Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 75(3) 2021 235–243

Two Step Calibration for Estimation of Finite Population Total under Two-Stage Sampling Design

Pradip Basak[1], Kaustav Aditya[2], Vandita Kumari[2] and Deepak Singh[2]
[1] *Uttar Banga Krishi Viswavidyalaya, Pundibari, Cooch Behar*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

Received 03 October 2021; Revised 18 April 2022; Accepted 24 April 2022

Calibration is a popular approach in sample surveys to produce efficient estimators of population parameter using population aggregates of auxiliary variable. However, many a times, such population aggregates of auxiliary variable is not available. Moreover, it may happen that there exists additional auxiliary variable which is less closely related to the study variable but having known population aggregates. Under such circumstances, information on both the auxiliary variables may be incorporated in the estimation process using two step calibration approach. For two-stage sampling design, efficient estimators of population total have been developed using two step calibration approach for the situation of unavailability of population aggregates of auxiliary variable for all the primary stage units (psu's) in the population. The approximate variance and the estimate of variance of the proposed calibration estimators have also been developed. Empirical results using both model-based and design-based simulations, with the latter based on real data set, show that the proposed calibration estimators illustrate superior performance than the existing estimators.

Keywords: Two step calibration, Population total, Two-stage sampling.

1. INTRODUCTION

Multi-stage sampling design is widely used in most of the large scale surveys to which two-stage sampling is the simplest case and frequently used in many real life surveys. For example, two-stage sampling design is often used in crop area estimation surveys in many developing countries. In this design, sample is selected in two stages. At the first stage, psu's are selected and at the second stage, second stage unit (ssu's) are drawn from each of the selected psu's. Usually, use of auxiliary information improves the estimator of population total under two-stage design (Sukhatme *et al.*, 1984). Many researchers have also used the predictive approach with availability of auxiliary information under two-stage design, see for example, Srivastava and Garg (2009), Sahoo *et al.* (2011) and references therein. Calibration approach proposed by Deville and Sarndal (1992) is widely used in sample surveys to produce efficient estimators of population parameter by incorporating auxiliary information at the estimation stage. Aditya *et al.* (2016) described regression type estimators of the population total for two-stage sampling using the calibration approach under the assumption that the population level auxiliary information is available at psu level. Mourya *et al.* (2016) developed calibration estimator for finite population total under two-stage sampling when the auxiliary information is available at the element level for the selected first-stage units in the random sample. Basak *et al.* (2017, 2018) developed calibration estimator of population regression coefficient under two-stage sampling design for different cases of availability of auxiliary information at psu and ssu level.

Biswas *et al.* (2020) developed calibration estimators of the finite population total under two stage sampling design assuming study variable is inversely related to the auxiliary variable and the population level auxiliary information is available at the second stage of selection. One of the assumptions of calibration approach is that population level information such as population aggregate is required for the auxiliary

Corresponding author: Pradip Basak
E-mail address: pradip@ubkv.ac.in

variable that is closely related to the study variable. In many practical applications, population aggregates of such auxiliary variable are often not available for all the psu's under two-stage sampling design. In such situations, so far, estimation of population total has been limited to the use of double sampling approach, see for example, Saini and Bahl (2012), Saini (2013) and references therein. Guha and Chandra (2019) proposed an improved estimator for the population total based on double sampling when auxiliary information is available for the first variable and not available for the second variable. Many often under two-stage design, there is availability of additional auxiliary variable which is less closely related to the study variable but having population aggregates at the psu level. The existing methods of survey estimation do not make proper use of such additional auxiliary information. Estevao and Särndal (2002) delineated the two step calibration approach to the additional auxiliary variable for estimating the population total. This paper considers the problem of estimating the population total under two-stage sampling design when population aggregates of auxiliary variables are not available and develops efficient estimators of population total using the information on additional auxiliary variable through two step calibration approach. In particular, it is assumed that population aggregates of auxiliary variable is unavailable for all the psu's in the population whereas for additional auxiliary variable, this information is available for all the psu's. Therefore, two different sampling designs are considered to address this situation and these are referred as Sampling Design 1 and 2 respectively.

The rest of this paper is organized as follows. Next Section describes the general notations used for the development of calibration estimators under two-stage sampling design. Section 3 presents the proposed estimators developed using two step calibration approach along with its approximate variance and the estimate of variance. Section 4 reports the results of the simulation studies to assess the empirical performance of the developed estimators. Finally, Section 5 provides the main concluding remarks.

2. PROPOSED ESTIMATORS

2.1 *Notations*

Let us consider a finite population *U=(1,2,…,k,…,N)* which is grouped into N_I clusters as $U_1, U_2, \ldots, U_i, \ldots, U_{N_I}$ with sizes of the clusters as $N_1, N_2, N_i, \ldots, N_{N_I}$ respectively. Thus, $U = \bigcup_{i=1}^{N_I} U_i$ and $N = \sum_{i=1}^{N_I} N_i$. These clusters are called psu's and the sampling units within the clusters (psu's) are called ssu's. At the first stage, a sample of psu's s_I of size n_I is selected from the population of psu's U_I of size N_I by using a suitable probability sampling scheme. Let, the first and second order inclusion probability at the first stage be π_{Ii} and π_{Iij} respectively. At the second stage, a sample of units s_i of size n_i is drawn from the i^{th} selected psu, U_i of size N_i, $\forall i \in s_I$ by using any probability sampling scheme. Thus, $s = \bigcup_{i=1}^{n_I} s_i$ and $n_s = \sum_{i=1}^{n_I} n_i$, where s is the two-stage sample and n_s is the two-stage sample size. Let, the first and second order inclusion probability at the second stage be $\pi_{k/i}$ and $\pi_{kl/i}$ respectively.

