Variance Estimation for the Census Transportation Planning Products with Perturbed American Community Survey Data

Jianzhu Li[1], Tom Krenzke[1], Mike Brick[1], David Judkins[1], Michael Larsen[2]
[1]Westat, 1600 Research Blvd., Rockville MD, 20850
[2]George Washington University, 6110 Executive Blvd, Ste 750, Rockville, MD 20852

Abstract

Census Transportation Planning Products (CTPP) are sets of tabulated data products designed for transportation planners. As the underlying data are moving from the Census Long Form sample to the smaller American Community Survey (ACS) five-year combined sample, disclosure risk becomes a non-avoidable concern, especially for small geographic areas. A perturbation approach was developed so that the CTPP products based on the perturbed data satisfy the transportation data user community's analytical needs while simultaneously satisfying the requirements set by the U.S. Census Bureau for reducing disclosure risk. This paper discusses the variance estimation on the CTPP tables using perturbed ACS data. The ACS uses the Successive Difference Replication (SDR) method for variance estimation because it has the advantage that the variance estimates can be computed regardless of the form of the statistics or the complexity of the design. However, the SDR estimator applied naively to the perturbed data does not account for the variance due to perturbation. As a remedy, an additional term was added to reduce the bias. The proposed estimators are compared with a few alternatives and evaluated through a simulation study.

Key Words: Balanced Repeated Replication, data perturbation, disclosure risk, perturbation variance, successive differences replication, synthetic data

1. Introduction

The Census Transportation Planning Products (CTPP) are sets of custom tabulations for the transportation community. A large volume of tables are generated by the U.S. Census Bureau at various geographical aggregations to support a wide range of transportation planning needs. In 2000, the data underlying the CTPP tables were based on the Census Long Form. Since then, the Long Form has been replaced by the American Community Survey (ACS) and the tables must now be run using the ACS. With the transition, the CTPP tables face serious cell suppression, especially for small geographical areas due to the smaller sample size in the ACS. Of course, the tables must satisfy the disclosure rules required by the Disclosure Review Board (DRB) at Census Bureau. The National Highway Cooperative Research Program (NCHRP) was concerned that large scale cell suppression would severely reduce the data usability, and therefore called for research on statistical disclosure control (SDC) that would provide transportation planners with high quality unsuppressed tables which strictly conform to the Census Bureau DRB's disclosure rules. Sponsored by the NCHRP and the National Academy of Sciences (NAS), Westat put together a research team that worked cooperatively to develop an

operationally feasible data perturbation approach to fulfill this goal. The proposed approach can be used to perturb the ACS data adequately so that tabulations based on the perturbed data will be approved for publication without any cell suppression.

This paper studies the variance estimation for the CTPP tables generated using the perturbed ACS data through the proposed perturbation approach. The usual ACS variance estimator was designed for estimates based on unperturbed ACS data so that perturbation variance is not naturally part of it. As a remedy, we propose to add an adjustment term to the usual ACS variance estimator to appropriately account for the error due to data perturbation. In Section 2, we briefly describe the disclosure risk elements in the CTPP tabulations, and the perturbation approach developed to reduce the risk. Variance estimators for estimates using perturbed data are presented and compared in Section 3, including an empirical example. In Section 4, we present the results from a simulation study which further evaluates the performance of different variance estimators. Final conclusions and remarks are given in Section 5.

2. Disclosure Risk and Perturbation Approach

The CTPP products include residence-based tables, workplace-based tables, and residence-to-workplace flow tables. The tables involve demographic variables (e.g., age of workers, minority status), socio-economic variables (e.g., household income, person earning, poverty status), and transportation variables (e.g., means of transportation, travel time, time leaving home), and show cell aggregates, means, and medians. The smallest geography for the tables will be at the Traffic Analysis Zone (TAZ) level, which is roughly similar to block groups. The new CTPP products will be processed from the 2006-2010 ACS combined sample, with a sample size that is only about half of the 2000 Census Long Form sample size. As a result, the disclosure risk becomes a serious concern since the smallest TAZs will have just 20-25 ACS sample workers. The tables of flows for each TAZ result in a large number of sample uniques.

A disclosure risk in the CTPP tables arises from the ability to link the tables to build microdata records (Krenzke and Hubble, 2009) with restricted geographical information, and match the records to the ACS Public Use Microdata Sample (PUMS) to obtain an additional 150 variables or so. Also, the sample uniques that are for scenarios such as long distance bicycle/walker commuters between two known TAZs are likely to be population uniques. The typical disclosure rules set up by the Census Bureau DRB would impact about 90 million CTPP tables in various geographical areas. These tables involve about 30 to 50 percent of all microdata in 90 percent of the TAZs.

