Heliyon 10 (2024) e25106

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Adjustment of model misspecification in estimation of population total under ranked set sampling through balancing

Shakeel Ahmed [a,*], Javid Shabbir [b], Huda M. Alshanbari [c], Safoora Anjum [d], Abd AL-Aziz Hosni EL-Bagoury [e]

[a] School of Natural Sciences, NUST, H-12 Islamabad, Pakistan
[b] Department of Statistics, University of Wah, Wah Cantt., Pakistan
[c] Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh 11671, Saudi Arabia
[d] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan
[e] Higher Institute of Engineering and Technology at Elmahala Elkobra, Egypt

ARTICLE INFO

Dataset link: https://www.wiley.com/en-in/Finite+Population+Sampling+and+Inference%3A+A+Prediction+Approach-p-9780471293415

Keywords:
Model-based approach
Superpopulation
Efficiency improvement
Parameter estimation
Basis function

ABSTRACT

In the model-based approach, researchers assume that the underlying structure, which generates the population of interest, is correctly specified. However, when the working model differs from the underlying true population model, the estimation process becomes quite unreliable due to misspecification bias. Selecting a sample by applying the balancing conditions on some functions of the covariates can reduce such bias. This study aims at suggesting an estimator of population total by applying the balancing conditions on the basis functions of the auxiliary character(s) for the situations where the working model is different from the underlying true model under a ranked set sampling without replacement scheme. Special cases of the misspecified basis function model, i.e. homogeneous, linear, and proportional, are considered and balancing conditions are introduced in each case. Both simulation and bootstrapped studies show that the total estimators under proposed sampling mechanism keep up the superiority over simple random sampling in terms of efficiency and maintaining robustness against model failure.

1. Introduction

The basic literature on survey sampling is categorized into two approaches, design-based and model-based inferences. The core difference between the two approaches is the randomness in the stochastic structure for statistical inference [1]. The design-based approach posses many appealing features and delineate better estimates of the parameters of interest in terms of efficiency but it disregards the importance of the model relationship of the study the auxiliary characters at the stage of estimation. Numerous varieties of design-based estimators for estimating population parameters have been constructed to reduce the bias and mean squared error (MSE) e.g. [2], [3], [4], and [5]. On the opposite, supporters of the model-based approach emphasize that randomization occurs due to the model error term. Therefore, it is not the required condition for a rigorous statistical inference [6]. In finite population sampling, the model-based approach has been an interesting source of discussion over the past 50 years. The concept of a model-based approach was, initially, suggested by [7] who employed a simple model for prediction of the non-sampled values and their total

* Corresponding author at: School of Natural Sciences, National University of Science and Technology, Islamabad, Pakistan.
E-mail address: shakeel.ahmed@sns.nust.edu.pk (S. Ahmed).

https://doi.org/10.1016/j.heliyon.2024.e25106
Received 3 May 2023; Received in revised form 31 December 2023; Accepted 20 January 2024

Available online 24 January 2024

which is regarded as unknown and stochastic. After that [8], [9], [10], [11] and [12] employed a model-based approach to predict the total of a finite population utilizing superpopulation model. This model helps in the sample selection, suggesting estimators, and improving the precision of the estimators.

The model-based inference, typically, assumes that the model which explains the behavior of the random phenomena under study is specified correctly. However, when the working model is incorrect, the inference about the parameter of interest can be affected due to the existence of bias in the estimators. For many years, the preservation against model misspecification has been a major concern of survey statisticians. In this direction, [13], [14], [15] and [16] suggested the balancing associated with a sample to prevent the inference from model misspecification. Later on, [6, Chapter 3] provided a detailed discussion on balance sampling for reducing the impact of the bias introduced because of model failure. A detailed literature on the model-based methods to choose a balanced sample can be obtained from [17] and [18]. [19] discussed many robust sample designs, and shown that under these designs, estimators of the population total are still approximately unbiased, even when the assumed model is misspecified. [20] has described the conditions that can be achieved on the auxiliary variables and on the inclusion probabilities to obtain an exactly balanced sample. [21] have proposed a sample balancing method in multi-way stratification layout and employed it to find sample sizes for domains which belong to different sub-populations. Several versions of the model-based estimators have been found utilizing the model relationship between the variable of interest and the predictors [see [22], [23], [24], [25], [26] and [27]]. Recently, [28] worked on a general model-based framework for estimation of an unknown population quantity under basis functions regression model (BFRM). The problems of subset selection with one predictor under an automated matrix approach, and ill-conditioning of regression models are also highlighted. [29] have incorporated idea of balancing assuming basis function regression model (BFRM).

In the same era, many survey sampling researchers have worked on improved methods of data collection. Among them ranked set sampling (RSS) technique is at least as efficient as simple random sampling (SRS) for obtaining the experimental data that are truly representative of the population under investigation. The idea of the ranked set sampling method, initially proposed by [30] to assess mean pasture yields, has recently been modified by numerous researchers to estimate the population parameters with improved efficiency. Several authors, such as [31], [32], [33], [34], [35] and [36] have worked on estimation of parameters under RSS. [37] has discussed RSS procedure with spline and penalized spline models (PSM) parametrically. Some new contributions on RSS can be found in [38], [5], and [39] and references are there in. All the above cited works covered the developments in ranked set sampling assuming with replacement sampling. Pioneer work on ranked set sampling without replacement can be found from [40]. Recently, [41] worked on model-based framework for estimating finite population totals under RSS. After that [42] worked on stratified judgment post-stratified sampling (SJPS) design which includes selecting a JPS from each stratum. In addition to stratification, the method induces extra ranking structure in the stratum samples. Inference was made under both design and model-based paradigms. More details on model-based inference under JPS are available in [43]. On similar line, [44] have extended the idea of model-based estimation to ranked set sampling by using a without replacement version and provided with an algorithm that ensures independent observations under RSSWOR. The method suggested in [44] works well when specified model is same as the true population model. In some cases such matching is not possible and we need to make adjustment during sample selection process. In this study, we introduce a ranked set sampling without replacement mechanism for estimation of the total of a finite population utilizing the model relationship of the response variable and some function of the auxiliary variable. We follow the algorithm given in [44] for selecting a RSSWOR. The idea of balancing is obtained from [6]. Section 2 delineates the mathematical results of the proposed estimation approach using the BFR model under SRSWOR. Section 3 covers the developed estimator of population total under misspecified BFRM and its special cases. Section 4 describes the population total estimation, when the model is misspecified with some special cases under RSSWOR. Finally, in Section 5, we present a simulation study and a bootstrap study to evaluate the properties of the estimators.

2. Basis function regression model under SRSWOR

Consider a population of size N indexed as $U = \{1, 2, 3, ..., N\}$ from a sub-population with the variable under study Y having finite lower order moments. In notations of matrices, let $\boldsymbol{y} = (y_i, i \in U)$ be the realization of the random vector $\boldsymbol{Y} = (Y_i, i \in U)$. Let s be a sample of size n chosen from U using SRSWOR and $\bar{s}$ be the set of the units that are not covered in s. The population vector of y can be written as $y = (y_s^t, y_{\bar{s}}^t)^t$, where y_s and $y_{\bar{s}}$ are, respectively, the vectors of n sampled and $(N-n)$ non-sampled values of y. We assume a basis function regression model (BFRM), where the response variable depends on some function of the predictor(s) called the basis functions and denoted by $\Phi(x)$. Let M be a linear basis function model defined as

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients, $\boldsymbol{\Phi}$ is the matrix of basis functions and $\boldsymbol{\zeta}$ is random error term with zero mean. Mathematically, we can write

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_0(x_1) & \Phi_1(x_1) & \cdots & \Phi_{M-1}(x_1) \\ \Phi_0(x_2) & \Phi_1(x_2) & \cdots & \Phi_{M-1}(x_2) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \Phi_0(x_N) & \Phi_1(x_N) & \cdots & \Phi_{M-1}(x_N) \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots\cdots & \beta_{M-1} \end{pmatrix}^t, \qquad \boldsymbol{\zeta} = \begin{pmatrix} \xi_1 & \xi_2 & \cdots\cdots & \xi_N \end{pmatrix}^t.$$

The function $E_M(Y|\Phi,\beta)=\boldsymbol{\Phi}\boldsymbol{\beta}$ is considered as a non-linear function of input variable but linear in parameters. The matrix of basis functions Φ and the matrix $\boldsymbol{\nu}$, a square matrix containing variances and covariances, partitioned as:

$$\boldsymbol{\Phi}=\begin{bmatrix}\boldsymbol{\Phi}_s\\ \boldsymbol{\Phi}_{\bar{s}}\end{bmatrix},\qquad \boldsymbol{\nu}=\begin{bmatrix}\boldsymbol{\nu}_{ss} & \boldsymbol{\nu}_{s\bar{s}}\\ \boldsymbol{\nu}_{\bar{s}s} & \boldsymbol{\nu}_{\bar{s}\bar{s}}\end{bmatrix},$$

where Φ_s and $\Phi_{\bar{s}}$ are the sub-matrices with respective orders $(n\times M)$ and $(N-n)\times M$.

Let $\tau_y=\boldsymbol{\gamma}^t y$ be a realization of the stochastic quantity $\boldsymbol{\gamma}^t Y$, where $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_i;i\in U)$ is the weight vector which is separated into sampled and non-sampled parts as $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_s^t,\boldsymbol{\gamma}_{\bar{s}}^t)^t$. The population total and mean are obtained by setting $\boldsymbol{\gamma}_i$ as 1 and $1/N$ respectively for all $i\in U$.