Let y and x be the study and auxiliary variable respectively. Here, it is assumed that there is availability of additional auxiliary variable z which is less linearly related to the study variable y but having population aggregates at the psu level. Let, y_{ik}, x_{ik} and z_{ik}, $\forall i \in s_I$, $k \in s_i$ be values of the variables corresponding to k-th unit of i-th selected psu. The population total of y is given by, $t_y = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{N_I} t_{iy}$, where $t_{iy} = \sum_{k=1}^{N_i} y_{ik}$ is the i^{th} psu total of y. Similarly, population total of x is given by $X = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik} = \sum_{i=1}^{N_I} X_i$, where $X_i = \sum_{k=1}^{N_i} x_{ik}$ is the i^{th} psu total of x and population total of z is $Z = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} = \sum_{i=1}^{N_I} Z_i$, where $Z_i = \sum_{k=1}^{N_i} z_{ik}$ is the i-th psu total of z.

With this, our interest is to estimate the population total, t_y. Following Särndal *et al.* (1992), the π-estimator of population total t_y under two-stage sampling design is given by

$$\hat{t}_{y\pi} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i} y_{ik}, \qquad (1)$$

where, $a_{Ii} = 1/\pi_{Ii}$ and $a_{k/i} = 1/\pi_{k/i}$.

Here, it is assumed that population aggregates of auxiliary variable x are unavailable for all psu's in the population, *i.e.*, X_i is unknown $\forall i = 1, 2, \ldots, N_I$

whereas for additional auxiliary variable *z*, this information is available for all the psu's, *i.e.* Z_i is known $\forall i = 1, 2, ..., N_I$.

2.2 Calibration estimation under Sampling Design 1

In Sampling Design 1, a two-phase sample is drawn at the psu level to collect population aggregates of auxiliary variable *x* for the first phase psu. The layout is described as below.

At first stage: $U_I(N_I)$

$\downarrow$

First phase: $s'_I(n'_I)$ observe X_i

$\downarrow$

Second phase: $s_I(n_I)$

At second stage:

$U_i(N_i)\forall i \in s_I$

$\downarrow$

$s_i(n_i)$ observe y_{ik}, x_{ik}, z_{ik}

Let, a_{1Ii} denotes the design weight at the first phase at psu level and a_{2Ii} denotes the conditional design weight at the second phase at psu level. Thus, $a_{1Ii} = 1/\pi_{1Ii}$ and $a_{2Ii} = 1/\pi_{2Ii}$ where, π_{1Ii} is the inclusion probability at first phase and π_{2Ii} is the conditional inclusion probability at second phase. Overall design weight for *i*-th psu is given by $a_{Ii} = a_{1Ii}a_{2Ii}$. Design weight at the second stage is given by $a_{k/i}$, where, $a_{k/i} = 1/\pi_{k/i}$. Thus, overall design weight corresponding to the *k*-th unit of *i*-th selected psu is given by $a_{ik} = a_{Ii}a_{k/i}$. Let, w_{1Ii} denotes the calibrated weight corresponding to the design weight at the first phase at psu level, a_{1Ii} and $w_{ik}^{(1)}$ denotes the overall calibrated weight corresponding to overall design weight, a_{ik}.

In the first step calibration, chi-square distance function measuring the distance between w_{1Ii} and a_{1Ii}, $\sum_{i=1}^{n'_I}\left(w_{1Ii} - a_{1Ii}\right)^2 / 2a_{1Ii}q_{1Ii}$ is minimized subject to the calibration constraints $\sum_{i=1}^{n'_I} w_{1Ii}Z_i = \sum_{i=1}^{N_I} Z_i = Z$ and $\sum_{i=1}^{n'_I} w_{1Ii} = N_I$. The first step calibrated weights, w_{1Ii} are obtained by using the Lagrangian multiplier approach.

Thus, the objective function for minimization is given by

$$\phi = \sum_{i=1}^{n'_I}\frac{\left(w_{1Ii} - a_{1Ii}\right)^2}{2a_{1Ii}q_{1Ii}} - \lambda'_1\left(\sum_{i=1}^{n'_I} w_{1Ii}Z_i - Z\right) - \lambda'_2\left(\sum_{i=1}^{n'_I} w_{1Ii} - N_I\right).$$

The first step calibrated weights, w_{1Ii} obtained by minimising this objective function subject to the calibration constraints are given as

$$w_{1Ii} = a_{1Ii}\left[1 + q_{1Ii}\left(\lambda'_1 Z_i + \lambda'_2\right)\right],$$

where,

$$\lambda'_1 = \frac{\hat{t}'_{q\pi}\left(Z - \hat{t}'_{z\pi}\right) - \hat{t}'_{qz\pi}\left(N_I - \hat{t}'_{\pi}\right)}{\hat{t}'_{q\pi}\hat{t}'_{qzz\pi} - \hat{t}'^2_{qz\pi}}, \quad \lambda'_2 = \frac{\hat{t}'_{qzz\pi}\left(N_I - \hat{t}'_{\pi}\right) - \hat{t}'_{qz\pi}\left(Z - \hat{t}'_{z\pi}\right)}{\hat{t}'_{q\pi}\hat{t}'_{qzz\pi} - \hat{t}'^2_{qz\pi}},$$

$$\hat{t}'_{qzz\pi} = \sum_{i=1}^{n'_I} a_{1Ii}q_{1Ii}Z_i^2, \quad \hat{t}'_{qz\pi} = \sum_{i=1}^{n'_I} a_{1Ii}q_{1Ii}Z_i, \quad \hat{t}'_{q\pi} = \sum_{i=1}^{n'_I} a_{1Ii}q_{1Ii},$$

$$\hat{t}'_{\pi} = \sum_{i=1}^{n'_I} a_{1Ii}, \text{ and } \hat{t}'_{z\pi} = \sum_{i=1}^{n'_I} a_{1Ii}Z_i.$$

Here, q_{1Ii} is an unknown positive constant and we have assumed $q_{1Ii} = 1$, as a particular case. The calibrated weights, w_{1Ii} obtained in the first step calibration are multiplied with the population aggregates of auxiliary variable (X_i) for the first phase psu to estimate the population total of *x* as $X^* = \sum_{i=1}^{n'_I} w_{1Ii}X_i$. The estimated population total of auxiliary variable, X^* is subsequently used in second step calibration as a constraint.

