3. $I(p:\pi) \geq 0$ with equality if and only if $\pi(w) = p(w)$

Application of MDI to CTA

1. MDI-CTA: given a distribution (original values in a table), select a combination of lower or upper safe values that yield minimum discriminant information
2. In principle, given a table with n sensitive cells, for each of the 2^n combinations, we would need to compute the discriminant information to find the MDI
3. Because of the limitations of computing resources, so many computations cannot be done in a timely manner
4. Therefore, we need heuristic steps

MDI-CTA Algorithm

Algorithm for a 3x3x3 table:

1. Within each row, for each combination of sensitive cells compute the discrimination information. This requires that we adjust the values of nonsensitive cells within that row (by making the values of the nonsensitive cells add up to the total of original values minus sensitive values in the row)
2. Choose the combination having the lowest value for the row
3. Repeat the steps above for each column and depth
4. The first heuristic solution is arrived at by majority rule:

   - for any cell, we choose a lower safe value if at least 2 of the dimensions had selected the lower safe otherwise we select the upper safe value; for even dimensions use a tie-breaker
   - we apply iterational proportional fitting (IPF) to obtain values for the non-sensitive cells

Improving the initial solution

1. Starting with this initial solution, we flip each of the sensitive cell values one at a time, use IPF to obtain values for the nonsensitive cells and compute the discriminant information. If the resulting discriminant information is minimum, we keep that combination. Otherwise, we discard it and keep the one we had previously
2. We continue this until the flipping lead to no changes in the discriminant information
3. The last value obtained is our solution

Illustration

Original table (sensitive cells marked yellow)

Col 1	Col 2	Col 3	Col 4	Col 5
4844.00	11958.00	10204.00	9100.00	25323.00
14628.00	16305.00	14984.00	3980.00	15565.00
12580.00	14464.00	20961.00	16993.00	9581.00
10282.00	7128.00	17178.00	21274.00	14893.00
21153.00	5088.00	20350.00	18186.00	5417.00

The first step is to find a local solution in each row and then in each column

Assume (1, 4) entry 9,100.00 requires 1,365.00 units adjustment
Assume (1, 5) entry 25,323.00 requires 3,798.00 units
Assume (4, 4) entry 21,274.00 requires 3,191.00 units
Assume (5, 2) entry 5,008.00 requires 764.00 units

How are solutions obtained?

Example row 1
In the first row, there are 4 possible combinations

Combination 1

4844.00	11958.00	10204.00	10465	21525.00
			+1365	-3798

Combination 2

4844.00	11958.00	10204.00	10465	29121
			+1365	+3798

Combination 3

4844.00	11958.00	10204.00	7735.00	29121
			-1365	+3798

Combination 4

4844.00	11958.00	10204.00	7735.00	21525.00
Deviation			- 1365	- 3798

Consider the 3rd combination

4844.00	11958.00	10204.00	7735.00	29121
Sum of original values of nonsensitive cells=27006			Sum of modified values for sensitive cells=36856	
Sum of the original values=61429				

To preserve the total within that row, we need to modify the original value of each nonsensitive cell by multiplying it by

Which yields:

4407.60	10880.69	9284.71	7735.00	29121

From these values, we compute the Kullback-Leibler for combination 3 in row 1:

$$K = 4407.6\log\left(\frac{4407.6}{4844}\right) + 10880.7\log\left(\frac{110880.7}{11958}\right) + 9284.7\log\left(\frac{9284.7}{10204}\right)$$
$$+7735\log\left(\frac{7735}{9100}\right) + 29121\log\left(\frac{29121}{25323}\right)$$

Performing the same operation for the other combination yield

For combination 1, K = 504.28
For combination 2, K = 897.55
For combination 4, K = 872.93

Hence in row 1, we choose combination 3

Initial solution for sensitive cells

After selecting the best combination in each row and column, we select for each sensitive cell whether to adjust up or down by majority rule

Row	Col	Org Data	Adjustment based on rows	Adjustment based on columns	Selection
01	04	9100.00	-	-	-
01	05	25323.00	+	+	+
04	04	21274.00	+	+	+
05	02	5088.00	+	+	+

Using IPF to adjust nonsensitive cells

Within each row, modify the nonsensitive cells so that sum of the modified values in that row equal the original total

Col 1	Col 2	Col 3	Col 4	Col 5
4844.00	11958.00	10204.00	7735.00	29121.00
14628.00	16305.00	14984.00	3980.00	15565.00
12580.00	14464.00	20961.00	16993.00	9581.00
10282.00	7128.00	17178.00	24465.00	14893.00
21153.00	5852.00	20350.00	18186.00	5417.00

In row 1, we need to modify each nonsensitive value by

$$\frac{\text{(Original row total - Sum sensitive Cells)}}{\text{Total original nonsensitive cell values}} =$$

$$\frac{61429 - 36856}{27006} = \frac{24573}{27006} = 0.91$$

In row 4, we multiply each nonsensitive cell by

$$\frac{70755 - 24465}{49841} = 0.93$$

In row 5, we multiply each nonsensitive cell by

$$\frac{70194 - 5852}{65106} = 0.99$$

This yields the table

4,407.60	10,880.69	9,284.71	7,735.00	29,121.00
14,628.00	16,305.00	14,984.00	3,980.00	15,565.00
12,580.00	14,464.00	20,961.00	16,993.00	9,581.00
9,618.92	6,668.32	16,070.20	24,465.00	13,932.56
20,904.78	5,852.00	20,111.20	17,972.59	5,353.43

Using above table (after adjusting nonsensitive values in each row), we adjust values of the nonsensitive cells in each column so that sum of values in each column add up to the original totals

