Austrian Journal of Statistics
July 2019, Volume 48, 43–57.
http://www.ajs.or.at/
doi:10.17713/ajs.v48i4.790

Evaluation of Synthetic Small-area Estimators Using Design-based Methods

Partha Lahiri
University of Maryland

Santanu Pramanik
National Council of Applied Economic Research

Abstract

The use of area-specific design-based mean squared error (MSE) to measure the uncertainty associated with synthetic and direct estimators is appealing since the same model-free criterion is applied. However, the small sample size is often a difficulty in obtaining a reliable estimator of the area-specific design-based MSE. Moreover, the area-specific design-based mean squared error estimator might yield undesirable negative values under certain circumstances. The existing solution to overcome the problem of small sample size is to consider average design-based MSE, average being taken over the available small areas. This may not solve the other problem of negative MSE. An alternative average design-based mean squared error estimator is proposed which always produces positive estimates. Simulation shows that this estimator performs better than the existing average design-based MSEs as it always produces positive estimates and accounts for the bias component usually present in synthetic estimators.

Keywords: bias, borrow strength, design-consistency, average MSE.

1. Introduction

Sample surveys designed to produce reliable estimates for a population cannot guarantee estimates of similar precision for small subgroups of the surveyed population, known as small areas or domains. For example, in the National Health and Nutrition Examination Survey (NHANES) III, overall sample size is too small to spread across the fifty states and the District of Columbia. Moreover, certain minority groups residing predominantly in certain states (e.g., California and Texas) are oversampled. Thus the NHANES III sample design results in small or no samples for the states, especially those states that do not have large populations for these minority groups. This problem of small or zero sample size in small areas prevents the use of usual direct survey estimates for small area parameters since the estimates are likely to be highly unreliable or unavailable.

In the absence of adequate direct information for small areas, it is customary to borrow strength from related sources to form indirect estimators that increase the effective sample size and hence reduce sampling errors of estimators. Such indirect estimators are usually based on implicit or explicit models that combine information from the sample survey, various administrative/census records, or previous surveys. There is a large volume of literature on indirect estimation for small areas. We refer to Rao and Molina (2015), Jiang and Lahiri (2006) and Pfeffermann (2013) for a detailed account on small area estimation.

Synthetic estimation, which applies to any probability and non-probability sample design, is an indirect method of borrowing strength from similar areas.There are a variety of synthetic estimators available in the literature. A synthetic estimator is not area specific in the study variable of interest. Synthetic methods are often employed in practice for their simplicity and ability to produce estimates for areas with no sample from the sample survey. When the survey does not provide any sample for many areas, a synthetic method may be appealing to public policy makers as the same estimation method is applied to all areas, irrespective of whether an area has sample or not.

Hansen, Hurwitz, and Madow (1953) presented an early example of a regression method to produce synthetic estimates of median number of radio stations heard during the day for over 500 counties of the United States. Ericksen (1974) developed a synthetic method, called the sample regression method, for estimating population changes of local areas. Nicholls (1977) considered such a regression method to produce synthetic estimates in Australia. Stasny, Goel, and Rumsey (1991) developed a regression-synthetic method for estimating county acreage of wheat using a non-probability sample of farms along with auxiliary data on planted acreage and district indicators. Marker (1999) and Rao and Molina (2015) presented more examples of synthetic small area estimators based on regression models.

Although regression models were used in producing synthetic estimates in earlier applications, more sophisticated models could be used. As an alternative to linear regression modeling, certain synthetic predictors have been proposed based on M-quantile regression (Chambers and Tzavidis 2006; Chambers, Chandra, and Tzavidis 2011; Fabrizi, Salvati, Pratesi, and Tzavidis 2014). The reweighting method (Schirm and Zaslavsky 1997), which uses a Poisson regression, can be also viewed as a synthetic method where the survey weights are synthetically determined to match certain known controls. A major advantage of such reweighting method is that for a given small area a single set of weights can be applied to estimate small area characteristics for a large number of variables. Elbers, Lanjouw, and Lanjouw (2003) developed a micro-simulation method for poverty mapping, commonly referred to as the ELL method or the World Bank method, using a linear mixed model to capture different design features of the survey design. Molina and Rao (2010) classified the ELL method as a synthetic method. A synthetic estimator may not use an explicit model. One such synthetic estimator is the well-known structure preserving estimator (SPREE), given in Chambers and Feeney (1977) and Purcell and Kish (1980).

In official statistics, a major advance on synthetic estimation, based on implicit models, came with an experiment conducted by researchers at the U.S. National Center for Health Statistics (NCHS) during the mid-1960's to estimate disability and other health characteristics for small areas (NCHS 1968). The U.S. Census Bureau followed the NCHS's lead in producing county level synthetic estimates of unemployment rates (Gonzalez and Hoza 1978). Researchers at the U.S. Census Bureau have been currently exploring the possibility of producing estimates of different components of the 2010 decennial census coverage errors using a synthetic method based on logistic regression.

Although a number of synthetic estimators have been suggested in the small area estimation literature, their evaluation received relatively less attention. The variance of a synthetic estimator, whether design-based or model-based, as a measure of evaluation could be misleading as it does not incorporate the bias component of the synthetic estimator, which could be more prominent than the variance component and does not vanish even for large samples. The design-based mean squared error (MSE) may be more acceptable in evaluating a synthetic estimator since it incorporates both the design-based variance and bias components and does not rely on the implicit or explicit model used to generate the synthetic estimator. The variance of a synthetic estimator can usually be estimated with precision using any variance estimation technique – design-based or model-based – since this involves variance estimation of parameter estimates (e.g., common regression coefficient or common mean) based on a large sample (usually whole sample, not area specific sample). While it is possible to obtain an unbiased or an approximately unbiased estimator of the design-based bias of a synthetic estimator, such an estimator is unreliable since it is based on a small sample size. For an illustration, the readers are referred to Example 4 of Section 4.

Recognizing the problem associated with the area specific design-based MSE estimation, Gonzalez and Waksberg (1973) introduced the concept of an average design-based MSE of a set of synthetic estimators and suggested its estimator. Their average design-based MSE does not use any implicit or explicit model used to generate synthetic estimates and is model-free. Thus different synthetic estimators and direct estimators can be compared using the same criterion. However, a major problem with the Gonzalez-Waksberg estimator is that it can produce undesirable negative average MSE estimates. In a recent study related to the coverage measurement program of the U.S. Census Bureau, the Gonzalez-Waksberg method frequently produced negative average design-based MSE estimates, which motivate our exploration of alternative estimators of average design-based MSE.

