Journal of Quantitative Economics
https://doi.org/10.1007/s40953-020-00226-8

ORIGINAL ARTICLE

District-Level Estimates of Poverty Incidence for the State of West Bengal in India: Application of Small Area Estimation Technique Combining NSSO Survey and Census Data

Hukum Chandra[1]

Accepted: 16 December 2020

Abstract
Despite having long term efforts, poverty is an important and persistent social issue in India. Existing data based on socio-economic surveys produce state and nationally representative poverty estimates but cannot be used directly to generate reliable disaggregate or local level estimates. The state and national level estimates often mask the variations at the local level which in turn restricts the effective implementation of policies related to poverty alleviation locally within and between administrative units. This paper uses the Household Consumer Expenditure Survey data of NSSO and link with the Population Census data to produce the reliable district-level estimates of poverty incidence in the rural areas of West Bengal in India. In particular, small area estimation (SAE) method is explored to generate reliable district-level poverty estimates. The results clearly indicate that the district-level estimates generated by model-based SAE method are precise and representative. A map showing how poverty incidence varies by district across the State of West Bengal is also produced. The estimates generated from this research are useful for meeting the data requirements for policy research and strategic planning by different international organizations and by Departments and Ministries in the Government of India.

Keywords Anti poverty · Poverty · Welfare · Well being

✉ Hukum Chandra
hchandra12@gmail.com

[1] ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012, India

Published online: 15 January 2021

Introduction

The sustainable development goals (SDGs), officially known as transforming our world: the 2030 agenda for sustainable development is a set of 17 "global goals". These are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. These goals build on the successes of the Millennium Development Goals. The SDG 1 is to "End poverty in all its forms everywhere". Poverty is more than the lack of income and resources to ensure a sustainable livelihood. Its manifestations include hunger and malnutrition, limited access to education and other basic services, social discrimination and exclusion as well as the lack of participation in decision-making. Economic growth must be inclusive to provide sustainable jobs and promote equality. Globally, the number of people living in extreme poverty has declined by more than half from 1.9 billion in 1990. However, 836 million people still live in extreme poverty. About one in five persons in developing regions lives on less than $1.25 per day. Southern Asia and sub-Saharan Africa are home to the overwhelming majority of people living in extreme poverty. High poverty rates are often found in small, fragile and conflict-affected countries. The all India poverty head count ratio has been brought down from 47% in 1990 to 21% in 2011–2012, nearly halved. Availability of reliable and timely statistics is very crucial for monitoring the progress of SDGs. West Bengal, after Bihar, the second-most densely populated state of India, is facing a poverty issue due to low development in agricultural and industrial sectors over time (Mitra 2016). For effective development in West Bengal in sense of poverty eradication, there is a crucial requirement to develop a focused scheme for poverty eradication. Timely and reliable disaggregated level statistics is essential for effective planning, implementation and monitoring of various government strategy. The disaggregate level statistics is also must for identifying the area more in requirement and for developing focused and target oriented intervention programs.

In India, though a lot of data are collected, processed and published for the country as a whole or for individual states, not much disaggregation of the data for sub-state level is done. The national sample survey office (NSSO) surveys are the main source of official statistics in India. A range of invaluable data at the macro level (e.g. state and national level) is produced through these surveys. However, the NSSO data cannot be used directly to produce reliable estimates at the micro level (e.g. district or further disaggregate level) due to small sample sizes. There is a rapidly growing demand of such disaggregate level statistics in India as the country is moving from centralized to more decentralized planning system. Therefore, the appropriate strategy, fund distribution and also monitoring of various plans is likely to affected due to the availability of reliable estimates at disaggregate level. At the same time it is also true that conducting district specific surveys is going to be very trivial and costly as well as time consuming job. Using the state level survey data to derive the estimates at district or further smaller level may result in small sample sizes leading to very unstable estimates (Rao and Molina 2015). A domain (or area) is regarded as "small" if the domain specific sample is insufficient to provide direct survey estimates with adequate precision and reliability. We refer such domain or

area as small domain or small area. In order to produce estimates for small areas with adequate levels of precision, it is standard to use model based estimation which utilizes information from outside areas with similar characteristics to the area of interest. The information from respondents who are outside the geographical area and other geographical characteristics are incorporated through the use of a statistical model. This technique is abbreviated as small area estimation (SAE) techniques which is a tool of the statistical sciences that combines survey sampling and finite population inference with statistical models. Most of the methods that have been developed for SAE are included in Rao and Molina (2015).

Based on the level of auxiliary information available from secondary data sources, SAE methods are categorized as based on area or aggregated level and unit level small area models. Area or aggregated level small area models are used when auxiliary information are available only at area level. They relate small area direct survey estimates to area-specific covariates (Fay and Herriot 1979). Unit level small area models, firstly suggested by Battese et al. (1988), relate the unit values of a study variable to unit-specific covariates. In this paper, we consider the area level version of small area model since covariates are available only at the area level. Here, variable of interest is discrete, in particular binary variable, therefore the model under consideration is area level logistic linear mixed model for SAE. We use SAE techniques to obtain model-based estimates of proportion of poor households (i.e. incidence of poverty) at district level in the State of West Bengal in India. This article, in particular, provides strategies how the existing large scale Household Consumer Expenditure Survey and census data can be combined to derive reliable small area estimates for various policy relevant parameters.