[6] defined a linear estimator for τ_y as $\hat{\tau}_y=\boldsymbol{g}_s^t\boldsymbol{y}_s$, where $\boldsymbol{g}_s=(g_i;i\in s)$ is the vector of wights to be optimized. Assuming (1), the typical prediction estimator, suggested by [10], is given by

$$\hat{\tau}_y=\boldsymbol{\gamma}_s^t\boldsymbol{y}_s+\boldsymbol{\gamma}_{\bar{s}}^t\left[\Phi_{\bar{s}}\hat{\boldsymbol{\beta}}+\boldsymbol{\nu}_{\bar{s}s}\boldsymbol{\nu}_{ss}^{-1}(\boldsymbol{y}_s-\Phi_s\hat{\boldsymbol{\beta}})\right], \tag{2}$$

where $\hat{\boldsymbol{\beta}}=(\Phi_s^t\boldsymbol{\nu}_{ss}^{-1}\Phi_s)^{-1}\Phi_s^t\boldsymbol{\nu}_{ss}^{-1}y_s$ is the weighted least square (WLS) of the vector $\boldsymbol{\beta}$. The error-variance of $\hat{\tau}_y$ in Equation (2), is given by

$$\begin{aligned}V_M\left(\hat{\tau}_y-\tau_y\right)=&\boldsymbol{\gamma}_{\bar{s}}^t\left(\boldsymbol{\nu}_{\bar{s}\bar{s}}-\boldsymbol{\nu}_{\bar{s}s}\boldsymbol{\nu}_{ss}^{-1}\boldsymbol{\nu}_{s\bar{s}}\right)\boldsymbol{\gamma}_{\bar{s}}+\boldsymbol{\gamma}_{\bar{s}}^t\left(\Phi_{\bar{s}}-\boldsymbol{\nu}_{\bar{s}s}\boldsymbol{\nu}_{ss}^{-1}\Phi_s\right)\left(\Phi_s^t\boldsymbol{\nu}_{ss}^{-1}\Phi_s\right)^{-1}\\ &\left(\Phi_{\bar{s}}-\boldsymbol{\nu}_{\bar{s}s}\boldsymbol{\nu}_{ss}^{-1}\Phi_s\right)^t\boldsymbol{\gamma}_{\bar{s}}.\end{aligned} \tag{3}$$

When the values on sampled and non-sampled units are assumed to be uncorrelated i.e. $\boldsymbol{\nu}_{\bar{s}s}=\boldsymbol{\nu}_{s\bar{s}}=0$, Equation (3) reduces to

$$\hat{\tau}_y=\boldsymbol{\gamma}_s^t\boldsymbol{y}_s+\boldsymbol{\gamma}_{\bar{s}}^t\Phi_{\bar{s}}\hat{\boldsymbol{\beta}} \tag{4}$$

with prediction error-variance,

$$V_M\left(\hat{\tau}_y-\tau_y\right)=\boldsymbol{\gamma}_{\bar{s}}^t\left(\boldsymbol{\nu}_{\bar{s}\bar{s}}+\Phi_{\bar{s}}\left(\Phi_s^t\boldsymbol{\nu}_{ss}^{-1}\Phi_s\right)^{-1}\Phi_{\bar{s}}^t\right)\boldsymbol{\gamma}_{\bar{s}}. \tag{5}$$

Further assuming i.i.d. error terms i.e. $\boldsymbol{\nu}_{ss}=\sigma^2 I_{ss}$ and $\boldsymbol{\nu}_{\bar{s}\bar{s}}=\sigma^2 I_{\bar{s}\bar{s}}$, the error-variance reduces to

$$V_M\left(\hat{\tau}_y-\tau_y\right)=\sigma^2\left[\boldsymbol{\gamma}_{\bar{s}}^t\boldsymbol{\gamma}_{\bar{s}}+\boldsymbol{\gamma}_{\bar{s}}^t\left(\Phi_s^t\Phi_s\right)^{-1}\Phi_{\bar{s}}^t\boldsymbol{\gamma}_{\bar{s}}\right],$$

where I_{ss} and $I_{\bar{s}\bar{s}}$ are the identity matrices of order n and $N-n$ respectively.

3. Estimation with misspecified BFRM under SRSWOR

The balanced sampling adjusts for the discrepancy that occurs when the working model on which the estimator is based deviates from the true underlying model [6]. Suppose the working model is M^* which is defined as

$$\boldsymbol{Y}=\boldsymbol{\Phi}^*\boldsymbol{\beta}^*+\boldsymbol{\zeta}. \tag{6}$$

Our aim is to obtain a methodology that makes an estimator $\hat{\tau}_y^*$ derived under a model M^* which is unbiased under the true model M defined in Equation (1). Following Equation (6), the general prediction estimator, is given by

$$\hat{\tau}_y^*=\boldsymbol{\gamma}_s^t y_s+\boldsymbol{\gamma}_{\bar{s}}^t\Phi_{\bar{s}}^*\hat{\boldsymbol{\beta}}^*, \tag{7}$$

where $\hat{\boldsymbol{\beta}}^*=\left(\Phi_s^{*t}\boldsymbol{\nu}_{ss}^{-1}\Phi_s^*\right)^{-1}\Phi_s^{*t}\boldsymbol{\nu}_{ss}^{-1}y_s$ is the WLS estimator of $\boldsymbol{\beta}^*$ (see details in [29]). The prediction bias of the estimator $\hat{\tau}_y^*$ given in Equation (7) under model M is expressed as

$$B_M\left(\hat{\tau}_y^*-\tau_y\right)=\boldsymbol{\gamma}_{\bar{s}}^t\left(\boldsymbol{\omega}^*\Phi_s-\Phi_{\bar{s}}\right)\boldsymbol{\beta}, \tag{8}$$

where $\boldsymbol{\omega}^*=\Phi_{\bar{s}}^*(\Phi_s^{*t}\boldsymbol{\nu}_{ss}^{-1}\Phi_s^*)^{-1}\Phi_s^{*t}\boldsymbol{\nu}_{ss}^{-1}$ is a $(N-n)\times n$ matrix of weights. The sample is $\boldsymbol{\omega}^*$-balanced on basis function when $\boldsymbol{\gamma}_{\bar{s}}^t\boldsymbol{\omega}^*\Phi_s=\boldsymbol{\gamma}_{\bar{s}}^t\Phi_{\bar{s}}$ and the bias in Equation (8) vanishes under the balancing condition [19]. The prediction error variance for a balanced sample is

$$V_M\left(\hat{\tau}_y^*-\tau_y\right)=\boldsymbol{\gamma}_{\bar{s}}^t\left(\Phi_{\bar{s}}^*(\Phi_s^{*t}\boldsymbol{\nu}_{ss}^{-1}\Phi_s^*)^{-1}\Phi_{\bar{s}}^{*t}+\boldsymbol{\nu}_{\bar{s}\bar{s}}\right)\boldsymbol{\gamma}_{\bar{s}}. \tag{9}$$

Assuming i.i.d. error terms i.e. $\boldsymbol{\nu}_{ss}=\sigma^2 I_{ss}$ and $\boldsymbol{\nu}_{\bar{s}\bar{s}}=\sigma^2 I_{\bar{s}\bar{s}}$, Equation (9) reduces to

$$V_M\left(\hat{\tau}_y^*-\tau_y\right)=\sigma^2\left(\boldsymbol{\gamma}_{\bar{s}}^t\Phi_{\bar{s}}^*(\Phi_s^{*t}\Phi_s^*)^{-1}\Phi_{\bar{s}}^{*t}\boldsymbol{\gamma}_{\bar{s}}+\boldsymbol{\gamma}_{\bar{s}}^t\boldsymbol{\gamma}_{\bar{s}}\right).$$

In the subsequent subsections, we discuss the idea of balancing under the most widely used population models as the special cases of the BFRM.

3.1. Case I: homogeneous population model

Under homogeneous population model the basis function matrix $\mathbf{\Phi}^*$ will be an N dimensional vector of $1'$s. The response on the ith population unit y under M^* is expressed as:

$$y_i = \beta_0 + \xi_i, \qquad \forall\, i \in U \tag{10}$$

where error terms ξ_i (for $i \in U$) are i.i.d. with mean zero and variance σ^2. The conditional mean, variance and the covariance are $E_M\left\{y_i|\Phi^*(x_i)\right\} = \beta_0$, $V_M\left\{y_i|\Phi^*(x_i)\right\} = \sigma^2$ and $Cov_M\left\{y_i, y_j|\Phi^*(x_i), \Phi^*(x_j)\right\} = 0$ (for $i \neq j$), respectively. A best linear unbiased estimator for β_0 is $\hat{\beta}_0 = \frac{\sum_{i\in s} y_i}{n}$, which yields an estimator of t_y, named as the expansion estimator, and is given by

$$\hat{t}_y^E = \sum_{i\in s} y_i + \sum_{i\in \bar{s}} \hat{\beta}_0 = N\bar{y}_s.$$

The expected prediction error (bias) of the expansion estimator under model M, is given in Equation (11)

$$B_M(\hat{t}_y^E - t_y) = N \sum_{l=0}^{M-1} \beta_l \left(\bar{\Phi}_{sl} - \bar{\Phi}_{Ul}\right), \tag{11}$$

where $\bar{\Phi}_{sl} = \frac{\sum_{i\in s} \Phi_l(x_i)}{n}$ and $\bar{\Phi}_{Ul} = \frac{\sum_{i\in U} \Phi_l(x_i)}{N}$ are means of the lth (for $l = 0, 1, 2, \cdots, (M-1)$) basis function for sample and population, respectively. A sample which satisfies the condition $\bar{\Phi}_s \approx \bar{\Phi}_U$ is known as a balanced sample. Under a balanced sample, the expansion estimator becomes unbiased even for misspecified working model. The error variance is obtained as

$$V_M(\hat{t}_y^E - t_y) = \frac{N^2}{n}(1-f)\sigma^2. \tag{12}$$

Equation (12) is equivalent to the variance expression of the designed-based total estimator in SRSWOR [2]. Equation (12) infers that model misspecification does not disturb the error variance of the total estimator when sample is balanced on the means of the basis functions.