In the second step calibration, chi-square distance function between $w_{ik}^{(1)}$ and a_{ik} is given by $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}\frac{(w_{ik}^{(1)} - a_{ik})^2}{2a_{ik}q_{ik}}$ and the calibration constraints are $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)}z_{ik} = Z$, $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)}x_{ik} = X^*$, and $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)} = N$. Thus, the objective function

$$\phi = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i}\frac{(w_{ik}^{(1)} - a_{ik})^2}{2a_{ik}q_{ik}} - \lambda_1\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)}z_{ik} - Z\right) - \lambda_2\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)}x_{ik} - X^*\right) - \lambda_3\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)} - N\right)$$

is minimized by using the Lagrangian multiplier approach to obtain the final calibrated weight, $w_{ik}^{(1)}$. Finally, the calibrated weights are obtained as

$$w_{ik}^{(1)} = a_{ik}\left[1 + q_{ik}\left(\lambda_1 z_{ik} + \lambda_2 x_{ik} + \lambda_3\right)\right],$$

where,

$$\lambda_1 = \frac{\left(X^* - \hat{t}_x\right)\left(\hat{t}_{qx}\hat{t}_{qz} - \hat{t}_{qxz}\hat{t}_q\right) + \left(Z - \hat{t}_z\right)\left(\hat{t}_{qxx}\hat{t}_q - \hat{t}_{qx}^2\right) + \left(N - \hat{t}\right)\left(\hat{t}_{qxz}\hat{t}_{qx} - \hat{t}_{qxx}\hat{t}_{qz}\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q + 2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz} - \hat{t}_{qxx}\hat{t}_{qz}^2 - \hat{t}_{qzz}\hat{t}_{qx}^2 - \hat{t}_q\hat{t}_{qxz}^2},$$

$$\lambda_2 = \frac{\left(X^* - \hat{t}_x\right)\left(\hat{t}_{qzz}\hat{t}_q - \hat{t}_{qz}^2\right) + \left(Z - \hat{t}_z\right)\left(\hat{t}_{qx}\hat{t}_{qz} - \hat{t}_{qxz}\hat{t}_q\right) + \left(N - \hat{t}\right)\left(\hat{t}_{qxz}\hat{t}_{qz} - \hat{t}_{qzz}\hat{t}_{qx}\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q + 2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz} - \hat{t}_{qxx}\hat{t}_{qz}^2 - \hat{t}_{qzz}\hat{t}_{qx}^2 - \hat{t}_q\hat{t}_{qxz}^2},$$

$$\lambda_3 = \frac{\left(X^* - \hat{t}_x\right)\left(\hat{t}_{qxx}\hat{t}_{qz} - \hat{t}_{qzz}\hat{t}_{qx}\right) + \left(Z - \hat{t}_z\right)\left(\hat{t}_{qxz}\hat{t}_{qx} - \hat{t}_{qxx}\hat{t}_{qz}\right) + \left(N - \hat{t}\right)\left(\hat{t}_{qzz}\hat{t}_{qxx} - \hat{t}_{qxz}^2\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q + 2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz} - \hat{t}_{qxx}\hat{t}_{qz}^2 - \hat{t}_{qzz}\hat{t}_{qx}^2 - \hat{t}_q\hat{t}_{qxz}^2}.$$

Here,

$$\hat{t}_{qzz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} z_{ik}^2 \ , \ \hat{t}_{qxx} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} x_{ik}^2 \ ,$$

$$\hat{t}_{qxz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} x_{ik} z_{ik} \ , \ \hat{t}_{qx} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} x_{ik} \ ,$$

$$\hat{t}_{qz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} z_{ik} \ , \ \hat{t}_q = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} q_{ik} \ , \ \hat{t}_x = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} x_{ik} \ ,$$

$$\hat{t}_z = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} z_{ik} \ \text{ and } \ \hat{t} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} \ .$$

In this case, q_{ik} is an unknown positive constant and we have taken $q_{ik} = 1,$ as a particular choice. Thus, the calibrated estimator of population total of y, $\hat{t}_{y\pi}^{c(1)}$ is given by

$$\hat{t}_{y\pi}^{c(1)} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(1)} y_{ik} \ . \tag{2}$$

Under this design, the usual double sampling ratio and regression estimator of population total is given by

$$\hat{t}_{y\pi}^{r(1)} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}\hat{t}'_{x\pi}, \tag{3}$$

$$\hat{t}_{y\pi}^{reg(1)} = \hat{t}_{y\pi} + b\left(\hat{t}'_{x\pi} - \hat{t}_{x\pi}\right), \tag{4}$$

where,

$$\hat{t}_{y\pi} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} y_{ik} \ , \ \hat{t}_{x\pi} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik} x_{ik} \ , \ \hat{t}'_{x\pi} = \sum_{i=1}^{n'_I} a_{1Ii} X_i$$

and

$$b = \frac{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)\left(y_{ik} - \hat{t}_{y\pi}/N\right)}{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)^2}.$$

