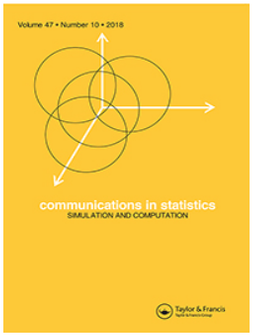

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Small area estimation combining data from two surveys

Sadikul Islam & Hukum Chandra

To cite this article: Sadikul Islam & Hukum Chandra (2019): Small area estimation combining data from two surveys, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1588308

To link to this article: https://doi.org/10.1080/03610918.2019.1588308

Published online: 03 Apr 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1588308

Small area estimation combining data from two surveys

Sadikul Islam and Hukum Chandra

ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

ABSTRACT

Many often two surveys conducted independently with same or different objectives, may have some auxiliary variables in common. The first survey, which has small sample size, collects both variable of interest as well as auxiliary variables. The second survey, relatively larger in sample size, has only some auxiliary variables in common to the first survey. A small area predictor is proposed by combining data from these two surveys. Empirical results show that the proposed small area predictor can lead to efficiency gains when two surveys are combined.

ARTICLE HISTORY
Received 26 April 2018
Accepted 24 February 2019

KEYWORDS
Non-nested surveys; small domain inference; borrowing strength; empirical predictor; mean squared error

AMS SUBJECT CLASSIFICATION
62D05

1. Introduction

In recent years, demands of small area statistics have increased tremendously. Small areas (or small domains) are subsets of the population with small sample sizes, so standard survey estimation methods (e.g. design-based direct estimators) for these areas, which only use information from the small area samples, are unreliable. Obviously, the problem can be overcome by increasing survey sample size, but this solution is usually not pursued because it consumes time and budget resources. As a result, small area estimation (SAE) methods have received significant attention due to their usefulness in producing robust and reliable small area estimates. In this context model-based SAE method that 'borrow strength' via statistical models are often used to produce reliable small area estimates. In particular, SAE methods use statistical models to predict estimates of interest for all the small areas, and a reliable estimate for one small area is obtained by 'borrowing strength' from sample survey data that are collected in other small areas, see (Rao and Molina 2015). SAE models produce good estimates, provided that good auxiliary variables are available and the model is correctly specified, see Pfeffermann (2013).

Often different agencies, departments and organizations conduct surveys from the same population independently for different or same purposes. If two surveys conducted independently have some variables in common, then it seems to be attractive to use data from both surveys for efficient survey estimation. The problem of combining data from two independent surveys to estimate totals at the population and large domain

CONTACT Hukum Chandra hchandra12@gmail.com ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi 110012, India.

levels are discussed in Zieschang (1990), Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004) and Wu (2004). These researchers considered that both the survey have common study variable as well as auxiliary variables to estimate totals at the population and large domain levels. Kim and Rao (2012) considered a design-based approach to combining information from two independent surveys with some common auxiliary variables. We consider a situation where the first survey, which has small sample size, collects both variable of interest as well as auxiliary variables whereas the second survey, relatively larger in sample size, has only some auxiliary variables in common. The two samples may be selected from possibly different sampling frames and also different sampling design. We use the term common auxiliary variables for those auxiliary variables observed in both surveys and their population totals are same. It is assumed that the observed auxiliary variables are comparable in the two surveys. The survey estimation based single survey with small sample size can lead to unstable estimates. In numerous practical situations, combining data from such type of surveys can be advantageous. If estimates are required in a short time period (i.e. quick estimates) and conducting a survey with large sample size is either difficult or not possible. For example, agricultural (e.g. crop yield estimation) and health surveys where collecting or recording data on variable of interest is either costly or time consuming or both, see Schenker and Raghunathan (2007).

Most of the research of combining sample survey data focuses on population and large domain estimation. The scarcity of data and inadequacy sample size are major problem for small areas as compared to the population level estimation. The problem increases further if the survey has a number of areas with no sample units (i.e. non-sample areas). Combining data from two independent surveys can be more attractive and advantageous for producing reliable small area estimates. The potential for efficient SAE by combining comparable information from multiple surveys has been recognized long back, see for example, Marker (2001) and Rao (2003, 23). Merkouris (2010), Lohr and Rao (2006), Elliott and Davis (2005) and Moriarity and Scheuren (2001) discussed the SAE by combining information from multiple surveys. Lohr and Prasad (2003) used multivariate models to combine information from several surveys. Ybarra and Lohr (2008) considered the SAE problem when the area-level auxiliary information has measurement errors. Manzi et al. (2011) used Bayesian hierarchical models to combine information from multiple surveys for SAE. Kim, Park, and Kim (2015) considered an area-level model approach for combining information from several sources in the context of SAE. Our approach extends the work of Maples (2017), which in turn extended the idea of Kim and Rao (2012) to combine data from two independent surveys for small area estimates. Maples considered SAE of proportions from binary variable under a generalized linear mixed model with logit link function (i.e. logistic linear mixed model). In contrast, we focus on estimation of small area means under a linear mixed model by combining information from two independent surveys.

In this article, we elaborate SAE method by combining data from two non-nested surveys when both surveys have some auxiliary variables in common. We adopt frequentist approach to SAE under a unit linear mixed model to combine information from two surveys. The values of auxiliary variables from both surveys are available for the sample units, but not accessible for the non-sample units. In addition, the aggregate values of auxiliary variables are known at population level, but area-specific population aggregates are not

available. We assume that a linear mixed model fitted to data from the first survey, with small sample size, is also a good working model for data from the second survey, which is larger in sample size. With this underlying assumption, we generate a synthetic data of proxy values for the unobserved study variable in second survey and then use the proxy data together with the associated survey weights, of second survey to define the empirical predictor of the small area means (Chandra, Sud, and Gharde 2015).

The rest of the article is organized as follows. In Sec. 2, we first introduce a linear mixed model and then describe the different small area predictors under this model. Mean squared error (MSE) estimation of the proposed small area predictor is developed in Sec. 3. The empirical performances of the different predictors are compared in Sec. 4, using both model-based and design-based simulations, with design-based simulations based on real data. Finally, Sec. 5 is devoted to concluding remarks.

2. Model and estimation of the small area mean

We assume that a finite population U containing N units can be partitioned into D non-overlapping domains $U_i(i = 1, .., D)$ such that $\cup_{i=1}^{D} U_i = U$. Following standard practice, we refer to these domains as small areas or just areas. We further assume that there is a known number N_i of population units in the area i such that $N = \sum_{i=1}^{D} N_i$. Let y_{ij} denote the value of the variable of interest y for unit j in area i. The area-specific mean of y for small area i is $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$. Let s denotes a sample of n units drawn from U by some specified sampling design, and assume that values of y are available for each of these n sample units. The non-sample component of U, containing $N-n$ units, is denoted by r. In what follows, we use a subscript of i to denote quantities specific to area i $(i = 1, ..., D)$. For example, s_i and r_i denote the n_i sample and N_i-n_i non-sample units respectively for area i. Let us assume that two surveys are conducted independently in the same population U, i.e. these surveys are non-nested. The first and second survey are denoted by $S_{(1)}$ and $S_{(2)}$ respectively. The samples of size $n_{(1)}$ and $n_{(2)}$ units from $S_{(1)}$ and $S_{(2)}$ are denoted by $s_{(1)}$ and $s_{(2)}$ respectively. The area specific sample sizes for $s_{(1)}$ and $s_{(2)}$ in area i are denoted by $n_{(1)i}$ and $n_{(2)i}$ respectively such that $n_{(1)} = \sum_{i=1}^{D} n_{(1)i}$ and $n_{(2)} = \sum_{i=1}^{D} n_{(2)i}$. It is assumed that sample size of the first survey, $S_{(1)}$ is much smaller than the second survey, $S_{(2)}$(i.e. $n_{(1)} \ll n_{(2)}$). Further, it is assumed that the smaller survey $S_{(1)}$ has collected both variable of interest y as well as set of auxiliary variables $\mathbf{x}$. The larger survey $S_{(2)}$ has not collected data on the variable of interest y but it has collected a set of auxiliary variables, common to the first survey. A subscript of (k) is used to denote the quantities associated with $k^{th}(k = 1, 2)$ survey. The sample and non-sample part of U, with respect to $S_{(k)}$, are denoted by $s_{(k)}$ and $r_{(k)}$ such that $U = s_{(k)} \cup r_{(k)}; k = 1, 2$. The area-specific $n_{(k)i}$ sample and $N_i-n_{(k)i}$ non-sample units, with respect to the sample $S_{(k)}$, are denoted by $s_{(k)i}$ and $r_{(k)i}$ respectively for area i. With this notation, the conventional design-based direct estimator (denoted by DIR) of area i mean, $\bar{Y}_i$ using data from the first survey, $S_{(1)}$ is defined as

$$\hat{\bar{Y}}_i^{DIR} = \sum_{j \in s_{(1)i}} w_{(1)ij}^{d} y_{ij}. \tag{1}$$

Here, $w^d_{(1)ij} = w^{*d}_{(1)ij} / \sum_{j \in s_{(1)i}} w^{*d}_{(1)ij}$ is a normalized survey (or sample) weight of $S_{(1)}$ for unit j in area i with $\sum_{j \in s_{(1)i}} w^d_{(1)ij} = 1$ and $w^{*d}_{(1)ij}$ is survey weight of $S_{(1)}$ for unit j in area i. The design-based variance of the direct estimator $\hat{\bar{Y}}^{DIR}_{(1)i}$ can be approximated by

$$\text{var}\left(\hat{\bar{Y}}^{DIR}_{i}\right) \approx \sum\nolimits_{j \in s_{(1)i}} w^d_{(1)ij}\left(w^d_{(1)ij} - 1\right)\left(y_{ij} - \hat{\bar{Y}}^{DIR}_{(1)i}\right)^2.$$

The expression for design-based variance estimator of the direct estimator is obtained from Särndal, Swensson, and Wretman (1992, 43, 185 and 391), with the simplifications $w^{*d}_{(1)ij} = 1/\pi_{(1)ij}$, $\pi_{(1)ij,ij} = \pi_{(1)ij}$ and $\pi_{(1)ij,ik} = \pi_{(1)ij}\pi_{(1)ik}, j \neq k$, where $\pi_{(1)ij}$ is the first order inclusion probability of unit j in area i in $S_{(1)}$ and $\pi_{(1)ij,ik}$ is the second order inclusion probability of units j and k in area i in $S_{(1)}$. Under simple random sampling, $w^{*d}_{(1)ij} = N_i n^{-1}_{(1)i}, w^d_{(1)ij} = n^{-1}_{(1)i}$ and the direct estimator (1) is area-specific sample mean, $\hat{\bar{Y}}^{DIR}_{i} = \bar{y}_{s_{(1)i}} = n^{-1}_{(1)i} \sum_{j \in s_{(1)i}} y_{ij}$. The area-specific direct estimator does not depend on an assumed model for its validity (Cochran 1977). The direct estimator (1) is unbiased but it is based on area-specific sample data. Unfortunately, the direct estimator becomes unstable when area specific sample size is small. Therefore, direct estimates are discouraged for the SAE. Further, due to high sampling variability, the confidence interval for these estimates are also very large. Furthermore, for small areas with no sample data, direct estimates cannot be used. In this context model-based SAE methods that 'borrow strength' via statistical models, can be used to produce reliable small area estimates (Rao 2003). These methods typically involve the use of indirect estimators based on suitable models. The SAE methods make use of explicit linking models based on random area-specific effects that take into account between areas variation beyond that is explained by auxiliary variables included in the model. Model-based small area estimators derived from unit level linear mixed models are widely used in SAE. This article also focuses around the model-based approach to estimation of the small area mean under the unit linear mixed model.

