REVSTAT – Statistical Journal
Volume 18, Number 2, April 2020, 223–235

ESTIMATION OF SMALL AREA TOTAL WITH RANDOMIZED DATA

Authors: Shakeel Ahmed
– Department of Statistics, Quaid-i-Azam University Islamabad,
Islamabad, Pakistan
sahmed@qau.edu.pk

Javid Shabbir
– Department of Statistics, Quaid-i-Azam University Islamabad,
Islamabad, Pakistan
javidshabbir@gmail.com

Sat Gupta
– Department of Mathematics and Statistics, University of North Carolina at Greensboro, North Carolina, USA
sngupta@uncg.edu

Frank Coolen
– Department of Mathematical Sciences, Durham University,
Durham, DH1 3LE, UK
frank.coolen@durham.ac.uk

Received: February 2019 Revised: September 2019 Accepted: September 2019

Abstract:

- In social surveys involving questions that are sensitive or personal in nature, respondents may not provide correct answers to certain questions asked by the interviewer. The impact of this non-response or inaccurate response becomes even more acute in the case of small area estimation (SAE) where we already have the problem of small sample size coming from the small area. To obtain a truthful response, we use randomized response techniques in each small area. We assume that a non-sensitive auxiliary variable, highly correlated with the study variable, is available. We use the word model in two senses — one in the context of population models, i.e. the relationship between the study variable and the auxiliary variable; and second, the scrambled response model. We focus on the problem of estimating small area total and examine its performance both theoretically and numerically.

Key-Words:

- *quantitative sensitive variable; randomized response models; small area estimation; superpopulation model.*

1. INTRODUCTION

In social sciences, responses on some stigmatizing variables are often needed to make inference about the behavior of some human populations. Examples of such situations are where questions are asked that are related to topics like tax evasion, use of illegal drugs, extra marital affairs, ethical issues, political affiliation, etc. In the case of stigmatizing study variables, non-sampling error may increase due to missing or false responses, which leads to biased estimates of population parameters such as mean, total or proportion. To reduce such bias in sample surveys, [34] proposed a randomized response technique (RRT) for obtaining more accurate estimates. A lot of research has been done for improving the original RRT model of [34]. Authors contributing in this area include [17], [18], [35], [6], [12], [22], [7], [3], [19, 20], [21] and [9, 10, 11]. In RRT literature, much more attention has been paid to design-based approach which assumes the population to consist of fixed constants. But in many real-life situations, population values are generated as realizations of a set of stochastic variables. Such population is called a superpopulation and the statistical models for such type of populations are called superpopulation models. Superpopulation models help in sample selection, constructing estimators for population parameters of interest, and enhancing the precision of estimates. A superpopulation model uses the relationship between the study variable and the auxiliary variable(s) to predict the population values for the non-sampled units assuming non-informative sampling approach. Under the framework of model-based inference, [14] dealt with the problem of estimation of a finite population mean or total. [27] and [8] attempted to obtain optimal model-unbiased estimators of the population mean and total using least squares estimation methods and the well-known Gauss–Markov theorem. Some discussion on model-based approach can be found in [2], [15], [16], [30, 31], [29], [28], and [33]. A detailed review of model-based estimation is also available in [32].

[13] and [24] have suggested post-censal estimates (estimates obtained immediately after census using the census results) for small areas and called it small area estimation (SAE). [23] dealt with labor force trend estimation for small areas. Work related to such methods can also be found in [25, 26] and [38]. More recently, [36] have considered estimation of uncertainty in spatial micro-simulation approaches for SAE. The main purpose of SAE is to overcome the problem of small sample when separate estimates for domains are needed. In this article, we develop some model-based estimators for small area totals assuming the study variable in each domain is sensitive. A generalized randomized response model has been used to collect information about the study variable. The rest of the article is structured as follows: an overview of SAE under direct response is considered in Section 2 with some superpopulation models. Section 3 extends the SAE given in Section 2 to randomized response models, assuming a sensitive quantitative study variable and non-sensitive auxiliary variable. Section 4 presents a numerical study based on two real life data sets. Some concluding remarks are provided in Section 5.

