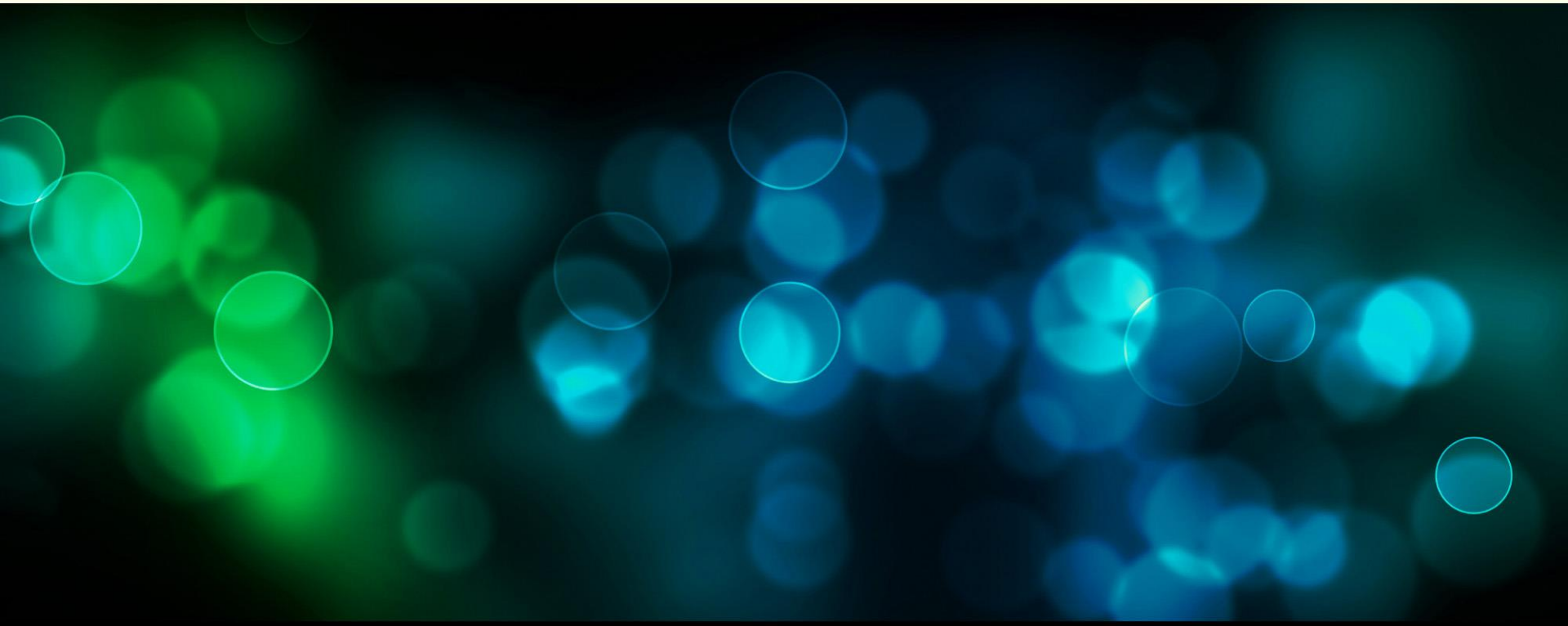
Government Data and Confidentiality: Compatible Companions With the Help of Statistical Disclosure Control



APDU Webinar
Tom Krenzke, Westat
Clara Reschovsky, Metropolitan Washington Council of Governments
September 25, 2013

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)

Webinar Goal and Outline

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

  → 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)
  - Protects records maintained on individuals
- HIPAA for patient privacy protections (OCR, 2012)
  - Safe harbor, Statistical expert review
- Office of Management and Budget (OMB, 1997)
  - Policies on confidentiality of statistical information
- Confidential Information Protection and Statistical Efficiency Act of 2002, Education Reform Act of 2002, US Patriot Act of 2001, etc.

