STUDY SERIES
(Statistics #2016-01)

NSF GRIP Report: American Community Survey Simulation Study

Claire McKay Bowen [1]

[1] Statistics Ph.D. Candidate, University of Notre Dame

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: October 25, 2016

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Contents

1 American Community Survey Data Summary **4**
1.1 Summary of Raw Data 4
1.1.1 Reduction of Data 4
1.2 Pre-processing the Data Details 5
1.2.1 Changing Continuous Variables to Categorical Variables 5
1.2.2 Reducing the Classes for Categorical Variables 5
1.2.3 Further Pre-Processing of the Data 6

2 Simulation Study **8**
2.1 Differential Privacy 8
2.2 Generation of Synthetic Data 9
2.2.1 Notation 9
2.2.2 Assumptions 9
2.2.3 Model-Based Differentially Private Data Synthesis 10
2.2.4 Proposed Model 10
2.2.5 Range of ϵ 11
2.2.6 Overall Simulation Steps 11

3 Simulation Results **12**
3.1 Combination Rules for Multiple Synthesis 12
3.2 Plots 13
3.3 Bias 13
3.4 Root Mean Squared Error 13
3.5 Coverage Probability 14
3.6 Empty Cells 14
3.7 Conclusion 14

A NSF GRIP Overview **19**
A.1 NSF GRIP Proposal 19
A.1.1 Overview 19
A.1.2 Research Proposal 21

B NSF GRIP Professional Development and Research **22**
B.1 Research Products 23
B.1.1 Dissertation 23

B.1.2 Research Presentations 23
B.1.3 Collaborative Research 23
B.2 Professional Development 23
B.2.1 Expanded Professional Networks 23
B.2.2 Presentation 24
B.2.3 Communication Skills 24

Purpose

Report Abstract

This report outlines the fellow's professional development and research at the US Census Bureau through the National Science Foundation Graduate Research Internship Program (NSF GRIP). The fellow investigated traditional multiple synthesis and a model-based differentially private data synthesis (MODIPS) methods to assess the feasibility and practicality of applying differential privacy on real-world data sets. In the Appendix, the fellow outlines the professional development in both collaborative and communication skills by expanding professional networks within and outside of the US Census Bureau, presenting research, and meeting with scientists.

US Census Bureau Mission Statement

The US Census Bureau mission "is to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly."

NSF GRIP Mission Statement

NSF GRIP "...provides professional development NSF GRFP (Graduate Research Fellowship Program) Fellows through internships developed in partnership with federal agencies. Through GRIP, Fellows participate in mission-related, collaborative research under the guidance of host research mentors at federal facilities and national laboratories. GRIP enhances the Fellows' professional skills, professional networks, and preparation for a wide array of career options. The sponsor agencies benefit by engaging Fellows in applied projects, helping to develop a highly skilled U.S. workforce in areas of national need."

Chapter 1

American Community Survey Data Summary

The raw data set is an American Community Survey (ACS) 2014 micro level data.

“ACS is an ongoing survey that provides vital information on a yearly basis about our nation and its people. Information from the survey generates data that help determine how more than $400 billion in federal and state funds are distributed each year.

Through the ACS, we know more about jobs and occupations, educational attainment, veterans, whether people own or rent their home, and other topics. Public officials, planners, and entrepreneurs use this information to assess the past and plan the future. When you respond to the ACS, you are doing your part to help your community plan hospitals and schools, support school lunch programs, improve emergency services, build bridges, and inform businesses looking to add jobs and expand to new markets, and more.”

1.1 Summary of Raw Data

The analysis was applied to a subset of the data, and this section details how the ACS data was reduced for analysis.

1.1.1 Reduction of Data

Real-life data poses additional challenges than simulated data such as missing or improper values (e.g. minors working several hours). Additionally, most differentially private data synthesis methods were applied to data sets with large number of observations and not parameters. The focus for the simulation study is to implement MODIPS to a data set with a large number of observations and a reasonable number of parameters (e.g. cross-tabulation less than a million). The data was reduced as follows (refer to Table 3.1 for variable notation):

1. Selected 19 variables (mixture of categorical and continuous)

2. Found completers from the 19 variables (2,927,202)
3. Removed observations, where the incomes did not equal the income total (1,883,531 observations)
4. Removed observations, who have non-positive income in SEM (1,883,333)
5. Removed observations, who are 18 years or older (1,857,505)

The subset chosen for the analysis has the following features (further details of the pre-processing is in Section 1.2:

- 1,857,505 observations
- 19 Variables
  - 13 Categorical
  - 6 Continuous
  - one of the categorical variables was originally continuous (CAGE)
- Continuous Variables: the 5 other continuous variables are income values such as Retirement Income (RET) or Self-Employed Income (SEM) and sum to Adjusted Total Income (ATI).
- Reduction of classes for categorical variables
  - Convert HIS to White, Black, Native America, Asian/Pacific Islander, Hispanic, Other (i.e. 79 to 6)
  - Convert OCC to general Occupation Codes (i.e. 540 to 14)
  - Convert POB to general Place of Birth Codes (i.e. 301 to 21)
  - Convert SCHL to smaller groups of education (i.e. 24 levels to 6 levels)

1.2 Pre-processing the Data Details

1.2.1 Changing Continuous Variables to Categorical Variables

The Age Variable (CAGE) was converted to a categorical variable by increments of 5 starting from 18 - 24, 25 - 29, 30 - 34, and so on till 65 +.

