Alternative Variance Estimators for Data Perturbed for Confidentiality Protection

Jianzhu Li,[1] Michael D. Larsen,[2] Tom Krenzke,[1] Laura Zayatz[3]
[1]Westat, 1600 Research Blvd, Rockville MD, 20850
[2]George Washington University, 6110 Executive Blvd, Ste. 750, Rockville, MD 20852
[3]U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233

Abstract

One method of protecting confidentiality of tabular data is to apply random perturbation on select variables in the underlying microdata. Perturbation variability needs to be appropriately accounted for in variance estimation for estimates derived from a data file altered through random perturbation. In previous work, we had studied methods for estimating variances using a single perturbed data set, and developed a variance estimator that incorporates a variance component associated with data perturbation. In this paper, we further explore three alternative approaches that can be considered in comparison to the initial estimator, with a goal of increasing the stability of the variance estimation, especially when estimates are extreme. The first alternative modifies the initial estimator through use of multiple perturbed data sets. The second alternative is a limited bootstrap approach that can be done by conducting the perturbation of the bootstrap samples multiple times, producing the replicate estimates, and subsequently computing the variance among the replicate estimates. The third alternative adjusts the initial estimator through the idea of small area estimation. Computational aspects of estimators are discussed. A simulation study was conducted to evaluate and compare the performance of the initial and alternative variance estimators using select variables in two test sites from the American Community Survey 2005-2009 sample data. The results are summarized in terms of the coverage rates and margin of errors of the estimators.

Key Words: Data perturbation, disclosure limitation, small area model, bootstrap, multiple perturbation

1. Introduction

One popular method of protecting confidentiality in tabular products is to apply random perturbation to the underlying microdata that are used to generate the tabular products. Perturbation is executed at the data preparation stage before the tabular products are generated. Unlike the Statistical Disclosure Control (SDC) treatments at the table level, this approach ensures consistency across tables because the original data are being altered. The successful application of data perturbation requires a thorough initial risk analysis beforehand to identify the variables and data values that are of high disclosure risk. After that, a random perturbation approach should be tailored to meet both goals of reducing the disclosure risk and maintaining the data usability. Last and importantly, the variance estimation approach for the tabular products must appropriately account for the additional variance caused by random perturbation. In previous work, we studied methods for estimating variances using a single perturbed data set, and developed a variance estimator that incorporates a variance component associated with data perturbation. In this paper, we further explore three alternative approaches that can be considered in comparison to the initial estimator, with a goal of increasing the stability of the variance estimation, especially when estimates are extreme. A simulation study was conducted along with the theoretical illustration to evaluate and compare the performance of the initial and alternative variance estimators using select variables in two test sites from the American Community Survey (ACS) 2005-2009 sample data. In Section 2, we introduce the background and motivation of our work, followed by a brief summary of findings in previous studies. The alternative variance estimators are discussed in Section 3. In Section 4 we describe the evaluation design and show the simulation results. A discussion and summary are given in Section 5.

2. Motivation and Previous Work

This study was motivated by Westat's work of perturbing the ACS data prior to generating the Census Transportation Planning Products (CTPP) for the American Association of State Highway and Transportation Officials (AASHTO). CTPP are large sets of tabulated data products involving dozens of variables generated by the U.S. Census Bureau at various geographical aggregations. These tables are designed to support a wide range of transportation planning needs. The data underlying the CTPP tables were changed from the Census Long Form to the ACS data in 2000. Due to the smaller size of the ACS, the CTPP tables were severely compromised because of the Census Bureau Disclosure Review Board's (DRB) extensive tabular data suppression rules in the new tables. In the NCHRP Project 08-79: Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules, Westat developed an operationally practical perturbation approach which can generate a mixed set of data with real values and randomly perturbed values. The perturbed data can be used to generate select sets of CTPP tables directly without table cells being suppressed. This enables transportation planners to make significantly better use of the ACS-based CTPP tables than they could otherwise do. The perturbation process involves four major steps: initial risk analysis, data perturbation, weight calibration, and risk and utility evaluation. Full descriptions of the methods can be found in the NCHRP final report on the study (Krenzke et al. 2011).

As an important component of the NCHRP 08-79, Westat also studied the methods for estimating variances associated with a single perturbed dataset (Krenzke et al. 2011, Li et al. 2011). The successive difference replication approach (see Fay and Train 1995; and Census Bureau 2009) was used to compute variances for the ACS estimates. Suppose $\hat{\theta}_0$ represents the ACS estimate of θ, and $\hat{\theta}_k$ is the ACS estimate of θ for replicate $k = 1,..,80$. Then the variance of $\hat{\theta}_0$ can be estimated as

$$\text{var}(\hat{\theta}_0) = \tfrac{4}{80}\textstyle\sum_{k=1}^{80}(\hat{\theta}_k - \hat{\theta}_0)^2. \qquad \text{(f1)}$$

Throughout this document, we use the formula numbers originally in Li et al. (2011). Treating the perturbed data as if they were observed and directly applying the usual ACS variance formula will result in the naïve estimator (f4):

$$\text{var}(\tilde{\theta}_0) = \tfrac{4}{80}\textstyle\sum_{\text{k}=1}^{80}(\tilde{\theta}_k - \tilde{\theta}_0)^2. \qquad \text{(f4)}$$

In the formula $\tilde{\theta}_0$ represents the CTPP perturbed estimate of θ and $\tilde{\theta}_{\mathrm{k}}$ is the estimate for replicate k. This estimator can be biased because it only accounts for the ACS sampling error, but not the variance component associated with the perturbation.

