Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 73(1) 2019 59–69

Small Domain Inference Combining Data from Two Independent Surveys*

Sadikul Islam[1] and Hukum Chandra[2]
[1]ICAR-Indian Institute of Soil and Water Conservation, Dehradun
[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 31 August 2018; Revised 02 March 2019; Accepted 05 March 2019

SUMMARY

Many often two surveys conducted independently, may have some auxiliary variables in common along with a set of extra variables that are not common to both the surveys. One survey, which is small in sample size but collects both variable of interest as well as a set of auxiliary variables. The another survey which is relatively larger in sample size, does not collects variable of interest but collects a set of auxiliary variables, common to the small survey. In addition, the large survey collects multiple response variables as well as set of auxiliary variables not common to the small survey. A small area predictor for small domain (or area) means is proposed by combining data from these two surveys using multipurpose weights. Empirical results from model-based as well as design-based simulations indicate that the proposed small area predictor that incorporates the additional auxiliary variables of the large survey along with the common auxiliary variables, provide better efficiency gain.

Keyewords: Combining data; Empirical predictor; Independent surveys; Non-sample area; Small domain; Multipurpose weight; Variable specific EBLUP weight.

1. INTRODUCTION

Sample survey is a cost effective approach for obtaining information on wide ranging of topics by observing a part of the population and making inferences about the population characteristics. Recent years demand for subpopulation or domain level estimates has increased tremendously. For example, estimates for small geographical areas like District, Tehsil or Village Panchayat etc, see Molina and Rao (2015). In real survey scenarios, frequently observed domain of interests are small domains, for which domain specific sample size is not large enough to produce reliable direct estimates. The indirect model based small area estimation (SAE) methods are popularly used to produce small area estimates (Molina and Rao 2015). But, further making improvement in reliability of SAE methods are still a challenging issue for survey statistician, due to scarcity in area specific sample data and increase in overall survey sample size is practically not feasible due to, budget and time constraints.

Often multiple agencies, government or private departments or organizations conduct surveys on same population independently for their own interest. Two surveys conducted independently on the same population can have one or more auxiliary variables in common along with set of variables that are not common for both the surveys. Further, it is also possible to have some linear relationships of variable of interest with the auxiliary variables. Hence, it seems to be attractive to utilize the data of both the surveys to improve precision in estimation. Different authors are already addressing the problem of survey data combining from two independent surveys having common variable of interest as well as auxiliary variables and considered estimation at population and large domain levels, see for example, Zieschang (1990), Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004), Wu (2004) and Kim and Rao (2012).

It is well known that the problem of scarcity and inadequacy of sample data is much more prominent at small domain level compared to the population level.

Corresponding author: Sadikul Islam
E-mail address: sadikul.islamiasri@gmail.com

*Paper presented in Dr. G.R. Seth Memorial Young Scientist Award Session on December 14, 2018 in the 72nd Annual Conference of the Indian Society of Agricultural Statistics held at ICAR-CIAE, Bhopal during December 13-15-2018.

Hence, combining data from two independent surveys can be advantageous to produce reliable small domain (or area) estimates. Several authors have already discussed different approach of combining survey data from multiple surveys for SAE, see for example Marker (2001), Moriarity and Scheuren (2001), Lohr and Prasad (2003), Rao (2003), Elliott and Davis (2005), Lohr and Rao (2006), Merkouris (2010), Ybarra and Lohr (2008) and Manzi *et al.* (2011). Kim *et al.* (2015) developed SAE approach for combining data from several sources using area level model. Maples (2017) extended the method of Kim and Rao (2012) for estimation of small area proportions from binary variable under logistic linear mixed model by combining data from two independent surveys. Islam *et al.* (2018) proposed small area estimator under a spatial dependent random effects model by combining data from two independent surveys. Recently, Islam and Chandra (2019) proposed estimation of small area means under a linear mixed model by combining information from two independent surveys. Maples (2017), Islam *et al.* (2018) and Islam and Chandra (2019) discussed SAE approach by considering the situation that one survey which is small in sample size, collects both variable of interest as well as auxiliary variables whereas the another survey, relatively larger in sample size, has only collects auxiliary variables common to small survey. In particular, they proposed SAE method, using common set of auxiliary information for combining data from two independent surveys and they does not consider any extra set of variables available for large survey not common to the small survey. In practice, survey with large size possible to have rich set of variables, of which a subset of variables possibly common to small survey. The sub-set of extra variables that are not common to small survey, can be both set of response variables as well as auxiliary variables of the large survey.

In this article, we extend the Islam and Chandra (2019) approach of SAE by combining data from two independent surveys. We considered that a survey with small in sample size that collects variable of interest as well auxiliary information and the another survey which is large in sample size does not collect the variable of interest but collect auxiliary variables common to the small survey. In addition, the large survey collects a set of extra response variables as well as auxiliary variables that are not common to the small survey. For developing small area estimators, it is assumed that the variable of interest is realization of linear mixed model with the common set of auxiliary variables and the response variables of the large survey are also realization of linear mixed with different parameter values with the extra set of auxiliary variables. Further, it is also considered that the area specific aggregate values (e.g. population means or totals) for all the auxiliary variables are not available but population level aggregate values of the auxiliary variables are available from different administrative or census sources. The rest of the article is organized as follows: In Section 2, we discussed the notations and different SAE methods. Further in sub-section 2.1 proposes an approach of SAE using multipurpose weight. The empirical evaluations of the proposed estimators are performed through model based as well as design based simulation studies in Section 3. Finally, concluding remarks are discussed in Section 4.

2. NOTATIONS AND SMALL AREA ESTIMATION

In this section, we define notations which are set out as follows. Let us assume that a population U consists of N finite population units. Further, assume that the population is composed of D non-overlapping small domains (or areas) denoted as U_d. Here, d is indexing D areas. It is assumed that the area specific population size is N_d which summed up to the whole population, $N = \sum_{d=1}^{D} N_d$ .We consider y_{dj} as the value of the variable of interest y for unit j in area d and the area-specific population mean for area d is $\bar{Y}_d = N_d^{-1} \sum_{j=1}^{N_d} y_{dj}$. Further, we assume that two surveys are conducted in the population U independently denoted as $S_{(I)}$ and $S_{(II)}$ m respectively. The sample sizes for $S_{(I)}$ and $S_{(II)}$ are $n_{(I)}$ and $n_{(II)}$ respectively, assuming $n_{(II)}$ is much larger than $n_{(I)}$. In the rest of the article, we also called $S_{(I)}$ and $S_{(II)}$ survey as small and large survey, respectively with same meaning. The terms related to sample and non-sample part of $S_{(I)}$ is denoted by the $s_{(I)}$ and $r_{(I)}$, respectively. Similarly, the terms related to sample and non-sample part of $S_{(II)}$ is denoted by the subscripts $s_{(II)}$ and $r_{(II)}$, respectively. The area-specific sample size for $S_{(I)}$ and $S_{(II)}$ are denoted by $n_{(I)d}$ and $n_{(II)d}$, ($d = 1, 2, \ldots, D$), respectively. The area-specific sample

sizes of both the surveys are summed up to overall sample size as $n_{(I)} = \sum_{d=1}^{D} n_{(I)d}$ and $n_{(II)} = \sum_{d=1}^{D} n_{(II)d}$. Now, it is considered that the small survey $S_{(I)}$ has collected both variable of interest *y* as well as vector of auxiliary variables $\mathbf{x}_{(c)}$ of order *P*. The larger survey $S_{(II)}$ does not collect the variable of interest *y* but it collects all the *P* auxiliary variables that are common to the small survey $S_{(I)}$ and in addition $S_{(II)}$ has collected *K* different response variables (denoted as $Y_1, Y_2, ..., Y_K$) as well as a vector of auxiliary variables $\mathbf{x}_{(uc)} = (X_1, ..., X_Q)$ of order *Q*, not collected in the small survey $S_{(I)}$. It is assumed that the variable of interest *y* is realization of unit level model linear mixed model, based on *P* common auxiliary variables $\mathbf{x}_{(c)}$ and further the *K* response variables are assumed to have same relationship with the set of *Q* extra auxiliary variables $\mathbf{x}_{(uc)}$, with different parameters value. In this section, we use an extra subscript $k (k = 1, ..., K)$ for indexing quantities associated with the response variable *k*.