3.2. Case II: linear basis function model

Assuming a linear basis function model of 1st order with intercept, the ith response on y can be expressed as:

$$y_i = \beta_0 + \beta_1\Phi_1(x_i) + \xi_i, \qquad \forall i \in U$$

The BLUEs for β_0 and β_1 are obtained under OLS technique and expressed as $\hat{\beta}_0 = \bar{y}_s - \hat{\beta}_1\bar{\Phi}_{s1}$ and $\hat{\beta}_1 = \frac{\sum_{i\in s}(y_i-\bar{y}_s)(\Phi(x_i)-\bar{\Phi}_s)}{\sum_{i\in s}(\Phi(x_i)-\bar{\Phi}_s)^2}$. After some algebraic operations, the estimator for population total t_y becomes

$$\hat{t}_y^L = N\left[\bar{y}_s + \hat{\beta}_1\left(\bar{\Phi}_{U1} - \bar{\Phi}_{s1}\right)\right], \tag{13}$$

where $\bar{\Phi}_{s1} = \frac{\sum_{i\in s}\Phi_1(x_i)}{n}$ and $\bar{\Phi}_{U1} = \frac{\sum_{i\in U}\Phi_1(x_i)}{N}$ are, respectively, the sample and population means of $\Phi_1(x)$. The prediction bias of the total estimator given in Equation (13) under linear regression model, is given by

$$B_M\left(\hat{t}_y^L - t_y\right) = N\left[\sum_{l=2}^{M-1} \beta_l\left(\bar{\Phi}_{sl} - \bar{\Phi}_{Ul}\right)\right]. \tag{14}$$

The misspecification bias under the working model is expressed in Equation (14) which can be minimized by selecting a sample such that the sample mean of the basis functions is nearly close to the population mean of the corresponding basis functions. Under balancing condition, the error variance of the total estimator is obtained as:

$$V_M(\hat{t}_y^L - t_y) = \frac{N^2}{n}(1-f)\sigma^2\left(1 + \frac{\left(\bar{\Phi}_{U1} - \bar{\Phi}_{s1}\right)^2}{(1-f)\varphi_s}\right), \tag{15}$$

where $\varphi_s = \frac{\sum_{i\in s}\left(\Phi_1(x_i)-\bar{\Phi}_{s1}\right)^2}{n}$ is the variance of basis function Φ_{s1} in the sample. It is observed that under first-order balancing, the variance of the total estimator under Equation (10) coincides with the variance of the expansion estimator given in Equation (15).

3.3. Case III: proportional basis function model

A proportional basis function model for the response on the ith population unit is expressed as

$$y_i = \beta_1 \Phi_1(x_i) + \Psi(x_i)\xi_i, \qquad \forall\, i \in U \tag{16}$$

where $\Psi(x_i)$ is some function of the auxiliary information. Equation (16) is converted to a model with homoscedastic error term

$$y_i^* = \beta_1 \Phi_1^*(x_i) + \xi_i, \tag{17}$$

where $y_i^* = \frac{y_i}{\Psi(x_i)}$, $\Phi_1^*(x_i) = \frac{\Phi_1(x_i)}{\Psi(x_i)}$. The WLS estimator for β_1 in Equation (17) is given by $\hat{\beta}_1 = \frac{\sum_{i\in s} \Phi_1^*(x_i) y_i^*}{\sum_{i\in s} \Phi_1^{*2}(x_i)}$. The estimator of total, is given by

$$\hat{t}_y^P = t_{ys} + \hat{\beta}_1 \sum_{i\in\bar{s}} \Phi_1^*(x_i). \tag{18}$$

The prediction bias of the proportional estimator given in Equation (18), is given by

$$B_M(\hat{t}_y^P - t_y) = (N-n) \sum_{l=0}^{M-1} \beta_l \Big[\sum_{i\in s} k_i^* \Phi_{sl}(x_i) - \frac{\bar{\Phi}_{\bar{s}l}}{\bar{\Phi}_{\bar{s}1}} \Big], \tag{19}$$

where $k_i^* = \big(\frac{\Phi_1^*(x_i)}{\Psi(x_i) \sum_{i\in s} \Phi_1^{*2}(x_i)} \big) \sum_{i\in\bar{s}} \Phi_1^*(x_i)$. The bias given in Equation (19) can be reduced by selecting a sample such that the difference on the right hand side of Equation (19) is minimum i.e. $\sum_{i\in s} k_i^* \Phi_{sl}(x_i) \approx \frac{\bar{\Phi}_{\bar{s}l}}{\bar{\Phi}_{\bar{s}1}}$. Under the balancing condition, the error variance is obtained as

$$V_M(\hat{t}_y^P - t_y) = \Big[\sum_{i\in s} k_i^{**2} \Psi^2(x_i) + \sum_{i\in\bar{s}} \Psi^2(x_i) \Big] \sigma^2, \tag{20}$$

where $k_i^{**} = (N-n)\, k_i^*$. The gamma population model given in [19, Chapter 5] is resulted by setting $\Psi(x_i) = x^{\gamma *}$ in Equation (16), where the quantity γ^* controls the variation in the response variable depends on the auxiliary variable. Similarly the ratio estimator is obtained by setting $\gamma^* = \frac{1}{2}$. For $\gamma^* = 0$, the given population model reduces to linear regression model with constant variance.

4. Estimation with misspecified BFRM under RSSWOR

After the pioneer work of [30], RSS attained much attention and has been applied in parameter estimation under with replacement settings. Due to its attractive feature of efficiency improvement in RSS, [44] introduced a new version of ranked set sampling for obtaining a without replacement sample. To add up in efficiency and to obtain robustness to model failure in finite population parameter estimation, in this article we propose the idea of balancing based on the means fo basis functions in RSSWOR suggested by [44]. For the ith judgment ordered random variable $Y_{[i]}$ for $i \in U$, let $\mu_{[i]}$ and $\sigma^2_{[i]}$ be, respectively, the mean and variance. Suppose $s_r = \{ y_{[1]1}, ..., y_{[m]1}; y_{[1]2}, ..., y_{[m]2}; .....; y_{[1]r}, ..., y_{[m]r} \}$ a ranked set sample of size $n = mr$, where m be the set size and r be the number of cycles. Let $\bar{s}_r$ be the set of index attached to the values of units that are not indexed in s_r. For a given ranked set sample s_r, we can rearrange the population vector as $y_r = (y_{s_r}^t, y_{\bar{s}_r}^t)^t$, where y_{s_r} and $y_{\bar{s}_r}$ be the vectors of mr sampled and $(N-mr)$ non-sampled values of the study variable respectively. Assuming judgment ranking which is done with respect to some covariate and the respective ordered values are subscripted as, $(y_{[i]}, x_{(i)})$ to denote judgment ranked units, we write the population BFRM of order (M) for y_r as follow

$$\boldsymbol{y}_r = \boldsymbol{\Phi}_r \boldsymbol{\beta} + \boldsymbol{\zeta}^*, \tag{21}$$

where $\boldsymbol{y}_r$ is the ranked response vector, $\boldsymbol{\Phi}_r$ is the ranked matrix containing basis functions, $\boldsymbol{\beta}$ is the vector of coefficients and $\boldsymbol{\zeta}^*$ is the vector of random error with mean zero vector and variance-covariance matrix $\boldsymbol{\nu}_r$. The function $E_M(Y|\Phi_r, \beta) = \boldsymbol{\Phi}_r \boldsymbol{\beta}$ is considered as non-linear function of regressors but the conditional mean is still linear in parameters. For estimating τ_y under Equation (21), the feature matrix Φ_r and covariance matrix $\boldsymbol{\nu}_r$ can be partitioned as:

$$\boldsymbol{\Phi}_r = \begin{bmatrix} \boldsymbol{\Phi}_{s_r} \\ \boldsymbol{\Phi}_{\bar{s}_r} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\nu}_r = \begin{bmatrix} \boldsymbol{\nu}_{s_r s_r} & \boldsymbol{\nu}_{s_r \bar{s}_r} \\ \boldsymbol{\nu}_{\bar{s}_r s_r} & \boldsymbol{\nu}_{\bar{s}_r \bar{s}_r} \end{bmatrix},$$

where $\boldsymbol{\Phi}_{s_r}$ and $\boldsymbol{\Phi}_{\bar{s}_r}$ are the sub-matrices of order $mr \times M$ and $(N-mr) \times M$, respectively. The matrix $\boldsymbol{\nu}_r$ is a square matrix consisting variances and covariances of the order statistics the sub-matrices $\boldsymbol{\nu}_{s_r \bar{s}_r}$ and $\boldsymbol{\nu}_{\bar{s}_r s_r}$ become null when RSSWOR given in [44] is applied for select the sample.