An estimator was developed to account for the additional variance due to data perturbation. It adds a term of squared difference between the ACS and perturbed estimates to the original ACS variance as follows:

$$\mathrm{var}\left(\tilde{\theta}_0\right) = \mathrm{var}(\hat{\theta}_0) + \left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2. \qquad \text{(f5)}$$

Assuming perturbation is independent of the sampling process, the formula (f5) is essentially the sum of the sampling variance and the perturbation variance. The original ACS estimate and variance, $\hat{\theta}_0$ and $\mathrm{var}(\hat{\theta}_0)$, can be computed from the unperturbed data during the process of generating the tabular products. Confidentiality is not a concern since the final users will not be able to separate $\mathrm{var}(\hat{\theta}_0)$ from $\mathrm{var}\left(\tilde{\theta}_0\right)$ or derive $\hat{\theta}_0$ from $\mathrm{var}\left(\tilde{\theta}_0\right)$ using the released tabular data. Evaluation results of a simulation study supported the use of the formula (f5) as the variance estimator for the CTPP tabulations. The Census DRB and the Census Bureau ACS Sample Design group also approved the decision of using the formula (f5) for variance estimation in the production process of the CTPP tables.

3. Alternative Variance Estimators

We extended the previous work to investigate alternative variance estimation methods that are feasible, acceptable to the user audience, as well as being accurate and stable. This research can directly benefit the Census Bureau and the users of its transportation products by further enabling effective use of perturbed ACS data for transportation planning purposes. Meanwhile, variance estimation, with perturbed data in general, is an important and interesting problem.

Three alternative approaches were considered in comparison to the variance estimation formula (f5), with a goal of increasing the stability of the variance estimation, especially when estimates are extreme.

1. The method (f5) used a single dataset in variance estimation. Methods suggested by Reiter (2003) in another context used multiple datasets to assess between perturbation variability. Following that idea, we developed a new variance estimation approach through multiple perturbations. This approach is numbered as (fn1), thereafter.
2. The second option is a limited bootstrap that could be done by perturbing the bootstrap samples of the original data, producing the replicate estimates, and subsequently measuring the variance among the estimates. This approach is numbered as (fn2), thereafter.
3. The third option is a small area estimation style estimator, which introduces a modest adjustment to (f5). The goal of the modest adjustment is to increase the stability of the variance estimator through borrowing strength and shrinkage estimation. This approach is numbered as (fn3), thereafter.

3.1 Multiple Perturbation Variance Estimator

The multiple perturbation approach replaces $\left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2$ in (f5) by an average of m, say 5, squared deviations from multiple perturbation datasets. The average of the squared deviations should be more stable than a single squared deviation. This adjustment procedure can be implemented by producing and utilizing multiple copies of the perturbed dataset. The formula for the multiple perturbation estimator is

$$v_{fn1}^{m} = \frac{4}{80}\sum_{k=1}^{80}(\hat{\theta}_k - \hat{\theta}_0)^2 + \frac{1}{m}\sum_{i=1}^{m}\left(\tilde{\theta}_0^i - \hat{\theta}_0\right)^2 \qquad \text{(fn1)}$$

where $\tilde{\theta}_0^i$ is the estimate of θ computed from the ith perturbed dataset. The estimate $\tilde{\theta}_0^i$ uses the full sample weights.

Increasing the number of replicates m should increase the stability of the variance estimator. Doing so, however, increases the computational burden. The trade-off in accuracy versus computing cost should be carefully evaluated before implementation.

3.2 Limited Bootstrap Variance Estimator

The limited bootstrap approach can be done by conducting the perturbation of the bootstrap samples multiple times, producing the replicate estimates and subsequently the variance among the estimates. This approach is computationally intensive and requires additional development and programming. The potential advantage of this approach is that the sampling and perturbation variability are captured simultaneously. A full bootstrap variance estimate based on 100s or 1000s of bootstrap replicates surely will be computationally impossible for the national ACS data due to the complexity of perturbation. In order to implement a bootstrap variance estimation methodology, a number of choices for defining the algorithm should be considered.

