Producing Government Data With Statistical Confidentiality Controls

by

Tom Krenzke, Westat
Ed Christopher, Federal Highway Administration

Sponsored by

Website
http://community.amstat.org/CPC/Home

December 17, 2015

What is the CTPP?

CTPP is an umbrella program of

1. **Data Products**
2. **Custom Data Tabulations**
3. **Training**
4. **Technical Assistance**
5. **Research**

for the transportation community

CTPP's main focus is currently on data from U.S. Census Bureau and the American Community Survey (ACS)

Census Transportation Planning Products

Historical Highlights (Transportations Special Tab)

1960	Where did you work Last week? Modes Uses
1970	More address detail, TAZ Geography, Special Tab (resident, workplace and flow tables)
1980	More detail (more modes, veh occupancy, travel time) , more tables (crosstabs), Census Bureau Staff added (JTW branch)
1990	Departure time added, processing improved, State and Urban products, PC extraction software
2000	Special Tab Grew, Disclosure Rules began
ACS	Disclosure Rules found steroids

Special Tabulation Size

	Buyers/Users	Direct Cost	Tables
1960	OMB	???	???
1970	112	$0.6 M	43
1980	152	$2.0 M	82
1990	All States and MPOs	$2.5 M	120
2000		$3.0 M	203
2005+	AASHTO Consolidated Purchase	$5.8 M	Multiple Products

2014 raising another $3M to support the program

50 States, DC, 483 Metropolitan Planning Organizations

5-year CTPP Data Product (2010 CTPP5)

Product Structure

3-Parts

Part 1- Residence

Part 2- Workplace

Part 3- Flows between Home and Work

On-Line Data Retrieval

Extraction Software

Raw Data Download

CTPP 5-Year Main Product

October 31, 2013

2006, '07, '08, '09, 2010

Small Areas
(TAD, Tract, TAZ, Block Group)

New TAZs and TADS

Modeled and Actual Data and Flows

Requires Disclosure Proofing

5

Unpublished Disclosure Rules

DRB Said…"Too many variables" crossed with Means of Transportation (Mode)

Total
Drove Alone
2 Person Carpool
3 Person Carpool
4 Person Carpool
5-6 Person Carpool
7+ Person Carpool
Bus/Trolley Bus
Streetcar/Trolley
Subway/Elevated
Railroad
Ferryboat
Bicycle
Walked
Taxicab
Motorcycle
Other Means
Worked at Home
18 Modes

- Age
- Class of Worker
- Disability status
- Earnings
- Household Income
- Poverty status
- Industry
- Occupation
- Length of U.S. residence
- Minority status (Y/N)
- Time Leaving Home
- Time Arriving (Part 2)
- Travel Time
- Vehicle Availability
- Workers in Household
- Age of Youngest Child

…makes for micro data record

…and with a micro data record you could identify an individual

(2010 CTPP5) Product Summary

Highlights	Low Lights
Based on CTPP2000 Tables Many NEW 1-way Tables More Age Tables Streamlined Race Tables More HH and HH Lifecycle Tables More Geographic breakdowns or levels New TADs Way more Flows Tables	Rounded Limited Number of Crosstabs with Mode -- Travel time -- Household income -- Vehicle availability -- Age -- Time leaving home -- Minority status -- Presence of children Tables will have Disclosure Proofing Large MOEs (@ 90%)

Introduction

- Statistical Disclosure Control (SDC) techniques
  - "… the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. SDC methods minimise the risk of disclosure to an acceptable level while releasing as much information as possible." – Hundepool et al. (2012)

Introduction (2)

- Provide practical insights for data producers of US government surveys to balance…
  - Risk
  - Utility
  - Operational feasibility

  → Duncan, Keller-McNulty, Stokes (2001)

- Outline
  1. Set the stage before data collection
  2. Get to the details during data collection
  3. Apply SDC after data collection
  4. SDC from a data user perspective