2.1. Estimation under a linear mixed model

Let $\mathbf{x}_{ij}$ denote the vector of values of p unit level auxiliary variables for unit j that are assumed to be predictive of y_{ij}. We assume that $\mathbf{x}_{ij}$ contains an intercept term as its first component. For making small area inference, we consider a unit level linear mixed model, in particular random intercepts model of form

$$y_{ij} = \mathbf{x}^T_{ij}\boldsymbol{\beta} + u_i + e_{ij}, j = 1, ..., N_i; i = 1, .., D, \tag{2}$$

where $\boldsymbol{\beta}$ is a p vector of regression coefficients (the fixed effects in the model), u_i denotes area-specific random effect and e_{ij} is an individual random effect. It is standard practice to model the random effects as Gaussian, and so we further assume that these effects are mutually independent between individuals and between areas, with $u_i \sim N(0, \sigma^2_u)$ and $e_{ij} \sim N(0, \sigma^2_e)$, see for example Battese, Harter, and Fuller (1988). It follows that $E(y_{ij}) = \mathbf{x}^T_{ij}\boldsymbol{\beta}$ and $Cov(y_{ij}, y_{ik}) = \sigma^2_u + \sigma^2_e I(j = k)$, where $I(A)$ is the indicator

function for the event A. The vector of parameters $\boldsymbol{\psi} = (\sigma_u^2, \sigma_e^2)^T$ are typically referred to as the variance components of (2). Throughout this paper, we assume that the sampling method is non-informative given the set of auxiliary variables, so the working model (2) holds for both sample and non-sample population units. The model (2) is fitted using sample data, $s_{(1)}$ and $\boldsymbol{\psi}$ is estimated using maximum likelihood (ML) or restricted ML (REML) estimation methods (Harville 1977). We use a 'hat' to denote an estimated quality. Given the estimated values $\hat{\boldsymbol{\psi}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$ of the variance components, we can obtain the estimate of fixed effects parameter $\hat{\boldsymbol{\beta}}$ and prediction of the random effects parameter $\hat{u}_i$. Under the model (2), using the estimated fixed and random effects, the empirical best linear unbiased estimator (EBLUP) of the mean of y in area i is $\hat{\bar{Y}}_i^{EBLUP} = N_i^{-1}[\sum_{j\in s_{(1)i}} y_{ij} + \sum_{j\in r_{(1)i}} (\mathbf{x}_{(1)ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i)]$, see Rao and Molina (2015). The EBLUP is widely used for SAE and under (2) it is proven to be efficient. However, the EBLUP requires the availability of auxiliary variables for non-sample units or at least area-specific population aggregate values of auxiliary variables. In many practical applications, area-specific population aggregate values of auxiliary variables (e.g. area-specific population means, $\bar{\mathbf{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_{ij};\ i = 1, \ldots, D$) are not known. As a consequence, it may not be possible to use the EBLUP estimator. As noted earlier in Sec. 1, the values of auxiliary variables are available for the sample units, but not accessible for the non-sample units. Also, the values of auxiliary variables are not available for area-specific population aggregate. This is most prevailing situation in many countries where Censuses are not regular or Censuses are regular but unit level auxiliary variables are not accessible. Sometimes, the area-specific population aggregates of auxiliary variables obtained from Census or administrative sources may not be consistent in definition (and also may not be coherence in time) with the auxiliary variables available in the sample. We consider two cases of availability of auxiliary variables and illustrate suitable small area estimators combining information from two surveys.

Case-I. Values of auxiliary variables are available for the sample units only. The working model (2) is fitted to sample data $s_{(1)}$ of the first survey $S_{(1)}$ and the estimate of fixed effects parameter and prediction of the random effects parameter are obtained. A plug-in empirical predictor for small area mean in area i denoted by EP1 (Chandra, Sud, and Gharde 2015) is defined as

$$\hat{\bar{Y}}_i^{EP1} = \sum_{j\in s_{(1)i}} w_{(1)ij}^d \hat{y}_{ij} = \hat{\bar{\mathbf{x}}}_{(1)i}^T \hat{\boldsymbol{\beta}} + \hat{u}_i, \tag{3}$$

where $\hat{y}_{ij} = \mathbf{x}_{(1)ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i$, $\hat{u}_i = \hat{\gamma}_i (\bar{y}_{s_{(1)i}} - \bar{x}_{s_{(1)i}}^T \hat{\boldsymbol{\beta}})$, $\hat{\gamma}_i = \hat{\sigma}_u^2 (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_{(1)i})^{-1}$, $\bar{\mathbf{x}}_{s_{(1)i}} = n_{(1)i}^{-1} \sum_{j=1}^{n_{(1)i}} \mathbf{x}_{(1)ij}$ and $\hat{\bar{\mathbf{x}}}_{(1)i} = \sum_{j=1}^{n_{(1)i}} w_{(1)ij}^d \mathbf{x}_{(1)ij}$ is design-based direct estimate of $\bar{\mathbf{x}}_i$ with $E_d(\hat{\bar{\mathbf{x}}}_{(1)i}) = \bar{\mathbf{x}}_i$. Here $E_d(.)$ denotes the expectation under a sampling design (d). Both the DIR estimator (1) and the EP1 predictor (3) are based on the first survey $S_{(1)}$ data alone, which is relatively very small in sample size. The second survey, which large in sample size, does not collect information on variable of interest y. In the spirit of Kim, Park, and Kim (2015), we combine the data from two surveys to define small area predictor, indirectly based on enhanced overall sample size. We assume that the working model (2) fitted to the data from first survey, is also a working model for the data from

second survey. The fitted model (2) is used to generate the proxy or synthetic values of y corresponding to the values of the auxiliary variables from second survey. In particular, we use parameter estimates from first survey to define the proxy or synthetic values of y for second survey, $\tilde{y}_{ij} = x_{(2)ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i, (j = 1, ..., n_{(2)i}; i = 1, .., D)$. An empirical predictor of small area mean in area i (denoted by EP2) using the auxiliary variables from second survey is obtained as

$$\hat{\bar{Y}}_i^{EP2} = \sum_{j \in s_{(2)i}} w_{(2)ij}^d \tilde{y}_{ij} = \hat{\bar{\mathbf{x}}}_{(2)i}^T \hat{\boldsymbol{\beta}} + \hat{u}_i, \tag{4}$$

where $\hat{\bar{\mathbf{x}}}_{(2)i}^T = \sum_{j \in s_{(2)i}} w_{(2)ij}^d \mathbf{x}_{(2)ij}$ is design-based direct estimate of $\bar{\mathbf{x}}_i$ with $E_d(\hat{\bar{\mathbf{x}}}_{(2)i}^T) = \bar{\mathbf{x}}_i$. Here $w_{(2)ij}^d = w_{(2)ij}^{*d} / \sum_{j \in s_{(2)i}} w_{(2)ij}^{*d}$ is normalized survey (or sample) weight of $S_{(2)}$ for unit j in area i with $\sum_{j \in s_{(1)i}} w_{(2)ij}^d = 1$ and $w_{(2)ij}^{*d}$ is survey weight of $S_{(2)}$ for unit j in area i. In case of simple random sampling, $w_{(2)ij}^d = N_i n_{(2)i}^{-1}$ and $w_{(2)ij}^{*d} = n_{(2)i}^{-1}$. Both EP1 and EP2 are using the synthetic rather actual values of y obtained under the working model (2), so they require a correction. This correction (or bias correction) is obtained from the first sample, $S_{(1)}$ since the values of y are available in $S_{(1)}$. The area-specific values of the bias correction are obtain as $\hat{B}_i = \sum_{j \in s_{(1)i}} w_{(1)ij}^d (y_{ij} - \hat{y}_{ij}); i = 1, .., D$. The area-specific values of bias correction can highly be unstable because area sample sizes of $S_{(1)}$ are very small. As a result, an average bias correction obtained as average of area-specific values of bias correction, $\hat{B} = D^{-1} \sum_{i=1}^D \hat{B}_i$ is computed. The bias-corrected version of EP1 and EP2 estimators are defined as $\hat{\bar{Y}}_{i,BC}^{EP1} = \hat{\bar{Y}}_i^{EP1} + \hat{B}$ and $\hat{\bar{Y}}_{i,BC}^{EP2} = \hat{\bar{Y}}_i^{EP2} + \hat{B}$ respectively.

In small area applications, there can be many areas without sample data, referred as non-sample areas. It is not possible to obtain direct estimates of small area quantities for non-sample areas. Let us assume that D_0 out of D areas are non-sample in the first survey, $S_{(1)}$ whereas these areas are sample areas in second survey, $S_{(2)}$. In this case, we cannot compute the DIR estimates and the EP1 estimates for these D_0 non-sample areas. The synthetic version of EP2 can still be obtained for these non-sample areas using auxiliary variables from the second survey, $S_{(2)}$. The synthetic predictor of mean for area i (denoted by SYN-EP2) is given by

$$\hat{\bar{Y}}_i^{SYN-EP2} = \hat{\bar{\mathbf{x}}}_{(2)i,out}^T \hat{\boldsymbol{\beta}}; i = 1, .., D_0, \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ is estimate of fixed effects parameter obtained by fitting the model (1) to the sample data from first survey, $s_{(1)}$ and $\hat{\bar{\mathbf{x}}}_{(2)i,out} = \sum_{j \in s_{(2)i}} w_{(2)ij}^d \mathbf{x}_{(2)ij}; i = 1, .., D_0$.