- Is the bootstrap drawn before or after perturbation of the ACS data? If it is before, it means that a perturbed ACS dataset and all perturbed bootstrap datasets need to be available to compute variances. If it is after, then the source of variability due to (random) perturbation logically is not captured by the bootstrap. But the amount of data that need to be made available, however, is much smaller.
- Should one draw the bootstrap sample from the original dataset with equal probabilities or with probabilities proportional to ACS sampling weights? In Rao and Wu (1988), bootstrap resampling is done with equal probabilities and weights are rescaled based on the number of times cases are selected.
- Should one draw independent bootstrap samples within strata or ignore ACS strata when drawing samples? One likely should respect the stratum structure in the original ACS sample. If not, one then introduces variability due to random sample sizes within strata. Sitter (1992) implemented Rao and Wu's (1988) rescaling bootstrap within strata (Lohr 1999; page 307). A question then is whether ACS stratum information is available on the ACS data file for this purpose. If not, then it cannot be used.
- Should one independently resample individuals, households, or clusters (of individuals or households)? If one resamples clusters, should one have resampling within sampled clusters as well? It would make sense that one should bootstrap PSUs instead of individuals directly, because the individuals are not independent of each other when there is a cluster structure organizing them. A

question then is whether appropriate cluster information is available on the ACS data file for this purpose.

- Should one rake the bootstrap weights to match ACS control totals? When the original ACS sample was drawn, it was drawn with the expectation that weights would be raked to control totals. That is, all possible ACS samples that would be considered would ultimately have weights adjusted so that they were within an accepted tolerance of raking control totals. As such, the weights for comparable samples should have the property of matching control totals. It would seem reasonable then that bootstrap samples should also have weights adjusted to match control totals.
- Should one use approximate Bayesian bootstrap or regular bootstrap? The bootstrap resamples with replacement from the distribution of observed values in order to mimic the process that yielded the original dataset. If the original dataset is small, then the resulting empirical distribution has a limited support set. Rubin (1981) proposed a Bayesian version of the bootstrap to address this situation in particular and a limitation of the bootstrap in general. An implementation of the Bayesian bootstrap in practice has become known as the approximate Bayesian bootstrap (ABB). Instead of using constant probabilities $1/_{\mathrm{n}}$ resampling, the ABB draws probabilities from a Dirichlet distribution for each bootstrap dataset. The varying probabilities are then used when drawing the bootstrap sample. The Bayesian bootstrap may not be feasible for the ACS data due to its large size and the number of additional complications (e.g., perturbation, computing time).

Decisions concerning the factors described above will to a large extent define the bootstrap replication process for the CTPP variance estimation based on perturbed ACS data.

3.2.1 Bootstrap before Perturbation Estimators

Let $\hat{\theta}_0$ be the ACS estimate. First, perturb the ACS data creating a perturbed dataset. Let $\tilde{\theta}_0$ be the estimate from the perturbed dataset. Next, independently bootstrap the ACS dataset K times and perturb each bootstrap sample. Rake each perturbed dataset to ACS control totals. Let $\tilde{\theta}_b^k$ be the estimate from the kth perturbed bootstrap sample, with the subscript b indicating "bootstrap before perturbation." Let $\bar{\tilde{\theta}}_b = \sum_{k=1}^{K} \tilde{\theta}_b^k / K$ be the average of the estimates from the K perturbed bootstrap samples. Bootstrap variance estimate centered at the average of the K perturbed bootstrap estimates is estimator (fn2 B.A): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_b^k - \bar{\tilde{\theta}}_b\right)^2$. A bootstrap variance estimate centered at the single perturbed dataset estimate is estimator (fn2 B.B): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_b^k - \tilde{\theta}_0\right)^2$. A bootstrap variance estimate centered at the ACS estimate is estimator (fn2 B.C): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_b^k - \hat{\theta}_0\right)^2$.

3.2.2 Bootstrap after Perturbation Estimators

Again, let $\hat{\theta}_0$ be the ACS estimate. The data are perturbed once, creating a perturbed dataset. Let $\tilde{\theta}_0$ be the estimate from the perturbed dataset. Independently bootstrap the perturbed dataset K times. Let $\tilde{\theta}_a^k$ be the estimate from the kth bootstrap sample of the perturbed dataset, with the subscript a indicating "bootstrap after perturbation." Let $\bar{\tilde{\theta}}_a = \sum_{k=1}^{K} \tilde{\theta}_a^k / K$ be the average of the estimates from the K bootstrap samples of the perturbed dataset. A bootstrap variance estimate centered at the average of the K bootstrap estimates is estimator (fn2 A.A): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_a^k-\bar{\tilde{\theta}}_a\right)^2$. A bootstrap variance estimate centered at the single perturbed dataset estimate is estimator (fn2 A.B): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_a^k-\tilde{\theta}_0\right)^2$. A bootstrap variance estimate centered at the ACS estimate is estimator (fn2 A.C): $\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\theta}_a^k-\hat{\theta}_0\right)^2$.

It is anticipated that centering at the ACS estimate will generate the largest variance, and centering at the average of bootstrap estimates will generate the smallest variance. Further it is anticipated that the variance estimates using bootstrapping after perturbation will be smaller than the variance estimates using bootstrapping before perturbation. As mentioned before, bootstrapping before perturbation probably will be the better option.