An estimator for τ_y, using general prediction theorem, given in [6], is obtained as follows

$$\hat{\tau}_{yr} = \boldsymbol{\gamma}_{s_r}^t y_{s_r} + \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r} \hat{\boldsymbol{\beta}}_{[rss]}, \tag{22}$$

where $\hat{\boldsymbol{\beta}}_{[rss]} = \big(\Phi_{s_r}^t \boldsymbol{\nu}_{s_r s_r}^{-1} \Phi_{s_r} \big)^{-1} \Phi_{s_r}^t \boldsymbol{\nu}_{s_r s_r}^{-1} y_{s_r}$ is the WLS estimator of $\boldsymbol{\beta}$ under RSSWOR. The prediction error of the estimator $\hat{\tau}_{yr}$ can be expressed as:

Algorithm 1 Proposed ranked set sample without replacement scheme.

1: Construct r random sub-populations from U i.e. $U_1, U_2, \cdots \cdot U_j \cdots \cdot, U_r$ of size N/r such that $\sum_{j=1}^{r} U_j = U$ and $N/r > m^2$. The division of population units into sub-population should be random and independent of the study variable to ensure independent selection from each sub-population.
2: Select m^2 units from the jth sub-population and partitioned into m sets each of size m for $j = 1, 2, 3, \cdots \cdot, r$.
3: Rank each set within itself and select the ith ranked unit from the ith set $i = 1, 2, 3, \cdots \cdot, m$ from each sub-population.
4: Repeat the Steps 1–3, R times and obtain R estimates from each RSSWOR samples.
5: Sort the R estimates in ascending order and divide into smaller groups and compute the balancing condition based on the sample means of the basis functions (see illustration of Steps 1–3 in Fig. 1 [44]).
6: Pick the sample which satisfies the balancing condition (e.g. the sample mean of lth basis function is very close to the corresponding population basis function).

$$\hat{\tau}_{yr} - \tau_y = \boldsymbol{\gamma}_{\bar{s}_r}^t \boldsymbol{\omega}_r y_{s_r} - \boldsymbol{\gamma}_{\bar{s}_r}^t y_{\bar{s}_r},$$

where $\boldsymbol{\omega}_r = \Phi_{\bar{s}_r} \left(\Phi_{s_r}^t \boldsymbol{\nu}_{s_r s_r}^{-1} \Phi_{s_r} \right)^{-1} \Phi_{s_r}^t \boldsymbol{\nu}_{s_r s_r}^{-1}$.

It is easy to show that $\hat{\tau}_{yr}$ is unbiased with prediction error variance as:

$$V_M \left(\hat{\tau}_{yr} - \tau_y \right) = V_M \left(\hat{\tau}_y - \tau_y \right) - \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r} \Phi_{s_r}^{-1} \Delta_{s_r}^t \Delta_{s_r} \left(\Phi_{s_r}^t \right)^{-1} \Phi_{\bar{s}_r}^t \boldsymbol{\gamma}_{\bar{s}_r}, \tag{23}$$

where $\boldsymbol{\nu}_{s_r s_r} = \sigma^2 I_{s_r s_r} - \Delta_{s_r}^t \Delta_{s_r}$, $\Delta_{s_r} = \left(\boldsymbol{\mu}_{s_r} - \boldsymbol{\mu} \right)$, $\boldsymbol{\mu}_{s_r}$ is the vector of population means for ranked data after random classification (see [44] for clarification) and $\boldsymbol{\mu}$ is the over all population mean. From Equations (5) and (23), it is clear that $\hat{\tau}_{yr}$ is always more efficient than the $\hat{\tau}_y$ which shows the superiority of under ranked set sampling over simple random sampling. The newly adopted ranked set sampling mechanism is illustrated in Algorithm 1.

We now discuss the problem of model misspecification bias in RSSWOR setting considering a working basis function model (M^*), which is different from the underlying true model (M). Suppose, we have the following working model (M^*).

$$\boldsymbol{y}_r = \boldsymbol{\Phi}_r^* \boldsymbol{\beta}^* + \boldsymbol{\zeta}^*. \tag{24}$$

Assuming single cycle i.e. $r = 1$ for making derivations simple, we write the estimator of population total under Equation (24) as follows

$$\hat{\tau}_{yr}^* = \boldsymbol{\gamma}_{s_r}^t y_{s_r} + \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r}^* \hat{\boldsymbol{\beta}}_{[rss]}^*, \tag{25}$$

where $\hat{\beta}_{[rss]}^* = \left(\Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1} \Phi_{s_r}^* \right)^{-1} \Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1} y_{s_r}$ is the weighted least square (WLS) estimator for the coefficient vector $\boldsymbol{\beta}^*$. We can write prediction error of $\hat{\tau}_{y[rss]}^*$ as

$$\hat{\tau}_{yr}^* - \tau_y = \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r}^* \left(\Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1} \Phi_{s_r}^* \right)^{-1} \Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1} y_{s_r} - \boldsymbol{\gamma}_{\bar{s}_r}^t y_{\bar{s}_r}.$$

The prediction bias under true model (M), is given by

$$B_M \left(\hat{\tau}_{yr}^* - \tau_y \right) = \boldsymbol{\gamma}_{\bar{s}_r}^t \left(\boldsymbol{\omega}_{[R]}^* \Phi_{s_r} - \Phi_{\bar{s}_r} \right) \boldsymbol{\beta}, \tag{26}$$

where $\boldsymbol{\omega}_r^* = \Phi_{\bar{s}_r}^* \left(\Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1} \Phi_{s_r}^* \right)^{-1} \Phi_{s_r}^{*t} \boldsymbol{\nu}_{s_r s_r}^{-1}$ is a $(N - m) \times M$. The sample is ω_r^* balanced on input variable(s) i.e. $\boldsymbol{\gamma}_{\bar{s}_r}^t \boldsymbol{\omega}_r^* \Phi_{s_r} = \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r}$. Under the balancing condition, the model bias given in Equation (26) vanishes and the prediction error variance reduces to

$$V_M \left(\hat{\tau}_{yr}^* - \tau_y \right) = V_M \left(\hat{\tau}_y^* - \tau_y \right) - \boldsymbol{\gamma}_{\bar{s}_r}^t \Phi_{\bar{s}_r}^* \Phi_{s_r}^{*-1} \Delta_{s_r}^t \Delta_{s_r} \left(\Phi_{s_r}^{*t} \right)^{-1} \Phi_{\bar{s}_r}^{*t} \boldsymbol{\gamma}_{\bar{s}_r}, \tag{27}$$

where $\boldsymbol{\nu}_{s_r s_r} = \sigma^2 I_{s_r s_r} - \Delta_{s_r}^t \Delta_{s_r}$ and $\Delta_{s_r} = \left(\mu_{s_r} - \mu \right)$. From Equations (9) and (27), it is clear that $\hat{\tau}_{yr}^*$ is more efficient than $\hat{\tau}_y$ which shows the superiority of the RSSWOR over SRSWOR for estimation of τ_y. The special cases of the BFRM under RSSWOR with balancing are discussed in following subsections.

4.1. Case I: homogeneous basis function model (HBFM)

The ordered population value of the study variable y, under HBFM is expressed as $y_{[i]} = \beta_0 + \xi_{[i]}$. The error term is approximately independently distributed with mean 0 and variance $\sigma_{[i]}^2$. It is assumed that the ranking is performed on some co-variate which is not related to $\xi_{[i]}$. We assume that $E_M \left\{ y_{[i]} | \Phi(x_{(i)}) \right\} = \beta_0$, $V_M \left\{ y_{[i]} | \Phi(x_{(i)}) \right\} = \sigma_{[i]}^2$ and $Cov_M \left\{ y_{[i]}, y_{[i']} | \Phi(x_{(i)}), \Phi(x_{(i')}) \right\} = 0$ for $i \neq i'$, where $y_{[i]}$ and $y_{[i']}$ are taken from different ranked sets. The expansion estimator $\hat{t}_{y[rss]}^E$ for the population total $t_y = \sum_{i \in s} y_{[i]} + \sum_{i \in \bar{s}} y_i$ under HBFM, is given by

$$\hat{t}_{y[rss]}^E = \sum_{i \in s_r} y_{[i]} + \sum_{i \in \bar{s}_r} \hat{\beta}_{0[rss]},$$

where $\hat{\beta}_{0[rss]} = \dfrac{\sum_{j=1}^{r} \sum_{i \in s_r} y_{[i]j}}{rm}$ is BLUP for β_0 under RSSWOR and the resulting total estimator is

$$\hat{t}_{yr}^{E} = N\frac{\sum_{j=1}^{r}\sum_{i\in s_r} y_{[i]j}}{rm}. \tag{28}$$

The prediction bias of the expansion estimator given in (28), after some simplification, is given by

$$\begin{aligned} B_M\left(\hat{t}_{yr}^{E} - t_y\right) &= E_M\left(\frac{N}{m}\sum_{i\in s_r} y_{[i]} - \sum_{i\in U} y_i\right) \\ &= N\sum_{q=0}^{M-1}\beta_q\left(\bar{\Phi}_{s_r q} - \bar{\Phi}_{U_q}\right) \end{aligned} \tag{29}$$

where $\bar{\Phi}_{s_r q} = \frac{\sum_{i\in s_r}\Phi_q(x_{(i)})}{m}$ and $\bar{\Phi}_{U_q} = \frac{\sum_{i\in U}\Phi_q(x_i)}{N}$, for $(q = 0, 1, 2, \cdots, M-1)$, are the sample and population means for the qth basis function of a balanced sample i.e. $\bar{\Phi}_{s_r} \approx \bar{\Phi}_U$. The error variance of $\hat{t}_{y[rss]}^{E}$, is given by

$$\begin{aligned} V_M(\hat{t}_{yr}^{E} - t_y) &= \frac{N}{m}(N-m)\sigma^2 - \left(\frac{N-m}{m}\right)^2\sum_{i\in s_r}\delta_{[i]}^2 \\ &= V_M(\hat{t}_{y}^{E} - t_y) - \left(\frac{N-m}{m}\right)^2\sum_{i\in s_r}\delta_{[i]}^2, \end{aligned} \tag{30}$$

where $\sigma_{[i]}^2 = \sigma^2 - \delta_{[i]}^2$ and $\delta_{[i]} = \mu_{[i]} - \mu$. Equation (30) shows that the total estimator under RSSWOR is at least as efficient as SRSWOR.