Case-II. Values of auxiliary variables are available for the sample units as well as for population aggregate. The area-specific aggregate values of auxiliary variables are not known. In this case, three small area estimators defined above, namely DIR, EP1 and EP2 can still be used. The aggregate values of auxiliary variables (e.g. population mean $\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^D \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ or total $\mathbf{t}_x = \sum_{i=1}^D \sum_{j=1}^{N_i} \mathbf{x}_{ij}$) are obtained from Census or administrative sources. We assume that definition of auxiliary variables, both in survey and Census or administrative sources, is same and consistent. Here, we examine few alternative small area predictors using this extra population level auxiliary information.

Chandra and Chambers (2009) introduced a model-based direct estimator (MBDE) for small area means under the linear mixed model (2). The MBDE of a small area mean improves upon the efficiency of the design-based direct estimator DIR by using sample weights that define the EBLUP for the population total (Royall and Cumberland 1978) under the same linear mixed model that underpins the EBLUP for small area mean. At this end, let $\mathbf{y}_i = (y_{i1}, .., y_{iN_i})^T$, $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iN_i})^T$ be a $N_i \times p$ matrix, $\mathbf{e}_i = (e_{i1}, ..., e_{iN_i})^T$, $\mathbf{z}_i = 1_{N_i}$ and 1_{N_i} is a vector of ones of length N_i. The population level version of the linear mixed model (2) is then expressed as

$$\mathbf{y}_U = \mathbf{X}_U\boldsymbol{\beta} + \mathbf{Z}_U\mathbf{u} + \mathbf{e}_U, \tag{6}$$

where $\mathbf{y}_U = (\mathbf{y}_1^T, ..., \mathbf{y}_D^T)^T$, $\mathbf{X}_U = (\mathbf{x}_1^T, ..., \mathbf{x}_D^T)^T$, $\mathbf{Z}_U = diag(\mathbf{z}_i = 1_{N_i}; 1 \le i \le D)$, $\mathbf{u} = (u_1, ..., u_D)^T$ and $\mathbf{e}_U = (\mathbf{e}_1^T, ..., \mathbf{e}_D^T)^T$. Since different areas are independent, the covariance matrix of $\mathbf{y}_U$ has block diagonal structure given by $\mathbf{V}_U = diag(\mathbf{v}_i; 1 \le i \le D)$ with $\mathbf{v}_i = Var(\mathbf{y}_i) = \sigma_u^2\mathbf{z}_i\mathbf{z}_i^T + \sigma_e^2\mathbf{I}_{N_i}$ and $\mathbf{I}_{N_i}$ is the identity matrix of order N_i. Using the estimated values $\hat{\boldsymbol{\psi}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$ of the variance components, the estimated covariance matrix is given by $\hat{\mathbf{V}}_U = diag(\hat{\mathbf{v}}_i; 1 \le i \le D)$, with $\hat{\mathbf{v}}_i = \hat{\sigma}_u^2\mathbf{z}_i\mathbf{z}_i^T + \hat{\sigma}_e^2\mathbf{I}_{N_i}$. Given a sample $s_{(1)}$ of size $n_{(1)}$ from this population, without loss of generality, we arrange the vector $\mathbf{y}_U$ so that its first $n_{(1)}$ elements correspond to the sample units, and then partition $\mathbf{y}_U$, $\mathbf{X}_U$, $\mathbf{Z}_U$ and $\mathbf{e}_U$ according to sample and non-sample units. We can therefore write (6) as follows:

$$\mathbf{y}_U = \begin{bmatrix} \mathbf{y}_{s_{(1)}} \\ \mathbf{y}_{r_{(1)}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{s_{(1)}} \\ \mathbf{X}_{r_{(1)}} \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_{s(1)} \\ \mathbf{Z}_{r(1)} \end{bmatrix}\mathbf{u} + \begin{bmatrix} \mathbf{e}_{s_{(1)}} \\ \mathbf{e}_{r_{(1)}} \end{bmatrix},$$

with variance matrix given by

$$\mathbf{V}_U = \begin{bmatrix} \mathbf{v}_{s_{(1)}s_{(1)}} & \mathbf{v}_{s_{(1)}r_{(1)}} \\ \mathbf{v}_{r_{(1)}s_{(1)}} & \mathbf{v}_{r_{(1)}r_{(1)}} \end{bmatrix}.$$

Thus, $\mathbf{X}_{s_{(1)}}$ represents the $n_{(1)} \times p$ matrix defined by the $n_{(1)}$ sample values of the auxiliary variable vector, while $\mathbf{v}_{s_{(1)}s_{(1)}} = diag\{\sigma_u^2 1_{n_{(1)i}} 1_{n_{(1)i}}^T + \sigma_e^2\mathbf{I}_{n_{(1)i}}; i = 1, ..., D\}$ is the $n_{(1)} \times n_{(1)}$ matrix of covariances of the response variable among the $n_{(1)}$ sampled units that make up the $n_{(1)} \times 1$ sample vector $\mathbf{y}_{s_{(1)}}$. Similarly, $\mathbf{v}_{s_{(1)}r_{(1)}} = diag\{\sigma_u^2 1_{n_{(1)i}} 1_{N_i - n_{(1)i}}^T; i = 1, ..., D\}$. Here, $\mathbf{I}_{n_{(1)i}}$ is the identity matrix of order $n_{(1)i}$ and $1_{n_{(1)i}}$ $(1_{N_i - n_{(1)i}})$ denotes a vector of ones of size $n_{(1)i}$ $(N - n_{(1)i})$. Under (6), the sample weights that define the EBLUP for the population total of y are

$$\mathbf{w}_{s_{(1)}}^{EBLUP} = \left(w_{(1)j}^{EBLUP}\right) = 1_{s_{(1)}} + \hat{\mathbf{H}}_{s_{(1)}}^T\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_{s_{(1)}}}\right) + \left(\mathbf{I}_{s_{(1)}} - \hat{\mathbf{H}}_{s_{(1)}}^T\mathbf{x}_{s_{(1)}}^T\right)\hat{\mathbf{v}}_{s_{(1)}s_{(1)}}^{-1}\hat{\mathbf{v}}_{s_{(1)}r_{(1)}}1_{r_{(1)}}, \tag{7}$$

where, as usual, a 'hat' denotes substitution of estimated variance components, and $\hat{\mathbf{H}}_{s_{(1)}} = (\mathbf{x}_{s_{(1)}}^T\hat{\mathbf{v}}_{s_{(1)}s_{(1)}}^{-1}\mathbf{x}_{s_{(1)}})^{-1}\mathbf{x}_{s_{(1)}}^T\hat{\mathbf{v}}_{s_{(1)}s_{(1)}}^{-1}$; $\mathbf{t}_x = \sum_{i=1}^D\sum_{j=1}^{N_i}\mathbf{x}_{ij} = N\bar{\mathbf{X}}$ and $\hat{\mathbf{t}}_{x_{s_{(1)}}} = \sum_{i=1}^D\sum_{j=1}^{n_{(1)i}}\mathbf{x}_{(1)ij} = n_{(1)}\bar{\mathbf{x}}_{s_{(1)}}$ are the vectors of population and sample totals of $\mathbf{x}$ respectively, $\mathbf{I}_{s_{(1)}}$ is the identity matrix of order $n_{(1)}$ and $1_{s_{(1)}}$ $(1_{r_{(1)}})$ denotes a vector of ones of size $n_{(1)}$ $(N - n_{(1)})$.

Following Chandra and Chambers (2009, 2011), the MBDE (denoted by MBDE) of area i mean of y is

$$\hat{\bar{Y}}_i^{MBDE} = \sum\nolimits_{j\in s_{(1)i}} \tilde{w}_{(1)ij}^{EBLUP} y_{ij}, \tag{8}$$

where $\tilde{w}_{(1)ij}^{EBLUP} = w_{(1)ij}^{EBLUP} / \sum_{j\in s_{(1)i}} w_{(1)ij}^{EBLUP}$. The MBDE of a small area mean is a weighted average of the sample values from the area, defined by using sample weights derived under a population level linear (6). It is noteworthy that the sample weights used to define the MBDE 'borrow strength' via a model that explicitly allows for small area effects. Therefore, the MBDE is expected to be better than the DIR. Replacing the survey weights in EP1 by the EBLUP sample weights (7), an empirical predictor (denoted by EP3) of area i mean of y based on $s_{(1)}$ is defined as

$$\hat{\bar{Y}}_i^{EP3} = \sum\nolimits_{j\in s_{(1)i}} \tilde{w}_{(1)ij}^{EBLUP} \hat{y}_{ij} = \left(\hat{\bar{\mathbf{x}}}_{(1)i}^{EBLUP}\right)^T \hat{\boldsymbol{\beta}} + \hat{u}_i, \tag{9}$$

where $\hat{\bar{\mathbf{x}}}_{(1)i}^{EBLUP} = \sum_{j\in s_{(1)i}} \tilde{w}_{(1)ij}^{EBLUP} \mathbf{x}_{(1)ij}$ with $E_d(\hat{\bar{\mathbf{x}}}_{(1)i}^{EBLUP}) = \bar{\mathbf{x}}_i$. The EP3 is defined by using EBLUP sample weights (7) which 'borrow strength' via model (6). As the MBDE estimates are efficient than the design-based direct estimates, we expect the improved efficiency of the EP3 as compared to the EP1. Let us consider a similar sample and non-sample decomposition of population units with respect to sample $s_{(2)}$ and assume that the model (6) is also working model for sample $s_{(2)}$. Using the estimated variance component parameters of model (6) based on $s_{(1)}$, the sample weights of $s_{(2)}$ that define the EBLUP for the population total of y are given as

$$\mathbf{w}_{s_{(2)}}^{EBLUP} = \left(w_{(2)j}^{EBLUP}\right) = 1_{s_{(2)}} + \hat{\mathbf{H}}_{s_{(2)}}^T \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_{s_{(2)}}}\right) + \left(\mathbf{I}_{s_{(2)}} - \hat{\mathbf{H}}_{s_{(2)}}^T \mathbf{x}_{s_{(2)}}^T\right) \hat{\mathbf{v}}_{s_{(2)}s_{(2)}}^{-1} \hat{\mathbf{v}}_{s_{(2)}r_{(2)}} 1_{r_{(2)}}. \tag{10}$$

