Survey Methodology, December 1994
Vol. 20, No. 2, pp. 185–192
Statistics Canada

Competitors to Genuine πps Sample Designs: A Comparison

OLIVER SCHABENBERGER and TIMOTHY G. GREGOIRE[1]

ABSTRACT

Without-replacement list sampling with probability proportional to some measure of element size has not enjoyed much application in forestry because of the difficulty of implementing such sample strategies, that have been termed πps designs to distinguish without-replacement sampling from the well-known with-replacement pps designs. In this contribution, an exact πps strategy (Sunter's variant 2), an approximate πps design (Sunter's variant 1) and the Rao-Hartley-Cochran random group method are examined and the variances of the respective estimators for total bole volume are computed for four tree populations. The results indicate that compared to the Rao-Hartley-Cochran design Sunter's variant 1 in general leads to higher precision if the relationship between auxiliary information x_k and target characteristic y_k is *loose* but is sensitive to the ordering of the sampling frame, whereas the Rao-Hartley-Cochran design does not require the sampling frame to be ordered at all and appears to be superior if strong linear relationships between x_k and y_k are present.

KEY WORDS: Probability proportional to size sampling; Fixed sample size; Approximate πps designs; Empirical comparison.

1. INTRODUCTION

Rao (1978) classifies methods for unequal probability sampling without replacement in two broad categories, (i) sampling schemes, where the inclusion probabilities π_k are proportional to the characteristic of interest, y_k, and the Horvitz-Thompson π estimator $\hat{t}_\pi$ is utilized; (ii) schemes that entertain statistics other than the Horvitz-Thompson estimator. Strategies in (i) are termed IPPS (inclusion probability proportional to size) and members of (ii) non-IPPS designs. In recent literature, *e.g.*, Särndal *et al.* (1992), selection probabilities when sampling with-replacement are denoted p, whereas their counterparts when sampling without replacement are denoted π. We therefore call sampling designs in (i) genuine πps strategies in this paper. Both, IPPS and non-IPPS designs have in common, that under exact proportionality, *i.e.*, $\pi_k \propto y_k$ and $n(s) \equiv n\{\text{constant}\}$, it is implied that $\text{Var}(\hat{t}) \equiv 0$ where $\hat{t}$ is the respective estimator used. For this reason, it seems appealing to draw a sample without replacement where $\pi_k \propto y_k$ and to keep the sample size fixed at the same time. Our interest in these methods concerns their utility to sampling needs in forestry.

Several exact πps designs are available, Rao (1978) gives an in depth account and discussion. Their implementation however is often a non-trivial task and numerically cumbersome for sample sizes usually encountered in forestry practice. Many of these exact πps strategies require enumeration of all possible samples or use algorithms that become increasingly prohibitive as n increases. A simple design, which is feasible for $n \leq 10$ is described by Sampford (1967).

In forestry, however, the number of samples to be drawn at any stage of a survey is oftentimes much larger, even after stratification. Consequently, one either approximates the πps selection process in a manner that allows the inclusion probabilities to be computed exactly, or approximates second-order inclusion probabilities π_{kl} in a design that ensures an exact πps selection. Rao, Hartley and Cochran (1962) described a non-IPPS design, also known as the random group method, that has gained considerable attention (see also Rao 1966, 1978). It is not a πps design, since it utilizes an estimator other than $\hat{t}_\pi$ to ensure zero variance when the π_k are proportional to y_k, but is of remarkable simplicity. An approximate πps design of the first kind is Sunter's method (Sunter 1977a, 1977b). These two designs are referred to in what follows as RHC and SUN1. Sunter (1986, 1989) described an exact πps strategy that can be applied if certain stipulated conditions about the ordering of the sampling frame are met and the possible samples can be enumerated to obtain π_{kl} for some pairs of elements. To avoid enumeration we use an approximation to these π_{kl}. This scheme will be called variant 2 or SUN2 in what follows.