3.3 Small-Area Estimation Style Variance Estimator

We modified the adjustment $\left(\tilde{\theta}_0-\hat{\theta}_0\right)^2$ in (f5) to increase its stability. The idea originates from the small area estimation models. Small area estimation often takes advantage of similarity among small areas through formulas that “borrow strength" and “shrink" estimates toward a common value. The amount of shrinkage typically increases with the variance of the estimate in question. The essential idea of a modified estimator, (fn3), as an adjustment to (f5), is to replace the adjustment $\left(\tilde{\theta}_0-\hat{\theta}_0\right)^2$ by an adjustment term influenced by shrinkage, $\left(\tilde{\theta}_{0\mathrm{p}}-\hat{\theta}_0\right)^2$, where $\tilde{\theta}_{0\mathrm{p}}$ is a modified estimate. The letter p is used to denote the suggestion that the modified estimate for use in variance estimation be a certain *posterior mean* estimate. This is not suggesting a replacement of the estimate based on perturbed data by a shrunken estimate, but rather to use the modified estimate arrived at through shrinkage in the adjustment term.

The formula (fn3) is expected to improve coverage of confidence intervals. The estimates impacted by shrinkage will be less variable than the raw estimates. As a result, the adjustment factors $\left(\tilde{\theta}_{0\mathrm{p}}-\hat{\theta}_0\right)^2$ might be more stable than $\left(\tilde{\theta}_0-\hat{\theta}_0\right)^2$. In particular, small adjustments might be made a little bigger (thereby increasing coverage levels) while large adjustments might be made a little smaller (thereby reducing interval lengths). The adjustments make more of a difference at lower levels of aggregation.

Details for two versions of (fn3) are presented below. The subscript 0, which denotes that an estimate is computed using the full sample weights, is omitted from the formulas in this section. In the first version, let $\bar{\hat{\theta}}$ be the average of ACS estimates, $\hat{\theta}_j$, for parameters $j=1,\dots,J$:

$$\bar{\hat{\theta}}=\frac{1}{J}\sum_{j=1}^{J}\hat{\theta}_j.$$

Let $s_{\bar{\hat{\theta}}}^2$ be the variance of ACS estimates, $\hat{\theta}_j$, for parameters $j=1,\dots,J$.

$$s_{\bar{\hat{\theta}}}^2=\frac{1}{J-1}\sum_{j=1}^{J}\left(\hat{\theta}_j-\bar{\hat{\theta}}\right)^2$$

Consider the following model, independently for $j = 1, \dots, J$:

$$\tilde{\theta}_j \sim N\left(\theta_j, v_{f5}\left(\tilde{\theta}_j\right)\right)$$

and

$$\theta_j \sim N\left(\bar{\hat{\theta}}, s^2_{\bar{\hat{\theta}}}\right).$$

For parameters $j = 1, \dots, J$, the posterior mean (indicated by subscript p) for parameter θ_j is

$$\tilde{\theta}_{jp}(1) = \frac{v_{f5}\left(\tilde{\theta}_j\right) * \bar{\hat{\theta}} + s^2_{\bar{\hat{\theta}}} * \tilde{\theta}_j}{v_{f5}\left(\tilde{\theta}_j\right) + s^2_{\bar{\hat{\theta}}}}.$$

The first version of variance estimator (fn3, version 1) is then, for parameter $j = 1, \dots, J$:

$$v_{fn3.1,j}\left(\tilde{\theta}_j\right) = v_{f1}\left(\hat{\theta}_j\right) + \left(\tilde{\theta}_{jp}(1) - \hat{\theta}_j\right)^2.$$

The formula (f1) can be used to compute $v_{f1}\left(\hat{\theta}_j\right)$. This is variance estimator (fn3.1).

In the second version of the modified variance estimation (fn3, version 2), the average of the estimates from the perturbed datasets are used as the “empirical Bayesian” point estimate of the “prior” mean. Thus, this average changes by perturbed dataset. Let $\bar{\tilde{\theta}}$ be the average of estimates $\tilde{\theta}_j$ for parameters $j = 1, \dots, J$ from the perturbed datasets:

$$\bar{\tilde{\theta}} = \frac{1}{J} \sum_{j=1}^{J} \tilde{\theta}_j.$$

Let $s^2_{\bar{\tilde{\theta}}}$ be the variance of estimates, $\tilde{\theta}_j$, for parameters $j = 1, \dots, J$:

$$s^2_{\bar{\tilde{\theta}}} = \frac{1}{J-1} \sum_{j=1}^{J} \left(\tilde{\theta}_j - \bar{\tilde{\theta}}\right)^2$$

Consider the following model, independently for $j = 1, \dots, J$:

$$\tilde{\theta}_j \sim N\left(\theta_j, v_{f5}\left(\tilde{\theta}_j\right)\right)$$

and

$$\theta_j \sim N\left(\bar{\tilde{\theta}}, s^2_{\bar{\tilde{\theta}}}\right).$$