The outline of this paper is as follows. In Section 2, we introduce a few notations. In Section 3, we discuss synthetic estimation. In Section 4, we review the design-based MSE estimators proposed by Gonzalez and Waksberg (1973) and Marker (1995) and explain the issues associated with estimating design-based MSE for an individual small area. We then introduce two average design-based MSE estimators in Section 5. Our proposed average design-based MSE estimators produce strictly positive estimates and design-consistent for a large number of small areas in the group. We compare proposed average MSE estimators with other rival estimators using real life data analyses (Section 6) and a Monte Carlo simulation (Section 7). Chambers *et al.* (2011) proposed an estimator of the area specific MSE of their proposed synthetic estimator based on the estimation of variance and bias under certain modeling assumptions. It is not, however, clear how their method can be applied to a wide range of synthetic predictors proposed in the literature. In contrast, we proposed estimators of average design-based MSEs for the purpose of evaluation of different small area estimators and not for any specific inferential purpose such as constructing confidence intervals. Our method can be used to evaluate any synthetic estimators, including that of Chambers and Tzavidis (2006). We conclude the paper with a conclusion section (Section 8) where we summarized our research findings and discuss future research in this area.

2. A table of notations

We use the following notations throughout the paper (unless otherwise defined):

m: number of small areas,

U: set of units in the finite population of interest, which contains m small areas $U_i, i = 1, \ldots, m$,

N_i: size of U_i, $(N = \sum_{i=1}^{m} N_i)$,

y_k: value of a characteristic (discrete or continuous) of interest for the kth unit in U,

$Y_i = \sum_{k \in U_i} y_k$, ith area population total,

$\bar{Y}_i = \frac{Y_i}{N_i}$, ith area population mean,

s: set of units in the sample,

s_i: set of units in s that belong to area i (with fixed size n_i),

w_k: sampling weight associated with unit $k \in s$.

Using the notations above, we define survey-weighted direct estimators of $\bar{Y}_i$ as $\hat{\bar{Y}}_i^D = \frac{\sum_{k \in s_i} w_k y_k}{\sum_{k \in s_i} w_k}$. Throughout the paper, we assume that $\hat{\bar{Y}}_i^D$ is design-unbiased or approximately so. An example of such $\hat{\bar{Y}}_i^D$ would be the Horvitz-Thompson estimator of the small area mean with w_k fixed and $\sum_{k \in s_i} w_k = N_i$, the known fixed population size for the ith small area.

3. Synthetic estimation

We describe two different synthetic estimators - one based on an area level model and the other based on an unit level model. In either case, a fixed effects or mixed model could be used, but the synthetic estimator will not be area specific with respect to the study variable y.

3.1. Area level model

We first fit a model that relates survey-weighted means $\hat{\bar{Y}}_i^D$ to a set of auxiliary variables at the small area level and then use fitted values for the areas as synthetic estimates of the small area means.

Example 1: In obtaining synthetic estimators of small area means $\bar{Y}_i$, we follow two steps:

Step 1: Fit a multiple regression model: $\hat{\bar{Y}}_i^D = \mathbf{X}_i^T \boldsymbol{\beta} + \xi_i$, where $\mathbf{X}_i$ is a vector of known auxiliary variables for area i; $\boldsymbol{\beta}$ is a vector of unknown regression coefficients; $\{\xi_i,\ i = 1, \cdots, m\}$ are uncorrelated errors with means 0 and known variances σ^2.

Step 2: A synthetic estimator of $\bar{Y}_i$ is then given by $\hat{\bar{Y}}_i = \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = (\sum_{j=1}^{m} \mathbf{X}_j \mathbf{X}_j^T)^{-1} \sum_{j=1}^{m} \mathbf{X}_j \hat{\bar{Y}}_j^D$$

is the ordinary least squares estimator of $\boldsymbol{\beta}$.

Example 2: The regression model in Example 1 does not incorporate sampling errors of direct estimates $\hat{\bar{Y}}_i^D$. A more reasonable model in this case would be the following linear mixed non-normal model proposed by Lahiri and Rao (1995) as an extension of the well-known Fay-Herriot model (Fay and Herriot 1979): $\hat{\bar{Y}}_i^D = \mathbf{X}_i^T \boldsymbol{\beta} + v_i + e_i,\ i = 1, \cdots, m$, where $\{v_i,\ i = 1, \cdots, m\}$ and $\{e_i,\ i = 1, \cdots, m\}$ are uncorrelated with $v_i \sim (0, A)$ and $e_i \sim (0, V_i^D)$, and sampling variances V_i^D are estimated using certain external information using the Generalized Variance Function (GVF) method (Otto and Bell 1995). The above linear mixed model motivates the following synthetic estimator of $\bar{Y}_i$: $\hat{\bar{Y}}_{iw} = \mathbf{X}_i^T \hat{\beta}_w$, where

$$\hat{\beta}_w = \left(\sum_{j=1}^{m} \frac{\mathbf{X}_j \mathbf{X}_j^T}{\hat{A} + V_j^D} \right)^{-1} \sum_{j=1}^{m} \frac{\mathbf{X}_j \hat{\bar{Y}}_j^D}{\hat{A} + V_j^D}$$

is the weighted least squares estimator of $\boldsymbol{\beta}$ and $\hat{A}$ is a consistent estimator of A. For example, $\hat{A}$ can be an ANOVA estimator (Lahiri and Rao 1995).

3.2. Unit level model

We fit an unit level model to the entire survey data in order to establish a relationship between the target variable and a set of auxiliary variables and then obtain the synthetic estimators of the small area means as $\hat{\bar{Y}}_i = N_i^{-1} \sum_{k \in U_i} \hat{y}_k$, where $\hat{y}_k$ is the fitted value of y_k from the model. Depending on the model, we may or may not need the auxiliary variables for all units in the population. For example, if we fit a multiple linear regression model, we need the small area population means for these auxiliary variables to estimate $\bar{Y}_i$. On the other hand, for a non-linear model like the logistic model, we need the auxiliary variables for all units of the population.