2. SAE UNDER DIRECT RESPONSE

Consider a finite population $U = \{U_1, U_2, ..., U_N\}$ of N units as a realization of a superpopulation with variable of interest y, and auxiliary variable x. For a specific sup-population A_k, also known as "small area", let d_{ki} be an area specific binary variable, for $k = 1, 2, 3, ...m$ and $i = 1, 2, ...N$, such that $d_{ki} = 1$ if U_i belongs to A_k, and zero otherwise. Further, let $N_k = \sum_U d_{ki}$ be the size of the k-th sub-population or k-th small area (usually unknown), $T_{yk} = \sum_U d_{ki}\, y_i$ and $T_{xk} = \sum_U d_{ki}\, x_i$ be the population totals, $\mu_{yk} = \frac{T_{yk}}{N_k}$ and $\mu_{xk} = \frac{T_{xk}}{N_k}$ be the population means, and $\sigma^2_{yk} = \frac{1}{N_k} \sum_U d_{ki}(y_i - \mu_{yk})^2$ and $\sigma^2_{xk} = \frac{1}{N_k} \sum_U d_{ki}(x_i - \mu_{xk})^2$ be the population variances of the study variable and the auxiliary variable respectively in the k-th area. The notation $\sum_U$ is used for summing the values over U. Also, let the covariance between the study variable and the auxiliary variable in the k-th area be $\sigma_{yxk} = \frac{1}{N_k} \sum_U d_{ik}\,(y_i - \mu_{yk})\,(x_i - \mu_{xk})$. Suppose that s is a member of the set S of all possible samples that can be drawn from U using simple random sampling without replacement (SRSWOR) scheme with size n, and $\bar{s}$ consists of all those elements of U that are not selected in sample s. The population total for the study variable, quantity of interest or estimand, in k-th area can then be expressed as $T_{yk} = \sum_s d_{ki}\, y_i + \sum_{\bar{s}} d_{ki}\, y_i$. A predictor for T_{yk} is obtained as follows:

$$\hat{T}_{yk} = \sum_s d_{ki}\, y_i + \sum_{\bar{s}} d_{ki}\, \hat{y}_i\,. \tag{2.1}$$

The main problem is to find $\hat{y}_i$ for $U_i \in \bar{s}$. The predictor $\hat{y}_i$ is obtained assuming different superpopulation models. We consider three most widely used population models:

1. Homogenous Population Model (HPM): $y = \mu_{yk} + \varepsilon$,
2. Linear Population Model (LPM): $y = \alpha_k + \beta x + \varepsilon$,
3. Ratio Population Model (RPM): $y = \gamma x + x^{1/2}\varepsilon$,

for $k = 1, 2, ..., m$, where ε is the stochastic error term which has mean 0 and a constant variance σ^2. Also, μ_{yk} and α_k are mean effects in k-th area and β and γ are the coefficients of the regression line of y on x for the whole population for the cases with and without intercepts. In model based approach, these parameters are termed as superpopulation parameters.

2.1. Homogeneous Population Model (HPM)

In case of HPM, a BLUP for μ_{yk}, obtained by minimizing the residual sum of square $\sum_s d_{ki}(y_i - \mu_{yk})^2$ is $\bar{y}_k = \frac{1}{n_k} \sum_s d_{ki}\, y_i$, which yields an estimator for T_{yk} given by

$$\hat{t}_{k\mathrm{h}} = \sum_s d_{ki}\, y_i + \sum_{\bar{s}} d_{ki}\, \bar{y}_k = \frac{N}{n} \sum_s d_{ki}\, y_i\,. \tag{2.2}$$

The sub-script 'h' is used to indicate that the superpopulation model is homogeneous. It is straight forward to show that $\hat{t}_{k\mathrm{h}}$ is an unbiased estimator of population total T_{yk} with variance given by

$$\mathrm{Var}(\hat{t}_{k\mathrm{h}}) = \lambda\big[\theta_k\, \sigma^2_{yk} + \theta_k(1 - \theta_k)\, \mu^2_{yk}\big]\,, \tag{2.3}$$

where $\theta_k = \frac{N_k}{N}$ is the population proportion of the units belonging to k-th small area, and $\lambda = \frac{N(N-N)}{n}$. For proof readers can see [5, p. 156–160].

2.2. Linear Population Model (LPM)

Now consider LPM for finding $\hat{y}_i$, $U_i \in \bar{s}$. The BLUP for α_k and β are obtained by minimizing the sum of squared prediction errors for specific areas, i.e.

$$\text{SSPE} = \sum_s d_{ki}(y_i - \alpha_k - x_i\beta)^2 .$$

These are given by $\hat{\alpha}_k = \bar{y}_k - \hat{\beta}\bar{x}_k$ and $\hat{\beta} = \frac{\sum_s d_{ki}(y_i - \bar{y}_k)(x_i - \bar{x}_k)}{\sum_s d_{ki}(x_i - \bar{x}_k)^2}$, where $\bar{y}_k$ and $\bar{x}_k$ are the sample means corresponding to k-th small area. The estimator of T_{yk} under LPM is given by

$$\hat{t}_{k\text{lr}} = \sum_s d_{ki}\, y_i + \sum_{\bar{s}} d_{ki}\big(\hat{\alpha}_k + \hat{\beta} x_i\big) .$$

After some simplifications and using assumption from [5], i.e. $\frac{N_k}{N} \approx \frac{n_k}{n}$, we get

$$\hat{t}_{k\text{lr}} = \frac{N}{n} t_{yk} + \hat{\beta}\left(T_{xk} - \frac{N}{n} t_{xk}\right), \tag{2.4}$$

where $t_{yk} = \sum_s d_{ki}\, y_i$ and $t_{xk} = \sum_s d_{ki}\, x_i$ are the sample totals for k-th small area. Further, $\hat{\beta}$ given in (2.4) is based on local (area specific) observations only, which do not account for relationship between the variables for the entire population. To overcome this deficiency, different area level models have been proposed in literature. For simplicity, we assume that the regression coefficient β of y on x is known for the whole population. For known β, we have

$$\hat{t}_{k\text{lr}} = \frac{N}{n} t_{yk} + \beta\left(T_{xk} - \frac{N}{n} t_{xk}\right). \tag{2.5}$$