The design-based direct estimator (DIR) of area *d* mean, $\bar{Y}_d$ using data from the small survey, $S_{(I)}$ is $\hat{\bar{Y}}_d^{DIR} = \sum_{j \in s_{(I)i}} w_{(I)dj}^{dw} y_{dj}$ Särndal *et al.* (1992). Here, $w_{(I)dj}^{dw} = w_{(I)dj}^{*dw} \Big/ \sum_{j \in s_{(I)d}} w_{(I)dj}^{*dw}$ is a normalized survey design weight of and $w_{(I)dj}^{*dw}$ is survey design weight of $S_{(I)}$ for unit *j* in area *d*. Following Chandra and Chambers (2009, 2011), the MBDE (denoted by MBDE) of area *d* mean of *y* is defined as

$$\hat{\bar{Y}}_d^{MBDE} = \sum_{j \in s_{(I)d}} \tilde{w}_{(I)dj}^{EBLUP} y_{dj} , \quad (1)$$

with, $\tilde{w}_{(I)dj}^{EBLUP} = w_{(I)dj}^{EBLUP} / \sum_{j \in s_{(I)d}} w_{(I)dj}^{EBLUP}$. Here,

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

with $\hat{\mathbf{H}}_{s_{(I)}} = \left(\mathbf{x}_{(c)s_{(I)}}^{T} \hat{\mathbf{v}}_{s_{(I)}s_{(I)}}^{-1} \mathbf{x}_{(c)s_{(I)}}\right)^{-1} \mathbf{x}_{(c)s_{(I)}}^{T} \hat{\mathbf{v}}_{s_{(I)}s_{(I)}}^{-1}$. Here $\hat{\mathbf{v}}_{s_{(1)}s_{(1)}}$ denotes estimate of variances between sampled units, $\hat{\mathbf{v}}_{s_{(I)}r_{(I)}}$ denotes estimate of variances between sampled and non-sampled units, $\mathbf{t}_{x_{(c)}} = \sum_{d=1}^{D} \sum_{j=1}^{N_d} \mathbf{x}_{(c)dj}$, $\hat{\mathbf{t}}_{x_{(c)s_{(I)}}} = \sum_{d=1}^{D} \sum_{j=1}^{n_{(I)d}} \mathbf{x}_{(I)dj} = n_{(I)} \bar{\mathbf{x}}_{(c)s_{(I)}}$, $\mathbf{I}_{s_{(I)}}$ is the identity matrix of order $n_{(I)}$, $\mathbf{1}_{s_{(I)}}$ and $\mathbf{1}_{r_{(I)}}$ denotes a vector of ones of size $n_{(I)}$ and $(N - n_{(I)})$, respectively.

Table 1. Description about the information available for the survey $S_{(I)}$ and $S_{(II)}$.

Variables	$S_{(I)}$	$S_{(II)}$
Response variable	*Y* (Our variable of interest)	$Y_1, Y_2, ..., Y_K$
Auxiliary variable	$\mathbf{x}_{(c)}$ vector of length *P*	$\mathbf{x}_{(c)}$ vector of order *P*
		$\mathbf{x}_{(uc)}$ vector of order *Q*

To this end, let us assume a unit level linear mixed model of form

$$y_{dj} = \mathbf{x}_{(c)dj}^{T} \boldsymbol{\beta} + u_d + e_{dj} , \; j = 1, ..., N_d ; d = 1, .., D , \quad (2)$$

where, $\boldsymbol{\beta}$ is a *P* vector of regression coefficients, u_d denotes area-specific random effect for area *d*, e_{dj} is an individual random effect for unit *j* in area *d*. It is assumed that u_d and e_{dj} are independent and separately follow normal distribution with zero mean and constant variances σ_u^2 and σ_e^2, respectively (Battese *et al.*, 1988). The model (2) is fitted using the data of small survey $S_{(I)}$ and the parameters of the model are estimated using maximum likelihood (ML) or restricted ML (REML) estimation methods (Harville 1977). The vector of estimated parameters of (2) is denoted as $(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$. Under this scenario, Chandra *et al.* (2015) discussed an empirical predictor for small area mean in area *d* denoted by EP1 is defined as

$$\hat{\bar{Y}}_d^{EP1} = \bar{\mathbf{x}}_{(c)s(I)d} + \hat{u}_d , \quad (3)$$

where,

$\hat{u}_d = \hat{\gamma}_d (\bar{y}_{s_{(I)d}} - \bar{x}_{(c)s_{(I)d}}^{T} \hat{\beta})$, $\hat{\gamma}_d = \hat{\sigma}_u^2 (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_{(I)d})^{-1}$, $\bar{\mathbf{x}}_{(c)s_{(I)d}} = n_{(I)d}^{-1} \sum_{j=1}^{n_{(I)d}} \mathbf{x}_{(c)(I)dj}$. Islam and Chandra (2019) discussed an empirical predictor (denoted by EP3) of area *d* mean of *y* based on $S_{(I)}$ data is defined as

$$\hat{\bar{Y}}_d^{EP3} = (\hat{\bar{\mathbf{x}}}_{(I)d}^{EBLUP})^T \hat{\boldsymbol{\beta}} + \hat{u}_d , \quad (4)$$

where, $\hat{\bar{\mathbf{x}}}_{(I)d}^{EBLUP} = \sum_{j \in s_{(I)d}} \tilde{w}_{(I)dj}^{EBLUP} \mathbf{x}_{(I)dj}$. Islam and Chandra (2019) proposed an empirical predictor of small area mean in area *d* (EP2) using design weight

as well as synthetic values of y of second survey $S_{(II)}$ is

$$\hat{\bar{Y}}_d^{EP2} = \hat{\bar{\mathbf{x}}}_{(c)(II)d}^T \hat{\boldsymbol{\beta}} + \hat{u}_d \,, \tag{5}$$

where, $\hat{\bar{\mathbf{x}}}_{(II)d}^T = \sum_{j\in s_{(II)d}} w_{(II)dj}^{dw} \mathbf{x}_{(c)(II)dj}$ is design-based direct estimate of $\bar{\mathbf{x}}_{(c)d} = N_d^{-1} \sum_{d=1}^{D} \sum_{j=1}^{N_d} \mathbf{x}_{(c)dj}$, $w_{(II)dj}^{dw} = w_{(II)dj}^{*dw} \Big/ \sum_{j\in s_{(II)d}} w_{(II)dj}^{*dw}$ is normalized survey weight of $S_{(II)}$ for unit j in area d and $w_{(II)dj}^{*dw}$ is design weight of large survey $S_{(II)}$ for unit j in area d. Further, Islam and Chandra (2019) developed the empirical predictor (EP4) of area d mean using synthetic values of variable study and EBLUP weights of $_{(II)}$ is defined as

$$\hat{\bar{Y}}_d^{EP4} = (\hat{\bar{\mathbf{x}}}_{(c)(II)d}^{EBLUP})^T \hat{\boldsymbol{\beta}} + \hat{u}_d \,, \tag{6}$$

where, $\tilde{w}_{(II)dj}^{EBLUP} = w_{(II)dj}^{EBLUP} / \sum_{j\in s_{(II)d}} w_{(II)dj}^{EBLUP}$ and $\hat{\bar{\mathbf{x}}}_{(c)(II)d}^{EBLUP} = \sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{EBLUP} \mathbf{x}_{(c)(II)dj}$.