In India, several researchers have deliberated on disaggregate level poverty estimation from Household Consumer Expenditure Survey data. Chaudhuri and Gupta (2009) illustrated district level poverty estimation using Household Consumer Expenditure Survey 2004–2005 data of NSSO. But, this study has indicated the limitation of large sampling variability of the estimates for some districts due to inadequacy of sample size. Coondoo et al. (2011) suggested an approach for generating micro level poverty indicators for two states of India namely, West Bengal and Madhya Pradesh using Household Consumer Expenditure Survey data. This approach is based on subgroup decomposable property of poverty measure where sub-state level estimates are obtained by solving a system of linear equations. Major demerit of this approach is that it belongs to the class of synthetic indirect method of SAE. Synthetic estimators are known to be biased due to homogeneity assumption between the domains of interest (Rao and Molina 2015). Chauhan et al. (2016) elaborated intra and inter-regional disparities in poverty and inequality using three quinquennial rounds of Household Consumer Expenditure Survey data of NSSO over two decades (1993–2012). Mohanty et al. (2016) described similar type of analysis where the poverty estimates are provided at district level within the region. However, in both of these studies poverty indicators are estimated fitting regression based fixed effect model. As a consequent, these approaches fail to capture dissimilarities across areas and this limitation is particularly being addressed by SAE approach which takes into account the variability between areas using the random area-specific effects in the model. The rationale behind using SAE approach in this

article specifically motivated from the issue of glittering poverty in most of the parts of rural India and ineffectiveness of traditional direct estimation and synthetic estimation approach invoked in various studies stated above in measuring the poverty proportions at disaggregate or local levels. The rest of the article is organised as follows. "Data Sources and Model Specification" describes the data used for the analysis and model fitting. "Methodology" provides a brief overview of the methodology used for analysis. "Results and Discussions" discusses the diagnostic procedures for examining the model assumptions, validating the small area estimates and describes the results. "Concluding Remarks" finally set out the main conclusions.

Data Sources and Model Specification

This section presents the basic sources of data i.e. survey data and the auxiliary data used to estimate the poverty incidence at district level. The incidence of poverty is defined as proportion of poor households, i.e. head count ratio (HCR). The HCR is poverty indicator or incidence measures the frequency of households under poverty line. Two types of variables are require for SAE analysis, the variable of interest for which small area estimates are required is drawn from the Household Consumer Expenditure Survey 2011–2012 of NSSO for rural areas of the state of West Bengal. The sampling design used in the Household Consumer Expenditure Survey 2011–2012 of NSSO for rural areas is stratified multi-stage random sampling design with district as strata, villages as first stage units (FSU) and households as the second (or ultimate) stage units (SSU). The list of 2001 census villages (or villages) are used as the sampling frame for selection of FSUs. The sample of FSUs (villages) are selected using probability proportional to size with replacement sampling, size being the population of the village as per Census 2001. In case of large FSUs, one intermediate stage of sampling is the selection of two 'hamlet-groups' from each FSU. In particular, if the population of the selected FSU is equal to 1200 or more, it is divided into a suitable number of 'hamlet-groups'. In such cases, two 'hamlet-groups' are selected, the first with maximum percentage share of population and the second from the remaining hamlet-groups by simple random sampling without replacement (SRSWOR). Subsequently, listing and selection of the households is done independently in the two selected hamlet-groups. The households listed in the selected FSU/hamlet-group are stratified into three second stage strata (SSS). The sample households (SSUs) are selected by SRSWOR from each SSS. A total of 3568 households were surveyed from the 18 districts of the West Bengal. Since West Bengal having 19 districts in the map of India, but this study has been carried out only take for 18 districts of West Bengal, because one of the district of West Bengal i.e. Kolkata come out under the well-developed area. In Kolkata the rural area does not exists, so in this study we does not consider Kolkata for the measure of incidence of poverty. The district-wise sample size varied from minimum 64 to maximum 320 with average of 198 (Table 2). Therefore, it is difficult to occur reliable estimates and their standard errors at district level. In such situations we should use SAE procedure to develop the reliable estimates for the districts having small sample data;

for more detail we can see, Pfeffermann (2002) and Rao and Molina (2015). The target variable used for the study is poor households. The poverty line has been used to identify whether given household is poor or not. A household having monthly per capita income consumer expenditure below the state's poverty line (Rs. 778) is categorized as poor household. The poverty line used in this study is same as those of year 2011–2012, given by then planning commission, Govt. of India.

The auxiliary (covariates) variables used in this analysis are drawn from the Population Census 2011. These auxiliary variables are only available as counts at district level, and these variables are reported separately for rural and urban areas of the district. There are approximately 50 such covariates available from Population Census 2011 to consider for small area modelling, and these are related to rural areas only, not for the entire district. We therefore carried out a preliminary data analysis in order to define appropriate covariates for SAE modelling, using principal component analysis (PCA) to derive composite scores for selected groups of variables. We first divide the selected number of auxiliary variables in three groups and then develop composite scope or index for each of these groups of auxiliary variables using PCA. The first composite or PCA score (denoted by g_1) is based on gender-wise literacy rate and gender-wise proportion of worker population. The first principal component or composite score for first group of PCA (g_{11}) explains about 55.65% of the variation in the dataset while adding the second component (g_{12}) explains about 83%. The second set of composite or PCA score (g_2) is based on following variables; gender-wise proportion of main worker population, gender-wise proportion of main cultivator population and gender-wise proportion of main agricultural labourers population. The first principal component (g_{21}) for second set of PCA explains about 45% of the variation in the dataset, while adding the second component (g_{22}) explains about 69%. This further enhanced to about 81% when the third component (g_{23}) is included. Finally, the third set of PCA or composite score (g_3) is derived from gender-wise proportion of marginal cultivation population and gender-wise proportion of marginal agriculture labourers population. The first principal component (g_{31}) for third set of PCA explains about 57.34% of the variation in the dataset, while adding the second component (g_{32}) explains 71.89%. The composite scores (i.e. g_{11}, g_{12}, g_{21}, g_{22}, g_{23}, g_{31} and g_{32}) obtained from three groups of variables are then considered as auxiliary variables. We fit a generalized linear model between district-specific direct survey estimates of proportion of poor households as response variable and set of composite scores (i.e. g_{11}, g_{12}, g_{21}, g_{22}, g_{23}, g_{31} and g_{32}) as auxiliary variables. This model is fitted using the *glm()* function in R and specifying the family as "*Binomial*" and the district-specific sample sizes as the weight. The purpose here is build a good explanatory and predictive model based on the available auxiliary data. Table 1 presents the results of the models fitted using different combination of auxiliary variables. We consider the best fitting generalized linear model using the values of AIC and residual deviance.

