Can a Synthetic Data Approach Applied to High Risk Data Result in Usable Data with a Very Low Risk?

Application to the Federal Employee Viewpoint Survey

CDAC/FCSM Workshop

New Advances in Disclosure Limitation

Bureau of Labor Statistics

September 27, 2017

Taylor Lewis[1], U.S. Office of Personnel Management

Tom Krenzke[2], Westat

[1] The opinions, findings, and conclusions expressed in this presentation are those of the author and do not necessarily reflect those of the U.S. Office of Personnel Management.

[2] The author would like to acknowledge Jane Li and Lin Li for their assistance in carrying out this research.

Outline

I. Background on Traditional Disclosure Avoidance Strategies and Those Applied to the Federal Employee Viewpoint Survey (FEVS)

II. FEVS Synthetic Data Application
   - Methodology
   - Data Utility Assessments
   - Risk Assessments

III. Summary and Further Research Questions

Traditional Strategies Reducing Disclosure Risk

- Information reduction:
  - Top coding → capping ages at “60+”
  - Rounding → converting income into ranges
  - Dropping variables
  - Separate files with separate sets of variables
  - Sampling
  - Suppression → deleting certain values

- Data perturbation:
  - Swapping values across two or more records
  - Noise infusion (e.g., adding random errors) – generally more applicable for continuous variables

Problems with Traditional Strategies

- Information reduction:
  - Dropping variables degrades overall data utility
  - Combining/collapsing may hide key relationships in data
  - Suppression might produce data that are not missing completely at random (MCAR) (Little and Rubin, 2002) and can reduce precision of estimates

- Data perturbation:
  - Data swapping maintains (unweighted) marginal distributions, but analyses involving the swapped and un-swapped variables jointly can be distorted (Reiter, 2012)
  - Noise infusion can also attenuate correlations and distort relationships amongst variables

Synthetic Data to the Rescue?

- First proposed by Rubin (1993), generating synthetic data is a promising (and rapidly evolving) methodology that addresses many of the traditional strategies' limitations

- Premise: model the observed data and use that model to produce plausible substitute values

- Two types of synthetic data:
  - Fully synthetic data (Raghunathan et al., 2003) – all values are synthesized
  - Partially synthetic data (Reiter, 2003) – only some values are synthesized (either a portion of variables, a portion of records, or some combination of both)

Synthetic Data: Advantages and Disadvantages

- Key advantage of fully synthetic data: because no actual values are released, disclosure risk is extremely low
- Attempts to match synthetic data with external databases for purposes of disclosure are pointless
- Partially synthetic data better maintains relationships in the data, but increases disclosure risk
- Key disadvantage of synthetic data: relationships omitted from model will not appear in the synthetic data → it is only possible for analysts to rediscover what is accounted for by the synthesis models

Background on the FEVS

- The Federal Employee Viewpoint Survey (FEVS) is an annual, Web-based survey of full-time, permanent, non-seasonal federal employees administered by the U.S. Office of Personnel Management (OPM)
- As of 2016 FEVS: sample size ~900,000; 80+ agencies participating; response rate just under 50%
- Instrument consists mainly of attitudinal items (e.g., perceptions of leadership, job satisfaction) on a Likert-type scale, but also about a dozen potentially observable demographics
- Highly detailed individual-level, work-unit information is provided by agencies for sampling/reporting purposes

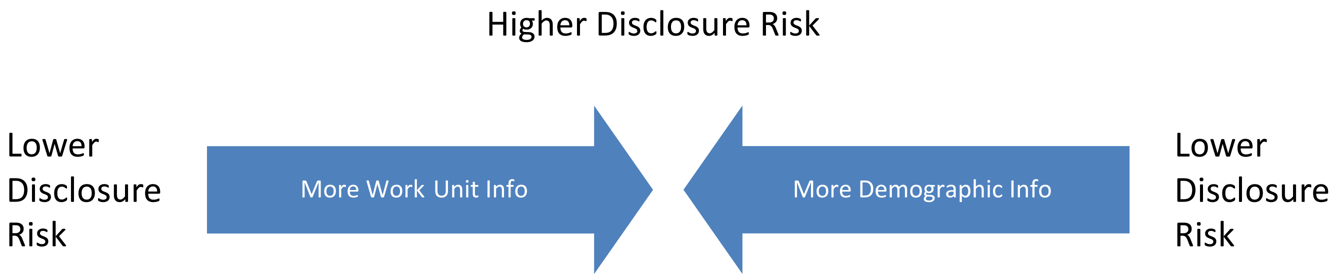
FEVS Data Releases

- After extensive reporting phase, three public-release data files (PRDFs) are made available (see https://www.fedview.opm.gov/2015/EVSDATA/):
  1. General (excluding LGBT item)
  2. LGBT (including LGBT item, fewer variables, common variables recoded to deter merging with general file)
  3. Trend (all prior general PRDFs stacked and coded forward to most recent FEVS)

- Privacy Act statement assures respondents “In any public release of survey results, no data will be disclosed that could be used to identify specific individuals”

Striking a Compromise

- In FEVS, work-unit information and observable demographics compete against each other with respect to disclosure risk



- For lower disclosure risk, one could release complete work-unit detail but no demographics, or vice versa → neither is ideal

- More appropriate approach is to strike a compromise, with the end goal to minimize disclosure risk while maximizing data utility