Here,

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

where $\hat{\mathbf{H}}_{s_{(II)}} = \left(\mathbf{x}_{(c)s_{(II)}}^T \hat{\mathbf{v}}_{s_{(II)}s_{(II)}}^{-1} \mathbf{x}_{(c)s_{(II)}}\right)^{-1} \mathbf{x}_{(c)s_{(II)}}^T \hat{\mathbf{v}}_{s_{(II)}s_{(II)}}^{-1}$, $\hat{\mathbf{v}}_{s_{(II)}s_{(II)}}$ denotes estimate of variances between sampled units, $\hat{\mathbf{v}}_{s_{(II)}r_{(II)}}$ denotes estimate of variances between sampled and non-sampled units, $\mathbf{t}_{x_{(c)}} = \sum_{d=1}^{D} \sum_{j=1}^{N_d} \mathbf{x}_{dj}$, $\hat{\mathbf{t}}_{x_{(c)s_{(II)}}} = \sum_{d=1}^{D} \sum_{j=1}^{n_{(II)d}} \mathbf{x}_{(c)(II)dj} = n_{(II)} \bar{\mathbf{x}}_{(c)s_{(II)}}$, $\mathbf{I}_{s_{(II)}}$ is the identity matrix of order $n_{(II)}$, $\mathbf{1}_{s_{(II)}}$ and $\mathbf{1}_{r_{(II)}}$ denotes a vector of ones of size $n_{(II)}$ and $(N - n_{(II)})$ respectively.

2.1 Small area estimation using multipurpose weight

Different estimators described in the previous section are either based on small survey data $S_{(I)}$ or combining data from two surveys using common auxiliary variables $\mathbf{x}_{(c)}$ and the estimators based on combining data from two independent surveys are not incorporating the extra set of variables, collected for larger survey $S_{(II)}$ only. Following Chandra and Chambers (2009), we propose a multipurpose weight based SAE method by incorporating extra set of variables [$Y_k, (k = 1,2,...,K)$ and $\mathbf{x}_{(uc)} = (x_{1(uc)},...,x_{Q(uc)})$] of large survey $S_{(II)}$ along with set of common auxiliary variables $\mathbf{x}_{(c)} = (x_{1(c)},...,x_{P(c)})$. For developing small area inference, again we assume that the K response variables individually holds model (2) with $\mathbf{x}_{(uc)}$, although with different parameter values of the form

$$\mathbf{y}_{kU} = \mathbf{x}_{(uc)U} \boldsymbol{\beta}_k + \mathbf{Z}_U \mathbf{u}_k + \mathbf{e}_{kU} \,, \tag{7}$$

where, $\mathbf{y}_{kU} = (\mathbf{y}_{k1}^T,\ldots,\mathbf{y}_{kD}^T)^T$, $\mathbf{x}_{(uc)U} = (\mathbf{x}_{(uc)1}^T,...,\mathbf{x}_{(uc)Q}^T)^T$, $\mathbf{Z}_U = diag(\mathbf{z}_d = \mathbf{1}_{N_d}; 1 \le d \le D)$, $\mathbf{u}_k = (u_{k1},\ldots,u_{kD})^T$ and $\mathbf{e}_{kU} = (\mathbf{e}_{k1}^T,\ldots,\mathbf{e}_{kD}^T)^T$. Since different areas are independent, the covariance matrix of k^{th} response variable $\mathbf{y}_{kU}$ has block diagonal structure is given as $\mathbf{V}_{kU} = diag(\mathbf{v}_{kd}; 1 \le d \le D)$ with $\mathbf{v}_{kd} = Var(\mathbf{y}_{kd}) = \sigma_{ku}^2 \mathbf{z}_d \mathbf{z}_d^T + \sigma_{ke}^2 \mathbf{I}_{N_d}$ and $\mathbf{I}_{N_d}$ is the identity matrix of order N_d. Using the estimated values of the variance components $\hat{\sigma}_{ku}^2$ and $\hat{\sigma}_{ke}^2$, the estimated covariance matrix is given by $\hat{\mathbf{V}}_{kU} = diag(\hat{\mathbf{v}}_{kd}; 1 \le d \le D)$, with $\hat{\mathbf{v}}_{kd} = \hat{\sigma}_{ku}^2 \mathbf{z}_d \mathbf{z}_d^T + \hat{\sigma}_{ke}^2 \mathbf{I}_{N_d}$. Given a sample of $S_{(II)}$ from this population, without loss of generality, we arrange the vector $\mathbf{y}_{kU}$ so that its first $n_{(II)}$ elements correspond to the sample units, and then partition $\mathbf{y}_{kU}$, $\mathbf{x}_{(uc)U}$, $\mathbf{Z}_{kU}$ and $\mathbf{e}_{kU}$ according to sample and non-sample units. Therefore, we can write (7) as follows:

$$\mathbf{y}_{kU} = \begin{bmatrix} \mathbf{y}_{ks_{(II)}} \\ \mathbf{y}_{kr_{(II)}} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{(uc)s_{(II)}} \\ x_{(uc)r_{(II)}} \end{bmatrix} \boldsymbol{\beta}_k + \begin{bmatrix} \mathbf{Z}_{s_{(II)}} \\ \mathbf{Z}_{r_{(II)}} \end{bmatrix} \mathbf{u}_k + \begin{bmatrix} \mathbf{e}_{ks_{(II)}} \\ \mathbf{e}_{kr_{(II)}} \end{bmatrix},$$

with variance matrix given by $\mathbf{V}_{kU} = \begin{bmatrix} \mathbf{v}_{ks_{(II)}s_{(II)}} & \mathbf{v}_{ks_{(II)}r_{(II)}} \\ \mathbf{v}_{kr_{(II)}s_{(II)}} & \mathbf{v}_{kr_{(II)}r_{(II)}} \end{bmatrix}$.