The proposed calibrated estimator of population total have non-linear form. Therefore, Taylor series linearization approach is used to derive an approximate variance of the estimator. Following Taylor series linearization methodology the approximate variance of the developed calibrated estimator of population total of y, $\hat{t}_{y\pi}^{c(1)}$ is given by

$$V(\hat{t}_{y\pi}^{c(1)}) = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{t_{Ei}}{\pi_{Ii}}\frac{t_{Ej}}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{E_{k/i}}{\pi_{k/i}}\frac{E_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta'_{Iij}\frac{X_i}{\pi'_{Ii}}\frac{X_j}{\pi'_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{x_k}{\pi_{k/i}}\frac{x_l}{\pi_{l/i}}\right]$$

where,

$$\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij} \ , \ \Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i} \ ,$$

$$\Delta'_{Iij} = \pi'_{Iij} - \pi'_{Ii}\pi'_{Ij} \ ,$$

$$t_{Ei} = Y_i - B_{iyx.z}X_i - B_{iyz.x}Z_i \ , \ Y_i = \sum_{k=1}^{N_i} y_{ik} \ ,$$

$$B_{iyx.z} = \frac{\left\{\left(\sum_{i=1}^{N_I} Y_i X_i\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{N_I} Y_i Z_i\right)\left(\sum_{i=1}^{N_I} X_i Z_i\right)\right\}}{\left\{\left(\sum_{i=1}^{N_I} X_i^2\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{N_I} X_i Z_i\right)^2\right\}},$$

$$B_{iyz.x} = \frac{\left\{\left(\sum_{i=1}^{N_I} Y_i Z_i\right)\left(\sum_{i=1}^{N_I} X_i^2\right) - \left(\sum_{i=1}^{N_I} Y_i X_i\right)\left(\sum_{i=1}^{N_I} X_i Z_i\right)\right\}}{\left\{\left(\sum_{i=1}^{N_I} X_i^2\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{N_I} X_i Z_i\right)^2\right\}},$$

$$E_{k/i} = y_{ik} - B_{ikyx.z}x_{ik} - B_{ikyz.x}z_{ik} \ ,$$

$$B_{ikyx.z} = \frac{\left\{\left(\sum_{k=1}^{N_i} y_{ik}x_{ik}\right)\left(\sum_{k=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} y_{ik}z_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{N_i} x_{ik}^2\right)\left(\sum_{i=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)^2\right\}},$$

$$B_{ikyz.x} = \frac{\left\{\left(\sum_{k=1}^{N_i} y_{ik}z_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}^2\right) - \left(\sum_{k=1}^{N_i} y_{ik}x_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{N_i} x_{ik}^2\right)\left(\sum_{k=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)^2\right\}}.$$

The variance estimator of the calibrated estimator $\hat{t}_{y\pi}^{c(1)}$ is obtained as

$$\hat{V}(\hat{t}_{y\pi}^{c(1)}) = \sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{t}_{Ei}}{\pi_{Ii}}\frac{\hat{t}_{Ej}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\hat{E}_{k/i}}{\pi_{k/i}}\frac{\hat{E}_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{n'_I}\sum_{j=1}^{n'_I}\breve{\Delta}'_{Iij}\frac{X_i}{\pi'_{Ii}}\frac{X_j}{\pi'_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{x_k}{\pi_{k/i}}\frac{x_l}{\pi_{l/i}}\right]$$

where, $\breve{\Delta}_{Iij} = \dfrac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}$, $\breve{\Delta}_{kl/i} = \dfrac{\pi_{kl/i} - \pi_{k/i}\pi_{l/i}}{\pi_{kl/i}}$,

$$\breve{\Delta}'_{Iij} = \frac{\pi'_{Iij} - \pi'_{Ii}\pi'_{Ij}}{\pi'_{Iij}},$$

$$\hat{t}_{Ei} = \hat{t}_{iy\pi} - \hat{B}_{iyx.z}X_i - \hat{B}_{iyz.x}Z_i\,,\ \hat{t}_{iy\pi} = \sum_{k=1}^{n_i} a_{k/i}y_{ik}\,,$$

$$\hat{B}_{iyx.z} = \frac{\left\{\left(\sum_{i=1}^{n_1} a_{Ii}\hat{t}_{iy\pi}X_i\right)\left(\sum_{i=1}^{N_1} Z_i^2\right) - \left(\sum_{i=1}^{n_1} a_{Ii}\hat{t}_{iy\pi}Z_i\right)\left(\sum_{i=1}^{n'_1} a'_{Ii}X_iZ_i\right)\right\}}{\left\{\left(\sum_{i=1}^{n'_1} a'_{Ii}X_i^2\right)\left(\sum_{i=1}^{N_1} Z_i^2\right) - \left(\sum_{i=1}^{n'_1} a'_{Ii}X_iZ_i\right)^2\right\}},$$

$$\hat{B}_{iyz.x} = \frac{\left\{\left(\sum_{i=1}^{n_1} a_{Ii}\hat{t}_{iy\pi}Z_i\right)\left(\sum_{i=1}^{n'_1} a'_{Ii}X_i^2\right) - \left(\sum_{i=1}^{n_1} a_{Ii}\hat{t}_{iy\pi}X_i\right)\left(\sum_{i=1}^{n'_1} a'_{Ii}X_iZ_i\right)\right\}}{\left\{\left(\sum_{i=1}^{n'_1} a'_{Ii}X_i^2\right)\left(\sum_{i=1}^{N_1} Z_i^2\right) - \left(\sum_{i=1}^{n'_1} a'_{Ii}X_iZ_i\right)^2\right\}},$$

$$\hat{E}_{k/i} = y_{ik} - \hat{B}_{ikyx.z}x_{ik} - \hat{B}_{ikyz.x}z_{ik}\,,$$

$$\hat{B}_{ikyx.z} = \frac{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)^2\right\}},$$

$$\hat{B}_{ikyz.x} = \frac{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)^2\right\}}.$$

2.3 Calibration estimation under Sampling Design 2

Here, a two-phase sample is drawn at the psu level, and at the ssu level, a first phase sample is drawn from first phase psu to collect unit level information on auxiliary variable *x* and *z*, whereas from second phase psu, a second phase sample is drawn to collect unit level information on study variable *y* only. The layout is described as below.