Current Disclosure Avoidance Methods

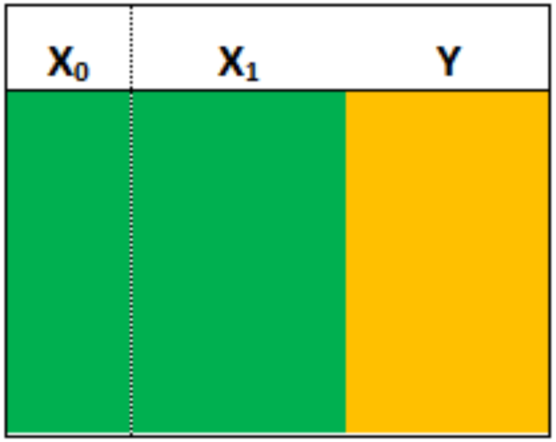
- Detailed in technical report (OPM, 2015):
  - Separate LGBT file
  - Starting point for work units: agency request, so long as at least 250 respondents
  - Certain variables removed and/or combined (e.g., minority status)
  - Categories collapsed for other variables

- Exhaustive tabulations assessment (ETA) (Krenzke et al., 2014) systematically examines all possible demographic combinations within a work unit, flagging records posing a disclosure risk

- Work unit identifiers with > 25% records flagged are set to missing, and ETA is done once more; for records still flagged (~8000 in FEVS 2016), only one of four "core" demographics (gender, age group, supervisory status, and minority status) is maintained

A Partially Synthetic Approach

Schematic Representation of Original Data Set:



Component	Contents
X_0	Unique Respondent ID Work Unit ID *(2 variables)*
X_1	Core Survey Items Analysis Weight *(84 variables)*
Y	HQ/Field Indicator Age Group Gender Education Level Intent to Leave Retirement Horizon Prior Military Status Race/Ethnicity Supervisory Status Agency/Federal Tenure Sexual Orientation Disability Status Telework Frequency *(15 variables)*

Premise: leave $\mathbf{X_0}$ and $\mathbf{X_1}$ intact, but model relationship between $\mathbf{X_1}$ and $\mathbf{Y}$ (independently within work units), and use to derive substitute values for $\mathbf{Y}$

A Partially Synthetic Approach (2)

- Fifteen variables comprising **Y** synthesized sequentially a la Raghunathan et al. (2001) using “synthpop” R package (Nowok et al., 2015)

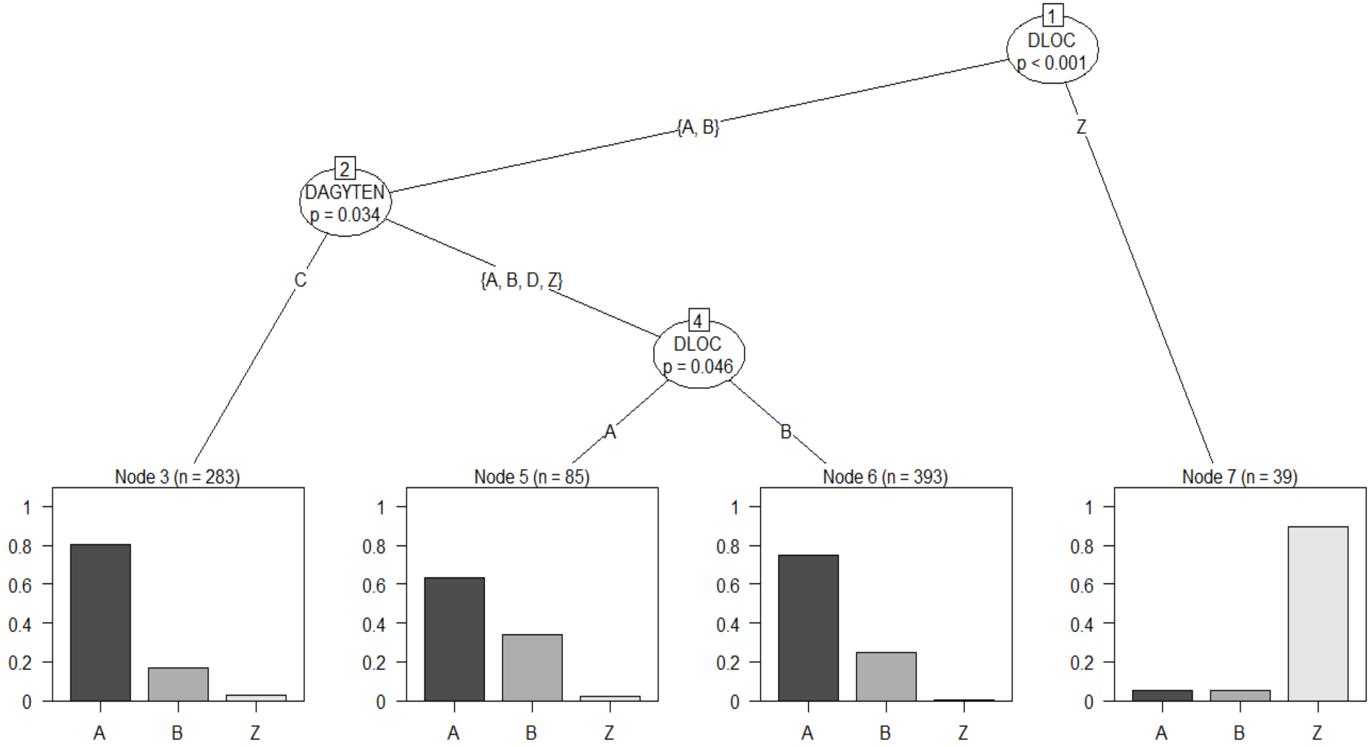
- Nonparametric “ctree” method used exclusively, based on classification and regression trees (CART) (Breiman et al., 1984) – successive binary splits partition data set into cells

- Within a cell, values are synthesized randomly in proportion to their occurrence in the observed data

- Created *M* = 3 implicates, not for variance estimation purposes per se, but to rule out deterministic relationships