The results in Table 1 show that the model with three significant auxiliary variables g_{11}, g_{21}, g_{22} with residual deviance and AIC values as 83.35 and 179.33 respectively is better working model (i.e. Model 5). These three auxiliary variables $g_{11,}$ g_{21} and g_{22} are then used in SAE. In the chosen model, the auxiliary variable g_{11} (i.e. the composite index derived from gender-wise literacy rate and gender-wise proportion

Table 1 Model parameters for the different generalised linear models for proportion of poor households

Model 1		Model 2		Model 3		Model 4		Model 5	
Intercept	−1.794***	Intercept	−1.790***	Intercept	−1.779***	Intercept	−1.783***	Intercept	−1.778***
g_{11}	−0.370*	g_{11}	−0.477**	g_{11}	−0.410**	g_{11}	−0.322**	g_{11}	−0.284**
g_{12}	−0.090	g_{21}	0.306*	g_{21}	0.320*	g_{21}	0.226*	g_{21}	0.212*
g_{21}	0.191	g_{22}	−0.261**	g_{22}	−0.212**	g_{22}	−0.196**	g_{22}	−0.181**
g_{22}	−0.217*	g_{23}	0.109^{+}	g_{23}	0.093^{+}	g_{31}	−0.042		
g_{23}	0.108^{+}	g_{31}	−0.081	g_{31}	−0.020				
g_{31}	−0.079	g_{32}	0.099						
g_{32}	0.107								
Residual deviance	76.89		79.08		80.14		83.00		83.35
AIC	180.87		181.07		180.12		180.98		179.33

*$p<0.05$; **$p<0.01$; ***$p<0.001$

$^{+}p<0.10$

of worker population) is strongly significant as a predictor for proportion of poor household. The negative coefficient of g_{11} is justified since poverty rate decreases with increases in the literacy rate and worker population. The districts with higher scores are more likely to have lower rates in terms of poverty incidence. Further, first and second composite scores (i.e. g_{21}, g_{22}) of g_2 are also significant. Therefore, we use the model 5, for reasons of simplicity, interpretability and model efficiency since adding more variables introduces instability into the SAE, see Pfeffermann (2002) and Ybarra and Lohr (2008).

Methodology

In this section we illustrate the theoretical framework used to produce small area estimates of poverty incidence and their measure of precision. The details presented here are followed from Chandra et al. (2011) and Johnson et al. (2010) and references therein. Let us assume a finite population U of size N and a sample s of size n is drawn from this population with a given survey design. We assume that this population is consist of D small areas (or simply areas) $U_d (d = 1, \ldots, D)$ such that $U = \bigcup_{d=1}^{D} U_d$ and $N = \sum_{d=1}^{D} N_d$. Throughout, we use a subscript d to index the quantities belonging to area $d (d = 1, \ldots, D)$, where D is the number of areas in the population. Here, $D = 18$ districts of the West Bengal are small areas of interest. The subscript s and r are used for denoting the quantities related to the sample and non-sample parts of the population. So that n_d and N_d represent the sample and population (i.e., number of households in sample and population) sizes in district d, respectively. Let s_d denotes the part of sample from area d such that $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. Let y_{di} denotes the value of target variable of interest y for unit i in area d. Let assume that the variable of interest y is binary and the target is the estimation of population counts $y_d = \sum_{i \in U_d} y_{di}$ or population proportions $P_d = N_d^{-1} \left(\sum_{i \in U_d} y_{di} \right)$ in area d. The design-based direct survey estimator of proportion of poor household in area d is given by $\hat{p}_d^{Direct} = \left(\sum_{i \in s_d} w_{di} \right)^{-1} \left(\sum_{i \in s_d} w_{di} y_{di} \right)$, where w_{di} is the survey weight associated with household i in area d. Assuming that joint inclusion $1/w_{di,d'j} = 0$ for $d \neq d'$ or $i \neq j$, the estimate of design variance of $\hat{p}_d^{Direct}$ is $v(\hat{p}_d^{Direct}) = \left(\sum_{i \in s_d} w_{di} \right)^{-2} \left\{ \sum_{i \in s_d} w_{di} (w_{di} - 1)(y_{di} - \hat{p}_d^{Direct})^2 \right\}$, see for example Särndal et al. (1992).

Let us denote by y_{sd} and y_{rd} the sample and non-sample counts of poor households in area (or district) d. The sample count y_{sd} has a Binomial distribution with parameters n_d and π_d, denoted by $y_{sd} \sim Bin(n_d, \pi_d)$, where π_d is the probability of a poor household in area d, often termed as the probability of a 'success'. Similarly, $y_{rd} \sim Bin(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent Binomial variables with π_d being a common success probability. Here we assume that only aggregated level data is available for the modelling. For example, from survey data y_{sd} and from secondary data sources $\mathbf{x}_d$ the p-vector of the covariates are available for area d. Following Chandra et al. (2001) and Johnson et al. (2010), the model linking the probabilities of success π_d with the covariates $\mathbf{x}_d$ is the generalised linear

mixed model (GLMM) with logit link function, i.e. logistic linear mixed model given by

$$logit(\pi_d) = \ln\left\{\pi_d(1-\pi_d)^{-1}\right\} = \eta_d = \mathbf{x}_d^T\boldsymbol{\beta} + u_d, \tag{1}$$

where $\boldsymbol{\beta}$ is the p-vector of regression coefficient often known as fixed effect parameters and u_d is the area-specific random effect that accounts for between area dissimilarity beyond that explained by the auxiliary variables included in the in the fixed part of the model. We assume that u_d's are independent and normally distributed with mean zero and constant variance φ. Under model (1), we get, $\pi_d = \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)\left\{1 + \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)\right\}^{-1}$. It is noteworthy that model (1) relates the area level proportions to area level covariates. This type of model is often referred to as 'area-level' model in SAE terminology, see for example Fay and Herriot (1979) and Rao (2003). Area level model was originally proposed by Fay and Herriot (1979). The Fay and Herriot method for SAE is based on area level linear mixed model and their approach is applicable to a continuous variable. This model is not applicable for discrete. The model (1) on the other hand is a special case of a GLMM with logit link function and suitable binary variable. Here,

$$y_{ds}|\, u_d \sim \text{Binomial}\left(n_d, \frac{\exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}\right) \text{and}$$

$$y_{dr}|\, u_d \sim \text{Binomial}\left(N_d - n_d, \frac{\exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}\right).$$