Särndal *et al.* (1992) describe the SUN1 and RHC strategies as entailing some loss of efficiency compared to corresponding πps designs, but no assessment of their comparative efficiency is provided. To our knowledge, none is extant; yet in light of the practical advantages offered by these designs, a comparative assessment would be helpful.

[1] Oliver Schabenberger and Timothy G. Gregoire, Department of Forestry, Section Forest Biometrics, College of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, U.S.A.

The purpose of this study is to compare the performance of the three strategies empirically, using data from forestry field studies and sampling intensities up to 10% which involve reasonably large samples.

The designs SUN1, SUN2, and RHC are appropriate if one has access to a list of population elements from which the sample can be drawn. A complete enumeration of the target characteristic y_k is not anticipated, but the probabilities of inclusion may be made proportional to an auxiliary variable x_k. That is, having complete knowledge about x_k prior to sampling, where it is surmised that x_k is roughly proportional to y_k, we try to achieve $\pi_k \propto x_k$ while $n \equiv$ constant.

In forestry such auxiliary information oftentimes is an easily obtainable characteristic of tree size such as height h, diameter at breast height d, or a combination thereof, which can be used to sample efficiently for bole volume or biomass, y. For example, the geometry of tree stems suggests relationships between d, h, and the volume contained in the tree bole that can be exploited in sampling. In the present investigation, the target parameter is the total bole volume per unit area or in an entire forest stand. In practice, some form of multistage sampling would be used, but for sake of exposition the present comparison includes single stage sampling only.

For the RHC and SUN designs, the auxiliary variables d, d^2, d^2h and the tree sequence number were used. The sequence number was chosen as an auxiliary variable since in the absence of ordering by size it is clearly unrelated to the target characteristic. It should indicate the sensitivity of competing strategies to uninformative auxiliary information (*cf.* Rao 1966).

All designs were investigated with samples of intensity 1%, 2%, 5%, and 10%. The performance of the different sampling designs was gauged in terms of the variance of each estimator of $t = \sum_{k=1}^{N} y_k$. Ratio-of-means estimation following simple random sampling was used as a benchmark, since it utilizes the same auxiliary information. The variances of the sample designs described in the following section were compared to the mean square error of the ratio-of-means estimator (ROM), evaluated using the second order delta method approximation in Sukhatme *et al.* (1984).

2. SAMPLE DESIGNS

2.1 Sunter's Design, Variant 1

Sunter initially proposed two different approximate πps designs: one relaxes the requirement of proportionality of inclusion probabilities π_k for a subset of the population, the other allows for some variation in sample size (Sunter 1977a, 1977b; Schreuder *et al.* 1990). In order that precision not be unduly sacrificed, it is assumed in the latter case that the variance of $n(s)$ is small, while in the first case that altering some π_k is not too serious. In this study only the first method was used since the RHC design operates with fixed sample size, too, and it is the comparative feasibility of the Sunter and RHC designs that prompted this study. Särndal *et al.* (1992) describe the allocation of the sample and the computation of the inclusion probabilities in detail. For part of the population, $\pi_k \propto x_k$ where x_k is the auxiliary information available for the k-th subject (or record). Let k^* denote an element in the ordered population. Then for all elements where $k < k^*$ selection is carried out proportional to x_k. The process ends if a total sample of size n is allocated or if $k = k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ where $t_k = \sum_{j \geq k} x_j$. In the latter case, the remaining samples are selected according to the list-sequential scheme of Bebbington (1975) among those elements for which $k \geq k^*$. As Sunter points out, this sampling scheme has the advantage that only one pass through the sampling frame is necessary. Moreover, the first and second order inclusion probabilities can be computed during this pass through the file. Since the design ensures that $\pi_{kl} > 0 \; \forall \; k, l$; $\pi_k\pi_l - \pi_{kl} > 0 \; \forall \; k, l$ and n is fixed, the non-negative Yates-Grundy estimator of variance can be readily computed. The first order inclusion probabilities are obtained as $\pi_k = nx_k/T_N$ if $k < k^*$ and $\pi_k = n\bar{x}_k/T_N$ if $k \geq k^*$ where $T_N = \sum_{k=1}^{N} x_k$ and $\bar{x}_{k^*} = t_{k^*}/(N - k^* + 1)$. Expressions for the second order inclusion probabilities are given in Särndal *et al.* (1992).

