Ann Inst Stat Math (2019) 71:63–91
https://doi.org/10.1007/s10463-017-0623-z

Two-stage cluster samples with ranked set sampling designs

Omer Ozturk[1]

Received: 19 September 2016 / Revised: 14 April 2017 / Published online: 24 October 2017

Abstract This paper draws statistical inference for population characteristics using two-stage cluster samples. Cluster samples in each stage are constructed using ranked set sample (RSS), probability-proportional-to-size sample, or simple random sample (SRS) designs. Each RSS sampling design is implemented with and without replacement policies. The paper constructs design-unbiased estimators for population mean, total, and their variances. Efficiency improvement of all sampling designs over SRS sampling design is investigated. It is shown that the efficiency of the estimators depends on the intra-cluster correlation coefficient and choice of sampling designs in stage I and II sampling. The paper also constructs an approximate confidence interval for the population mean (total). For a fixed cost, the optimal sample sizes for stage I and stage II samples are constructed by maximizing the information content of the sample. The proposed sampling designs and estimators are applied to California School District Study and Ohio Corn Production Data.

Keywords Intra-cluster correlation coefficient · Adjusted R^2 · Cluster sample · Finite population correction · Without replacement

Electronic supplementary material The online version of this article (doi:10.1007/s10463-017-0623-z) contains supplementary material, which is available to authorized users.

✉ Omer Ozturk
ozturk.4@osu.edu

[1] Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

1 Introduction

Populations having hierarchical structure are very common in many data collection procedures. A typical cluster sampling involves two sets of populations: a population of primary units and a population of secondary units within each primary unit, where similar observations in a secondary population are nested within a primary unit. This hierarchical structure creates homogeneity and dependence among the secondary population units. Hence, sample observations from clustered populations are not independent. Even though the correlation structure is not the main interest in most cases, its presence should be addressed with appropriate modeling techniques to remove its impact on statistical inference. Otherwise, dependencies may lead to erroneous statistical inference.

Clustered populations are very common in survey sampling studies. A particular example can be given in school district study, where primary interest is in certain aspects of school success, such as achievement score, dropout rates. In this case, different school districts form clusters (primary population) and schools within a school district constitute a secondary population. The schools within a school district are likely to be similar and dependent since they are in the same neighborhood and share similar resources.

Another example can be given in agricultural survey, where the primary interest is the estimation of crop production in a state. In this case, the counties are the primary (clusters) population and farms within counties are the secondary populations. Again farms within counties are more alike than those between counties due to common environmental factors.

In a finite population setting, appropriate sampling procedures for clustered populations are well studied in the literature. Standard text books, Lohr (1999) and Thompson (2002), provide detailed developments for cluster and two-stage samples in simple random sampling setting. In these settings, due to positive correlation within-cluster observations, two-stage cluster sample estimators require larger sample sizes than a simple random sample (SRS) estimator to achieve the same precision. On the other hand, data collection in a two-stage cluster sample is easier and estimators have higher precision per dollar spent on each unit.

In recent years, to increase the information content of cluster and two-stage cluster samples, alternative cost-efficient sampling procedures are suggested in Wang et al. (2016), Nematollahi et al. (2008) and Sud and Mishra (2006). These alternative sampling designs induce additional structure in a dataset through the relative position (rank) of a unit in a small set of size H. The main idea in this cost-efficient sampling method, called ranked set sampling (RSS), is introduced in McIntyre (2005) to reduce the sampling cost in pasture yields. The method uses ranking at the time of sampling before a measurement is made to determine the relative position (rank) of the unit to be measured in a small set. This position information provides additional auxiliary information to stratify the data, increases the information content of the sample, and reduces the sampling cost.

Construction of an RSS sample in the context of agricultural survey in a county involves the following steps. One begins by selecting H sets of H farms from the population of all farms in the county. The crop production is not measured for these

farms, but rather auxiliary information is gathered such as farm size, acre of land planted etc. This auxiliary information is used to rank each set of H farms from the smallest perceived value of crop production to the largest such value. Next, for each set $i = 1, \ldots, H$, the farm with judgment rank i is selected and its crop production is measured, while the other $H - 1$ farms in the set are discarded. Thus, from the original set of H^2 farms, only H are actually measured, whereas the remaining $H(H - 1)$ are used to facilitate ranking and determine which farms to measure. This procedure is repeated d times to achieve an overall sample size of $n = Hd$. The resulting sample consists of H judgment classes, each of size d. In recent years, research in ranked set sampling has drawn considerable attention in the literature. Up-to-date references in RSS can be found in Wolfe (2012) and Hollander et al. (2014) and references therein.

Wang et al. (2016) considered RSS design in a cluster randomized design to estimate the treatment effects in a two sample problem. They used RSS in mixed effect model assuming cluster effect is random. They showed that the use of RSS at cluster level has much bigger impact on efficiency than using RSS in within-cluster level. Nematollahi et al. (2008) in a finite population setting used RSS only in the second stage of a two-stage sampling with replacement policies. Since they use RSS design only in the second stage with replacement, the efficiency improvement of their estimator with respect to a two-stage SRS sample estimator was minimal. Sud and Mishra (2006) in a finite population setting also used a two-stage cluster sample with ranked set sampling design under the assumption that the cluster population sizes are all equal.

The use of RSS design along with existing sampling designs are not very common in the literature with a few exceptions. Muttlak and McDonald (1992) incorporated RSS sampling design with a line intersect method. Sroka (2008) used it in stratified sampling by constructing RSS sample from each stratum. On the other hand, there has been a considerable attention in using RSS design in a finite population setting. Patil et al. (1995) used ranked set sample to estimate population mean for a population of size N when the sample is constructed without replacement. Deshpande et al. (2006) expanded the without replacement policy in Patil et al. (1995) into three different designs, design-0, design-1 and design-2, and constructed confidence intervals for population quantiles. The design-0 constructs the sample with replacement, design-1 constructs the sample by replacing only the unmeasured units, and design-2 constructs the sample by replacing none of the units regardless of whether they were measured or not. Computation of inclusion probabilities and the construction of Horwitz–Thompson-type estimators are investigated in Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2007, 2008), Jafari Jozani and Johnson (2011, 2012), Gokpinar and Ozdemir (2010), Ozturk and Jafari Jozani (2013), Frey (2011) and Ozturk (2014, 2016a). A comprehensive up-to-date literature review in RSS can be found in recent review paper in Wolfe (2012).

In this paper, we provide a comprehensive procedure to provide a unified framework in two-stage cluster sampling in a finite population setting. Each stage can be constructed with three sampling designs, RSS with replacement (D_1), RSS without replacement (D_2), and SRS without replacement (D_3), yielding nine different sampling designs. If the cluster population sizes are not equal, stage one sample may be constructed with probability-proportional-to-size (PPS) sampling (D_4). Section 2 provides a brief description for each one of the designs D_i; $i = 1, \ldots, 4$. Section 3

introduces two-stage cluster sampling designs. Section 4 introduces estimators for population mean and total. We show that these estimators are design unbiased. Section 5 investigates the efficiency of the estimators constructed from nine different designs. If the cluster population sizes equal, we show that the design using D_2 in both stages dominates all the other eight designs in efficiency. Section 6 provides sample size estimation to minimize the variance of the estimator under certain cost model of two-stage cluster sample designs. Section 7 constructs design-unbiased estimators for the variances of estimators of the population mean and total. Section 8 provides empirical evidence for the efficiency of the estimators under various degree of ranking information and dependence structure. Section 9 uses the proposed estimator to estimate the corn production in Ohio and school successes in California. Finally Sect. 10 provides some concluding remarks. The detailed proofs of the theorems are provided in supplementary material.

2 Sampling designs

In this section, we provide some preliminary results for the four sampling designs (D_1, D_2, D_3, D_4) in a finite population setting. Let $\mathcal{P} = \{u_1, \ldots, u_N\}$ be the population, where the population size N is a known integer. We denote the characteristic of interest as Y. The values of Y in the population will be denoted as $y_i; i = 1, \ldots, N$. We assume that $y_i; i = 1, \ldots, N$, are distinct and ordered, $y_{(1)} < \cdots < y_{(N)}$, where $y_{(i)}$ is the ith largest value of Y in the population. Population mean and variance will be denoted by $\bar{y}_N$ and T^2, respectively

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad T^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2.$$

We consider four different sampling designs from population $\mathcal{P}$, RSS with replacement (D_1), RSS without replacement (D_2), SRS without replacement (D_3), and PPS (D_4) design.

RSS with replacement (D_1) Let H and d be two integers denoting the set and cycle sizes. This design selects $n = Hd$ sampling units from $\mathcal{P}$. To construct the sample, one needs to select H units from the population at random without replacement, rank them for their perceived y values based on available auxiliary information, and identify and measure the unit that corresponds to the smallest perceived y values, $y_{r_{[1]}}$, where $r_{[1]}$ identifies one of the units in the population having the smallest perceived rank ($r_{[1]}$) in the set. In other words, $r_{[1]} = i$ for the ith unit in $\mathcal{P}$ for some i. Once $r_{[1]}$ is identified, all H units including the measured one are returned back into the population. One then selects another set of size H and ranks them. This time, the unit corresponding to the second perceived smallest y value, $y_{r_{[2]}}$, is identified and measured. Again once $r_{[2]}$ is identified, all H units are returned into the population. The process is continued in this way, until H units, $r_{[1]}, \ldots, r_{[H]}$, are identified and measured from the population. This is called a cycle. To increase the sample size, this process is repeated d times to obtain an RSS sample of size $n = dH$,

$$D_1 = \{r_{1[1]}, \ldots, r_{d[1]}, \ldots, r_{1[H]}, \ldots, r_{d[H]}\},$$

where $r_{j[h]}$ $(1 \leq r_{j[h]} \leq N)$ is a random variable representing the population identification number (population size N) of a sampling unit that corresponds to the rank h in a set of size H in cycle j. The measured values of the characteristic Y on the sample units in sample D_1 are given by

$$y_{D_1} = \{y_{r_{1[1]}}, \ldots, y_{r_{d[1]}}, \ldots, y_{r_{1[H]}}, \ldots, y_{r_{d[H]}}\}.$$

Since each set is constructed after replacing the previous set of units back in the population, the same unit can be selected more than once. The subscript "1" in D_1 is used to indicate that RSS is constructed with replacement. Since sample is constructed with replacement, random variables $r_{j[h]}$s are all independent. The square brackets in $r_{j[h]} \in D_1$ are used to indicate that ranking procedure may not be accurate. In the farm example, the crop production of farms may be ranked based on the size of the farms. In this case, we may expect that the amount of crop production is in the same order as the farm size. If the order of crop production is different from the order of farm size this may lead to ranking error. The use of square bracket denotes this type of ranking error. If the ranking procedure is accurate, we replace the square brackets with the round parenthesis and write $r_{j(h)}$. This design is equivalent to design-0 in Deshpande et al. (2006).

RSS without replacement (D_2) Construction of this sample follows the same steps as in D_1 except that none of the H units are returned into the population before selecting the next set in all cycles. The selection of sets and construction of cycles are not independent in Design D_2. Thus, the same unit can not appear in the sample more than once. This sample (design) is denoted with D_2, where subscript "2" indicates that sample is constructed without replacement. In sample D_2, even though $r_{j[h]}$s are obtained by ranking the units in different sets, they are not independent due to without replacement selection. This design is equivalent to design-2 in Deshpande et al. (2006).

SRS without replacement (D_3) This design constructs a simple random sample without replacement from a finite population. This sample will be denoted by

$$D_3 = \{s_1, \ldots, s_n\}, \quad y_{D_3} = \{y_{s_1}, \ldots, y_{s_n}\},$$

where n is the sample size. Again random variables y_{s_j}; $j = 1, \ldots, n$, are not independent.

Probability-proportional-to-size (PPS) sampling (D_4) Sample is constructed using probability sampling. Selection probabilities are proportional to the size of the units. In this design, we denote the first- and second-order inclusion probabilities with π_i and $\pi_{i,j}$, respectively, $i, j = 1, \ldots, N, i \neq j$.

Note that construction of designs D_1 and D_2 requires using a ranking procedure to rank sampling units in a set of size H. Unless stated otherwise, it is assumed that ranking procedure may have ranking error under certain consistency principal. The ranking procedure will be called consistent if the same ranking mechanism is used in all sets and ranking mechanism assigns a rank (with possible error) to each unit in the set. Without loss of generality, we use design D_k and sample D_k interchangeably to denote the data and sampling procedures for $k = 1, \ldots, 4$.