Here, $\mathbf{x}_{(uc)s_{(II)}}$ is a matrix order of $n_{(II)} \times Q$, containing the values of the auxiliary variables. $\mathbf{v}_{ks_{(II)}s_{(II)}} = diag\{\sigma_{ku}^2 \mathbf{1}_{n_{(II)d}} \mathbf{1}_{n_{(II)d}}^T + \sigma_{ke}^2 \mathbf{I}_{n_{(II)d}}; d = 1,...,D\}$ is the $n_{(II)} \times n_{(II)}$ matrix of covariances of the response variable among the $n_{(II)}$ sampled units of $S_{(II)}$. Similarly, $\mathbf{v}_{ks_{(II)}r_{(II)}} = diag(\sigma_{ku}^2 \mathbf{1}_{n_{(II)d}} \mathbf{1}_{N_d - n_{(II)d}}^T; d = 1,...,D)$ is a matrix of order $n_{(II)} \times (N - n_{(II)})$ represents the covariances of the response variable for sampled and non-sampled units for the survey $S_{(II)}$. Here, $\mathbf{I}_{n_{(II)d}}$ is the identity matrix of order $n_{(II)d}$, $\mathbf{1}_{n_{(II)d}}$ and $\mathbf{1}_{N_d - n_{(II)d}}$ denotes a vector of ones of size $n_{(II)d}$ and $N - n_{(II)d}$ respectively. We fit the model (7) using sample data of $S_{(II)}$ for each of K response variables independently

with respect to auxiliary variables $\mathbf{x}_{(uc)}$ and parameters are estimated for each response variable y_k denoted as $\hat{\boldsymbol{\beta}}_k$ $\hat{\sigma}^2_{ku}$ and $\hat{\sigma}^2_{ke}$, (k=1,2,…,K). We use these parameter estimates for estimation of covariance matrix denoted by $\hat{\mathbf{V}}_{kU} = diag(\hat{\mathbf{v}}_{kd}; 1 \le d \le D)$, where $\hat{\mathbf{v}}_{kd} = \hat{\sigma}^2_{k,u}\mathbf{z}_d\mathbf{z}_d^T + \hat{\sigma}^2_{k,e}\mathbf{I}_{N_d}$. Now, we define the sample weights that define the variable specific EBLUP weight under (7), for the population total of response variable y_k $(k = 1, 2\ldots K)$:

$$\begin{aligned}\mathbf{w}^{EBLUP}_{ks_{(II)}} &= (w^{EBLUP}_{(II)kdj}) \\ &= \mathbf{1}_{s_{(II)}} + \hat{\mathbf{H}}^T_{ks_{(II)}}(\mathbf{t}_{\mathbf{x}_{(uc)}} - \hat{\mathbf{t}}_{\mathbf{x}_{(uc)s_{(II)}}}) + \\ &\quad (\mathbf{I}_{s_{(II)}} - \hat{\mathbf{H}}^T_{ks_{(II)}}\mathbf{x}^T_{(uc)s_{(II)}})\hat{\mathbf{v}}^{-1}_{ks_{(II)}s_{(II)}}\hat{\mathbf{v}}_{ks_{(II)}r_{(II)}}\mathbf{1}_{r_{(II)}} \quad (8)\end{aligned}$$

Here, $\hat{\mathbf{H}}_{ks_{(II)}} = \left(\mathbf{x}^T_{(uc)s_{(II)}}\hat{\mathbf{v}}^{-1}_{ks_{(II)}s_{(II)}}\mathbf{x}_{(uc)s_{(II)}}\right)^{-1}\mathbf{x}^T_{(uc)s_{(II)}}\hat{\mathbf{v}}^{-1}_{ks_{(II)}s_{(II)}}$; $\mathbf{I}_{s_{(II)}}$ is the identity matrix of order $n_{(II)}$; $\mathbf{1}_{s_{(II)}}$ and $\mathbf{1}_{r_{(II)}}$ are denoted as vector of ones of size $n_{(II)}$ and $N - n_{(II)}$, respectively, $\mathbf{t}_{\mathbf{x}_{(uc)}} = \sum_{d=1}^{D}\sum_{j=1}^{N_d}\mathbf{x}_{(uc)dj}$ and $\hat{\mathbf{t}}_{\mathbf{x}_{(uc)}} = \sum_{d=1}^{D}\sum_{j=1}^{n_{(2)d}}\mathbf{x}_{(uc)dj}$ are the vectors of population and sample totals of $\mathbf{x}_{(uc)}$, respectively. In $S_{(II)}$ the variable of interest y is not collected, so it is not possible to use the extra set of covariates $\mathbf{x}_{(uc)}$ directly to for SAE. For utilising the extra set of variables, a common multipurpose weights for all the K response variables is developed using the parameter estimates $\hat{\boldsymbol{\beta}}_k$ $\hat{\sigma}^2_{ku}$ and $\hat{\sigma}^2_{ke}$, (k=1, 2… K) of fitted (7) and variable specific EBLUP weights (8). Following Islam and Chandra (2019) generate synthetic values of variable of interest y using fitted (2) corresponding to common auxiliary variables $\mathbf{x}_{(c)}$ of large survey $S_{(II)}$. The developed multipurpose weight of $S_{(II)}$ is used as a plugged in weight for synthetic y for developing small area means estimator.

Following Chandra and Chambers (2009) the optimal set of multipurpose weights using the K response variables and the extra set of covariates $\mathbf{x}_{(uc)}$, collected in $S_{(II)}$ denoted as $w^{MP}_{s_{(II)}} = \left\{w^{MP}_{(II)dj} \in S_{(II)}\right\}$. In what follows, we describe the derivations for construction of multipurpose weights. Let the population total of y_k is denoted by $T_k = \mathbf{1}_N^T\mathbf{y}_{kU}$ and the estimator of this based on multipurpose weight, denoted by $\hat{T}_k = (\mathbf{w}^{MP}_{ks_{(2)}})^T\mathbf{y}_k$. The weights $\mathbf{w}^{MP}_{s_{(2)}}$ are derived based on two criteria are as follows:

(a) $E\left(\hat{T}_k - T_k\right) = 0$ for each value of k, (b) $\sum_k \psi_k Var\left(\hat{T}_k - T_k\right)$ is minimized at $w^{MP}_{s_{(2)}}$ where $Var\left(\hat{T}_k - T_k\right)$ is prediction variance, $\sum_k \psi_k Var\left(\hat{T}_k - T_k\right)$ is called the ψ-weighted total prediction variance. Here ψ is a nonnegative scalar quantity specified by user and ψ denotes the relative importance associate with each response variable in such a way that $\sum_{k=1}^{K}\psi_k = 1$. The weights $w^{MP}_{s_{(2)}}$ is called ψ-optimal if the two criteria (a) and (b) are fulfilled. Following Chandra and Chambers (2009) the expression of optimal multipurpose sample weights for K response variables are defined as

$$\begin{aligned}\mathbf{w}^{MP1}_{s_{(II)}} = \left(\tilde{w}^{MP1}_{(II)dj}\right) &= \mathbf{1}_{s_{(II)}} + \mathbf{H}^T_{1s_{(II)}}\left(\mathbf{t}_{\mathbf{x}_{(uc)}} - \hat{\mathbf{t}}_{\mathbf{x}_{(uc)s_{(II)}}}\right) + \\ &\quad \left(\mathbf{I}_{s_{(II)}} - \mathbf{H}^T_{1s_{(II)}}\mathbf{x}^T_{(uc)s_{(II)}}\right)\mathbf{U}_1^{-1}\mathbf{W}_1\mathbf{1}_{r_{(II)}}. \quad (9)\end{aligned}$$

Here, $\mathbf{H}_{1s_{(II)}} = \left(\mathbf{x}^T_{(uc)s_{(II)}}\mathbf{U}_1^{-1}\mathbf{x}_{(uc)s_{(II)}}\right)^{-1}\mathbf{x}^T_{(uc)s_{(II)}}\mathbf{U}_1^{-1}$, $\mathbf{U}_1 = diag(U_{1d}; 1 \le d \le D)$ and $\mathbf{W}_1 = diag(W_{1d}; 1 \le d \le D)$

where $U_{1d} = \sum_{k=1}^{K}\psi_k\mathbf{v}_{ks_{(II)}s_{(II)},d} = \sum_{k=1}^{K}\psi_k(\sigma^2_{u,k}\mathbf{z}_d\mathbf{z}_d^T + \sigma^2_{e,k}\mathbf{I}_{N_d})$ and $W_{1d} = \sum_{k=1}^{K}\psi_k\mathbf{v}_{ks_{(II)}r_{(II)},d} = \sum_{k=1}^{K}\psi_k(\sigma^2_{u,k}\mathbf{1}_{n_{(II)d}}\mathbf{1}^T_{N_d - n_{(II)d}})$.