Visualization of CTREE Method

- Key advantage of CART (Reiter, 2005): find and exploit only most important relationships from a large pool of potential predictors → in example below, only the HQ/field duty station indicator (DLOC) and agency tenure (DAGYTEN) are needed for synthesizing gender



Benefits Relative to Current Methods

- Dramatically reduced disclosure risk

- More works can be identified
  - 368 vs 181 distinct work units

- More demographic information can be included
  - 15 vs 11 variables
  - 48 vs 31 total demographic variable categories

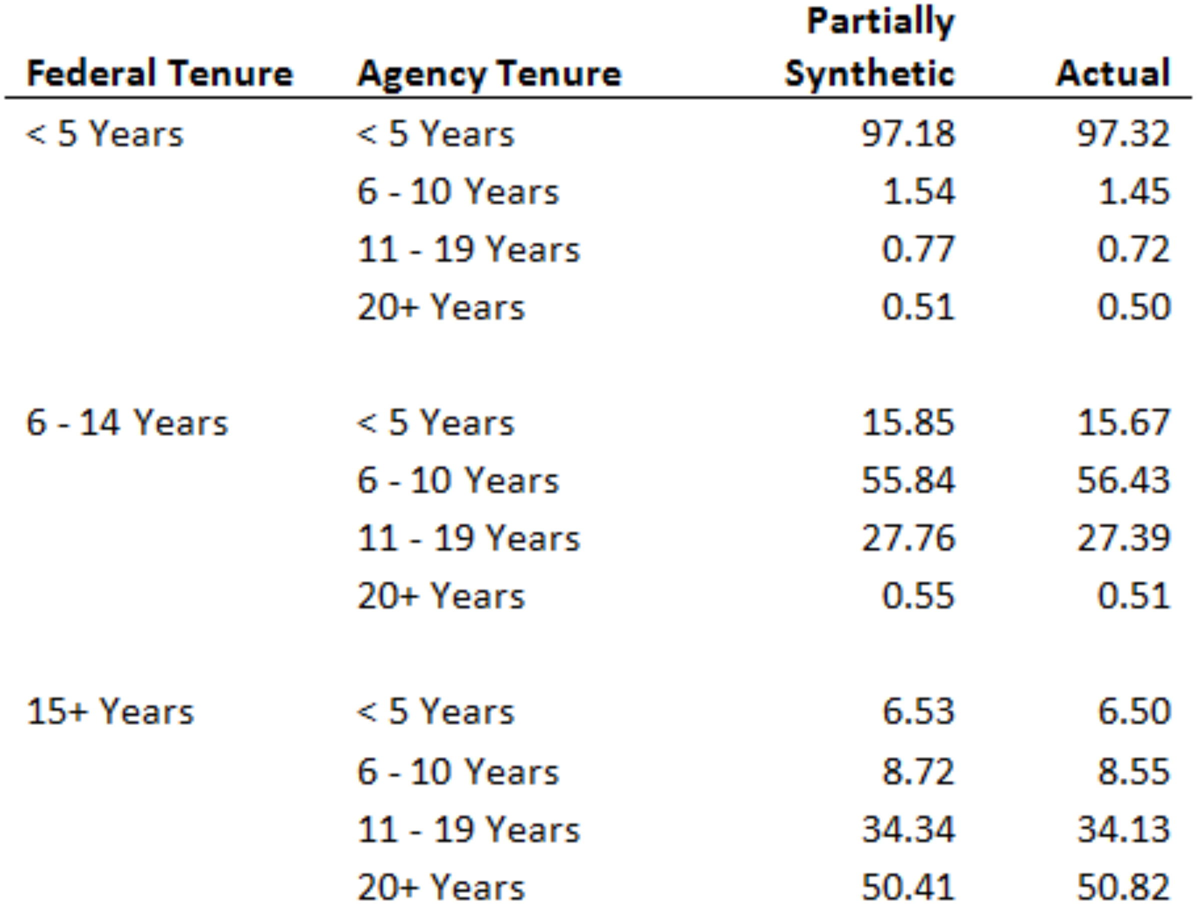
- No need for separate LGBT file

- Key downside: no guarantee synthetic data results match those that would be generated with the actual data

Results: Univariate Marginal Distributions

	Partially Synthetic		Actual	
Variable	Count	Percent	Count	Percent
Telework Status				
Telework	207,930	53.29	207,774	53.28
No Telework - Barrier	137,280	35.19	137,102	35.15
No Telework - Choice	44,951	11.52	45,123	11.57
Missing	17,628		17,790	
Headquarters vs. Field Duty Station				
Headquarters	156,170	40.37	156,217	40.40
Field	230,645	59.63	230,420	59.60
Missing	20,974		21,152	
Supervisory Status				
Non-Supervisor/Team Leader	303,972	78.10	303,683	78.06
Supervisor/Manager/Senior Leader	85,251	21.90	85,358	21.94
Missing	18,566		18,748	