For these four sampling designs, $D_k; k = 1, \ldots, 4$, sample averages (or the estimate of the population mean) will be denoted by

$$y_{\bar{D}_k} = \frac{1}{n} \sum_{a_i \in D_k} y_{a_i}, k = 1, 2, 3, \quad y_{\bar{D}_4} = \frac{1}{N} \sum_{a_i \in D_4} \frac{y_{a_i}}{\pi_{a_i}}.$$

In this notation, for example, the estimator $y_{\bar{D}_1}$ can be written in an explicit form as $y_{\bar{D}_1} = \frac{1}{dH} \sum_{h=1}^{H} \sum_{j=1}^{d} y_{r_{j[h]}}$. We note that even though $y_{\bar{D}_1}$ and $y_{\bar{D}_2}$ have the same expression, they have different distributional properties due to the construction of the samples. The notation $\bar{D}_k$ is used to indicate that the average of the characteristic Y is taken over the sample units in design D_k. The proof of the following Lemma is given in Ozturk (2016b).

Lemma 1 *Under a consistent ranking scheme, the estimators $y_{\bar{D}_k}, k = 1, \ldots, 4$, are unbiased for population mean and have variances*

$$\mathrm{Var}(y_{\bar{D}_1}) = \sigma^2_{\bar{D}_1} = \frac{1}{n} \left(T^2 - \frac{1}{H} \sum_{h=1}^{H} (\bar{y}_{[h]} - \bar{y}_N)^2 \right) = \frac{1}{nH} \sum_{h=1}^{H} T^2_{[h]},$$

$$\mathrm{Var}(y_{\bar{D}_2}) = \sigma^2_{\bar{D}_2} = \left(\frac{(N-1-n)T^2}{n(N-1)} - \frac{1}{nH} \sum_{h=1}^{H} (\bar{y}_{[h]} - \bar{y}_N)^2 - \frac{1}{nH} \sum_{h=1}^{H} T_{[h,h]} \right)$$

$$\mathrm{Var}(\bar{D}_3) = \sigma^2_{\bar{D}_3} = \frac{(N-n)T^2}{n(N-1)},$$

$$\mathrm{Var}(\bar{D}_4) = \frac{1}{N^2} \left\{ \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j \right\}$$

where

$$\bar{y}_{[h]} = \sum_{j=1}^{N} P(y_{r_{1[h]}} = y_j; r_{1[h]} \in D_k) y_j,$$

$$T^2_{[h]} = \sum_{j=1}^{N} P(y_{r_{1[h]}} = y_j; r_{1[h]} \in D_k) y_j^2 - \bar{y}^2_{[h]}; k = 1, 2,$$

$$T_{[h,h]} = \sum_{j=1}^{N} \sum_{t \neq j}^{N} P\left(y_{r_{1[h]}} = y_j, y_{r_{2[h]}} = y_t; (r_{1[h]}, r_{2[h]}) \in D_2 \right) y_j y_t - \bar{y}^2_{[h]}.$$

Under certain conditions, these three variances are ordered.

Corollary 1 *Assume that $T_{[h,h']} \leq 0$ for all $h, h' = 1, \ldots, H$. Then the design D_2 is more efficient than design D_1 and design D_3. Hence, the inequalities $\sigma^2_{\bar{D}_2} \leq \sigma^2_{\bar{D}_1}$ and $\sigma^2_{\bar{D}_2} \leq \sigma^2_{\bar{D}_3}$ hold. If $\frac{\sum_{h=1}^{H} T^2_{[h]}}{HT^2} \leq fpc$, then $\sigma^2_{\bar{D}_1} \leq \sigma^2_{\bar{D}_3}$, where $fpc = \frac{N-n}{N-1}$ is finite population correction factor.*

The assumption $T_{(h,h')} < 0$ holds under perfect ranking since sample units are selected without replacement. Hence, order statistics are negatively correlated. We expect that negative correlation also holds under any reasonable ranking methods. The inequality $\sigma^2_{D_1} \leq \sigma^2_{D_3}$ depends on the quality of ranking information. Under perfect ranking, this inequality holds. Under completely random ranking, the inequality is reversed $\sigma^2_{D_1} \geq \sigma^2_{D_3}$. In this case, there is no ranking information to stratify the sample and sample units in design D_1 are independent, while units in design D_3 are negatively correlated. Hence, due to negative correlations, $\sigma^2_{D_1} \geq \sigma^2_{D_3}$ for finite N and $\sigma^2_{D_1} = \sigma^2_{D_3}$ for large N.

Corollary 2 *Under perfect ranking, the marginal and joint probability mass functions of* $y_{r_{1[h]}}$ *and* $y_{r_{1[h]}}, y_{r_{2[h]}}$ *are given in* Ozturk (2016b)

$$\begin{aligned}
\beta_{D_k}(y_i, h) &= P\left(y_{r_{1(h)}} = y_i; r_{1(h)} \in D_k\right) = \frac{\binom{i-1}{h-1}\binom{N-i}{H-h}}{\binom{N}{H}}; \; y_i \in \mathcal{P}; \; k = 1, 2, \\
\beta_{D_1}(y_i, y_j, h, h') &= P\left(y_{r_{1(h)}} = y_i, y_{r_{2(h')}} = y_j; (r_{1(h)}, r_{2(h')}) \in D_1\right) \\
&= \beta_{D_1}(i, h)\beta_{D_1}(i, h'); \; (y_i, y_j) \in \mathcal{P}, \\
\beta_{D_2}(i, j, h, h') &= P\left(y_{r_{1(h)}} = y_i, y_{r_{2(h')}} = y_j; (r_{1(h)}, r_{2(h')}) \in D_2\right) \\
&= \sum_{\lambda=0}^{j-i-1} \frac{I(i < j)\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}\binom{j-1-h-\lambda}{h'-1}\binom{N-j-H+\lambda+h}{H-h'}}{\binom{N}{H}\binom{N-H}{H}},
\end{aligned}$$

where $I(a < b)$ *is one (zero) if* $a < b$ $(a > b)$ *and* $(y_i, y_j) \in \mathcal{P}$.

It is now clear that $\sigma^2_{\bar{D}_k}; \; k = 1, 2$ can be computed under perfect ranking using Corollary 2.

3 Two-stage cluster sample

In this section, we consider two-stage cluster sampling, where population of primary sampling units contain N clusters, $\mathcal{P}^I = \{u_1, \ldots, u_N\}$. Population of secondary sampling units for each one of the cluster u_i in $\mathcal{P}^{\mathcal{I}}$ contains M_i units $\mathcal{P}^{II}_{u_i} = \{u_1, \ldots, u_{M_{u_i}}\}$. We assume that N and $M_i; \; i = 1, \ldots, N$, are known integers. The value of the characteristic of interest Y of a unit u_j in cluster $u_i, u_j \in \mathcal{P}^{II}_{u_i}$ and $u_i \in \mathcal{P}^I$, is denoted by y_{u_i,u_j}. If we drop u_j in this notation, y_{u_i} denotes the population total of cluster u_i with respect to characteristic Y. Without loss of generality, we assume that population units are ordered in $\mathcal{P}^I$ and $\mathcal{P}^{II}_{u_i}$ for each u_i, i.e, $y_{(1)} < \cdots < y_{(N)}$ and $y_{(i,1)} < \cdots < y_{(i,M_i)}$ for $i = 1, \ldots, N$.

Two-stage cluster sample can be constructed by using any one of the designs $D_k; \; k = 1, \ldots, 4$, in stage I, and any one of the designs $D_k; k = 1, 2, 3$ in stage II, yielding 12 different sampling designs. The stage I sample can be constructed from population $\mathcal{P}^I$ with design $D_k, k = 1, 2, 3$. In this case, the samples from population $\mathcal{P}^I$ will be denoted with D^I_k to identify that they are constructed from the cluster population $\mathcal{P}^I$.

The set, cycle and sample sizes will be denoted by H_1, d_1, and n_1, respectively. For example, sample of selected units for design D_1^I in stage I are denoted with

$$D_1^I = \left\{ r_{1[1]}, \ldots, r_{d_{1[H_1]}} \right\},$$

where $r_{j[h]}$ $(1 \leq r_{j[h]} \leq N)$ is the selected population unit $r_{j[h]}$ in $\mathcal{P}^I$ that corresponds to the rank h in a set of size H_1 in cycle j. Similar notation can also be written for other sampling designs, D_2^I, D_3^I, and D_4^I.

In stage I sampling, y_i represents the population total of cluster i of the characteristic Y and may not be available. On the other hand, y_is are usually proportional to cluster sizes M_i, $i = 1, \ldots, N$. Thus, in the construction of designs D_1^I and D_2^I, ranking procedure can be performed based on either M_is or previous census results. This may introduce some ranking error, but the sampling designs produce design-unbiased estimator and improved efficiency for estimation of population mean and total as long as we have a consistent ranking scheme. If the cluster sizes (M_i) are available and they are not all equal, design D_3^I can be replaced with design D_4^I.

After the stage I sample is constructed, say D_k^I, we construct a stage II sample from population $\mathcal{P}_{a_i}^{II}$ for each one of the selected sample unit $a_i \in D_k^I$. In stage II, we can again use any one of the sampling designs D_q; $q = 1, 2, 3$. These samples will be denoted with $D_q^{II}(a_i)$; $q = 1, 2, 3$, for each a_i, $a_i \in D_k^I$; $k = 1, \ldots, 4$. In the second stage, the set, cycle, and sample sizes will be denoted by H_{2,a_i} , d_{2,a_i} and n_{2,a_i}, respectively. The sample units of second stage for design D_2 from population $\mathcal{P}_{a_i}^{II}$ can be written as

$$D_2^{II}(a_i) = \left\{ r_{1[1]}, \ldots, r_{d_{2,a_i}[H_{2,a_i}]} \right\}; a_i \in D_k^I; k = 1, \ldots, 4,$$

where $r_{j[h]}$ is the selected unit from the secondary population $\mathcal{P}_{a_i}^{II}$ having rank h in a set of size H_{2,a_i} in cycle j. If there is no confusion, to simplify the notation we drop the terms in the round parentheses in $D_q^{II}(a_i)$ and write D_q^{II}, $q = 1, 2, 3$.

In a two-stage cluster sample, the sample observations on selected units will be denoted with $y_{D_k^I, D_q^{II}}$, $k = 1, \ldots, 4$; $q = 1, 2, 3$. For example, for $k = 1$, $q = 1$, and equal set and cycle sizes $(H_{2,a_i} \equiv H_2, d_{2,a_i} \equiv d_2)$, $y_{D_1^I, D_1^{II}}$ is given by

$$y_{D_1^I, D_1^{II}} = \left\{ y_{r_{1[1]}, r_{1[1]}}, \ldots, y_{r_{1[1]}, r_{d_2[H_2]}}, \ldots, y_{r_{d_1[H_1]}, r_{1[1]}}, \ldots, y_{r_{d_1[H_1]}, r_{d_2[H_2]}} \right\},$$

where $y_{r_{i[h]}, r_{j[k]}}$ is the value of Y on unit $r_{j[k]}$ in cluster population $\mathcal{P}_{r_{i[h]}}^{II}$, $r_{i[h]} \in D_1^I$ and $r_{j[k]} \in D_1^{II}(r_{i[h]})$. If $k = 2$, $q = 3$ and $n_{2,a_i} \equiv n_2$, $y_{D_2^I, D_3^{II}}$ becomes

$$y_{D_1^I, D_3^{II}} = \left\{ y_{r_{1[1]}, s_1}, \ldots, y_{r_{1[1]}, s_{n_2}}, \ldots, y_{r_{d_1[H_1]}, s_1}, \ldots, y_{r_{d_1[H_1]}, s_{n_2}} \right\}.$$

4 Estimators for population mean and total

We now construct design-unbiased estimator for population mean and total. The population mean ($\bar{y}$) and total (t) of a two-stage cluster populations are given by

$$\bar{y} = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{i,j}, \quad t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{i,j}, \quad \bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i.$$

Let

$$y_{\bar{D}_k^I, \bar{D}_q^{II}} = \frac{1}{\bar{M}} \frac{1}{n_1} \sum_{a_i \in D_k^I} M_{a_i} y_{a_i, \bar{D}_q^{II}} = \frac{1}{\bar{M} n_1} \sum_{a_i \in D_k^I} M_{a_i} \frac{1}{n_{2,a_i}} \sum_{b_j \in D_q^{II}} y_{a_i, b_j}; \quad k, q = 1, 2, 3$$

$$t_{\bar{D}_k^I, \bar{D}_q^{II}} = \frac{N}{n_1} \sum_{a_i \in D_k^I} M_{a_i} y_{a_i, \bar{D}_q^{II}} = \frac{N}{n_1} \sum_{a_i \in D_k^I} M_{a_i} \frac{1}{n_{2,a_i}} \sum_{b_j \in D_q^{II}} y_{a_i, b_j}; \quad k, q = 1, 2, 3,$$

and

$$y_{\bar{D}_4^I, \bar{D}_q^{II}} = \frac{1}{\sum_{i=1}^{N} M_i} \sum_{a_i \in D_4^I} \frac{M_{a_i} y_{a_i, \bar{D}_q^{II}}}{\pi_{a_i}}, \quad t_{\bar{D}_4^I, \bar{D}_q^{II}} = \sum_{a_i \in D_4^I} \frac{M_{a_i} y_{a_i, \bar{D}_q^{II}}}{\pi_{a_i}} \quad q = 1, 2, 3,$$

where $n_1 = d_1 H_1$ and $n_{2,a_i} = d_{2,a_i} H_{2,a_i}$. We note that $y_{\bar{D}_3^I, \bar{D}_3^{II}}$ and $t_{\bar{D}_3^I, \bar{D}_3^{II}}$ are the estimators of population mean and total based on two-stage cluster samples, where SRS design is used in both stages.