The empirical predictor (denoted by EP.MP1) of area d mean based on the multipurpose weight (9) and synthetic y values of $S_{(II)}$ is defined as

$$\begin{aligned}\hat{\bar{Y}}^{EP.MP1}_d &= \sum_{j\in s_{(II)d}}\tilde{w}^{MP1}_{(II)dj}(\mathbf{x}^T_{(c)dj}\hat{\boldsymbol{\beta}} + \hat{u}_d) \\ &= (\sum_{j\in s_{(II)d}}\tilde{w}^{MP1}_{(II)dj}\mathbf{x}_{(c)dj})^T\hat{\boldsymbol{\beta}} + \hat{u}_d \\ &= (\hat{\bar{\mathbf{x}}}^{MP1}_{(II)d})^T\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad (10)\end{aligned}$$

where, $\tilde{w}^{MP1}_{(II)dj} = w^{MP1}_{(II)dj} / \sum_{j\in s_{(II)d}} w^{MP1}_{(II)dj}$ and $\hat{\bar{\mathbf{x}}}^{MP1}_{(II)i} = \sum_{j\in s_{(II)d}}\tilde{w}^{MP1}_{(II)dj}\mathbf{x}_{(II)dj}$ with $E_d(\hat{\bar{\mathbf{x}}}^{MP1}_{(II)d}) = \bar{\mathbf{x}}_d$.

Further, following Chandra and Chambers (2009) we also discuss the second method of deriving multipurpose weights based on "importance averaging" of the variable-specific EBLUP sample weights (8) across all K response variables is defined as:

$$\tilde{\mathbf{w}}^{MP2}_{s_{(II)}} = \sum_{k=1}^{K}\psi_k\mathbf{w}^{EBLUP}_{ks_{(II)}} \quad (11)$$

Here, ψ_k is importance of the response variable y_k in such a way that, $\sum_{k=1}^{K}\psi_k = 1$. The empirical predictor (denoted by EP.MP2) of area d mean based on the multipurpose weights MP2 (11) and the synthetic values of $S_{(II)}$ is defined as

$$\begin{aligned}\hat{\bar{Y}}_d^{EP.MP2} &= \sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP2}(\mathbf{x}_{(c)dj}^T\hat{\boldsymbol{\beta}} + \hat{u}_d) \\ &= (\sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP2}\mathbf{x}_{(c)dj})^T\hat{\boldsymbol{\beta}} + \hat{u}_d \\ &= (\hat{\bar{\mathbf{x}}}_{(II)d}^{MP2})^T\hat{\boldsymbol{\beta}} + \hat{u}_d, \end{aligned} \quad (12)$$

where, $\tilde{w}_{(II)dj}^{MP2} = w_{(II)dj}^{MP2} / \sum_{j\in s_{(II)d}} w_{(II)dj}^{MP2}$ and $\hat{\bar{\mathbf{x}}}_{(II)i}^{MP2} = \sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP2}\mathbf{x}_{(c)(II)dj}$ with $E_d(\hat{\bar{\mathbf{x}}}_{(II)d}^{MP2}) = \bar{\mathbf{x}}_d$.

The area with no sample data or non-sample area (out of sample area) is very common challenge for any small area estimation method. Hence without addressing the applicability of the develop estimators for non-sample areas, purpose cannot be fulfilled. Recently, Islam and Chandra (2019) proposed synthetic estimators SYN.EP2 and SYN.EP4 for small area mean estimation for non-sampled areas. Hence following the approach of Islam and Chandra (2019) develop modified version of EP.MP1 and EP.MP2 estimators that can produced estimates for non-sample areas. Here, we assume that the small survey has D_{ns} non-sample areas out of D areas but for large survey data for non-sample areas are collected. We generate the synthetic y values corresponding to $\mathbf{x}_{(c)s_{(II)}}$ data of $S_{(II)}$ for D_{ns} areas which are non-sampled in $S_{(I)}$. Similarly, multipurpose weights are developed for D_{ns} areas non-sample for $S_{(I)}$. The proposed synthetic predictor EP.MP1 (denoted by EP.MP1.Syn) of mean for non-sampled area d is given by

$$\begin{aligned}\hat{\bar{Y}}_d^{EP.MP1.Syn} &= (\sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP1}\mathbf{x}_{(c)dj})^T\hat{\boldsymbol{\beta}} \\ &= (\hat{\bar{\mathbf{x}}}_{(II)d,out}^{MP1})^T\hat{\boldsymbol{\beta}};\ d = 1,..,D_{ns}, \end{aligned} \quad (13)$$

where $\hat{\bar{\mathbf{x}}}_{(II)d,out}^{MP1} = ((\sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP1}\mathbf{x}_{(c)dj})\hat{\boldsymbol{\beta}})^T\hat{\boldsymbol{\beta}};\ d = 1,..,D_{ns}$, D_0 denotes number of non-sampled areas in $S_{(I)}$. The synthetic version of EP.MP1 (denoted by EP.MP1. Syn) of mean for non-sampled area d is given by

$$\begin{aligned}\hat{\bar{Y}}_d^{EP.MP2.Syn} &= (\sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP2}\mathbf{x}_{(c)dj})^T\hat{\boldsymbol{\beta}} \\ &= (\hat{\bar{\mathbf{x}}}_{(II)d,out}^{MP2})^T\hat{\boldsymbol{\beta}};\ d = 1,..,D_{ns}, \end{aligned} \quad (14)$$

where, $\hat{\bar{\mathbf{x}}}_{(II)d,out}^{MP2} = (\sum_{j\in s_{(II)d}} \tilde{w}_{(II)dj}^{MP2}\mathbf{x}_{(c)dj})^T\hat{\boldsymbol{\beta}};\ d = 1,..,D_{ns}$, D_{ns} denotes number of non-sampled areas in small survey $S_{(I)}$.

3. EMPIRICAL EVALUATIONS OF THE PROPOSED ESTIMATORS

In this section, the performances of the proposed small area estimators are evaluated through two type of simulation studies. The first one is model based simulation where samples are drawn from a hypothetical populations, generated through statistical models and the second one is design based simulation where samples are drawn from a population generated through real survey data. Different small area mean estimators used for simulation studies are as follows: DIR, MBDE, EP1, EP2, EP3, EP4, EP.MP1, EP.MP2, EP.MP1.Syn, EP.MP2.Syn, SYN.EP2 and SYN.EP4 and these estimators are already discussed in section 2. The two performance criteria are used in the simulation studies, (i) Average percentage relative bias (RB) and (ii) Average percentage relative root mean squared error (RRMSE) and the expressions RB and RRMSE are given as:

$$RB(m) = \underset{d}{mean}\left\{\bar{m}_d^{-1}T^{-1}\sum_{t=1}^{T}(\hat{m}_{dt} - m_{dt})\right\}\times 100 \text{ and}$$

$$RRMSE(m) = \underset{d}{mean}\left\{\sqrt{T^{-1}\sum_{t=1}^{T}\left(\frac{\hat{m}_{dt} - m_{dt}}{m_{dt}}\right)^2}\right\}\times 100,$$

where the subscript d indexes the small areas and the subscript t indexes the T Monte Carlo simulations, with m_{dt} denoting the true area d mean at simulation t, with predicted value $\hat{m}_{dt}$ and the average true area d mean over T simulations is $\bar{m}_d = T^{-1}\sum_{t=1}^{T} m_{dt}$. The discussed expressions of RB and RRMSE are for model based simulation. In design based simulation formula is same as model based simulation but only difference that the term m_{dt} is replaced by m_d, since the population is assumed as fixed for design based simulation. The whole simulation process is independently repeated T times.