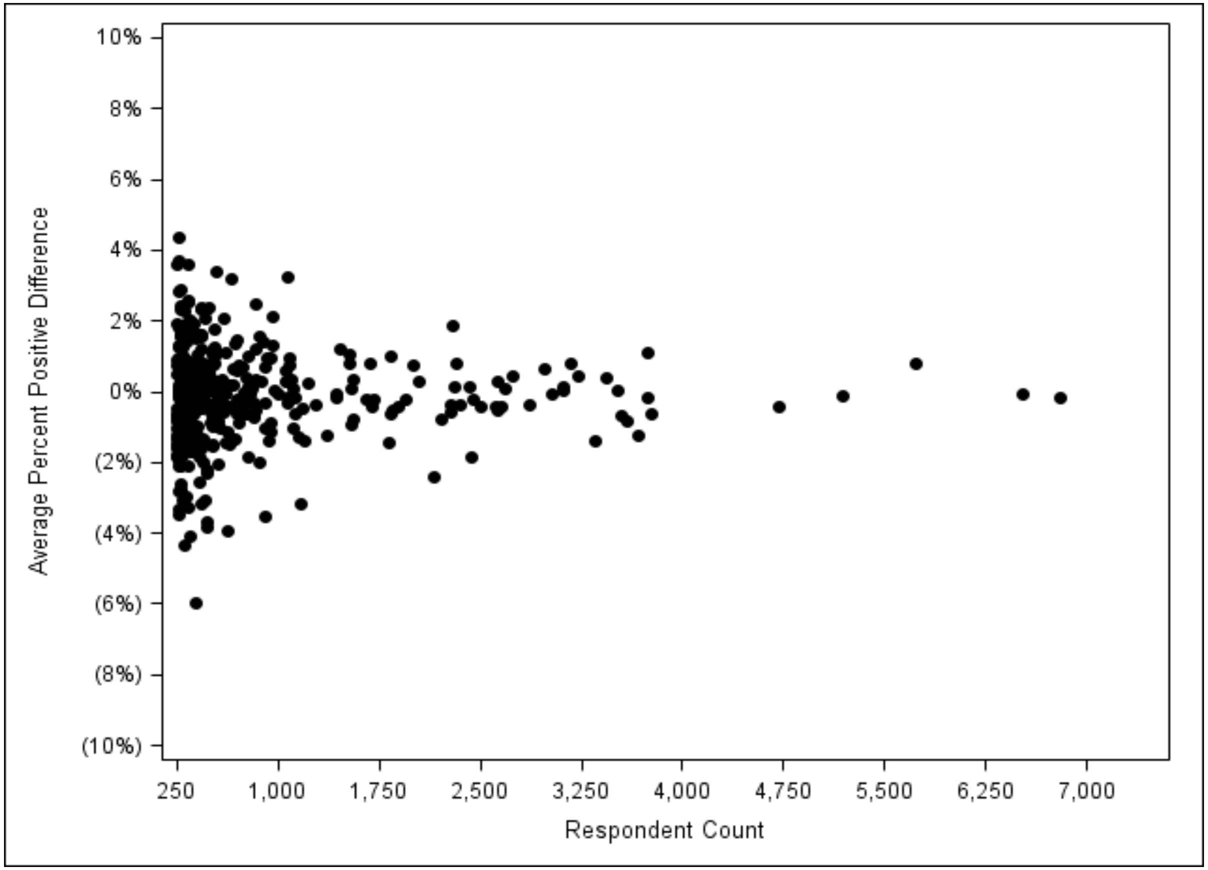
Results: Bivariate Marginal Distributions



Federal Tenure	Agency Tenure	Partially Synthetic	Actual
< 5 Years	< 5 Years	97.18	97.32
	6 - 10 Years	1.54	1.45
	11 - 19 Years	0.77	0.72
	20+ Years	0.51	0.50
6 - 14 Years	< 5 Years	15.85	15.67
	6 - 10 Years	55.84	56.43
	11 - 19 Years	27.76	27.39
	20+ Years	0.55	0.51
15+ Years	< 5 Years	6.53	6.50
	6 - 10 Years	8.72	8.55
	11 - 19 Years	34.34	34.13
	20+ Years	50.41	50.82