4.2. Case-II: linear basis function model (LBFM)

The response on the ith ordered population unit under a LBFM, is given by

$$y_{[i]} = \beta_0 + \beta_1\Phi_1(x_{(i)}) + \xi_{[i]} \tag{31}$$

with $E_M\left\{y_{[i]}|\Phi(x_{(i)})\right\} = \beta_0 + \beta_1\Phi_1(x_{(i)})$. Under (31), we obtain the following estimator for population total

$$\begin{aligned} \hat{t}_{yr}^{L} &= \sum_{i\in s_r} y_{[i]} + \sum_{i\in\bar{s}_r}\left(\hat{\beta}_{0[rss]} + \hat{\beta}_{1[rss]}\Phi_1(x_{(i)})\right) \\ &= N\left[\bar{y}_{s_r} + \hat{\beta}_{1[rss]}\left(\bar{\Phi}_{U_1} - \bar{\Phi}_{s_r 1}\right)\right], \end{aligned} \tag{32}$$

where $\hat{\beta}_{1[rss]}$ is the BLUE of β_1 under RSSWOR. The prediction bias of the total estimator in Equation (32) is given as:

$$\begin{aligned} B_M\left(\hat{t}_{yr}^{L} - t_y\right) &= E_M\left[\frac{N}{m}\sum_{i\in s_r} y_{[i]} + N\hat{\beta}_{1[rss]}\left(\bar{\Phi}_{U_1} - \bar{\Phi}_{s_r 1}\right) - \sum_{i\in U} y_i\right] \\ &= N\sum_{q=2}^{M-1}\beta_q\left(\bar{\Phi}_{s_r q} - \bar{\Phi}_{U_q}\right), \qquad q = 2, 3, \cdots, M-1. \end{aligned} \tag{33}$$

The bias given in Equation (33) can be reduced to zero by selecting a balanced sample under RSSWOR i.e. $\bar{\Phi}_{s_r} \approx \bar{\Phi}_U$. Under balancing condition, the estimator under LBFM reduces to expansion estimator. The prediction error variance of the total estimator under RSSWOR, is given by

$$V_M(\hat{t}_{yr}^{L} - t_y) = V_M\left[\frac{(N-m)}{m}\sum_{i\in s_r} y_{[i]} + N\hat{\beta}_{1[rss]}\left(\bar{\Phi}_{U_1} - \bar{\Phi}_{s_r 1}\right) - \sum_{i\in\bar{s}_r} y_i\right].$$

After some simplification, we get

$$\begin{aligned} V_M(\hat{t}_{yr}^{L} - t_y) &= V_M(\hat{t}_{y}^{L} - t_y) - \left(\frac{N-m}{m}\right)^2\sum_{i\in s_r}\delta_{[i]}^2 - N^2\left(\bar{\Phi}_{U_1} - \bar{\Phi}_{s_r 1}\right)^2 \\ &\quad \times\frac{\sum_{i\in s_r}\left(\Phi_1(x_{(i)}) - \bar{\Phi}_{s_r 1}\right)^2\delta_{[i]}^2}{m^2\varphi_{s_r}^2}, \end{aligned} \tag{34}$$

where $\varphi_{s_r} = \frac{1}{m}\sum_{i\in s_r}\left(\Phi_1(x_{(i)}) - \bar{\Phi}_{rs1}\right)^2$. Equation (34) indicates that prediction error variance is reduced by selecting a balanced sample under RSSWOR and also in this situation, variance of LBFM reduces to variance of the HPM. This provides that $\hat{t}_{yr}^{L}$ is at least as efficient as its counterpart under SRSWOR.

4.3. Case-III: proportional basis function model (PBFM)

The PBFM for the response on the ith ordered population value can be modeled as

$$y_{[i]} = \beta_1 \Phi_1(x_{(i)}) + \Psi(x_{(i)})\xi_{[i]} \quad \text{for} \quad i \in U \tag{35}$$

where $\Psi(x_{(i)})$ is some function of the ordered values of the auxiliary variable. Following transformation on Equation (35)is used to make error terms homoscedastic

$$y^*_{[i]} = \beta_1 \Phi^*_1(x_{(i)}) + \xi_{[i]}, \tag{36}$$

where $y^*_{[i]} = \frac{y_{[i]}}{\Psi(x_{(i)})}$, and $\Phi^*_1(x_{(i)}) = \frac{\Phi_1(x_{(i)})}{\Psi(x_{(i)})}$.

The error term $\xi_{[i]}$ is approximately independently distributed with $\mu_{[i]} = 0$ and variance $\sigma^2_{[i]}$. The estimator $\hat{t}^P_{yr}$ for population total under RSSWOR is expressed as follows:

$$\hat{t}^P_{yr} = \sum_{i \in s_r} y_{[i]} + \hat{\beta}_{1[rss]} \sum_{i \in \bar{s}_r} \Phi^*_1(x_{(i)})$$

The BLUE for β_1 under RSSWOR is obtained as $\hat{\beta}_{1[rss]} = \frac{\sum_{j=1}^{r} \sum_{i \in s_r} \Phi^*_1(x_{(i)j}) y^*_{[i]j}}{\sum_{j=1}^{r} \sum_{i \in s_r} \Phi^{*2}_1(x_{(i)j})}$. For $r = 1$, the estimator is written as

$$\hat{t}^P_{yr} = \sum_{i \in s_r} y_{[i]} + \sum_{i \in \bar{s}_r} \Phi^*_1(x_{(i)}) \frac{\sum_{i \in s_r} \Phi^*_1(x_{(i)}) y^*_{[i]}}{\sum_{i \in s_r} \Phi^{*2}_1(x_{(i)})}.$$

The expression for prediction bias under RSSWOR is given in Equation (37)

$$\begin{aligned} B_M\left(\hat{t}^P_{yr} - t_y\right) &= E_M\left[(N-m)\sum_{i \in s_r} k^*_{(i)} y_{[i]} - \sum_{i \in \bar{s}_r} y_i\right] \\ &= (N-m)\sum_{q=0}^{M-1} \beta_q \left[\sum_{i \in s_r} k^*_{(i)} \Phi_{sq}(x_{(i)}) - \frac{\bar{\Phi}_{\bar{s}_r q}}{\bar{\Phi}_{\bar{s}_r 1}}\right], \qquad q = 0, 1, 2, \cdots \cdot M-1, \end{aligned} \tag{37}$$

where $k^*_{(i)} = \left[\frac{\Phi^*_1(x_{(i)})}{\Psi(x_{(i)}) \sum_{i \in s_r} \Phi^{*2}_1(x_{(i)})}\right] \sum_{i \in \bar{s}_r} \Phi^*_1(x_{(i)})$. Unbiasedness can be achieved by selecting a sample which satisfies the condition i.e. $\sum_{i \in s_r} k^*_{(i)} \Phi_{sq}(x_{(i)}) - \frac{\bar{\Phi}_{\bar{s}_r q}}{\bar{\Phi}_{\bar{s}_r 1}}$. The prediction error variance of the total estimator is expressed as

$$\begin{aligned} V_M(\hat{t}^P_{yr} - t_y) &= \left[\sum_{i \in s_r} k^{**2}_{(i)} \Psi^2(x_{(i)}) + \sum_{i \in \bar{s}_r} \Psi^2(x_{(i)})\right]\sigma^2 - \sum_{i \in s_r} k^{**2}_{(i)} \Psi^2(x_{(i)}) \delta^2_{[i]} \\ &= V_M(\hat{t}^P_y - t_y) - \sum_{i \in s_r} k^{**2}_{(i)} \Psi^2(x_{(i)}) \delta^2_{[i]}, \end{aligned} \tag{38}$$

where $k^{**}_{(i)} = (N-m)\, k^*_{(i)}$. Equation (38) shows the supremacy of the predictive estimator $\hat{t}^P_{yr}$ over its counterpart under SRSWOR (see Equation (20)).

5. Empirical studies

We conduct two empirical studies to evaluate the performance of the proposed estimators of the finite population total under RSSWOR. For this purpose, firstly, a simulation study is conducted using hypothetically generated population and then we provide a bootstrap study using real-world data.