3.1 Model-based Simulation Study

In the model based simulations, we generate population data using linear mixed model particularly the random intercepts model of form

$$y_{kdj} = \alpha_k + \beta_{k1}x_{(c)dj} + \beta_{k2}x_{(uc)1dj} + \beta_{k3}x_{(uc)2dj} + u_d + e_{dj},$$
$$d = 1,....D; j = 1,..., N_d; k = 1,2,3. \quad (11)$$

The parameters of the model (11) are described in Table 2.

Table 2. Description of parameters set up of the model (11).

Response variable	Model Intercept	Auxiliary Variable Common to Both Surveys	Auxiliary Variable Collected in Large Survey Only	
		$x_{(c)}$	$x_{(uc)1}$	$x_{(uc)2}$
y_1	$\alpha_1 = 350$	$\beta_{11} = 1.5$	$\beta_{12} = 1.2$	$\beta_{13} = 1.4$
y_2	$\alpha_2 = 250$	$\beta_{21} = 1.2$	$\beta_{22} = 1.1$	$\beta_{23} = 1.5$
y_3	$\alpha_3 = 500$	$\beta_{31} = 1.7$	$\beta_{32} = 1.0$	$\beta_{33} = 1.6$

Here, $x_{(c)dj}$ and $x_{(uc)1dj}(d = 1,....D; j = 1,..., N_d)$ are generated using chi-squared distribution with degrees of freedom 20 and 10 respectively and $x_{(uc)2dj}(d = 1,....D; j = 1,..., N_d)$ is generated through normal distribution with zero mean and 36 variance. The random area effects u_d and individual effects e_{dj} are independently drawn from $N(0, \sigma_u^2)$ and $N(0, 94.09)$ distributions, respectively. The simulation studies when normality assumptions of random components in (11) hold, expressed as Simulation 1. We use two values of area effects variance σ_u^2 as 10.40 and 23.52, so that intra area correlation coefficients are $\rho = 0.10$ (simulation is denoted as Simulation1-I) and 0.20 (simulation is denoted as Simulation1-II), respectively. Total 25 small areas are considered and six combinations of area-specific sample sizes for the small and large samples are taken, $(n_{(I)d}, n_{(II)d}) = (2,25)$, $(2,50)$, $(4,25)$, and $(4,50)$, respectively. Here, the stratified random sampling design is used for sample selection procedure. It is assumed that the small survey $S_{(I)}$ collects variable of interest y_1 and auxiliary variable $x_{(c)}$. The large survey $S_{(II)}$ does not collects y_1 but collects auxiliary variables $x_{(c)}$, $x_{(uc)1}$ and $x_{(uc)2}$ as well as extra two response variables y_2 and y_3. The whole process of generating population data to calculation of small area means estimates is repeated independently $T = 2000$ times. Table 3 presents the average values of percentage relative biases and percentage relative root mean squared errors of the different small area estimators in Simulation 1.

Further, another simulation study is performed, assuming that the normality assumptions of random effect components u_d and e_{dj} in (11) does not hold denoted by Simulation 2. In the Simulation $2u_d$ and e_{dj} are generated through chi-squared distribution independently. The Simulation 2 is further divided into two subsections, Simulation 2-I and Simulation 2-II. In Simulation 2-I, u_d and e_{dj} are independently generated through chi-squared distribution with 1 and 5 degree of freedom, respectively. Similar way, in Simulation 2-II, u_d and e_{dj} are independently generated through chi-squared distribution with 2 and 5 degree of freedom, respectively. Table 4 presents the average values of percentage relative biases and percentage relative root mean squared errors of the different small area estimators under Simulation 2.

The non-sample area case is also considered in the model based simulation study denoted as Simulation 3. In Simulation 3, all conditions of Simulation 1 are hold, exception is the small survey $S_{(I)}$ collects data for 20 areas and 5 areas out of 25 areas are non-sampled, whereas the large survey $S_{(II)}$ collects data for all the 25 areas. The purpose of the simulation study is to observe the performance of the proposed estimators namely EP.MP1.Syn, EP.MP2.Syn under non-sampled areas. The results of Simulation 3 are summarized in Table 5.

In Table 3, our discussions focus on the two developed estimators namely, EP.MP1 and EP.MP2. It is observed that relative bias is not really an issue in Table 3, as the biases of all the estimators are almost negligible and of the same order and magnitude for all set of sample sizes. But, the difference in performances are identified with respect to average percentage RRMSE. The results in Table 3 shows that the estimators based on combining data from two independent surveys (EP2, EP4, EP.MP1 and EP.MP2) has smaller average percentage RRMSE value than the estimators based on small the survey $S_{(I)}$ data. Further, it is found that EP.MP1 and EP.MP2 has lesser average percentage RRMSE value than EP2 and EP4 estimators. The estimators (EP.MP1 and EP.MP2) that

combing data from two independent surveys, utilizing the extra set of variables of the large survey $S_{(II)}$ along with common auxiliary variable, further gain in efficiency with respect to smaller value of average percentage RRMSE. The results of the EP.MP1 and EP.MP2 are consistent over all the four sample size combinations in Table 3. Again, when the intraclass correlation coefficient value is increases from 0.1 to 0.2 that is Simulation 1-I to Simulation 2-II, the EP.MP1 and EP.MP2 outperform the other estimators. The proposed EP.MP1 and EP.MP2 estimators are perform at par for throughout in Table 3. Hence, we can conclude that the proposed EP.MP1 and EP.MP2 estimators are most out performer in Table 3.

Table 4 shows the average percentage relative bias and average relative RMSE of the different estimators under Simulation 2. Similar to Table 3 relative bias is not really an issue, as the biases of all the estimators are almost negligible and of the same order and

Table 3. Values of average percentage relative biases (RB) and average percentage relative root mean squared errors (RRMSE) of the different estimators exists under in Simulation 1.