1.2.2 Reducing the Classes for Categorical Variables

Hispanic Origin

Convert to a general race variable

- 1 = White
- 2 = Black
- 3 = Native American
- 4 = Asian or Pacific Islander
- 5 = Hispanic
- 6 = Other

Occupation Code

Convert to a general occupation type.

Place of Birth

Convert to a general regions of the world.

- 1 = United States
- 2 = US Island and Puerto Rico
- 3 = Northern Europe
- 4 = Western Europe
- 5 = Southern Europe
- 6 = Eastern Europe
- 7 = Eastern Asia
- 8 = South Central Asia
- 9 = South Eastern Asia
- 10 = Western Asia
- 11 = Northern America (not US)
- 12 = Latin America
- 13 = Caribbean
- 14 = South America
- 15 = Eastern Africa
- 16 = Middle Africa
- 17 = Northern Africa
- 18 = Southern Africa
- 19 = Western Africa
- 20 = Oceania
- 21 = At Sea/Abroad, Not Reported

Education Attainment

Convert to a general education level

- 1 = Less than high school
- 2 = HS or GED equivalent
- 3 = Some College
- 4 = Associate
- 5 = Bachelors
- 6 = Advance Degree

1.2.3 Further Pre-Processing of the Data

Converted the data to be completely discrete. All the income variables except AWAG and ATI were converted to 0 or 1 (0 = none, 1 = some income). ATI was removed entirely, reducing the data set to 18 variables. In addition, AWAG was converted to the following levels:

- 1 = $\$0$
- 2 = $\$0 < \text{AWAG} \leq \$60{,}000$

- 3 = \$60,000 $<$ AWAG $\leq$ \$250,000
- 4 = \$250,000 $<$ AWAG

Chapter 2

Simulation Study

The ultimate goal is to release data publicly without revealing personal information about the participants in the data. Current data privacy methods for releasing data sets often do not either quantify how much privacy is being "leaked" or make several assumptions on the behavior and/or knowledge of a data intruder, who seeks to expose personal information about the data set.

In the last decade, the computer science community has created a concept called differential privacy that provides strong privacy guarantee in mathematical terms that quantifies the privacy risk without making assumptions about the background knowledge of data intruders. Although computer scientists and statisticians have conducted extensive research on applying and improving differential privacy methods, very little research is on applying differential privacy methods on real-world data sets.

This simulation study is to help assess the utility and inferential properties of the sanitized data from a traditional multiple synthesis (a data privacy method that does not quantify privacy risk) and differential privacy approach called model-based differentially private data synthesis (MODIPS) on a real-world data set. As outlined from the previous chapter, the data set is a K-cross tabulated data set from ACS.

2.1 Differential Privacy

Conceptually, differential privacy is a condition on data privacy algorithms that examines all possible versions of a person in a data set when a query or question is submitted to the algorithm. The query result should not depend on whether or not a person is part of the data set. Simply, no matter the person or question, differential privacy ensures that a person's information does not contribute significantly to the final query result based on a quantified bound, $\epsilon \in (0, \infty)$. When ϵ is large, more information will be "leaked". When ϵ is small, less information is leaked. Ideally, ϵ should be as close to 0 as possible.

To guarantee very little information is leaked, the inference on the query results will be poor, so some information must be leaked. Since ϵ is used in terms of privacy leakage that could be infinite, the reader should think of ϵ as a score from 1 (or very close to 0) to 10. e.g. if a differentially private algorithm requires $\epsilon \geq 8$, a lot of information is being leaked.

For the mathematical details of differential privacy, the fellow suggests the reader to read the following:

- Dwork et al. (2006) for the original proposal of differential privacy
- Abowd and Vilhuber (2008) for a simple application of differential privacy
- Wasserman and Zhou (2010) for a rigorous statistical overview and application of differential privacy
- Bowen and Liu (2016) for applications of several differential private data synthesis techniques

2.2 Generation of Synthetic Data

This section outlines the mathematical details of how the synthetic data is generated.

2.2.1 Notation

- N is the number of observations in the data ($N =$ 1,857,505)
- K is the number of classes total from the cross tabulation of the data ($K =$ 300,625)
- $\boldsymbol{n}$ be a $K \times 1$ vector of the number of observations in each cell of the cross-tabulation
- $\boldsymbol{\pi}$ be a $K \times 1$ vector of proportion of observations in each cell of the cross-tabulation
- p is the number of attributes in the data ($p = 18$)
- s is the sufficient statistic from the model
- m is the number of multiple imputations
- $\mathbf{x}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{x}}^*$ be the raw data, synthetic data, and the differentially private synthetic data, respectively

2.2.2 Assumptions

1. **The raw data set's proportions, $\boldsymbol{\pi}$ used as the true parameters:** The statistics from the raw data set will be used as the true proportions for the simulation.
2. $\boldsymbol{N^* = N}$: Our goal is to replicate the original data set with a privacy guarantee, so the released synthetic data set will be the same sample size as the original.

3. **Zero count cells will remain zero count:** We will treat all empty cells as population zeros. We do this, because 1) we have no further information other than the original sample to contribute to the DIPS process and 2) the model we are using contains components on the interactions between the continuous and categorical variables, so empty cells will not contribute information as well. i.e. If $\pi_k = 0$, then $x_k = 0$ for $k = 1, ..., K$.
4. **The raw data is complete and accurate:** Filling in the missing values in the original data set will create inaccuracies. However, since the primary goal of this work is to access the feasibility of implementing MODIPS to a real data set, we will assume that the complete data is the true data.
5. **IID Data Sampling:** We will assume that the data is sampled from an infinite population since our focus is on how we can protect the participants in the data with differential privacy.