As an alternative to cell suppressions, we developed a general approach to perturbing the ACS microdata (see Krenzke et al., 2011a and 2011b). The Census Bureau DRB approved this approach since the tables generated from the perturbed data have substantially reduced disclosure risk. As a result, the tables are not subject to cell suppression prior to publication. The approach begins with an initial risk analysis to flag high risk values by forming the tables and applying the disclosure rules. Next, in the data replacement step, the high risk values are perturbed using either a semi-parametric approach or a constrained hotdeck approach, depending upon the variable types. Both approaches change the high risk values slightly while maintaining the associations between variables in TAZs with medium to large sample sizes. After the data replacement step, a weighting calibration process called raking (Deming and Stephan,

1940) is applied to bring consistency between certain ACS estimates and estimates based on perturbed data at the Public Use Microdata Area (PUMA) level. An evaluation was conducted that concluded that the perturbation approach achieves a good balance on retaining data utility and reducing disclosure risk.

3. Variance Estimation with Perturbed Data

The ACS has very complex survey design and weighting adjustment process (U.S. Census Bureau, 2009). To approximate the variances of the estimates under this design, the ACS implements the Successive Difference Replication (SDR) approach (Wolter, 1984; Fay & Train, 1995; Judkins, 1990). The SDR approach is designed to be used with systematic samples such as ACS, for which the sample is selected from a frame sorted by geographic ordering. Its main advantage is that the variance estimates can be computed for all sorts of statistics despite the complexity in the sampling or weighting procedures.

Suppose $\hat{\theta}$ represents the ACS estimate of a statistic, θ, using the ACS full sample weight, and $\hat{\theta}_k$ is the ACS estimate of θ using the kth set of ACS replicate weights (See U.S. Census Bureau, 2009, Chapter 12, for the formation of the replicate weights). Then the variance of $\hat{\theta}$ can be estimated using the SDR formula as

$$\mathrm{var}\left(\hat{\theta}\right) = \frac{4}{80}\sum_k \left(\hat{\theta}_k - \hat{\theta}\right)^2 \tag{1}$$

In this research we focus on developing a variance estimator for the subset of the CTPP tables that will be generated using the perturbed ACS data. They are referred to as perturbed tables or perturbed estimates. A variance estimator that can appropriately estimate the variance of the perturbed estimates should capture two components, variance due to sampling error, and variance due to perturbation given an ACS sample.

One simple idea for estimating the variance of the perturbed estimates would be just treating the perturbed data as if they were not perturbed and applying the ACS formula naively. This naïve variance estimator is

$$\mathrm{var}\left(\hat{\varphi}\right) = \frac{4}{80}\sum_k \left(\hat{\varphi}_k - \hat{\varphi}\right)^2 \tag{2}$$

where $\hat{\varphi}$ is the estimate based on perturbed data and raked full sample weight, and $\hat{\varphi}_k$ is the estimate based on perturbed data and the kth set of raked replicate weight (Note that after data perturbation the raking process is applied to both the full sample weight and each set of the replicate weights). This variance estimator reduces to the usual ACS estimator in equation (1) when the perturbation error is small enough to ignore. It can be shown that, under certain assumptions, the estimated variance from a replication method based on the perturbed data, in expectation with respect to perturbation, is not very different from the variance based on the unperturbed data. In the Appendix we prove this using the Balanced Repeated Replication (BRR) estimator as an example. The SDR estimator, and other replication estimators, should have similar properties though it is more difficult to work through the theoretical proof. More general, the naïve estimator is likely to be negatively biased since variance due to data perturbation is not accounted for. The bias could be serious if the amount of perturbation is moderate to large.