This leads to

$$E\left(y_{sd}\,|\,u_d\right) = n_d \frac{\exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)} \text{ and } E\left(y_{rd}\,|\,u_d\right) = \left(N_d - n_d\right) \frac{\exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)}.$$

Note that the estimation of fixed effect parameters $\boldsymbol{\beta}$ and area specific random effects u_d's uses the data from all small areas or districts. We use an iterative procedure that combines the penalized quasi-likelihood (PQL) estimation of $\boldsymbol{\beta}$ and $\mathbf{u} = (u_1, \ldots, u_D)^T$ with restricted maximum likelihood (REML) estimation of ϕ to estimate these unknown parameters. Detailed description of the approach can be followed from Breslow and Clayton (1993), Manteiga et al. (2007) and Saei and Chambers (2003). The total population counts, i.e. the total number of poor households in district d can be expressed as $y_d = y_{sd} + y_{rd}$, where the first term y_{sd}, the sample count is known whereas the second term y_{rd}, the non-sample count, is unknown. Therefore, a plug-in empirical predictor (EP) of the population count in area d is defined as

$$\hat{y}_d^{EP} = y_{sd} + \hat{E}\left(y_{rd}\,|\,u_d\right) = y_{sd} + \left(N_d - n_d\right) \frac{\exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)}. \tag{2}$$

An estimate of the poverty incidence in an area, i.e. the proportion of poor household in small area d is obtained as

$$\hat{p}_d^{EP} = \frac{\hat{y}_d^{EP}}{N_d} = \frac{1}{N_d}\left[y_{sd} + \left(N_d - n_d\right)\frac{\exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1+\exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)}\right]. \tag{3}$$

For area with zero sample sizes (i.e. non-sampled areas), the conventional approach for estimating area proportions or counts is synthetic estimation, based on a suitable GLMM fitted to the data from the sampled areas. Under model (1), for non-sampled areas, the synthetic type predictor of total population count for small area d is obtained as $\hat{y}_d^{SYN} = N_d\left[\exp(\mathbf{x}_{d,out}^T\hat{\boldsymbol{\beta}})\left\{1+\exp(\mathbf{x}_{d,out}^T\hat{\boldsymbol{\beta}})\right\}^{-1}\right]$, where $\mathbf{x}_{d,out}$ denote the vector of covariates associated with non-sampled area d. Similarly, the proportion of poor household, in an area is $\hat{p}_d^{SYN} = \exp(\mathbf{x}_{d,out}^T\hat{\boldsymbol{\beta}})\left\{1+\exp(\mathbf{x}_{d,out}^T\hat{\boldsymbol{\beta}})\right\}^{-1}$.

The mean squared error (MSE) estimates are computed to assess the reliability of estimates and also to construct the confidence interval for the small area estimates. Estimation of the MSE of the EP (3) is followed from development reported in Saei and Chambers (2003), Manteiga et al. (2007), Johnson et al. (2010) and references therein. Let us denote by $\hat{\mathbf{V}}_{sd} = diag\left\{n_d\hat{p}_d^{EP}(1-\hat{p}_d^{EP})\right\}$ and $\hat{\mathbf{V}}_{rd} = diag\left\{(N_d-n_d)\hat{p}_d^{EP}(1-\hat{p}_d^{EP})\right\}$, the diagonal matrices defined by the corresponding variances of the sample and non-sample part respectively. Similarly,$\mathbf{A} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}$, $\mathbf{B} = \left\{diag(N_d^{-1})\right\}\left\{\hat{\mathbf{V}}_{rd}\mathbf{X} - \mathbf{A}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\right\}$ and $\hat{\boldsymbol{\Sigma}} = \left(\hat{\phi}^{-1}\mathbf{I}_D + \hat{\mathbf{V}}_{sd}\right)^{-1}$, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_D^T)^T$ is a $D \times p$ matrix and $\mathbf{I}_D$ is an identity matrix of order D. We further define $\hat{\mathbf{V}}_{(1)} = \left\{\mathbf{X}^T\hat{\mathbf{V}}_{sd}\mathbf{X} - \mathbf{X}^T\hat{\mathbf{V}}_{sd}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\right\}^{-1}$ and $\hat{\mathbf{V}}_{(2)} = \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\hat{\mathbf{V}}_{(1)}\mathbf{X}^T\hat{\mathbf{V}}_{sd}^T\hat{\boldsymbol{\Sigma}}$. With these notations, assuming model (1) holds, an approximate MSE estimate of (3) is given by

$$mse(\hat{p}_d^{EP}) = m_1(\hat{\phi}) + m_2(\hat{\phi}) + 2m_3(\hat{\phi}). \tag{4}$$

The first two components m_1 and m_2 constitute the largest part of the overall MSE estimates in Eq. (4). These are the MSE of the best linear unbiased predictor type estimator when ϕ is known, see Saei and Chambers (2003). The third component m_3 of the MSE estimate (4) is the variability due to the estimate of ϕ. The analytical expression of these components of MSE estimate are $m_1(\hat{\phi}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}\mathbf{A}^T$, $m_2(\hat{\phi}) = \mathbf{B}\hat{\mathbf{V}}_{(1)}\mathbf{B}^T$, and $m_3(\hat{\phi}) = trace\left(\hat{\nabla}_i\hat{\boldsymbol{\Sigma}}^{+}\hat{\nabla}_j^T v(\hat{\phi})\right)$, where $\hat{\boldsymbol{\Sigma}}^{+} = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}^T$. Here $v(\hat{\phi})$ is the asymptotic covariance matrix of the estimate of variance component $\hat{\phi}$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\phi}$. This this also depends upon whether we are use ML or REML estimate for $\hat{\phi}$. In this paper, we use REML estimate for variance component $\hat{\phi}$, and then $v(\hat{\phi}) = 2\left(\hat{\phi}^{-2}(D-2a_1) + \hat{\phi}^{-4}a_{11}\right)^{-1}$ with $a_1 = \hat{\phi}^{-1}\,trace(\hat{\mathbf{V}}_{(2)})$ and $a_{11} = trace(\hat{\mathbf{V}}_{(2)}\hat{\mathbf{V}}_{(2)})$. Let us write $\Delta = \mathbf{A}\hat{\boldsymbol{\Sigma}}$ and $\hat{\nabla}_i = \partial(\Delta_i)/\partial\phi\Big|_{\phi=\hat{\phi}} = \partial(\boldsymbol{A}_i\hat{\boldsymbol{\Sigma}})/\partial\phi\Big|_{\phi=\hat{\phi}}$, where $\boldsymbol{A}_i$ is the ith row of the matrix $\mathbf{A}$. The