2.2.3 Model-Based Differentially Private Data Synthesis

The MODIPS approach is based in a Bayesian modeling framework and releases m multiple sets of surrogate copies of the original data $\mathbf{x}$ to account for the uncertainty of the synthesis model. An illustration of the MODIPS algorithm is given in Figure 2.1. In summary, the MODIPS approach first constructs an appropriate Bayesian model from the original data and identified the Bayesian sufficient statistics $\mathbf{s}$ associated with the model. The posterior distribution of $\boldsymbol{\theta}$ can then be represented as $f(\boldsymbol{\theta}|\mathbf{s})$. The MODIPS approach then sanitizes $\mathbf{s}$ with privacy budget ϵ/m, where m is the number of surrogate data sets to release. Denote the sanitized $\mathbf{s}$ by $\mathbf{s}^*$. Synthetic data $\tilde{\mathbf{x}}$ is simulated given $\mathbf{s}^*$ by first drawing $\boldsymbol{\theta}^*$ from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{s}^*)$, and then simulating $\tilde{\mathbf{x}}^*$ from $f(\mathbf{x}|\boldsymbol{\theta}^*)$. The procedure is repeated m times to generate m surrogate data sets.

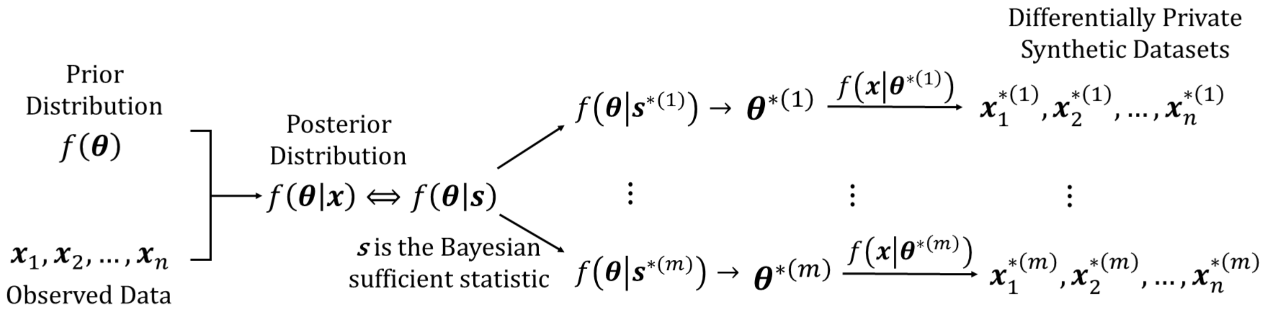
Differentially Private Synthetic Datasets
Prior Distribution
$f(\boldsymbol{\theta})$
$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$
Observed Data
Posterior Distribution
$f(\boldsymbol{\theta}|\boldsymbol{x}) \Leftrightarrow f(\boldsymbol{\theta}|\boldsymbol{s})$
$\boldsymbol{s}$ is the Bayesian sufficient statistic
$f(\boldsymbol{\theta}|\boldsymbol{s}^{*(1)}) \rightarrow \boldsymbol{\theta}^{*(1)} \xrightarrow{f(\boldsymbol{x}|\boldsymbol{\theta}^{*(1)})} \boldsymbol{x}_1^{*(1)}, \boldsymbol{x}_2^{*(1)}, \dots, \boldsymbol{x}_n^{*(1)}$
$\vdots$
$f(\boldsymbol{\theta}|\boldsymbol{s}^{*(m)}) \rightarrow \boldsymbol{\theta}^{*(m)} \xrightarrow{f(\boldsymbol{x}|\boldsymbol{\theta}^{*(m)})} \boldsymbol{x}_1^{*(m)}, \boldsymbol{x}_2^{*(m)}, \dots, \boldsymbol{x}_n^{*(m)}$

Figure 2.1: The MODIPS algorithm.

2.2.4 Proposed Model

Since the data are converted to all categorical, Multinomial-Dirichelet model is used. $m = 5$ surrogate data sets were imputed.

For the traditional MS approach, a non-informative Dirichlet prior was applied on $\boldsymbol{\pi}$, $f(\boldsymbol{\pi}') = \mathrm{D}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\} = 1/2$ and its posterior distribution was $f(\boldsymbol{\pi}'|\boldsymbol{n}) = D(\boldsymbol{\alpha}')$, where $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{n}$ and $\boldsymbol{n} = \{n_1, ..., n_K\}$. Then, $\tilde{\mathbf{x}}$ was drawn from $f(\tilde{\mathbf{x}}|\boldsymbol{\pi}') = \text{Multinom}(N, \boldsymbol{\pi})$.

For the MODIPS approach, the Bayesian sufficient statistics based from the traditional MS model was $\mathbf{s} = \boldsymbol{n}$ and is sanitized using the Laplace mechanism, $\mathbf{s}^* = \boldsymbol{n}^*$, where the global sensitivity was 1 for $\boldsymbol{n}$; i.e.

$$e_k \overset{iid}{\sim} \text{Lap}(0, \epsilon^{-1}), n_k^* = n_k + e_k, \quad p_k = \frac{n_k^*}{\sum_k n_k^*}$$

where n_k^* is the sanitized cell count and p_k^* is the sanitized cell proportion for $k = 1, ..., K$. Note that n_k^* can be negative if the regular Laplace mechanism is used. A common post-hoc processing approach is replacing any negative n_k^* with 0. For the simulation study, any negative counts were replaced with 0 and then the number of observations was calculated based on the sanitized proportions to ensure $\sum_k n_k^* = N$.