where, $\hat{\mathbf{H}}_{s_{(2)}} = \left(\mathbf{x}_{s_{(2)}}^T \hat{\mathbf{v}}_{s_{(2)}s_{(2)}}^{-1} \mathbf{x}_{s_{(2)}}\right)^{-1} \mathbf{x}_{s_{(2)}}^T \hat{\mathbf{v}}_{s_{(2)}s_{(2)}}^{-1}$, $\hat{\mathbf{t}}_{x_{s_{(2)}}} = \sum_{i=1}^{D} \sum_{j=1}^{n_{(2)i}} \mathbf{x}_{(2)ij} = n_{(2)} \bar{\mathbf{x}}_{s_{(2)}}$, $\mathbf{I}_{s_{(2)}}$ is the identity matrix of order $n_{(2)}$ and $1_{s_{(2)}}$ ($1_{r_{(2)}}$) denotes a vector of ones of size $n_{(2)}$ ($N-n_{(2)}$). Here, $\hat{\mathbf{v}}_{s_{(2)}s_{(2)}} = diag\{\hat{\sigma}_u^2 1_{n_{(2)i}} 1_{n_{(2)i}}^T + \hat{\sigma}_e^2 \mathbf{I}_{n_{(2)i}}; i = 1, ..., D\}$ and $\hat{\mathbf{v}}_{s_{(2)}r_{(2)}} = diag\{\hat{\sigma}_u^2 1_{n_{(2)i}} 1_{N_i - n_{(2)i}}^T; i = 1, ..., D\}$, where $\mathbf{I}_{n_{(2)i}}$ is the identity matrix of order $n_{(2)i}$ and $1_{n_{(2)i}}$ ($1_{N_i - n_{(2)i}}$) denotes a vector of ones of size $n_{(2)i}$ ($N-n_{(2)i}$). The empirical predictor (denoted by EP4) of area i mean based on $S_{(2)}$ and using the EBLUP weight (10) is defined as

$$\hat{\bar{Y}}_i^{EP4} = \sum\nolimits_{j\in s_{(2)i}} \tilde{w}_{(2)ij}^{EBLUP} \tilde{y}_{ij} = \left(\hat{\bar{\mathbf{x}}}_{(2)i}^{EBLUP}\right)^T \hat{\boldsymbol{\beta}} + \hat{u}_i, \tag{11}$$

where $\tilde{w}_{(2)ij}^{EBLUP} = w_{(2)ij}^{EBLUP} / \sum_{j\in s_{(2)i}} w_{(2)ij}^{EBLUP}$ and $\hat{\bar{\mathbf{x}}}_{(2)i}^{EBLUP} = \sum_{j\in s_{(2)i}} \tilde{w}_{(2)ij}^{EBLUP} \mathbf{x}_{(2)ij}$ with $E_d(\hat{\bar{\mathbf{x}}}_{(2)i}^{EBLUP}) = \bar{\mathbf{x}}_i$.

Both EP3 and EP4 are defined using the synthetic values obtained under (6), so a bias will obviously be introduced. In both EP3 and EP4 predictors, the bias corrections are applied. As noticed earlier in case-I, there can be situations where some of the areas are non-sample in the first survey, $S_{(1)}$ but these are sample areas in the second survey, $S_{(2)}$. The small area predictors such as the MBDE given by (8) and the EP3 given by (9) cannot be used for such non-sample areas. However, the synthetic version of EP4

(denoted by SYN-EP4) of population mean for non-sample area i is given by

$$\hat{\bar{Y}}_i^{SYN-EP4} = \left(\hat{\bar{\mathbf{x}}}_{(2)i,out}^{EBLUP}\right)^T \hat{\boldsymbol{\beta}}; i = 1, .., D_0, \tag{12}$$

where $\hat{\bar{\mathbf{x}}}_{(2)i,out}^{EBLUP} = \sum_{j \in s_{(2)i}} \tilde{w}_{(2)ij}^{EBLUP} \mathbf{x}_{(2)ij}; i = 1, .., D_0$ and $\hat{\boldsymbol{\beta}}$ is estimate of fixed effects parameter of the model (2) fitted to $S_{(1)}$.

3. MSE estimation

The MSE is the most common measure of accuracy in SAE. MSE estimation of the EP1 follows along the same lines as reported in Chandra, Sud, and Gharde (2015) and references therein. This Section develops an approach for estimating the MSE of the EP2. Here, we first obtain an approximation of the MSE of EP2 and then define estimate of this MSE. The MSE estimators of EP3 and EP4 are obtained directly from the MSE of EP1 and EP2 respectively, replacing the survey weights by the EBLUP sample weights used in defining these estimators. The MSE estimator of the MBDE is given in Chandra and Chambers (2009). Here, following McCulloch and Searle (2001, 300), a standard result is adopted to define the conditional expectation, $E(.) = E_d\{E_\xi(.|d)\}$ under both design (d) and model (ξ). Further, for known variance components $\boldsymbol{\psi} = (\sigma_u^2, \sigma_e^2)^T$ of the model (2), $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator with $E_\xi(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = 0$ whereas for the predictor of random effects, $E_\xi(\hat{u}_i - u_i) = 0$, see McCulloch and Searle (2001, 168–171). The expectation of prediction error $(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i)$ is then obtained by taking the expectations under both design (d) and model (ξ) as

$$\begin{aligned} E\left(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i\right) &= E_d\left\{E_\xi\left(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i\right)|\mathbf{x}_{s_{(2)}}\right\} \\ &= E_d\left[E_\xi\left\{\hat{\bar{\mathbf{x}}}_{(2)i}^T\hat{\boldsymbol{\beta}} + \hat{u}_i - \left(\bar{\mathbf{x}}_i^T\boldsymbol{\beta} + u_i\right)\right\}|\mathbf{x}_{s_{(2)}}\right] \\ &= E_d\left[\left\{\left(\hat{\bar{\mathbf{x}}}_{(2)i} - \bar{\mathbf{x}}_i\right)^T\boldsymbol{\beta} + E_\xi(\hat{u}_i - u_i)\right\}|\mathbf{x}_{s_{(2)}}\right] = E_d\left\{\left[\left(\hat{\bar{\mathbf{x}}}_{(2)i} - \bar{\mathbf{x}}_i\right)^T\boldsymbol{\beta}\right]|\mathbf{x}_{s_{(2)}}\right\} = 0. \end{aligned}$$

Similarly, under both design (d) and model (ξ), the variance of prediction error $(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i)$ is

$$\begin{aligned} Var(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i)) &= E_d\{Var_\xi(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i))|\mathbf{x}_{s_{(2)}}\} + Var_d\{E_\xi(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i))|\mathbf{x}_{s_{(2)}}\} \\ &= E_d\{Var_\xi(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i)|\mathbf{x}_{s_{(2)}}\} + Var_d\{(\hat{\bar{\mathbf{x}}}_{(2)i} - \bar{\mathbf{x}}_i)^T\boldsymbol{\beta}\} \\ &= E_d\{Var_\xi(\hat{\bar{Y}}_i^{EP2} - \bar{Y}_i)|\mathbf{x}_{s_{(2)}}\} + \boldsymbol{\beta}^T Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})\beta. \end{aligned}$$

Here, $Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})$ is the design based variance of $\hat{\bar{\mathbf{x}}}_{(2)i}^T = \sum_{j \in s_{(2)i}} w_{(2)ij}^d \mathbf{x}_{(2)ij}$. The second term, $\boldsymbol{\beta}^T Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})\beta$ is variability due to use of estimate $\hat{\bar{\mathbf{x}}}_{(2)i}$ of $\bar{\mathbf{x}}_i$ in EP2. Under the working model (2), when model parameters are known, the EP2 is like best linear unbiased estimator (BLUP) type estimator. The term $\boldsymbol{\beta}^T Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})\beta$ disappears when $\bar{\mathbf{x}}_i$ is known. The first term on the right hand side of this prediction error variance is directly followed using results reported in Rao and Molina (2015), Prasad and Rao (1990), Datta and Lahiri (2000), Chandra, Sud, and Gharde (2015) and references therein. Under (2), the MSE estimator of the EP2 is of form

$$\begin{aligned} MSE(\hat{\bar{Y}}_i^{EP2}) &= E_d\{M_{1i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}}) + M_{2i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}}) + M_{3i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}})\} + \boldsymbol{\beta}^T Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})\beta \\ &= M_{1i} + M_{2i} + M_{3i} + M_{4i}, \end{aligned} \tag{13}$$

where,

$$\begin{aligned} M_{1i} &= E_d\{M_{1i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}})\} = \sigma_u^2(1-\gamma_i), \\ M_{2i} &= E_d\{M_{2i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}})\} = (\hat{\bar{\mathbf{x}}}_{(2)i}^T - \gamma_i \bar{\mathbf{x}}_{s_{(2)i}}^T) Var_\xi(\hat{\boldsymbol{\beta}})(\hat{\bar{\mathbf{x}}}_{(2)i}^T - \gamma_i \bar{\mathbf{x}}_{s_{(2)i}}^T)^T, \\ M_{3i} &= E_d\{M_{3i}(\boldsymbol{\psi}|\mathbf{x}_{s_{(2)}})\} = tr\left[\left(\frac{\partial \gamma_i}{\partial \boldsymbol{\psi}}\right)\mathbf{v}_{s_{(1)}s_{(1)i}}^{-1}\left(\frac{\partial \gamma_i}{\partial \boldsymbol{\psi}}\right)^T Var_\xi(\boldsymbol{\psi})\right] \text{and} \\ M_{4i} &= \boldsymbol{\beta}^T Var_d(\hat{\bar{\mathbf{x}}}_{(2)i})\boldsymbol{\beta}. \end{aligned}$$