An alternative straightforward solution for estimating the error component due to perturbation given an ACS sample is to add an adjustment term comprised of the squared difference between the ACS and perturbed estimates, $\left(\hat{\varphi}-\hat{\theta}\right)^2$. If there is no perturbation bias ($E_p(\hat{\varphi}|s)=\hat{\theta}$, where E_p means expectation with respect to random perturbation conditional on the sample *s*), then the expectation of this adjustment term is essentially the perturbation variance $\left(E_p\left(\left(\hat{\varphi}-\hat{\theta}\right)^2|s\right)=\sigma_{ps}^2\right)$. If there is some perturbation bias, then this term estimates the mean squared error. Adding the adjustment term to the naïve estimator (2) gives an estimator we call the naïve-with-adjustment estimator,

$$\text{var}\left(\hat{\varphi}\right)=\frac{4}{80}\sum\nolimits_k\left(\hat{\varphi}_k-\hat{\varphi}\right)^2+\left(\hat{\varphi}-\hat{\theta}\right)^2. \tag{3}$$

An alternative estimator to formula (3) is to add the adjustment term to the usual ACS estimator (1). We call this the ACS-with-adjustment estimator. Assuming the perturbation is independent of the sampling process, the estimator given in (4) is essentially the sum of sampling variance and perturbation variance,

$$\text{var}\left(\hat{\varphi}\right)=\frac{4}{80}\sum\nolimits_k\left(\hat{\theta}_k-\hat{\theta}\right)^2+\left(\hat{\varphi}-\hat{\theta}\right)^2. \tag{4}$$

The microdata underlying the CTPP tabulations will not be released, which prevents the users from deriving the unperturbed estimate $\hat{\theta}$ by separating the two components in formulae (3) and (4). While the ACS estimate is well protected against being disclosed, using it in the variance formulae enables us to estimate the perturbation variance from only one perturbed dataset (in other words multiple perturbed data are not needed).

We begin our investigation of these alternative variance estimators by giving an empirical example. Figure 1 shows the estimated standard errors (SEs) of the county-level mean travel time for workers who drove alone, using the ACS data from 2005 to 2009. The computations were based on the original ACS and the perturbed dataset using the proposed perturbation approach for the test site Atlanta. The horizontal axis represents the 20 counties in Atlanta. They are sorted in an increasing order of the estimated SEs based on the usual ACS estimator (1). The SEs computed from the ACS estimator (1) and the naïve estimator (2) are very similar, and generally smaller than the SEs computed from the naïve with adjustment estimator (3) and the ACS-with-adjustment estimator (4). The estimated SEs computed from (3) and (4) account for the difference in the point estimates from the original and the perturbed data. This second term was moderate or large for some of the counties, but small or close to zero for others. We suspect this occurred because the post-perturbation raking attenuates the difference in the estimates due to perturbation. Although travel time was one of the raking dimensions at the PUMA level, the county-level estimates based on the perturbed data were not fully aligned with the estimates based on the original ACS data, especially in large PUMAs containing a few counties.

4. Simulation

To further evaluate the proposed variance estimators, a simulation study was conducted. In the simulation, the perturbation approaches developed by the research team were

applied to the data from one test site, Olympia, and the perturbations were applied 1,000 times independently. From each of the 1,000 independent perturbed datasets, the mean travel time for workers who drove alone within each Combined TAZ (CTAZ) was calculated; a CTAZ contained at least 300 workers living in the area. The variances were computed using three different estimators: naïve estimator (2), naïve-with-adjustment estimator (3), and ACS-with-adjustment estimator (4).

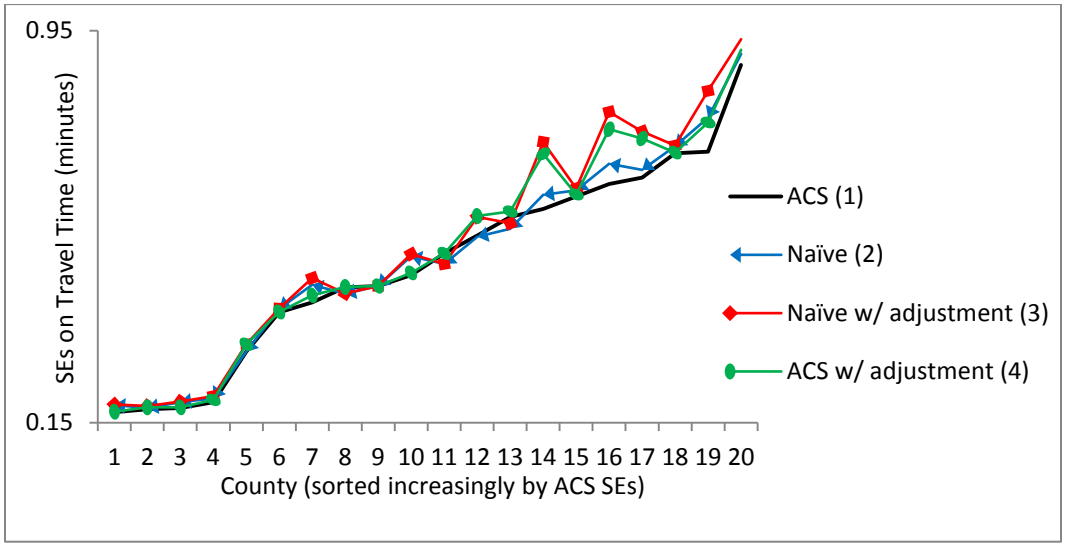

Figure 1: Estimated standard errors (in minutes) of the mean travel time for workers who drove alone in 20 counties in Atlanta: ACS 2005-2009.