Similar to the traditional MS approach, a non-informative Dirichlet prior was applied on $\boldsymbol{\pi}$, $f(\boldsymbol{\pi}) = \mathrm{D}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\} = 1/2$ and its posterior distribution was $f(\boldsymbol{\pi}^*|\boldsymbol{n}^*) = D(\boldsymbol{\alpha}'^*)$, where $\boldsymbol{\alpha}'^* = \boldsymbol{\alpha} + \boldsymbol{n}^*$ and $\boldsymbol{n}^* = \{n_1^*, ..., n_K^*\}$. Then, $\tilde{\mathbf{x}}^*$ was drawn from $f(\tilde{\mathbf{x}}^*|\boldsymbol{\pi}^*) = \text{Multinom}(N, \boldsymbol{\pi}^*)$.

2.2.5 Range of ϵ

We will test for the "ideal" level of ϵ (differential privacy) by exploring $\epsilon = e^{-6}, e^{-5}, ..., e^4$ on the cross tabulation of the categorical variables. Bias, Root Mean Squared Error, and Coverage Probability will be calculated for comparison.

2.2.6 Overall Simulation Steps

1. Find proportions of the cross tabulation of the raw data set, $\boldsymbol{\pi}$.
2. Use $\boldsymbol{\pi}$ to generate "original" (ORI) data set drawn from $\boldsymbol{x} \sim \text{Multinom}(N, \boldsymbol{\pi})$.
3. Apply Multinomial-Dirichelet Model for multiple synthesis (MS) to obtain $\tilde{\boldsymbol{x}}$ and model-based differentially private data synthesis (MODIPS) to obtain $\tilde{\boldsymbol{x}}^*$.
4. Gather statistical inferences on ORI, MS, and MODIPS.
5. Repeat Steps 2 - 4 500 times.

Chapter 3

Simulation Results

3.1 Combination Rules for Multiple Synthesis

In the case of releasing multiple sets of synthetic data valid approaches are needed to combine the inferences from the multiple sets to yield an overall estimate. Suppose the parameter of interest is β, where β is a scalar. Denote the estimate of β in the j^{th} synthetic data by $\hat{\beta}_j$ and the associated standard error by v_j. The final inferences on β are obtained via

$$\begin{aligned} \bar{\beta} &= m^{-1}\textstyle\sum_{j=1}^{m}\hat{\beta}_j & (3.1)\\ T &= m^{-1}B+W & (3.2)\\ (\bar{\beta}-\beta)T^{-1/2} &\sim t_{\nu=(m-1)(1+mW/B)^2}, & (3.3)\end{aligned}$$

where $B=\sum_{i=1}^{m}(\hat{\beta}_j-\bar{\beta})^2/(m-1)$ (between-set variability) and $W=m^{-1}\sum_{j=1}^{m}v_j^2$ (average per-set variability).

The variance combination rule given in Equation (3.2) was proposed first by Reiter (2003) for dealing with inferences in the context of partial synthesis without DP. Liu (2016) proved that the combination rule in Equation (3.2) also applies in the case of the MODIPS approach and the only difference is what the between-set variability B is comprised of. B in the MODIPS approach comprises of additional variability than traditional multiple synthesis – sanitizing s. Due to the extra sanitization step of s in the MODIPS approach as compared to the traditional MS approach, B in the former will be larger than in the latter, leading to less precise estimate on β; a price paid for DP guarantee. The variance combination rule given in Equations (3.1) to (3.3) also applies in the MD synthesizer and other DIPS approaches that rely on multiple set releases to account for synthesis model uncertainty, though the variability source contributing to the between-set B might differ. In the context of the simulation study, $\beta=p_k$, $\hat{\beta}_j=\hat{p}_{k,j}$, $\bar{\beta}=\bar{p}_k$, and $v_j^2=\hat{p}_{k,j}(1-\hat{p}_{k,j})N^{-1}$.

3.2 Plots

Figures 3.1 and 3.2 summarize the results from the simulation study.

3.3 Bias

$\bar{p}_{k,i}$, the estimated proportion of a cell from the ith iteration, was calculated from Equation 3.1, and averaged across the 500 repeats:

$$\bar{p}'_k = \sum_{i=1}^{500} \frac{\bar{p}_{k,i}}{500}, \text{ for } k = 1..., K$$

For Figure 3.1, the average($|bias|$) was calculated from

$$\bar{\boldsymbol{B}}' = \sum_{k=1}^{K} \frac{|B'_k|}{K} = \sum_{k=1}^{K} \frac{|\bar{p}'_k - \pi_k|}{K}$$

For Figure 3.2, the plot was based on the 10%, 25%, 50%, 75%, and 90% quantiles of $\bar{\boldsymbol{B}}'$.

3.4 Root Mean Squared Error

The estimated proportions of each cell, $\bar{p}'_k$, were converted to the number of observations, $\bar{n}'_k$. The variability of the estimated observations in the 500 repeats was calculated from

$$V'_k = \frac{\sum_{i=1}^{500} \left((\bar{n}'_k - \bar{n}_{k,i})^2 \right)}{500 - 1}, \text{ for } k = 1..., K$$

Then, the root mean squared error was calculated by

$$R'_k = \sqrt{V'_k + (B'_k)^2}$$

For Figure 3.1, the average($RMSE$) was calculated from

$$\bar{\boldsymbol{R}} = \sum_{k=1}^{K} \frac{R'_k}{K}$$

For Figure 3.2, the plot was based on the 10%, 25%, 50%, 75%, and 90% quantiles of $\bar{\boldsymbol{R}}'$.