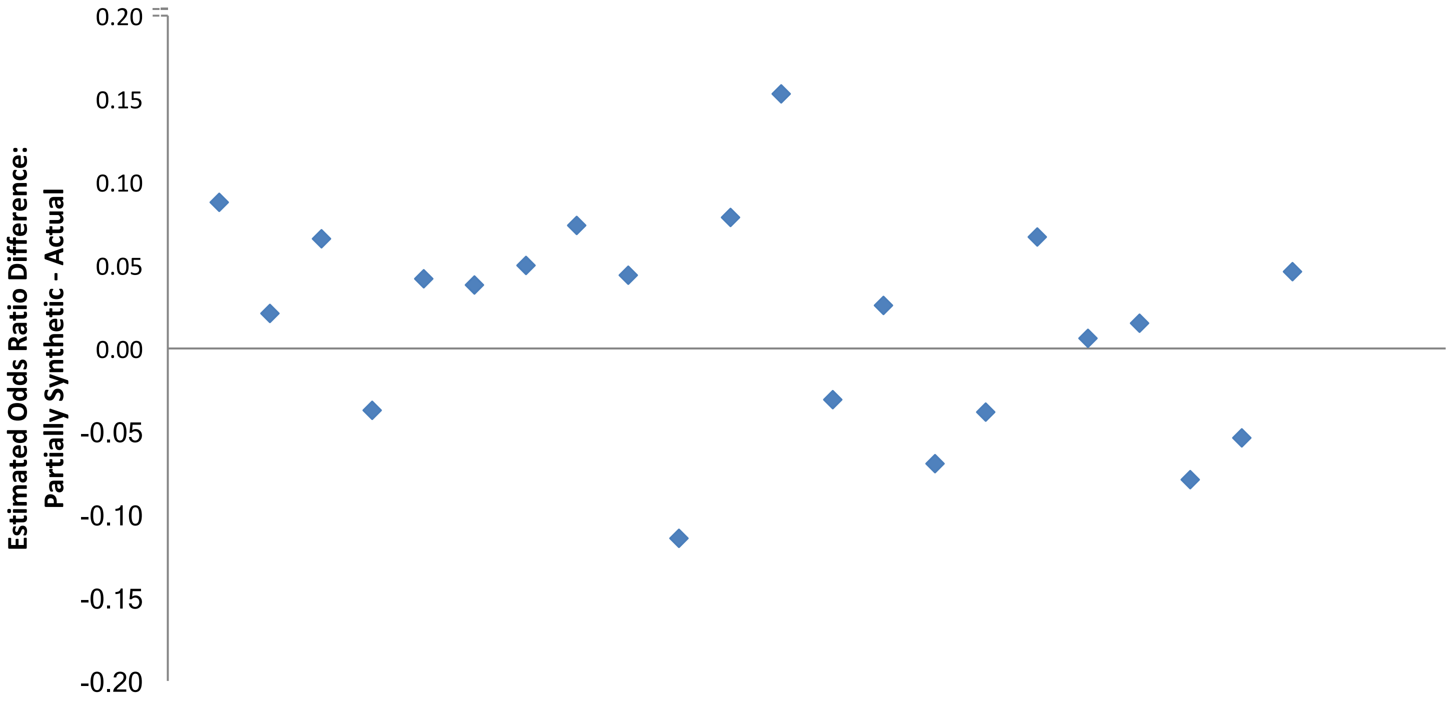
Results: Point Estimate Differences

Average Percent Positive Difference for 2016 FEVS Demographic Categories within a Work Unit: Partially Synthetic Data vs Actual Data



Results: Model Parameter Differences

Estimated Odds Ratio Differences for the Multinomial Logistic Regression Model Discussed in Whitford and Lee (2015):



Risk Assessments

- Traditional public-release data file (PRDF)
- Partially synthetic PRDF

- Re-identification -- Hundepool et al. (2012)
  - Achieved by an intruder when comparing a target individual in a sample with an available list of units (external file) that contains individual identifiers (e.g., name and address), plus a set of identifying variables
  - Occurs when the unit in the released file and a unit in the external file belong to the same individual in the population

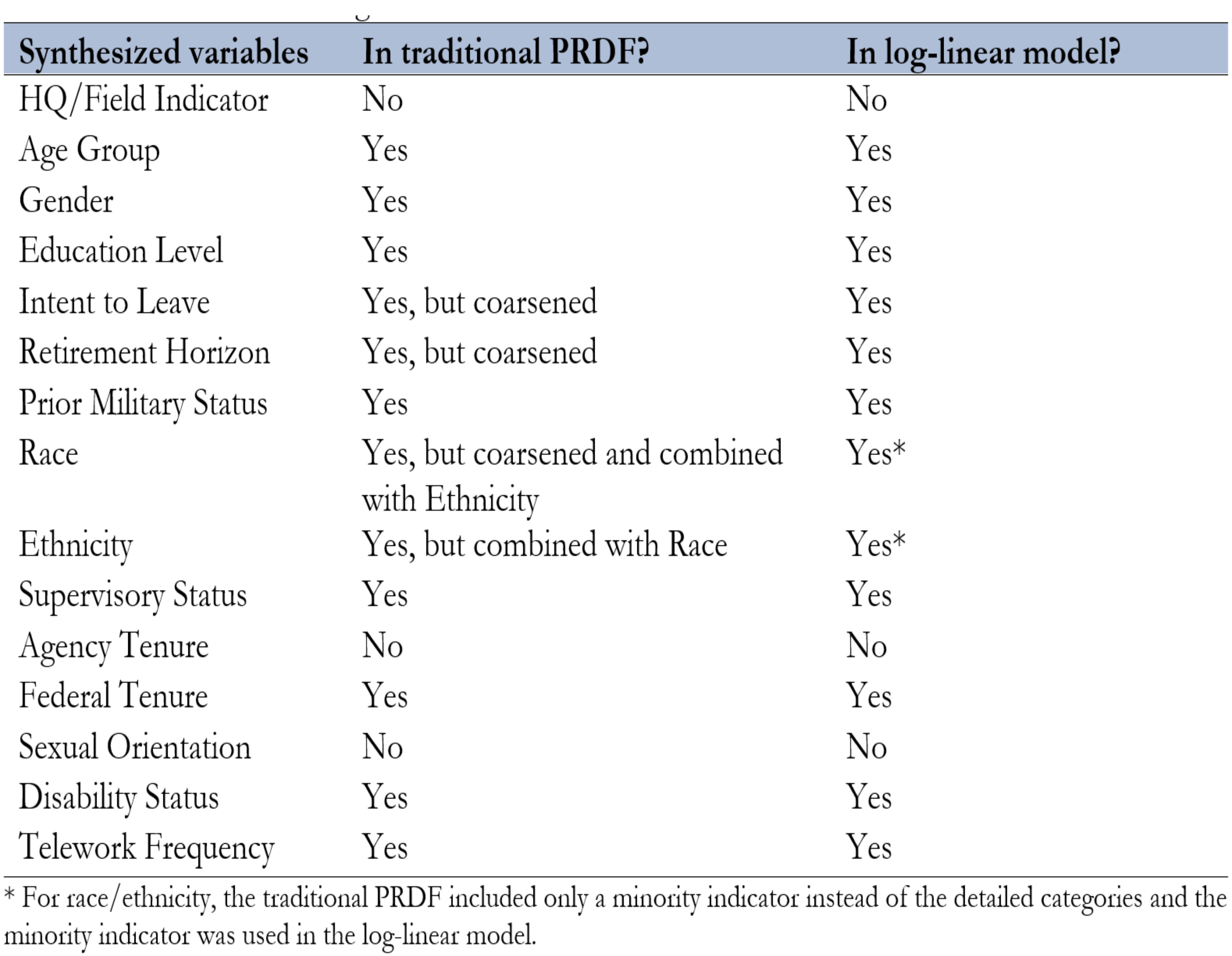
Risk Elements

- Questionnaire items
  - 15 indirect identifiers
  - Mostly attitudinal
- Work unit
- High sampling rate
  - Sampling rate equal to 1
- About 50 percent response rate