Lemma 2 *For any consistent ranking scheme,* $y_{\bar{D}_k^I, \bar{D}_q^{II}}$ *and* $t_{\bar{D}_k^I, \bar{D}_q^{II}}; k = 1, \ldots, 4, q = 1, 2, 3$*, are design unbiased for* $\bar{y}$ *and* t

$$E(y_{\bar{D}_k^I, \bar{D}_q^{II}}) = \bar{y} \text{ and } E(t_{\bar{D}_k^I, \bar{D}_q^{II}}) = t.$$

We now consider the variances of these estimators. For notational convenience, we define the population mean and variance of cluster (primary unit) u_i in population $\mathcal{P}^I$

$$\bar{y}_{u_i} = \frac{1}{M_{u_i}} \sum_{j=1}^{M_{u_i}} y_{u_i, j}, \quad T_{u_i}^2 = \frac{1}{M_{u_i}} \sum_{j=1}^{M_{u_i}} (y_{u_i, j} - \bar{y}_{u_i})^2, \quad u_i \in \mathcal{P}^I.$$

From Lemma 1, for a given value of $a_i \in D_k^I$, the conditional variance of the sample average of a second-stage sample for designs $D_q^{II}; q = 1, 2, 3$, are given by

$$\begin{aligned}
\mathrm{Var}(y_{a_i,\bar{D}_1^{II}}|a_i) &= \sigma^2_{a_i,\bar{D}_1^{II}} = \frac{1}{n_{2,a_i}}\left(T^2_{a_i} - \frac{1}{H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}}(\bar{y}_{[h]|a_i} - \bar{y}_{a_i})^2\right)\\
&= \frac{1}{n_{2,a_i}H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}} T^2_{[h]|a_i},\\
\mathrm{Var}(y_{a_i,\bar{D}_2^{II}}|a_i) &= \sigma^2_{a_i,\bar{D}_2^{II}} = \frac{(M_{a_i}-1-n_{2,a_i})T^2_{a_i}}{n_{2,a_i}(M_{a_i}-1)} - \frac{1}{n_{2,a_i}H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}}(\bar{y}_{[h]|a_i} - \bar{y}_{a_i})^2\\
&\quad - \frac{1}{n_{2,a_i}H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}} T_{[h,h]|a_i}\\
&= \frac{-T^2_{a_i}}{M_{a_i}-1} + \frac{1}{n_{2,a_i}H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}} T^2_{[h]|a_i} - \frac{1}{n_{2,a_i}H_{2,a_i}}\sum_{h=1}^{H_{2,a_i}} T_{[h,h]|a_i},\\
\mathrm{Var}(y_{a_i,\bar{D}_3^{II}}|a_i) &= \sigma^2_{a_i,\bar{D}_3^{II}} = \frac{(M_{a_i}-n_{2,a_i})T^2_{a_i}}{n_{2,a_i}(M_{a_i}-1)},
\end{aligned}$$

where

$$\begin{aligned}
\bar{y}_{[h]|a_i} &= \sum_{j=1}^{M_{a_i}} P(y_{a_i,r_{1[h]}} = y_{a_i,j})y_{a_i,j},\\
T^2_{[h]|a_i} &= \sum_{j=1}^{M_{a_i}} P(y_{a_i,r_{1[h]}} = y_{a_i,j})y^2_{a_i,j} - \bar{y}^2_{[h]|a_i}\\
T_{[h,h]|a_i} &= \sum_{j=1}^{M_{a_i}}\sum_{t\neq j}^{M_{a_i}} P(y_{a_i,r_{1[h]}} = y_{a_i,j}, y_{a_i,r_{2[h]}} = y_{a_i,t})y_{a_i,j}y_{a_i,t} - \bar{y}^2_{[h]|a_i}.
\end{aligned}$$

Note that the expression $\bar{y}_{[h]|a_i}$ is the conditional mean of the hth judgment order statistic $y_{a_i,r_1[h]}$ in sample D_1^{II} or D_2^{II} given cluster (primary unit) a_i. The expression $T_{[h,h]|a_i}$ is the conditional covariance between judgment order statistics $y_{a_i,r_1[h]}$ and $y_{a_i,r_2[h]}$ in design D_2^{II} given cluster a_i. In design D_1^{II}, $T_{[h,h]|a_i}$ is zero, since sample is selected with replacement, and sample observations are independent.

Theorem 1 *Under any consistent ranking scheme, the variance of estimators* $y_{\bar{D}_k^I,\bar{D}_q^{II}}$ *and* $t_{\bar{D}_k^I,\bar{D}_q^{II}}$ *are given by*

$$\sigma^2_y(\bar{D}_k^I, \bar{D}_q^{II}) = \mathrm{Var}(y_{\bar{D}_k^I,\bar{D}_q^{II}}) = \frac{1}{n_1 N\bar{M}^2}\sum_{i=1}^{N} M_i^2\sigma^2_{i,\bar{D}_q^{II}} + \frac{\sigma^2_{\bar{D}_k^I}}{\bar{M}^2}; \quad k,q = 1,2,3,$$

$$\sigma_y^2(\bar{D}_4^I, \bar{D}_q^{II}) = \mathrm{Var}(y_{\bar{D}_4^I, \bar{D}_q^{II}}) = \frac{1}{(N\bar{M})^2} \sum_{i=1}^{N} \frac{M_i^2 \sigma^2_{i,\bar{D}_q^{II}}}{\pi_i} + \frac{\sigma^2_{\bar{D}_4^I}}{(N\bar{M})^2}; \quad q = 1, 2, 3,$$

and

$$\sigma_t^2(\bar{D}_k^I, \bar{D}_q^{II}) = \mathrm{Var}(t_{\bar{D}_k^I, \bar{D}_q^{II}}) = \frac{N}{n_1} \sum_{i=1}^{N} M_i^2 \sigma^2_{i,\bar{D}_q^{II}} + N^2 \sigma^2_{\bar{D}_k^I}; \quad k, q = 1, 2, 3,$$

$$\sigma_t^2(\bar{D}_4^I, \bar{D}_q^{II}) = \mathrm{Var}(t_{\bar{D}_4^I, \bar{D}_q^{II}}) = \sum_{i=1}^{N} \frac{M_i^2 \sigma^2_{i,\bar{D}_q^{II}}}{\pi_i} + \sigma^2_{\bar{D}_k^I}; \quad q = 1, 2, 3,$$

where $\sigma^2_{\bar{D}_k^I}$ *is given in Lemma* 1 *with* $\bar{D}_k^I = \bar{D}_k$*;* $k = 1, \ldots, 4$.

Corollary 3 (i) *If judgment order statistics are negatively correlated in design* D_2^I *and design* D_2^{II}*, then* $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II}) \leq \sigma_y^2(\bar{D}_k^I, \bar{D}_q^{II})$; $k, q = 1, 2, 3$.
(ii) *If the following inequality holds*

$$\frac{1}{N} \sum_{i=1}^{N} \frac{M_i^2 T_i^2}{n_{2,i}} \left\{ \frac{\sum_{h=1}^{H_{2,i}} T^2_{[h]|i}}{H_{2,i} T_i^2} - \mathrm{fpc}_i \right\} + T^2 \left\{ \frac{\sum_{h=1}^{H_1} T_{[h]}}{H_1 T^2} - \mathrm{fpc} \right\} \leq 0,$$

then $\sigma_y^2(\bar{D}_1^I, \bar{D}_1^{II}) \leq \sigma_y^2(\bar{D}_3^I, \bar{D}_3^{II})$*, where* $\mathrm{fpc}_i = (M_i - n_{2,i})/(M_i - 1)$ *and* $\mathrm{fpc} = (N - n_1)/(N - 1)$ *are the finite population correction factors.*

Expression in the curly brackets in Corollary 3 can be written in terms of the efficiency of RSS sample means. Let $\mathrm{RE}_i = \mathrm{var}(\bar{y}_{\mathrm{RSS,i}})/\mathrm{var}(\bar{y}_{\mathrm{SRS,i}})$ be the relative efficiency of RSS sample mean with respect to SRS sample mean both with replacement sampling from cluster i. Then the first curly bracket in Corollary 3 becomes $\mathrm{RE}_i - \mathrm{fpc}_i$. For large M_i (infinite population), $\mathrm{RE}_i - \mathrm{fpc}_i$ is always negative or zero. For finite populations, this difference depends on quality of ranking information. Under random ranking $\mathrm{RE}_i - \mathrm{fpc}_i \geq 0$, under perfect ranking $\mathrm{RE}_i - \mathrm{fpc}_i \leq 0$. Similar inequalities can be established for the expression in second curly bracket as well. Corollary 3 indicates that two-stage RSS estimator $y_{\bar{D}_1^I, \bar{D}_1^{II}}$ can be less efficient than two-stage SRS estimator $y_{\bar{D}_3^I, \bar{D}_3^{II}}$ in finite population setting if the quality of ranking information is poor.

5 Efficiency of the two-stage cluster sample estimators

An estimator constructed from a two-stage SRS cluster sample usually yields less precision than the one would be obtained from an SRS sample. In general, the efficiency depends on the relative magnitude of between- and within-cluster variation. If the between-cluster variation is higher than the within-cluster variation, a two-stage SRS cluster sample estimator is less efficient than a usual SRS estimator. To study the impact of intra-cluster correlation coefficient (ICC) on the proposed estimators, we assume that the population sizes in all clusters are equal ($M_i = M, i = 1, \ldots, N$).

For unequal cluster population sizes, one can replace ICC with adjusted R^2 to account the impact of the homogeneity of cluster populations. Readers are referred to Sect. 9 for unequal cluster sizes. We also use the same set and cycle sizes in stage II sampling ($H_2 \equiv H_{2,i}$ and $d_2 \equiv d_{2,i}$). Under these assumptions, the ICC is given by

$$\text{ICC} = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}, \quad \text{SSW} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{i,j} - \bar{y}_i)^2, \quad \text{SSTO} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{i,j} - \bar{y})^2,$$

where SSW and SSTO are the sum of square within-cluster and sum of square total errors. It is clear that $-1/(M-1) \leq \text{ICC} \leq 1$. In a two-stage SRS setting, the estimators $y_{\bar{D}_3^I, \bar{D}_3^{II}}$ and $t_{\bar{D}_3^I, \bar{D}_3^{II}}$ are less (more) efficient if ICC > 0 (ICC < 0) (Chapter 5 in Lohr 1999). ICC provides a measure of homogeneity in clusters. The positive values indicate that within-cluster units are similar since they share similar environmental factor. Hence, SSW tends to be smaller with respect to SSTO, yielding positive ICC. The negative values of ICC indicates that within-cluster units are more dispersed than the units selected at random from the entire population so that the ratio SSW/SSTO ≈ 1. This leads to ICC < 0.