Example 3: Let y_k be a binary variable associated with an attribute and consider estimation of the total number of units in the ith small area satisfying the attribute. Assume the following logistic model: logit $[P(y_k = 1)] = \mathbf{x}_k^T\boldsymbol{\beta} + \xi_k$, where ξ_k are iid with means 0 and variances σ^2. A synthetic estimator of $\bar{Y}_i$ is then obtained using the following two simple steps:

Step 1: Obtain the usual survey-weighted estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, of the logistic model using the entire survey data.

Step 2: A synthetic estimator of $\bar{Y}_i$ is then given by $\hat{\bar{Y}}_i = N_i^{-1}\sum_{k\in U_i}\hat{y}_k$, where $\hat{y}_k = \frac{\exp(\mathbf{x}_k^T\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_k^T\hat{\boldsymbol{\beta}})}$, the fitted value of y_k from the logistic model.

In the U.S. Census Coverage Measurement (CCM) program, logistic models have been explored in order to estimate the three necessary probabilities of the logistic regression dual system estimator (Mule 2010). The CCM program also evaluated the performance of average MSE estimators for the synthetic estimates of the number of erroneous enumeration using the 2000 census coverage data. Their logistic model included main effects due to different factors such as census region, place size grouping, tenure, race/ethnic origin domain, sex as well as different interactions such as domain by tenure, tenure by sex, among others. Alternatively one can consider a more realistic mixed model that incorporates random effects to account for clustering and/or small area specific random effects to account for between area variability not explained by the fixed effects. For such a mixed model, a synthetic estimator would be derived using the implied marginal model as in section 3.1 and hence would not be small area specific in the study variable y.

4. Design-based mean squared error of synthetic estimators

We define the design-based mean squared error (MSE) of a synthetic estimator $\hat{\bar{Y}}_i$ of the ith small area mean $\bar{Y}_i$ as

$$\text{MSE}(\hat{\bar{Y}}_i) \equiv M_i = E(\hat{\bar{Y}}_i - \bar{Y}_i)^2 = V_i + B_i^2;\ i = 1, \cdots, m \tag{1}$$

where $V_i = V(\hat{\bar{Y}}_i)$, variance of $\hat{\bar{Y}}_i$, $B_i = E(\hat{\bar{Y}}_i) - \bar{Y}_i$, bias of $\hat{\bar{Y}}_i$, and the expectations and variances are with respect to the sample design.

The variances V_i are generally small as they are based on the total sample size, which is usually large. But the term B_i^2 is independent of the sample size and its magnitude depends on the validity of the synthetic assumption that generates the synthetic estimators.

Example 4: Stratified simple random sampling

We consider a stratified simple random sampling design in which a simple random sample of same size n is selected from each stratum. We assume strata and small areas to be identical. A direct estimator of $\bar{Y}_i$ is given by $\hat{\bar{Y}}_i^D = \bar{y}_i$, the sample mean for area i. For illustration, we consider the overall sample mean, for example, $\hat{\bar{Y}}_i = \bar{y} = m^{-1}\sum_{i=1}^m \bar{y}_i$, as a synthetic estimator of $\bar{Y}_i$.

Ignoring the finite population correction (fpc), we have

$$V(\hat{\bar{Y}}_i^D) \equiv V_i^D \approx \frac{S_i^2}{n} = O\left(\frac{1}{n}\right),\ \text{Bias}(\hat{\bar{Y}}_i^D) \equiv B_i^D = 0, \tag{2}$$

$$V(\hat{\bar{Y}}_i) \equiv V_i \approx \frac{\overline{S^2}}{nm} = O\left(\frac{1}{nm}\right),\ \text{Bias}(\hat{\bar{Y}}_i) \equiv B_i = \bar{Y} - \bar{Y}_i, \tag{3}$$

where $\bar{Y}$ is the overall population mean; S_i^2 is the population element variance for area i and $\overline{S^2} = m^{-1}\sum_{i=1}^{m} S_i^2$.

The direct estimator $\hat{\bar{Y}}_i^D$ is design-unbiased. For a large within area sample size n, $\hat{\bar{Y}}_i^D$ and the associated variance estimator $\hat{V}_i^D = s_i^2/n$, where s_i^2 is the sample variance, are also design-consistent. While design-consistency property is necessary, it is not sufficient in a small area estimation problem as n is typically small. On the other hand, the synthetic estimator is design-biased, and the extent of the bias depends on the validity of the synthetic assumption $\bar{Y}_i = \bar{Y}$; $i = 1, \ldots, m$. The sample size has no impact on B_i, but V_i is typically small since nm is large.

In the MSE expression Eq. (1), the variance component is likely to be negligible relative to the squared bias component. Hence, it is important to obtain a reliable estimate of the squared bias component of the MSE. For large m, a design-consistent estimator of V_i is $\hat{V}_i = \hat{\overline{S^2}}/(nm)$, where $\hat{\overline{S^2}} = m^{-1}\sum_{i=1}^{m} s_i^2$. We can estimate B_i by an unbiased estimator $\hat{B}_i = \bar{y} - \bar{y}_i$, but this estimator is unreliable due to the high variability of $\bar{y}_i$, unless the within area sample size n is large.

A naïve estimator of M_i is given by $\hat{M}_i^{\text{Naïve}} = \hat{V}_i$; $i = 1, \ldots, m$, where $\hat{V}_i$ is a design-consistent estimator of V_i, for large m. For large m and usual regularity conditions, $E(\hat{M}_i^{\text{Naïve}}) \approx M_i - B_i^2$, and thus $\hat{M}_i^{\text{Naïve}}$ underestimates the true design-based MSE, irrespective of the sample size, and the extent of the underestimation depends on the validity of the synthetic assumption that generates the synthetic estimators.

Using simple algebra (Rao 2003, p. 52), we have

$$\text{MSE}(\hat{\bar{Y}}_i) = E(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - V_i^D + 2\text{Cov}(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) \tag{4}$$

$$= E(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - V(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D) + V_i. \tag{5}$$

Assume $\text{Cov}(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) \approx 0$ and $\text{E}(\hat{\bar{Y}}_i^D) = \bar{Y}_i$, where $X \approx Y$ means $X - Y$ tends to zero under certain regularity conditions, for large m. This assumption holds in many situations. For Example 4, $\text{Cov}(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) = O\left(\frac{1}{mn}\right)$ and $\text{E}(\hat{\bar{Y}}_i^D) = \bar{Y}_i$. Under these assumptions, Eq. (4) becomes $\text{MSE}(\hat{\bar{Y}}_i) \approx E(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - V_i^D$, which motivates the Gongalez-Waksberg (GW) estimator of M_i:

$$\hat{M}_i^{\text{GW}} = (\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - \hat{V}_i^D, \tag{6}$$

where $\hat{V}_i^D$ is an usual design-based estimator of V_i^D (Gonzalez and Waksberg 1973). Since within area sample size is typically small, this is a highly unstable estimator of M_i. Moreover, estimates $\hat{M}_i^{\text{GW}}$ could be negative.