Here $Var_\xi(\hat{\boldsymbol{\beta}}) = (\mathbf{x}_{s_{(1)}}^T \mathbf{v}_{s_{(1)}s_{(1)}}^{-1} \mathbf{x}_{s_{(1)}})^{-1}$ and $Var_\xi(\hat{\boldsymbol{\psi}})$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\psi}}$. Both $Var_\xi(\hat{\boldsymbol{\beta}})$ and $Var_d(\hat{\boldsymbol{\psi}})$ are defined under the model (2) fitted to sample data from first survey $S_{(1)}$. In MSE (13), the first term M_{1i} is the leading term which accounts for the variability of the BLUP type of estimator when all the parameters are known and is of order $o(1)$. The second term M_{2i} due to estimating the fixed effects $\boldsymbol{\beta}$ for given $\mathbf{x}_{s_{(2)}}$, is of order $o(D^{-1})$ for large D. Finally, the third term M_{3i} is of the same order of magnitude as M_{2i} and hence is also dominated by M_{1i}, see Prasad and Rao (1990) and Rao and Molina (2015). The plug-in estimates of these components of MSE (13) are obtained by substituting the estimated values of the variance components under the model (2). Following the approximations described in Prasad and Rao (1990), Datta and Lahiri (2000) and Chandra, Sud, and Gharde (2015), estimator of the MSE (14) is given by

$$mse(\hat{\bar{Y}}_i^{EP2}) = \hat{M}_{1i} + \hat{M}_{2i} + 2\hat{M}_{3i} - C_i^T(\hat{\boldsymbol{\psi}})\nabla M_{1i}(\hat{\boldsymbol{\psi}}) + \hat{\boldsymbol{\beta}}^{\mathrm{T}} v_d(\hat{\bar{\mathbf{x}}}_{(2)i})\hat{\boldsymbol{\beta}}, \tag{14}$$

where

$$\begin{aligned} \hat{M}_{1i} &= M_{1i}\left(\hat{\boldsymbol{\psi}}|\mathbf{x}_{s_{(2)}}\right) = \hat{\sigma}_u^2(1-\hat{\gamma}_i), \\ \hat{M}_{2i} &= M_{2i}\left(\hat{\boldsymbol{\psi}}|\mathbf{x}_{s_{(2)}}\right) = \left(\hat{\bar{\mathbf{x}}}_{(2)i}^T - \hat{\gamma}_i \bar{\mathbf{x}}_{s_{(2)i}}^T\right) v_\xi(\hat{\boldsymbol{\beta}}) \left(\hat{\bar{\mathbf{x}}}_{(2)i}^T - \hat{\gamma}_i \bar{\mathbf{x}}_{s_{(2)i}}^T\right)^T \\ \text{with } v_\xi(\hat{\boldsymbol{\beta}}) &= (\mathbf{x}_{s_{(1)}}^T \hat{\mathbf{v}}_{s_{(1)}s_{(1)}}^{-1} \mathbf{x}_{s_{(1)}})^{-1}, \\ \hat{M}_{3i} &= M_{3i}\left(\hat{\boldsymbol{\psi}}|\mathbf{x}_{s_{(2)}}\right) = tr\left[\left(\frac{\partial \gamma_i}{\partial \boldsymbol{\psi}}\right)_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \hat{\mathbf{v}}_{s_{(1)}s_{(1)i}}^{-1} \left(\frac{\partial \gamma_i}{\partial \boldsymbol{\psi}}\right)_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}^T v_\xi(\hat{\boldsymbol{\psi}})\right], \text{and} \\ \hat{M}_{4i} &= \hat{\boldsymbol{\beta}}^{\mathrm{T}} v_d(\hat{\bar{\mathbf{x}}}_{(2)i})\hat{\boldsymbol{\beta}}, \text{ with } v_\xi(\hat{\boldsymbol{\psi}}) = I^{-1}(\hat{\boldsymbol{\psi}}). \end{aligned}$$

Here, a 'hat' on a quantity denotes a corresponding plug-in estimator of that quantity. For example, $\hat{M}_{1i}$, $\hat{M}_{2i}$, $\hat{M}_{3i}$ are obtained from M_{1i}, M_{2i}, M_{3i} respectively, replacing $\boldsymbol{\psi}$ by $\hat{\boldsymbol{\psi}}$. The multiplier of "2" in $\hat{M}_{3i}$ term arise because the plug-in estimator of M_{1i} is biased low, with $E_\xi(\hat{M}_{1i}) \cong M_{1i} - M_{3i} + C_i^T(\hat{\boldsymbol{\psi}})\nabla M_{1i}(\hat{\boldsymbol{\psi}})$, see (Rao 2003, 104). Overall, the order of the bias in MSE estimate (14) is $o(D^{-1})$ since $\hat{M}_{1i}$ and $\hat{M}_{3i}$ have biases of order $o(D^{-1})$. The MSE (14) is therefore an approximately model unbiased estimator in the sense that its bias is of order $o(D^{-1})$ and considered as a second order approximation. In MSE (14),

Table 1. Description of different small area estimators.

Case	Estimator	Description
	DIR	Direct estimator (1) with sample weights $w_{(1)ij}$ based on first survey, $S_{(1)}$
I	EP1	Predictor (3) with sample weights $w_{(1)ij}$ based on first survey, $S_{(1)}$
	EP2	Predictor (4) with sample weights $w_{(2)ij}$ based on second survey, $S_{(2)}$
	SYN-EP2	Synthetic predictor (5) with sample weights $w_{(2)ij}$ based on second survey ,$S_{(2)}$
	MBDE	MBDE (9) with sample weights (8) based on first survey, $S_{(1)}$
II	EP3	Predictor (10) with sample weights (8) based on first survey, $S_{(1)}$
	EP4	Predictor (12) with sample weights (11) based on second survey, $S_{(2)}$
	SYN-EP4	Synthetic predictor (13) with sample weights (11) based on second survey, $S_{(2)}$

$$C_i^T(\hat{\boldsymbol{\psi}}) = \frac{1}{2D}\left\{ I^{-1}(\hat{\boldsymbol{\psi}}) \underset{1\le j\le D}{col}\, tr\left[\left(\sum_i x_{s_{(1)i}}^T \mathbf{v}_{s_{(1)i}s_{(1)i}}^{-1} x_{s_{(1)i}} \right)^{-1} \left(\sum_i x_{s_{(1)i}}^T \mathbf{v}_{s_{(1)i}s_{(1)i}}^{(j)} x_{s_{(1)i}} \right) \right] \right\}$$

is the bias in estimating the variance components $\hat{\boldsymbol{\psi}}$, with

$$\mathbf{v}_{s_{(1)i}s_{(1)i}}^{(j)} = \partial \mathbf{v}_{s_{(1)i}}^{-1} / \partial \boldsymbol{\psi}_j = -\mathbf{v}_{s_{(1)}s_{(1)i}}^{-1} \left(\partial \mathbf{v}_{s_{(1)}s_{(1)i}}^{-1} / \partial \boldsymbol{\psi}_l \right) \mathbf{v}_{s_{(1)}s_{(1)i}}^{-1}; j = 1, 2$$

and $I^{-1}(\hat{\boldsymbol{\psi}})$ is the inverse of information matrix $I(\hat{\boldsymbol{\psi}})$, and $\nabla M_{1i}(\hat{\boldsymbol{\psi}})$ is the first order derivative of M_{1i} with respect to $\boldsymbol{\psi}$ at $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}$. The bias expression $C_i^T(\hat{\boldsymbol{\psi}})$ is negligible when REML method is used for estimating the variance components $\boldsymbol{\psi}$, while this bias is of order $o(D^{-1})$ when ML method. So, the term $C_i^T(\hat{\boldsymbol{\psi}})\nabla M_{1i}(\hat{\boldsymbol{\psi}})$ is included in the MSE estimator (14) when ML method is used for estimating $\boldsymbol{\psi}$.

4. Empirical evaluations

In this Section, we use simulation studies to illustrate the empirical performance of the different small area estimators defined in the preceding Sections. The different small area estimators are described in Table 1. We carried out two types of simulation studies. In the first one, the properties of the small area estimators have been assessed by using model-based simulation to generate artificial population and sample data. In this case, at each simulation population data are first generated under the model and a single sample is then taken from this simulated population by stratified simple random sampling without replacement with small area as strata. These data are then used to compare the performances of the different estimators. The second type of simulation study is design-based, with population data derived from a real survey dataset (i.e. a real sample data was used to construct an artificial finite population). The performance of these small area estimators are evaluated in the context of repeated sampling from a real population using realistic sampling methods. In this case, a real survey data is first used to simulate a population, and this fixed population is then repeatedly sampled according to a pre-specified design.

The performance of the different estimators in the simulation studies is evaluated by computing for each small area the average percentage relative bias (RB), the average percentage relative root mean squared error (RRMSE) and the average percentage relative efficiency (RE) as

$$RB(m) = \underset{i}{mean}\left\{ \bar{m}_i^{-1} R^{-1} \sum\nolimits_{r=1}^{R} (\hat{m}_{ir} - m_{ir}) \right\} \times 100,$$

$$RRMSE(m) = \underset{i}{mean}\left\{\sqrt{R^{-1}\sum_{r=1}^{R}\left(\frac{\hat{m}_{ir}-m_{ir}}{m_{ir}}\right)^2}\right\} \times 100, \text{ and}$$

$$RE(m) = \frac{RMSE(Direct\ \ estimator)}{RMSE(Proposed\ \ small\ \ area\ \ estimator)} \times 100, \text{ with}$$

$$RMSE(m) = \underset{i}{mean}\left\{\sqrt{R^{-1}\sum_{r=1}^{R}(\hat{m}_{ir}-m_{ir})^2}\right\} \times 100.$$

Here the subscript i indexes the small areas and the subscript r indexes the R Monte Carlo simulations, with m_{ir} denoting the true area i mean at simulation r, with predicted value $\hat{m}_{ir}$, $\bar{m}_i = R^{-1}\sum_{r=1}^{R} m_{ir}$. Note that in the design-based simulations since the population is a fixed quantity so, $m_{ir} = m_i$. The whole process of calculation of small area estimates is independently replicated R times.

In addition to these we also perform the simulation studies to measure the performance of the MSE estimators of the small area mean predictors. The performance of MSE estimators in the simulation studies is evaluated by computing for each small area the average percentage relative bias (RB) and coverage rate (CR) of nominal 95 percent confidence intervals defined as follows:

$$RB(M) = \underset{i}{mean}\left\{M_i^{-1}R^{-1}\sum_{r=1}^{R}\left(\hat{M}_{ir} - M_i\right)\right\} \times 100 \text{ and}$$

$$CR(M) = \underset{i}{mean}\left\{R^{-1}\sum_{r=1}^{R} I\left(|\hat{m}_{ir} - m_{ir}| \le 1.96\sqrt{\hat{M}_{ir}}\right)\right\}.$$

Here $\hat{M}_{ir}$ denotes the estimate of MSE estimator in area i at simulation r and M_i denotes the actual MSE in area i.

4.1. Model-based simulation study

In the model-based simulations, we set number of areas as $D = 30$. Population size for each area are kept fixed over simulations as $N_i = 500$ and total population size is $N = 15000$. We consider two sets of model-based simulation, namely SIM1 and SIM2. The first model-based simulation (denoted by SIM1) is based on population data generated under the linear mixed model (2). In particular, population values for y are generated under the random intercepts model of form $y_{ij} = 500 + 1.5x_{ij} + u_i + e_{ij}$, where $x_{ij}(i = 1, ...D; j = 1, ..., N_i)$ drawn from a chi squared distribution with 20 degree of freedom. The random area effects u_i and individual effects e_{ij} are independently drawn from $N(0, \sigma_u^2)$ and $N(0, \sigma_e^2 = 94.09)$ distributions respectively. We use two values of area effects variance $\sigma_u^2 = 10.40$ and 23.52 so that intra area correlation are $\rho = \sigma_u^2(\sigma_u^2 + \sigma_e^2)^{-1} = 0.10$ and 0.20 respectively. The model-based simulation, SIM1 for $\rho = 0.10$ and 0.20 are referred as SIM1-A and SIM1-B respectively. Two samples of size $n_{(1)} = 90$ (first survey) and $n_{(2)} = 600$ (second survey) are selected from each simulated

Table 2. Values of percentage relative biases (RB), percentage relative root mean squared errors (RRMSE) and percentage relative efficiencies (RE) of the different estimators in SIM1 of model based simulations. The values are averaged over 30 small areas.