Note that when the cluster populations have the same size, it is easy to show that

$$T^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \frac{M^2}{N}\sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2 = \frac{M}{N}\text{SSB},$$
$$\sum_{i=1}^{N} T_i^2 = \sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}(y_{i,j} - \bar{y}_i)^2 = \frac{1}{M}\text{SSW},$$

where SSB is sum of square between-cluster errors. Using these expression, one can rewrite $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ as follows

$$\begin{aligned}
\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II}) &= \frac{-\sum_{i=1}^{N} T_i^2}{n_1 N(M-1)} - \frac{T^2}{M^2(N-1)} \\
&\quad + \frac{1}{n_1 N n_2 H_2}\sum_{i=1}^{N}\sum_{h=1}^{H_2}(T_{[h]|i}^2 - T_{[h,h]|i}) + \frac{1}{M^2 n_1 H_1}\sum_{h=1}^{H_1}(T_{[h]}^2 - T_{[h,h]}) \\
&= \frac{\text{SSW}}{n_1 N M(M-1)} + \frac{SSB}{MN(N-1)} + \frac{1}{n_1 N n_2 H_2}\sum_{i=1}^{N}\sum_{h=1}^{H_2}(T_{[h]|i}^2 - T_{[h,h]|i}) \\
&\quad + \frac{1}{M^2 n_1 H_1}\sum_{h=1}^{H_1}(T_{[h]}^2 - T_{[h,h]}). \qquad (1)
\end{aligned}$$

It is clear that $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ depends on ranking mechanism through unknown population characteristics such as variances and covariance of judgment order statistics. We first express these variances and covariances as functions of relative efficiencies of RSS and SRS sample mean estimators. Let $W_1 = \sum_{i=1}^{N} y_{i,\bar{D}_2^{II}}$ and $W_{\text{SRS}} = \sum_{i=1}^{N} y_{i,\bar{D}_3^{II}}$.

We now express $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ as a function of ICC and efficiencies of stage I and stage II RSS sampling designs.

Lemma 3 *Let* $\eta_1 = \text{var}(y_{\bar{D}_2^I})/\text{var}(y_{D_3^I})$ *and* $\eta_2 = var(W_1)/\text{var}(W_{SRS})$. *Under any consistent ranking scheme*

$$\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II}) = \frac{1}{n_1}\left\{\frac{\eta_1 \text{MSB}}{M} - \frac{\eta_2 \text{MSW}}{M}\right\} + \frac{1}{n_1 n_2}\eta_2 \text{MSW} - \frac{\eta_1 \text{MSB}}{NM},$$

and

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})}{\text{MST}} = \frac{\eta_1(N-n_1)}{n_1 NM}\frac{\{(M-1)\text{ICC}+1\}(NM-1)}{M(N-1)} + \frac{\eta_2(M-n_2)}{n_1 n_2}\frac{(1-\text{ICC})(NM-1)}{NM^2},$$

where MSW *and MST are mean square of within errors and between errors, respectively.*

It is now clear that $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ is a function of ICC, η_1 and η_2. Under perfect ranking, $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ can be computed using Lemma 1 and Corollary 2. Under imperfect ranking, the variance depends on ranking mechanism through η_1 and η_2. Both η_1 and η_2 are 1 under random ranking and decreases with the improvement in ranking quality. Lower bounds of η_1 and η_2 depend on the set sizes, cycle sizes, and the population.

For wide range of choices of η_1 and η_2, Fig. 1 shows surface plots for $\sigma_{\bar{y}}^2$, when ICC $= 0.05$ and ICC $= 0.5$ for fixed values of $N = M = 50, n_1 = n_2 = 4$. Panel (a) in Fig. 1 indicates that $\sigma_{\bar{y}}^2$ is a decreasing function of η_2, while it is relatively constant with respect to η_1. Hence, when ICC is small, the quality of ranking information in stage II sampling is much more important than the quality of ranking information in stage I sampling. The opposite is true in panel (b), where ICC $= 0.5$. The quantity $\sigma_{\bar{y}}^2$ is decreasing function of η_1 while it is relatively constant in η_2. In this case, the quality of ranking information in stage I sampling is much more important than the quality of ranking information in stage II sampling.

We now investigate the relative efficiency of the estimator $y_{\bar{D}_k^I, \bar{D}_q^{II}}$ with respect to the estimator $y_{\bar{D}_2^I, \bar{D}_2^{II}}$. Let

$$RE_{k,q} = \frac{\text{Var}(y_{\bar{D}_k^I, \bar{D}_q^{II}})}{\text{Var}(y_{\bar{D}_2^I, \bar{D}_2^{II}})} = \frac{\sigma_y^2(\bar{D}_k^I, \bar{D}_q^{II})}{\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})}; \quad k, q = 1, 2, 3.$$

The values of $\text{RE}_{k,q}$ greater than one indicate that the estimator $y_{\bar{D}_2^I, \bar{D}_2^{II}}$ outperforms the estimator $y_{\bar{D}_k^I, \bar{D}_q^{II}}$. Efficiencies are computed under a perfect ranking scheme with stage I population size $N = 100$ and stage II population sizes $M_i = 50; i = 1, \ldots, N$. The stage I population is generated from discrete normal population with

$$c_i = Q(i/(N+1), 100, \tau_I^2); i = 1, \ldots, N,$$

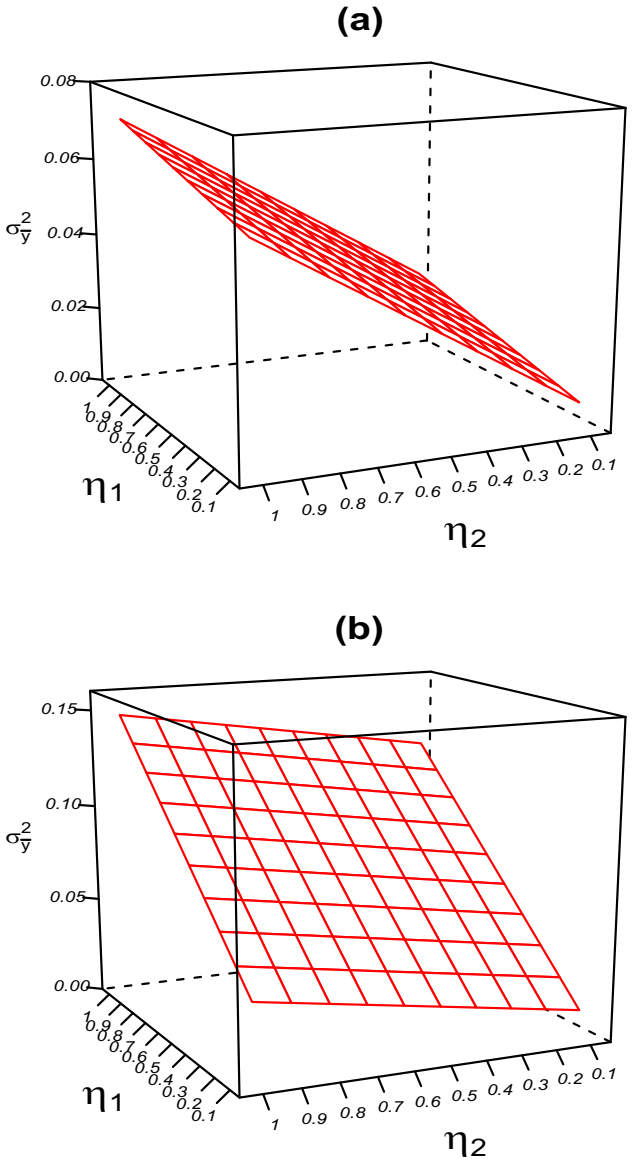

Fig. 1 Scaled variance of $y_{\bar{D}_2^I, \bar{D}_2^{II}}$ ($\sigma^2_{\bar{y}} = \sigma^2_y(\bar{D}_2^I, \bar{D}_2^{II})$/MST) with respect to η_1 and η_2 with $N = M = 50$, $n_1 = n_2 = 4$ for **a** ICC $= 0.05$ and **b** ICC $= 0.5$

where $Q(p_i, \mu, \tau_I^2)$ is a quantile function of normal distribution with mean μ and variance τ_I^2. For each value of c_i; $i = 1, \ldots, N$, the second stage populations are generated by

$$y_{i,j} = c_i + Q(j/(M+1), 10, \tau_{II}^2);\ j = 1, \ldots, M.$$

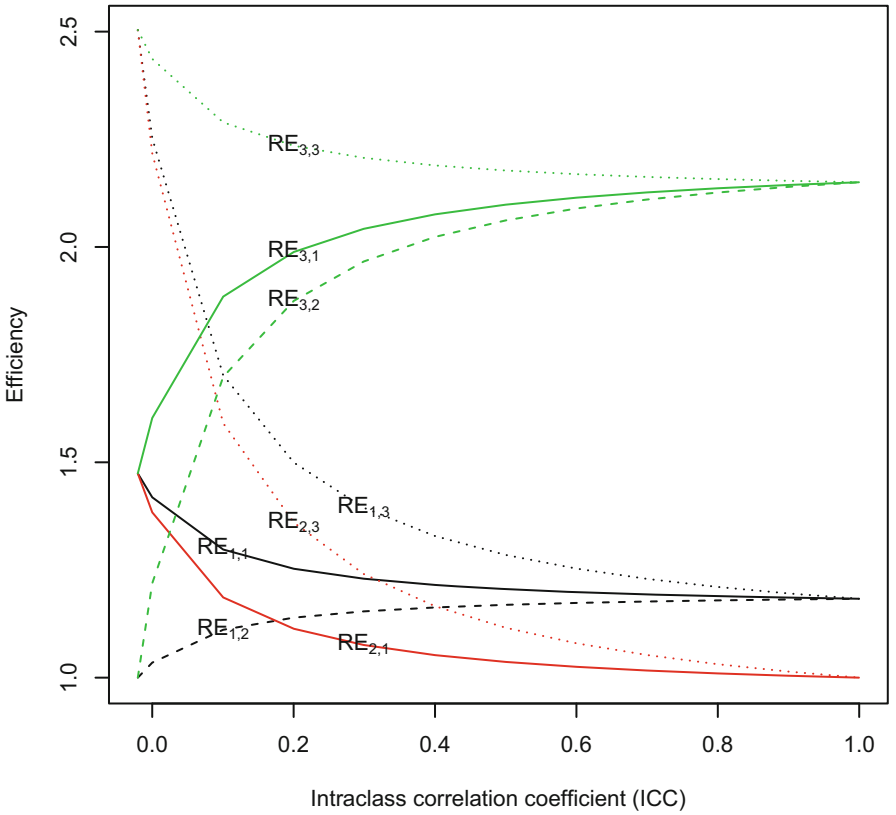

Fig. 2 Relative efficiencies of the estimator $y_{\bar{D}_k^I,\bar{D}_q^{II}}$ with respect $y_{D_2^I,\bar{D}_2^{II}}$, $\mathrm{RE}_{k,q} = \frac{\sigma_y^2(\bar{D}_k^I,\bar{D}_q^{II})}{\sigma_y^2(\bar{D}_2^I,\bar{D}_2^{II})}$

The variances τ_I^2 and τ_{II}^2 are selected in such a way that the desired ICC values would be produced. For example, for ICC $= 0.5$, we used $\tau_I = 10$ and $\tau_{II} = 10.09$.

Figure 2 presents the efficiency curves of the estimators. Efficiencies are computed under perfect ranking using Corollary 2. It is clear that all efficiency curves are greater than one indicating that the estimator $y_{\bar{D}_2^I,\bar{D}_2^{II}}$ outperforms all the other eight estimators. It is also clear that all curves lie below the curve of $\mathrm{RE}_{3,3}$. Hence, use of RSS design in estimator $y_{\bar{D}_k^I,\bar{D}_q^{II}}$ yields higher efficiencies than the two-stage SRS cluster sample estimator $y_{\bar{D}_3^I,\bar{D}_3^{II}}$. The impact of the designs D_k^I and D_q^{II} on the efficiency depends on the magnitude of the ICC. For a fixed design D_k^I in stage I sampling (say D_3^I), efficiency curves for all designs D_q^{II} in stage II sampling (say green curves) converge to the same value as ICC approaches to 1. For a fixed design D_q^{II} in stage II sampling (say D_3^{II}), efficiency curves (say dotted lines) are all equal for designs D_k^I, $k = 1, 2, 3$, when ICC $= -1/(M-1)$ and diverges when ICC increases. These findings are consistent with the main features of the surface plots in Fig. 1, where ranking information (or use of RSS design) plays important role in stage I (stage II) sampling when ICC is large (small). We also note that the efficiency curves $\mathrm{RE}_{k,q}$

reverses the concavity of $\text{RE}_{q,k}$ and one of the pairs becomes preferred, depending on the concavity switch.

Figure 2 indicates that the design D_2^I, D_2^{II} would be optimal for any ICC, $-1/(M-1) \leq \text{ICC} \leq 1$. If the ICC > 0.4, the design D_2^I, D_q^{II} does not lose too much efficiency for $q = 1, 3$. In this case, either D_2^I, D_1^{II} or D_2^I, D_3^{II} can be used in practice.

