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

Lawrence H. Cox
LCOX@CDC.GOV

UNECE/Eurostat work session on statistical data confidentiality
Manchester, UK
17-19 Dec 2007

Views expressed are solely those of the author

OBJECTIVE

To examine and compare the effects on data quality of two SDL methods for tabular data based on controlled tabular adjustment (**CTA**)

- quality-preserving CTA (**QP-CTA**)
  Cox and Kelly, *Monographs of Official Statistics* (2002)
  Cox, Kelly and Patil, *LNCS 3050* (2004)

- minimum discrimination information CTA (**MDI-CTA**)
  Cox, Orelien and Shah, *LNCS 4302* (2006)

LOCAL & GLOBAL QUALITY

Local Data Quality

- masked cell values are close (as possible) to original data values
- time trends of individual cell values preserved

Global Data Quality

- parameters of original distribution preserved
- shape of original distribution preserved
- multivariate relationships preserved

REF: Cox, *Proc. 56th Session ISI* (2007)

SDL FOR TABULAR DATA

Tabular System: **Tx = v**

- **T** = tabular equations
  coefficients = 0, 1, possibly one -1 per row
- **x** = variable representing values of tabular cells
- **v** = constant value assigned to equation (often, 0)
- **a** = original data, so **Ta = v**

Sensitive Cells: cells failing a *disclosure rule*, such as

p-percent Rule: cell is sensitive if cell value minus second largest contribution is within p-percent of largest contribution

Complementary Cell Suppression (CCS): suppresses sensitive cells plus as many nonsensitive cells as necessary to ensure no interval estimate of suppressed sensitive data violates disclosure rule

Operational/Quality Characteristics of CCS

- NP-Hard problem
- leaves holes in data
- data NOT missing at random: thwarts analysis

CONTROLLED TABULAR ADJUSTMENT

Basic CTA Methodology

- replaces sensitive cell values with *safe values*: values outside the *protection interval* determined by the disclosure rule
- also adjusts some or all nonsensitive cell values to restore additivity to tabular system **Tx** = **v**
- nonsensitive adjustments typically small
- ease-of-use: CTA is unquestionably an improvement over CCS

MILP for Basic CTA

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

$$\mathrm{Ty} = \mathrm{T}(y^+ - y^-) = 0$$

$$p_i(1 - I_i) \le y_i^- \le m_i(1 - I_i), \quad p_i I_i \le y_i^+ \le m_i I_i$$

I_i binary, $i = 1, ..., s$

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

s = number of sensitive cells; n = number of cells
p_i = *lower/upper protection limit* for sensitive cell i
m_i = upper bound on adjustment to sensitive cell i
e_i = bound on adjustment to nonsensitive cell i
(often, e_i =measurement error)

$y_i = y_i^+ - y_i^-$ = (net) *adjustment* to cell value a_i
$\boldsymbol{a} + \boldsymbol{y}$= *adjusted (masked) data*
Other objective functions possible/useful

ILLUSTRATION OF CTA

(Nearly) Actual 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**

4x9 Table With (Protection Limits): 7 Sensitive Cells

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**

Table 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	**6575**
617	196	1899	1095	2172	3825	4372	260	797	**15233**
0	26	0	12	0	0	65	0	180	**283**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Table After Controlled Tabular Adjustment

OPERATIONAL AND QUALITY CHARACTERISTICS OF BASIC CTA

Operational

- preserves additivity
- far fewer (*s)* binary variables than CCS (*n*-*s*)
- heuristics enable solutions based on LP relaxation

Quality

- capacities on cell adjustments control local quality
- proper objective functions encourage global quality

QUALITY-PRESERVING CTA (QP-CTA)

Method

Define a linear functional

$$L(y)=(\sum_{i=1}^{n} a_i(y_i^+ - y_i^-))/Var(a)$$

and adjoin to the constraint system

$$L(y)=0$$

- if constraint $L(y)=0$ is infeasible, then incorporate $\min|L(y)|$ into the objective
- constraint forces adjustments to be orthogonal to data
- $L(y)=Cov(a,y)/Var(a)$, hence constraint forces $Cov(a,y)=0$

Quality-Preserving Capabilities of QP-CTA

Univariate (one original data set **a**)

Mean

- because additivity is preserved, means along tabular equations (rows, cols, etc.) are preserved
- other means can be preserved by incorporating appropriate constraints