Table 1 shows the relative difference between the average of the 1,000 perturbed estimates and the ACS estimates for each CTAZ, as well as the ratios of the average standard errors from (3) and (4) to the standard error from the usual ACS estimator (1). The perturbation noise is generally no more than two percent of the ACS estimates in most of the CTAZs, but reaches three percent in CTAZ 17 and 5 percent in CTAZ 16. A majority of the perturbed standard errors from the ACS-with-adjustment estimator (4) are 2 to 9 percent higher than those from the usual ACS estimator. In CTAZ 16, the standard error from (4) is 28 percent higher than that from (1). The perturbed standard errors from (3) are similar to those from (4) but they can sometimes be lower than the ACS estimated standard errors. It appears that the data perturbation process only adds a small amount of noise to the ACS data for large CTAZs, which makes the perturbed estimates deviate only slightly from the original estimates. The impact in small areas such as TAZs was not evaluated because confidentiality concerns mandated that these be substantially perturbed.

We computed coverage rates to evaluate whether the variance estimators appropriately account for both the sampling error and the perturbation error. Coverage rates summarize how well the constructed confidence intervals covers the true values through independently repeated sampling and perturbation processes. However, the true values were not available in this study, since Olympia ACS data was just one sample. Therefore, instead of drawing repeated ACS samples, we drew the simulated true values (mean travel time) for individual CTAZs from a normal distribution with the ACS point estimate as the mean and the ACS variance estimate as the variance, assuming the ACS point and variance estimates from the unperturbed data for each CTAZ were

approximately unbiased. We computed, on average, how likely the confidence intervals based on the perturbed estimates contained the randomly drawn true values. The results are presented in Figure 2. Each boxplot is based on 22 averages (one for each CTAZ).

Table 1: Relative Difference between ACS and Perturbed Estimates, and Ratios of Standard Errors from (3) and (4) to that from Usual ACS Estimator (1), by CTAZ: ACS 2005-2009, Olympia

CTAZ	*Relative Diff between perturbed estimate and ACS estimate*	*Ratio of standard errors: (3)/(1)*	*Ratio of standard errors: (4)/(1)*
1	0.01	1.03	1.06
2	0.00	0.94	1.03
3	0.01	1.00	1.05
4	0.02	1.08	1.08
5	0.01	0.93	1.03
6	0.01	1.09	1.06
7	0.01	1.04	1.04
8	0.02	1.09	1.10
9	0.01	1.09	1.04
10	0.01	1.07	1.06
11	0.01	1.09	1.08
12	0.02	1.18	1.10
13	0.00	1.08	1.02
14	0.01	1.07	1.04
15	0.01	1.21	1.06
16	0.05	1.30	1.28
17	0.03	1.12	1.09
18	0.01	1.18	1.09
19	0.01	0.88	1.03
20	0.01	1.16	1.04
21	0.01	1.04	1.06
22	0.00	1.08	1.05

The coverage for the naïve estimator (2) was always lower than the coverage based on the naïve-with-adjustment estimator (3). The coverage rates from the naïve estimator were acceptable for some CTAZs, but could be lower than the nominal rates for majority of the CTAZs, and even fell below 90 percent occasionally. This clearly showed that the naïve estimator (2) did not capture the variance due to perturbation appropriately.

The coverage rates were very close to the nominal 95 percent when the ACS-with-adjustment estimator (4) was used to estimate the variance. The performance of the confidence intervals based on the naïve-with-adjustment estimator (3) was also good, but slightly less stable than those based on (4). This reflects the instability of the naïve estimator in estimating the sampling error since the second components in (3) and (4) are identical.