At first stage: $U_I(N_I)$

↓

First phase: $s'_I(n'_I)$

↓

Second phase: $s_I(n_I)$

At second stage:

From first phase psu | **From second phase psu**

$U_i(N_i)\forall i \in s'_I$ | $U_i(N_i)\forall i \in s_I$

↓ $n'_i > n_i$ ↓

$s'_i(n'_i)$ observe x_{ik}, z_{ik} | $s_i(n_i)$ observe y_{ik}

Let, a_{1ik} denotes the design weight corresponding to the units in the first phase sample drawn from first phase psu's and a_{2ik} denotes the design weight corresponding to the units in the second phase sample drawn from second phase psu's. Overall design weight, a_{ik} is given by $a_{ik} = a_{1ik}a_{2ik}$. Let, w_{1ik} denotes the calibrated weight corresponding to the design weight of the first phase units drawn from first phase psu, a_{1ik} and $w_{ik}^{(2)}$ denotes the overall calibrated weight corresponding to the overall design weight, a_{ik}.

First Step Calibration

In the first step calibration, the chi-square distance function measuring the distance between w_{1ik} and a_{1ik} is given by $\sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i}\frac{(w_{1ik}-a_{1ik})^2}{2a_{1ik}q_{1ik}}$. Here the calibration constraints are, $\sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik}z_{ik} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} = Z$ and $\sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik} = N$. Thus, the objective function

$$\phi = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i}(w_{1ik}-a_{1ik})^2/2a_{1ik}q_{1ik} - \lambda'_{1ik}\left(\sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik}z_{ik} - Z\right) - \lambda'_{2ik}\left(\sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik} - N\right)$$

is minimized by using the method of Lagrange multiplier to obtain the first step calibrated weight, w_{1ik}. The first step calibrated weights are obtained as

$$w_{1ik} = a_{1ik}\left[1 + q_{1ik}\left(\lambda'_{1ik}z_{ik} + \lambda'_{2ik}\right)\right],$$

where, $\lambda'_{1ik} = \frac{\hat{t}'_q\left(Z-\hat{t}'_z\right) - \hat{t}'_{qz}\left(N-\hat{t}'\right)}{\hat{t}'_q\hat{t}'_{qzz} - \hat{t}'^2_{qz}}$,

$$\lambda'_{2ik} = \frac{\hat{t}'_{qzz}\left(N-\hat{t}'\right) - \hat{t}'_{qz}\left(Z-\hat{t}'_z\right)}{\hat{t}'_q\hat{t}'_{qzz} - \hat{t}'^2_{qz}}\,,\ \hat{t}'_{qzz} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}q_{1ik}z_{ik}^2\,,$$

$$\hat{t}'_{qz} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}q_{1ik}z_{ik}\,,\ \hat{t}'_q = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}q_{1ik}\,,$$

$$\hat{t}'_z = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}z_{ik} \text{ and } \hat{t}' = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}\,.$$

Here, q_{1ik} is an unknown positive constant and we have assumed $q_{1ik} = 1$. The weights, w_{1ik} obtained in the first step calibration are used to estimate the population total of auxiliary variable as $X^{**} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik}x_{ik}$ which is required as a constraint in the second step calibration.

Second Step Calibration

From second step calibration, overall calibration weight ($w_{ik}^{(2)}$) is obtained. Thus, the chi-square distance function between $w_{ik}^{(2)}$ and a_{ik} is given by $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}\frac{(w_{ik}^{(2)}-a_{ik})^2}{2a_{ik}q_{ik}}$ and the calibration constraints are,

$$\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} z_{ik} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik} z_{ik} = Z,$$

$$\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} x_{ik} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik} x_{ik} = X^{**} \text{ and}$$

$$\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} w_{1ik} = N.$$

The objective function for minimization is given by

$$\phi = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i}(w_{ik}^{(2)}-a_{ik})^2/2a_{ik}q_{ik} - \lambda_{1ik}\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} z_{ik} - Z\right) - \lambda_{2ik}\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} x_{ik} - X^{**}\right) - \lambda_{3ik}\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} - N\right).$$

Finally, the calibrated weights are obtained as

$$w_{ik}^{(2)} = a_{ik}\left[1+q_{ik}\left(\lambda_{1ik}z_{ik}+\lambda_{2ik}x_{ik}+\lambda_{3ik}\right)\right],$$

where,

$$\lambda_{1ik} = \frac{\left(X^{**}-\hat{t}_x\right)\left(\hat{t}_{qx}\hat{t}_{qz}-\hat{t}_{qxz}\hat{t}_q\right)+\left(Z-\hat{t}_z\right)\left(\hat{t}_{qxx}\hat{t}_q-\hat{t}_{qx}^2\right)+\left(N-\hat{t}\right)\left(\hat{t}_{qxz}\hat{t}_{qx}-\hat{t}_{qxx}\hat{t}_{qz}\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q+2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz}-\hat{t}_{qxx}\hat{t}_{qz}^2-\hat{t}_{qzz}\hat{t}_{qx}^2-\hat{t}_q\hat{t}_{qxz}^2},$$

$$\lambda_{2ik} = \frac{\left(X^{**}-\hat{t}_x\right)\left(\hat{t}_{qzz}\hat{t}_q-\hat{t}_{qz}^2\right)+\left(Z-\hat{t}_z\right)\left(\hat{t}_{qx}\hat{t}_{qz}-\hat{t}_{qxz}\hat{t}_q\right)+\left(N-\hat{t}\right)\left(\hat{t}_{qxz}\hat{t}_{qz}-\hat{t}_{qzz}\hat{t}_{qx}\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q+2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz}-\hat{t}_{qxx}\hat{t}_{qz}^2-\hat{t}_{qzz}\hat{t}_{qx}^2-\hat{t}_q\hat{t}_{qxz}^2},$$

$$\lambda_{3ik} = \frac{\left(X^{**}-\hat{t}_x\right)\left(\hat{t}_{qxx}\hat{t}_{qz}-\hat{t}_{qzz}\hat{t}_{qx}\right)+\left(Z-\hat{t}_z\right)\left(\hat{t}_{qxz}\hat{t}_{qx}-\hat{t}_{qxx}\hat{t}_{qz}\right)+\left(N-\hat{t}\right)\left(\hat{t}_{qzz}\hat{t}_{qxx}-\hat{t}_{qxz}^2\right)}{\hat{t}_{qxx}\hat{t}_{qzz}\hat{t}_q+2\hat{t}_{qxz}\hat{t}_{qx}\hat{t}_{qz}-\hat{t}_{qxx}\hat{t}_{qz}^2-\hat{t}_{qzz}\hat{t}_{qx}^2-\hat{t}_q\hat{t}_{qxz}^2}.$$