5.1. Simulation study

For the purpose of efficiency comparisons, we conduct a Monte Carlo (MC) experiment by creating a hypothetical population with $N = 1000$ data points. Following [44], the values on the auxiliary character (x) are obtained assuming a gamma distribution with different sets of parameters a and b. The values of $\boldsymbol{y}$ are obtained using the relationship $\boldsymbol{y} = \boldsymbol{x} \ominus +\epsilon$, where $\epsilon \sim (0, \sigma^2 I_N)$ is randomly generated error term and $\ominus$ is computed as the averaged eigen vector corresponding to the eigen values of the data matrix $H = x^t x$ that are greater than unity. We select an SRS and a RSSWOR each with total units $n = mr = 10$, 20, 25, 40, 50, and 80, where $m = 2$, 5, 8 sizes and $r = 5$, 10. The sampling process is replicated $\eta = 20{,}000$ times to evaluate the behavior of the estimators and their mutual comparison. The mean squared prediction error (MSPE) of the proposed estimators and the corresponding estimators with balancing restrictions are obtained as follows

$$MSPE_{BI.rss} = \sum_{c=1}^{\eta} \left\{ \frac{(\hat{t}^I_{yrc} - t_y)^2}{\eta} \right\}, \qquad I = E, P, L \tag{39}$$

Table 1
Simulated MSEs of total estimator under SRSWOR and RSSWOR with $M = 3$.

r	m	$RE_{BE.rss}$	$RE_{BP.rss}$	$RE_{BL.rss}$	$RE_{BE.rss}$	$RE_{BP.rss}$	$RE_{BL.rss}$
		$M = 3$			$M = 5$		
				G(3,1)			
	2	1.0491	1.0424	1.0421	1.0808	1.0818	1.0847
5	5	1.1225	1.1242	1.1259	1.2069	1.2048	1.1998
	8	1.2969	1.2628	1.2808	1.6285	1.6124	1.5640
	2	1.2112	1.1968	1.2023	1.0934	1.0901	1.0805
10	5	1.1158	1.1004	1.1003	1.1690	1.1645	1.1509
	8	1.2975	1.2082	1.2237	1.4171	1.4137	1.4040
				G(3,2)			
	2	1.0888	1.0904	1.0947	1.1184	1.1208	1.1259
5	5	1.1025	1.1063	1.1247	1.2096	1.2081	1.2051
	8	1.1725	1.1652	1.1615	1.4610	1.4461	1.4049
	2	1.1468	1.1439	1.1495	1.1715	1.1682	1.1654
10	5	1.2029	1.2104	1.2194	1.1790	1.1743	1.1623
	8	1.2303	1.2175	1.2098	1.1400	1.1351	1.1268
				G(5,2)			
	2	1.0170	1.0122	1.0168	1.2606	1.2625	1.2675
5	5	1.0646	1.0478	1.0447	1.1599	1.1603	1.1574
	8	1.0814	1.0740	1.0658	1.4963	1.4920	1.4852
	2	1.0997	1.1027	1.0990	1.0382	1.0310	1.0348
10	5	1.0903	1.0939	1.0916	1.2006	1.1961	1.1863
	8	1.2327	1.2254	1.2205	1.3012	1.3106	1.3051
				G(5,4)			
	2	1.0628	1.0641	1.0672	1.1871	1.1745	1.1663
5	5	1.0418	1.0192	1.0160	1.1260	1.1281	1.1262
	8	1.1270	1.1313	1.1244	1.3109	1.3068	1.2992
	2	1.0533	1.0558	1.0574	1.0550	1.0561	1.0577
10	5	1.2274	1.2297	1.2171	1.0748	1.0644	1.0594
	8	1.1906	1.1808	1.1773	1.0871	1.0845	1.0765

$$MSPE_{BI.srs} = \sum_{c=1}^{\eta} \left\{ \frac{(\hat{t}^{I}_{yc} - t_y)^2}{\eta} \right\}, \qquad I = E, P, L \tag{40}$$

with a relative efficiency as

$$RE_{BI.rss} = \frac{MSPE_{BI.srs}}{MSE_{BI.rss}}, \qquad I = E, P, L$$

The absolute biases (ABs) under balanced sample technique under RSSWOR are calculated as

$$AB_{BI.rss} = \sum_{c=1}^{\eta} \left| \frac{(\hat{t}^{I}_{yrc} - t_y)}{\eta} \right|, \qquad I = E, P, L \tag{41}$$

For comparison, the ABs are also obtain under SRSWOR of size n as:

$$AB_{BI.srs} = \sum_{c=1}^{\eta} \left| \frac{(\hat{t}^{I}_{yc} - t_y)}{\eta} \right|, \qquad I = E, P, L \tag{42}$$

The results of simulation study for different choices of Gamma distribution $G(a;b)$ are given in Table 1. Relative efficiency (RE) of the total estimator under balanced RSSWOR with respect to their counterparts are obtained in Table 1. The RSSWOR estimator performs better when set size is relatively larger. Similarly, RE values are higher for higher order polynomial basis function models. Further, the RE values are higher for choice of shape parameter of the Gamma distribution a. The absolute biases (ABs) of the total estimators under SRSWOR and RSSWOR are reported in Tables 2 and 3 for $M = 3$ and $M = 5$ respectively. The total estimators under RSSWOR yield relatively smaller ABs than that of the estimators under SRSWOR. Comparing ABs in Tables 2 and 3 one can conclude that the constrain of balancing is more effective with lower order polynomial basis function models. Similarly, the values of ABs for different choices of the parameters of gamma distribution can be compared by observing Tables 2 and 3. ABs of the total estimator in RSSWOR tend to decrease with increase in set size m and the number of cycles r too, which indicates that our proposed estimator performs better under model misspecification

Table 2
Simulated ABs of total estimator under SRSWOR and RSSWOR with $M = 3$.

t	m	$AB_{BE.srs}$	$AB_{BP.srs}$	$AB_{BL.srs}$	$AB_{BE.rss}$	$AB_{BP.rss}$	$AB_{BL.rss}$
				G(3,1)			
5	2	66.8657	72.0003	78.2543	58.6901	62.8968	67.5400
10		40.3001	49.1976	53.4408	24.5476	32.5703	38.5398
5	5	20.1106	30.1876	36.5044	11.9226	18.9517	20.1343
10		9.6356	12.6104	14.9427	4.0930	7.6618	8.0473
5	8	12.9411	16.9934	21.3778	7.0693	8.3414	10.5984
10		6.8300	8.3676	10.8993	2.0317	4.0343	4.9982
				G(3,2)			
5	2	27.8350	28.6832	32.7945	19.1209	19.5267	20.3128
10		17.1977	20.5302	24.3838	10.1298	13.1127	17.5576
5	5	10.2189	13.9700	17.2160	6.9432	8.3357	9.3555
10		5.1276	6.8604	8.6562	0.8678	2.9407	3.1512
5	8	7.7666	5.0500	9.3713	1.3776	3.7934	5.8796
10		1.2689	2.9954	4.5702	0.3311	0.4659	0.8682
				G(5,2)			
5	2	19.4222	28.8517	33.3209	10.5543	19.9300	25.5437
10		11.7477	17.3109	20.0423	7.8265	10.1012	13.4287
5	5	7.2266	11.9065	18.3234	4.4209	6.2111	7.7354
10		5.8535	7.3711	8.4976	2.8780	4.1609	4.5025
5	8	6.4175	9.2900	11.4509	3.6634	7.0676	5.3942
10		4.1888	5.2754	6.8710	0.9876	0.4460	1.3543
				G(5,4)			
5	2	12.3006	16.9863	19.9821	7.1368	9.1630	10.6479
10		6.2535	8.3733	11.5766	3.8410	5.0946	6.8834
5	5	4.7820	5.9207	8.6599	2.5495	3.5129	4.1620
10		2.8951	4.1012	5.9011	0.7632	0.4058	1.0006
5	8	3.8451	6.9086	7.3425	0.0943	1.5589	1.9028
10		0.5092	1.8525	3.5061	0.0159	0.8881	0.0303

Table 3
Simulated ABs of total estimator under SRSWOR and RSSWOR with $M = 5$.

r	m	$AB_{BE.srs}$	$AB_{BP.srs}$	$AB_{BL.srs}$	$AB_{BE.rss}$	$AB_{BP.rss}$	$AB_{BL.rss}$
				G(3,1)			
5	2	3655.2512	3711.4980	4077.9060	2104.2515	2309.1798	2671.1745
10		3060.7540	3289.4904	3754.9755	1888.7451	1990.3500	2210.0817
5	5	1866.3167	2107.8587	2619.0132	1100.4511	1490.0865	1610.9230
10		928.8201	1161.2856	1700.6612	546.5607	712.2954	901.0886
5	8	1224.0187	1446.2904	2057.2836	615.5572	857.9963	1088.6778
10		176.3144	280.9307	473.1444	74.1323	119.2898	241.2923
				G(3,2)			
5	2	435.3540	469.3012	495.3176	259.5445	288.0545	302.0144
10		167.1598	183.5406	215.3223	92.0376	101.1512	143.6133
5	5	113.1509	130.0323	164.0114	67.3387	84.7843	100.1067
10		52.4399	68.9607	104.5009	32.4143	40.5731	71.8876
5	8	63.5800	82.1165	129.8623	36.0677	46.4209	78.8656
10		15.1812	26.9074	54.6509	8.3840	15.8922	31.7555
				G(5,2)			
5	2	1510.3540	1557.3012	1654.3176	825.8523	874.0756	901.9987
10		1110.1598	1151.5406	1250.3223	676.6017	697.3065	712.7732
5	5	627.1509	693.0323	854.0114	324.2975	401.8634	484.0146
10		250.4399	281.9607	412.5009	141.9864	197.2886	245.6687
5	8	304.5800	365.1165	522.8623	171.9111	228.5843	335.8843
10		172.1812	215.9074	309.6509	87.5912	115.8388	173.1566
				G(5,4)			
5	2	94.5793	97.7126	104.0087	46.7772	59.3593	73.8702
10		64.9339	67.6973	81.4644	33.5258	42.7515	60.9946
5	5	35.5339	40.6358	50.6530	20.4888	26.2799	35.9581
10		17.1437	21.8906	27.8021	9.5260	12.9178	17.3328
5	8	21.8644	26.2070	36.9291	12.4457	14.2724	20.4412
10		7.2685	11.0959	18.2345	3.2359	6.5321	10.8713