5. Conclusions and Remarks

The ACS-with-adjustment estimator outperforms the other estimators for variance estimation on the perturbed estimates. It uses the original ACS full sample estimates and replicate estimates, as well as the perturbed full sample estimates. This estimator is actually computationally simpler and more stable than the naïve-with-adjustment estimator. The adjustment term, the squared difference between the original and

perturbed estimates, serves as an appropriate estimate of the conditional perturbation variance. It ensures that the confidence intervals constructed on the perturbed estimates and variances have the coverage rates close to the nominal, even when the perturbation process may have introduced some bias to the estimates. A disadvantage of using this adjustment term for estimating the error due to perturbation is that it is only based on one set of perturbed data. Using multiple independently perturbed datasets for perturbation variance estimation would effectively improve the stability. However, generating multiple perturbed datasets could dramatically increase the time and effort in the data perturbation process given the large sample size of ACS and the perturbation approach being used.

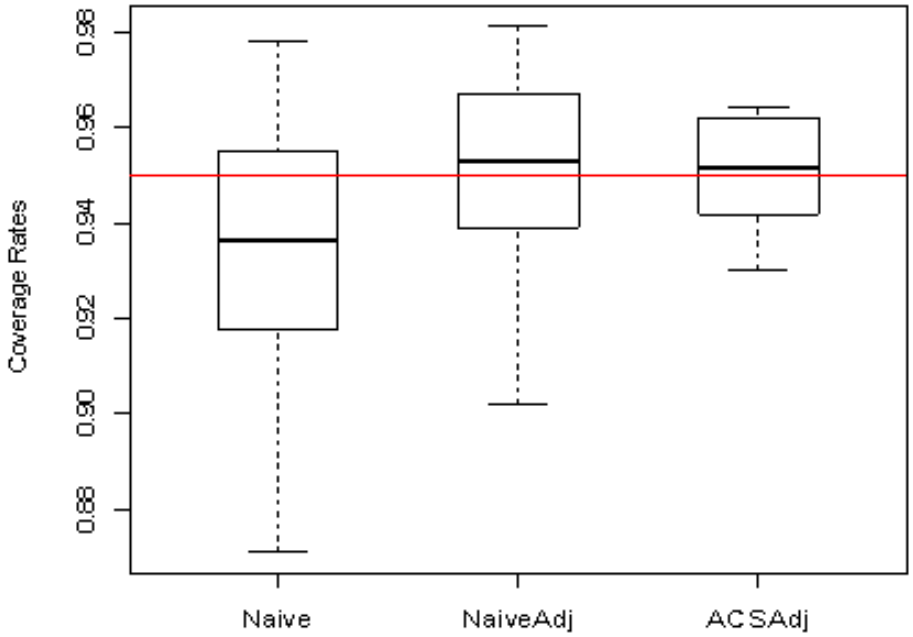

Figure 2: Coverage rates of confidence intervals based on three variance estimators (left: Naïve estimator; middle: Naïve-with-adjustment estimator; right: ACS-with-adjustment estimator)

Reiter (2003) discusses generating multiple datasets with partial synthesis to facilitate variance estimates that account for between dataset error variance. Assume perturbations are made independently for $i = 1, \ldots, m$ to yield m different perturbed datasets. Let $\hat{\varphi}^i$ denote the CTPP perturbed estimate of θ based on the ith perturbed data and $v(\hat{\varphi}^i)$ denote the estimated variance of $\hat{\varphi}^i$, treating the ith perturbed data as being unperturbed (e.g., computed using the naïve estimator). Under certain regularity conditions, the analyst can obtain valid inferences for θ by combining $\hat{\varphi}^i$ and $\hat{v}(\hat{\varphi}^i)$ as follows:

$$\bar{\hat{\varphi}} = \frac{1}{m}\sum_i \hat{\varphi}^i ,$$

$$\text{var}\left(\bar{\hat{\varphi}}\right) = \frac{1}{m}\sum_i \hat{v}\left(\hat{\varphi}^i\right) + \frac{1}{m}\frac{1}{m-1}\sum_i \left(\hat{\varphi}^i - \bar{\hat{\varphi}}\right)^2 \qquad (5)$$

The point estimate is the average of the m perturbed estimates, $\bar{\hat{\varphi}}$. The variance of $\bar{\hat{\varphi}}$ is the sum of two components, with the first term estimating the sampling error and the

second term estimating the perturbation variance, or the variation between the perturbed estimates. This set of estimators is designed for publishing multiple perturbed datasets for which the analysts will be able to conduct any types of analyses that they desire. For the CTPP products Census Bureau will only release a set of pre-defined tabulations, but not microdata.