Risk Assessment on Traditional PRDF

- File risk measure
  - The expected number of population uniques given the sample uniques
    - $Risk = \sum_{SU} E(F_k = 1 | f_k = 1)$
      - where *SU* is the set of sample uniques, f_k is the sample frequency in cell *k*, and F_k is the population frequency in cell *k*
      - F_k must be estimated
        - » Estimated using loglinear models with the Skinner and Shlomo (2008) approach
          - Stabilizes estimate
          - Uses weights
- Work unit and 12 select indirect identifiers
  - Low number of missing values
  - Highly identifiable

Risk Assessment on Traditional PRDF (2)

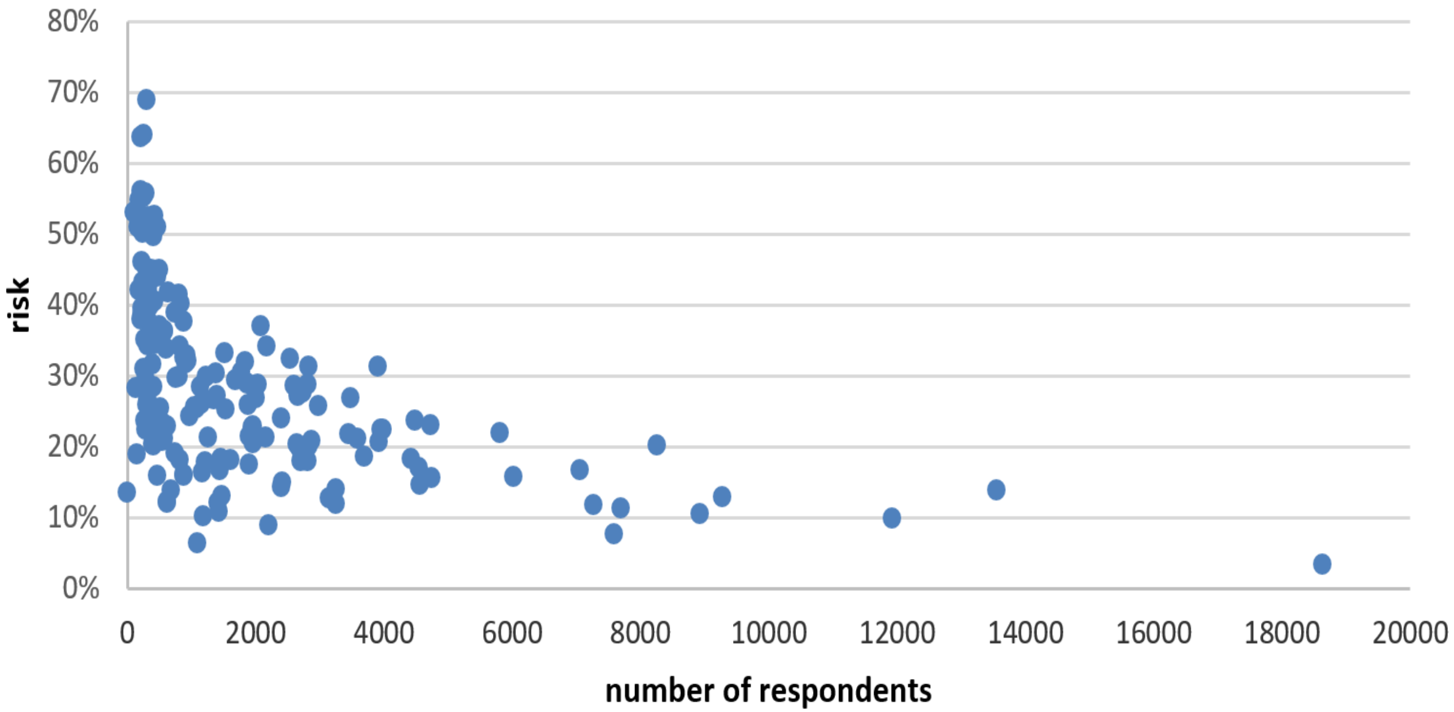


Synthesized variables	In traditional PRDF?	In log-linear model?
HQ/Field Indicator	No	No
Age Group	Yes	Yes
Gender	Yes	Yes
Education Level	Yes	Yes
Intent to Leave	Yes, but coarsened	Yes
Retirement Horizon	Yes, but coarsened	Yes
Prior Military Status	Yes	Yes
Race	Yes, but coarsened and combined with Ethnicity	Yes*
Ethnicity	Yes, but combined with Race	Yes*
Supervisory Status	Yes	Yes
Agency Tenure	No	No
Federal Tenure	Yes	Yes
Sexual Orientation	No	No
Disability Status	Yes	Yes
Telework Frequency	Yes	Yes

* For race/ethnicity, the traditional PRDF included only a minority indicator instead of the detailed categories and the minority indicator was used in the log-linear model.