The sub-script 'lr' is used to denote that the underlying model is linear. For known β, $\hat{t}_{k\text{lr}}$ is unbiased for T_{yk} with variance given by

$$\text{Var}(\hat{t}_{k\text{lr}}) = \lambda\big(\sigma^{*2}_{yk} + \beta^2\sigma^{*2}_{xk} - 2\beta\sigma^{*}_{yxk}\big) , \tag{2.6}$$

where $\sigma^{*2}_{yk} = \theta_k\sigma^2_{yk} + \theta_k(1-\theta_k)\mu^2_{yk}$, $\sigma^{*2}_{xk} = \theta_k\sigma^2_{xk} + \theta_k(1-\theta_k)\mu^2_{xk}$ and $\sigma^{*}_{yxk} = \theta_k\sigma_{yxk} + \theta_k(1-\theta_k)\mu_{yk}\mu_{xk}$. The value of β that minimizes the variance is $\beta_{\text{opt}} = \frac{\sigma^{*}_{yxk}}{\sigma^{*2}_{xk}}$. The corresponding minimum variance of $\hat{t}_{k\text{lr}}$ is given by

$$\text{Var}(\hat{t}_{k\text{lr}})_{\text{opt}} = \lambda\left(1 - \rho^{*2}_{yxk}\right)\sigma^{*2}_{yk} , \tag{2.7}$$

where $\rho^{*}_{yxk} = \frac{\sigma^{*}_{yxk}}{\sigma^{*}_{yk}\sigma^{*}_{xk}}$. From Equations (2.7) and (2.3), it is obvious that $\hat{t}_{k\text{lr}}$ is always more efficient than $\hat{t}_{k\text{h}}$ for any linear relationship between y and x.

2.3. Ratio Population Model (RPM)

For situations when there is a proportional relationship between the survey variable and the auxiliary variables, the RPM [32] is often preferred as the working model. RPM is given by

$$y = \gamma x + x^{1/2}\varepsilon\,. \tag{2.8}$$

The estimator for γ which minimizes the sum of squared errors, i.e. $\mathrm{SSE}^* = \sum_s d_{ki}\big(\frac{y_i - x_i\gamma}{x_i^{1/2}}\big)^2$, is given by $\hat{\gamma} = \frac{\sum_s d_{ki}\, y_i}{\sum_s d_{ki}\, x_i}$. Now consider

$$\hat{t}_{k\mathrm{r}} = \sum_s d_{ki}\, y_i + \sum_{\bar{s}} d_{ki}(\hat{\gamma} x_i) \tag{2.9}$$

as an estimator of T_{yk}. The sub-script 'r' is used to denote the ratio population model for the response variable. After simplification and assuming $\frac{N_k}{N} \approx \frac{n_k}{n}$, we get

$$\hat{t}_{k\mathrm{r}} = \frac{\sum_s d_{ki}\, y_i}{\sum_s d_{ki}\, x_i} \sum_{\bar{s}} d_{ki}\, x_i = \frac{N}{n}\left[t_{yk}\,\frac{n\,\mu_{xk}}{t_{xk}}\right]. \tag{2.10}$$

The bias and MSE respectively, of $\hat{t}_{k\mathrm{r}}$, are given by

$$\mathrm{Bias}(\hat{t}_{k\mathrm{lr}}) \cong \frac{\lambda}{N}\,\mu_{yk}\big(C^{*2}_{xk} - C^{*}_{yxk}\big) \tag{2.11}$$

and

$$\mathrm{MSE}(\hat{t}_{k\mathrm{r}}) \cong \lambda\mu^2_{yk}\big(C^{*2}_{yk} + C^{*2}_{xk} - 2\,C^{*}_{yxk}\big)\,, \tag{2.12}$$

where $C^{*2}_{yk} = \frac{\sigma^{*2}_{yk}}{\mu^2_{yk}}$, $C^{*}_{xk} = \frac{\sigma^{*2}_{xk}}{\mu^2_{xk}}$ and $C^{*}_{yxk} = \frac{\sigma^{*}_{yxk}}{\mu_{yk}\,\mu_{yk}}$. From (2.3) and (2.12), it can be inferred that $\mathrm{MSE}(\hat{t}_{k\mathrm{r}}) \leq \mathrm{Var}(\hat{t}_{k\mathrm{h}})$ if $\rho^{*}_{yxk} \geq \frac{1}{2}\frac{C^{*}_{xk}}{C^{*}_{yk}}$.

