QUANTITATIVE METHODS FOR BALANCING THE DATA QUALITY-DATA CONFIDENTIALITY TRADEOFF

Lawrence H. Cox
National Center for Health Statistics
LCOX@CDC.GOV

First Annual Conference on Quantitative Methods & Statistical Applications in Defense and National Security
Panel on Privacy Protection in an Era of Data Mining and Record Linkage
RAND Corporation
Santa Monica, CA
Feb 15-16, 2006

The confidentiality protection scenario

* most research/application of methods for *confidentiality protection*, aka *statistical disclosure limitation* (**SDL**), have arisen/resulted from **official statistics**

* national statistical office (NSO) collects data from individual persons/businesses/organizations for the purpose of creating and releasing/sharing a data set or summary (recently: making available a statistical data base query system)

* NSO collects these data under a pledge not to divulge directly or indirectly (confidential) data pertaining to these "individuals" to *unauthorized third parties*
    - **pledge** often required and backed by law
    - regarded as ethical conduct by statistics profession
    - in interests of NSO vis a vis respondent cooperation
    - **unauth. third parties** include legitimate data users
    - known as *the intruder*

* NSO must abbreviate/modify/treat/substitute/..... original/derived data to meet confidentiality requirements

* in Olden Times (pre-1970-1980)
    - **confidentiality protection** was blunt/unsystematic
    - lacked methodological/evaluation framework
    - trumped other considerations
    - paid little (formal) attention to data quality/utility

Confidentiality protection methods

* abbreviate data, e.g., fewer categories
* suppress data, e.g., delete selected data
* reduce data granularity, e.g., round/aggregate data
* statistically modify data, e.g., random noise/perturbation
* REALLY modify data, e.g., data swapping/encrypt files
* model data; release model/model outputs (*synthetic data*)
* *old lock & key*, e.g., research data centers

These methods

* have been developed/published/discussed/refined/tried
* revealed complex statistical/mathematical/computational issues & connections leading to important research
* until recently, paid little attention/not connected to data quality & utility

How to connect/combine data quality & data confidentiality?

* via inference

* properly confidentialized data meet standards
  that set an upper limit the extent to which the intruder can
    - infer/deduce the identity of individuals
    - infer/deduce individual data

* released data of proper quality meet standards
  that set a lower limit on the extent to which
    - statistical inferences based on released data are "equivalent to" statistical inferences based on original data
    - are consistent with/conform to other released data

An example: Controlled tabular adjustment (**CTA**)

* confidentiality concerns for tabular data result in NSO declaring the values of certain tabular cells to be *sensitive* and not releasable
    - small frequencies in contingency tables, e.g., count = 1 or 2
    - dominated cell values in aggregate data, e.g., sales total dominated by values of 1 or 2 firms
* previous approaches would suppress sensitive cells plus other cells to sufficiently hide sensitive values
* this is very difficult to do mathematically and creates many holes in the data that thwart statistical analysis
* **CTA**
    - replaces sensitive values with *safe values* (viz., values sufficiently far from true values)
    - adjusts other values to restore additive relationships
    - based on mathematical programming
    - result is a fully populated, additive set of tabulations

Quality-preserving controlled tabular adjustment (**QPCTA**)

* treats the combined data quality/confidentiality problem using linear constraints
* enforces all additive relationships
* keeps changes to individual nonsensitive cells small, e.g., within measurement error
* minimizes "distance" from original data, e.g., sum of absolute adjustments to individual cell values
* assures correlation = 1 between adjusted & original data
* preserves mean and variance-covariance relationships (correlations, regression coefficients) across variables
* based on keeping adjustments **orthogonal** to original data and to each other by means of linear functionals
    - original data = **a**
    - adjusted data = **a + y** (**y** = net adjustments)
    - incorporate constraint: $\mathbf{L(y) = \text{sum } a_i y_i = 0}$

Reference

LH Cox, JP Kelly and R Patil. Balancing quality and confidentiality for multivariate tabular data. **Privacy in Statistical Data Bases, Lecture Notes in Computer Science 3050** (J. Domingo-Ferrer and V. Torra, eds.). Berlin: Springer-Verlag, 2004, 87-98.

(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**

4x9 Table of Magnitude Data & Protection Limits for (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**

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

167	309	1284	579	4485	3981	2442	1150	91	**14488**
56	1479	172	667	1006	327	1667	1178	31	**6583**
617	202	1899	1098	2177	3825	4372	260	821	**15271**
0	52	0	24	0	0	81	0	100	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

After Optimal Quality-Preserving Controlled Tabular Adjustment

min {total abs. adjust. | net deviation ≤ 25%, additive, marginals} = 282 Units

Connnections to this conference, NDHS, etc., or lack thereof

NSO perspectives that appear different from NDHS needs

Statistical disclosure limitation

* typically applied to data products disseminated widely
* seeks to protect individuals from discovery/disclosure
* *protect* mostly means hiding or obfuscating information that **from some perspective** is salient/outlying/anomalous
* perspectives include
    - the data: at a glance or via mathematical analysis
    - insider knowledge
    - other data sources/files, e.g., via record linkage

Data quality in official statistics

* mostly means preserving statistical inference & structure
* often concerned with period-to-period consistency
* more focused on average/typical values than extremes
* may seek to preserve some individual values, but
* mostly not concerned about indiv. values/clusters/bumps

NSO perspectives that appear to align with NDHS needs

Statistical disclosure limitation

* develop mathematical methods for “disaggregating” data
* highly computationally intensive
* connected to record linkage
* subject to constraints expressed mathematically
* blends algebraic/combinatorial/statistical methods
* analogs/connections to cryptology
* skill set of researchers

Data quality

* relies on identifying & preserving the data “baseline”
* based on notion of nearness/distance of data & features
* can incorporate probabilistic structure & reasoning
* can incorporate constraints on local & global deviations
* examination/combination of multiple data sources
    - evaluation
    - distributed computing