Marker (1995) obtained an area specific design-based MSE estimator for small area total estimation. Such a formula can be extended when the parameters of interest are the small area means. Using Eq. (1) and the synthetic assumption $B_i^2 = K$; $i = 1, \ldots, m$, one obtains the Marker's estimator as

$$\hat{M}_i^{\text{M}} = \hat{V}_i + \hat{K}, \tag{7}$$

where

$$\hat{K} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i^D - \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i. \tag{8}$$

Marker's method produces reasonable MSE estimators only when the variance is a substantial proportion of the MSE and the variances differ across small areas (Marker 1995, p. 74; Rao 2003, p. 53). Unlike the Gonzalez-Waksberg MSE estimator, Marker's MSE estimator can be used even when there is no sample for a given small area - one needs to ignore that area in estimating the average squared bias. However, stability of $\hat{M}_i^{\text{M}}$ is achieved at the expense of introducing bias due to the synthetic assumption. Moreover, like the Gonzalez-Waksberg estimator, it can also yield negative MSE estimate since average bias-squared estimate ($\hat{K}$) can be negative.

5. Estimation of average design-based mean squared error

Following Gonzalez and Waksberg (1973), we define the average design-based mean squared error (AMSE) of a set of m synthetic estimators of small area population means as

$$\text{AMSE} \equiv M = \bar{V} + \eta, \tag{9}$$

where $\bar{V} = m^{-1}\sum_{i=1}^{m} V_i$ and $\eta = m^{-1}\sum_{i=1}^{m} B_i^2$. In some cases, it may be meaningful to consider a weighted average of MSEs in Eq. (1) with respect to known weights. The methodology proposed can be extended to this more general case, but we sacrifice the generalizability for simplicity in exposition.

A naïve AMSE estimator is given by

$$\hat{M}^{\text{Naïve}} = m^{-1}\sum_{i=1}^{m} \hat{V}_i = \hat{\bar{V}}, \tag{10}$$

where $\hat{V}_i$ is a design-consistent estimator of V_i; $i = 1, \cdots, m$. For large m and usual regularity conditions, $E(\hat{M}^{\text{Naïve}}) \approx M - \eta$, and thus $\hat{M}^{\text{Naïve}}$ underestimates the true AMSE, irrespective of the sample size, and the extent of the underestimation depends on the validity of the synthetic assumption that generates the synthetic estimators.

By averaging $\hat{M}_i^{\text{GW}}$ (Eq. 6) over small areas, we get the Gonzales-Waksberg (GW) AMSE estimator as

$$\hat{M}^{\text{GW}} = F_1 - F_2, \tag{11}$$

where

$$F_1 = m^{-1}\sum_{i=1}^{m}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2, \; F_2 = m^{-1}\sum_{i=1}^{m}\hat{V}_i^D. \tag{12}$$

The average of Marker's design-based MSE formula is identical to the Gonzalez-Waksberg average design-based MSE estimator $\hat{M}^{\text{GW}}$. The GW estimator of average design-based MSE is design-consistent for large m, under suitable regularity conditions, but it can produce undesirable negative estimate of average design-based MSE.

It is interesting to note that, Gonzales-Waksberg average design-based MSE estimator can be obtained by maximizing the following function:

$$f(M; \hat{M}^{\text{GW}}) = M^{-\frac{1}{2}} \exp\left(-\frac{\hat{M}^{\text{GW}}}{2M}\right) \tag{13}$$

with respect to M in the range $(-\infty, \infty)$. Depending on the value of $\hat{M}^{\text{GW}}$, the function $f(M; \hat{M}^{\text{GW}})$ can take maximum in the negative half of the real line. Maximizing $f(M; \hat{M}^{\text{GW}})$ in the range $M \in (0, \infty)$ does not solve the problem as the maximum could be attained on the boundary at zero.

To overcome the potential problem of non-positivity of GW estimator, we resort to the approach introduced by Yoshimori and Lahiri (2014). We propose the following adjustment to $\hat{M}^{\text{GW}}$:

$$\hat{M}^{\text{GW-adj}} = \begin{cases} \hat{M}^{\text{GW}}; & \text{if } \hat{M}^{\text{GW}} > 0 \\ \hat{M}^{\text{GW}*}; & \text{otherwise,} \end{cases} \tag{14}$$

where $\hat{M}^{\text{GW}*}$ is a maximizer of the adjusted function $a(M)f(M; \hat{M}^{\text{GW}})$ with respect to M in the range $(0, \infty)$, and $a(M) = \left[\arctan\left(M\sum_{j=1}^{m}\frac{1}{M+V_j^D}\right)\right]^{\frac{1}{m}}$.

For large m, $\log[a(M)f(M; \hat{M}^{\text{GW}})] = \log[f(M; \hat{M}^{\text{GW}})] + \log[a(M)] \approx \log[f(M; \hat{M}^{\text{GW}})]$. Hence, $\hat{M}^{\text{GW-adj}}$ is close to $\hat{M}^{\text{GW}}$ (the design-consistent GW estimator), illustrating the design-consistency of $\hat{M}^{\text{GW-adj}}$. Since $a(0) = 0$, $\hat{M}^{\text{GW-adj}}$ is always strictly positive. The adjustment factor $a(M)$

was proposed earlier by Yoshimori and Lahiri (2014) in adjusting the residual maximum likelihood (REML) estimator of the unknown variance component of the Fay-Herriot model so that the resulting adjusted REML estimator retains asymptotic properties of the REML while yielding a strictly positive estimate.