Under imperfect ranking, the efficiencies depend on ranking quality through η_1 and η_2. For moderately large N and M, the following approximation provides some insight about the design choices in stage I and II sampling

$$\frac{\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})}{\text{MST}} \approx \frac{\eta_1 \text{ICC}}{n_1} + \frac{\eta_2(1-\text{ICC})}{n_1 n_2}. \tag{2}$$

It is clear that Eq. (2) is an increasing function of ICC if $\eta_1 n_2 > \eta_2$. Hence, the variance of $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ is an increasing function of ICC. The approximation in Eq. (2) indicates that the first term is dominant in the variance of the two-stage cluster sample mean estimator when ICC is large. The quantity η_1 is the efficiency measure of RSS design in stage I. To achieve a substantial reduction in the variance, we should select η_1 as small as possible (RSS should be as efficient as possible in stage I). The design D_2^I would be an optimal choice in stage I sampling. Since the contribution of the second term to variance is relatively small, the efficiency of RSS procedure in stage II is not as crucial as the efficiency of RSS procedure in stage I. For example, if the ranking procedure is completely random in stage II, Eq. (2) reduces to

$$\frac{\sigma_y^2(\bar{D}_2^I, \bar{D}_3^{II})}{\text{MST}} \approx \frac{\eta_1 \text{ICC}}{n_1} + \frac{(1-\text{ICC})}{n_1 n_2}.$$

The loss of efficiency due to using design D_3^{II} in stage II would be $(1-\eta_2)(1-\text{ICC})/(n_1 n_2)$, where $\eta_2 \leq 1$. It is clear that the loss of efficiency is not very large for large ICC.

If ICC is small, the first term in Eq. (2) becomes negligible. To reduce the variance, η_2 should be as small as possible (RSS should be as efficient as possible in stage II sampling). In this case, design D_2^{II} would be optimal choice in stage II.

6 Sample size determination

One of the objectives of a sampling design is to maximize the information content of a sample for a fixed cost. Since our sampling design involves two-stage sampling, where information content of the sample in each stage is different, it is important to determine sample size n_1 and n_2 in stage I and II sampling for a fixed cost. In this section, we again consider cluster population sizes are all equal $M_i = M$ for for the brevity of the presentation. The total cost C_T of the sample will be denoted by

$$C_T = c_0 + n_1 c_1 + n_1 n_2 c_2,$$

where c_0 is start up cost, c_1 and c_2 are the cost per unit in stage I and II sampling, respectively. Under this cost model, we wish to minimize the $\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II})$ with respect to n_1 and n_2

$$\sigma_y^2(\bar{D}_2^I, \bar{D}_2^{II}) = \frac{1}{n_1}\{\eta_1 \text{MSB} - \eta_2 \text{MSW}\} + \frac{1}{n_1 n_2}\eta_2 \text{MSW} - \frac{\eta_1 \text{MSB}}{NM}$$
$$= \frac{A_1}{n_1} + \frac{A_2}{n_1 n_2} - \frac{\eta_1 T^2}{M^2(N-1)}.$$

Minimizer of the above function with respect to n_1 and n_2 subjected to constraint of cost function is given by

$$n_2 = \sqrt{\frac{c_1 A_2}{c_2 A_1}} = \sqrt{\frac{c_1 M \eta_2 \text{MSW}}{c2(-\eta_2 \text{MSW} + \eta_1 \text{MSB})}}, \quad n_1 = \frac{C_T - c_0}{c_1 + n_2 c_2}.$$

The optimal value of n_2 in terms of intra-cluster correlation coefficient can be written as

$$n_2 = \sqrt{\frac{\eta_2 c_1 M(N-1)(1-\text{ICC})}{c_2\{-(1-\text{ICC})\eta_2(N-1) + N\{(M-1)\text{ICC}+1\}\eta_1\}}}$$

for $\text{ICC} > \frac{(\eta_2-\eta_1)N-\eta_2}{\eta_2(N-1)+N(M-1)\eta_1}$. The homogeneity of cluster population can equivalently be measured by adjusted R^2, $R^2 = 1 - \text{MSW}/\text{MST}$. Note that $\text{MSB}/\text{MST} = 1 + N(M-1)R^2/(N-1)$. Using these two equalities, the optimal n_2 is given by

$$n_2 = \sqrt{\frac{c_1 M \eta_2 (1-R^2)(N-1)}{c_2\left\{-\eta_2(1-R^2)(N-1) + \eta_1\left\{N-1+N(M-1)R^2\right\}\right\}}}$$

for $R^2 > (\eta_2 - \eta_1)(N-1)/\{\eta_2(N-1) + \eta_1 N(M-1)\}$.

In certain setting, there may not be available auxiliary observations for ranking in stage two populations. In this case, ranking may not be possible and one may have to use design D_3^{II}. The optimal sample size to minimize the variance of $\sigma_y^2(\bar{D}_2^I, \bar{D}_3^{II})$ can be derived in similar fashion. Using the same approaches that we used in the construction of Eq. (1) and in the proof of Lemma 3, $\sigma_y^2(\bar{D}_2^I, \bar{D}_3^{II})$ can be written as

$$\sigma_y^2(\bar{D}_2^I, \bar{D}_3^{II}) = \frac{M\sum_{i=1}^N T_i^2}{n_2 n_1 N(M-1)} + \frac{1}{n_1}\left\{\frac{\eta_1 N T^2}{M^2(N-1)} - \frac{\sum_{i=1}^N T_i^2}{N(M-1)}\right\} - \frac{\eta_1 T^2}{(N-1)M^2}.$$

The minimizer of the above expression subject to cost constraint is given by

$$n_2 = \sqrt{\frac{M c_1(\text{MSW})}{c_2(\eta_1 \text{MSB} - \text{MSW})}}, \quad n_1 = \frac{C_t - c_0}{c_1 + c_2 n_2}.$$

The optimal n_2 in terms of R^2 and ICC are given by

$$n_2 = \sqrt{\frac{c_1 M(1-R^2)(N-1)}{c_2[\eta_1\{N-1+N(M-1)R^2\}-(1-R^2)(N-1)]}},$$
$$R^2 > \frac{(N-1)(1-\eta_1)}{\eta_1 N(M-1)+(N-1)}$$

and

$$n_2 = \sqrt{\frac{c_1(N-1)M(1-\text{ICC})}{c_2[N\eta_1\{1+(M-1)\text{ICC}\}-(N-1)(1-\text{ICC})]}},$$
$$\text{ICC} > \frac{N(1-\eta_1)-1}{\eta_1 N(M-1)+(N-1)}.$$

To determine optimal n_2, we need a measure of homogeneity of cluster populations. This could either be obtained from intra-cluster correlation coefficient or from adjusted R^2. These values could be available from previous studies or can be obtained with a small pilot study. The quantities η_1 and η_2 may depend on set sizes, population and ranking process. They may also depend on cycle size if the population size is relatively small. Patil et al. (1995) showed that η_1 and η_2 are smaller (RSS is more efficient) in finite population than their values in infinite population. In infinite population setting, Takahasi and Wakimoto (1968) established that

$$\frac{2}{H_1+1} \le \eta_1 \le 1 \text{ and } \frac{2}{H_2+1} \le \eta_2 \le 1.$$

In finite population setting, lower bounds would be even smaller than these lower bounds. We may use these lower bounds as a rough estimate for η_1 and η_2.

A close inspection of n_2 indicates that larger ICC (or R^2) allocates more resources to sample from clusters and less from secondary population. In the extreme case ICC $= 1$, $n_2 = 0$. In this case, all units within a cluster have the same value (cluster mean). One should then sample from stage I population instead of using two-stage cluster sample. Otherwise, measuring more than one units per cluster costs extra time and money without increasing the information content of the sample.

7 Variance estimates

We construct unbiased estimators for the variances of the two-stage cluster sample means. In this section, we assume that set $H_{2,i}$ and cycle sizes $d_{2,i}$ in stage II samples could be different. We need some preliminary results. Let

$$A_{a_i,\tilde{D}_q^{II}} = \frac{1}{2H_{2,a_i}^2 d_{2,a_i}(d_{2,a_i}-1)} \sum_{j=1}^{d_{2,a_i}} \sum_{t\neq j}^{d_{2,a_i}} \sum_{h=1}^{H_{2,a_i}} \left\{y_{a_i,r_{j[h]}} - y_{a_i,r_{t[h]}}\right\}^2;$$

$$a_i \in D_k^I; k = 1, \ldots, 4;\ q = 1, 2,$$

$$B_{a_i,\bar{D}_2^{II}} = \frac{1}{2H_{2,a_i}^2 d_{2,a_i}^2} \sum_{j=1}^{d_{2,a_i}} \sum_{t=1}^{d_{2,a_i}} \sum_{h=1}^{H_{2,a_i}} \sum_{h' \neq h}^{H_{2,a_i}} \left\{ y_{a_i,r_{j[h]}} - y_{a_i,r_{t[h']}} \right\}^2; \quad a_i \in D_k^I; k = 1, 2, 3,$$

$$C_{a_i,\bar{D}_3^{II}} = \left(1 - \frac{n_{2,a_i}}{M_{a_i}}\right) \frac{1}{n_{2,a_i}(n_{2,a_i} - 1)} \sum_{s_j \in D_3^{II}} \left\{ y_{a_i,s_j} - y_{a_i,\bar{D}_3^{II}} \right\}^2; \quad a_i \in D_k^I; k = 1, 2, 3.$$

Lemma 4 *For any $a_i \in D_k^I; k = 1, 2, 3$, let*

$$\hat{\sigma}^2_{a_i,\bar{D}_1^{II}} = \frac{A_{a_i,\bar{D}_1^{II}}}{d_{2,a_i}}, \quad \hat{\sigma}^2_{a_i,\bar{D}_3^{II}} = C_{a_i,\bar{D}_3^{II}}, \text{ and}$$

$$\hat{\sigma}^2_{a_i,\bar{D}_2^{II}} = \frac{A_{a_i,\bar{D}_2^{II}}}{d_{2,a_i}} - \frac{A_{a_i,\bar{D}_2^{II}} + B_{a_i,\bar{D}_2^{II}}}{M_{a_i}}.$$

Under any consistent ranking scheme, the following equalities hold

$$E\left(\hat{\sigma}^2_{a_i,\bar{D}_q^{II}} | a_i\right) = \sigma^2_{a_i,\bar{D}_k^{II}}; \quad q = 1, 2, 3; \quad a_i \in D_k^I; \quad k = 1, 2, 3.$$

Lemma 4 provides an unbiased estimator for the conditional variance of a stage II sample mean, $y_{a_i,\bar{D}_q^{II}}; q = 1, 2, 3$, given that the cluster a_i is sampled in stage I. The construction of an unbiased estimator of the variance of the two-stage sample mean and total requires additional notation. Let

$$A^*_{\bar{D}_k^I,\bar{D}_q^{II}} = \frac{1}{2d_1(d_1 - 1)H_1^2} \sum_{i=1}^{d_1} \sum_{j \neq i}^{d_1} \sum_{h=1}^{H_1} \left\{ M_{r_{i[h]}} y_{r_{i[h]},\bar{D}_q^{II}} - M_{r_{i[h]}} y_{r_{j[h]},\bar{D}_q^{II}} \right\}^2; \ k = 1, 2; q = 1, 2, 3,$$

$$B^*_{\bar{D}_2^I,\bar{D}_q^{II}} = \frac{1}{2d_1^2 H_1^2} \sum_{i=1}^{d_1} \sum_{j=1}^{d_1} \sum_{h=1}^{H_1} \sum_{h' \neq h}^{H_1} \left\{ M_{r_{i[h]}} y_{r_{i[h]},\bar{D}_q^{II}} - M_{r_{j[h']}} y_{r_{j[h']},\bar{D}_q^{II}} \right\}^2; q = 1, 2, 3,$$

$$s^{*2}_y = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left\{ M_{s_i} y_{s_i,\bar{D}_q^{II}} - \frac{1}{n_1} \sum_{j=1}^{n_1} M_{s_j} y_{s_j,\bar{D}_q^{II}} \right\}^2,$$

$$\hat{\sigma}^2_{\bar{D}_4^I} = \sum_{a_i \in D_4^I} \frac{1 - \pi_{a_i}}{\pi^2_{a_i}} z^2_{a_i,\bar{D}_q^{II}} + \sum_{(a_i,a_j) \in \bar{D}_q^{II}} \sum_{j \neq i} \left(\frac{\pi_{a_i a_j} - \pi_{a_i}\pi_{a_j}}{\pi_{a_i}\pi_{a_j}} \right) \frac{z_{a_i,\bar{D}_q^{II}} z_{a_j,\bar{D}_q^{II}}}{\pi_{a_i a_j}},$$

where $z_{a_i,\bar{D}_q^{II}} = M_{a_i} y_{a_i,\bar{D}_q^{II}}$.