Set the Stage (2)

Review Relevant Agency Standards and Practices

- Census Bureau documents used by the Disclosure Review Board
  - http://www.census.gov/srd/sdc/
  - http://www.census.gov/srd/sdc/FR_23693-94.pdf
- National Center for Education Statistics Standards
  - http://nces.ed.gov/statprog/2002/std4_2.asp
- National Center for Health Statistics
  - 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 (3)

- Other helpful resources
  - Confidentiality and Data Access Committee (CDAC)
    - Established by OMB and FCSM
    - http://www.fcsm.gov/committees/cdac/about.html
  - ASA Committee on Privacy and Confidentiality
    - http://www.amstat.org/committees/pc/

Set the Stage (4)

Establish the Modes and Access Levels of Dissemination

- Some modes and levels of access
  - Restricted use file (RUF)
  - Remote access to RUF (e.g., NCHS)
    - Agency analysts review output, and provide results
  - 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 FactFinderStatic tables, tables in reports

Get to Details During Data Collection

Draft written plan for review by agency DRB

- Initial risk analysis
  - — Identify risk scenarios and high risk values
    - ✦Risk scenarios -- El Emam et al. (2009)
      - ◆ *Prosecutor – Looking for a specific person*
      - ◆ *Journalist – Not looking for a specific person*
- Data coarsening plans
- Perturbation plans
- Impact evaluation

- Approval of plan by DRB needed around the end of data collection

Get to Details (2)

Computer Programs

- Develop programs for risk assessment, treatments, impact
- Review existing software… some examples
  - Risk assessment
    - Mu-Argus (mainly developed at Statistics Netherlands)
    - *InitialRisk (NCES)*
    - SUDA (University of Manchester)

Get to Details (3)

Computer Programs (Continued)

- Existing software examples
  - Perturbation
    - Data swapping
      - *Data Swapping ToolKit (National Institute of Statistical Sciences)*
      - *DataSwap (NCES)*
        - *Includes impact analysis*
    - Mu-Argus (mainly developed at Statistics Netherlands)
      - *E.g., rank swapping, Noise added to weights*
    - IVEWare (Survey Research Center, Institute for Social Research, University of Michigan)

Apply SDC After Data Collection

- General goals for applying Statistical Disclosure Control (SDC) treatments
  - Balance risk reduction with retention of data utility
  - Keep in mind operations and timelines
    - E.g., File management
    - E.g., Perturbation of target variables that are used in weighting process need to be done before weighting starts
- Components of the SDC process depend on modes and level of access

Apply SDC – PUF

General Process

- Pre-SDC steps
- Identify high risk values
- Coarsen values (recode) and determine variables to suppress
- Determine need for further SDC treatment (random perturbation)
- Evaluate the risk level and data utility

Apply SDC – PUF: Pre-SDC Steps

Identify

- Variables to treat (target variables)
  - — Typically factual identifying characteristics
- Item types (continuous, unordered categorical…)
- Missing value codes
  - — 7, 8, 9s

Apply SDC – PUF: Identify High Risk Values

Sources of Risk – External Files

- Record linkage
  - Exact matching and Statistical matching
  - 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)
    - Others

Apply SDC – PUF: Identify High Risk Values (2)

- Risks within the internal data set
  - Personal identifiable information (PII)
  - Sample design and weighting variables
  - Geographic detail
  - Contextual variables
  - Outliers (continuous variables, spatial)
- Run summary statistics for each variable
  - Start making data coarsening decisions
- Review responses to open ended questions

Apply SDC – PUF: Identify High Risk Values (3)

Review Combinations of Variables

- Only a few variables are needed to identify a sample unique
  - Exhaustive tabulations identify sample uniques or sparse combinations of variables
    - k-anonymity (Sweeney, 2002)
  - Special Unique Detector Algorithm (SUDA) (Elliot, 2002)
- Re-identification risk
  - Log-linear models (Skinner and Shlomo, 2008)
  - Mu-Argus – sampling weights incorporated

Apply SDC – PUF: Identify High Risk Values (4)

- Possibly use risk measures when statistical matching risk assessment to many publicly available files is problematic when attempting to cover all possible files
- Agencies may want to use risk measures to
  - Re-assess their current confidentiality rules
  - Set risk thresholds for their studies