One downside of $\hat{M}^{\text{GW-adj}}$ is that it may be less than the naïve estimator $\hat{M}^{\text{Naïve}}$, which does not incorporate the bias component. We propose an adjustment to the Marker's area specific design-based MSE estimator as

$$\hat{M}_i^{\text{M-adj}} = \hat{V}_i + \hat{K}^{\text{M-adj}}, \tag{15}$$

where

$$\hat{K}^{\text{M-adj}} = \begin{cases} \hat{K}; & \text{if } \hat{K} > 0 \\ \hat{K}^*; & \text{otherwise,} \end{cases} \tag{16}$$

where $\hat{K}^*$ is a maximizer of $a(K)f(K;\hat{K})$ with respect to $K \in (0,\infty)$. This leads to the following average design-based MSE estimator:

$$\hat{M}^{\text{M-adj}} = m^{-1}\sum_{i=1}^{m}\hat{M}_i^{\text{M-adj}} \tag{17}$$

By construction, $\hat{M}^{\text{M-adj}}$ is strictly positive and always larger than the naïve estimator. Also, using the previous arguments, it can be shown that $\hat{K}^{\text{M-adj}} \approx \hat{K}$, for large m, so that $\hat{M}^{\text{M-adj}} \approx \hat{M}^{\text{GW}}$, showing the design-consistency of $\hat{M}^{\text{M-adj}}$. Proposed average design-based MSE estimators should work as long as there are a large number of similar small areas with positive sample sizes within the group.

6. Data analysis

We use state level data for the years 1991, 1993, and 1997 from the Small Area Income and Poverty Estimates (SAIPE) program of the U.S. Census Bureau in order to compare the following four estimators of average design-based MSE: naïve ($\hat{M}^{\text{Naïve}}$, Eq. 10), GW ($\hat{M}^{\text{GW}}$, Eq. 11), GW-adj ($\hat{M}^{\text{GW-adj}}$, Eq. 14), and Marker-adj ($\hat{M}^{\text{M-adj}}$, Eq. 17). For details about the SAIPE program, we refer to Citro and Kalton (2000). The parameters of interest are proportions of 5-17 year old children in poverty for the fifty states and the District of Columbia for the years 1991, 1993, and 1997. The survey-weighted proportions $\hat{\bar{Y}}_i^D$ are obtained using the Current Population Survey (CPS) data. The sampling variance estimates $\hat{V}_i^D$ are obtained using a Generalized Variance Function (GVF) method (Otto and Bell 1995). The Census Bureau now uses the American Community Survey (ACS) data as the direct source of survey data to produce state level SAIPE estimates. Also, the Census Bureau does not use synthetic estimators for state level SAIPE. Nonetheless, we find it convenient to illustrate the performances of different average design-based MSE estimators using the old SAIPE data.

We obtain synthetic estimates using an area level multiple regression synthetic model described in Example 1 of Section 3.1. As for the auxiliary variables $\mathbf{X}_i$, we use Internal Revenue Service (IRS) data, food stamp data and census residuals, which were used in the SAIPE program. First we form four distinct groups of states so that the number of states within a group is about the same (13 states in each of three groups and 12 in the other). The small areas in a given group are similar in terms of the sampling variance estimates $\hat{V}_i^D$. For a given year the reliability of direct estimates varies more across groups than within a group. For example, direct estimates for the states within the first group are much more reliable than those in the second group.

We present our data analysis in Table 1. The GW average design-based MSE estimator (which is same as Marker average design-based MSE estimator) frequently produces negative estimates. In fact, average bias-squared estimator as suggested by Marker (1995), turned out

Table 1: Evaluation of synthetic estimator using different average MSE estimators: Small area income and poverty estimates data analysis

		Average MSE Estimates of Synthetic Estimator				Coefficient of Variation	
Grouping of states by $\hat{V}_i^D$	No. of states	Naïve	GW	GW-adj	Marker-adj	Direct	Marker-adj
			Results based on 1991 SAIPE Data				
[1.94,8.19)	13	0.5145	1.1261	1.1261	1.1261	0.1139	0.0596
[8.19,11.77)	12	0.6078	-4.0785	0.0001	0.6079	0.1912	0.0478
[11.77,14.09)	13	0.8387	-3.9421	0.0001	0.8387	0.2202	0.0548
[14.09,30.01]	13	0.7292	-8.4406	0.0001	0.7293	0.2115	0.0416
			Results based on 1993 SAIPE Data				
[2.15,7.96)	13	0.8486	0.5305	0.5305	0.8487	0.1114	0.0500
[7.96,12.18)	12	0.8245	-0.4405	0.0001	0.8246	0.2328	0.0661
[12.18,14.67)	13	1.4091	3.6617	3.6617	3.6617	0.2217	0.1066
[14.67,38.22]	13	1.3113	18.4013	18.4013	18.4013	0.1856	0.1926
			Results based on 1997 SAIPE Data				
[2.34,7.86)	13	0.5519	-0.0589	0.0001	0.5520	0.1314	0.0447
[7.86,11.01)	12	0.6097	-1.5812	0.0001	0.6097	0.2439	0.0619
[11.01,13.65)	13	0.7961	-4.7722	0.0001	0.7962	0.2003	0.0528
[13.65,30.81]	13	1.1976	-2.5704	0.0001	1.1976	0.2090	0.0550

to be negative for all groups and areas (not shown in the Table). The GW-adj estimator overcomes the drawback of yielding negative estimates, but it suffers from a severe underestimation problem as their magnitudes are less than the naïve average MSE estimates in many cases across years. The Marker-adj estimates are always positive and greater than naïve average design-based MSE estimates. Large differences between the naïve average design-based MSE estimates and the Marker-adj estimates, for some of the groups (e.g., group 1 in 1991 and group 3 & 4 in 1993), are probably indicative of large biases arising from the violation of the synthetic assumption.

In the last two columns of Table 1, we evaluate direct and synthetic estimators of $\bar{Y}_i$ based on the estimated coefficients of variation (CV). CV direct is square root of average $\hat{V}_i^D$ divided by average of $\hat{Y}_i^D$, average being taken over the small areas within a group. To evaluate synthetic estimator, we have considered only Marker-adj average design-based MSE estimator as this is the only average design-based MSE estimator that produces positive estimates and at the same time does not suffer from underestimation problem. The CV in the last column is defined as square root of $\hat{M}^{\text{M-adj}}$ divided by average of synthetic estimates, $\hat{Y}_i$. The CV of the synthetic estimator is always less than that of direct estimator. This establishes the superiority of synthetic estimator (having reasonable measure of uncertainty) over direct estimator when we do not have enough sample in some of the small areas.

