Optimal cluster selection probabilities to estimate the finite population distribution function under PPS cluster sampling

José A. Mayor Gallego *

Abstract

We study the estimation of the finite population distribution function under several sampling strategies based on a PPS cluster sampling, that is to say, with cluster selection probabilities proportional to size.

For the estimation of population means and totals, is well-known that this type of strategies give good results if the cluster selection probabilities are proportional to the total of the study variable or to a related auxiliary variable, over the cluster. We prove that, for the estimation of the distribution function using cluster sampling, this solution in general is not good and, under an appropriate criteria, we obtain the optimal cluster selection probabilities in order to minimize the variance of the estimation.

We apply our methodology for two classical PPS sampling strategies: the sampling with replacement with the Hansen-Hurwitz estimator, and the random groups sampling procedure with the Rao-Hartley-Cochran estimator. We will present a small simulation to compare the efficiency of this approach with other methods.

Key words: finite population distribution function, selection probabilities proportional to size, optimal cluster selection probabilities, random groups.

AMS classification: 62D05

*Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla. Facultad de Matemáticas. Tarfia s/n, 41012 Sevilla, España. e-mail:mayor@cica.es

1 Introduction

Usually, the theory of sampling from finite population is centered on the point estimation of some parameters as the finite population means, variances and ratios. In this paper, we consider the estimation of a functional parameter, the distribution function in relation to a numerical variable, defined over the population.

In literature we can find different approaches to this estimation problem. Chambers and Dunstan (1986) assume a model-based approach to develop an estimating procedure. Kuk (1988) studies several estimators of the distribution function under sampling with unequal probabilities, proportional to an auxiliary variable and Rao, Kovar and Mantel (1990) by means of the auxiliary information. In the same lines, we also can cite the papers of Chambers et al. (1992), Rao (1994), and Welsh and Ronchetti (1998).

We propose an alternative approach based on the application of an average-type criterion to the mean square error of the distribution function estimation, in order to find the more appropriate selection probabilities of the clusters.

In a general frame, let us consider a finite population $U = \{1, 2 \dots N\}$ and let Y denote the numerical survey variable of interest. Let Y_i be the value of Y for the ith population element, with ordered values $0 \leq Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}$. The aim is to estimate the distribution function of the Y variable,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i, +\infty)}(t)$$

where $I_{[Y_i,+\infty)}(t),\ i \in U$ are the indicator functions of the $[Y_i, +\infty)$ intervals.

If we assume that s is a sample obtained from U whit a sampling design $(S, p(\cdot))$, and $\widehat{F}(t)$ is an estimator of $F(t)$, the classical way to measure the precision of this estimator is to study of the mean square error,

$$\mathrm{MSE}[\widehat{F}(t)] = \sum_{s \in S} (\widehat{F}(t) - F(t))^2 \, p(s)$$

Note that the MSE is a real function with different values depending on t, therefore it is not possible to use this function for a direct comparison. An alternative way to evaluate the discrepancy between $F(t)$ and $\widehat{F}(t)$ is

to consider the quantity,

$$I(s) = \int_{Y_{(1)}}^{Y_{(N)}} (\widehat{F}(t) - F(t))^2 dt$$

assuming that $(\widehat{F}(t) - F(t))^2$ is integrable over $[Y_{(1)}, Y_{(N)}]$. Thus, we can compute the expected discrepancy,

$$\begin{aligned} E[I(s)] &= \sum_{s \in S} I(s)p(s) = \sum_{s \in S} \left(\int_{Y_{(1)}}^{Y_{(N)}} (\widehat{F}(t) - F(t))^2 dt \right) p(s) \\ &= \int_{Y_{(1)}}^{Y_{(N)}} \left(\sum_{s \in S} (\widehat{F}(t) - F(t))^2 \, p(s) \right) \, dt = \int_{Y_{(1)}}^{Y_{(N)}} \mathrm{MSE}[\widehat{F}(t)] \, dt \\ &\triangleq \| \mathrm{MSE}[\widehat{F}(t)] \|_1 \end{aligned}$$