There are two drawbacks to applying (5) directly to the CTPP products. First, (5) does not use the original ACS estimates in variance estimation because they are unknown to the analysts. The variance estimates can be biased if in expectation the perturbation noise is not zero. But for the CTPP tables both the point estimates and the variances are produced by the Census Bureau for whom the ACS estimates are available. Moreover, there is no disclosure concern associated with using the original ACS estimates since the users have no way to separate the ACS estimates from the overall variance estimates. Therefore, using the ACS estimates in variance estimation for the perturbed estimates is safe, and can improve the variance estimation due to both sampling error and perturbation error. Second, in formula (5), the perturbation noise can be attenuated through averaging across multiple perturbed estimates. The proposed perturbation approach intends to change the high risk values slightly for the purpose of retaining the data utility. Using the average as the point estimate may not reduce the disclosure risk adequately.

Other than directly applying formula (5) to the CTPP products, we may just borrow the idea of estimating the perturbation variance through multiple perturbed data for improving stability. We can use one perturbed estimate, say $\hat{\varphi}^1$, as the point estimate for publication, and estimate its variance as

$$\mathrm{var}\left(\hat{\varphi}^1\right) = \frac{4}{80}\sum\nolimits_k \left(\hat{\theta}_k - \hat{\theta}\right)^2 + \frac{1}{m}\sum\nolimits_i \left(\hat{\varphi}^i - \hat{\theta}\right)^2,$$

where the first term is the usual ACS estimator, the best available for estimating the sampling error, and the second term is the variation between the *m* perturbed estimates. Again, the feasibility of using multiple perturbed data for variance estimation heavily depends on the efficiency of the data perturbation process and the table generating process. The current plan is to create single perturbed data for generating the CTPP tables to assure efficiency by sacrificing some degree of stability in estimating the perturbation variance.

6. Appendix

Assume the BRR half samples (Wolter, 2007) are fully orthogonal, we have $\hat{\theta} = \frac{1}{B}\sum_b \hat{\theta}_b$ and $\hat{\varphi} = \frac{1}{B}\sum_b \hat{\varphi}_b$ in a perturbed dataset, where *B* is the number of BRR replicate weights. We further assume that perturbation does not introduce any bias, i.e., $E_p\hat{\varphi} = \hat{\theta}$ and $E_p\hat{\varphi}_b = \hat{\theta}_b$. Taking the expectation with respect to perturbation *p*, we obtain

$$E_p \frac{1}{B}\sum\nolimits_b \left(\hat{\varphi}_b - \hat{\varphi}\right)^2$$

$$= E_p \frac{1}{B}\sum\nolimits_b \left(\hat{\varphi}_b^2 - \hat{\varphi}^2\right)$$

$$= \frac{1}{B}\sum_b \left(\hat{\theta}_b^2 + \text{var}_p\left(\hat{\varphi}_b\right) - \hat{\theta}^2 - \text{var}_p\left(\hat{\varphi}\right) \right)$$

Since the half samples are orthogonal, we have $\frac{1}{B}\sum_b var_p(\hat{\varphi}_k) = var_p(\hat{\varphi})$. Then,

$$E_p \frac{1}{B}\sum_b \left(\hat{\varphi}_b - \hat{\varphi}\right)^2 = \frac{1}{B}\sum_b \left(\hat{\theta}_b - \hat{\theta}\right)^2 .$$

If the BRR replicate weights have been adjusted to account for nonresponse or poststratification, the above equation may only approximately hold.

References

Deming, W.E. and F.F. Stephan (1940), On a least square adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, 11, 427–444.

Fay, R. and Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *JSM Proceedings*, Section on Government Statistics. Alexandria, VA. 154-159.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6 (3), 223-239.

Krenzke, T. and Hubble, D. (2009). Toward quantifying disclosure risk for area-level tables when public microdata exists. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 4707-4717.

Krenzke, T., Li, J., Freedman, M. Judkins, D., Hubble, D., Roisman, R., and Larsen, M. (2011a). Producing transportation data products from the American Community Survey that comply with disclosure rules. Final report prepared for the National Cooperative Highway Research Program Transportation Research Board of The National Academies.

Krenzke, T., Li, J., Judkins, D., and Larsen, M. (2011). Evaluating a constrained hotdeck to perturb American Community Survey Data for the Census Transportation Planning Products. *Proceedings of the Joint Statistical Meetings*, American Statistical Association Section on Survey Research Methods.

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

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

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79, 781-790.

Wolter, K. M. (2007). *Introduction to Variance Estimation*, Second Edition, New York: Springer-Verlag.