WP 22
The Case For—Or Against—
Hybrid SDL Methods

Lawrence H. Cox
National Institute of Statistical Sciences
USA

UNECE Work session on statistical data confidentiality
Tarragona, Spain
October 24-26, 2011

INTRODUCTION

Hybrid SDL method

- combination of two or more SDL methods
- in a specified order

Examples

- microaggregation + perturbation
- CCS (suppression) + CTA
- CTA + CCS (suppression)
- synthesis + sampling
- swapping + (CTA, CCS, perturbation, …)
- MASSC

Motivations

- primary method provides insufficient protection
- secondary method to enhance/restore data quality/utility degraded by primary method
- confuse the intruder, thereby enhancing protection

Issues raised by hybrids

- need to enhance protection provided by primary method? by how much? how much incremental protection is provided by secondary method?
- incremental effects (+/-) of secondary method on data quality, analyzability and usability?
- how transparent is the hybrid?
- can the hybrid be expressed/evaluated analytically?
- does the result justify the complexity?
- is confusion a valid criterion to defend SDL?

This paper

- preliminary discussion of these issues
- based on 5 papers involving hybrids
  - # 3 UNECE, PSD proceedings
  - # 1 European Q-conference
  - # 1 JASA

THE PAPERS

Tabular SDL

- Castro and Giessing (2006)
    - CTA + CCS (suppression)
- Better and Kelly (2010)
    - swapping + CTA

Microdata SDL

- Flossman and Lechner (2006)
    - perturbation + blanking (suppression)
- Oganian and Karr (2006)
    - microaggregation + perturbation
- Dreschler and Reiter (2010)
    - synthesis + sampling

Relevant But Not Discussed

- RTI MASSC (2004): *Proc. 2003 UNECE*

TABULAR SDL PAPERS

Castro and Giessing (2006)

- perform CTA
- it is possible that full protection—particularly at higher levels/larger cells—cannot be achieved due to conflicts with a priori capacity constraints (including zero restrictions) on adjustments to nonsensitive cells
- apply RCTA where feasible; CCS on the remainder
- NOTE: CTA was always RCTA—viz., zero-restrictions and capacity constraints are an essential component of (QP-)CTA to preserve *local quality*
    - # Cox and Dandekar (2004): *Proc. 2002 FCSM*
    - # Cox and Kelly (2004): *Proc 2003 UNECE*
    - # Cox et al. (2004): *PSD 2004--LNCS 3050*

My Comments

- SDL in tabular data—magnitude/establishment data in particular—is driven by a disclosure rule: *linear sensitivity measure* (Cox 1981)
  - # count data: t-threshold rule
  - # magnitude data: p-percent rule
- rule quantifies minimal acceptable protection interval $P = [L_x , U_x]$
- both CTA and CCS methodologies/computations are driven by P
- CTA cell adjustment capacities are intended to control local quality
- there are times when local quality and global protection are in conflict
- there are strategies for dealing with that situation e.g., Cox and Dandekar (2004)
- CTA was developed to move beyond destruction caused by suppression
- falling back from CTA to cell suppression only confuses the user and degrades quality and usability

Better and Kelly (2010)

- perform swapping (matching) by solving assignment problem based on “optimal” weights--details of weighting proprietary
- other details sketchy--appears matching is based on optimizing or controlling some statistical criterion (unspecified) using metaheuristics
- p-values (sic) are mentioned
- if SDL is not completely successful (criteria unspecified), apply CTA

My Comments

- objectives quite unclear
- what weakness of CTA motivates this method?
- does quality mean conformity to certain predefined estimates? if so, CTA may be modified to do this
- swapping is “forever”, requires microdata, and can have unintended consequences—why do it?
- swapping at low levels can be weak SDL
- swapping at high levels can destroy quality
- what motivates doing more than CTA (alone)?
- authors confuse meaning of p-values: where is the statistical evaluation of quality?
- is this complexity motivated by a data protection or data quality need or simply to create something different?
- nearly total lack of transparency: proprietary?

MICRODATA SDL PAPERS

Flossman and Lechner (2006)

- perturbation + blanking (suppression)
- perturbation works well for smaller values but perturbations drawn from a single (additive) distribution become decreasingly effective as values increase in size; also, effectiveness of reidentification thru matching is enhanced in the presence of multiple perturbed variables
- motivation is analogous to Castro-Giessing
- methods for analysis are provided

My Comments

- as with tabular data, suppression thwarts analysis
- as with tabular data, perturbation/adjustment becomes more difficult at higher levels/larger cells
- perturbation methods adjusted to different scales or based on multiplicative/logarithmic models may be worth investigating

Oganian and Karr (2006)

- microaggregation + perturbation
- microaggregation is performed for SDL
- perturbation is performed to enhance variance attenuated by microaggregation (restore quality)
- perturbation also enhances SDL, viz., against matching, but this is secondary

My Comments

- this is a coherent hybrid
- motivation(s) for secondary method are clear
    # restore variance attenuated by primary
    # provide additional SDL to thwart reidentification via matching
- statistical properties of hybrid analyzable
- transparent, or potentially so

Dreschler and Reiter (2010)

- synthesis + sampling for a large (census) file
- synthesis for SDL
- sampling to enhance SDL and provide manageable, statistically representative file(s) (single or multiply imputed) for analysis and public use
- methods for analysis provided

My Comments

- synthesis is an established methodology for SDL with known/knowable quality characteristics
- synthesis at the census/large file level provides for richer models and enhanced quality (as measured by conformity to original distributions)
- sampling enables creation of manageable, analyzable masked files (for public use) with discoverable statistical properties
- single or multiply imputed masked files possible
- methods for analysis are available

CONCLUDING COMMENTS

Desiderata: SDL methods should

- quantify data protection
- quantify data quality
- preserve data utility
- strive to achieve maximum transparency

This is difficult to achieve even for single, analytically tractable methods such as

- rounding/perturbation
- CTA
- CCS
- probabilistic swapping/shuffling

Hybrid methods may only to muddy the waters—confuse the intruder—and in so doing likely also confuse the analyst and may fail to protect

Must be wary of *developing something different* only *to develop something different*

Emergence of commercial software complicates and confounds statistical evaluation and transparency