As such, we can search the more appropriate sampling designs minimizing $\| \mathrm{MSE}[\widehat{F}(t)] \|_1$. Next, we particularize this approach for two sampling strategies under two-stage cluster sampling

2 Cluster sampling strategy based on the Hansen-Hurwitz estimator

Let us suppose that the population is clustered and let us denote $C_1, \ldots, C_M$ as the clusters, and $U_c = \{1, 2, \ldots, M\}$ as the cluster population. Let us suppose that n clusters are drawn with replacement and selecting probabilities p_i, $i \in U_c$, and s_c is the clusters sample. Subsampling is done independently each time a cluster is selected. Then, we can use the following estimator for the distribution function,

$$\widehat{F}_{\mathrm{HHC}}(t) = \frac{1}{n} \sum_{i \in s_c} \frac{\widehat{F}_i(t)}{p_i}$$

where,

$$F_i(t) = \frac{1}{N} \sum_{k \in C_i} I_{[Y_k, +\infty)}(t)$$

and $\widehat{F}_i(t)$ denotes an estimation of $F_i(t)$.

This estimator is an adaptation to the distribution function and to cluster sampling of the classical Hansen-Hurwitz estimator for population means and totals, Hansen and Hurwitz (1943). See Särndal et al. (1992, p.151). If $\widehat{F}_i(t),\ i \in U_c$ are unbiased estimators then $\widehat{F}_{\mathrm{HHC}}(t)$ is unbiased, and its variance is,

$$\begin{aligned} V[\widehat{F}_{\mathrm{HHC}}(t)] &= \frac{1}{n}\left[\sum_{i\in U_c}\frac{F_i^2(t)}{p_i} - F^2(t)\right] + \frac{1}{n}\sum_{i\in U_c}\frac{V[\widehat{F}_i(t)]}{p_i} \\ &= \frac{1}{n}\sum_{i\in U_c}\frac{F_i^2(t)+V[\widehat{F}_i(t)]}{p_i} - \frac{1}{n}F^2(t) \end{aligned}$$

In order to apply the minimizing criterion formulated in the above section, we compute,

$$\begin{aligned} \|\,\mathrm{MSE}[\widehat{F}_{\mathrm{HHC}}(t)]\|_1 &= \int_{Y_{(1)}}^{Y_{(N)}} V[\widehat{F}_{\mathrm{HHC}}(t)]\,dt \\ &= \int_{Y_{(1)}}^{Y_{(N)}}\left[\frac{1}{n}\sum_{i\in U_c}\frac{F_i^2(t)+V[\widehat{F}_i(t)]}{p_i} - \frac{1}{n}F^2(t)\right]dt \\ &= \frac{1}{n}\sum_{i\in U_c}\frac{1}{p_i}\int_{Y_{(1)}}^{Y_{(N)}}\left[F_i^2(t)+V[\widehat{F}_i(t)]\right]dt \\ &\quad -\frac{1}{n}\int_{Y_{(1)}}^{Y_{(N)}}F^2(t)\,dt = \frac{1}{n}\sum_{i\in U_c}\frac{A_i}{p_i} - C(Y) \end{aligned}$$

where,

$$A_i = \int_{Y_{(1)}}^{Y_{(N)}}\left[F_i^2(t)+V[\widehat{F}_i(t)]\right]dt \qquad \forall i \in U_c$$

and $C(Y)$ does not depend on the probabilities p_i.