3. SAE UNDER RANDOMIZED RESPONSE TECHNIQUE

When the study variable is of sensitive nature, it is difficult to obtain 100% response through direct response method. For improved response rate in such situations, survey statisticians prefer to use RRT. Assuming quantitative study variable, and following [11], we use the following scrambled response model

$$z = a y + b\,, \tag{3.1}$$

where y is the sensitive study variable which follows one of the population models given in Section 2, a and b are two uncorrelated scrambling variables with means μ_a and μ_b, and variances σ^2_a and σ^2_b respectively. Further, a and b are independent of the study variable y. Note that respondents from each small area use the same scrambling variables a and b whose distributions are unknown to the interviewer while the means and variances are known. Taking expectation of Equation (3.1) with respect to randomization mechanism, we have $E_R(z) = \mu_a\, y + \mu_b$. The transformed scrambled response is obtained as $y = \frac{E_R(z) - \mu_b}{\mu_a}$. A sample unbiased estimate for y is $\tilde{y} = \frac{z - \mu_b}{\mu_a}$.

3.1. Homogeneous Population Model (HPM)

When the underlying population model is homogeneous, i.e. when there is no covariate affecting the outcome variable, a BLUP for the superpopulation parameter μ_{yk} is $\tilde{\bar{y}}_k = \tilde{t}_{yk}/n_k$ which yields an estimator for T_{yk} given by

$$\tilde{t}_{k\text{h}} = \sum_s d_{ki}\,\tilde{y}_i + \sum_{\bar{s}} d_{ki}\,\tilde{\bar{y}}_k = n_k\,\tilde{\bar{y}}_k + (N_k - n_k)\,\tilde{\bar{y}}_k = \frac{N}{n}\,\tilde{t}_{yk}\,, \tag{3.2}$$

where $\tilde{t}_{yk} = \sum_s d_{ki}\,\tilde{y}_i$. We assume that the sampling weights for the whole sample and the sample within k-th domain are same, i.e. $\frac{N_k}{N} \approx \frac{n_k}{n}$. It is easy to show that $\tilde{t}_{k\text{h}}$ is an unbiased estimator of population total T_{yk} with variance

$$\text{Var}(\tilde{t}_{k\text{h}}) = \lambda\left(\theta_k\,\tilde{\sigma}^2_{yk} + \theta_k(1-\theta_k)\,\mu^2_{yk}\right), \tag{3.3}$$

where $\tilde{\sigma}^2_{yk} = \text{Var}(\tilde{y}_i\,|\,d_{ki}=1) = \frac{1}{\mu_a^2}\,\text{Var}(z_i\,|\,d_{ki}=1)$, and

$$\begin{aligned}\text{Var}(z_i\,|\,d_{ki}=1) &= V_s\big\{E_R(z_i\,|\,d_{ki}=1)\big\} + V_R\big\{E_S(z_i\,|\,d_{ki}=1)\big\}\\ &= E_s\left(\sigma_a^2\,y_i^2 + \sigma_b^2\,|\,d_{ki}=1\right) + V_s\left(\mu_a\,y_i + \mu_b\,|\,d_{ki}=1\right)\\ &= \sigma_a^2\,\mu_{2,yk} + \sigma_b^2 + \mu_a^2\,\sigma^2_{yk}\,,\end{aligned} \tag{3.4}$$

where E_s and V_s are the expectation and variance with respect to the data generating mechanism. Also $\mu_{2,yk}$ is the second order raw moment for k-th area. Using value of $\tilde{\sigma}^2_{yk}$ from (3.3), we get

$$\begin{aligned}\text{Var}(\tilde{t}_{k\text{h}}) &= \lambda\left(\theta_k\,\sigma^2_{yk} + \theta_k(1-\theta_k)\,\mu^2_{yk} + \theta_k\,\psi^2_{yk}\right),\\ \text{Var}(\tilde{t}_{k\text{h}}) &= \text{Var}(\hat{t}_{k\text{h}}) + \lambda\left(\theta_k\,\psi^2_{yk}\right),\end{aligned} \tag{3.5}$$

where $\psi^2_{yk} = \frac{1}{\mu_a^2}\left(\sigma_a^2\,\mu_{2,yk} + \sigma_b^2\right)$. It is observed from (3.5) that $\text{Var}(\tilde{t}_{k\text{h}})$ is always larger than $\text{Var}(\hat{t}_{k\text{h}})$ as the second term is positive. For detailed derivation, see [1]. The $\text{Var}(\tilde{t}_{k\text{h}})$ decreases with decrease in variance of the scrambled variables but this leads to reduction in respondent's privacy as well. Hence, the variance of the scrambled response models should be of a reasonable size resulting in a proper tradeoff between respondent's privacy and the efficiency of the proposed estimators.

To improve efficiency for a fixed level of privacy protection, we use model relationship between the available auxiliary variable and the study variable. Subsections 3.2 and 3.3 cover linear and ratio population models respectively that utilize the relationship between the variables at unit level to increase efficiency.