Lemma 5 *Let*

$$\hat{\sigma}^2_y(\bar{D}_1^I, \bar{D}_q^{II}) = \frac{1}{d_1 \bar{M}^2} A^*_{\bar{D}_1^I,\bar{D}_q^{II}}$$

$$\hat{\sigma}^2_y(\bar{D}_2^I, \bar{D}_q^{II}) = \left\{ \frac{1}{d_1 \bar{M}^2} A^*_{\bar{D}_2^I,\bar{D}_q^{II}} - \frac{A^*_{\bar{D}_2^I,\bar{D}_q^{II}} + B^*_{\bar{D}_2^I,\bar{D}_q^{II}}}{N\bar{M}^2} \right\} + \frac{1}{N\bar{M}^2 n_1} \sum_{a_i \in D_2^I} M^2_{a_i} \hat{\sigma}^2_{a_i,\bar{D}_q^{II}}$$

$$\hat{\sigma}_y^2(\bar{D}_3^I, \bar{D}_q^{II}) = \frac{N-n_1}{N\bar{M}^2}\frac{s^{*2}_y}{n_1} + \frac{1}{N\bar{M}^2 n_1}\sum_{s_i\in D_3^I}^{n_1} M_{s_i}^2 \hat{\sigma}^2_{s_i,\bar{D}_q^{II}},$$

$$\hat{\sigma}_y^2(\bar{D}_4^I, \bar{D}_q^{II}) = \frac{\hat{\sigma}^2_{\bar{D}_4^I}}{(N\bar{M})^2} + \sum_{a_i\in D_4^I} \frac{M_{a_i}^2 \hat{\sigma}^2_{a_i,\bar{D}_q^{II}}}{(N\bar{M})^2\pi_{a_i}}.$$

Then $\hat{\sigma}_y^2(\bar{D}_k^I, \bar{D}_q^{II})$ *is an unbiased estimator for* $\sigma_y^2(\bar{D}_k^I, \bar{D}_q^{II})$*;* $k = 1, \ldots, 4; q = 1, 2, 3.$

Unbiased estimator for the variance of the estimator $t_{\bar{D}^I,\bar{D}^{II}}$ is obtained by $N^2\bar{M}^2\hat{\sigma}_y^2(\bar{D}^I, \bar{D}^{II})$. We note that these estimators are unbiased for any consistent ranking scheme. It does not rely on perfect ranking of within set units.

Unbiased estimator of variance allows us to construct confidence intervals for the population mean and total. Using normal approximation, approximate confidence interval is given by

$$y_{\bar{D}_k^I,\bar{D}_q^{II}} \pm t_{df,1-\alpha/2}\hat{\sigma}_y(\bar{D}_k^I, \bar{D}_q^{II});\ k = 1, \ldots, 4;\ q = 1, 2, 3, \tag{3}$$

where $t_{df,a}$ is the ath upper quantile of t-distribution with df degrees of freedom. A reasonable choice for df can be obtained from Satterthwaite approximation. We first rewrite the $\mathrm{Var}(y_{\bar{D}_k^I,\bar{D}_q^{II}})$ in Theorem 1

$$\sigma_y^2(\bar{D}_k^I, \bar{D}_q^{II}) = \begin{cases} \frac{\sigma^2_{\bar{D}_k^I}}{\bar{M}^2} + \frac{\sum_{i=1}^N M_i^2\sigma^2_{i,\bar{D}_q^{II}}}{n_1 N\bar{M}^2} = \sigma^2_{y,1}(\bar{D}_k^I, \bar{D}_q^{II}) + \sigma^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II});\ k = 1, 2, 3 \\ \frac{\sigma^2_{\bar{D}_4^I}}{(N\bar{M})^2} + \sum_{i=1}^N \frac{M_i^2\sigma^2_{i,\bar{D}_q^{II}}}{(N\bar{M})^2\pi_i} = \sigma^2_{y,1}(\bar{D}_4^I, \bar{D}_q^{II}) + \sigma^2_{y,2}(\bar{D}_4^I, \bar{D}_q^{II}). \end{cases}$$

One may interpret $\sigma^2_{y,1}(\bar{D}_k^I, \bar{D}_q^{II})$ and $\sigma^2_{y21}(\bar{D}_k^I, \bar{D}_q^{II})$ as the contribution of between- and within-cluster variation to the variance of the estimator $y_{\bar{D}_k^I,\bar{D}_q^{II}}$. Using Lemmas 3 and 4, unbiased estimators of $\sigma^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II})$ and $\sigma^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II})$ are given by

$$\hat{\sigma}^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II}) = \frac{1}{n_1^2\bar{M}^2}\sum_{a_i\in D_k^I} M_{a_i}^2\hat{\sigma}^2_{a_i,\bar{D}_q^{II}};\quad k = 1, \ldots, 4;\ q = 1, 2, 3$$

$$\hat{\sigma}^2_{y,1}(\bar{D}_k^I, \bar{D}_q^{II}) = \hat{\sigma}^2_y(\bar{D}_k^I, \bar{D}_q^{II}) - \hat{\sigma}^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II});\quad k = 1, \ldots, 4;\ q = 1, 2, 3.$$

We now provide the Satterthwaite approximation for the degrees of freedom

$$df = \frac{\left\{\hat{\sigma}^2_{y,1}(\bar{D}_k^I, \bar{D}_q^{II})/n_1 + \hat{\sigma}^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II})/n_2\right\}^2}{\left(\hat{\sigma}^2_{y,1}(\bar{D}_k^I, \bar{D}_q^{II})/n_1\right)^2/(n_1-1) + \left(\hat{\sigma}^2_{y,2}(\bar{D}_k^I, \bar{D}_q^{II})/n_2\right)^2/(n_2-1)}.$$

The df can be rounded to the nearest integer if desired.

8 Empirical evidence

We now look at the efficiency of the estimators as a function of the set size and quality of ranking information in a simulation study. Two-stage cluster population again constructed from discrete normal distribution with $N = M = 120$ and ICC $= 0.5$. In the simulation study, the set sizes $H_1 = H_2$ are taken to be 2, 3, 4, 5. The cycle sizes are selected as $d_1 = 4$ and $d_2 = 3$. Simulation size is taken to be 20,000.

The quality of ranking information is modeled through the additive perceptual error model in Dell and Clutter (1972). This model for a set of size H selects H units from the population, $\boldsymbol{y} = (y_1, \ldots, y_H)$, and generates H random numbers, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_H)$, from a normal distribution with mean zero and variance σ_ϵ^2. It is assumed that y and ϵ are independent. These two vectors are added to form an additive model, $\boldsymbol{v} = \boldsymbol{y} + \boldsymbol{\epsilon}$. The vector $\boldsymbol{v}$ is sorted from smallest to largest and their ranks are assigned as judgment ranks to the component of vector $\boldsymbol{y}$. The quality of ranking information in stage I and stage II sampling is controlled by the correlation coefficient between random variables v and Y, $\rho = \text{corr}(Y, v)$. The ranking quality of a two-stage cluster sample is denoted with ρ^I and ρ^{II}, where ρ^I and ρ^{II} are the correlation coefficients in perceptual error model in stage I and stage II sampling, respectively. The simulation study considered $\rho^I = \rho^{II} = 0, 0.2, \ldots, 1$.

Figure 3 compares the variances of the estimators $y_{\bar{D}_2^I, \bar{D}_2^{II}}$, $y_{\bar{D}_3^I, \bar{D}_2^{II}}$ and $y_{\bar{D}_2^I, \bar{D}_3^{II}}$ with the variance of the estimator $y_{\bar{D}_3^I, \bar{D}_3^{II}}$. Efficiency curves are the ratio of the variances of these three estimators with respect to $\sigma_y^2(\bar{D}_3^I, \bar{D}_3^{II})$. Hence, the curves lying above 1 indicate that the corresponding estimator is more efficient than SRS-based two-stage cluster sample estimator, $y_{\bar{D}_3^I, \bar{D}_3^{II}}$. The first panel in Fig. 3 compares the variance of $y_{\bar{D}_2^I, \bar{D}_2^{II}}$ with the variance of $y_{\bar{D}_3^I, \bar{D}_3^{II}}$. It is clear that the new estimator is better than SRS-based estimator for all ρ. Improvement in efficiency increases substantially with set size H. For example, if $H = 5$, the proposed estimator is roughly four times more efficient than the estimator $y_{\bar{D}_3^I, \bar{D}_3^{II}}$ for $\rho > 0.8$.

The second panel compares $y_{\bar{D}_2^I, \bar{D}_3^{II}}$ with $y_{\bar{D}_3^I, \bar{D}_3^{II}}$. In this panel, we also observe a similar pattern as in the previous panel with a slightly smaller efficiency improvement. This clearly shows that if we do not use RSS design in stage II, since ICC is relatively large (ICC $= 0.5$) the loss of efficiency is not very significant which is consistent with our earlier findings. In fact, for ICC $= 0.5$, the third panel clearly shows that use of RSS design in stage I is crucial for the efficiency of two-stage cluster sample mean. Use of RSS in second stage does not provide significant improvement on the efficiency of the estimator.

Table 1 presents coverage probabilities of the confidence intervals in Eq. (3) based on estimators $y_{\bar{D}_2^I, \bar{D}_2^{II}}$, $y_{\bar{D}_2^I, \bar{D}_3^{II}}$, $y_{\bar{D}_3^I, \bar{D}_2^{II}}$, $y_{\bar{D}_3^I, \bar{D}_3^{II}}$, and SRS when intra-cluster correlation coefficient is 0.1 and 0.5. In this part of the simulation we again used the same simulation parameters as in Fig. 3. Table 1 shows that the empirical coverage probabilities are reasonably close to the nominal coverage probability 0.95. This may be anticipated from the fact that underlying population is generated from discrete normal distribution. We also performed simulations with skewed and heavy tailed populations and observed that simulated coverage probabilities are smaller than nominal value 0.95. Due to space limitation, these results are not reported in this paper.

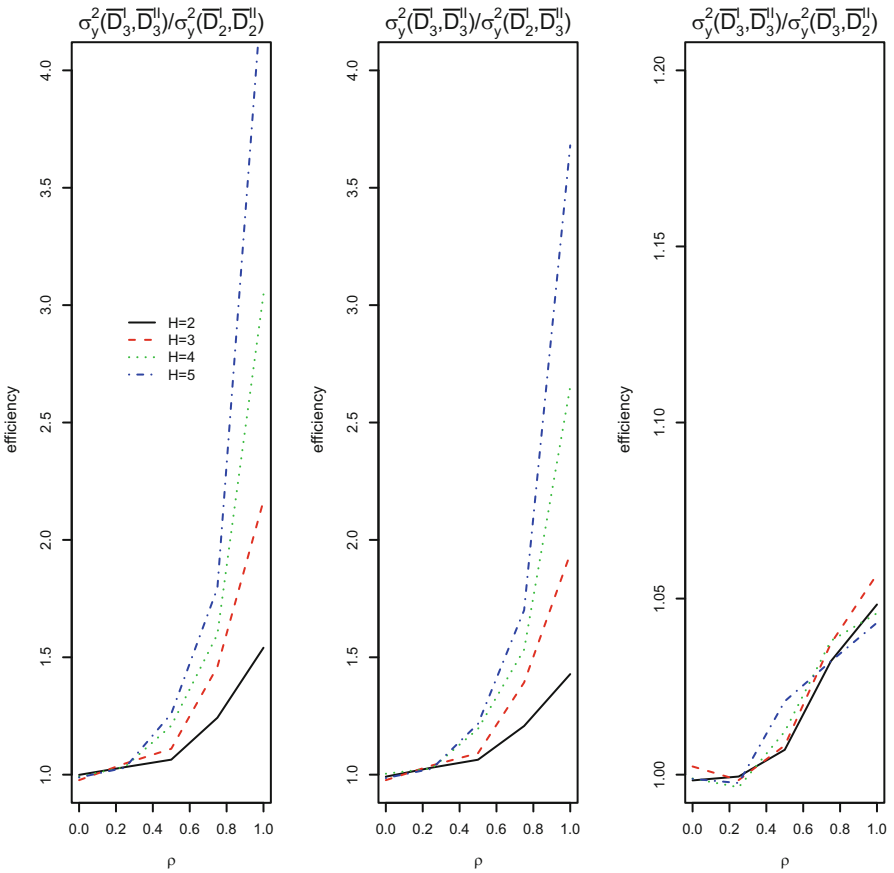

Fig. 3 Efficiency of two-stage cluster mean estimators for different values of set sizes ($H = H_1 = H_2$) and quality of ranking information ($\rho = \rho^I = \rho^{II}$). The efficiencies are defined as the ratio of variances at the top of each panel, ICC = 0.5, and $d_1 = d_2 = 4$

9 Examples

In this section, we apply the proposed estimators to two different populations; Ohio farm population and California school district population. In these examples, the cluster sizes, $M_i; i = 1, \ldots, N$ are available and differ significantly. Hence, we replace SRS (D_3^I) design with PPS (D_4^I) design.