Set the Stage Before Data Collection

- Motivation -- Laws
  - Privacy Act of 1974 (Section 552a)
  - HIPAA for patient privacy protections (OCR, 2012)
  - Office of Management and Budget (OMB, 1997)
- Relevant Agency Standards and Practices
  - Census Bureau Disclosure Review Board (DRB)
    - http://www.census.gov/srd/sdc/
    - http://www.census.gov/srd/sdc/FR_23693-94.pdf
  - National Center for Education Statistics (NCES) Standards
    - http://nces.ed.gov/statprog/2002/std4_2.asp
  - National Center for Health Statistics (NCHS)
    - http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf
  - Federal Committee on Statistical Methodology Working Paper 22 (FCSM, 2005)
    - http://www.fcsm.gov/working-papers/SPWP22_rev.pdf

Set the Stage Before Data Collection (2)

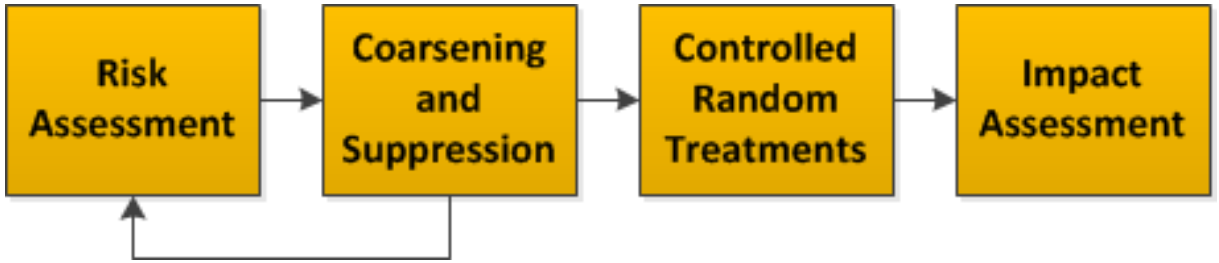
- Establish the Modes and Access Levels of Dissemination
  - Some modes and levels of access
    - Restricted use file (RUF)
    - Public use file (PUF)
    - Remote access to RUF (e.g., NCHS)
    - Real-time on-line analytic system (OAS) from a RUF
    - OAS from a PUF
      - *Census Bureau's DataFerrett*
    - OAS from static tables
      - *Census Bureau's American FactFinder*
    - Static tables
      - *CTPP*
      - *Reports*

Get to Details During Data Collection

- Written plan for SDC treatments
  - Approval by the DRB needed by end of data collection
- Get familiar with data
  - Variables to treat (target variables)
  - Item types (continuous, unordered categorical…)
  - Missing value codes
  - Imputed values
- Plan operations
  - Data flow
  - Timelines (weighting process)
  - Computer programs

Apply SDC After Data Collection

- General goals for applying SDC treatments
  - Balance risk reduction with retention of data utility, while optimizing operations and timelines
- SDC process



- Components of the SDC process depend on modes and level of access

Apply SDC – PUF: Risk Assessment

- Risk scenarios (El Emam et al., 2009)
- Risks within the internal data set, indirect identifiers
  - Personal identifiable information (PII)
  - Sample design and weighting variables
  - Geographic detail
  - Demographics
  - Contextual variables
  - Outliers (continuous variables, spatial)
- Review responses to open ended questions

Apply SDC – PUF: Risk Assessment (2)

- Combinations of variables (Sweeney, 2002)
- Re-identification risk
  - Probability that a sample unique is unique in the population
- Example approaches
  - Exhaustive n-way tabulations (*InitialRisk, NCES)*
  - Special Unique Detector Algorithm (SUDA) (Elliot, 2002)
  - Log-linear models (Skinner and Shlomo, 2008)
  - Mu-Argus (mainly developed at Statistics Netherlands)

Apply SDC – PUF: Risk Assessment (3)

- Sources of Risk – External Files
  - Exact and statistical matching (record linkage) on common indirect variables to obtain PII
  - Summary in Winkler (1993)
    - https://www.census.gov/srd/papers/pdf/rr93-8.pdf
  - Diniz da Silva, et al. (2010), evaluation of…
    - Link Plus (CDC)
    - RELAIS (ISTAT)
    - FEBRL (Australian National University and the New South Wales Dept of Health)
  - Fine-grained Record Integration and Linkage Tool (FRIL) from CDC