Thus, the calibrated estimator of population total of y, $\hat{t}_{y\pi}^{c(2)}$ is given by

$$\hat{t}_{y\pi}^{c(2)} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}^{(2)} y_{ik}\,. \tag{5}$$

Under this design, the usual double sampling ratio and regression estimator of population total is given by

$$\hat{t}_{y\pi}^{r(2)} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}\hat{t}''_{x\pi}, \tag{6}$$

$$\hat{t}_{y\pi}^{reg(2)} = \hat{t}_{y\pi} + b\left(\hat{t}''_{x\pi}-\hat{t}_{x\pi}\right), \tag{7}$$

where, $\hat{t}''_{x\pi} = \sum_{i=1}^{n'_I}\sum_{k=1}^{n'_i} a_{1ik}x_{ik}$.

The approximate variance of the calibrated estimator $\hat{t}_{y\pi}^{c(2)}$ using Taylor series linearization method is obtained as

$$V(\hat{t}_{y\pi}^{c(2)}) = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{t_{Ei}}{\pi_{Ii}}\frac{t_{Ej}}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{E_{k/i}}{\pi_{k/i}}\frac{E_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta'_{Iij}\frac{X_i}{\pi'_{Ii}}\frac{X_j}{\pi'_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi'_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta'_{kl/i}\frac{x_k}{\pi'_{k/i}}\frac{x_l}{\pi'_{l/i}}\right]$$

where, $\Delta'_{kl/i} = \pi'_{kl/i} - \pi'_{k/i}\pi'_{l/i}$.

Variance estimator of the calibrated estimator $\hat{t}_{y\pi}^{c(2)}$ is obtained as

$$\hat{V}(\hat{t}_{y\pi}^{c(2)}) = \sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{t}'_{Ei}}{\pi_{Ii}}\frac{\hat{t}'_{Ej}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\hat{E}'_{k/i}}{\pi_{k/i}}\frac{\hat{E}'_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{n'_I}\sum_{j=1}^{n'_I}\breve{\Delta}'_{Iij}\frac{\hat{t}'_{ix\pi}}{\pi'_{Ii}}\frac{\hat{t}'_{jx\pi}}{\pi'_{Ij}} + \sum_{i=1}^{n'_I}\frac{1}{\pi'_{Ii}}\sum_{k=1}^{n'_i}\sum_{l=1}^{n'_i}\breve{\Delta}'_{kl/i}\frac{x_k}{\pi'_{k/i}}\frac{x_l}{\pi'_{l/i}}\right]$$

where, $\breve{\Delta}'_{kl/i} = \frac{\pi'_{kl/i}-\pi'_{k/i}\pi'_{l/i}}{\pi'_{kl/i}}$,

$$\hat{t}'_{Ei} = \hat{t}_{iy\pi} - \hat{B}'_{iyx.z}\hat{t}'_{ix\pi} - \hat{B}'_{iyz.x}Z_i\,,\quad \hat{t}'_{ix\pi} = \sum_{k=1}^{n'_i} a_{k/i}x_{ik}\,,$$

$$\hat{B}'_{iyx.z} = \frac{\left\{\left(\sum_{i=1}^{n_I} a_{Ii}\hat{t}_{iy\pi}\hat{t}'_{ix\pi}\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{n_I} a_{Ii}\hat{t}_{iy\pi}Z_i\right)\left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'_{ix\pi}Z_i\right)\right\}}{\left\{\left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'^2_{ix\pi}\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'_{ix\pi}Z_i\right)^2\right\}},$$

$$\hat{B}'_{iyz.x} = \frac{\left\{\left(\sum_{i=1}^{n_I} a_{Ii}\hat{t}_{iy\pi}Z_i\right)\left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'^2_{ix\pi}\right) - \left(\sum_{i=1}^{n_I} a_{Ii}\hat{t}_{iy\pi}\hat{t}'_{ix\pi}\right)\left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'_{ix\pi}Z_i\right)\right\}}{\left\{\left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'^2_{ix\pi}\right)\left(\sum_{i=1}^{N_I} Z_i^2\right) - \left(\sum_{i=1}^{n'_I} a'_{Ii}\hat{t}'_{ix\pi}Z_i\right)^2\right\}},$$

$$\hat{E}'_{k/i} = y_{ik} - \hat{B}'_{ikyx.z}x_{ik} - \hat{B}'_{ikyz.x}z_{ik}\,,$$

$$\hat{B}'_{ikyx.z} = \frac{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}z_{ik}\right)^2\right\}},$$

$$\hat{B}'_{ikyz.x} = \frac{\left\{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}z_{ik}\right)\right\}}{\left\{\left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n'_i} a'_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n'_i} a'_{k/i}x_{ik}z_{ik}\right)^2\right\}}.$$

3. EMPIRICAL EVALUATIONS

This Section summarizes the simulation studies conducted to evaluate the empirical performance of the developed estimators. Two types of simulation studies, namely design based simulation and model based simulation are considered. In the case of design based simulation, real survey dataset is used as a finite population. From this fixed population, repeated random samples are drawn with assumed sampling design. In the second type of simulation study, a synthetic population is generated and then a sample is drawn, and process is repeated several times. In this case, at each simulation run population data is first generated under the model and then a single sample is drawn from this simulated population.