		SIM1-A			SIM1-B		
$n_{(1)i}, n_{(2)i}$	Predictor	RB	RRMSE	RE	RB	RRMSE	RE
3,20	DIR	0.008	1.48	100	0.008	1.48	100
	EP1	0.008	1.20	124	0.008	1.28	116
	EP2	0.008	0.72	205	0.008	0.85	174
	MBDE	0.007	1.47	101	0.008	1.47	100
	EP3	0.007	1.18	125	0.007	1.27	117
	EP4	0.009	0.72	206	0.009	0.85	174
3,50	DIR	0.008	1.48	100	0.008	1.48	100
	EP1	0.008	1.19	124	0.008	1.26	117
	EP2	0.006	0.65	229	0.006	0.78	188
	MBDE	0.006	1.47	100	0.006	1.47	100
	EP3	0.006	1.18	126	0.006	1.26	118
	EP4	0.007	0.64	229	0.007	0.78	188
5, 20	DIR	0.022	1.14	100	0.022	1.14	100
	EP1	0.022	0.96	119	0.022	1.02	112
	EP2	0.014	0.66	173	0.014	0.75	153
	MBDE	0.015	1.13	101	0.015	1.14	101
	EP3	0.015	0.95	121	0.015	1.01	113
	EP4	0.015	0.66	174	0.015	0.74	153
5, 50	DIR	0.000	1.14	100	0.000	1.14	100
	EP1	0.000	0.96	119	0.000	1.02	112
	EP2	−0.005	0.59	192	−0.005	0.69	165
	MBDE	−0.003	1.13	100	−0.003	1.13	100
	EP3	−0.003	0.95	120	−0.003	1.01	113
	EP4	−0.003	0.59	193	−0.003	0.69	166

population, with area sample sizes of first and second survey as $n_{(1)i} = 3$ and $n_{(2)i} = 20$ respectively. Sampling is via stratified random sampling, with the strata defined by the small areas. The sample values of y for the first sample only and the sample values of x for both samples obtained in each simulation are then used to estimate the small area means using the different predictors. The process of generating population and sample data, estimation of parameters and calculation of small area estimates is independently replicated $R = 1000$ times. The SIM1 simulations are further repeated with different combination of sample sizes for first and second samples. In particular, three additional combinations of sample sizes are: $(n_{(1)}, n_{(2)}) = (90, 1500), (150, 600)$ and $(150, 1500)$. The corresponding area-specific sample sizes for the first and second samples are: $(n_{(1)i}, n_{(2)i}) = (3, 50)$, $(5, 20)$ and $(5, 50)$ respectively. In these model-based simulations, we consider two cases of the availability auxiliary information. These are: (i) Case-I, auxiliary variables are available for the sample units only. (ii) Case-II, auxiliary variables are available for the sample units, in addition, the population totals of auxiliary variables are also known. The performance of the different estimators across the different simulations is assessed by computing the average values of their area specific percentage relative bias, percentage relative root mean squared error and percentage relative efficiency. Table 2 presents the average values of the different estimators from SIM1 simulations. Boxplots of area-specific values of relative root mean squared errors of the different estimators generated from SIM1-A simulation for two sample sizes $(n_{(1)i} = 3, n_{(2)i} = 20)$ and $(n_{(1)i} = 5, n_{(2)i} = 20)$ are depicted in Figure 1.

Conditions for the second model-based simulation study (denoted by SIM2) are the same as in the first simulations (SIM1), with the exception that the area level random

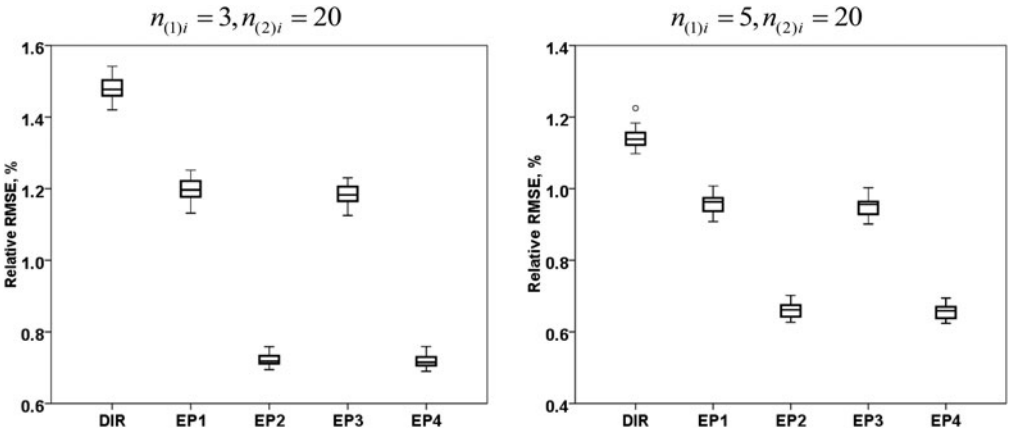

Figure 1. Boxplots of area-specific values of relative RMSE of the different small area predictors in SIM1-A of model based simulations.

effects and the individual level random effects are independently drawn from mean corrected chi-square distributions respectively. In SIM2 set of simulations, we examined the robustness of the different estimators to model misspecification. Here, the random area effects u_i and individual effects e_{ij} are independently drawn from chi squared distribution with d_1 and $d_2 = 5$ degree of freedom respectively. The values of $d_1 = 1$ and 2 are chosen such that intra-area effects are 0.17 and 0.29 respectively. Corresponding to these two values of the intra-area effects 0.17 and 0.29, simulation sets are denoted by SIM2-A and SIM2-B respectively. The SIM2 simulations are replicated with four different combinations of sample sizes as described in SIM1. Table 3 presents the average values of percentage relative biases, percentage relative root mean squared errors and relative efficiencies of the different estimators from the model based simulation SIM2.

The SIM1 simulations are further replicated by considering a situation where few areas in the first survey are non-sample (denoted by SIM3 simulations). In particular, in the first sample, $s_{(1)}$, there are 25 sample and $D_0 = 5$ non-sample areas. In contrast, in the second sample, $s_{(2)}$, all the $D = 30$ areas are sample areas. The whole process of generating the population to calculation of small area estimates are independently replicated $R = 1000$ times. The purpose of SIM3 simulation studies is to examine the performance of the proposed small area estimators namely SYN-EP2 and SYN-EP4 for non-sample areas. The corresponding results from SIM3 simulations are summarized in Table 4.

The performance of proposed MSE estimator is also examined in SIM1 simulations. The average values of empirical true root MSE, estimated root MSE, percentage relative bias and coverage rate of nominal 95% confidence interval from SIM1 simulations are presented in Table 5. These results are averaged over 30 small areas. The bias-corrected version of the different estimators are also evaluated in both SIM1 and SIM2 sets of simulation studies. However, empirical performance of the bias-corrected estimators are almost same as that of the without bias-corrected version. So, the results of different predictors reported in this paper related to without bias-correction.

Table 2 shows the average relative bias, average relative RMSE and average relative efficiency of the different estimators investigated in simulation set SIM1. In Table 2, we

Table 3. Values of percentage relative biases (RB), percentage relative root mean squared errors (RRMSE) and percentage relative efficiencies (RE) of the different estimators in SIM2 of model based simulations. The values are averaged over 30 small areas.

		SIM2-A			SIM2-B		
$n_{(1)i}, n_{(2)i}$	Predictor	RB	RRMSE	RE	RB	RRMSE	RE
3,20	DIR	−0.005	1.08	100	0.005	1.061	100
	EP1	−0.005	1.05	103	0.005	1.053	101
	EP2	−0.003	0.45	240	0.002	0.428	248
	MBDE	−0.004	1.07	101	0.000	1.049	101
	EP3	−0.001	1.04	105	0.001	1.040	102
	EP4	−0.001	0.45	243	0.001	0.423	251
3,50	DIR	0.010	1.09	100	−0.001	1.051	100
	EP1	0.010	1.06	103	−0.001	1.042	101
	EP2	0.000	0.33	327	0.000	0.301	349
	MBDE	0.001	1.08	101	−0.003	1.041	101
	EP3	0.001	1.04	105	−0.003	1.032	102
	EP4	0.000	0.33	331	−0.001	0.297	354
5,20	DIR	0.000	0.841	100	0.005	0.818	100
	EP1	0.000	0.821	102	0.005	0.813	101
	EP2	−0.002	0.442	190	−0.001	0.421	195
	MBDE	−0.002	0.831	101	−0.002	0.807	101
	EP3	−0.002	0.810	104	−0.010	0.802	102
	EP4	−0.002	0.437	193	−0.001	0.415	197
5,50	DIR	0.015	0.843	100	0.006	0.817	100
	EP1	0.015	0.824	102	0.006	0.812	101
	EP2	−0.001	0.312	270	0.000	0.285	286
	MBDE	0.002	0.830	102	−0.001	0.807	101
	EP3	0.002	0.810	104	−0.001	0.801	102
	EP4	0.001	0.309	273	0.000	0.282	290

observe a marginal gain in efficiency (in terms of lower relative root MSE) for the different estimators defined for case-II (MBDE, EP3 and EP4) as compared to their corresponding estimators of case-I (DIR, EP1 and EP2). The estimators defined for case-II uses extra auxiliary information in the form of population aggregates. This clearly illustrates that additional auxiliary information is not providing any significant gain in efficiency. Hereafter, our discussion focus on three estimators defined for case-I, i.e. DIR, EP1 and EP2 estimators. The differences in performance between the various estimators in Table 2 are essentially as one would expect. Relative bias is not really an issue, as the biases of all the estimators are almost negligible and of the same order of magnitude for all set of sample sizes. The proposed small area estimator combining information from two surveys is the most efficient in terms of relative RMSE. Here we see a substantial gain in efficiency (as measured by a lower relative RMSE) when the EP2 is compared to the EP1. The relative efficiency of the EP2 as compared to the EP1 is higher for smaller value of intra area effect (SIM1-A). The relative RMSE of all the estimators increases (or relative performance decreases) slightly when intra-area correlation increases from 0.10 to 0.20. Further, for a given sample size of smaller survey ($n_{(1)}$), as the sample sizes of larger survey ($n_{(2)}$) increases, the relative efficiency of the EP2 increases (or relative RMSE decreases) as compared to the EP1. The boxplot in Figure 1 also confirms the conclusions from Table 2. The results in Table 2 and Figure 1 clearly indicate that combining data from two surveys provide a significant gain in SAE.