Apply SDC – PUF: Coarsening and Suppression

- Coarsening
  - Categories – combine categories
  - Continuous variables -- specified categories
    - Top-codes
- Variable suppression
  - Open-ended items
  - Items with 2 categories where one is sparse
- Rerun risk assessment
- Controlled random treatments
  - E.g., American Community Survey Public Use File
    - Perturbation
    - Subsampling

Apply SDC – PUF: Controlled Random Treatments

- Goals
  - Maintain the true underlying distribution
  - Preserve structured patterns
  - Minimize Mean Square Error = Variance + $Bias^2$
- Identify treatment rate and target variables
- Gather predictor variables
- Some slippery slopes
  - Treating each item independently
  - Treating without best predictors available
  - Treating without attention to missing value codes

Apply SDC – PUF: Controlled Random Treatments (2)

- Rank swapping (Greenberg, 1987)
  - Records close in rank on a sorted variable are designated as pairs for swapping values
  - Software – Mu-Argus
- Data swapping (Summary in Fienberg, 2005)
  - General steps
    - Select target records
    - Find swapping partners by matching on characteristics
    - Swap data values
  - Software
    - Data Swapping ToolKit, by NISS
    - *DataSwap,* by NCES

Apply SDC – PUF: Controlled Random Treatments (3)

- Parametric
  - Model-based multivariate sequential replacement
    - IVEWare, Raghunathan et al. (2001)
- Semi-parametric
  - Model-assisted constrained hot deck (more later)
- Non-parametric approaches
  - Classification trees, Reiter (2005) and Dreschler and Reiter (2011)
- Other approaches
  - Data shuffling (Muralidhar and Sarathy, 2006)
  - FCSM (2005)
  - Spatial (Wang and Reiter (2012); Paiva et al (2013))

Apply SDC – PUF: Controlled Random Treatments (4)

- Account for treatment error component in variances
  - — Multiple imputation approach (Summary by Reiter, 2009)
- Some diagnostics
  - — Frequencies, Skip pattern checks, Mean within table cells, Correlations, Scatterplots, Regression coefficients
  - — Global utility measures (Woo, et al 2009)

Apply SDC – Online Analytic Systems

- Provides data (estimates) to the public
  - — Generate from public microdata – no issues
  - — Generate from restricted microdata
- OAS real-time tabulators
  - — Developing Microdata Analysis System (Freiman et al., 2011) at Census Bureau
  - — Developing Online Analytic Real-time System (Gentleman, 2011) at NCHS
  - — Australian system (Tam, 2011)

Apply SDC – OAS (2)

- Intruder attacks
  - Table differencing

Universe 1

	B	
A	1	2
1	10	10
2	10	10

—

Universe 2

	B	
A	1	2
1	10	10
2	10	9

=

Difference

	B	
A	1	2
1	0	0
2	0	1

  - Averaging
  - Link implicit tables
    - ✦Obtain characteristics of sliver
  - Record linkage
    - ✦Use the sliver's characteristics to match to public files to attach small geography from OAS

Apply SDC – OAS (3)

- Treat underlying microdata
- Real-time system approaches
  - Threshold rules and table denials
  - Post-tabular adjustments or dynamic subsampling
  - Rounding
- Risks, properties, and approaches summarized in Krenzke et al. (2013b)

Apply SDC – Static Tables

- Census Transportation Planning Products
  - Pre-specified tables
  - Generated from 2006-2010 American Community Survey data
  - Tables to be generated from the ACS 5-year data
    - Residence
      - *Means of Transportation (MOT)*
      - *Demographics variables*
    - Workplace
    - Flows
      - *E.g., Mean travel time*