$n_{(I)d}, n_{(II)d}$	Predictor	Simulation 1-I		Simulation 1-II	
		RB	RRMSE	RB	RRMSE
	DIR	-0.007	2.11	-0.007	2.11
	EP1	-0.007	1.62	-0.007	1.74
	EP2	-0.002	1.13	-0.002	1.73
2,25	MBDE	-0.010	2.11	-0.010	2.11
	EP3	-0.009	1.60	-0.009	1.31
	EP4	-0.007	1.12	-0.007	1.30
	EP.MP1	-0.002	0.88	-0.002	1.16
	EP.MP2	-0.002	0.88	-0.002	1.16
	DIR	0.007	2.10	0.007	2.10
	EP1	0.007	1.60	0.007	1.73
	EP2	0.001	0.92	0.001	1.71
4,50	MBDE	0.000	2.09	0.000	2.09
	EP3	-0.001	1.58	0.000	1.13
	EP4	0.001	0.92	0.001	1.12
	EP.MP1	0.001	0.86	0.001	1.11
	EP.MP2	0.001	0.86	0.001	1.11
	DIR	-0.001	1.63	-0.001	1.63
	EP1	-0.001	1.30	-0.001	1.40
	EP2	-0.013	1.06	-0.013	1.39
4,25	MBDE	-0.007	1.63	-0.007	1.63
	EP3	-0.007	1.28	-0.007	1.19
	EP4	-0.007	1.06	-0.007	1.18
	EP.MP1	-0.014	0.83	-0.014	1.07
	EP.MP2	-0.014	0.83	-0.014	1.07
	DIR	-0.001	1.64	-0.001	1.64
	EP1	-0.001	1.30	-0.001	1.41
	EP2	-0.003	0.84	-0.003	1.40
4,50	MBDE	-0.003	1.63	-0.003	1.63
	EP3	-0.003	1.29	-0.003	1.00
	EP4	-0.002	0.84	-0.002	1.00
	EP.MP1	-0.003	0.79	-0.003	0.99
	EP.MP2	-0.003	0.79	-0.003	0.99

Table 4. Values of average percentage relative biases (RB) and average percentage relative root mean squared errors (RRMSE) of the different estimators exists in Simulation 2.

$n_{(I)d}, n_{(II)d}$	Predictor	Simulation 2-I		Simulation 2-II	
		RB	RRMSE	RB	RRMSE
	DIR	0.000	1.71	-0.006	1.68
2,25	EP1	0.000	1.44	-0.006	1.47
	EP2	-0.006	0.86	-0.003	0.91
	MBDE	-0.006	1.69	-0.011	1.66
	EP3	-0.006	1.42	-0.012	1.44
	EP4	-0.004	0.85	-0.009	0.90
	EP.MP1	-0.005	0.46	-0.003	0.57
	EP.MP2	-0.005	0.46	-0.003	0.57
	DIR	-0.019	1.70	0.011	1.69
2,50	EP1	-0.019	1.43	0.011	1.47
	EP2	-0.001	0.67	0.001	0.73
	MBDE	-0.009	1.69	0.001	1.68
	EP3	-0.009	1.41	0.001	1.45
	EP4	-0.001	0.66	0.001	0.73
	EP.MP1	-0.001	0.44	0.001	0.55
	EP.MP2	-0.001	0.44	0.001	0.55
	DIR	-0.001	1.30	-0.008	1.30
4,25	EP1	-0.001	1.11	-0.008	1.15
	EP2	0.002	0.83	-0.002	0.88
	MBDE	0.000	1.29	-0.006	1.29
	EP3	0.000	1.10	-0.006	1.14
	EP4	0.002	0.82	-0.005	0.87
	EP.MP1	0.002	0.43	-0.002	0.54
	EP.MP2	0.002	0.43	-0.002	0.54
	DIR	0.009	1.31	-0.005	1.32
4,50	EP1	0.009	1.12	-0.005	1.16
	EP2	0.004	0.65	-0.006	0.70
	MBDE	0.005	1.30	-0.003	1.31
	EP3	0.004	1.11	-0.003	1.15
	EP4	0.005	0.64	-0.002	0.69
	EP.MP1	0.004	0.41	-0.006	0.51
	EP.MP2	0.004	0.41	-0.006	0.51

magnitude for all set of sample sizes. In Table 4 results are also similar to Table 3 with respect to RRMSE. The estimators based on combining data from two independent surveys (EP2, EP4, EP.MP1 and EP.MP2) has smaller average percentage RRMSE value than the estimators based on small the survey $S_{(I)}$ data only. Further, it is found that EP.MP1 and EP.MP2 has lesser average percentage RRMSE value than EP2 and EP4 estimators. Hence, the performances of the developed estimators are unaltered when the normal distribution of random effect components of (11) are replaced by chi-squared distribution. Table 4 shows that the performances of EP.MP1 and EP.MP2 are consistently hold for all the four sample size combinations as well as for bot h Simulation 2-I and Simulation 2-II similar to Table 3. Hence, we can conclude that the proposed EP.MP1 and EP.MP2 estimators are most out performer in Table 4.

In Table 5 shows the results of Simulation 3 for sample as well as non-sample areas. Here, 5 areas of the small survey $S_{(I)}$ are taken as non-sample areas and rest 20 areas are in sampled. In Table 5, the results for sample areas as well as non-sample areas are averaged over 20 sample areas and 5 non-sample areas, respectively. The results for sample areas are similar to the results of Table 3 and Table 4. The estimators (SYN.EP2, SYN.EP4, EP.MP1.Syn and EP.MP2.Syn) applied to non-sample areas are performing more or less similar to sample areas. The results obtained in Table 5 shows that EP.MP1. Syn and EP.MP2.Syn estimators are noteworthy. The values of average percentage RRMSE of the EP.MP1. Syn and EP.MP2.Syn are smaller than SYN.EP2 and SYN.EP4 estimators. Further, the performances of two proposed synthetic estimators, EP.MP1.Syn and EP.MP2.Syn are at par. The performances of EP.MP1. Syn and EP.MP2.Syn are consistent for all sample size pairs as well as for both Simulation 3-I and Simulation 3-II. This clearly shows an evidence that the proposed synthetic estimator EP.MP1.Syn and EP.MP2.Syn has potential to generate the reliable estimates for non-sample areas.

3.2 Design-based Simulation Study

Design based simulation studies are conducted to support the performances of the proposed small area mean estimators for real survey data. The simulation studies are conducted using the data of Australian

Table 5. Values of average percentage relative biases (RB) and average percentage relative root mean squared errors (RRMSE) of the different estimators in Simulation 1 ($S_{(I)}$: 20 sample + 5 non-sample areas; $S_{(II)}$: all 25 sample areas).

$(n_{(I)d}, n_{(II)d})$	Areas	Predictor	Simulation 3				
			RB	RRM-SE	$(n_{(I)d}, n_{(II)d})$	RB	RRM-SE
2, 25	Sampled	DIR	-0.008	2.28	4, 25	-0.001	1.76
		EP1	-0.008	1.75		-0.001	1.40
		EP2	-0.002	1.22		-0.016	1.14
		MBDE	-0.012	2.28		-0.008	1.76
		EP3	-0.011	1.73		-0.008	1.38
		EP4	-0.008	1.21		-0.008	1.14
		EP.MP1	-0.002	0.95		-0.017	0.90
		EP.MP2	-0.002	0.95		-0.017	0.90
	Non-sampled	SYN.EP2	-0.002	1.46		-0.016	1.37
		SYN.EP4	-0.008	1.45		-0.008	1.37
		EP.MP1.Syn	-0.002	1.05		-0.017	1.08
		EP.MP2.Syn	-0.002	1.05		-0.017	1.08
2, 50	Sampled	DIR	0.008	2.27	4, 50	-0.001	1.77
		EP1	0.008	1.73		-0.001	1.40
		EP2	0.001	0.99		-0.004	0.91
		MBDE	0.000	2.26		-0.004	1.76
		EP3	-0.001	1.71		-0.004	1.39
		EP4	0.001	0.99		-0.002	0.91
		EP.MP1	0.001	0.93		-0.004	0.85
		EP.MP2	0.001	0.93		-0.004	0.85
	Non-sampled	SYN.EP2	0.001	1.09		-0.004	1.09
		SYN.EP4	0.001	1.09		-0.002	1.09
		EP.MP1.Syn	0.001	1.02		-0.004	1.02
		EP.MP2.Syn	0.001	1.02		-0.004	1.02