MSE estimates of the synthetic predictor $\hat{p}_d^{SYN}$ are a special case of (4) when $n_d = 0$ and it is given by $mse(\hat{p}_d^{SYN}) = \left[diag\left\{\hat{p}_d^{SYN}(1-\hat{p}_d^{SYN})\right\}\right]\hat{\phi}\mathbf{I}_D\left[diag\left\{\hat{p}_d^{SYN}(1-\hat{p}_d^{SYN})\right\}\right]^T$.

Results and Discussions

In SAE application, generally two types of diagnostics measures are suggested and used, the model diagnostics and the diagnostics for the small area estimates, see for example Brown et al. (2001). The model diagnostics are applied to verify the assumptions of underlying model, i.e. how well working model is fitted to data. On the other hand, the small area estimate diagnostics provide an indication of the reliability (and validity) of the model-based small area estimates. In model (1), the random district specific effects are assumed to have an independent and identical normal distribution with mean zero and fixed variance ϕ. If the model assumptions are satisfied then the district level residuals from model (1) are expected to be randomly (i.e., pattern less) distributed and not significantly different from the regression line $y=0$. Histogram and q–q plot are also used to examine the normality assumption. Figure 1 shows the distribution of the district level residuals (left hand side plot), histogram of the district level residuals (centre plot) and normal q–q plot of the district level residuals (right hand side plot). From Fig. 1, it can be seen that district level residuals are randomly distributed and the line of fit does not significantly differ from the line $y=0$. It is verify that through histogram and q–q plot, the random district effects are normal distributed. Therefore all the assumptions for the model diagnostics are completely satisfied for the data.

To evaluate the reliability and the validity of the small area estimates second set of diagnostics is used. Such diagnostics are suggested by Brown et al. (2001). In SAE, model-based small area estimates should be consistent with unbiased direct survey estimates, whenever these are known and also be more accurate than direct survey estimates. The values for the model-based small area estimates derived from the fitted model should provide an approximation to the direct survey estimates that is consistent with these values being close to the expected values of the direct estimates. The model-based small area estimates should have mean squared errors significantly lower than

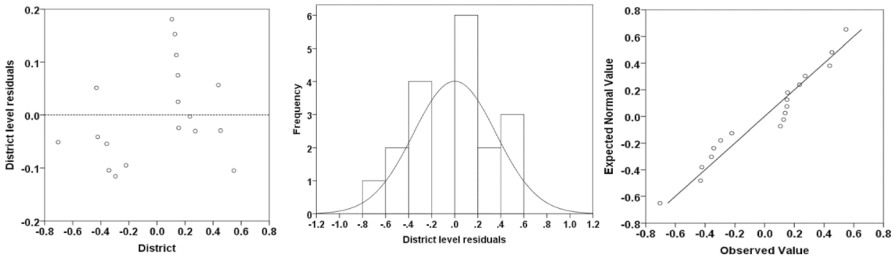

Fig. 1 Distribution of the district level residuals (left), histogram of the district level residuals (centre) and normal q–q plot of the district level residuals (right) for the model-based small area estimates generated by the EP (2)

the variances of corresponding direct survey estimates. We deliberate three commonly used measures namely the bias diagnostics, percent coefficient of variation (CV) and the 95% confidence intervals for small area estimates diagnostics. For the 95% confidence interval we examine the width of the interval for the direct survey estimates compared to the model-based estimates generated by the EP (2). For more precise estimates, we expect the width of the confidence interval to be narrower.

We employ the bias diagnostics to inspect if the small area estimates are less extreme when compared to the direct survey estimates. In addition, if direct survey estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If small area estimates are close to the true values the regression of the direct estimates on the model-based estimates should be equivalent. We plot direct survey estimates on y-axis and model-based small area estimates generated by the EP method on x-axis and we look for divergence of regression line from $y=x$ and test for intercept = 0 and slope = 1. The bias scatter plot of the direct survey estimates against the model-based small area estimates for EP is given in Fig. 2. The bias diagnostics plot in Fig. 2 indicates that the small area estimates generated by the EP method are less extreme when compared to the direct survey estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. That is, the estimates of poverty incidence generated by EP method lies along the line y = x for most of the districts which indicates that they are approximately design unbiased. Although, the results for bias test (i.e. intercept = 0 and slope = 1) are not reported, but the test support the conclusion from Fig. 2.

We also use goodness of fit (GoF) diagnostic. This diagnostic tests whether the direct and model-based estimates generated by the EP are statistically different. The null hypothesis is that the direct and model-based estimates are statistically equivalent. The alternative is that the direct and model-based estimates are statistically different. The GoF diagnostic is computed using the following Wald statistic for EP estimate:

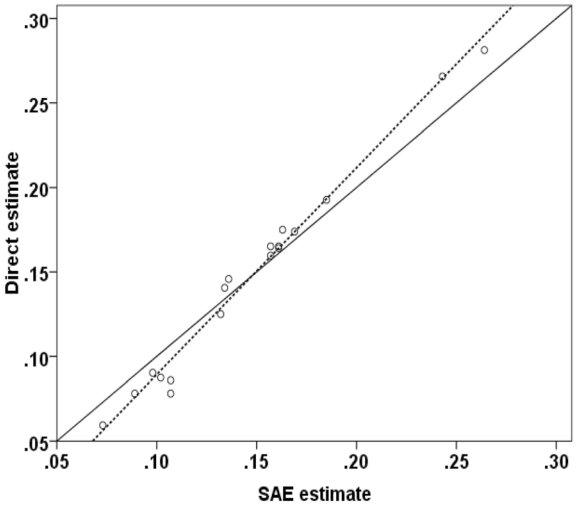

Fig. 2 Bias diagnostics plot with $Y=X$ line (solid line) and regression line (dotted line) for the model-based small area estimates generated by the EP (2)

$$W = \sum_d \left\{ \frac{\left(\hat{p}_d^{Direct} - \hat{p}_d^{EP}\right)^2}{v(\hat{p}_d^{Direct}) + mse(\hat{p}_d^{EP})} \right\}.$$

The value from the test statistic is compared against the value from a chi square distribution with D degrees of freedom. For our analysis, this is the chi square value with $D = 18$ degrees of freedom which is 9.39 at 5% level of significance. For EP, the value of Wald statistic is W = 2.52. A smaller value (less than 9.39 in this case) indicates no statistically significant difference between model-based estimates generated by the EP and direct survey estimates. The diagnostic results clearly show that EP estimates are consistent with direct survey estimates.

We also examine the aggregation of direct and model-based EP estimate at state level. In small area applications, aggregation of model-based small area estimates at higher level is always desirable. The NSSO always expects that the small area (i.e. district) estimates are aggregated to higher (state) level estimate. We compute state level incidence of poverty by aggregating the direct survey estimates as Direct estimate $= \sum_d \left(N_d \times \hat{p}_d^{Direct}\right) \Big/ \sum_d N_d$ and the model-based estimates as Model-based estimate $= \sum_d \left(N_d \times \hat{p}_d^{EP}\right) \Big/ \sum_d N_d$. The state level estimate of incidence of poverty by direct and EP methods are 0.1392 and 0.1397 respectively. As one expects, model-based district level estimate are aggregated well to state level direct estimate.

We use the percent CV to assess the comparative precision of model-based small area estimates (EP) and direct survey estimates. The CVs show the sampling variability as a percentage of the estimate. The also CV provides a measure of relative errors, and gives an indication of the precision of the model-based estimates when compared with the direct estimates. The percent CV of an estimate $\hat{p}_d$ is defined as $CV(\hat{p}_d) = \left(se(\hat{p}_d)/\hat{p}_d\right) \times 100$, where $se(\hat{p}_d) = \sqrt{mse(\hat{p}_d)}$ is the estimate of standard error of $\hat{p}_d$ and $mse(\hat{p}_d)$ is the estimate of MSE($\hat{p}_d$). Estimates with large CVs are considered unreliable (i.e. smaller is better). But, there are no internationally accepted tables available that allow us to judge what is "too large". Different organization used different cut off for CV to release their estimate for the public use. For example, Office for National Statistics, United Kingdom has cut off CV value of 20% for acceptable estimates. The % CV of direct and EP estimates are given in Table 2. Figure 3 presents the district-wise distribution of % CV for the model-based estimates and direct estimates. The results in Table 2 and district-wise values in Fig. 3 clearly show that direct survey estimates for small area poverty incidence are unstable with CV varies from 12.80 to 43.52% with average of 19.75. The % CV of EP ranges from 12.96 to 32.24% with average of 17.53%. The results in Fig. 3 and Table 2 clearly reveal the model-based small area estimates generated by the EP (SAE estimates) are reliable than the direct estimates.

The district-wise 95 percent confidence intervals (95% CIs) of the EP (SAE estimates) and the direct estimates are shown in Fig. 4. It is important to note that the 95% CIs for the direct estimates are calculated assuming a simple random sample generated the sample proportions. This ignores the effects of differential weighting and clustering within districts that would further inflate the true standard errors of the direct estimates. The standard errors of the direct estimates are too large and

Table 2 Distribution of district-wise sample sizes (n_d), estimates of poverty incidence (estimate) along 95% confidence interval (95% CI) and percentage coefficient of variation (% CV) generated by direct survey estimate (direct estimate) and model-based small area estimate (SAE estimate) for West Bengal

District	n_d	Direct estimate				SAE estimate			
		Estimate	95% CI		% CV	Estimate	95% CI		%CV
			Lower	Upper			Lower	Upper	
Darjiling	64	0.08	0.01	0.14	43.52	0.11	0.04	0.17	32.24
Jalpaiguri	192	0.13	0.08	0.17	19.20	0.13	0.09	0.18	17.44
Koch Bihar	128	0.14	0.08	0.20	22.04	0.13	0.08	0.19	20.30
Uttar Dinajpur	128	0.27	0.19	0.34	14.68	0.24	0.17	0.31	14.61
Dakshin Dinajpur	96	0.15	0.08	0.22	24.69	0.14	0.08	0.20	22.54
Maldah	184	0.17	0.12	0.23	16.10	0.17	0.12	0.22	15.20
Murshidabad	280	0.18	0.13	0.22	13.14	0.16	0.12	0.21	13.30
Birbhum	160	0.09	0.04	0.13	25.14	0.10	0.06	0.14	20.80
Barddhaman	288	0.09	0.06	0.12	18.83	0.10	0.07	0.13	16.77
Nadia	224	0.17	0.12	0.21	15.14	0.16	0.11	0.20	14.80
North 24 Parganas	256	0.08	0.05	0.11	21.76	0.09	0.06	0.12	18.46
Hugli	224	0.17	0.12	0.21	15.14	0.16	0.12	0.21	14.30
Bankura	192	0.19	0.14	0.25	14.53	0.19	0.13	0.24	14.40
Puruliya	160	0.28	0.21	0.35	12.80	0.26	0.20	0.33	12.96
Haora	288	0.16	0.12	0.20	13.77	0.16	0.12	0.20	13.36
South 24 Parganas	128	0.09	0.04	0.13	29.09	0.11	0.06	0.15	22.12
Paschim Medinipur	320	0.06	0.03	0.09	21.90	0.07	0.05	0.10	18.38
Purba Medinipur	256	0.16	0.12	0.21	14.02	0.16	0.12	0.20	13.61
Minimum	64	0.06	0.01	0.09	12.80	0.07	0.04	0.10	12.96
Average	198	0.15	0.10	0.20	19.75	0.15	0.10	0.19	17.53
Maximum	320	0.28	0.21	0.35	43.52	0.26	0.20	0.33	32.24