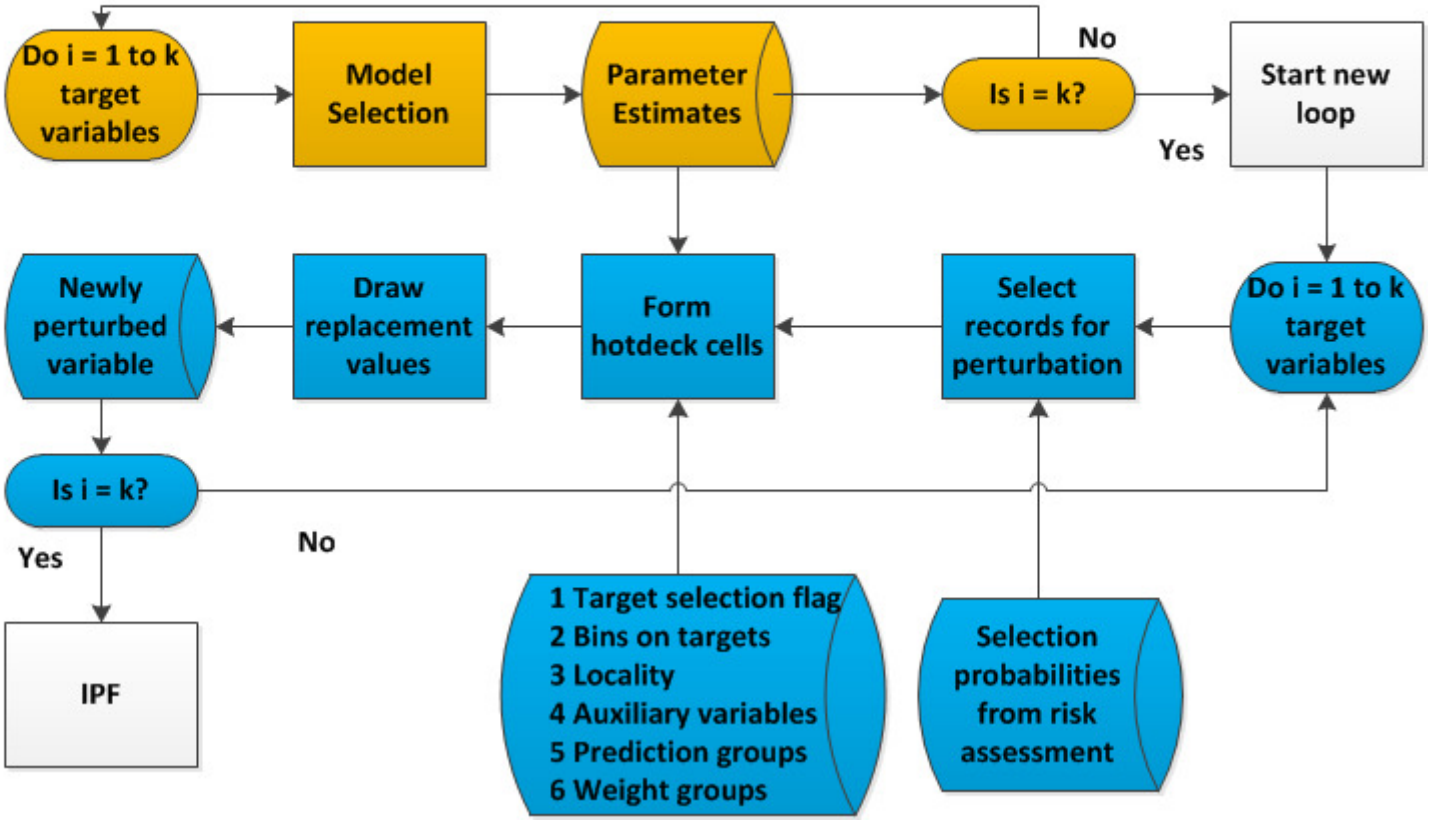
Apply SDC – Static Tables (2)

- CTPP (continued)
  - Set A tables
    - Very few Census Bureau DRB rules
    - Generated tables from original ACS microdata
  - Set B tables
    - Census Bureau DRB rules
    - Generated tables from perturbed microdata
    - Rules lifted
  - Tables have been published
  - Model-assisted constrained hotdeck (Krenzke et al., 2011, 2013a)

Apply SDC – Static Tables (3)