For parameters $j = 1, \dots, J$, the posterior mean (indicated by subscript p) for parameter θ_j is:

$$\tilde{\theta}_{jp}(2) = \frac{v_{f5}(\tilde{\theta}_j)*\bar{\bar{\tilde{\theta}}}+s^2_{\bar{\bar{\tilde{\theta}}}}*\tilde{\theta}_j}{v_{f5}(\tilde{\theta}_j)+s^2_{\bar{\bar{\tilde{\theta}}}}}.$$

The second version of variance estimator (fn3, version 2) is then, for parameter $j = 1, \ldots,$ J:

$$v_{\mathrm{fn}3.2,j}(\tilde{\theta}_j) = v_{f1}(\hat{\theta}_j) + (\tilde{\theta}_{jp}(2) - \hat{\theta}_j)^2$$

This is variance estimator (fn3.2).

A challenge comes in choosing which adjustments to include in a model together for the purpose of shrinkage. A small set of table entries could be defined by a single table. A medium set of table entries could be defined by a group of closely related tables. A large set of table entries could be defined by a broader group of related tables. A second challenge comes in making these choices when estimates are produced for different levels of aggregation. Should estimates and variance estimates from different levels of aggregation be used together in one hierarchical model? Simulation results from two different aggregations are presented later in this paper. Finally, the methodology described above is designed for estimators that have a normal distribution. For binary variables, one could consider transformations to increase the correspondence of the sampling distributions of estimators to the normality assumption. Alternatively, one could consider different small area models for non-normal variables.

4. Simulation Study

A simulation study with the aim of evaluating and comparing the performance of different variance estimators was conducted at the U.S. Census Bureau using the internal five-year ACS sample data from 2005–2009.

4.1 Simulation Design

The design of the simulation study is illustrated in the following aspects.

4.1.1 Test Sites

The input data were subset to two test sites: Olympia (OLY) and Atlanta (ATL). Due to the computational intensity of the simulation program (in SAS), the Atlanta data were subset to two neighboring counties: Henry and Clayton.

4.1.2 Level of Aggregations

Two levels of aggregations, combined traffic analysis zones (CTAZs), were created and used in the evaluation process of this simulation study: (1) CTAZ300, formed by combining traffic analysis zones until there were at least 300 sampled workers who lived in this area, and (2) CTAZ50, formed by combining traffic analysis zones until there were at least 50 sampled workers who lived in this area. The Olympia test site contains 22 CTAZ300s and 87 CTAZ50s. The Atlanta test site contains more records than Olympia and has 33 CTAZ300s and 105 CTAZ50s.

4.1.3 Outcome Variables

The outcome variables are travel time (variable name: JWMN) and minority status (variable name: MINORITY). For each CTAZ300 and CTAZ50, we computed the mean of travel time for workers who drove alone and the percentage of the minority population.

4.1.4 Perturbation

We implemented the data perturbation techniques developed during the NCHRP 08-79 project (see details in Krenzke et al. 2011), which combine the constrained hot deck and the semi-parametric perturbation approaches into one processing step. The constrained hot deck was used for ordered variables, whereas the semi-parametric was used for unordered categorical or binary variables. Partial replacement was done for each of the target variables. At the end of the data replacement, a raking procedure was run to adjust the full sample and replicate weights.

4.1.5 Number of Iterations

Two simulation programs were run for this study: the one for the limited bootstrap estimators used 300 iterations, with each involving 40 bootstrap samples; the other for the rest of the variance estimators used 400 iterations, with each involving five independently perturbed datasets for the estimator (fn1). The number of iterations was limited by the computational burden since in each single iteration the data replacement and raking process had taken a large amount of time. Even fewer iterations were executed for the limited bootstrap estimators to ensure that a reasonable number of bootstrap samples could be drawn per replicate.

4.1.6 Summary Measures

For each variance estimation method, we computed the coverage rates, the average margin of error (MOE), and the variation in the MOE from the simulated data. The coverage indicator is a binary variable that signifies whether a constructed confidence interval covers the true population value (mean travel time or percentage of minority population). The true population values were not available since ACS data were just one sample. A solution was to use the simulated true values for individual CTAZs which were drawn from a normal distribution with the ACS estimates as the mean and the variance. The coverage rate was computed as the mean of coverage indicators across all iterations. The average MOE was computed as 1.96 multiplied by the average standard errors across iterations. The coverage rates and length of MOEs indicated whether the confidence intervals based on examined variance estimators cover the true values at the nominal level. The variance of the MOE showed the stability of the variance estimators.

All except for the limited bootstrap estimators were evaluated for two test sites by two analysis variables by two geographical levels. The limited bootstrap estimators were only evaluated for travel time in CTAZ300 using the Olympia data due to its computational intensity.

4.2 Simulation Results

The simulation results from both test sites and both levels of aggregations reach similar conclusions. Only the results from the Olympia test site are presented in this paper. Sections 4.2.1 and 4.2.2 summarize the results for the analysis variables JWMN and MINORITY, respectively, at the CTAZ300 level. The results at the lower aggregation CTAZ50 are shown in Appendix A.