United States Department of Agriculture's (USDA) National Quarterly Agricultural Survey We use 1992 Ohio Corn production data provided by the Ohio Agricultural Statistics Department in its county estimation program. This dataset includes responses from farms in the USDA National Quarterly Agricultural Survey and from farms responding to the Ohio supplemental survey, Husby et al. (2005). Further information on USDA survey can be found in Iwig (1993). The population contains 6346 farms in 88 counties. We considered these 88 counties as stage I population and the farms within each county as stage II population. On each farm, there were five variables:

Table 1 Coverage probabilities of the confidence intervals $y_{\bar{D}_k^I,\bar{D}_q^{II}} \pm t_{df,0.975}\hat{\sigma}_y(\bar{D}_k^I, \bar{D}_q^{II})$, based on clustered discrete normal population, $H_1 = H_2 = H, d_1 = 4, d_2 = 3, N = M = 120$, and simulation size is 20,000

			ICC = 0.10					ICC = 0.5				
ρ^I	ρ^{II}	H	$y_{\bar{D}_2^I,\bar{D}_2^{II}}$	$y_{\bar{D}_2^I,\bar{D}_3^{II}}$	$y_{\bar{D}_3^I,\bar{D}_2^{II}}$	$y_{\bar{D}_3^I,\bar{D}_3^{II}}$	SRS	$y_{\bar{D}_2^I,\bar{D}_2^{II}}$	$y_{\bar{D}_2^I,\bar{D}_3^{II}}$	$y_{\bar{D}_3^I,\bar{D}_2^{II}}$	$y_{\bar{D}_3^I,\bar{D}_3^{II}}$	SRS
0.00	0.00	2	0.951	0.950	0.955	0.960	0.948	0.930	0.930	0.937	0.900	0.950
0.00	0.00	3	0.945	0.900	0.948	0.950	0.948	0.940	0.936	0.940	0.946	0.948
0.00	0.00	4	0.940	0.944	0.946	0.900	0.948	0.940	0.937	0.950	0.948	0.950
0.00	0.00	5	0.940	0.942	0.950	0.946	0.952	0.900	0.941	0.950	0.947	0.950
0.25	0.20	2	0.950	0.952	0.960	0.955	0.950	0.933	0.932	0.900	0.938	0.950
0.25	0.25	3	0.900	0.944	0.950	0.951	0.950	0.939	0.940	0.943	0.942	0.900
0.25	0.25	4	0.943	0.944	0.900	0.946	0.950	0.939	0.940	0.944	0.940	0.950
0.25	0.25	5	0.938	0.940	0.948	0.949	0.900	0.941	0.940	0.946	0.950	0.950
0.50	0.50	2	0.950	0.950	0.955	0.950	0.949	0.931	0.900	0.939	0.940	0.950
0.50	0.50	3	0.942	0.940	0.949	0.950	0.950	0.930	0.934	0.943	0.900	0.949
0.50	0.50	4	0.940	0.900	0.942	0.940	0.948	0.940	0.937	0.950	0.948	0.951
0.50	0.50	5	0.940	0.942	0.945	0.900	0.950	0.940	0.939	0.950	0.946	0.950
0.75	0.75	2	0.950	0.952	0.950	0.955	0.951	0.900	0.929	0.940	0.936	0.950
0.75	0.80	3	0.940	0.943	0.950	0.950	0.950	0.932	0.930	0.900	0.941	0.950
0.75	0.75	4	0.900	0.944	0.940	0.948	0.950	0.934	0.930	0.945	0.944	0.900
0.75	0.75	5	0.938	0.944	0.900	0.947	0.950	0.934	0.930	0.950	0.950	0.951
1.00	1.00	2	0.947	0.950	0.951	0.953	0.900	0.927	0.930	0.940	0.940	0.950
1.00	1.00	3	0.938	0.950	0.945	0.950	0.951	0.922	0.900	0.943	0.940	0.949
1.00	1.00	4	0.934	0.940	0.940	0.950	0.950	0.920	0.921	0.948	0.900	0.949
1.00	1.00	5	0.917	0.900	0.945	0.950	0.949	0.910	0.912	0.950	0.947	0.950

corn production (bushels, Y), farm size (acreage, Z_1), group size (Z_2), acre planted (Z_3), and acre harvested (Z_4). Our interest lies in estimation of mean (or total) corn production in Ohio using these 6346 farms as a hierarchical population. We use designs D_1, D_2 and D_4 in stage I, and designs D_1, D_2 and D_3 in stage II sampling. Design D_2^{II} with set size $H_2 = 3$ and cycle size $d_2 = 2$ requires population size to be at least 18. Since the number of farms in some counties was very small, we removed all counties having less than 20 farms from the population. The size of stage I population is then reduced from $N = 88$ to $N = 73$ counties. The number of farms in these 73 counties, $M_i; i = 1, \ldots, N$, varied between 21 and 248. The correlation coefficient between the number of farms in counties and county totals of corn production was $\rho^I = 0.741$, where $\rho^I = \text{cor}(\boldsymbol{M}, \boldsymbol{Y})$ and $\boldsymbol{M}$, $\boldsymbol{Y}$ are the vectors of number of farms and total corn production in counties, respectively. Since ρ^I is relatively large, we use M_is to rank the counties for their total corn production in stage I sampling. In stage II sampling, we used auxiliary variables $Z_r; r = 1, 2, 3, 4$, to rank the farms for their corn production. The correlation coefficient between Y and Z_r varies from county to county. Average correlation coefficients between the corn production and auxiliary variables Z_r on farms over all counties are $\bar{\rho}_1^{II} = 0.854$, $\bar{\rho}_2^{II} = 0.619$, $\bar{\rho}_3^{II} = 0.975$, $\bar{\rho}_4^{II} = 0.982$, where $\bar{\rho}_r^{II} = \sum_{i=1}^{N} \text{cor}(\boldsymbol{Y}_i, \boldsymbol{Z}_{r,i})/N$, $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_{r,i}$ are vectors that contain corn production and values of auxiliary variable Z_r in county i. Since farm sizes are not equal in counties, we use adjusted R^2 to measure the homogeneity within-county farms. For 73 counties, adjusted R^2 was 4.5% which indicates that within-county variation is nearly equal to between county variation. We expect that using RSS design in stage II sampling would have big impact on the efficiency of the two-stage RSS estimator.

We performed a simulation study to investigate the efficiency of the proposed estimators $y_{\bar{D}_k^I, \bar{D}_q^{II}}; k = 1, 2, 4, q = 1, 2, 3$ based on nine different designs. Datasets are generated using two-stage sampling from the farm population in 73 counties in 1992 Ohio USDA dataset. Set and cycle sizes are taken to be $H_1 = 2, 3, 4$, $d_1 = 3$ for stage I sampling, and $H_2 = 3$, $d_2 = 2$ for stage II sampling. For design D_4^I and D_3^{II} sample sizes are matched with $n_1 = H_1 d_1$ and $n_2 = H_2 d_2$. Simulation size is taken to be 100,000.

Figure 4 plots the standard deviation of the estimator $y_{\bar{D}_k^I, \bar{D}_q^{II}}$ $(\sigma_y(\bar{D}_k^I, \bar{D}_q^{II}))$ with respect the auxiliary variables Z_r, $r = 1, 2, 3, 4$. The estimators with designs (D_k^I, D_q^{II}) are identified in the legends of the graph. We note that scales of y-axes are different in three panels. The panels 1, 2, and 3 are for the estimators with set sizes $H_1 = 2, 3, 4$ in stage I sampling, respectively. It is clear from these panels that $\sigma_y(\bar{D}_k^I, \bar{D}_q^{II})$ are decreasing functions of set size H_1 as expected. Within each panel, we see the impact of the quality of ranking information in stage II sampling. The ranking quality of auxiliary variables Z_3 ($\bar{\rho}_3 = 0.975$) and Z_4($\bar{\rho}_4 = 0.984$) are about the same. Hence, they yield roughly the same efficiency. The ranking quality of auxiliary variable Z_2($\bar{\rho}_2 = 0.619$) is not as good as the ranking qualities of the other three auxiliary variables. The estimators constructed based on auxiliary variable Z_2 yield the largest standard deviation among the four auxiliary variables. The efficiencies of the estimators based on auxiliary variable Z_1($\bar{\rho}_1 = 0.854$) are in between the efficiencies of estimators based on auxiliary variables Z_2 and Z_3, Z_4. All three panels in Fig. 4

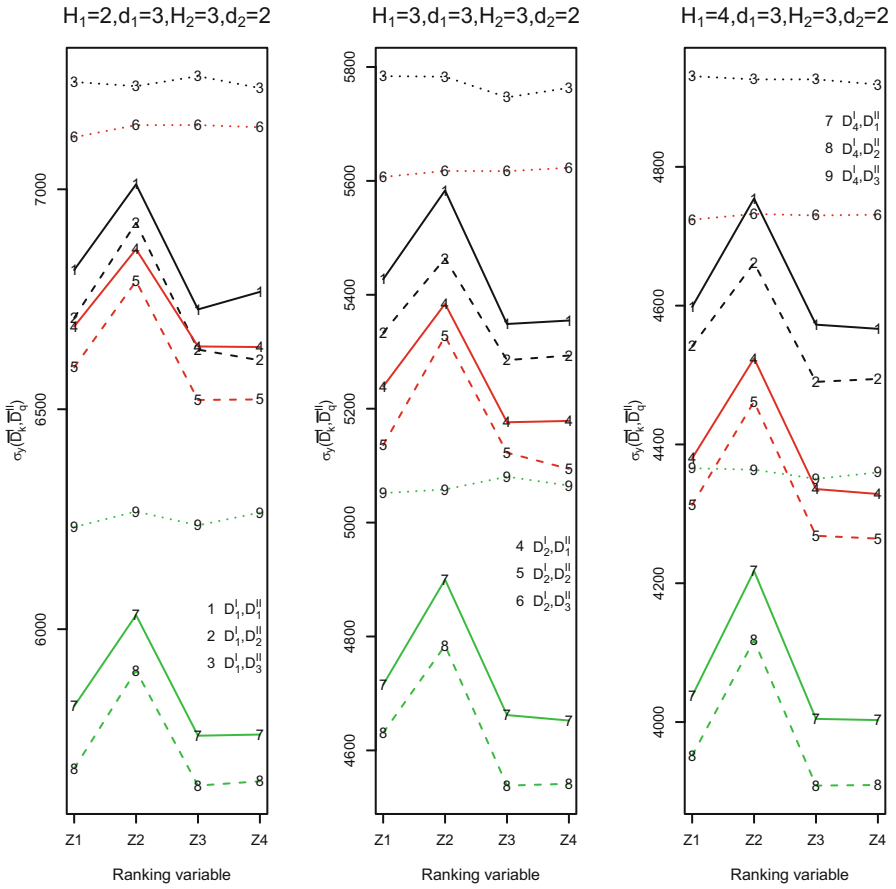

Fig. 4 Standard deviation of the estimators $y_{\bar{D}_k^I,\bar{D}_q^{II}}$ $(\sigma_y(\bar{D}_k^I, \bar{D}_q^{II}))$ for corn production in 1992 USDA Ohio survey data. The designs D_k^I, D_q^{II} are identified on the legends of the graph, $\rho^I = 0.741$, and adjusted $R^2 = 4.5\%$

indicate that PPS (design D_4^I) sampling and RSS sampling (D_2^{II}) are optimal in stages I and II , respectively. These findings are consistent with the surface plot in panel (a) in Fig. 1, where ICC $= 0.05$ and the efficiency increases with the quality of ranking information in stage II sampling.

The coverage probabilities of the confidence intervals in equation (3) are presented in Table 2. Note that in the simulation study the sample sizes are small and underlying population is strongly skewed right. Hence, the coverage probabilities are usually smaller than the nominal coverage probability 0.95. On the other hand, they are closer to the nominal value for large sample sizes. For example, the coverage probabilities are 0.88 when $H_1 = 2, d_1 = 3$ and increased to a value around 0.92 when $H_1 = 4, d_1 = 3$.