3.5 Coverage Probability

For the original data, the 95% confidence interval was obtained by

$$L = \left(p_k - t \cdot \sqrt{u}, p_k + t \cdot \sqrt{u}\right), \quad C_k = I_L(\bar{\pi}_k)$$

where $u = p_k(1 - p_k)N^{-1}$.

The within-set variance of $\hat{p}_{k,j}$ was $v_j^2 = \hat{p}_{k,j}(1 - \hat{p}_{k,j})N^{-1}$ in each set for $j = 1, ..., 5$. Then, Equations (3.1) to (3.3) were applied to obtain the 95% confidence interval and calculate the CP of $\bar{p}_{k,j}$. i.e.

$$L = \left(\bar{p}_k - t_\nu \cdot \sqrt{T}, \bar{p}_k + t_\nu \cdot \sqrt{T}\right), \quad C'_k = I_L(\bar{\pi}_k)$$

For Figure 3.1, the average 95% CP was calculated from

$$\bar{C}'_k = \sum_{k=1}^{K} \frac{C'_k}{K}$$

For Figure 3.2, the plot was based on the 10%, 25%, 50%, 75%, and 90% quantiles of $\bar{\boldsymbol{C}}'$.

3.6 Empty Cells

From Figure 3.1, many of the cells were empty when generating the ORI data set (25.7% empty cells). The large number of empty cells was caused by the data set being sparse. However, both MS and MODIPS had fewer empty cells; $\approx$ 5.2% when ϵ was large. While on average each MS imputed data set had over 30% empty cells, the cells are averaged over the m imputed data sets, resulting in fewer empty cells. MODIPS had very few empty cells ($>$ 1%) when ϵ was small and increased to 5.1% as ϵ was larger. This result is caused by MODIPS evenly distributing the N observations in each cell to guarantee a high level of privacy for smaller values of ϵ.

3.7 Conclusion

The following conclusions were drawn from the simulation study:

- $\epsilon \approx 2.7$ to 7.4 provides low bias and RMSE for MODIPS with decent coverage probability. Around $\epsilon = 4$ should be used on the ACS data set to provide a privacy guarantee while still maintaining sufficient statistical inferences.
- As ϵ increases, MODIPS results approach MS results.
- ORI, MS, and MODIPS have high over coverage of the 95% confidence interval. This result might be caused by the large amount of empty cells and the method of estimating

the 95% confidence interval does not account for it. However, since the over estimation extends to the original data, the method that constructed the confidence interval may be inappropriate for this data set.

- Future research needs to be conducted on how the inference changes with different priors for MS and MODIPS (non-informative priors) and on a full mixture data set (both continuous and categorical).

Column #	Name	Details	Num. or Char.	Values	Range or # of Classes	Notes
1	CIT	Citizenship	Char.	1, 2, 3, 4, 5	5	
2	HIS	Hispanic Origin	Char.	1, 2, 3, 4, 5, 6	6	
3	MAR	Marital Status	Char.	1, 2, 3, 4, 5	5	
4	MIL	Served in Armed Forces	Char.	1, 2, 3, 4	4	
5	OCC	Occupation Code	Char.	1 - 14	14	
6	OI	Other Income Amount	Num.		0 - 800	99.99% are 0's
7	POB	Place of Birth	Char.	1 - 21	21	
8	RET	Retirement Income	Num.		0 - 800	99.99% are 0's
9	SCHL	Educational Attainment	Char.	1 - 6	6	
10	SEM	Self-Employment Income	Num.		0 - 800	99.99% are 0's
11	SEX	Gender	Char.	1, 2	2	
12	SS	Social Security or Railroad Retirement Income	Num.		0 - 800	99.99% are 0's
13	ATI	Adjusted Total Income	Num.		0 - 1,016,168	All Incomes and Wage 6.16% are 0's
14	AWAG	Adjusted Wages/Salary Income	Num.		0 - 1,016,167	7.54% are 0's
15	HICOV	Any Health Insurance Coverage	Char.	1, 2	2	
16	POV	In Poverty	Char.	0, 1	2	
17	CAGE	Calculated Age	Char.	1 - 10	10	
18	VETSTAT	Veteran Status	Char.	1, 2, 3	3	
19	HHT	House Hold Family Type	Char.	1 - 7	7	

Table 3.1: Summary of all the variables under consideration.