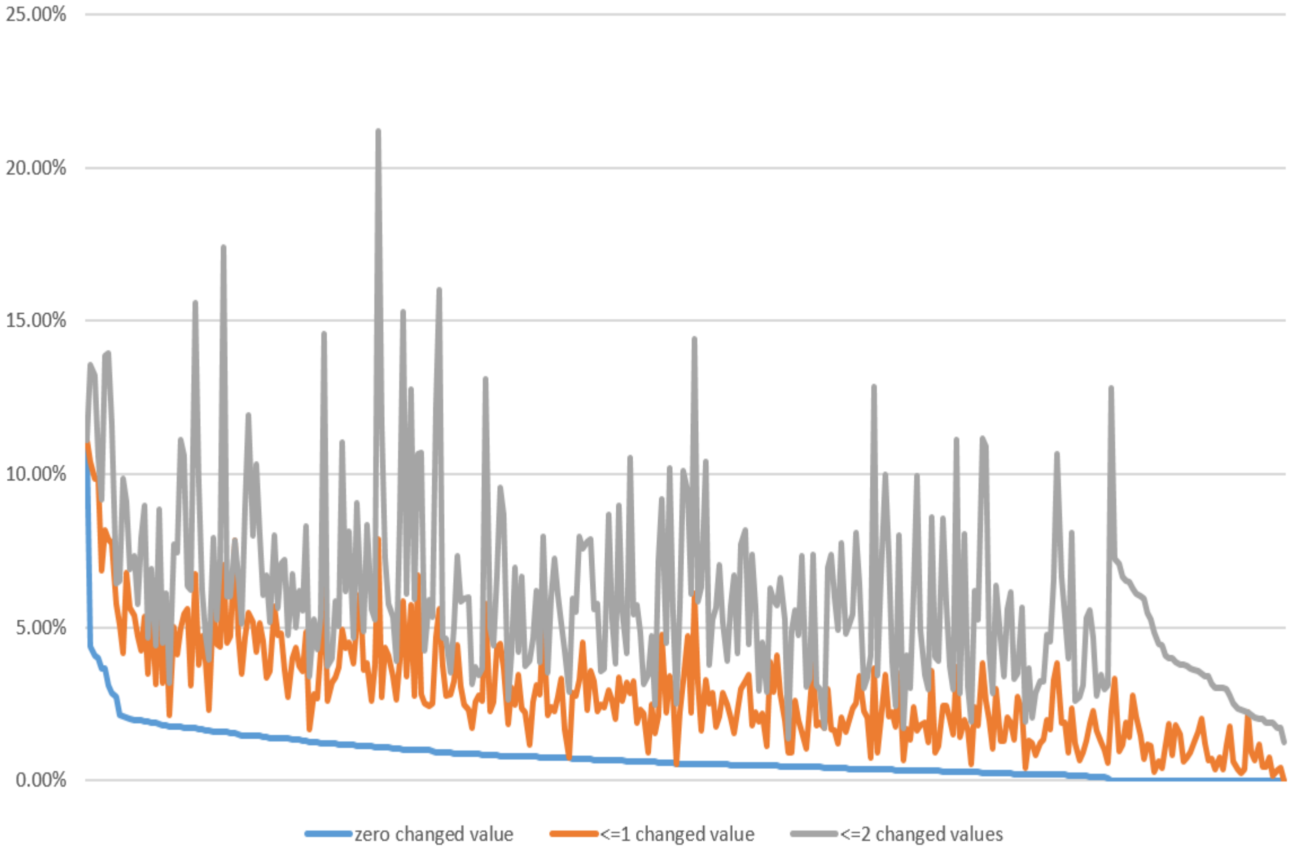
Risk Assessment on Traditional PRDF: Results

- File risk measure was computed for each work unit
  - Ranged from 3% to 69% across all work units (or combined work units) with a median of 26% and a mean of 28%
  - High risk



Risk Assessment on Partially Synthetic PRDF

- First look: Percentage of changed values among the 15 indirect identifiers, by work unit



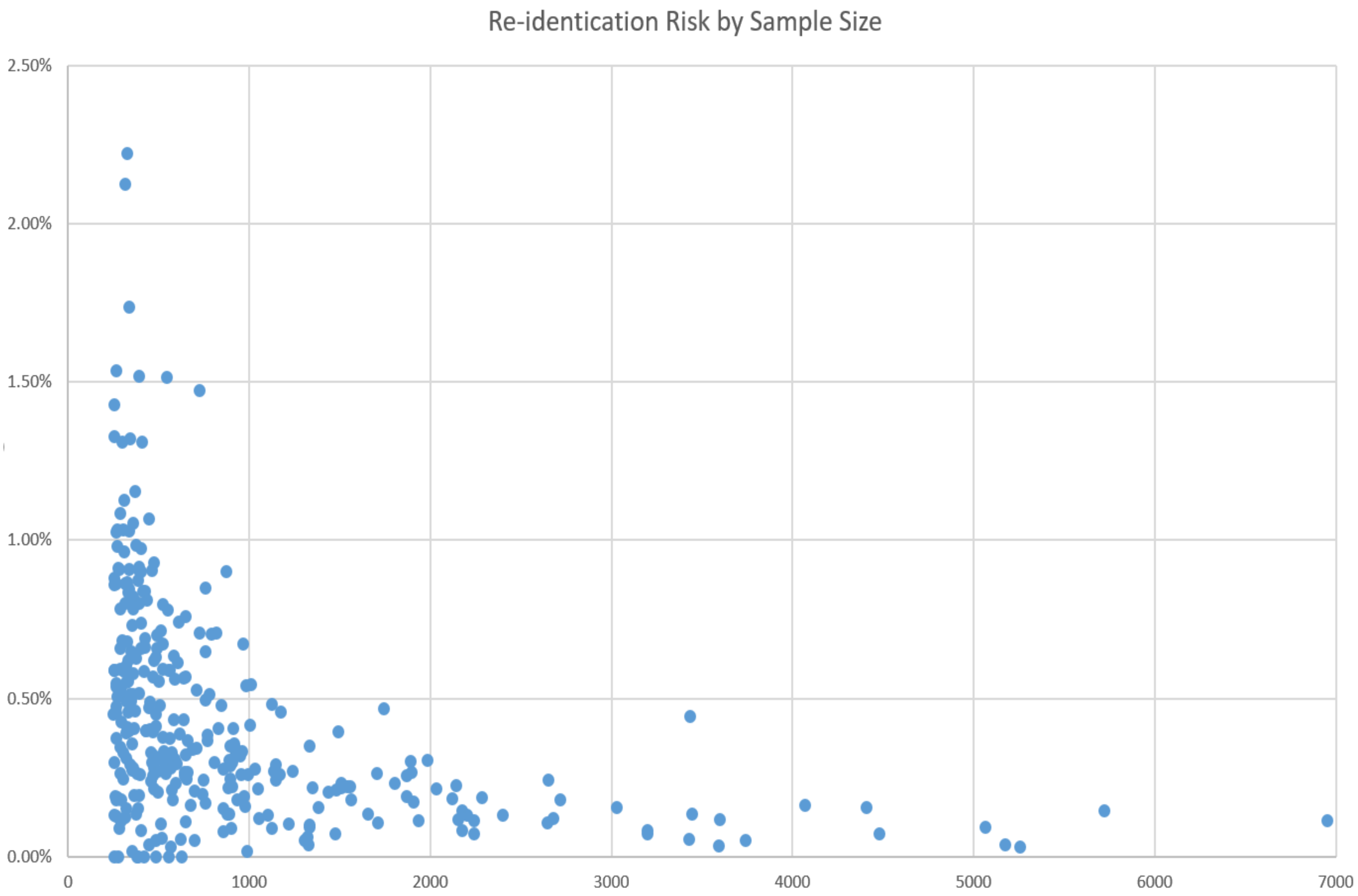
Risk Assessment on Partially Synthetic PRDF (2)