Apply SDC – PUF: Coarsening and Suppression

- Recodes
  - Categories – Combine categories
  - Continuous variables -- specified categories
    - Top-codes -- Weighted average for those over cutpoint
- Variable suppression
  - Open-ended items
  - Items with 2 categories where one is sparse
- After coarsening, rerun risk assessment
- If risks remain, consider further SDC treatments
  - E.g., American Community Survey Public Use File
    - Random perturbation
    - Subsampling

Apply SDC – PUF: Random Perturbation

- Treatment goals
  - Maintain the true underlying distribution of the data (point estimates, variances, multivariate associations, shape of the data)
  - Preserve structured patterns
  - In general, minimize
    - Mean Square Error = Variance + Bias^2
- Pre-treatment steps
  - Identify perturbation rate, data patterns, distributions of variables
  - Create a pool of predictor variables
- Some slippery slopes
  - Perturb each item independently
  - Perturb without best predictors available
  - Perturb without attention to missing value codes

Apply SDC – PUF: Random Perturbation (2)

Examples of Random Perturbation Approaches

- Rank swapping (Greenberg, 1987)
  - Records close in rank on a sorted variable are designated as pairs for swapping. Typically the sorting variable is the target variable.
- Data swapping (Summary in Fienberg, 2005)
  - Used by Census Bureau, NCES for example
  - Select target record
  - Find swapping partner by matching on characteristics
  - Swap data values

Apply SDC – PUF: Random Perturbation (3)

Examples (Continued)

- Parametric approaches
  - Approaches implemented in IVEWare, Raghunathan et al. (2001) – Multivariate sequential replacement
    - E.g., For continuous variables, fit a linear regression with draws of random normal deviates to add noise for data replacement
- Semi-parametric approaches
  - Model-assisted constrained hot deck (MACH) (Krenzke et al, 2013a) -- Multivariate sequential replacement
- Each application is different
  - Other approaches
    - FCSM (2005)
    - Data shuffling (Muralidhar and Sarathy, 2006)

Apply SDC – PUF: Random Perturbation (4)

- Account for perturbation error component in variances
  - Multiple imputation approach (Summary by Reiter, 2009)
- Some diagnostics
  - Before and after perturbation
    - Frequencies
    - Skip pattern checks
    - Mean within table cells
    - Correlations
    - Scatterplots
    - Regression coefficients

Apply SDC – OAS

- OAS
  - Provides data (estimates) to the public
- Generate from public microdata – no issues
- Generate from restricted microdata
  - Risk of forming a pseudo-microdata record gathered from OAS estimates and attach restricted data to other files
- OAS examples
  - Developing Microdata Analysis System (Freiman et al., 2011) at Census Bureau
  - Developing On-line Analytic Real-time System (Gentleman, 2011) at National Center for Health Statistics
  - Australian system (Tam, 2011)

Apply SDC – OAS (2)

- Table differencing
  - Two slightly different universes are queried
    - Both pass the threshold rules
    - Difference between the tables
      - *'Sliver', Implicit table*
- Link implicit tables
  - Obtain characteristics of sliver
- Record linkage
  - Use characteristics to match to public files to attach small geography from OAS

Apply SDC – OAS (3)

Protections and Treatments

- Perturb underlying microdata
- Real-time system approaches
  - Threshold rules
    - Unweighted count needs to be greater than X
    - Universe, Table marginal, Cells
  - Post-tabular adjustments or dynamic subsampling
  - Rounding
  - Summarized in Krenzke et al. (2013b)