The problem is then formulated as the determination of the p_i minimizing the above expression. Therefore, the optimization problem is,

$$\min_{p_1,\ldots,p_M}\sum_{i\in U_c}\frac{A_i}{p_i} \quad \text{subject to} \quad \sum_{i\in U_c}p_i = 1, \qquad p_i > 0\ \forall i \in U_c$$

To solve this problem, we apply the Cauchy-Schwartz inequality,

$$\begin{aligned}\sum_{i\in U_c}\frac{A_i}{p_i} &= \left(\sum_{i\in U_c}\frac{A_i}{p_i}\right)\left(\sum_{i\in U_c}p_i\right)\\ &\geq \left(\sum_{i\in U_c}\left(\frac{A_i}{p_i}\right)^{1/2}p_i^{1/2}\right)^2 = \left(\sum_{i\in U_c}A_i^{1/2}\right)^2\end{aligned}$$

and the equality holds if and only if,

$$\left[\frac{A_i/p_i}{p_i}\right]^{1/2} = \text{constant} \qquad \forall i\in U_c$$

therefore, the optimal cluster selection probabilities are,

$$p_i \propto \sqrt{A_i} \qquad \forall i\in U_c$$

and it follows that,

$$p_i = \frac{\sqrt{A_i}}{\sum_{i\in U_c}\sqrt{A_i}} \qquad \forall i\in U_c$$

Next, we study the A_i quantities. Let us suppose that subsampling is done in the cluster C_i and a sample s_i is obtained by means of a sampling design d_i, without replacement and with associated inclusion probabilities $\pi^i_{kl},\ k,l\in C_i$, with $i\in U_c$. To estimate $F_i(t)$, we use the Horvitz-Thompson estimator,

$$\widehat{F}_i(t) = \frac{1}{N}\sum_{k\in s_i}\frac{I_{[Y_k,+\infty)}(t)}{\pi^i_k}$$

This estimator is unbiased, therefore,

$$A_i = \int_{Y_{(1)}}^{Y_{(N)}}\left[F_i^2(t)+V[\widehat{F}_i(t)]\right]dt = \int_{Y_{(1)}}^{Y_{(N)}}E\left[\widehat{F}_i^2(t)\right]dt \qquad \forall i\in U_c$$

and using that,

$$E\left[\widehat{F}_i^2(t)\right] = \frac{1}{N^2}\sum_{k,l\in C_i}\frac{\pi^i_{kl}}{\pi^i_k\pi^i_l}\,I_{[Y_k,+\infty)}(t)I_{[Y_l,+\infty)}(t)$$

we get,

$$A_i = \int_{Y_{(1)}}^{Y_{(N)}} E\left[\widehat{F}_i^2(t)\right] dt = \frac{1}{N^2} \int_{Y_{(1)}}^{Y_{(N)}} \sum_{k,l \in C_i} \frac{\pi_{kl}^i}{\pi_k^i \pi_l^i} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\, dt$$

$$= \frac{1}{N^2} \sum_{k,l \in C_i} \frac{\pi_{kl}^i}{\pi_k^i \pi_l^i} \left(Y_{(N)} - \max\{Y_k, Y_l\}\right)$$

In practice the Y variable is not available but a suitable way to compute the optimal p_i is based on the existence of an auxiliary variable, X, entirely controlled and related to the Y variable by means of the following general super-population model,

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \beta > 0, \quad E_s[\varepsilon_i] = 0 \qquad \forall i \in U$$

Thus, we replace the above minimization problem by,

$$\min_{p_1,\ldots,p_M} E_s\left[\sum_{i \in U_c} \frac{A_i}{p_i}\right] \quad \text{subject to} \quad \sum_{i \in U_c} p_i = 1, \qquad p_i > 0 \ \forall i \in U_c$$

and using the inequality $E_s[\max\{Y_k, Y_l\}] \geq \max\{E_s[Y_k], E_s[Y_l]\}$, we have,

$$E_s\left[\sum_{i \in U_c} \frac{A_i}{p_i}\right] = \frac{1}{N^2} \sum_{i \in U_c} \frac{1}{p_i} \sum_{k,l \in C_i} \frac{\pi_{kl}^i}{\pi_k^i \pi_l^i} E_s\left[Y_{(N)} - \max\{Y_k, Y_l\}\right]$$