3.2. Linear Population Model (LPM)

Assuming LPM, we find the predicted transformed scrambled response $\tilde{y}_i$, $U_i \in \bar{s}$. The BLUP for α_k and β are obtained by minimizing the sum of squared errors for the k-th area as follows:

$$\text{SSE} = \sum_s d_{ki}\,\tilde{e}_i^2 = \sum_s d_{ki}(\tilde{y}_i - \alpha_k - x_i\beta)^2\,,$$

where $\tilde{\hat{\alpha}}_k = \tilde{\bar{y}}_k - \tilde{\hat{\beta}}\,\bar{x}_k$ and $\tilde{\hat{\beta}} = \frac{\sum_s d_{ki}(\tilde{y}_i - \tilde{\bar{y}}_k)\,(x_i - \bar{x}_k)}{\sum_s d_{ki}(x_i - \bar{x}_k)^2}$. The predictive estimator under LPM using transformed scrambled response is given by

$$\tilde{t}_{k\mathrm{lr}} \;=\; \sum_s d_{ki}\,\tilde{y}_i + \sum_{\bar{s}} d_{ki}\big(\tilde{\hat{\alpha}}_k + \tilde{\hat{\beta}}\,x_i\big)\,. \tag{3.6}$$

After some simplification, we get

$$\tilde{t}_{k\mathrm{lr}} \;=\; \frac{N}{n}\,\tilde{t}_{yk} + \tilde{\hat{\beta}}\left(T_{xk} - \frac{N}{n}\,t_{xk}\right).$$

By same argument as given in Subsection 2.2, we have

$$\tilde{t}_{k\mathrm{lr}} \;=\; \frac{N}{n}\,\tilde{t}_{yk} + \beta\left(T_{xk} - \frac{N}{n}\,t_{xk}\right). \tag{3.7}$$

For known β, $\tilde{t}_{k\mathrm{lr}}$ is unbiased for T_{yk}, with variance given by

$$\mathrm{Var}(\tilde{t}_{k\mathrm{lr}}) \;=\; \lambda\big(\tilde{\sigma}^{*2}_{yk} + \beta^2\,\sigma^{*2}_{xk} - 2\,\beta\,\sigma^{*}_{yxk}\big)\,. \tag{3.8}$$

The optimum value of β is $\beta_{\mathrm{opt}} = \frac{\sigma^*_{yxk}}{\sigma^{*2}_{xk}}$ with corresponding design optimum variance

$$\mathrm{Var}(\tilde{t}_{k\mathrm{lr}})_{\mathrm{opt}} \;=\; \lambda\big(1 - \tilde{\rho}^{*2}_{yxk}\big)\,\tilde{\sigma}^{*2}_{yk}\,, \tag{3.9}$$

where $\tilde{\rho}^*_{yxk} = \frac{\sigma^*_{yxk}}{\tilde{\sigma}^*_{yk}\,\sigma^*_{xk}}$. Equation (3.9) shows that $\tilde{t}_{k\mathrm{lr}}$ is always more efficient than $\tilde{t}_{k\mathrm{h}}$ for any correlation between y and x.

3.3. Ratio Population Model (RPM)

For the situation when there is a proportional relationship between the sensitive study variable, and the auxiliary variable whose values are available for all population units and the variance of the survey variable is also proportional to the auxiliary variable, the RPM is often preferred. Consider (3.1), where y follows the ratio population model. The estimator for γ which minimizes the sum of squared errors, i.e. $\mathrm{SSE}^* = \sum_s d_{ki}\big(\frac{\tilde{y}_i - x_i\gamma}{x_i^{1/2}}\big)^2$, is given by $\tilde{\hat{\gamma}} = \frac{\sum_s d_{ki}\,\tilde{y}_i}{\sum_s d_{ki}\,x_i}$. Consider the prediction problem as follows

$$\tilde{t}_{k\mathrm{r}} \;=\; \sum_s \tilde{d}_{ki}\,\tilde{y}_i + \sum_{\bar{s}} d_{ki}\big(\tilde{\hat{\gamma}}\,x_i\big)\,. \tag{3.10}$$

After simplification, we get

$$\tilde{t}_{k\mathrm{r}} \;=\; \frac{\sum_s d_{ki}\,\tilde{y}_i}{\sum_s d_{ki}\,x_i}\,\sum_{\bar{s}} d_{ki}\,x_i \;=\; \frac{N}{n}\left[\tilde{t}_{yk}\,\frac{n\,\mu_{xk}}{t_{xk}}\right]. \tag{3.11}$$