In the simulation studies, the following estimators of population total under two-stage sampling design are considered.

i) π-estimator, $\hat{t}_{y\pi}$ (denoted as Est- π),

ii) Double sampling ratio estimator based on design 1 and 2, $\hat{t}_{y\pi}^{r(1)}$ and $\hat{t}_{y\pi}^{r(2)}$ (denoted as RAT1 and RAT2),

iii) Double sampling regression estimator based on design 1 and 2, $\hat{t}_{y\pi}^{reg(1)}$ and $\hat{t}_{y\pi}^{reg(2)}$ (denoted as REG1and REG2), and

iv) Developed calibrated estimator based on design 1 and 2, $\hat{t}_{y\pi}^{c(1)}$ and $\hat{t}_{y\pi}^{c(2)}$ (denoted as CAL1 and CAL2 respectively).

3.1 *Design based simulation*

The real survey dataset of 284 municipalities of Sweden popularly referred to as 'MU284 population' is used for design based simulation study. The 284 municipalities are divided into 50 clusters comprising of 5 to 9 municipalities. These 50 clusters are nothing but psu's and municipalities within the clusters are referred to as ssu's. The dataset contains information on multiple variables among which three variables are selected for the present study. Here, the variable revenues from the 1985 Municipal taxation (RMT85, measured in millions of kronor, *y*) is used as study variable. The aim is to estimate total revenues from the 1985 Municipal taxation using 1985 population (P85, in thousands, *x*) as the auxiliary variable and number of seats in the municipal council (S82, *z*) as the additional auxiliary variable. The correlations between the variables are presented in Table 1.

Table 1. Correlation between variables in MU284 data

Variables	RMT85 (*y*)	P85 (*x*)	S82 (*z*)
RMT85 (*y*)	1	0.96	0.58
P85 (*x*)	0.96	1	0.69
S82 (*z*)	0.58	0.69	1

From this population, for sampling design 1, in the first stage, 30 and 20 psu's are selected at first and second phase respectively. In the second stage, 2 units are drawn from each of the 20 selected psu's. In each stages, sample selection is done by simple random sampling without replacement (SRSWOR). For sampling design 2, in the first stage, 30 and 20 psu's are selected at first and second phase respectively. In the second stage, 4 units are selected from each of the 30 selected psu's at first phase and 2 units are drawn from each of the 20 selected psu's at second phase respectively. Here, $n_I' = 30$, $n_I = 20$, $n_i' = 4$, $n_i = 2$, and $n_s = 40$. Then the various estimators of the population total are computed using this sample data. The Monte Carlo simulation was run *M*=5000 times. Simulation studies are carried out in R software. The performance of the estimators are evaluated by percentage absolute relative bias (ARB) and percentage relative root mean squared error (RRMSE), defined by

$$ARB(\hat{T}) = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{\hat{T}_i - T}{T}\right| \times 100 \text{ and}$$

$$RRMSE(\hat{T}) = \sqrt{M^{-1}\sum_{i=1}^{M}\left(\frac{\hat{T}_i - T}{T}\right)^2} \times 100,$$

where $\hat{T}_i$ denotes the estimated value of population total at simulation run *i*, with true value *T* and *M* denotes the number of simulation run. The values of percentage absolute relative bias and percentage relative root mean square error of different estimators are reported in Table 2.

Table 2. Percentage absolute relative bias (ARB, %) and percentage relative root mean square error (RRMSE, %) of different estimators considered in design based simulation

Estimator	ARB, %	RRMSE, %
Est- π	27.63	33.40
RAT1	13.14	16.14
REG1	12.35	15.41
CAL1	10.62	13.50
RAT2	14.32	17.63
REG2	14.25	17.05
CAL2	11.36	14.59

The results from Table 2 show that the values of both percentage absolute relative bias and relative root mean square error are higher for π-estimator as compared to all other estimators considered in the simulation studies. For sampling design 1, the CAL1 estimator shows the better performance in terms of both the criteria followed by the double sampling regression and ratio estimator. For sampling design 2, the values of percentage absolute relative bias are lower for the CAL2 estimator followed by REG2 and RAT2 estimators. In the case of percentage relative root mean square error, it is highest for RAT2 and least for CAL2. Therefore, in terms of both the criteria CAL2 shows the better performance for sampling design 2. However, if we compare the estimators from both the sampling designs, then CAL1 gives the better performance.

3.2 *Model based simulation*

In this simulation study, a finite population of size *N*=1000 units is considered. The variable of interest y is generated from the model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_e^2)$, and $\beta_0 = \beta_1 = 1$ (assumed). The auxiliary variable x is generated from, $x_i = \alpha_0 + \alpha_1 z_i + \delta_i$, where $\delta_i \sim N(0,1)$, and $\alpha_0 = \alpha_1 = 1$ (assumed). Here, the additional auxiliary variable z is generated using Chi-squared distribution *i.e.* $z_i \sim \chi_p^2$. We have chosen $\sigma_e^2 = 2$ and $p = 5$ such that correlation coefficient between the variables in the generated population is given by