4.2.1 Results for JWMN

Table 1 summarizes coverage rates for 22 CTAZ300 locations. A majority of CTAZ300 locations have coverage rates within 2.5 percent of the desired nominal 95 percent. The (f4) approach has lower coverage rates compared to other estimators with some CTAZ300 locations having coverage below 92.5 percent or even below 90 percent. Estimators (f5) and (fn1) have quite similar performance. Estimators (fn3.1) and (fn3.2) have the highest percent of cases with coverage above 95 percent. They also have four cases each with coverage above 97.5 percent.

Table 1: Coverage Rate for 22 CTAZ300 Test Sites in Olympia for Variable JWMN

Olympia # of 22 CTAZ300 for JWMN Coverage Rate	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
<0.90	1	0	0	0	0
0.90, <0.925	4	0	0	0	0
0.925, <0.95	8	10	8	6	6
0.95, <0.975	9	12	14	12	12
0.975, 1.00	0	0	0	4	4
0.925, <0.975	17	22	22	18	18
Percent of CTAZ with coverage 0.925, <0.975	77%	100%	100%	82%	82%

Table 2 summarizes average MOE for 22 CTAZ300 locations for variable JWMN. All MOEs are compared to that for (f4), which is the smallest. The smallest MOE for (f4) probably contributes to its lower confidence coverage. Estimators (fn3.1) and (fn3.2) have the highest average MOE, which contributes to their slightly higher confidence coverage. The MOE is quite similar for (f5) and (fn1), but (f5) is slightly smaller.

Table 2: Margin of Error for 22 CTAZ300 Locations in Olympia for Variable JWMN

Olympia MOE in 22 CTAZ300 for JWMN	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Average MOE	2.61	2.71	2.72	2.82	2.82
Average MOE relative to (f4)	1.00	1.04	1.04	1.08	1.08

Table 3 summarizes standard deviations of MOEs across 22 CTAZ300 locations. To be clear, each of 22 CTAZ300s has MOEs computed for 400 replicates. Each set of 400 replicates produces a standard deviation (SD) of MOE. The 22 standard deviations are summarized below by their mean and first, second, and third quartiles. It is important to note that the SDs of the (fn1)'s MOE are about 50 percent less than those of (f5). Estimator (fn1) is more stable since it is based on multiple perturbed datasets. The SDs of the (fn3)'s MOE are larger than those of (f5) at the CTAZ300 level, while the results at the lower aggregation CTAZ50 in Appendix A show that the SDs of the (fn3)'s MOE can be smaller than those of (f5). This indicates the adjustment terms based on the small area models can be more stable at lower geographical level.

Table 3: Standard Deviation of Margin of Error Estimates for 22 CTAZ300 Locations in Olympia for Variable JWMN

Olympia – Standard Deviation of MOE in 22 CTAZ300 for JWMN	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Mean of SD of MOE	0.18	0.16	0.07	0.20	0.20
1st quartile of SD of MOE	0.11	0.11	0.05	0.11	0.11
median of SD of MOE	0.15	0.13	0.06	0.17	0.18
3rd quartile of SD of MOE	0.25	0.18	0.08	0.23	0.23

The six bootstrap variance estimation variations were applied only in Olympia at the CTAZ300 level on the variable JWMN. Results are presented in the three tables below. As can be seen in Table 4, coverage is not very good: several CTAZ300 locations have under coverage using all estimators.

Table 4: Coverage Rate for 22 CTAZ300 Locations in Olympia for Variable JWMN using Bootstrap Variance Estimation

Olympia # of 22 CTAZ300 JWMN Coverage Rate	*Variance Estimators for Bootstrap*					
	A.A	*A.B*	*A.C*	*B.A*	*B.B*	*B.C*
<0.90	8	7	6	7	7	6
0.90, <0.925	5	5	4	6	4	4
0.925, <0.95	4	4	6	4	5	5
0.95, <0.975	3	4	4	3	4	5
0.975, 1.00	2	2	2	2	2	2
0.925,<0.975	7	8	10	7	9	10
Percent of CTAZ with coverage 0.925, <0.975	32%	36%	45%	32%	41%	45%

Table 5 shows the average MOE across bootstrap alternatives. They are comparable to the MOE of (f4).

Table 5: Margin of Error for 22 CTAZ300 Test Sites in Olympia for Variable JWMN using Bootstrap Variance Estimation

Olympia MOE in 22 CTAZ300 for JWMN	*Variance Estimators for Bootstrap*					
	A.A	*A.B*	*A.C*	*B.A*	*B.B*	*B.C*
Average MOE	2.53	2.62	2.66	2.56	2.66	2.69
Average MOE relative to (f4)	0.97	1.00	1.02	0.98	1.02	1.03

Table 6 presents standard deviation of MOE estimates for the six bootstrap variance estimation alternatives. The first three bootstrap alternatives that introduce less variability into the bootstrap procedure produce less variable MOE estimates. This makes sense given the different amount of variability captured by the two classes of procedures (bootstrap after perturbation versus bootstrap before perturbation).