7. Simulation study

We consider $m = 20$ small areas, each containing $N_i = 500$ units. We generate the finite population based on the following nested error regression model (Battese, Harter, and Fuller 1988):

Level 1: $y_{ij}|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma_e^2),\ j = 1, \ldots, N_i,$

Level 2: $\theta_i|\mu, \sigma_v^2 \overset{iid}{\sim} N(\mu, \sigma_v^2),\ i = 1, \ldots, m.$

To generate the finite population, we assume $\mu = 10$ and consider three different combinations of variance components as $(\sigma_v^2,\ \sigma_e^2) = (1, 50), (10, 50)$ and $(10, 1)$, which correspond to variance component ratio $\lambda \equiv \sigma_v^2/\sigma_e^2 = 0.02, 0.2$ and 10, respectively. Once we generate the finite population, we assume it to be fixed, as is the usual practice in a design-based framework. We treat design strata as small areas. The first scenario $(\sigma_v^2,\ \sigma_e^2) = (1, 50)$ reflects that the between area variability is much smaller compared to the within area variability. Hence, the synthetic assumption might work reasonably well. On the other hand, the second and third choices indicate large differences among small areas. To estimate the small area means $\bar{Y}_i$, we draw a stratified simple random sample of size $n_i = 8$ from each stratum. For the evaluation purpose, we repeat the sampling procedure 10,000 times. For this simulation, we consider $\hat{\bar{Y}}_i^D = \bar{y}_i$, the sample mean for area i; $\hat{\bar{Y}}_i = \bar{y}_w = \sum_{i=1}^m \frac{N_i}{N}\bar{y}_i$; $\hat{V}_i^D = \left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}$, where s_i^2 is the sample variance for area i; $\hat{V}_i = \sum_{i=1}^m \frac{N_i^2}{N^2}\hat{V}_i^D$. Synthetic estimators and their variances are constant across small areas.

Table 2: Evaluation of average mean squared error (MSE) estimates of synthetic estimator for small area means

	Naïve	GW	GW-adj	Marker-sim-adj	Marker-adj
Scenario 1: $\sigma_v^2 = 1, \sigma_e^2 = 50, \lambda = 0.02$					
Bias	-0.793	-0.662	0.348	0.190	0.548
Variance	0.001	5.376	2.099	2.357	1.798
MSE	0.630	5.814	2.220	2.393	2.098
Scenario 2: $\sigma_v^2 = 10, \sigma_e^2 = 50, \lambda = 0.2$					
Bias	-6.876	-0.638	-0.606	-0.614	-0.593
Variance	0.001	12.810	12.385	12.480	12.230
MSE	47.285	13.217	12.753	12.857	12.581
Scenario 3: $\sigma_v^2 = 10, \sigma_e^2 = 1, \lambda = 10$					
Bias	-6.636	-0.009	-0.009	-0.009	-0.009
Variance	0.000	0.164	0.164	0.164	0.164
MSE	44.033	0.164	0.164	0.164	0.164

Based on our simulation, we compare the performances of five different average MSE estimators of the synthetic estimator of small area means $\bar{Y}_i$; $i = 1, \ldots, m$. Four of them have already been discussed in Section 6: naïve ($\hat{M}^{\text{Naïve}}$, Eq. 10), GW ($\hat{M}^{\text{GW}}$, Eq. 11), GW-adj ($\hat{M}^{\text{GW-adj}}$, Eq. 14), and Marker-adj ($\hat{M}^{\text{M-adj}}$, Eq. 17). Marker-sim-adj is another variation of Marker's average design-based MSE estimator, defined based on Eq. 15 with $\hat{K}^{\text{M-sim-adj}} = \hat{K}$, if $\hat{K}$ is positive; and 0 otherwise. Since Marker-adj outperforms Marker-sim-adj in terms of lower simulated MSE (Table 2, all three scenarios), we do not include Marker-sim-adj in the boxplots in order to have less congested graphs.

In Figure 1, we display boxplots of average MSE estimates obtained after applying four different methods under scenario 1. As mentioned earlier, under this scenario the synthetic assumption might work reasonably well, which is evident from Figure 1 and Table 2 (scenario 1). The naïve average MSE estimates of the synthetic estimator do not deviate much from the true average MSE (dotted line in Figure 1), but it does underestimate the true average MSE consistently. Even under scenario 1, the naïve average MSE estimator has higher bias than that of any other estimators (Table 2, scenario 1). But the bias is only slightly higher and the variability is much less than for other estimators and consequently in terms of MSE, the naïve average MSE estimator performs better than the alternatives.

The GW method yields a large percentage (36.1%) of negative estimates, as is evident from