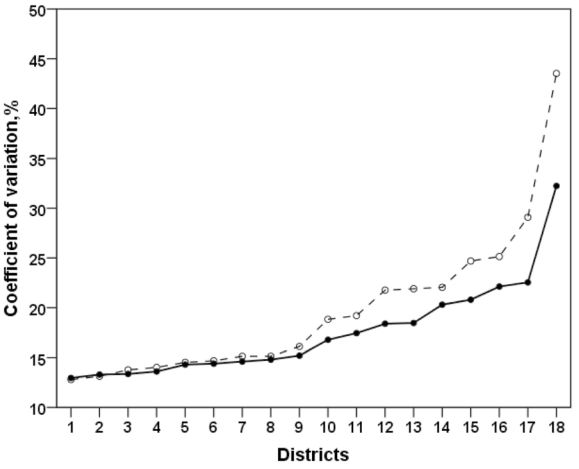

Fig. 3 District-wise percentage coefficient of variation for the direct (dash line, °) and model based small area estimate (solid line, •). The values are shown in increasing order of percentage coefficient of variation of the direct estimates

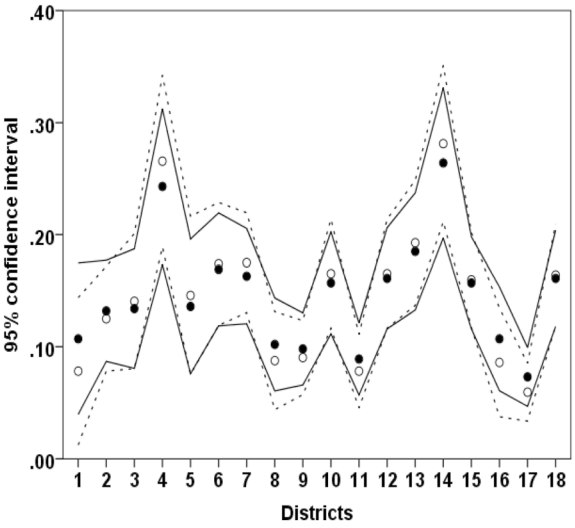

Fig. 4 District-wise 95% confidence interval (lower and upper) for the direct (dash line) versus model based small area estimates (solid line). Direct estimate (dash line, °) and model-based small area estimate (solid line, •)

therefore the estimates are unreliable. In Fig. 4 and Table 2, we observe that 95% CIs for the direct estimates are wider than the 95% CIs for the EP estimates. It indicates that the 95% CIs for the EP estimates are more precise and contain both direct and EP estimates of incidence of poverty.

In Table 2, we also present the district-wise estimates of poverty incidence, i.e. proportion of poor household. Here, for ease of understanding we interpret the results in terms of percentages and not proportions, and focus on interpreting the model-based estimates of the small areas. The spatial mapping of district-wise poverty incidence or proportion of poor household generated by the model-based method for the state of West Bengal is also shown in Fig. 5. This spatial map is very useful in identifying the districts and areas with low and high level of poverty incidence in the state. The spatial inequality in distribution of poverty incidence is demonstrated through this map. In simple words we can say that the degree of inequality with respect to distribution of proportion of poor households in different districts is displayed in this map. The estimates in Table 2 and map in Fig. 5 confirm that there is inequality in the distribution of proportion of poor households in the different districts in West Bengal. The district-wise poverty incidence produced by EP method in rural areas of West Bengal ranges from 7 to 26% with average of 15%.

From Fig. 5, it is observe that Paschim Medinipur has (7%) the lowest whereas Purulia (26%) has the highest incidence of poverty in West Bengal. Darjiling, Birbhum, Barddhaman, North 24 Parganas have small poverty incidence that lying between 9 and 11% whereas Uttar Dinajpur and Purulia have the highest rate of poverty incidence that lying between 23 and 26%. It is noted that the poverty incidence is not clustered in a region rather it is distributed throughout the State. The district level estimates as well as spatial map of poverty rate expected to deliver precious information to policy-analyst and decision makers for finding the areas and districts demanding more attention for development in the state as well as monitoring the progress of SDG1. This application and description of methodology can also be