The California Department of Education (CDE) Academic Performance Index (API) Data The California Department of Education (CDE) calculates the Academic Performance Index (API) of schools in school districts. The API is a single number on

Table 2 Coverage probabilities of the confidence intervals, $y_{\bar{D}_k^I,\bar{D}_q^{II}} \pm t_{df,0.975}\hat{\sigma}_y(\bar{D}_k^I, \bar{D}_q^{II})$, for population mean of 1992 USDA Ohio corn production data, $H_2 = 3$, $d_2 = 2$

Est.	$H_1 = 2, d_1 = 3$				$H_1 = 3, d_1 = 3$				$H_1 = 4, d_1 = 3$			
	Z_1	Z_2	Z_3	Z_4	Z_1	Z_2	Z_3	Z_4	Z_1	Z_2	Z_3	Z_4
$y_{\bar{D}_1^I,\bar{D}_1^{II}}$	0.887	0.879	0.888	0.884	0.910	0.900	0.905	0.908	0.918	0.919	0.920	0.919
$y_{\bar{D}_1^I,\bar{D}_2^{II}}$	0.887	0.880	0.883	0.887	0.901	0.899	0.904	0.904	0.913	0.909	0.917	0.916
$y_{\bar{D}_1^I,\bar{D}_3^{II}}$	0.877	0.877	0.874	0.879	0.898	0.897	0.899	0.894	0.909	0.911	0.909	0.911
$y_{\bar{D}_2^I,\bar{D}_1^{II}}$	0.886	0.877	0.887	0.887	0.908	0.901	0.904	0.907	0.919	0.916	0.919	0.917
$y_{\bar{D}_2^I,\bar{D}_2^{II}}$	0.883	0.878	0.888	0.886	0.903	0.902	0.907	0.907	0.920	0.918	0.919	0.918
$y_{\bar{D}_2^I,\bar{D}_3^{II}}$	0.878	0.871	0.878	0.876	0.903	0.903	0.904	0.904	0.918	0.919	0.919	0.918
$y_{\bar{D}_3^I,\bar{D}_1^{II}}$	0.902	0.890	0.901	0.901	0.920	0.915	0.918	0.916	0.927	0.922	0.928	0.927
$y_{\bar{D}_3^I,\bar{D}_2^{II}}$	0.900	0.893	0.900	0.899	0.920	0.912	0.917	0.920	0.927	0.925	0.928	0.929
$y_{\bar{D}_3^I,\bar{D}_3^{II}}$	0.890	0.888	0.888	0.887	0.912	0.910	0.909	0.909	0.925	0.921	0.926	0.919
srs	0.878	0.881	0.883	0.879	0.896	0.895	0.896	0.896	0.907	0.906	0.906	0.907

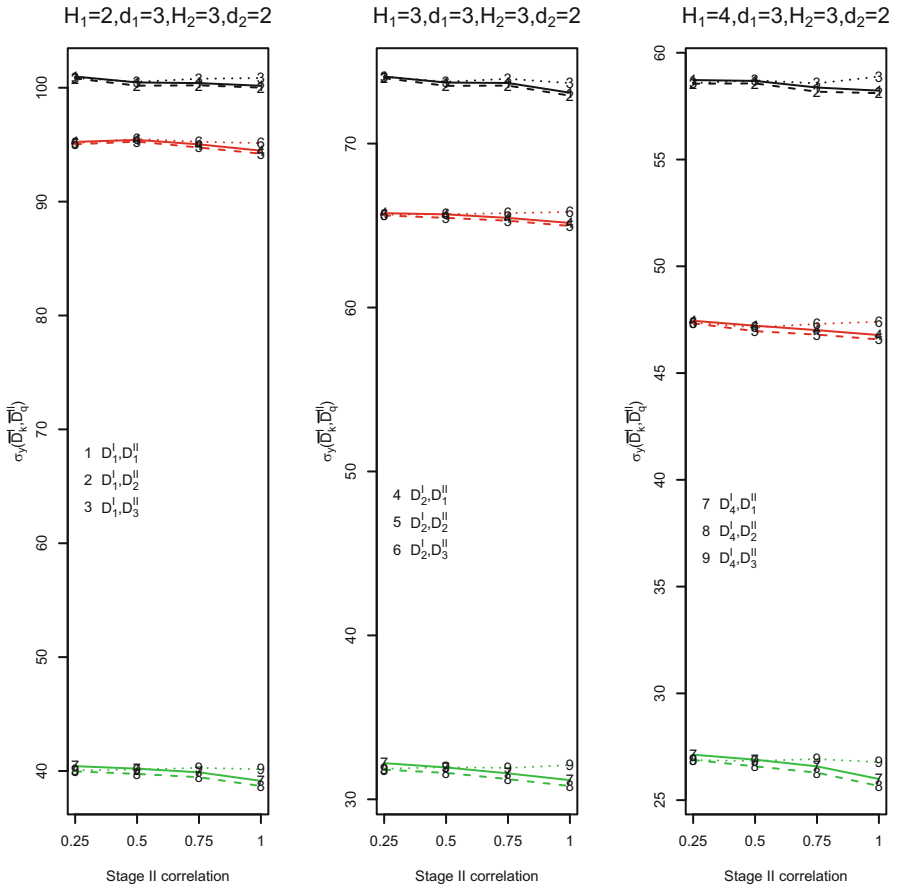

Fig. 5 Standard deviation of the estimators $y_{\bar{D}_k^I,\bar{D}_q^{II}}$ $(\sigma_y(\bar{D}_k^I, \bar{D}_q^{II}))$ for California School District study. The designs D_k^I, D_q^{II} are identified on the legends of the graph, $\rho^I = 0.995$, and adjusted $R^2 = 43.69\%$

a scale of 200–1000 that indicates how well students in a school or district performed on the test given in spring of previous year. It is calculated using results of the STAR (Standardized Testing and Reporting) program and the California High School Exit Exam (CAHSEE). The CDE disseminates the results directly to schools and districts as well as posting them on their website (http://api.cde.ca.gov).

We consider another example for two-stage cluster sampling using 2000 API scores in California school district study. We use publicly available API dataset in the *R* package `survey`. The data set shows a hierarchical structure where schools are nested within school districts and contains information for all schools with at least 100 students. We again removed all school districts having less than 21 and greater than 99 schools. The remaining $N = 59$ school districts constitute the stage I population. The number of schools in most of the remaining school districts was between 20 and 50. Unlike USDA Ohio farm population, API data had relatively large adjusted $R^2 = 50.12\%$ indicating that within district schools are more homogeneous

than between school districts. The correlation coefficient between the district size $(M_i; i = 1, \ldots, 59)$ and total API scores in districts was $\rho^I = 0.940$. Thus we use the number of schools in districts to rank the total of API scores in stage I sampling. Using API dataset, we performed another simulation study with $H_1 = 2, 3, 4, d_1 = 3$ in stage I sampling and $H_2 = 3$ and $d_2 = 2$ in stage II sampling. Simulation study is performed using designs D_1^I, D_2^I, D_4^I in stage I and designs D_1^{II}, D_2^{II}, D_3^{II} in stage II. Ranking in stage II sampling is performed based on Dell and Clutter model with $\rho^{II} = 0.25, 0.50, 0.75, 1.00$. Simulation size is taken to be 100,000.

Figure 5 presents the standard deviation of the estimators $y_{\bar{D}_k^I, \bar{D}_q^{II}}$ with respect to stage II correlation coefficient ρ_r^{II}, $r = 1, \ldots, 4$. The design choices of each line is given in the legends of Fig. 5. It is clear that the estimators using design D_4^I in stage I sampling (lines 8, 7 and 9) are significantly better than the other two sets of estimators using design D_1^I and D_2^I (lines 1, 2, 3, 4, 5, 6). Even though designs D_1^I, D_2^I use ranked set sampling procedure, its efficiency is not significantly better than PPS design (D_4^I).

The quality of ranking information in stage II sampling does not have a significant impact on efficiency since lines (7, 8, 9) in each panel are roughly parallel to horizontal axes. They do not change significantly with ρ_r^{II}. This is also consistent with the main features in panel (b) of surface plot 1, where ICC $= 0.5$ and efficiency is roughly constant with respect to the quality of ranking information in stage II sampling.

10 Concluding remarks

Cluster sampling is used commonly in large surveys when the population has a nested structure. The estimates obtained from cluster samples usually have larger variance than the estimate obtained from the same number of observation using an SRS. Since the population units are naturally organized in hierarchical structure in cluster sampling, it is less expensive to sample from clusters. Hence, it can provide more precision per dollar spent. This observational economy is further improved using RSS designs in a two-stage cluster sampling. In a finite population framework, RSS samples can be constructed with or without replacement. This paper establishes important connections between efficiency and design parameters, such as ICC, set sizes, quality of ranking information in RSS designs. These connections provide useful guidelines for data analyst to design a two-stage cluster sample using RSS designs. We show that if the cluster population sizes are available and not all equal, the PPS (D_4) and RSS (D_2) designs are optimal in stage I and stage II sampling, respectively. On the other hand, if the cluster population sizes are not available and/or are all equal, the design D_2 in stage I (stage II) provides significant improvement in efficiency when ICC is large (small). The paper also provides estimators and confidence intervals for population mean and total. Estimators are design unbiased and do not require distributional assumptions on the underlying population.

References

Al-Saleh, M. F., Samawi, H. M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, *16*, 198–209.

Dell, T. R., Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, *28*, 545–555.
Deshpande, J. V., Frey, J., Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, *13*, 25–40.
Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference*, *141*, 3632–3639.
Gokpinar, F., Ozdemir, Y. A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, *39*, 89–95.
Hollander, M., Wolfe, D. A., Chicken, E. (2014). *Nonparametric statistical methods* (3rd ed.). New York: Wiley.
Husby, C. E., Stasny, E. A., Wolfe, D. A. (2005). An application of ranked set sampling for mean and median estimation using USDA crop production data. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*, 354–373.
Iwig, W. C. (1993). The National Agricultural Statistics Service County Estimates Program. In *Indirect Estimators in Federal Programs, Statistical Policy Working Paper 21, Report of the Federal Committee on Statistical Methodology, Subcommittee on Small Area Estimation, Washington, DC* (pp. 7.1–7.15).
Jafari Jozani, M., Johnson, B. C. (2011). Design based estimation for ranked set sampling in finite population. *Environmental and Ecological Statistics*, *18*, 663–685.
Jafari Jozani, M., Johnson, B. C. (2012). Randomized nomination sampling in finite populations. *Journal of Statistical Planning and Inference*, *142*, 2103–2115.
Lohr, S. (1999). *Sampling: Design and analysis*. New York: Duxbury Press.
McIntyre, G. A. (2005). A method of unbiased selective sampling using ranked-sets. *The American Statistician*, *59*, 230–232.
Muttlak, H. A., McDonald, L. L. (1992). Ranked set sampling and the line intercept method: A more efficient procedure. *Biometrical Journal*, *34*, 329–346.
Nematollahi, N., Salehi, M. M., Aliakbari Saba, R. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics—Theory and Methods*, *37*, 2402–2415.
Ozdemir, Y. A., Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, *36*, 89–99.
Ozdemir, Y. A., Gokpinar, F. (2008). A new formula for inclusion probabilities in median ranked set sampling. *Communications in Statistics—Theory and Methods*, *37*, 2022–2033.
Ozturk, O. (2014). Estimation of population mean and total in finite population setting using multiple auxiliary variables. *Journal of Agricultural, Biological and Environmental Statistics*, *19*, 161–184.
Ozturk, O. (2016a). Estimation of a finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological and Environmental Statistics*, *21*, 181–202.
Ozturk, O. (2016b). Statistical inference based on judgment post-stratified samples in finite population. *Survey Methodology*, *42*, 239–262.
Ozturk, O., Jafari Jozani, M. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data Analysis*, *69*, 122–132.
Patil, G. P., Sinha, A. K., Taillie, C. (1995). Finite population correction for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, *47*, 621–636.
Sroka, C. J. (2008). Extending ranked set sampling to survey methodology. Ph.D. thesis, Department of Statistics, Ohio State University.
Sud, V., Mishra, D. C. (2006). Estimation of finite population mean using ranked set two stage sampling designs. *Journal of the Indian Society of Agricultural Statistics*, *60*, 108–117.
Takahasi, K., Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, *20*, 1–31.
Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.
Wang, X., Lim, J., Stokes, L. (2016). Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. *Journal of the American Statistical Association*, *516*, 1576–1590.
Wolfe, D. A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *ISRN Probability and Statistics*,. doi:10.5402/2012/568385.