Table 3 presents the simulation results of the different estimators from simulation set SIM2. The conclusions from the results in Table 3 are almost same as observed in

Table 4. Values of percentage relative biases (RB), percentage relative root mean squared errors (RRMSE) and percentage relative efficiencies (RE) of the different estimators in SIM3 of model based simulations [$S_{(1)}$: 25 sample + 5 non-sample areas; $S_{(2)}$: all 30 sample areas]. The values are averaged over 25 areas for sample and 5 areas for non-sample results.

			SIM1-A			SIM1-B		
$n_{(1)i}, n_{(2)i}$	Areas	Predictor	RB	RRMSE	RE	RB	RRMSE	RE
3,20	Sample	DIR	0.006	1.47	100	0.006	1.47	100
		EP1	0.006	1.20	123	0.006	1.28	115
		EP2	0.014	0.73	201	0.014	0.86	170
		MBDE	0.008	1.47	100	0.008	1.47	100
		EP3	0.009	1.19	124	0.008	1.27	116
		EP4	0.011	0.73	202	0.011	0.86	171
	Non-sample	SYN-EP2	0.008	0.76	–	0.006	1.02	–
		SYN-EP4	0.004	0.76	–	0.002	1.02	–
3,50	Sample	DIR	0.016	1.47	100	0.016	1.47	100
		EP1	0.016	1.19	123	0.016	1.27	115
		EP2	0.009	0.66	224	0.009	0.80	184
		MBDE	0.009	1.46	100	0.009	1.46	100
		EP3	0.009	1.18	124	−0.005	1.26	116
		EP4	0.010	0.66	224	0.010	0.80	184
	Non-sample	SYN-EP2	0.001	0.70	–	−0.003	0.98	–
		SYN-EP4	0.001	0.70	–	−0.003	0.98	–
5,20	Sample	DIR	0.017	1.14	100	0.017	1.14	100
		EP1	0.017	0.96	118	0.017	1.02	111
		EP2	0.008	0.68	168	0.008	0.76	149
		MBDE	0.007	1.13	101	0.007	1.13	101
		EP3	0.006	0.95	119	0.006	1.01	112
		EP4	0.006	0.67	169	0.006	0.76	150
	Non-sample	SYN-EP2	0.012	0.76	–	0.014	1.03	–
		SYN-EP4	0.010	0.76	–	0.012	1.03	–
5,50	Sample	DIR	−0.005	1.14	100	−0.005	1.14	100
		EP1	−0.005	0.96	119	−0.005	1.02	112
		EP2	−0.006	0.60	191	−0.006	0.69	165
		MBDE	−0.006	1.13	101	−0.006	1.13	101
		EP3	−0.006	1.09	105	−0.006	1.01	113
		EP4	−0.007	0.60	192	−0.006	0.69	165
	Non-sample	SYN-EP2	−0.003	0.68	–	−0.006	0.96	–
		SYN-EP4	−0.004	0.67	–	−0.007	0.96	–

Table 5. Values of empirical true root MSEs (TRMSE), estimated root MSEs (ERMSE), percentage relative biases (RB) and coverage rates (CR) of nominal 95% confidence interval from SIM1 simulations. The values are averaged over 30 small areas.

		SIM1-A				SIM1-B			
$n_{(1)i}, n_{(2)i}$	Predictor	TRMSE	ERMSE	RB	CR	TRMSE	ERMSE	RB	CR
3,20	EP1	40.25	42.25	5.22	0.91	45.82	45.90	0.38	0.91
	EP2	14.62	16.27	11.47	0.96	20.39	19.94	−2.03	0.95
	EP3	39.46	42.84	8.77	0.91	45.11	46.54	3.38	0.91
	EP4	14.47	16.27	12.73	0.96	20.24	19.94	−1.26	0.95
3,50	EP1	39.57	42.47	7.57	0.91	44.94	46.11	2.85	0.91
	EP2	11.71	13.59	16.33	0.96	17.27	17.24	−0.02	0.95
	EP3	38.80	43.09	11.31	0.91	44.27	46.81	5.98	0.91
	EP4	11.66	13.59	16.86	0.96	17.22	17.24	0.29	0.95
5,20	EP1	25.74	26.47	3.08	0.93	29.17	29.47	1.29	0.93
	EP2	12.27	12.74	4.22	0.95	15.73	15.76	0.54	0.95
	EP3	25.22	26.65	5.93	0.93	28.69	29.67	3.64	0.94
	EP4	12.12	12.74	5.49	0.95	15.58	15.76	1.46	0.95
5,50	EP1	25.72	26.36	2.71	0.93	29.13	29.38	1.08	0.93
	EP2	9.82	9.95	1.49	0.95	13.31	12.99	−2.23	0.95
	EP3	25.22	26.53	5.43	0.93	28.68	29.57	3.32	0.93
	EP4	9.74	9.95	2.29	0.95	13.23	12.99	−1.66	0.95

Table 2 and Figure 1. Again, we observe that for the fixed sample size in first survey, increase in area specific sample sizes in second survey, the relative RMSE reduces and relative efficiency increases consistently for the EP2 as compare the EP1. However, in deviation to the SIM1 results in Table 2, the relative performance (in terms of lower relative RMSE) as well as the relative efficiency of all the estimators increases when intra-area correlation increases from 0.10 to 0.20.

The results in Table 4 generated from SIM3 simulations indicate some intuition of performance of the proposed estimator for non-sample areas. In this case, SIM1 simulations are replicated, except that in first survey 5 areas are taken as non-sample areas, i.e. only 25 areas are in sample. In Table 4, the results for sample areas are the average values of 25 sample areas. The relative performances of all the estimators for sample areas in Table 4 are identical to the results in Table 2. The results for non-sample areas generated by the SYN-EP2 (or the SYN-EP4) method are noteworthy. These results are average over 5 non-sample areas. The performances of two synthetic estimators, SYN-EP2 or SYN- EP4 are at par. Both biases and RMSEs of the SYN-EP2 (or the SYN-EP4) estimator for non-sample areas have almost same magnitude as of the EP2 estimator for sample areas. This clearly shows an evidence that the proposed synthetic estimator has potential to generate the reliable estimates for non-sample areas.

We now turn to an examination of performance of the MSE estimation investigated in the simulation. In Table 5, we see that the MSE estimator (14) for the EP2 behaves in exactly the same way as the corresponding MSE estimator for the EP4. The empirical true root MSE of the EP2 is much smaller than the EP1, which shows a gain in efficiency by combining data from two surveys. This happens because the EP2 estimator is based on enhanced effective sample size. The results in Table 5 further shows that the empirical true root MSEs and the estimated root MSEs obtained using the MSE estimator (14) for the EP2 are very close. The relative bias of the MSE estimator decreases with increase in sample sizes. It also decreases when intra-area correlation increases from 0.10 to 0.20. The use of MSE estimates to calculate 'normal theory based' confidence intervals is common practice. We combine the EP2 estimate with the MSE estimate to generate 'normal theory based' confidence intervals as the estimate plus or minus twice its corresponding estimated root MSE. Table 5 shows that the actual coverage rates achieved by these intervals, are generally 95 per cent for the MSE of the EP2. Overall, the proposed MSE estimator tracts the true MSE of the EP2 reasonably well, with a good coverage property.

4.2. Design-based simulation study

Design-based simulations complement model-based simulations for SAE since they allow us to evaluate the performance of SAE methods in the context of a real population and realistic sampling methods. From a finite population perspective we believe that this type of simulation constitutes a more practical and appropriate representation of the SAE problem. The design-based simulations are based on real survey data collected in the 1995–96 Australian Agricultural Grazing Industry Survey (AAGIS) conducted by the Australian Bureau of Agricultural and Resource Economics. In the original sample consists of 759 farms from 12 regions (the small areas of interest),

Table 6. Region-wise sample size ($n_{(2)i}$) for large sample in AAGIS data.

Region	$n_{(2)i}$	Region	$n_{(2)i}$
1	88	7	63
2	44	8	42
3	73	9	87
4	54	10	47
5	58	11	76
6	84	12	43

which make up the wheat-sheep zone for Australian broad acre agriculture. This original sample data is used to generate a population of size $N = 39,562$ farms by re-sampling the original AAGIS sample of size 759 farms with probability proportional to a farm's sample weight, see Chandra et al. (2012). From this (fixed) population, $R = 1000$ independent samples of $n_{(2)} = 759$ farms are selected using stratified random sampling, with regions are strata and with stratum sample allocations the same as in the original AAGIS sample. This sample is treated as the larger samples (second sample, $s_{(2)}$ of size $n_{(2)} = 759$). Table 6 shows the region-specific sample sizes $n_{(2)i}$ of 12 regions or areas.

The sample size for smaller sample, also referred as the first sample, $s_{(1)}$ is chosen as $n_{(1)} = 60$ with area-specific sample sizes, $n_{(1)i} = 5$. Again, $R = 1000$ independently stratified random samples of $n_{(1)} = 60$ farms are selected from same the simulated population (fixed), with regions are strata and with stratum sample size as $n_{(1)i} = 5$. In the simulation studies, two additional sample sizes, $n_{(1)} = 120$ and 240 with area-specific sample sizes as $n_{(1)i} = 10$ and 20 respectively are also considered for smaller sample, $s_{(1)}$. With same procedure, $R = 1000$ independent samples for these two sample sizes are also selected from this (fixed) simulated population. The variable of interest is the total cash costs (TCC), and the target is the average value of TCC in each region. A range of potential explanatory variables are available for building a working small area model. The covariates used in the fixed part of this working model returned an R^2 value of 0.40; they are land area, four identifiers for the five industries (i.e. specialist crop farms, mixed livestock and crop farms, sheep specialists, beef specialists, and mixed sheep and beef farms), number of closing stock-beef, number of closing stock-sheep and quantity of harvested wheat. The results for the design-based simulations using the AAGIS data presented in Table 7 and in Figure 2.