Table 4
Bootstrapped MSPE of total estimators under SRSWOR and RSSWOR.

r	m	$MSPE_{BE.srs}$	$MSPE_{BP.srs}$	$MSPE_{BL.srs}$	$MSPE_{BE.rss}$	$MSPE_{BP.rss}$	$MSPE_{BL.rss}$
5	4	912.5090	901.3022	906.8722	884.9288	876.3699	868.9105
8		632.8460	625.6318	616.4567	570.1571	546.0445	524.2228
5	6	742.3634	736.5551	732.0162	618.6125	611.2861	603.4957
8		432.1373	428.5528	426.4057	385.6711	376.1896	360.272
5	8	525.6232	516.6093	506.6172	469.9386	459.9073	434.9171
8		315.2605	311.7891	307.9543	279.2272	275.6586	267.4239

Table 5
Bootstrapped ABs of total estimator under SRSWOR and RSSWOR.

r	m	$AB_{BE.srs}$	$AB_{BP.srs}$	$AB_{BL.srs}$	$AB_{BE.rss}$	$AB_{BP.rss}$	$AB_{BL.rss}$
5	4	643.0678	703.1282	931.7880	376.5418	466.0784	597.7331
8		452.2373	616.5886	796.0038	265.6684	389.3196	474.7331
5	6	498.8261	697.3275	827.1205	283.9132	409.6674	515.2632
8		245.7981	410.0384	498.8969	92.897	167.4791	197.3432
5	8	325.2421	503.5297	622.8771	158.1708	217.0325	305.4045
8		84.75926	119.6282	205.4875	46.3246	78.5706	99.3762

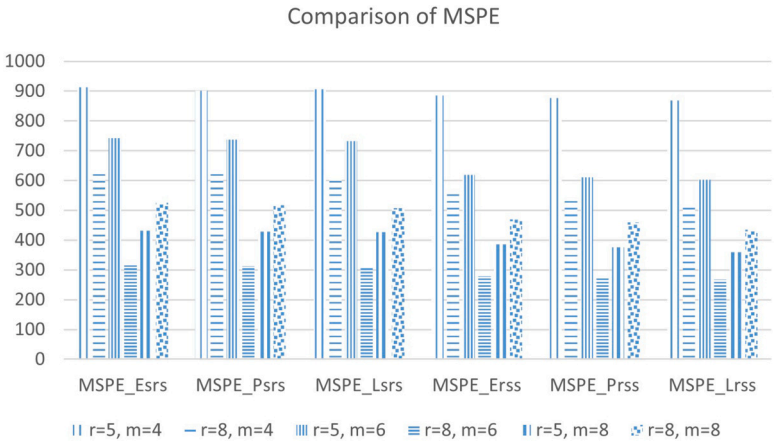

Fig. 1. Comparison of MSPE for total estimators for different choices of r, and m with different models.

5.2. Bootstrapped study

To evaluate the performance and applicability of the proposed estimators with balanced sampling strategy under RSSWOR, we consider a data set given in [6, Appendix B-2, Page 424]. The data consists of $N = 393$ hospitals which is considered as the population to be sampled for bootstrapping. The number of beds in each hospital is taken as the auxiliary variable, and the number of patients discharged is taken as the study variable. The formula used for the mean squared error and absolute bias are same as given in Equations (39), (40) and (41), (42) respectively. See [6] about the detail on variable and relationship between the study variable and the auxiliary variable.

The results computed from the real data set (hospital data) are reported in Tables 4 and 5. Three simple basis function models are used to obtain results under ranked set sampling under respective balancing conditions (see conditions obtained under Cases I, II and III in Section 4). The procedure given in Algorithm 1 is repeated $R = 20{,}000$ times and the MSPE and absolute biases (ABs) are obtained using Equations (39) and (41).

Tables 4–5 give bootstrapped results of different estimators under HPM, LBFM, and PBFM. The MSPE values in Table 4 are reported after dividing by 10^6. It can be seen that the MSE values reduce under RSSWOR using the balanced sampling technique as compared to SRSWOR. Table 5 shows that the ABs under RSSWOR using the balanced sample rapidly go down and each estimator is unbiased where the sample mean is nearer to the population mean. Both simulation and bootstrapped studies provide evidence of superiority of ranked set sampling both in terms of AB and MSE. The results also suggest that it is more easy to get balancing conditions under RSSWOR as in ranked set sampling we consider many sets and observe the values of the auxiliary variables and rank the data accordingly which can be used for balancing. For visual comparison the MSPE and AB values are displayed in Figs. 1 and 2 respectively.

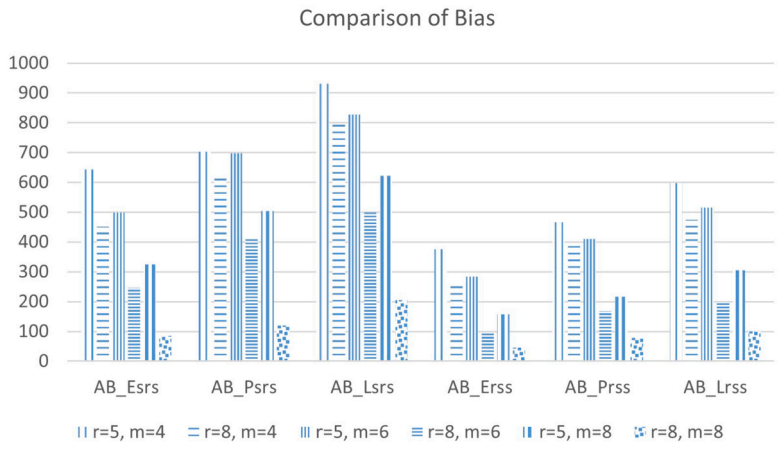

Fig. 2. Comparison of AB for total estimators for different choices of r, and m with different models.

6. Conclusion

Model-based robust estimators of finite population total, assuming the superpopulation setting, are discussed under the misspecified basis function regression model in SRSWOR and RSSWOR. Some well known estimators are identified as the special cases of the proposed estimators. The expressions for prediction error, bias, and mean squared error of the proposed estimators are derived. The results indicate that the proposed technique based on balanced sampling under RSSWOR works well in case of model misspecification in terms of bias reduction. Both mathematical expressions and empirical study keep up the superiority of the total estimators with balancing conditions under RSSWOR over SRSWOR. The results suggested that through proper selection of a sample and estimator, estimation becomes robust against model failure. Hence, the suggested estimators can be used in the estimation of any linear combination of the study variable including mean, total and proportions. It is also applicable in public health to check the prevalence of disease, where demographic information can be used as the auxiliary data. The results also suggest that it is more easy to get balancing conditions under RSSWOR as in ranked set sampling, we consider many sets and observe the values of the auxiliary variables and rank the data accordingly which can be used for balancing. In RSSWOR, we assume that the error term is random with a zero mean and constant variance but this assumption is very weak as the mean of the error may not be zero for a ranked set. The idea can be extended to other ranked set sampling schemes and we can work with the same idea under Bayesian framework.

CRediT authorship contribution statement

Shakeel Ahmed: Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Conceptualization. **Javid Shabbir:** Visualization, Validation, Supervision. **Huda M. Alshanbari:** Writing – review & editing, Validation, Funding acquisition. **Safoora Anjum:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Abd AL-Aziz Hosni EL-Bagoury:** Writing – review & editing, Validation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset associated with this manuscript is available in [6, Appendix B-2, Page 424].

Download link: https://www.wiley.com/en-in/Finite+Population+Sampling+and+Inference%3A+A+Prediction+Approach-p-9780471293415.

Acknowledgement

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R 299), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] Carl-Erik Särndal, Ib Thomsen, Jan M. Hoem, D.V. Lindley, O. Barndorff-Nielsen, Tore Dalenius, Design-based and model-based inference in survey sampling [with discussion and reply], Scand. J. Stat. (1978) 27–52, https://www.jstor.org/stable/4615682.

[2] W.G. Cochran, The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce, J. Agric. Sci. 30 (2) (1940) 262–275, https://doi.org/10.1017/S0021859600048012.

[3] Lakshmi N. Upadhyaya, Housila P. Singh, Use of transformed auxiliary variable in estimating the finite population mean, Biom. J. 41 (5) (1999) 627–636, https://doi.org/10.1002/(SICI)1521-4036(199909)41:5<627::AID-BIMJ627>3.0.CO;2-W.