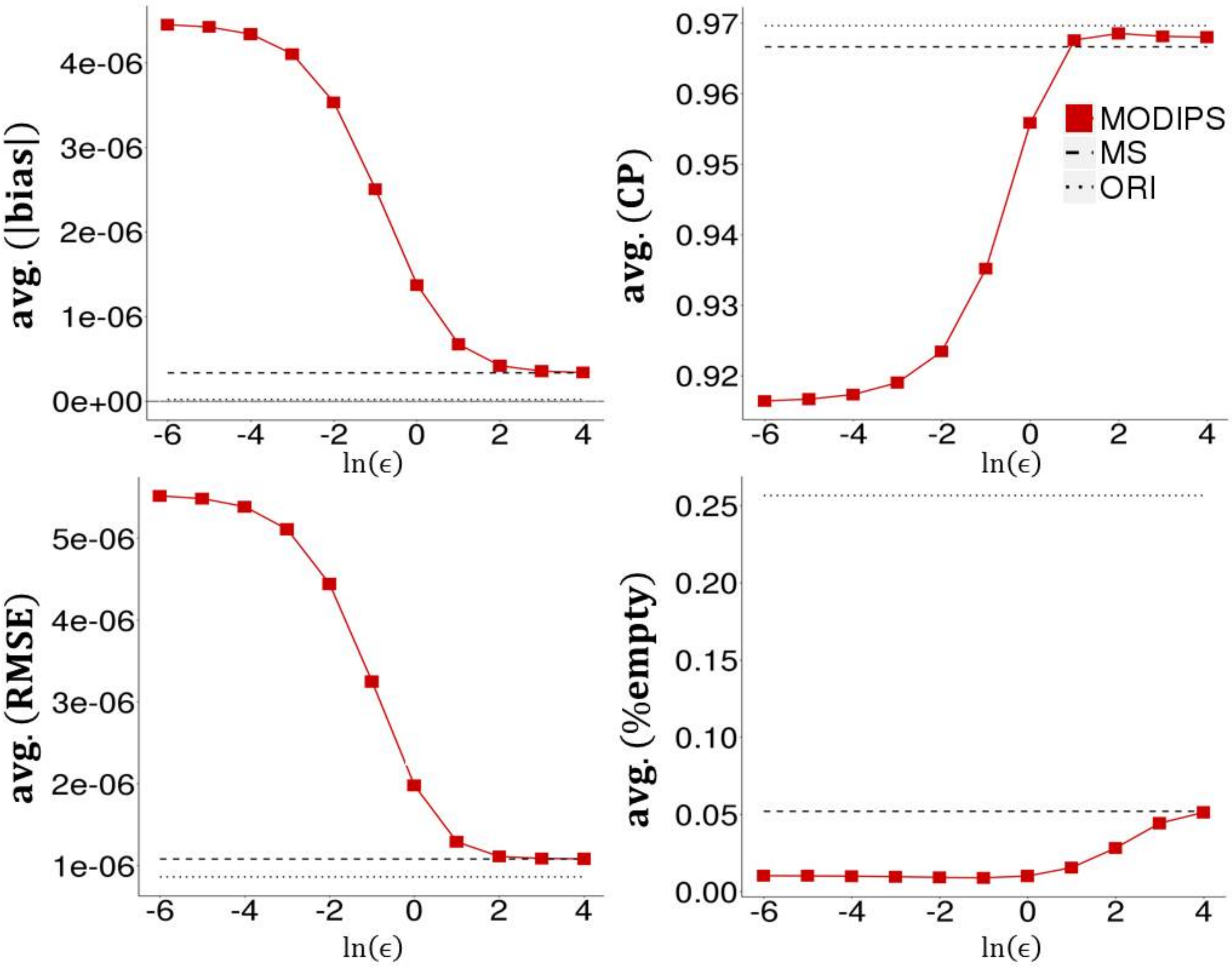

Figure 3.1: Statistical inferences of the average of π_k for $k = 1, ..., K$.