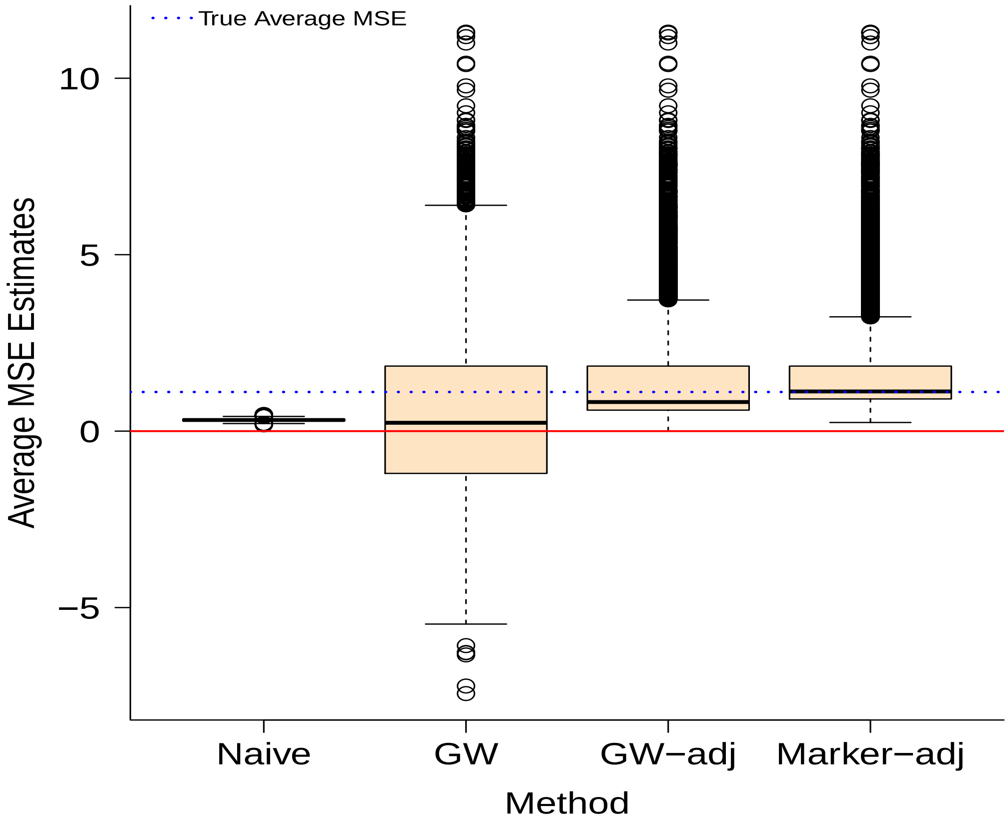

Figure 1: Boxplots of average mean squared error (MSE) estimates of synthetic estimator for scenario 1 ($\sigma_v^2 = 1,\ \sigma_e^2 = 50$). The horizontal blue dotted line represents true average MSE.

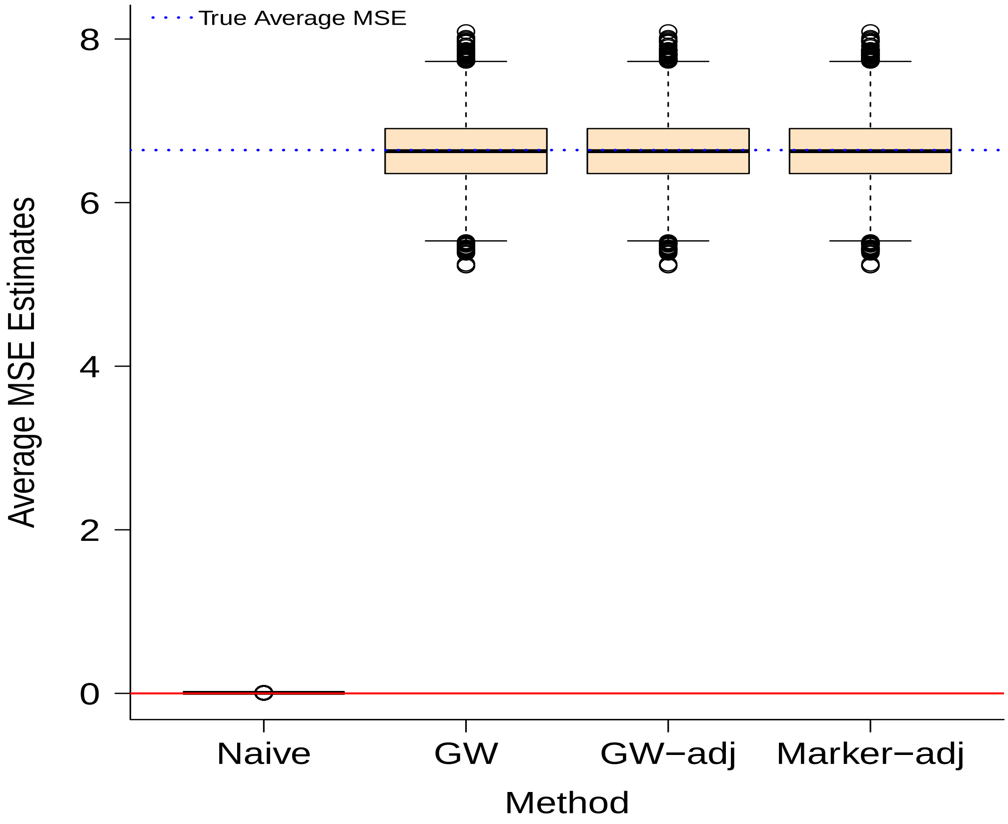

Figure 2: Boxplots of different average mean squared error (MSE) estimates of synthetic estimator for scenario 3 ($\sigma_v^2 = 10,\ \sigma_e^2 = 1$). The horizontal blue dotted line represents true average MSE.

the boxplot (Figure 1). The estimates of squared bias term $\hat{K}$ (Eq. 8) are also negative for similar percentage of samples. The GW average MSE estimator has much more uncertainty than any other average MSE estimators. This makes the GW estimator worst in terms of MSE measure. Both GW-adj and the Marker-adj average MSE estimators always produce positive estimates, satisfying a highly desirable criteria of MSE estimator. The Marker-adj average MSE estimator has slightly lower mean squared error than that of the GW-adj and the former is always larger than the naïve average MSE estimator, unlike the GW-adj average MSE estimator.

In Figure 2, we display boxplots of average MSE estimates obtained after applying four different methods under scenario 3 ($\sigma_v^2 = 10,\ \sigma_e^2 = 1$). Under this scenario, synthetic assumption is more likely to be violated as the between-area variability is much larger than within-area variability. Consequently, the naïve average MSE estimator severely underestimates the true average MSE of the synthetic estimator as it ignores the bias component. The mean squared error of this estimator is much larger than that of other three estimators (Table 2, scenario 3). The GW estimator does not produce any negative estimates, neither does the squared bias term $\hat{K}$ (Eq. 8). All the three other average MSE estimators (GW, adjusted GW and adjusted Marker) perform quite similarly in terms of bias, variance and MSE (Table 2, scenario 3). In essence, under scenario 3, all the three estimators become identical which is not the case under scenario 2 ($\sigma_v^2 = 10,\ \sigma_e^2 = 50$). Under scenario 2, in a very small percentage of samples (less than 1%), $\hat{M}^{\mathrm{GW}}$ and $\hat{K}$ produce negative estimates. Marker-adj AMSE estimator performs better than the alternatives as it has the lowest simulated MSE (Table 2, scenario 2).

8. Discussion

In the absence of enough sample size within small area, synthetic estimators are commonly used by various government organizations while producing official statistics. Although a number of synthetic estimators have been suggested in the small area estimation literature, their evaluation received relatively less attention. The focus of the paper is design-based evaluation of synthetic estimators. Area-specific MSE estimation of synthetic estimator can be problematic because of small sample size within area. Hence, we revisited the available method of using average MSE (average being taken over similar small areas) as the measure of uncertainty, to overcome the problem of unreliable area-specific MSE estimates. Innovation achieved in the paper is to ensure that the proposed average MSE estimators always produce positive estimates and do not suffer from the underestimation problem.