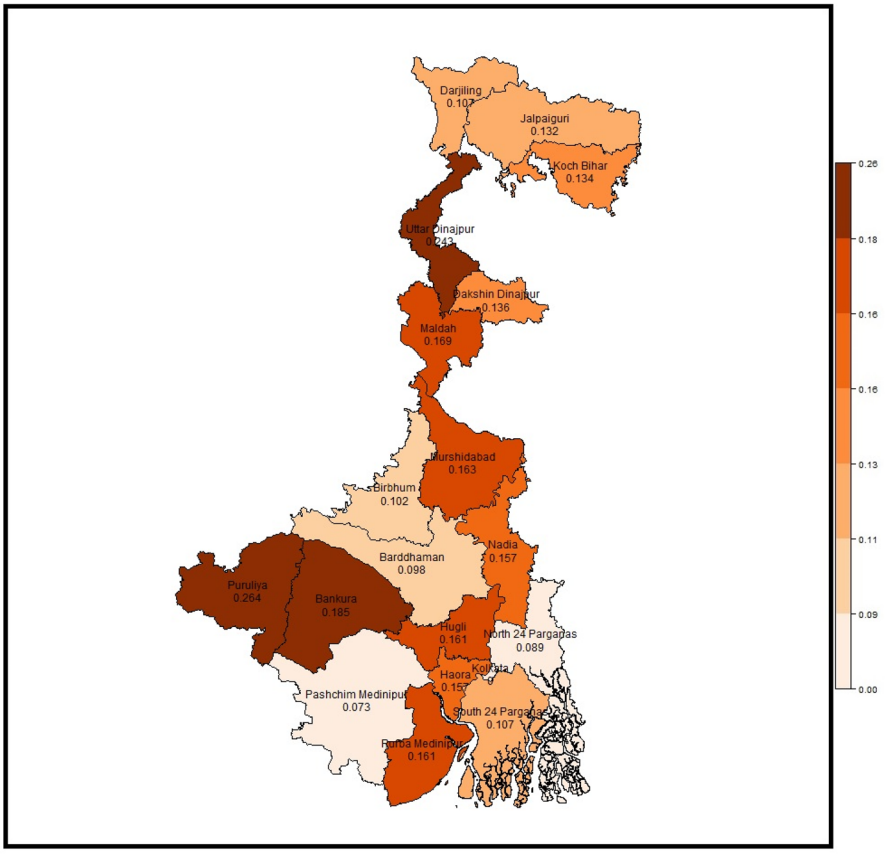

Fig. 5 Poverty mapping generated for the state of West Bengal

used as a guidelines for other application of SAE in different survey data as well as data from other countries. It is evident that model-based SAE method brings gain in efficiency in district level estimates. The SAE can be used as cost effective and efficient approach for generating reliable micro level statistics from existing survey data and using auxiliary information from different published sources. The results clearly indicates the advantage of using SAE technique to cope up the small sample size problem in producing the estimates or reliable confidence intervals.

Concluding Remarks

For policy formulation, planning, allocation of funds, monitoring and evaluation of programmes, there is the need for statistics at the local level where programmes are designed and implemented. In India, in recent year there has been lot of emphasis on decentralised of governance, and therefore need for such

disaggregate level estimate becomes important and unavoidable. Censuses are usually limited as they tend to focus mainly on the basic socio-demographic and economic data and not available for every time period. On the other hand, country is fortunate to have regular NSSO survey for generating number of socio-economic indicators. The NSSO surveys are aimed to generate estimates at national and state level. They do not provide sub-state level statistics. National and state level estimates, generated from the NSSO surveys, mask variations at the district level and render little information for local level planning and allocation of resources. In this article, we illustrate an application of SAE techniques to generate reliable and representative statistics on poverty incidence at district level for rural areas of the state of West Bengal by linking data from the Household Consumer Expenditure Survey 2011–2012 of NSSO and the Census 2011. The diagnostic measures used for examining the validity and reliability of the model-based estimates clearly confirm that generated estimates have reasonably good precision. The SAE method has also generated reliable estimates for the districts with very small sample sizes. The results clearly show the advantage of using SAE technique to cope up the small sample size problem in producing the reliable district level estimates and confidence intervals. The district-wise estimates and spatial map of poverty incidence generated by SAE technique reveals striking differences and point to specific geographical areas where intervention should be strengthened. The estimates are helpful in identifying the districts/regions with lower and higher level of incidence of poverty. It is expected that the estimates generated using SAE technique in general and from this research in particular should be useful for for meeting the data requirements for policy research and strategic planning including monitoring the progress of sustainable development goals by different international organizations and by Departments and Ministries in the Government of India.

Acknowledgements The author would like to acknowledge the valuable comments and suggestions of the Editor and the referee. These led to a considerable improvement in the paper. The work was carried out under the ICAR-National Fellow Project at ICAR-IASRI, New Delhi, India.

References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83: 28–36.

Breslow, N.E., and D.G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88: 9–25.

Brown, G., Chambers, R., Heady, P., and D. Heasman. 2001. Evaluation of small area estimation methods - an application to unemployment estimates from the UK LFS. In: Proceedings of statistics canada symposium 2001. Achieving data quality in a statistical agency: a methodological perspective.

Chandra, H., N. Salvati, and U.C. Sud. 2011. Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal Applied Statistics* 38 (11): 2413–2432.

Chaudhuri, S., and N. Gupta. 2009. Levels of living and poverty patterns: a district-wise analysis for India. *Economic and Political Weekly* XLIV (9): 94–110.

Chauhan, R.K., S.K. Mohanty, S.V. Subramanian, J.K. Parida, and B. Padhi. 2016. Regional estimates of poverty and inequality in India, 1993–2012. *Social Indicators Research* 127: 1249–1296.

Coondoo, D., A. Majumder, and S. Chattopadhyay. 2011. District-level poverty estimation: a proposed method. *Journal of Applied Statistics* 38 (10): 2327–2343.

Fay, R.E., and R.A. Herriot. 1979. Estimation of income from small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association* 74: 269–277.

Johnson, F.A., H. Chandra, J. Brown, and S. Padmadas. 2010. Estimating district-level births attended by skilled attendants in ghana using demographic health survey and census data: an application of small area estimation technique. *Journal Official Statistics* 26 (2): 341–359.

Manteiga, G.W., M.J. Lombardìa, I. Molina, D. Morales, and L. Santamarìa. 2007. Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computation Statistics and Data Analysis* 51: 2720–2733.

Mitra, S. 2016. Poverty in West Bengal: a review of recent performance and programmes. In *Understanding development-an Indian perspective on legal and economic policy*, ed. S. Banerjee, V. Mukherjee, and S.K. Haldar. New Delhi: Springer.

Mohanty, S.K., D. Govil, R.K. Chauhan, R. Kim, and S.V. Subramanian. 2016. Estimates of poverty and inequality in the districts of India, 2011–2012. *Journal of Development Policy and Practice* 1 (2): 142–202.

Pfeffermann, D. 2002. Small area estimation: new developments and directions. *International Statistical Review* 70: 125–143.

Rao, J.N.K. 2003. *Small area estimation*. New York: John Wiley and Sons.

Rao, J.N.K., and I. Molina. 2015. *Small area estimation*, 2nd ed. New York: John Wiley and Sons.

Saei, A., and R. Chambers. 2003. Small area estimation under linear and generalized linear mixed models with time and area effects. Working Paper M03/15. Southampton Statistical Sciences Research Institute, University of Southampton, U.K.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model assisted survey sampling*. New York: Springer.

Ybarra, L.M.R., and S. Lohr. 2008. Small area estimation when auxiliary information is measured with error. *Biometrika* 95 (4): 919–931.