The bias and MSE of $\tilde{t}_{k\mathrm{r}}$ are given by

$$\mathrm{Bias}(\tilde{t}_{k\mathrm{lr}}) \;\cong\; \frac{\lambda}{N}\,\mu_{yk}\big(C^{*2}_{xk} - C^*_{yxk}\big) \tag{3.12}$$

and

$$\mathrm{MSE}(\tilde{t}_{k\mathrm{r}}) \;\cong\; \lambda\mu^2_{yk}\big(\tilde{C}^*_{yk} + \tilde{C}^*_{xk} - 2\,C^*_{yxk}\big)\,, \tag{3.13}$$

where $\tilde{C}^{*2}_{yk} = \frac{\tilde{\sigma}^{*2}_{yk}}{\mu^2_{yk}}$ and $C^*_{yxk} = \frac{\sigma^*_{yxk}}{\mu_{yk}\,\mu_{xk}}$. Equation (3.12) shows that the use of RRT to collect response on the dependent variable does not affect the bias of ratio estimator. From (3.5) and (3.13), it can be inferred that $\mathrm{MSE}(\tilde{t}_{k\mathrm{r}}) \leq \mathrm{Var}(\tilde{t}_{k\mathrm{h}})$ if $\tilde{\rho}^*_{yxk} \geq \frac{1}{2}\,\frac{C^*_{xk}}{\tilde{C}^*_{yk}}$.

4. NUMERICAL STUDY

For numerical validation of our proposed estimators, two real life data sets, one with two small areas and the other with three small areas, are used. The detailed descriptions along with summary statistics of the populations are given in following subsections.

Blood transfusion data

The data are taken from [37], where F, the frequency of donations, is the study variable, T (Time in months since first donation) is taken as the covariate, and a binary variable representing whether he/she donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood) is taken as the area membership variable.

Players head circumference data

This data is taken from [4] which contains physical measures of $N = 90$ players forming three groups, i.e. high school football players (Group 1), college football players (Group 2) and Non-football players (Group 3), each having 30 students. The three groups represent the small areas. The study variable y and the auxiliary variable x respectively are jaw width and ear-to-top-of-head measurement of players. The scrambling variables a and b are generated from Uniform distributions with different ranges.

Table 1: Summary statistics.

Parameter	**Data 1**		**Data 2**		
k	1	2	1	2	3
θ_k	0.7620	0.2380	0.3333	0.3333	0.3333
μ_{yk}	4.8018	7.7978	13.0833	10.0800	10.9467
μ_{xk}	4.8018	7.7978	14.7333	13.4533	13.6967
σ^2_{yk}	22.5318	64.5916	1.0876	1.1520	1.4577
σ^2_{xk}	605.4251	558.3500	0.8920	0.5702	0.3921
σ_{yxk}	76.3885	140.5756	0.5402	0.0870	0.0870
ρ_{yxk}	0.6540	0.7402	0.3333	0.3333	0.3333

Table 1 provides the summary statistics for the data sets. The theoretical results (TR) are obtained using Variance/MSE expressions given in Section 2. The simulated results (SR) are obtained using following algorithm:

1. Select a simple random sample of size n (100 and 30 for Populations I and II respectively) without replacement from the populations described above and stratify the populations according to the domain membership variable d_k.

2. Record information y and x for all small areas after generating values of scrambling variables a and b from uniform distribution with different ranges.
3. Calculate the values of small area estimators under direct and randomized response technique.
4. Repeat Steps 1–3 50000 times and obtain the simulated Variance, MSE and PRE.

The PRE in Table 2 are computed as $\mathrm{PRE_r} = \frac{\mathrm{Var}(\hat{t}_{k\mathrm{h}})}{\mathrm{MSE}(\hat{t}_{k\mathrm{r}})}$ and $\mathrm{PRE_{lr}} = \frac{\mathrm{Var}(\hat{t}_{k\mathrm{h}})}{\mathrm{Var}(\hat{t}_{k\mathrm{lr}})}$ for $\hat{t}_{k\mathrm{r}}$ and $\tilde{t}_{k\mathrm{lr}}$ are respectively while $\mathrm{PRE_h}$ is 100 for $\hat{t}_{k\mathrm{h}}$. Table 2 gives the theoretical and simulated PREs of the small area total estimators for different domains under direct response (without using randomized response techniques) with both data sets. PREs in Tables 3 and 4 are obtained in similar manner using the Variances and MSEs under RRT. The theoretical and simulated values of PRE are reported in Tables 2–4 with notations TR and SR respectively.

Table 2: PREs of the SAE under direct response.

		Type	$\mathrm{PRE_h}$	$\mathrm{PRE_r}$	$\mathrm{PRE_{lr}}$
Data I	$k = 1$	TR	100	215.864	216.839
		SR	100	217.230	218.106
	$k = 2$	TR	100	378.592	379.123
		SR	100	370.443	375.775
Data II	$k = 1$	TR	100	13853.214	13862.216
		SR	100	12993.771	14382.960
	$k = 2$	TR	100	5134.249	5137.867
		SR	100	4855.831	5352.376
	$k = 3$	TR	100	6770.974	6770.974
		SR	100	6371.760	7076.799

From Table 2, one can infer that for both data sets, total estimators under RPM and LPM (see the last two columns) which utilize auxiliary information provide smaller variance than the MSE of Total estimator under HPM. Further, estimator obtained through LPM outperforms the other two competitors in all cases.