In Table 7, we again notice that the estimators defined under case-II practically offer no gain over the estimators defined under case-I. So, our discussion concentrates around three estimators only, i.e. DIR, EP1 and EP2. Figure 2 also demonstrates the boxplots of region-specific values for DIR, EP1 and EP2. The results in Table 7 reveal that the EP2 has lowest relative bias than both the EP1 and DIR estimators, except for $n_{(1)i} = 10$. Further, the EP2 has minimum relative RMSE and maximum efficiency than the EP1 and the DIR. The results in Table 7 clearly provide an encouraging performance of the proposed EP2 estimator. Figure 2 shows the boxplots of region-specific values of actual relative RMSE for DIR, EP1 and EP2 (all expressed in percentage terms) obtained in design-based simulations. Again, the EP2 outperforms EP1 and DIR in terms of the distribution of relative RMSE between regions for all the sample sizes of first (i.e. smaller) sample. Generally, the results set out in Table 7 and Figure 2 support the conclusion that the combining data from two surveys improves SAE, with the

Table 7. Values of percentage relative biases (RB), percentage relative root mean squared errors (RRMSE) and percentage relative efficiencies (RE) of the different estimators from design based simulations using the AAGIS data. The values are averaged over 12 regions.

	$n_{(1)i} = 5$			$n_{(1)i} = 10$			$n_{(1)i} = 20$		
Predictor	RB	RRMSE	RE	RB	RRMSE	RE	RB	RRMSE	RE
DIR	1.36	57.35	100	−0.30	37.29	100	0.19	27.59	100
EP1	1.56	54.09	106	−0.09	34.87	107	0.22	25.97	106
EP2	0.11	30.81	186	−0.55	23.73	157	−0.02	21.01	131
MBDE	1.61	60.65	95	0.51	38.71	96	1.20	30.66	90
EP3	1.68	56.98	101	0.77	35.14	106	1.31	27.71	100
EP4	0.30	29.68	193	−0.32	22.88	163	0.25	20.00	138

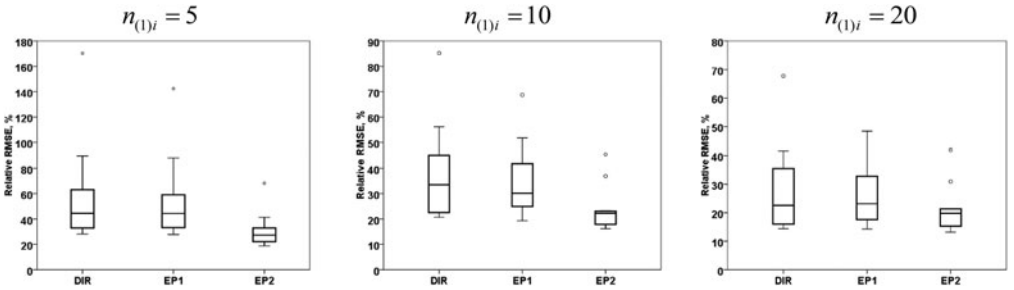

Figure 2. Boxplots of region-specific values of relative RMSE of the different predictors from design based simulations using the AAGIS data.

proposed EP2 emerging as the best performing of the methods that we investigated in the empirical evaluations.

5. Concluding remarks

This paper investigates SAE method by combining data from two independent surveys. The empirical results, based both on simulated data under the model as well as on real data, clearly indicate that combining information from two surveys can bring significant gains in SAE efficiency. The conclusions are essentially unaffected when we carried out similar simulations using chi-square (i.e. non-Gaussian) distributed random effects. We also suggest a method for estimating the MSE of the proposed small area estimator and demonstrate empirically that it performs well. In addition, we consider the non-sampled case, i.e. where the area of interest has no sample data in the main survey, and therefore no direct estimate. Synthetic estimation is the standard approach here, and we develop a synthetic version of proposed estimator using data from both surveys. The empirical results provide evidence that this estimator can generate reliable estimates for non-sampled areas.

There are several issues that still need to be explored in the context of using unit level models for combining data from two surveys. One important practical issue in this regard relates to the model robustness and validity for the second survey. A further issue relates to the link between the survey data and the spatial information. In this paper, we have assumed that different areas are independent to each other. However, it is often reasonable to assume that the effects of neighboring areas are correlated, with

the correlation decaying to zero as the distance between these areas increases (Chandra, Salvati, and Chambers 2007). The method needs to be extended under a spatial version of small area model. Authors are currently working on some of these issues.

In many countries, the crop cutting experiment (CCE) surveys are used for producing crop yield estimate. Although the CCE technique is an objective method of assessment of crop yield, the procedure of conduct of CCE is tedious and time consuming. As a result, most of the countries are looking the alternative method for crop yield estimation. But, the CCE technique is still most popular and practiced method for generating the crop yield estimates because of the nonexistence of reliable alternative. Due to this and some other factors, in every country, there is a felt need to reduce the sample size drastically in CCE surveys so that the volume of work can be reduced and managed properly. However, reduction in sample size can have a direct bearing on the standard error of the estimator. The reduction in sample size can be more alarming when this survey is also used for producing small area estimates. This is one of the most important and outstanding problem in agricultural statistics. The proposed method can directly be applied and hence recommended for disaggregate level crop yield estimation. In particular, CCE survey with very small size should be conducted where observation on yield (i.e. study variable) along with some auxiliary variables such as farmer's eye observation, seed rate, plant density, fertilizer rate and expert assessment of yield should be recorded. The second survey with relatively large sample size should be conducted where some of the auxiliary variables in common should be collected. These two surveys should be combined using the suggested SAE method. This statement clearly shows the usefulness of the proposed method as a prospect SAE method for real life application.

Acknowledgments

The authors would like to acknowledge the valuable comments and suggestions of the Editor and an anonymous referee. These led to a considerable improvement in the paper.

References

Battese, G. E., R. M. Harter, and W. A. Fuller. 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. doi:10.1080/01621459.1988.10478561.

Chandra, H., and R. Chambers. 2011. Small area estimation under transformation to linearity. *Survey Methodology* 37 (1):39–51. doi:10.1002/bimj.201300233.

Chandra, H., and R. Chambers. 2009. Multipurpose weighting for small area estimation. *Journal of Official Statistics* 25 (3):379–95. doi: ro.uow.edu.au/infopapers/3326.

Chandra, H., N. Salvati, and R. Chambers. 2007. Small area estimation for spatially correlated populations- a comparison of direct and indirect model-based estimators. *Statistics in Transition* 8 :331–50. doi: ro.uow.edu.au/infopapers/3022.

Chandra, H., N. Salvati, R. Chambers, and N. Tzavidis. 2012. Small area estimation under spatial nonstationarity. *Computational Statistics and Data Analysis* 56 (10):2875–88. doi:10.1016/j.csda.2012.02.006.

Chandra, H., U. C. Sud, and Y. Gharde. 2015. Small area estimation using estimated population level auxiliary data. *Communications in Statistics-Simulation and Computation* 44 (5):1197–209. doi:10.1080/03610918.2013.810255.

Cochran, W. G. 1977. *Sampling techniques*. 3rd ed. New York: John Wiley and Sons.
Datta, G. S., and P. Lahiri. 2000. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* 10 (2):613–27. doi: jstor.org/stable/24306735.
Elliott, M. R., and W. W. Davis. 2005. Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (3):595–609. doi:10.1111/j.1467-9876.2005.05459.x.
Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72 (358):320–38. doi:10.1080/01621459.1977.10480998.
Hidiroglou, M. A. 2001. Double sampling. *Survey Methodology* 27 (2):143–54.
Kim, J. K., S. Park, and S. Kim. 2015. Small area estimation combining information from several sources. *Survey Methodology* 41 :21–36.
Kim, J. K., and J. N. K. Rao. 2012. Combining data from two independent surveys: A model assisted approach. *Biometrika* 99 (1):85–100. doi:10.1093/biomet/asr063.
Lohr, S., and N. G. N. Prasad. 2003. Small area estimation with auxiliary survey data. Canadian Journal of Statistics 31 (4):383–96. doi:10.2307/3315852.
Lohr, S., and J. N. K. Rao. 2006. Estimation in multiple-frame surveys. *Journal of the American Statistical Association* 101 (475):1019–30. doi:10.1198/016214506000000195.
Marker, D. A. 2001. Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology* 27 (2):183–8.
Manzi, G., D. J. Spiegelhalter, R. M. Turner, J. Flowers, and S. G. Thompson. 2011. Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (1):31–50. doi:10.1111/j.1467-985X.2010.00648.x.
McCulloch, C. E., and S. R. Searle. 2001. *Generalized, linear and mixed models*. New York: John Wiley and Sons.
Maples, J. 2017. Improving small area estimate of disability: Combining the American communities survey with the survey of income and program participation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(4):1211–27. doi: doi:10.1111/rssa.12310.
Merkouris, T. 2010. Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society. Series B* 68 :509–21. doi: org/10.1111/j.1467-9868.2009.00724.x.
Merkouris, T. 2004. Combining independent regression estimators from multiple surveys. *Journal of American Statistical Association* 99 (468):1131–9. doi:10.1198/016214504000000601.
Moriarity, C., and F. Scheuren. 2001. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* 17 :407–22.
Pfeffermann, D. 2013. New important developments in small area estimation. *Statistical Science* 28 (1):40–68. doi:10.1214/12-STS395.
Prasad, N. G. N., and J. N. K. Rao. 1990. The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* 85 (409):163–71. doi:10.1080/01621459.1990.10475320.
Rao, J. N. K. 2003. *Small area estimation*. New York: Wiley.
Rao, J. N. K., and I. Molina. 2015. *Small area estimation*. 2nd ed. New Jersey: John Wiley and Sons.
Renssen, R. H., and N. Nieuwenbroek. 1997. Aligning estimates for common variables in two or more sample surveys. *Journal of American Statistical Association* 92 (437):368–75. doi:10.1080/01621459.1997.10473635.
Royall, R. M., and W. G. Cumberland. 1978. Variance estimation in finite population sampling. *Journal of the American Statistical Association* 73 (362):351–8. doi:10.1080/01621459.1978.10481581.
Särndal, C. E., B. Swensson, and J. H. Wretman. 1992. *Model assisted survey sampling*. New York: Springer-Verlag.

Schenker, N., and T. Raghunathan. 2007. Combining informationfrom multiple surveys to enhance estimation of measures of health. *Statistics in Medicine* 26 (8):1802–11. doi:10.1002/sim.2801.

Wu, C. 2004. Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics* 32 (1):15–26. doi:10.2307/3315996.

Ybarra, L. M. R., and S. L. Lohr. 2008. Small area estimation when auxiliary information is measured with error. *Biometrika* 95 (4):919–31. doi:10.1093/biomet/asn048.

Zieschang, K. D. 1990. Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association* 85 (412):986–1001. doi:10.1080/01621459.1990.10474969.