$$\leq \frac{\beta}{N^2} \sum_{i \in U_c} \frac{1}{p_i} \sum_{k,l \in C_i} \frac{\pi_{kl}^i}{\pi_k^i \pi_l^i} \left(X_{(N)} - \max\{X_k, X_l\}\right)$$

$$= \frac{\beta}{N^2} \sum_{i \in U_c} \frac{B_i}{p_i}$$

where $X_{(N)}$ is the maximum value of the X variable and,

$$B_i = \sum_{k,l \in C_i} \frac{\pi_{kl}^i}{\pi_k^i \pi_l^i} \left(X_{(N)} - \max\{X_k, X_l\}\right) \qquad \forall i \in U_c$$

therefore we obtain the auxiliary optimization problem,

$$\min_{p_1,\ldots,p_M} \sum_{i \in U_c} \frac{B_i}{p_i} \quad \text{subject to} \quad \sum_{i \in U_c} p_i = 1, \qquad p_i > 0 \ \forall i \in U_c$$

and it now follows that the optimal probabilities are given by,

$$p_i = \frac{\sqrt{B_i}}{\sum_{i \in U_c} \sqrt{B_i}} \qquad \forall i \in U_c$$

The final values of these selection probabilities depend on the π^i_{kl}, that is to say, on the sampling designs $d_i,\ i \in U_c$. Next, we study a particular and usual case based on the application of simple random sampling in each cluster.

2.1 Simple random subsampling

Let suppose that in each cluster C_i a sample s_i is obtained by means of simple random sampling of n_i units, then,

$$\pi^i_k = \frac{n_i}{N_i},\ \forall k \in C_i \qquad \pi^i_{kl} = \frac{n_i(n_i-1)}{N_i(N_i-1)},\ \forall k,l \in C_i,\ k \neq l$$

and we obtain,

$$\begin{aligned} B_i &= \sum_{k,l \in C_i} \frac{\pi^i_{kl}}{\pi^i_k \pi^i_l} \left(X_{(N)} - \max\{X_k, X_l\} \right) = \sum_{k \in C_i} \frac{N_i}{n_i} \left(X_{(N)} - X_k \right) \\ &\quad + \sum_{\substack{k,l \in C_i \\ k \neq l}} \frac{N_i(n_i-1)}{n_i(N_i-1)} \left(X_{(N)} - \frac{1}{2}X_k - \frac{1}{2}X_l - \frac{1}{2}|X_k - X_l| \right) \\ &= N_i^2 \left(X_{(N)} - \overline{X}_i - \frac{n_i - 1}{2n_i N_i (N_i - 1)} \sum_{k,l \in C_i} |X_k - X_l| \right) \end{aligned}$$

where $\overline{X}_i$ is the mean of the auxiliary variable X over the cluster C_i. Therefore, the optimal probabilities are,

$$p_i \propto N_i \sqrt{X_{(N)} - \overline{X}_i - \frac{n_i - 1}{2n_i N_i (N_i - 1)} \sum_{k,l \in C_i} |X_k - X_l|}$$

Note that the quantity,

$$\frac{1}{N_i(N_i-1)}\sum_{k,l\in C_i}|X_k-X_l|$$

is a dispersion measure of the cluster C_i. Therefore we conclude that the optimal selection probabilities are directly proportional to cluster sizes, N_i, but if these sizes are similar, the probabilities are bigger for the clusters with minor total value of the variable, and/or less dispersed.

This result seems to be in contrast to the classical approach to reduce the sampling error estimating population totals and means, based on taking selection probabilities proportional to the total of the study variable or to the auxiliary variable on each cluster, in order to reduce the variance due to the first stage, but note that these problems have very different objectives.