Table 6: Standard Deviation of Margin of Error Estimates for 22 CTAZ300 Test Sites in Olympia for Variable JWMN using Bootstrap Variance Estimation

Olympia – Standard Deviation of MOE in 22 CTAZ300 for JWMN	*Variance Estimators for Bootstrap*					
	A.A	*A.B*	*A.C*	*B.A*	*B.B*	*B.C*
Mean of SD of MOE	0.26	0.27	0.25	0.32	0.30	0.32
1st quartile of SD of MOE	0.21	0.23	0.21	0.26	0.24	0.26
Median of SD of MOE	0.27	0.29	0.25	0.32	0.29	0.30
3rd quartile of SD of MOE	0.31	0.33	0.31	0.38	0.36	0.36

A major conclusion of this part of the study is that it is likely that more than 40 bootstrap replicates will be required in order to stabilize MOE estimates and improve coverage levels.

4.2.2 Results for MINORITY

Methods were also implemented on the binary variable MINORITY, which takes values of “Yes” and “No.” Coverage rates for CTAZ300 locations are given in Table 7. Results for (f5) and (fn1) are very comparable and quite good. Results for this variable using other estimators are a little bit worse than they were for JWMN. For the (fn3.1) and (fn3.2) estimators, this is possibly because MINORITY is a binary variable and the normality assumption in the small area estimation model does not hold very well. In particular, the variances in binomial models depend on the means, so one could consider implementing a transformation to stabilize the variance (e.g., the arcsin square root variance stabilizing transformation).

Table 7: Coverage Rates for 22 CTAZ300 Locations in Olympia for Variable MINORITY

Olympia # of 22 CTAZ300 for MINORITY Coverage Rate	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
<0.90	10	0	0	2	2
0.90, <0.925	7	0	0	1	0
0.925, <0.95	3	7	8	10	10
0.95, <0.975	2	15	14	7	8
0.975, 1.00	0	0	0	2	2
0.925, <0.975	5	22	22	19	20
Percent of CTAZ with coverage 0.925, <0.975	23%	100%	100%	86%	91%

A summary of the MOEs for variable MINORITY in 22 CTAZ300 locations is given in Table 8. The MOEs are much smaller due to the scale of MINORITY (binary) versus JWMN (continuous). As before, (f4) is the smallest on average. The other estimators have larger average MOEs. As with JWMN, for variable MINORITY the estimators (fn3.1) and (fn3.2) are impacted by a few outliers.

Table 8: Margin of error for 22 CTAZ300 locations in Olympia for variable MINORITY

Olympia MOE in 22 CTAZ300 for MINORITY	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Average MOE	0.0512	0.0599	0.0602	0.0600	0.0601
Average MOE relative to (f4)	1.000	1.169	1.175	1.172	1.173

Table 9 presents standard deviations of MOE across 22 CTAZ300 locations in Olympia. Qualitatively, the results are similar to those of variable JWMN: estimator (fn1) has a smaller variability in MOE estimates.

Table 9: Standard Deviation of Margin of Error Estimates for 22 CTAZ300 Locations in Olympia for Variable MINORITY

Olympia – Standard Deviation of MOE in 22 CTAZ300 for MINORITY	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Mean of SD of MOE	0.004	0.009	0.006	0.008	0.008
1st quartile of SD of MOE	0.003	0.006	0.004	0.005	0.005
Median of SD of MOE	0.003	0.008	0.005	0.007	0.007
3rd quartile of SD of MOE	0.004	0.010	0.007	0.009	0.010

For both analysis variables and both geographical areas, a common pattern is seen. Estimators (f5) and (fn1) have similar coverage rates, and estimators (fn3.1) and (fn3.2) have similar coverage rates. Estimator (f4) has lower coverage rates compared to other estimators. It also has the smallest estimated margin of error, which probably contributes to lower confidence coverage. Estimators (fn3.1) and (fn3.2) have the highest average margin of error, which contributes to their slightly higher confidence coverage. Margin of error is quite similar for (f5) and (fn1). However, the standard deviations of the (fn1)'s MOE are about 50 percent less than those of (f5). The higher stability of (fn1) comes from averaging adjustment terms from multiple perturbed datasets. The SDs of the (fn3)'s MOE can be either larger or smaller than those of (f5).

5. Summary

For both test sites, both levels of aggregation and both analysis variables, we found that the gain from using estimators (f5) and (fn1) accounting for perturbation errors over the other options is quite obvious in terms of better coverage rates and, particularly in the case of (fn1), smaller standard deviation for margins of error. If computing resources are available, then the variance estimator (fn1) is recommended over (f5). This statement is due to the fact that (fn1) provides stable estimates of margins of error through its averaging of multiple independent squared deviations, whereas (f5) relies on a single squared deviation to account for variation due to perturbation. Estimator (f5) worked well and remains an acceptable choice.