Based on our simulation results under three different scenarios, which cover a broad spectrum of characteristics of small areas under a finite population set up, we can conclude that the Marker-adj average MSE estimator has the potential to be used as a measure of evaluation for synthetic estimators. It always produces positive estimates, accounts for the bias usually present in a synthetic estimator and produces estimates which are always higher than the corresponding naïve average MSE estimates. This average MSE estimator is useful for comparing different direct and synthetic small area estimators since the same criterion is used for all estimators. For the computation of average design-based MSE, the groups of small areas can be formed in different meaningful ways so that the information lost due to grouping can be minimized. Once the most reasonable synthetic estimator is chosen based on the average design-based MSE, one may use the underlying model that motivates the synthetic estimator for area-specific inferential purposes. Extension of the methodology to cover other indirect estimators such as the traditional composite estimators (Rao and Molina 2015, section 4) and empirical/hierarchical Bayes estimators is currently under investigation.

Acknowledgements

The authors thank William R. Bell, Vincent T. Mule and Patrick J. Cantwell of the U.S.Census

Bureau for making some useful comments related to the U.S. Census CCM program. The first author's research was supported in part by National Science Foundation grant SES-0851001 and a U.S. Census Bureau contract YA132309CN0057.

References

Battese GE, Harter RM, Fuller WA (1988). "An Error-components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association*, **83**(401), 28–36.

Chambers R, Chandra H, Tzavidis N (2011). "On Bias-Robust Mean Squared Error Estimation for Pseudo-Linear Small Area Estimators." *Survey Methodology*, **37**(2), 153–170.

Chambers R, Tzavidis N (2006). "M-Quantile Models for Small Area Estimation." *Biometrika*, **93**(2), 255–268.

Chambers RL, Feeney GA (1977). "Log Linear Models for Small Area Estimation." *Paper presented at the Joint Conference of the CSIRO Division of Mathematics and Statistics and the Australian Region of the Biometrics Society, Newcastle, Australia*, pp. 29 August–2 September. Biometrics Abstract No. 2655.

Citro C, Kalton G (eds.) (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond.* Panel on Estimates of Poverty for Small Geographic Area, Committee on National Statistics, Washington, DC: National Academy Press.

Elbers C, Lanjouw JO, Lanjouw P (2003). "Micro–Level Estimation of Poverty and Inequality." *Econometrica*, **71**(1), 355–364. ISSN 1468-0262.

Ericksen EP (1974). "A Regression Method for Estimating Population Changes of Local Areas." *Journal of the American Statistical Association*, **69**(348), 867–875.

Fabrizi E, Salvati N, Pratesi M, Tzavidis N (2014). "Outlier Robust Model-Assisted Small Area Estimation." *Biometrical Journal*, **56**(1), 157–175.

Fay RE, Herriot RA (1979). "Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association*, **74**(366, part 1), 269–277. ISSN 0003-1291.

Gonzalez ME, Hoza C (1978). "Small-area Estimation with Application to Unemployment and Housing Estimates." *Journal of the American Statistical Association*, **73**(361), 7–15. ISSN 0162-1459.

Gonzalez ME, Waksberg J (1973). "Estimation of the Error of Synthetic Estimates." In *first meeting of the international Association of Survey Statisticians, Vienna, Austria*, volume 18, p. 25.

Hansen MH, Hurwitz WN, Madow WG (1953). *Sample Survey Methods and Theory, vol 1.* Wiley-Interscience.

Jiang J, Lahiri P (2006). "Mixed Model Prediction and Small Area Estimation." *Test*, **15**(1), 1–96. ISSN 1133-0686.

Lahiri P, Rao JNK (1995). "Robust Estimation of Mean Squared Error of Small Area Estimators." *Journal of the American Statistical Association*, **90**(430), 758–766. ISSN 0162-1459.

Marker DA (1995). *Small Area Estimation: A Bayesian Perspective.* Ph.D. thesis, University of Michigan.

Marker DA (1999). "Organization of Small Area Estimators Using a Generalized Linear Regression Framework." *Journal of Official Statistics*, **15**, 1–24.

Molina I, Rao JNK (2010). "Small Area Estimation of Poverty Indicators." *Canadian Journal of Statistics*, **38**(3), 369–385.

Mule VT Jr (2010). "U.S. Census Coverage Measurement Survey Plans." In *American Statistical Association, Proceedings of the Survey Research Methods Section.*

NCHS (1968). "Synthetic State Estimates of Disability." *PHS Publication, National Center for Health Statistics*, **1759**.

Nicholls A (1977). "A Regression Approach to Small Area Estimation." *Australian Bureau of Statistics, Canberra.*

Otto MC, Bell WR (1995). "Sampling Error Modelling of Poverty and Income Statistics for States." In *American Statistical Association, Proceedings of the Section on Government Statistics*, pp. 160–165.

Pfeffermann D (2013). "New Important Developments in Small Area Estimation." *Statistical Science*, **28**(1), 40–68.

Purcell NJ, Kish L (1980). "Postcensal Estimates for Local Areas (or Domains)." *International Statistical Review/Revue Internationale de Statistique*, pp. 3–18.

Rao JNK (2003). *Small Area Estimation.* Wiley Series in Survey Methodology. Wiley-Interscience [John Wiley & Sons]. ISBN 0-471-41374-7.

Rao JNK, Molina I (2015). *Small Area Estimation.* John Wiley & Sons.

Schirm AL, Zaslavsky AM (1997). "Reweighting Households to Develop Microsimulation Estimates for States." In *American Statistical Association, Proceedings of the Survey Research Methods Section.*

Stasny E, Goel P, Rumsey D (1991). "County estimates of wheat production." *Survey Methodology*, **17**(2), 211–225.

Yoshimori M, Lahiri P (2014). "A New Adjusted Maximum Likelihood Method for the Fay–Herriot Small Area Model." *Journal of Multivariate Analysis*, **124**, 281–294.

Affiliation:

Partha Lahiri
Joint Program in Survey Methodology
University of Maryland
College Park, MD 20742, USA
E-mail: plahiri@survey.umd.edu
URL: https://jpsm.umd.edu/

Santanu Pramanik
National Council of Applied Economic Research (NCAER)
11 Indraprastha Estate
New Delhi-110 002, INDIA
E-mail: spramanik@ncaer.org
URL: http://www.ncaer.org/

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 48 *Submitted:* 2018-01-31
July 2019 *Accepted:* 2018-06-23