On the other hand, note also that the estimator $\widehat{F}_{\mathrm{HHC}}(t)$ in general is not a distribution function. For example, under the simple random subsampling hypothesis we have,

$$\widehat{F}_i(t)=\frac{1}{N}\sum_{k\in s_i}\frac{I_{[Y_k,+\infty)}(t)}{n_i/N_i}$$

therefore,

$$\widehat{F}_{\mathrm{HHC}}(+\infty)=\frac{1}{nN}\sum_{i\in s_c}\frac{N_i}{p_i}\neq 1\quad\text{in general}$$

that is to say, the ultimate value is not necessarily unity, however this deviation is minimized under our approach, minimizing $\|\operatorname{MSE}[\widehat{F}_{\mathrm{HHC}}(t)]\|_1$.

It also is possible to avoid this drawback using alternative estimators, for example ratio type estimators. In general these estimators are biased and the study of the sampling error is more complex.

3 Strategy based on the Rao-Hartley-Cochran estimator

This strategy, suggested in J.N.K. Rao et al. (1962), is a grouping method based on to subdivide the U_c population with M clusters, at random into n sub-population or groups, $G_1, \ldots, G_n$, of pre-determined sizes $M_1, \ldots, M_n$, with $\sum_k M_k = M$, and to draw one cluster per group using PPS method of sampling, independently within each group.

Thus, the complete sampling process has two stages. In the first stage the clusters are drawn in two phases, that is to say, the random grouping phase and the random selecting phase. The second stage is the subsampling of the clusters.

According to the PPS method, given the random grouping, the chance of selecting the ith unit in the sample is,

$$P_i = \frac{p_i}{\sum\limits_{j \in G_k} p_j} \quad \text{if } i \in G_k, \qquad k = 1, \ldots, n, \quad i = 1, \ldots, M$$

where $p_i > 0,\ i \in U_c$, is a given probability distribution over U_c. Using the notation as the above section, we have,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i, +\infty)}(t) = \sum_{i \in U_c} \frac{1}{N} \sum_{k \in C_i} I_{[Y_k, +\infty)}(t) = \sum_{i \in U_c} F_i(t)$$

Then, if $\widehat{F}_i(t),\ i \in U_c$, are unbiased estimators, an unbiased estimator of the distribution function, based on the population total estimator suggested by Rao, Hartley and Cochran under this sampling scheme, is given by,

$$\widehat{F}_{\mathrm{RHCC}}(t) = \sum_{i \in s_c} \frac{\widehat{F}_i(t)}{P_i}$$

where s_c is the selected sample of clusters. To compute the variance of this estimator, we will denote by E_1 and V_1 the operators of expectation and variance with respect to the first stage sampling design, that is to say, the sampling of clusters, and by E_2 and V_2 the corresponding operators for the second stage sampling design, that is to say, the subsampling in the clusters. Thus,

$$V[\widehat{F}_{\mathrm{RHCC}}(t)] = V_1\, E_2 \left[\widehat{F}_{\mathrm{RHCC}}(t)\right] + E_1\, V_2 \left[\widehat{F}_{\mathrm{RHCC}}(t)\right]$$

and using the well-known expression for the variance of the Rao-Hartley-Cochran estimator, we obtain for the first term,

$$\begin{aligned} V_1\, E_2\left[\widehat{F}_{\mathrm{RHCC}}(t)\right] &= V_1\left[\sum_{i\in s_c} E_2\left[\frac{\widehat{F}_i(t)}{P_i}\right]\right] \\ &= V_1\left[\sum_{i\in s_c}\frac{E_2[\widehat{F}_i(t)]}{P_i}\right] = V_1\left[\sum_{i\in s_c}\frac{F_i(t)]}{P_i}\right] \\ &= \frac{\sum_{k=1}^{n} M_k^2 - M}{M^2 - M}\left[\sum_{i\in U_c}\frac{F_i^2(t)}{p_i} - F^2(t)\right] \end{aligned}$$

In order to develop the second term, let us denote by E_g the expectation with respect to the grouping phase, and for E_c the expectation with respect to the selection cluster phase.