In the future, we could further investigate the performance of (fn1) and (f5) in the following aspects:

- Compare (fn1) and (f5) estimators in terms of coverage, margin of error, and stability of variance estimates on more test sites, on more variables, and at different levels of aggregations.
- Evaluate the computational issues regarding (f5) and (fn1) on a much broader, and perhaps national, scale.
- For (fn1), examine the issue of the number of replications ($\mathrm{M} = 5{,}10{,}20, \ldots$) in terms of the performance of the multiple perturbation estimator.

In addition, further study can be carried out on the small area estimation style estimators and the limited bootstrap approach. For the small area estimation style estimators, we could examine the model and potential variance stabilizing transformations especially for the estimator of proportions based on binary data. Additionally, we could examine the role of outliers on the modeling. It was noted that a few CTAZ locations appeared to have outlying results. Robust shrinkage methods or shrinkage methods that push adjustments (through strategically developed prior distributions) toward the (f5) adjustment might improve performance. For the limited bootstrap approach, we could examine the issue of the number of replicates.

References

Fay, R. and Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. Joint Statistical Meetings Proceedings of the Section on Survey Research Methods of the American Statistical Association.

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. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.

Li, J., Krenzke, T., Brick, M., Judkins, D., and Larsen, M. (2011). Variance estimation for Census Transportation Planning Products with perturbed American Community Survey data. Joint Statistical Meetings Proceedings of the Section on Survey Research Methods of the American Statistical Association.

Lohr S. L. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. Survey Methodology, 29, 181-188.

Rao, J. N. K. and Wu, C. F. J. (1988), Resampling inference with complex survey data, Journal of the American Statistical Association, 83(401), 231–241.

Rubin, D. B. (1981). The Bayesian Bootstrap. Annals of Statistics, 9, 130-134.

Sitter, R. R. (1992). A resampling procedure for complex survey data. Journal of the American Statistical Association, 87, 755-765.

U.S. Census Bureau (2009). Design and methodology: American Community Survey. http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

Appendix A: Simulation Results for 87 CTAZ50 Locations in Olympia

Table A-1: Coverage Rate for 87 CTAZ50 Locations in Olympia for Variable JWMN

Olympia # of 87 CTAZ50 for JWMN Coverage Rate	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
<0.90	20	0	0	2	2
0.90, <0.925	16	2	3	4	5
0.925, <0.95	33	33	36	26	26
0.95, <0.975	17	50	46	40	39
0.975, 1.00	1	2	2	15	15
0.925, <0.975	50	83	82	66	65
Percent of CTAZ with coverage 0.925, <0.975	57%	95%	94%	76%	75%

Table A-2: Margin of Error for 87 CTAZ50 Locations in Olympia for Variable JWMN

Olympia *MOE in 87 CTAZ50 for JWMN*	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Average MOE	4.98	5.42	5.44	5.78	5.76
Average MOE relative to (f4)	1.00	1.09	1.09	1.16	1.16

Table A-3: Standard Deviation of Margin of Error Estimates for 87 CTAZ50 Locations in Olympia for Variable JWMN

Olympia – Standard Deviation of MOE in 87 CTAZ50 for JWMN	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Mean of SD of MOE	0.55	0.41	0.19	0.41	0.41
1st quartile of SD of MOE	0.27	0.23	0.11	0.18	0.18
Median of SD of MOE	0.41	0.33	0.16	0.32	0.32
3rd quartile of SD of MOE	0.65	0.48	0.24	0.47	0.47

Table A-4: Coverage Rate for 87 CTAZ50 Locations in Olympia for Variable MINORITY

Olympia # of 87 CTAZ50 for MINORITY Coverage Rate	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
<0.90	35	1	1	4	3
0.90, <0.925	18	0	1	7	9
0.925, <0.95	23	30	33	32	30
0.95, <0.975	11	55	52	35	34
0.975, 1.00	0	1	0	9	11
0.925, <0.975	34	85	85	67	64
Percent of CTAZ with coverage 0.925, <0.975	39%	98%	98%	77%	74%

Table A-5: Margin of Error for 87 CTAZ50 Locations in Olympia for Variable MINORITY

Olympia MOE in 87 CTAZ50 for MINORITY	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Average MOE	0.0978	0.1147	0.1153	0.1168	0.1172
Average MOE relative to (f4)	1.000	1.173	1.179	1.194	1.198

Table A-6: Standard Deviation of Margin of Error Estimates for 87 CTAZ50 Locations in Olympia for Variable MINORITY

Olympia – Standard Deviation of MOE in 87 CTAZ50 for MINORITY	*Variance Estimators*				
	(f4)	*(f5)*	*(fn1)*	*(fn3.1)*	*(fn3.2)*
Mean of SD of MOE	0.010	0.017	0.012	0.015	0.015
1st quartile of SD of MOE	0.006	0.010	0.007	0.007	0.007
Median of SD of MOE	0.008	0.014	0.010	0.012	0.012
3rd quartile of SD of MOE	0.013	0.021	0.015	0.018	0.019