Variance

- $|Var(a+y)/Var(a)-1|=|2L(y)+Var(y)/Var(a)|$
- $Var(y)/Var(a)$ is typically small
- so, $L(y)=0$ preserves variance

Correlation/regression of original/masked data

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

Multivariate (two or more related original data sets **a**, **b**)

- must also preserve $Cov(a,b)=Cov(a+y,b+z)$
- analogous technical results

Operational/Quality Characteristics of QP-CTA

Operational

Pro

- preserves additivity
- relies on standard LP software
- constraints, objective easily modified
- typically computationally efficient
- robust over tabular structure, dimension, size
- can be performed in a multivariate setting
- does not require that marginal totals be fixed

Con

- solvable exactly under specific conditions, but in general heuristics assign safe values
- objective/heuristics not statistically based

Data quality

Local Quality

- local adjustments can be minimized
- can exempt structural zeroes/other values from change
- can adjust nonstructural zero away from zero

Global Quality

- can minimize global/average distance
- preserves univariate/multivariate properties for standard and arbitrary subsets

MINIMUM DISCRIMINATION INFORMATION CTA (MDI-CTA)

Kullback-Leibler Minimum Discrimination Information

- measures distance btwn 2 statistical distributions defined on a probability space Ω
- first P is known and second Q^* is closest to P in MDI within a class of distributions
- $Q^* = \arg\min\{I(Q:P) = \sum_{\omega\in\Omega} Q(\omega)\log(\frac{Q(\omega)}{P(\omega)})\}$
- P = original distribution (table)
- class = tables satisfying specified marginal totals (*minimal sufficient statistics* = MSS)
- iterative proportional fitting (**IPF**) computes unique minimal MDI solution
- IPF permits fixing a subset of the cell values
  * sensitive cells set at selected safe values
  * structural zeroes

MDI-CTA

- arbitrary choice of safe sensitive cell values
- conditional on choice and MSS, IPF computes minimal MDI solution
- heuristic updates choice to improve MDI
- terminate when MDI btwn original & adjusted tables is statistically insignificant

Operational/Quality Characteristics of MDI-CTA

Operational

Pro

- preserves additivity
- relies on standard statistical algorithms available as software
- typically computationally efficient
- objective/heuristics tied to statistical criteria

Con

- heuristic assigns safe sensitive cell values
- not easily applied to arbitrary tabular structure
- no clear generalization to multivariate setting
- requires (some) marginal totals be fixed

Data quality

Local Quality

- can exempt structural zeroes/other values from change, but
  - * nonstructural zeroes fixed at zero
  - * no control on extent of local changes

Global Quality

- preserves original distribution, not just selected parameters or statistics

QUALITY COMPARISON OF QP-CTA AND MDI-CTA

Local Data Quality

Concerned with changes to cell values and trends

LOCAL QUALITY CHARACTERISTIC	QP	MDI
Preserve sensitive values close as possible	Y	Y
Preserve nonsensitive values closely	Y	N
Can exempt values from change	Y	Y
Can adjust nonstructural zeroes	Y	M
Preserves additivity	Y	Y
Can adjust MSS marginal totals	Y	N
Can preserve cell value trends	Y	N

M: Y if an epsilon-method is adopted

Global Data Quality

Concerned with changes to original distribution

GLOBAL QUALITY CHARACTER.	**QP**	**MDI**
Preserves univariate parameters/statistics	Y	*
Preserves distributional shape	M	Y
Preserves multivariate parameters/relations	Y	N

*: likely Y as MDI preserves shape

M: Y if changes to nonsensitive cells are

- within measurement error
- not overwhelmed by changes to sensitive cells

Operational Characteristics

Concerned with problem types solvable, computability, ease-of-use

OPERATIONAL CHARACTERISTIC	QP	MDI
Applicable to arbitrary tabular structure	Y	N
Requires heuristics for most problems	Y	Y
Can change marginal totals	Y	N
Easy to implement for complex structures	Y	N
Computationally efficient	Y	Y

CONCLUDING COMMENTS

QP-CTA and MDI-CTA are two methods based on controlled tabular adjustment for producing a quality-preserving, disclosure-limited set of tabulations from an original set of tabulations that contains disclosure

We presented mathematical/statistical models for applying these methods and examined their respective strengths and limitations operationally and for preserving data quality

We observe that often a strength of one method is a weakness of the other, and vice-versa, which motivated our comparison of their quality characteristics