For every cluster i belonging to the group G_k, let W_i^k stand for the indicator random variable defined as $W_i^k = 1$ if i is selected, $W_i^k = 0$, otherwise. Note that $E_c[W_i^k] = P_i$.

And for every pair of clusters $i, j \in U_c$, let δ_{ij}^k stand for the indicator random variable defined as $\delta_{ij}^k = 1$ if $i, j \in G_k$, $\delta_{ij}^k = 0$, otherwise. Note that the group G_k is a simple random sample, therefore we have $E_g[\delta_{ij}^k] = M_k/M$ if $i = j$, and $E_g[\delta_{ij}^k] = M_k(M_k - 1)/(M(M-1))$ if $i \neq j$.

Applying the above defined random variables we have,

$$E_1\, V_2\left[\widehat{F}_{\mathrm{RHCC}}(t)\right]$$

$$= E_1\left[\sum_{i\in s_c} V_2\left[\frac{\widehat{F}_i(t)}{P_i}\right]\right] = E_1\left[\sum_{i\in s_c}\frac{V_2[\widehat{F}_i(t)]}{P_i^2}\right] = E_g\, E_c\left[\sum_{i\in s_c}\frac{V_2[\widehat{F}_i(t)]}{P_i^2}\right]$$

$$= E_g\, E_c\left[\sum_{k=1}^{n}\sum_{i\in G_k}\frac{V_2[\widehat{F}_i(t)]}{P_i^2} W_i^k\right] = E_g\left[\sum_{k=1}^{n}\sum_{i\in G_k}\frac{V_2[\widehat{F}_i(t)]}{P_i^2} E_c[W_i^k]\right]$$

$$= E_g\left[\sum_{k=1}^{n}\sum_{i\in G_k}\frac{V_2[\widehat{F}_i(t)]}{P_i}\right] = E_g\left[\sum_{k=1}^{n}\sum_{i\in G_k}\frac{V_2[\widehat{F}_i(t)]}{p_i}\sum_{j\in G_k} p_j\right]$$

$$= E_g\left[\sum_{k=1}^{n}\sum_{i,j\in G_k} V_2[\widehat{F}_i(t)]\,\frac{p_j}{p_i}\right] = E_g\left[\sum_{k=1}^{n}\sum_{i,j\in U_c} V_2[\widehat{F}_i(t)]\,\frac{p_j}{p_i}\,\delta_{ij}^k\right]$$

$$= \sum_{k=1}^{n}\sum_{i\in U_c} V_2[\widehat{F}_i(t)]\,E_g[\delta_{ii}^k] + \sum_{k=1}^{n}\sum_{\substack{i,j\in U_c\\ i\neq j}} V_2[\widehat{F}_i(t)]\,\frac{p_j}{p_i}\,E_g[\delta_{ij}^k]$$

$$= \sum_{k=1}^{n}\sum_{i\in U_c} V_2[\widehat{F}_i(t)]\,\frac{M_k}{M} + \sum_{k=1}^{n}\sum_{\substack{i,j\in U_c\\ i\neq j}} V_2[\widehat{F}_i(t)]\,\frac{p_j}{p_i}\,\frac{M_k(M_k-1)}{M(M-1)}$$

$$= \sum_{i\in U_c} V_2[\widehat{F}_i(t)] + \sum_{k=1}^{n}\frac{M_k(M_k-1)}{M(M-1)}\sum_{i\in U_c} V_2[\widehat{F}_i(t)]\,\frac{1-p_i}{p_i}$$

$$= \frac{\sum_{k=1}^{n} M_k^2 - M}{M^2-M}\sum_{i\in U_c}\frac{V_2[\widehat{F}_i(t)]}{p_i} + \left[1-\frac{\sum_{k=1}^{n} M_k^2 - M}{M^2-M}\right]\sum_{i\in U_c} V_2[\widehat{F}_i(t)]$$