- Re-identification risk
  - What is the expected number of correct matches?
    - Raw-to-Raw
    - Synthetic-to-Raw
- Raw-to-Raw -- Exact matching
  - For each work unit, the match was conducted on 15 indirect identifiers
    - Some multiple records with the same subgroup
    - Example: If 3 records have the same characteristics, then risk is a 1 in 3 chance of matching correctly
  - On average, 88 percent matched correctly, ranging from 57 percent to 98 percent across work units
  - Did not account for approximate 50 percent response rate, which lowers the risk value

Risk Assessment on Partially Synthetic PRDF (3)

- Synthetic-to-Raw -- Probability-based matching
  - Used Westat's *WesLink* SAS macro, based on log-likelihood estimation
  - Identify group of best matches for each record, given the work unit and the 15 indirect identifiers
    - Threshold is set to minimize false positives and false negatives
  - Probability of correct match computed for each individual record
    - If the true record was among the best matches
      - Probability of correct match = 1 / (# of best matches)
    - If the true record was not among the best matches
      - Probability of correct match = 0
  - File risk was computed the average of the probabilities
    - 0.43 percent, not accounting for the response rate
    - 20 units ranged from 1.0 percent to 2.2 percent

Risk Assessment on Partially Synthetic PRDF (4)



Summary

- Partially synthetic 2016 FEVS data does not produce perfect replications of the actual data, but results are reasonably close and devoid of any systematic biases
- Differences tend to zero as sample sizes increase, as do measures of disclosure risk
- Of course, no guarantee all conceivable analyses will be as harmonious as those presented here
- Open question: is the extra noise a fair price to pay in exchange for more detailed demographic and work unit information, and dramatically reduced disclosure risk?

Further Research Questions

- Could data utility be increased if fewer values were synthesized?
  - Do not synthesize variables that are not highly identifiable (e.g., intention to leave)
  - Synthesize only a subset of variables for a subset of records with high disclosure risk
- Could the analysis weights be recalibrated in some way to make results more concomitant?
- Are there other solutions?
  - Remote access servers
  - Hybrid approach of both the traditional statistical disclosure limitation techniques (coarsening and suppression) and synthetic data

References

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

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., Li, J., and Li, L. (2014). “An Evaluation of the Impact of Missing Data on Disclosure Risk Measures,” Proceedings of the Joint Statistical Meetings, Alexandria, VA: American Statistical Association.

Nowok, B., Raab, G. and Dibben, C. (2015). "synthpop: Bespoke Creation of Synthetic Data in R," Package vignette available online at: http://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models,” *Survey Methodology*, **27**, pp. 85 – 95.

Raghunathan, T., Reiter, J., and Rubin, D. (2003). “Multiple Imputation for Statistical Disclosure Limitation,” *Journal of Official Statistics*, **19**, pp. 1 – 16.

References (2)

Reiter, J. (2003). “Inference for Partially Synthetic, Public Use Microdata Sets,” *Survey Methodology*, **29**, pp. 181 – 188.

Reiter, J. (2005). “Using CART to Generate Partially Synthetic Public Use Microdata,” *Journal of Official Statistics*, **21**, pp. 441 – 462.

Reiter, J. (2012). “Statistical Approaches to Protecting Confidentiality for Microdata and their Effects on the Quality of Statistical Inferences.” *Public Opinion Quarterly*, **76**, pp. 163–181

Rubin, D. (1993). “Discussion: Statistical Disclosure Limitation,” *Journal of Official Statistics*, **9**, pp. 461 – 468.

Skinner, C., and Shlomo, N. (2008). “Assessing Identification Risk in Survey Microdata Using Log-Linear Models,” *Journal of the American Statistical Association*, **103**, pp. 989 – 1001.

U.S. Office of Personnel Management. (2015). *Federal Employee Viewpoint Survey Technical Report*. Available online at: https://www.fedview.opm.gov/2015/Published/.

Whitford, A., and Lee, S.-Y. (2015). “Exit, Voice, and Loyalty with Multiple Exit Options: Evidence from the US Federal Workforce,” *Journal of Public Administrations Research and Theory*, **25**, pp. 373 – 398.