Tables 3 and Table 4 give a comparison of the three competing population models in term of PREs for Data I and Data II respectively under randomized response. Going from top to bottom in Tables 3 and 4, we observe that the PREs decrease with increase in variability in the scrambling variables. Also, comparing Table 2 with Tables 3 and 4, we can infer that the efficiency of the domain estimators decreases when using randomized response technique. But that is expected given that RRT introduces noise in the data. Without RRT, the real loss of efficiency will be much larger due to “invisible” response bias.

Table 3: PRE of the SAE under randomized response for Data I.

	a	b	Type	PRE_{h}	PRE_{r}	PRE_{lr}
$k = 1$	$U(2,3)$	$U(0,1)$	TR	100	210.572	211.480
			SR	100	210.726	211.501
		$U(0,5)$	TR	100	208.031	208.907
			SR	100	207.984	208.938
	$U(1,4)$	$U(0,1)$	TR	100	181.451	182.026
			SR	100	179.096	179.553
		$U(0,5)$	TR	100	180.063	180.625
			SR	100	177.685	178.235
$k = 2$	$U(2,3)$	$U(0,1)$	TR	100	363.439	363.921
			SR	100	355.997	360.266
		$U(0,5)$	TR	100	360.746	361.220
			SR	100	351.889	357.175
	$U(1,4)$	$U(0,1)$	TR	100	284.007	284.270
			SR	100	275.869	277.921
		$U(0,5)$	TR	100	282.689	282.949
			SR	100	273.902	276.442

Table 4: PRE of the SAE under randomized response for Data II.

	a	b	Type	PRE_{h}	PRE_{r}	PRE_{lr}
$k = 1$	$U(2,3)$	$U(0,1)$	TR	100	3740.45	3740.97
			SR	100	2624.11	2797.29
		$U(0,5)$	TR	100	3403.93	3404.35
			SR	100	2368.88	2524.28
	$U(1,4)$	$U(0,1)$	TR	100	631.56	631.57
			SR	100	435.21	460.29
		$U(0,5)$	TR	100	623.77	623.78
			SR	100	429.72	454.59
$k = 2$	$U(2,3)$	$U(0,1)$	TR	100	2578.63	2579.54
			SR	100	1991.16	2127.92
		$U(0,5)$	TR	100	2318.17	2318.90
			SR	100	1758.21	1875.70
	$U(1,4)$	$U(0,1)$	TR	100	593.55	593.59
			SR	100	424.67	447.19
		$U(0,5)$	TR	100	582.27	582.31
			SR	100	416.73	438.87
$k = 3$	$U(2,3)$	$U(0,1)$	TR	100	2930.02	2930.02
			SR	100	2203.50	2354.74
		$U(0,5)$	TR	100	2642.70	2642.70
			SR	100	1967.09	2095.90
	$U(1,4)$	$U(0,1)$	TR	100	608.22	608.22
			SR	100	432.67	455.85
		$U(0,5)$	TR	100	598.12	598.12
			SR	100	426.19	448.82

5. CONCLUSION

In this study, an attempt for obtaining separate total estimates for the sensitive study variable in each domain (small area) is made using the model relationship between the sensitive study variable and the auxiliary variable. It is observed that the small area total estimators under randomized response techniques possess larger variance (as they should) as compared to the estimators obtained through direct responses. As the privacy and efficiency move in opposite directions, one can't improve both at the same time. Our proposed estimators provide greater efficiency in estimating small area totals when an appropriate model relationship between the study variable and the auxiliary variable is used. Our numerical study with two real life data sets supports the theoretical findings. This is clear from the fact that both $\mathrm{PRE}_{\mathrm{r}}$ and $\mathrm{PRE}_{\mathrm{lr}}$ are greater than $\mathrm{PRE}_{\mathrm{h}}$.

ACKNOWLEDGMENTS

The authors would like to express their most sincere gratitude to the editor and the referees for their careful reading of the paper and for making many useful suggestions, all of which helped improve the presentation of the paper substantially.

REFERENCES

[1] Ahmed, S.; Shabbir, J. and Gupta, S. (2017). Use of scrambled response model in estimating the finite population mean in presence of non response when coefficient of variation is known, *Communication in Statistics – Theory and Methods*, **46**(17), 8435–8449.

[2] Basu, D. (1971). *An Essay on the Logical Foundations of Survey Sampling, Part I.* In "Foundations of Statistical Inference" (V.P. Godambe and D. Sprott, Eds.), Holt, Rinehart and Winston, Toronto, 203–233.

[3] Bar-Lev, S.K.; Bobovitch, E. and Boukai, B. (2004). A note on randomized response models for quantitative data, *Metrika*, **60**, 255–260.

[4] Bryce, G.R. (1980). "Some observations on the analysis of growth curves", Paper No. SD-025-R, Brigham Young University, Department of Statistics, **41**, 627–636.

[5] Chambers, R. and Clark, R. (2012). An introduction to model-based survey sampling with applications, *OUP Oxford*, **37**.