[4] Giancarlo Diana, Marco Giordan, Pier Francesco Perri, An improved class of estimators for the population mean, Stat. Methods Appl. 20 (2) (2011) 123–140, https://doi.org/10.1007/s10260-010-0156-6.

[5] Ehsan Zamanzade, M. Mahdizadeh, Using ranked set sampling with extreme ranks in estimating the population proportion, Stat. Methods Med. Res. 29 (1) (2020) 165–177, https://doi.org/10.1177/0962280218823793.

[6] Richard Valliant, Alan H. Dorfman, Richard M. Royall, Finite Population Sampling and Inference: A Prediction Approach, Number 04; QA276. 6, V3, John Wiley, New York, 2000, https://www.wiley.com/en-us/Finite+Population+Sampling+and+Inference%3A+A+Prediction+Approach-p-9780471293415.

[7] V.P. Godambe, A unified theory of sampling from finite populations, J. R. Stat. Soc. B 17 (2) (1955) 269–278, https://doi.org/10.1111/j.2517-6161.1955.tb00203.x.

[8] K.R.W. Brewer, Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process, Aust. J. Stat. 5 (3) (1963) 93–105, https://doi.org/10.1111/j.1467-842X.1963.tb00288.x.

[9] D. Basu, An essay on the logical foundations of survey sampling, part one*, https://api.semanticscholar.org/CorpusID:58791176, 2011.

[10] Richard M. Royall, The linear least-squares prediction approach to two-stage sampling, J. Am. Stat. Assoc. 71 (355) (1976) 657–664, https://doi.org/10.1080/01621459.1976.10481542.

[11] R.M. Royall, Robustness and optimal design under prediction models for finite populations, Surv. Methodol. 18 (2) (1992) 179–185, https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199200214488.

[12] Raymond L. Chambers, Alan H. Dorfman, Thomas E. Wehrly, Bias robust estimation in finite populations using nonparametric calibration, J. Am. Stat. Assoc. 88 (421) (1993) 268–277, https://doi.org/10.1080/01621459.1993.10594319.

[13] Richard M. Royall, Jay Herson, Robust estimation in finite populations I, J. Am. Stat. Assoc. 68 (344) (1973) 880–889, https://doi.org/10.1080/01621459.1973.10481440.

[14] A.J. Scott, K.R.W. Brewer, E.W.H. Ho, Finite population sampling and robust estimation, J. Am. Stat. Assoc. 73 (362) (1978) 359–361, https://doi.org/10.1080/01621459.1978.10481582.

[15] Richard M. Royall, Dany Pfeffermann, Balanced samples and robust Bayesian inference in finite population sampling, Biometrika 69 (2) (1982) 401–409, https://doi.org/10.1093/biomet/69.2.401.

[16] William G. Cumberland, R.M. Royall, Does simple random sampling provide adequate balance?, J. R. Stat. Soc. B 50 (1) (1988) 118–124, https://doi.org/10.1111/j.2517-6161.1988.tb01717.x.

[17] Jean-Claude Deville, Yves Tillé, Efficient balanced sampling: the cube method, Biometrika 91 (4) (2004) 893–912, https://doi.org/10.1093/biomet/91.4.893.

[18] Nedyalkova Desislava, Yves Tillé, Optimal sampling and estimation strategies under the linear model, Biometrika 95 (3) (2008) 521–537, https://doi.org/10.1093/biomet/asn027.

[19] Ray Chambers, Robert Clark, An Introduction to Model-Based Survey Sampling with Applications, vol. 37, OUP, Oxford, 2012, https://global.oup.com/academic/product/an-introduction-to-model-based-survey-sampling-with-applications-9780198566625?cc=pk&lang=en&.

[20] Jean-Claude Deville, Échantillonnage équilibré exact poisonnien, in: 8ème Colloque Francophone sur les Sondages, 2014, pp. 1–6, http://paperssondages14.sfds.asso.fr/submission_36.pdf.

[21] Piero Demetrio Falorsi, Paolo Righi, A unified approach for defining optimal multivariate and multi-domains sampling designs, in: Topics in Theoretical and Applied Statistics, Springer, 2016, pp. 145–152.

[22] Shweta Chauhan, B.V.S. Sisodia, Model based prediction of finite population total under super population model, J. Reliab. Stat. Stud. (2018) 57–68, https://journals.riverpublishers.com/index.php/JRSS/article/view/20871.

[23] Yuki Kawakubo, Genya Kobayashi, Small area estimation of general finite-population parameters based on grouped data, arXiv preprint, arXiv:1903.07239, 2019, https://doi.org/10.48550/arXiv.1903.07239.

[24] Conlet Kikechi, Simwa Onyino, Ganesh Pokhariyal, On prediction based robust estimators of finite population totals, pp. 101–107, https://www.mathsjournal.com/pdf/2019/vol4issue6/PartB/4-5-17-732.pdf, 12 2019.

[25] Shakeel Ahmed, Javid Shabbir, Sat Gupta, Frank Coolen, Estimation of small area total with randomized data, REVSTAT Stat. J. 18 (2) (2020) 223–235, https://doi.org/10.57805/revstat.v18i2.298.

[26] Razieh Jafaraghaie, Prediction of finite population parameters using parametric model under some loss functions, Commun. Stat., Theory Methods (2020) 1–20, https://doi.org/10.1080/03610926.2020.1801736.

[27] Isabel Molina, Malay Ghosh, Accounting for dependent informative sampling in model-based finite population inference, Test 30 (1) (2021) 179–197, https://doi.org/10.1007/s11749-020-00708-0.

[28] Shakeel Ahmed, Javid Shabbir, A novel basis function approach to finite population parameter estimation, Sci. Iran. (2021), https://doi.org/10.24200/SCI.2021.56353.4682.

[29] Safora Anjum, Javid Shabbir, Shakeel Ahmed, Model-based estimation for population total under model misspecification using the balanced sampling scheme, Commun. Stat., Simul. Comput. (2022) 1–12, https://doi.org/10.1080/03610918.2022.2053718.

[30] G.A. McIntyre, A method for unbiased selective sampling, using ranked sets, Aust. J. Agric. Res. 3 (4) (1952) 385–390, https://doi.org/10.1071/AR9520385.

[31] Hani M. Samawi, Hassen A. Muttlak, Estimation of ratio using rank set sampling, Biom. J. 38 (6) (1996) 753–764, https://doi.org/10.1002/bimj.4710380616.

[32] Carlos N. Bouza, Ranked set sub sampling the non response strata for estimating the difference of means, Biom. J. 44 (7) (2002) 903–915, https://doi.org/10.1002/1521-4036(200210)44:7<903::AID-BIMJ903>3.0.CO;2-A.

[33] Elizabeth J. Tipton Murff, Thomas W. Sager, The relative efficiency of ranked set sampling in ordinary least squares regression, Environ. Ecol. Stat. 13 (1) (2006) 41–51, https://doi.org/10.1007/s10651-005-5689-8.

[34] Omer Ozturk, Jayant V. Deshpande, Ranked-set sample nonparametric quantile confidence intervals, J. Stat. Plan. Inference 136 (3) (2006) 570–577, https://doi.org/10.1016/j.jspi.2004.07.011.

[35] Douglas A. Wolfe, Ranked set sampling: its relevance and impact on statistical inference, Int. Sch. Res. Not. (2012) 2012, https://doi.org/10.5402/2012/568385.

[36] Ehsan Zamanzade, Nasser Reza Arghami, Michael Vock, A parametric test of perfect ranking in balanced ranked set sampling, Commun. Stat., Theory Methods 43 (21) (2014) 4589–4611, https://doi.org/10.1080/03610926.2012.737495.

[37] M. Al-kadiri, Linear penalized spline model estimation using ranked set sampling technique, Hacet. J. Math. Stat. 46 (4) (2017) 669–683, https://doi.org/10.15672/HJMS.201510314219.

[38] M. Mahdizadeh, Ehsan Zamanzade, Reliability estimation in multistage ranked set sampling, REVSTAT Stat. J. 15 (4) (2017) 565–581, https://doi.org/10.57805/revstat.v15i4.227.

[39] M. Mahdizadeh, Ehsan Zamanzade, Smooth estimation of the area under the ROC curve in multistage ranked set sampling, Stat. Pap. 62 (4) (2021) 1753–1776, https://doi.org/10.1007/s00362-019-01151-6.

[40] G.P. Patil, A.K. Sinha, C. Taillie, Finite population corrections for ranked set sampling, Ann. Inst. Stat. Math. 47 (4) (1995) 621–636, https://doi.org/10.1007/BF01856537.

[41] Omer Ozturk, Konul Bayramoglu Kavlak, Model based inference using ranked set samples, Surv. Methodol. 44 (1) (2018) 1–17, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54925-eng.pdf?st=5Zy0Q9aM.

[42] Omer Ozturk, Konul Bayramoglu Kavlak, Statistical inference using stratified judgment post-stratified samples from finite populations, Environ. Ecol. Stat. 27 (1) (2020) 73–94, https://doi.org/10.1007/s10651-019-00435-2.
[43] Omer Ozturk, Konul Bayramoglu Kavlak, Model-based inference using judgement post-stratified samples in finite populations, Aust. N. Z. J. Stat. 63 (2) (2021) 377–393, https://doi.org/10.1111/anzs.12320.
[44] Shakeel Ahmed, Javid Shabbir, On use of ranked set sampling for estimating super-population total: gamma population model, Sci. Iran. 06 (2019), https://doi.org/10.24200/sci.2019.50976.1946.