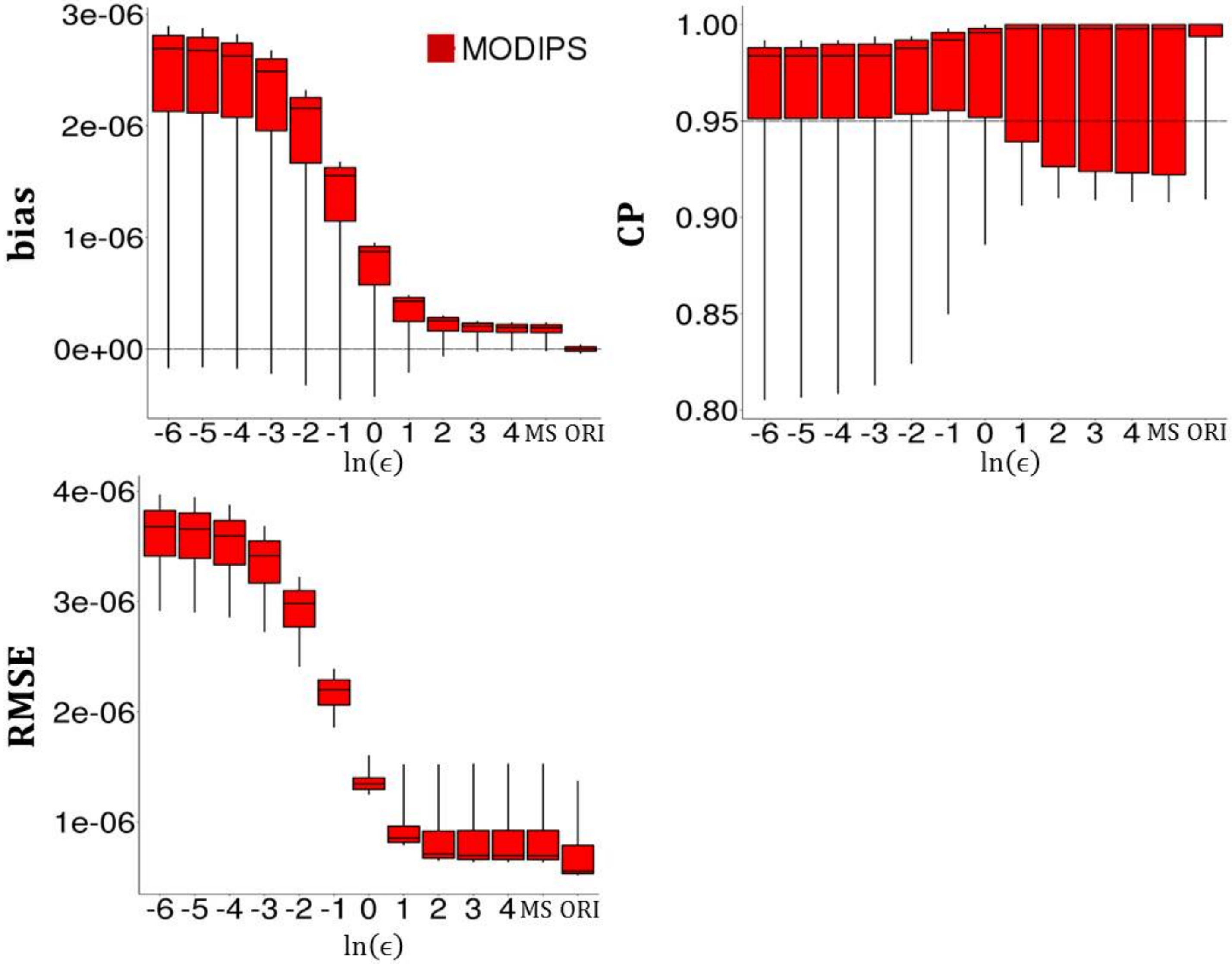

Figure 3.2: Statistical inferences of π_k for $k = 1, ..., K$, plotting the 10%, 25%, 50%, 75%, and 90% quantiles.

Appendix A

NSF GRIP Overview

Summary of the original NSF GRIP Proposal that was submitted December 4th, 2015 by the fellow and approved March 8th, 2016 by the NSF GRIP and the US Census Bureau.

A.1 NSF GRIP Proposal

A.1.1 Overview

The fellow proposes to intern at the U.S. Census Bureau and utilize their expertise and data in further developing current methodology in statistical disclosure limitation (SDL), i.e. methods of data privacy and confidentiality. Specifically, the fellow will investigate SDL techniques that preserve differential privacy — a condition on data-releasing algorithms that quantifies disclosure risk — such as the model-based differentially private data synthesis (MODIPS) (Bowen and Liu, 2016). This project will complete the fellow's thesis, which focuses on developing and validating MODIPS, by providing a case study that verifies the practical applications of MODIPS in real-world scenarios.

Intellectual Merit

The proposed project will expand the applications of differentially private data synthesis to real-world data sets by using an innovative method, modips. To the fellow's knowledge, there is only one application of differentially private data synthesis on US Census Data — OnTheMap, which uses U.S. Census Bureau commuter data (Machanavajjhala et al., 2008). With the U.S. Census Bureau's large collection of real-world data sets, the fellow will conduct case studies of incomparable precision and contemporary relevance. The fellow is fluent in R and experienced in applying differentially private data synthesis techniques.

Broader Impacts

Many entities such as statistical agencies, national security agencies, hospitals, and educational institutions will directly benefit from the proposed project by being able to share their data safely with collaborators and strike a balance between the privacy of the respondents and the efficiency and validity of the statistical inference. The U.S. Census Bureau “shares its expertise globally, and conducts work openly.” (U.S. Census Bureau’s mission statement).

A.1.2 Research Proposal

Background

In the era of information and technology, big data offers tremendous benefits for economics, national security, and other areas through data-driven decision making, insight discovery, and process optimization. However, a crucial concern is the extreme risk of exposing personal information of those who contribute to the data when sharing it among collaborators or releasing it publicly. An intruder could identify a participant by isolating the numerous connections to other contributors within the data set or linking to other public data sets — even with identifiers such as name and address removed. Some examples of identification breach include the genotype and HapMap linkage effort (Homer et al., 2008), the AOL search log release (Götz et al., 2012), and the Washington State health record (Sweeney, 2013).

Statisticians address privacy and confidentiality issues using statistical disclosure limitation (SDL). These techniques aim to provide a high level of privacy while minimizing information loss from data perturbation, thus allowing valid and integrated statistical analysis for public release. One SDL method is data synthesis, which generates a synthetic data set based on the model of the original data (Rubin, 1993; Little, 1993; Drechsler and Reiter, 2011; Drechsler, 2011).

However, the evaluation of disclosure risk on SDL methods often depends on the specific values in a given data set, as well as various assumptions regarding the background knowledge and behaviors of data intruders (McClure and Reiter, 2012; Manrique-Vallier and Reiter, 2012; Kim et al., 2015). One way to quantify disclosure risk without many such assumptions or limitations to a specific method is differential privacy (DP) (Dwork et al., 2006; Dwork, 2008, 2011). DP was originally developed for releasing summary statistics through queries submitted to a database, known as interactive privacy mechanism or query-based method.

Combining data synthesis and DP, Machanavajjhala et al. (2008) developed a differentially private mechanism, called OnTheMap, for worker commuter patterns data collected from various United States Census Bureau's data sets. The released synthetic data set shows points on a map that represent the commuting pattern of individuals from origin to destination within the US. Currently, OnTheMap is the only implementation of DP at the US Census Bureau.

Proposed Research Topic

U.S. Census Broad Topic: Simulation and Statistical Modeling
The fellow will investigate noise infusion for statistical disclosure control by applying various differentially private data synthesis techniques, including the fellow's model-based differentially private data synthesis (MODIPS), to real-world data sets (Bowen and Liu, 2016).