[6] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.

[7] Chaudhuri, A. and Roy, D. (1997). Model assisted survey sampling strategies with randomized response, *Journal of Statistical Planning Inference*, **60**, 61–68.

[8] Cumberland, W.G. and Royall, R.M. (1981). Prediction models and unequal probability sampling, *Journal of the Royal Statistical Society: Series B (Methodological)*, **43**(3), 353–367.

[9] Diana, G. and Perri, P.F. (2009). Estimating a sensitive proportion through randomized response procedures based on auxiliary information, *Statistical Papers*, **50**, 661–672.

[10] Diana, G. and Perri, P.F. (2010). New scrambled response models for estimating the mean of a sensitive quantitative character, *Journal of Applied Statistics*, **37**, 1875–1890.

[11] Diana, G. and Perri, P.F. (2011). A class of estimators for quantitative sensitive data, *Statistical Papers*, **52**, 633–650.

[12] Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data, *Journal of Statistical Planning Inference*, **7**, 307–316.

[13] Fay, R.E. and Herriot, R.A. (1979). Estimation of income for small places: An application of James Stein procedures to census data, *Journal of the American Statistical Association*, **74**, 268–277.

[14] Fuller, W.A. (1970). "Simple Estimators for the Mean of Skewed Populations", Technical report, Iowa State University, Dept. of Statistics.

[15] Godambe, V.P. (1955). A unified theory of sampling from finite populations, *Journal of the Royal Statistical Society: Series B*, **17**, 269–278.

[16] Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, I, *The Annals of Mathematical Statistics*, **36**, 1707–1722.

[17] Greenberg, B.G.; Abul-Ela, A.L.A.; Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework, *Journal of the American Statistical Association*, **64**, 520–539.

[18] Greenberg, B.G.; Kuebler, R.R.; Abernathy, J.R. and Horvitz, D.G. (1971). Application of the randomized response technique in obtaining quantitative data, *Journal of the American Statistical Association*, **66**, 243–250.

[19] Gupta, S.; Gupta, B. and Singh, S. (2002). Estimation of sensitive level of personal interview survey questions, *Journal of Statistical Planning Inference*, **100**, 239–247.

[20] Gupta, S.N.; Shabbir, J. and Sehra, S. (2010). Mean and sensitivity estimation in optional randomized response models, *Journal of Statistical Planning Inference*, **140**, 2870–2874.

[21] Huang, K.C. (2010). Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling, *Metrika*, **71**, 341–352.

[22] Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure, *Biometrika*, **77**, 439–442.

[23] Pfeffermann, D.; Bell, P. and Signorelli, D. (1996). Labour force trend estimation in small areas, *Proceedings of the Annual Research Conference, US Bureau of the Census*, 407–431.

[24] Purcell, N.I. and Kish, L. (1980). Postcensal estimates for local areas (or domains), *International Statistical Review*, **48**, 3–18.

[25] Rao, J. (1994). Small area estimation by combining time series and cross sectional data, *Canadian Journal of Statistics*, **22**, 511–528.

[26] Rao, J. (2003). *Small Area Estimation*, New York, Wiley.

[27] Royall, R.M. (1970). An old approach to finite population sampling theory, *Journal of the American Statistical Association*, **63**, 1269–1279.

[28] Royall, R.M. (1992). The model based (prediction) approach to finite population sampling theory, *Lecture Notes – Monograph Series*, **17**, 225–240.

[29] Sarndal, C.E.; Thomsen, I.; Hoem, J.M.; Lindley, D.V.; Barndorff-Nielsen, O. and Dalenius, T. (1978). Design-based and model-based inference in survey sampling [with discussion and reply], *Scandinavian Journal of Statistics*, 27–52.

[30] Smith, T.M.F. (1976). The foundations of survey sampling: A review, *Journal of the Royal Statistical Society: Series A*, **139**, 183–195.

[31] Smith, T.M.F. (1983). On the validity of inferences from non-random samples, *Journal of the Royal Statistical Society: Series A (General)*, **146**(4), 394–403.

[32] Valliant, R. (2009). Model-Based Prediction of Finite Population Totals, *Sample Surveys: Inference and Analysis*, **29B**, 23–31.

[33] Valliant, R.; Dorfman, A.H. and Royall, M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, New York.

[34] Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, **60**, 63–69.

[35] Warner, S.L. (1971). The linear randomized response model, *Journal of the American Statistical Association*, **66**(336), 884–888.

[36] Whitworth, A.; Carter, E.; Ballas, D. and Moon, G. (2017). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. Computers, *Environment and Urban Systems*, **63**, 50–57.

[37] Yeh, I.-C.; Yang, King-Jang and Ting, T.-M. (2008). Knowledge discovery on RFM model using Bernoulli sequence, *Expert Systems with Applications*, **36**, 5866–5871.

[38] You, Y. (2008). Small area estimation using area level models with model checking and applications, *Proceedings of the Survey Methods Section*, Statistical Society of Canada.