and putting together the components of the variance, we obtain,

$$V[\widehat{F}_{\mathrm{RHCC}}(t)] = \frac{\sum_{k=1}^{n} M_k^2 - M}{M^2-M}\left[\sum_{i\in U_c}\frac{F_i^2(t)}{p_i} - F^2(t)\right]$$

$$+\frac{\sum_{k=1}^{n} M_k^2 - M}{M^2-M}\sum_{i\in U_c}\frac{V_2[\widehat{F}_i(t)]}{p_i} + \left[1-\frac{\sum_{k=1}^{n} M_k^2 - M}{M^2-M}\right]\sum_{i\in U_c} V_2[\widehat{F}_i(t)]$$

If we compare this variance with $V[\widehat{F}_{\mathrm{HHC}}(t)]$, we see that they are very similar except for some terms not depending on the p_i probabilities. Therefore the optimal selection probabilities are the same, that is to say, in general,

$$p_i \propto \sqrt{\sum_{k,l\in C_i}\frac{\pi_{kl}^i}{\pi_k^i\pi_l^i}\left(X_{(N)} - \max\{X_k, X_l\}\right)} \qquad \forall i \in U_c$$

or,

$$p_i \propto N_i\sqrt{X_{(N)} - \overline{X}_i - \frac{n_i-1}{2n_iN_i(N_i-1)}\sum_{k,l\in C_i}|X_k - X_l|} \qquad \forall i \in U_c$$

if we use simple random sampling in each cluster.

Finally, in relation to the M_k values, it is well-know that the choice of $M_1, \ldots, M_n$ which would minimize the variance corresponds to,

$$M_1 = M_2 = \cdots, M_n = M/n \qquad \text{if } M \text{ is divisible by } n$$

and,

$$M_1 = M_2 = \cdots = M_\nu = m + 1,\ M_{\nu+1} = \cdots = M_n = m$$
$$\text{if } M = mn + \nu,\ 1 \leq \nu \leq n - 1$$

4 A Monte Carlo comparative study. Conclusions

In order to compare the performance of PPS sampling procedures described in the previous sections in relation to other sampling strategies, we have carried out a comparative study based on a simulation.

The study is done with the clustered population of Swedish municipalities given by Särndal et al. (1992, p. 660). This population, named Clustered MU284 population consisting of all municipalities and contains 50 cluster and 284 municipalities. The cluster sizes are between five and eight municipalities.

The study variable, named P85, is the 1985 population for each municipality, in thousand, and the auxiliary variable, named P75, is the 1975 population, in thousand. For these variables, the parameters of the model $Y_i = \alpha + \beta X_i + \varepsilon_i$, under the ordinary least square criterion, where R^2 denotes the coefficient of determination, are given in Table 1. and they show the validity of the model. It is interesting to note that the optimal cluster selection probabilities do not depend on these parameters, but the validity of the super-population model has influence on the goodness of the estimation and the variance. This situation is similar to the classical PPS methods to estimate means and totals.

α	β	R^2	Mean Residual
1.315	0.974	0.997	−2.71E−15

Table 1. Regression parameters of the P75 and P85 variables in the Clustered MU284 population, taking P85 as the dependent variable.

In order to compare the efficiency of the different strategies, we will use the following distance as a discrepancy measure,

$$d(\widehat{F}, F) = \left[\int_{Y_{(1)}}^{Y_{(N)}} \left(\widehat{F}(t) - F(t) \right)^2 \, dt \right]^{1/2}$$

Note that this distance is compatible with the criterion used to find the optimal selection probabilities. We will compute the average of this distance over 1000 random samples, with sizes $n = 4, 7, 10$ and 15 clusters, obtained in the first stage by different methods. These sample sizes represent a wide choice for a population with 50 clusters.