	Column 1	Column 2	Column3	Column 4	Column 5
Sum orig. cells	63,487.00	54,943.00	83,677.00	69,533.00	65,362.00
Sum sens. cells		5,852.00		32,200.00	29,121.00
Sum nonsen. cells	62,139.30	48,318.01	81,411.11	38,945.59	44,431.99
Multiply nonsen. by	1.02	**1.02**	**1.03**	**0.96**	**0.82**

Second Iteration of IPF

We repeat the process by adjusting the nonsensitive cells within each row from the resulting table

4503.19	11054.76	9543.13	7735.00	29121.00
14945.26	16565.85	15401.04	3815.20	14593.24
12852.84	14695.39	21544.40	16289.38	8982.84
9827.54	6775.00	16517.48	24465.00	13062.72
21358.17	5852.00	20670.95	17228.41	5019.21

IPF Solution

4399	10840	9334	7735	29121	**61429**
14968	16651	15439	3815	14589	**65462**
12885	14787	21619	16299	8989	**74579**
9846	6813	16567	24465	13064	**70755**
21389	5852	20718	17219	5016	**70194**
63487	**54943**	**83677**	**69533**	**70779**	**342409**

The marginal totals are preserved

PERFORMANCE OF THE ALGORITHM

1. We verify how good the solution is by generating at least 5,000 combinations at random and compare our solution against the lowest discriminant information from that sample
2. Simple linear regression parameters between the modified and original tables should yield $b_0 \approx 0$ and $b_1 \approx 1$
3. Formal tests such as Kolmogorov-Smirnov can be used to detect whether the original and modified values have the same statistical distribution

Comparison with a random sample

Table Dim	Perc Sen Cell	MDI for Solution	MDI from random sample (or all combinations) (Q2.5, Q97.5)
10x10	5%	67.82	67.82 (67.82, 85.16)
10x10	10%	1695.72	1617.17 (1695.93, 2130.84)
10x10	20%	201.25	200 (213.38, 366.91)
10x10	30%	191.55	181.13 (198.95, 308.56)
20x20x20	10%	24542.62	26790.78 (27177.76, 28002.92)
20x20x20	20%	25167.26	27750.5 (27824.3, 28678.9)
20x20x20	30%	75290.4	85086.48 (86221.08, 89707.66)
30x30		175.196	174.47 (177.90, 181.32)
13x13x13		158.87	163.045 (166.456, 373.301)

Green Color=All combinations were computed
Yellow Color=Example from Salazar
No. of random samples = 5000

These results show that the algorithm leads to a solution that's almost always better than selecting the best solution from a sample of 5,000 solutions

Preservation of original distribution

Table Dim.	Percent Sens. Cell	b_0 regress. adjusted on original	b_1	Correlation	Mean pct. chng. to non-sens cells (min, max)
10x10	5%	-0.02	1.02	0.99	-0.00 (-0.04, 0.03)
10x10	10%	0.02	0.98	0.99	-0.00 (-0.03, 0.04)
10x10	20%	-0.00	1.00	0.99	0.00 (-0.05, 0.05)
10x10	30%	0.00	1.00	0.95	-0.01 (-0.11, 0.13)
20x20x20	10%	-0.01	1.00	0.97	-0.00 (-0.06, 0.05)
20x20x20	20%	0.00	1.00	0.97	-0.00 (-0.05, 0.05)
20x20x20	30%	0.01	0.99	0.92	-0.00 (-0.09, 0.09)
30x30		0.00	1.00	1	0.00 (-0.00, 0.00)
13x13x13		0.00	1.00	1	0.00 (-0.00, 0.15)

Preservation of original distribution: statistical tests

Table Dim.	Percent Sens.	K-S p-values: adjust & orig. from same distrib. (unconditional)	Kuiper p-values (uncondit.)	Chi-square p-values (conditional)
10x10	5%	1.00	1.00	1.00
10x10	10%	1.00	1.00	0.00
10x10	10%	0.97	0.98	1.00
10x10	30%	0.97	0.91	0.87
20x20x20	10%	0.60	0.16	1.00
20x20x20	20%	0.51	0.21	1.00
20x20x20	30%	0.056	0.00	0.00
4x9		0.88	0.97	0.00
30x30		0.00	0.00	0.00
13x13x13		0.00	0.00	1.00

LIMITATIONS/FUTURE IMPROVEMENT

1. A more optimal solution could be found by replacing values of sensitive cells with a value beyond the lower or upper bound
2. Marginal totals are held fixed. Sometimes a better solution could be found by allowing fluctuations in the marginal total
3. Heuristics may need to be improved when the dimensions of the table are even
4. Changes sometimes should be allowed to zero cells
5. Once, we have arrived at a final solution, it would be ideal to determine how much better it is compared to the solution coming from the random sample or to compute the probability of obtaining a better solution
6. Software developed is limited to a 30x30x30 table. Future version of the program should attempt to make it functional at least for a county level data set (one dimension of the table with equal or greater to 3,000)

CONCLUDING COMMENTS

- We presented a new algorithm for CTA based on Kullback-Leibler MDI
- Advantages of the method
  - always a unique solution
  - additivity to marginals preserved
  - original distribution preserved
- Results show that the algorithm leads to a solution that preserves the statistical distribution of the original values
- Future improvement will seek to obtain a more optimal solution and quantify how good the solution obtained is

REFERENCE

Cox, LH, Orelien, JG and Shah, B. A method for preserving statistical distributions subject to controlled tabular adjustment. In: **Privacy and Statistical Data Bases 2006, Lecture Notes in Computer Science 4302** (J. Domingo-Ferrer, L. Franconi, eds.). Heidelberg: Springer-Verlag, 2006, 1-11.