Model-Assisted Constrained Hotdeck

Suppose there are k target variables



Summary

- Practical SDC tools
  - — Government data → public
- Data producers
  - — Production-oriented setting
- Recent applications
  - — Online analytic systems
  - — Perturbation through the CTPP application
- ***Over to Ed…***

References

- Diniz da Silva, A., SanAna Martins Romeo, O., Silva Soares, T. and Layter Xavier, V. (2010). Study of record linkage software for the 2010 Brazilian Census Post Enumeration Survey. *The Survey Statistician*. Book and Software Review, pp 31-39.
- Dreschler and Reiter (2011). An empirical evaluation of easily implemented non-parametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*. 55:3232 -3243.
- Duncan, G.T., Keller-McNulty, S. (2001). Disclosure risk vs. data utility: the R-U confidentiality map. Technical Report. Statistical Sciences Group. Los Alamos National Laboratory.
- El Emam, K., Dankar, F., Vaillancourt, R., Roffey, T., and Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian Journal of Hospital Pharmacy* 62(4):307-319.
- FCSM (2005). Report on Statistical Disclosure Methodology. Statistical Policy Working Paper 22 of the Federal Committee on Statistical Methodology, 2nd version. Revised by Confidentiality and Data Access Committee 2005, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget

References (2)

- Freiman, M., Lucero, J., Singh, L., You, J., DePersio, M., and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau. Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.
- Gentleman, J. F. (2011). A Real-Time Online System for Analyzing Restricted Data from the U.S. National Center for Health Statistics' National Health Interview Survey. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: http://isi2011.congressplanner.eu/pdfs/650208.pdf (Accessed February 1, 2013).
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.-P. (2012). Statistical Disclosure Control. Chichester, UK: John Wiley & Sons.
- Karr and Reiter (2014). Using statistics to protect privacy. Chapter 13 of *Privacy, Big Data and the Public Good: Frameworks for Engagement*. Edited by Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. Cambridge University Press.
- Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R., and Larsen, M. (2011). Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules. Contractor's Final Report for NCHRP 08-79. National Cooperative Highway Research Program, Transportation Research Board of the National Academies. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w180.pdf (accessed 12/8/2014)

References (3)

- Krenzke, T., Gentleman, J., Li, J., and Moriarity, C. (2013b). Addressing disclosure concerns and analysis demands in a real-time online analytic system. *Journal of Official Statistics*, Vol 29, No. 1, pp. 99-134.
- Krenzke, T., Li, J., and Zayatz, L. (2013a). Balancing Use of Weights, Predictions, and Locality Effects in a Model-Assisted Constrained Hot Deck Approach for Random Perturbation. Proceedings of the Joint Statistical Meetings, American Statistical Association.
- Muralidhar, K. and Sarathy, R. (2006). Data shuffling: A new masking approach for numerical data. *Management Science*, Vol. 52, No. 5 (May, 2006), pp. 658-670
- OCR (2012). Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Office of Civil Rights. Published on website November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
- Paiva, T., Chakraborty, A., Reiter, J., and Gelfand, A. (2013). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine*. DOI: 10.1002/sim.6078.

References (4)

- Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics,* 21(3), 441–462.
- Reiter, J. (2009). Multiple Imputation for Disclosure Limitation: Future Research Challenges. *Journal of Privacy and Confidentiality*, 1, No 2, pp 223-233.
- Tam, S. (2011). On-line Access of Micro-Data at the Australian Bureau of Statistics – Challenges and Future Directions. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: isi2011.congressplanner.eu/pdfs/650030.pdf (Accessed February 1, 2013).
- Wang and Reiter (2012). Multiple imputation for sharing precise geographies for public use data. *Annals of Applied Statistics*. 6(1): 229-252.
- Winkler, W. (1993). Matching and Record Linkage. U.S. Census Bureau.
- Woo, M., Reiter, J., Oganian, A., and Karr, A. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* 1:111-124.