Appendix B

NSF GRIP Professional Development and Research

Summary of the fellow's goals:

1. to implement model-based differentially private data synthesis (MODIPS – a differential privacy method that generates a synthetic data set) on a real-world data set
2. to include some of the research conducted at the US Census Bureau to the fellow's dissertation as a case study
3. to establish collaborations for long-term projects at the US Census Bureau
4. to attend US Census Seminars, learning current research at the US Census
5. to connect with other scientists for potential future collaborations

The fellow accomplished the goals by:

1. implementing MODIPS on a real-world data set (but only on discrete data)
2. including the research into the fellow's dissertation as an example application of MODIPS
3. meeting with several researchers at US Census Bureau and will continue communication well after the internship
4. attending over a dozen seminars during the internship, such as how LiDAR improves spatial statistics and what current data privacy methods are applied at Google
5. connecting with other scientists such as Dr. Daniel Kifer, Professor of Computer Science and Engineering at Pennsylvania State University, and Dr. Jerry Reiter, Professor of Statistics at Duke University.

B.1 Research Products

B.1.1 Dissertation

The fellow will incorporate the results from the internship into her dissertation as an example of a real-world application of MODIPS.

B.1.2 Research Presentations

The fellow will present some of the findings at conferences such as the Joint Statistical Meetings, the largest international statistics conference.

Dr. Bimal Sinha invited the fellow to present her dissertation work at the University of Maryland - Baltimore County.

B.1.3 Collaborative Research

The fellow will be working with Dr. Daniel Kifer on applying differentially private data synthesis techniques on other US Census Bureau data.

B.2 Professional Development

The fellow accomplished the following professional development activities:

B.2.1 Expanded Professional Networks

The fellow met with several scientists from the Center for Statistical Research and Methodology Group and data privacy community at the US Census Bureau:

- Dr. John Abowd, Associate Director of Research and Methodology and Chief Scientist
- Dr. Aref Dajani, Principal Researcher at the Center for Disclosure Avoidance Research
- Dr. Maria M. Garcia, Research Mathematical Statistician at the Center for Statistical Research and Methodology
- Dr. Daniel Kifer, Professor of Computer Science and Engineering at Pennsylvania State University
- Dr. Martin Klein, Principal Researcher at the Center for Statistical Research and Methodology
- Dr. Amy Lauger, Principal Researcher at the Center for Disclosure Avoidance Research

- Laura McKenna, Special Assistant to the Associate Director for Research and Methodology and Chair of the Disclosure Review Board
- Edward Porter, Researcher at the Center for Statistical Research and Methodology
- Dr. Jerry Reiter, Professor of Statistics at Duke University.
- Chad Russel, IT Specialist, Center for Statistical Research and Methodology
- Dr. Kimberly Sellers, Associate Professor in the Department of Mathematics and Statistics at Georgetown University and Principal Researcher, Center for Statistical Research and Methodology
- Dr. Bimal Sinha, Professor of Statistics at the University of Maryland - Baltimore County and Principal Researcher, Research Mathematical Statistician at the Center for Statistical Research and Methodology
- Dr. Tommy Wright, Chief of the Center for Statistical Research and Methodology

The fellow also networked outside of data privacy community in the Washington D.C. area by attending the NSF Division of Graduate Education Panel at the NSF headquarterss:

- Dr. Susan E. Brennan, Program Director, NSF Division of Graduate Education
- Dr. Erik Jones, NSF GRIP Director
- Dr. Joerg Schlatterer, NSF GROW Director
- Dr. Maya Wei-Haas, Assistant Editor at the Smithsonian, resulting in an invitation to speak at the fellow's university for the Regional Women in Science Conference (http://awis.nd.edu/wsc/).

B.2.2 Presentation

The fellow presented the concept of differential privacy to the Center for Statistical Research and Methodology September Group meeting, improving the group's understanding of differential privacy as the US Census Bureau moves toward on applying the differential privacy.

B.2.3 Communication Skills

The fellow met with several scientists at the US Census Bureau, improving verbal and written communication skills.

Bibliography

Abowd, J. M. and Vilhuber, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases*, pages 239–246. Springer.

Bowen, C. M. and Liu, F. (2016). Differentially private data synthesis methods. *arXiv preprint arXiv:1602.01063*.

Drechsler, J. (2011). *Synthetic datasets for Statistical Disclosure Control*. Springer, New York.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic data sets. *Computational Statistics and Data Analysis*, 55(12):461–468.

Dwork, C. (2008). Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, 4978:1–19.

Dwork, C. (2011). Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer.

Götz, M., Machanavajjhala, A., Wang, G., Xiao, X., and Gehrke, J. (2012). Publishing search logs - a comparative study of privacy guarantees. *IEEE Trans. Knowl. Data Eng.*, 24:5205325.

Homer, N., Szelinger, S., Redmann, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., and Craig, D. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167.

Kim, H. J., Karr, A. F., and Reiter, J. P. (2015). Statistical disclosure limitation in the presence of edit rules.

Little, R. (1993). Statistical analysis of masked data. *Journal of the Official Statistics*, 9:407–407.

Liu, F. (2016). Model-based differential private data synthesis. *arXiv preprint arXiv:1606.08052*.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 277–286. IEEE.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500):1385–1394.

McClure, D. R. and Reiter, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4(1):8.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468.

Sweeney, L. (2013). Matching known patients to health records in washington state data. *Available at SSRN 2289850*.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.