In spite of we have assumed in our theoretical study that each cluster selected in the first stage is subsampled in the second stage, for the simulation study we have supposed that each cluster selected in the first stage will be entirely studied in the second stage, that is to say, a special case of simple random subsampling with $n_i = N_i$. In this way the subsampling error is removed and the comparison is concentrated on the cluster sampling error.

Furthermore, in order to apply the random group strategy, the quantities $M_1, \ldots, M_n$, that is to say, the number of clusters in each random group, are taken according to the optimal choice to minimize the variance, mentioned in Section 3.

The results of the comparison are given in Table 2. For the studied sampling strategies, we has used the following abbreviated names,

1. PPS-HHC $\sqrt{\ }$. Cluster selection probabilities proportional to,

$$N_i \sqrt{X_{(N)} - \overline{X}_i - \frac{n_i - 1}{2 n_i N_i (N_i - 1)} \sum_{k,l \in C_i} |X_k - X_l|}$$

   for each cluster, with replacement in the first stage, and the Hansen-Hurwitz estimator.

2. PPS-HHC, $T(X)$. Cluster selection probabilities directly proportional to the totals of the auxiliary variable for each cluster, with replacement in the first stage and the Hansen-Hurwitz estimator.

3. PPS-RHCC $\sqrt{\ }$. Cluster selection probabilities proportional to,

$$N_i \sqrt{X_{(N)} - \overline{X}_i - \frac{n_i - 1}{2n_i N_i (N_i - 1)} \sum_{k,l \in C_i} |X_k - X_l|}$$

for each cluster, using the random groups sampling procedure and the Rao-Hartley-Cochran estimator.

4. PPS-RHCC, $T(X)$. Cluster election probabilities proportional to the totals of the auxiliary variable for each cluster, using the random groups sampling procedure and the Rao-Hartley-Cochran estimator.

5. SRSWR-HHC. Simple random sampling of clusters, with replacement and the Hansen-Hurwitz estimator.

6. SRSWOR-HTC. Simple random sampling of clusters, without replacement and the Horvitz-Thompson estimator.

Method	$n = 4$	$n = 7$	$n = 10$	$n = 15$
PPS $-$ HHC $\sqrt{\ }$	0.8046	0.6159	0.5186	0.4252
PPS $-$ HHC, $T(X)$	7.4283	5.9547	4.9698	4.1770
PPS $-$ RHCC $\sqrt{\ }$	0.7828	0.5754	0.4575	0.3555
PPS $-$ RHCC, $T(X)$	7.5403	5.1927	4.5543	3.4284
SRSWR $-$ HHC	2.2245	1.7710	1.4486	1.1886
SRSWOR $-$ HTC	2.1364	1.5577	1.2648	0.9908

TABLE 2. Comparative study for different strategies, using the Clustered MU284 population, P85 as study variable and P75 as auxiliary variable.

We see that the PPS methods, using the optimal selection probabilities, give the best results. As expected, the results are very poor if these methods are used with selection probabilities proportional to the totals of the auxiliary variable. Finally, the random sampling, with or without replacement, shows an intermediate efficiency, but it proves to be better than the PPS sampling with selection probabilities directly proportional to the totals of the auxiliary variable on each cluster.

To sum up, the results given for the empirical Monte Carlo study are in concordance with the theoretical results above obtained. This results show that we can consider our approach as a promising alternative.

Acknowledgments

The author thanks several anonymous referees for extensive and helpful comments. One of them has suggested the relation of the optimal selection probability p_i, under simple random subsampling, with the Gini coefficient of the cluster C_i.

References

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika.* **73**, 597-604.

Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika.* **79**, 577-582.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics.* **14**, 333-362.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika.* **75**, 97-103.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B.* **24**, 482-491.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika.* **77**, 365-375.

Rao, J.N.K. (1994). Estimating totals and distributions functions using auxiliary information in the estimation stage. *Journal of Official Statistics.* **10**, 153-166.

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

Welsh, A.H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society B.* **60**, Part 2, 413-428.