Apply SDC – Static Tables

Example -- Census Transportation Planning Products

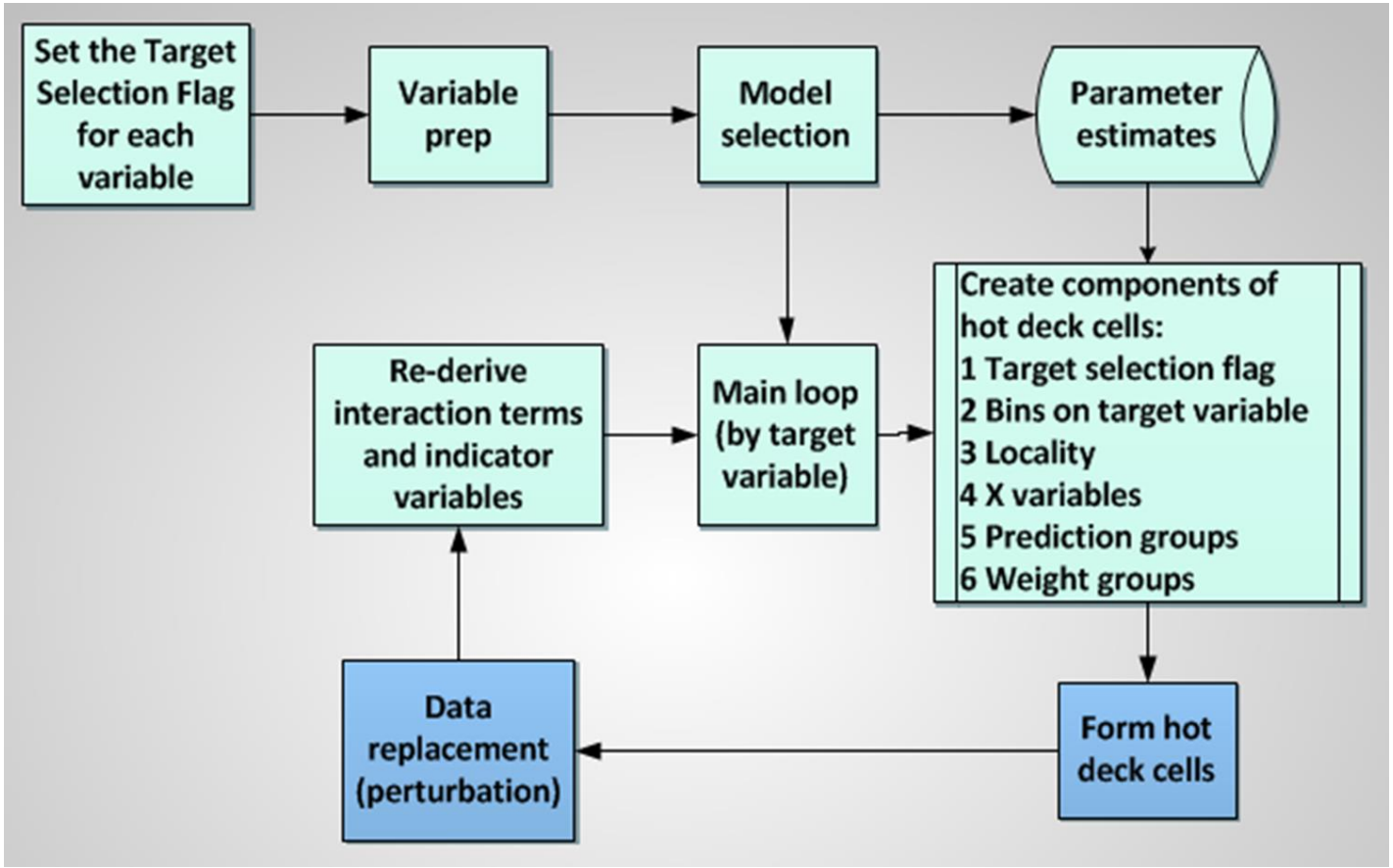
- Pre-specified tables
- Generated from 2006-2010 American Community Survey data
- Tables to be generated from the ACS 5-year data
  - Part 1 Residence
    - Means of Transportation (MOT)
    - Demographics variables
  - Part 2 Workplace
  - Part 3 Flows
    - E.g., Mean travel time

Apply SDC – Static Tables (2)

Example -- Census Transportation Planning Products (continued)

- Set A tables
  - No Census Bureau DRB rules
  - Generated tables from original ACS microdata
- Set B tables
  - Census Bureau DRB rules
  - Generated tables from perturbed microdata (via MACH approach)
  - Rules lifted
- Tables will be available in the very near future

Apply SDC – Static Tables (3)



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.
- 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.
- 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.

References (3)

- 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

References (4)

- 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).
- Winkler, W. (1993) Matching and Record Linkage. U.S. Census Bureau.

Thank You