The Outcome

Bottom Line

Perturbation is/was not a big deal (YET)

Too many other ACS issues to overcome

- **Large MOEs (w/ normal ACS)**
- **Multi-year Data**
- **Aggregating Variables and/or zones**
- **Current ACS vs CTPP**
- **Knowing what to use**

Table Summary using 5-year Table list

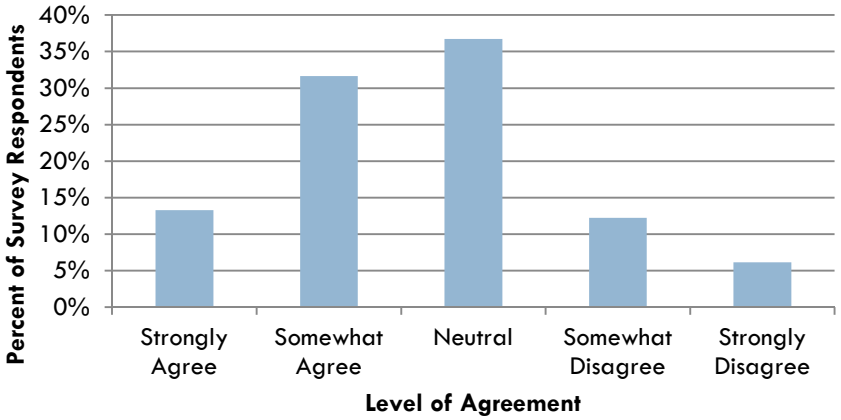
	TAZ/BG	Tract	TAD	Place	County	PUMA	State
Part 1							
Regular	111	→	→	→	→	→	→
Perturbed	77	→	→	→	→	→	→
Part 2							
Regular	50	→	→	→	→	→	→
Perturbed	65	→	→	→	→	→	→
Part 3							
Regular	2	→	→	→	→	→	→
Perturbed	38	→	→	→	→	→	→

Tables Using Perturbed Data Set

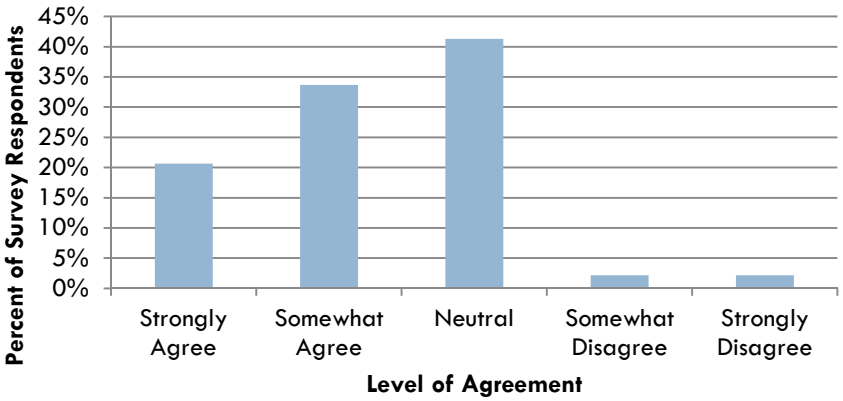
Means of transportation
Aggregate Travel Time
Aggregate HH Income
Almost all Part 3 Tables
Aggregate Vehicles Used
Mean HH Income
Aggregate Carpools

CTPP Based on Disclosure Proofed Data – Survey Findings

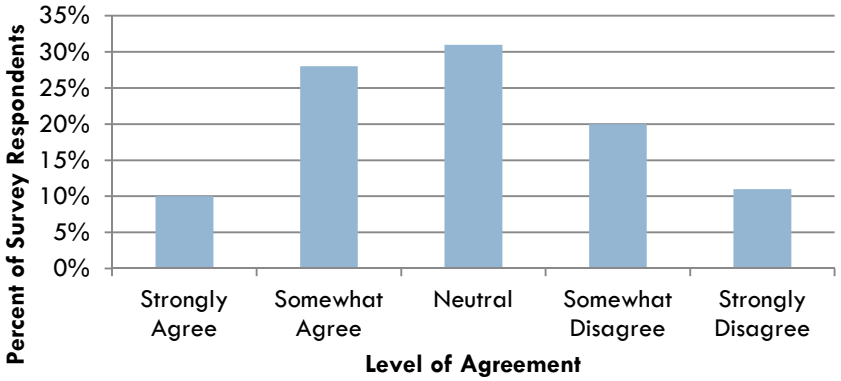
I understand the general methods used for disclosure proofing