Table 3. Correlation between variables in the generated population

Variables	*y*	*x*	*z*
y	1	0.85	0.60
x	0.85	1	0.70
z	0.60	0.70	1

These *N*=1000 units are divided into 50 clusters comprising of 20 units each. Thus, here $N_I = 50$ and $N_i = 20$. From this population, three different combinations of sample are drawn- i) $n_I' = 40$, $n_I = 30$, $n_i' = 15$, $n_i = 10$, $n_s = 300$, ii) $n_I' = 30$, $n_I = 20$, $n_i' = 15$, $n_i = 10$, $n_s = 200$, iii) $n_I' = 30$, $n_I = 20$, $n_i' = 10$, $n_i = 5$, $n_s = 100$. Sample are drawn by using SRSWOR at both the stages. Here also, we considered the same set of estimators that were included in the design based simulation study. The simulation is repeated to a total number of 5000 times. The performance of the estimators are evaluated by percentage absolute relative bias (ARB) and percentage relative root mean squared error (RRMSE), defined by

$$ARB(\hat{T}) = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{\hat{T}_i - T_i}{T_i}\right| \times 100 \text{ and}$$

$$RRMSE(\hat{T}) = \sqrt{M^{-1}\sum_{i=1}^{M}\left(\frac{\hat{T}_i - T_i}{T_i}\right)^2} \times 100,$$

where $\hat{T}_i$ denotes the estimated value of population total at simulation run *i*, with true value T_i and *M* denotes the number of simulation run. The values of percentage absolute relative bias and relative root mean square error of the estimators for different combinations of sample sizes are reported in the following table.

Table 4. Percentage absolute relative bias (ARB, %) and percentage relative root mean square error (RRMSE, %) of different estimators in model based simulation

Sample size	Estimator	ARB, %	RRMSE, %
$n_I' = 40$, $n_I = 30$, $n_i' = 15$, $n_i = 10$, $n_s = 300$	Est- π	8.57	10.77
	RAT1	7.64	10.09
	REG1	6.02	7.18
	CAL1	5.20	6.50
	RAT2	9.11	11.88
	REG2	7.64	8.06
	CAL2	5.57	6.76
$n_I' = 30$, $n_I = 20$, $n_i' = 15$, $n_i = 10$, $n_s = 200$	Est- π	12.68	16.02
	RAT1	10.95	13.34
	REG1	10.13	12.43
	CAL1	8.81	11.55
	RAT2	11.22	13.53
	REG2	11.15	13.41
	CAL2	9.68	12.01
$n_I' = 30$, $n_I = 20$, $n_i' = 10$, $n_i = 5$, $n_s = 100$	Est- π	16.68	20.62
	RAT1	14.95	18.06
	REG1	14.13	17.13
	CAL1	12.80	16.24
	RAT2	15.22	18.23
	REG2	15.15	18.11
	CAL2	13.68	16.71

The results from the model based simulations reported in Table 4 reveal the conclusions identical to design based simulation in Table 3. In particular, these results in Table 4 indicate two points. First, the developed calibrated estimators perform better than the existing estimators including double sampling ratio and regression estimator for both the sampling designs and different combinations of sample sizes in terms of percentage absolute relative bias and relative root mean square error. Second, among the developed calibrated estimators, the CAL1 performs better than the CAL2.

4. CONCLUSION

Calibration estimators of the population total have been developed under two-stage sampling design based on the situation of unavailability of auxiliary information at the psu level. Monte Carlo simulations based on both simulated and real dataset show the superiority of the proposed calibration estimators of the population total in comparison to the existing estimators such as Horvitz-Thompson type, double sampling ratio and regression estimators. Further, among the developed estimators, the calibration estimator based on the sampling design 1 performs better than the estimator based on sampling design 2. Therefore, the developed calibration estimators will produce reliable estimate of population parameters from the two-stage survey data in the situations of unavailability of auxiliary information at the psu level.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable comments and suggestions provided by the anonymous referee. The authors also gratefully acknowledge the valuable contributions made by Late Dr. Hukum Chandra, National Fellow and Principal Scientist, ICAR-IASRI, New Delhi.

REFERENCES

Aditya, K., Sud, U.C., Chandra, H. and Biswas, A. (2016). Calibration based regression type estimator of the population total under two stage sampling design. *Journal of the Indian Society of Agricultural Statistics*, **70(1)**, 19-24.

Basak, P., Sud, U.C. and Chandra, H. (2017). Calibration Estimation of Regression Coefficient for Two-stage Sampling Design. *Journal of the Indian Society of Agricultural Statistics*, **71(1)**, 1-6.

Basak, P, Sud, U.C. and Chandra, H. (2018). Calibration Estimation of Regression Coefficient for Two-stage Sampling Design using Single Auxiliary Variable. *Journal of the Indian Society of Agricultural Statistics*, **72(1)**, 1-6.

Biswas, A., Aditya, K., Sud, U.C. and Basak, P. (2020). Product type calibration estimators of finite population total under two stage sampling. *Journal of the Indian Society of Agricultural Statistics*, **74(1)**, 23-32.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* **87**, 376-382.

Estevao, V.M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official* Statistics, **18(2)**, 233-255.

Guha, S. and Chandra, H. (2020). Improved chain-ratio type estimator for population total in double sampling. *Mathematical Population Studies*, **27(4)**, 216-231.

Mourya, K.K., Sisodia, B.V.S. and Chandra, H. (2016). Calibration approach for estimating finite population parameters in two-stage sampling. *Journal of Statistical Theory and Practice*, **10(3)**, 550-562.

Sahoo, L.N., Sahoo, R.K., Senapati, S.C. and Mangaraj, A.K. (2011). A general class of estimators in two-stage sampling with two auxiliary variables. *Hacettepe Journal of Mathematics and Statistics*, **40(5)**, 757-765.

Saini, M. (2013). A class of predictive estimators in two-stage sampling when auxiliary character is estimated. *Indian Journal of Science and technology*, **6(3)**, 4213-4218.

Saini, M. and Bahl, S. (2012). Estimation of population mean in two stage design using double sampling for stratification and multi-auxiliary information. *International Journal of Computer Applications*, **47(9)**, 17-21.

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

Srivastava, M. and Garg, N. (2009). A general class of estimators of a finite population mean using multi-auxiliary information under two stage sampling scheme. *Journal of Reliability and Statistical Studies*, **2(1)**, 103-118.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press. (USA).