Agricultural Grazing Industry Survey (AAGIS), conducted by the Australian Bureau of Agricultural and Resource Economics in the year 1995-96. The original sample is of size 759 farms from 12 regions (the areas of interest). Following the process of Islam and Chandra (2019), the original sample is used to generate the population of size 39562 farms. The sample size for the large survey are taken as 759 farms (original sample) and three different area specific sample sizes 5, 10 and 15 are drawn for the small survey. The whole process from sample selection to estimation is repeated 2000 times. The variable of interest is assumed as the total cash costs (TCC), and the aim is to estimate region specific mean TCC value. In small survey that collects TCC as well as

auxiliary variables namely, number of closing stock-beef, number of closing stock-sheep and quantity of harvested wheat. The large survey that does not collects TCC but collects auxiliary variables common to the small survey. In addition, the large survey collects extra set of variables where, total cash receipts (TCR) and Debt taken as response variables and land area is taken as auxiliary variables, not common to small survey. The results of the design-based simulations are summarized in Table 6.

Table 6. Values of average percentage relative biases (RB) and average percentage relative root mean squared errors (RRMSE) of the different estimators under design based simulations using the AAGIS data.

Predictor	$n_{(1)d}=5$		$n_{(1)d}=10$		$n_{(1)d}=15$	
	RB	**RRMSE**	**RB**	**RRMSE**	**RB**	**RRMSE**
DIR	1.36	57.35	-0.30	37.29	0.19	27.59
EP1	4.58	51.96	2.75	32.67	2.63	24.61
EP3	4.47	51.06	2.62	32.64	2.17	24.33
MBDE	1.25	60.17	-0.36	38.01	-0.10	28.37
EP2	0.90	35.52	0.71	25.90	0.66	21.32
EP4	0.91	35.26	0.67	25.85	0.52	21.18
EP.MP1	0.52	32.70	0.36	24.22	0.30	20.00
EP.MP2	0.52	32.70	0.36	24.22	0.30	20.00

In Table 6, the results reveal that the EP.MP1 and EP.MP2 has lowest relative bias followed by EP4 and EP2 estimators. Further, the results of Table 6 with respect to average percentage RRMSE clearly support the results of model based simulations that EP.MP1 and EP.MP2 has minimum average percentage RRMSE followed by EP2 and EP4 estimators. Table 6 results provide an encouraging performance of the proposed EP.MP1 and EP.MP2 estimators. The results set out in Table 6 support the conclusion that the use of extra set of variables in combining data from two surveys along with common auxiliary variables further improves small area estimation. Hence, the proposed EP.MP1 and EP.MP2 emerging as the best performing of the methods that we investigated in the empirical evaluations.

4. CONCLUDING REMARKS

We developed SAE method by combining data from two independent surveys. The empirical results, based on simulated data as well as on real survey data, clearly indicate that combining information from two surveys can bring significant gains in SAE efficiency and the incorporation of extra set of variables of the large survey that is not common to small survey further gain in efficiency in small area estimation. The conclusions are same for estimate of non-sample areas using synthetic version of EP.MP1 and EP.MP2 as well as the situation when the normality assumption of random effect components of linear mixed model are replaced by chi-squared distribution.

ACKNOWLEDGEMENT

The first author gratefully acknowledges the financial support provided by a Ph.D. scholarship from the Ministry of Social Justice & Empowerment, Government of India.

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

REFERENCES

Battese, G.E., Harter, R.M. and Fuller, W.A., 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.

Chandra, H. and Chambers, R., 2011. Small area estimation under transformation to linearity. *Survey Methodology*, **37(1)**, 39-51.

Chandra, H. and Chambers, R., 2009. Multipurpose weighting for small area estimation. *Journal of Official Statistics*, **25(3)**, 379-395.

Chandra, H., Salvati, N., and Chambers, R., 2007. Small area estimation for spatially correlated populations-a comparison of direct and indirect model-based estimators. *Statistics in Transition*, **8**, 331-350.

Chandra, H., Salvati, N, Chambers, R. and Tzavidis, N., 2012. Small area estimation under spatial non-stationarity. *Computational Statistics and Data Analysis*, **56(12)**, 2875-2888.

Chandra, H., Sud, U.C. and Gharde, Y., 2015. Small area estimation using estimated population level auxiliary data. *Communications in Statistics-Simulation and Computation*, **44(5)**, 1197-1209.

Cochran, W.G., 1977. *Sampling Techniques*, 3rd edition, New York: John Wiley and Sons.

Datta, G.S. and Lahiri, P., 2000. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 613-627.

Elliott, M.R. and Davis, W.W., 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.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72(358)**, 320-338.

Hidiroglou, M.A., 2001. Double sampling. *Survey Methodology*, **27(2)** 143-154.

Islam, S. and Chandra, H., 2019. Small area estimation combining data from two surveys. *Communication in Statistics-Simulations and Computation.* In Press. (https://doi.org/10.1080/03610918.2019.1588308)

Islam, S., Chandra, H., Aditya, K. and Lal, S.B., 2018. Small area estimation under a spatial model using data from two surveys. *International Journal of Agricultural and Statistical Sciences*, **14(1)**, 231-237.

Kim, J.K., Park, S. and Kim, S., 2015. Small area estimation combining information from several sources. *Survey Methodology*, **41**, 21-36.

Kim, J.K. and Rao, J.N.K., 2012. Combining data from two independent surveys: a model assisted approach. *Biometrika*, **99**, 85-100.

Lohr, S. and Prasad, N.G.N., 2003. Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics*, **31**, 383-396.

Lohr, S. and Rao, J.N.K., 2006. Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, **101**, 1019-1030.

Marker, D. A., 2001. Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology*, **27(2)**, 183-188.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J. and Thompson, S.G., 2011. Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A*, **174**, 31-50.

Maples, J.J., 2017. Improving small area estimates of disability: combining the American

community survey with the survey of income and program participation. *Journal of Royal Statistical Society A,* **180(4)**:1211-1227.

McCulloch, C.E. and Searle, S.R., 2001. *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.

Merkouris, T., 2010. Combining information from multiple surveys by using regression for efficient small area estimation. *Journal of the Royal Statistical Society B*, **68**, 509-521.

Merkouris, T., 2004. Combining independent regression estimators from multiple surveys. *Journal of American Statistical Association*, **99**, 1131-1139.

Moriarity, C. and Scheuren, F., 2001. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, **17**, 407-422.

Rao, J.N.K. and Molina, I., 2015. *Small Area Estimation*. John Wiley & Sons. Inc., New Jersey, 2nd edition.

Renssen, R.H. and Nieuwenbroek, N., 1997. Aligning estimates for common variables in two or more sample surveys. *Journal of American Statistical Association*, **92**, 368-375.

Royall, R.M. and Cumberland, W.G., 1978. Variance estimation in finite population sampling. *Journal of the American Statistical Association*, **73**, 351-358.

Särndal, C.E., Swensson, B. and Wretman, J.H., 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schenker, N. and Raghunathan, T., 2007. Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, **26**, 1802-1811.

Wu, C., 2004. Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics*, **32**, 15-26.

Ybarra, L.M.R., and Lohr, S.L., 2008. Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919-931.

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.