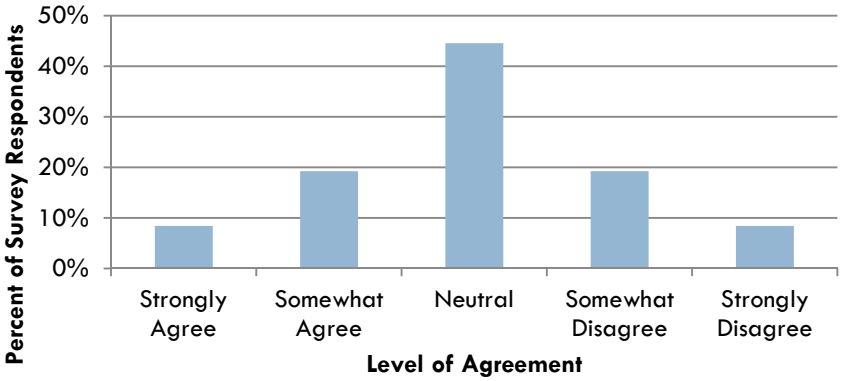
Having the disclosure proofed tables (B tables) is preferable to having tables with supressed values



Having both unmodified (A tables) and modified (B tables) is confusing

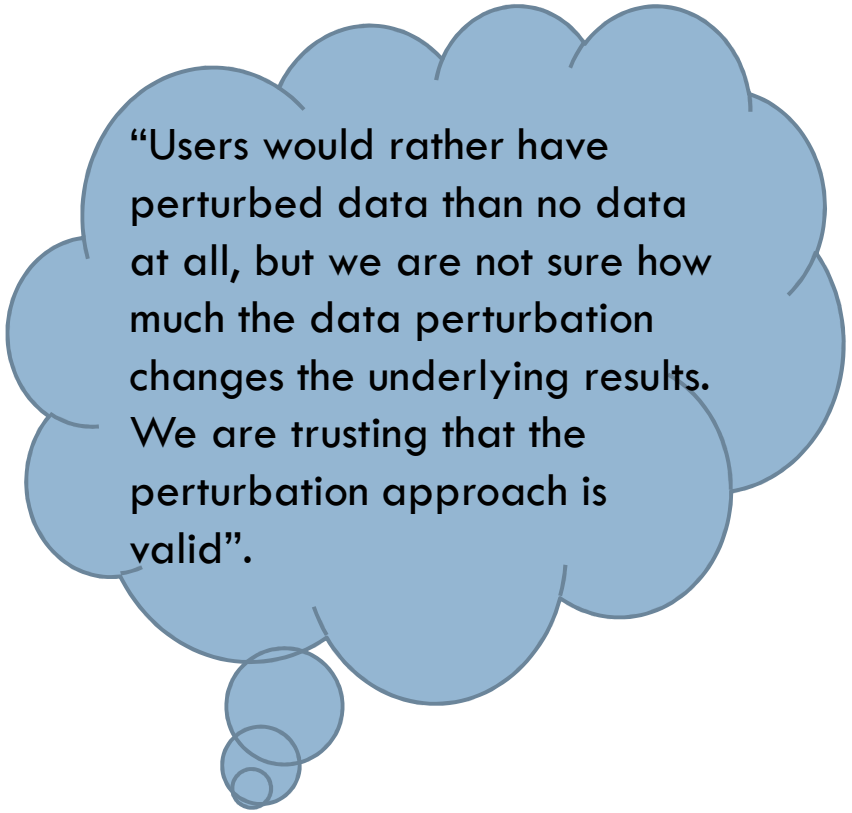


I use the disclosure proofed tables (B tables) without reservation



Peer Review: Disclosure Proofing

- Rather than documentation of the perturbation, participants would like to see simple results of comparisons between raw and disclosure proofed data



Questions

Presentations will be posted to Committee on Privacy and Confidentiality Website

http://community.amstat.org/CPC/Home