Consequently, the ordering of the population affects the performance of the SUN1 design, since the inclusion probabilities and therefore the variance depend on k^* (see (2) below). For large sample sizes the condition $k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ may be resolved in favor of $k^* = \min\{k: nx_k/t_k \geq 1\}$, which in turn may lead to a premature switch from πps to SRS sampling owing to the ordering of the sampling frame. Note that $x_k/t_k < x_{k'}/t_{k'}$ for $k' > k$ need not be true since if $x_k > x_{k+1}$ and $t_k > t_{k+1}$ it may well be that x_k/t_k is greater or smaller than x_{k+1}/t_{k+1}. It thus can happen that $nx_k > t_k$ and $nx_{k'} < t_{k'}$, for some k, k' where $k' > k$. In this case, that may occur rather frequently, it is unclear if the switch from πps to SRS should take place the first time $nx_k \geq t_k$ or not. Sometimes it may happen that for the first two or three elements of the population $nx_k \geq t_k$ but falls below t_k for the main portion of the sampling frame. This is especially the case when n is large and a few very big x_k appear on top of the population list. To stick to Sunter's rule in such a case would in essence be equivalent to drawing a simple random sample.

The π estimator for the population total can be computed as

$$\hat{t}_{\pi\text{SUN1}} = \sum_{k=1}^{N} \frac{y_k}{\pi_k} I_k, \qquad (1)$$

where I_k is the sample inclusion indicator function. The variance is obtained as

$$\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN1}}) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\mathrm{Cov}(I_k,I_l)\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2, \quad (2)$$

which is the Yates-Grundy form with $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ (Särndal *et al.* 1992). We use the notation $\mathrm{VAR}_{\mathrm{SUN1}}$ for (2) subsequently.

2.2 Sunter's Variant 2

In Sunter (1986, 1989) an exact πps design is described for samples of size $n > 2$. To fix ideas let $z_k = x_k/T_N$ and order the population such that

$$nz_k < Z_k,\ k = 1, \ldots, N - (n+1)$$

$$(n-k)z_1 < Z_k,\ l \geq k \geq N - n,$$

where $Z_k = \sum_{i=k}^{N} z_i$. Let m_k denote the number of samples out of n still to be drawn when arriving at the k-th population element u_k. Given that the two conditions are met, the following algorithm selects an exact πps sample. For u_k, $P(u_k \mid m_k) = nz_k/Z_k$ until $m_k = 0$ or $m_k = N - k$; in the latter case discard one of the remaining units with probability $1 - (m_k z_l/Z_k)$ and retain the others.

It is not always possible to order the population such that the above conditions are met. Sunter (1986) describes an algorithm that checks, whether the ordering is possible. The inclusion probabilities are

$$\begin{aligned}\pi_k &= nz_k \\ \pi_{kl} &= n(n-1)z_k z_l \gamma_k \quad k \leq N - n - 1,\ l > k,\end{aligned} \quad (3)$$

where

$$\gamma_k = \frac{1}{Z_{k+1}}\left(1-\frac{z_1}{Z_2}\right)\cdots\left(1-\frac{z_{k-1}}{Z_k}\right), \quad k = 2, \ldots, N-(n+1).$$

The remaining second-order inclusion probabilities, namely π_{kl} for $l > k > N - n$ have to be obtained from enumeration of possible samples which is likely to be infeasible. Sunter argues that (3) gives a good approximation for those pairs of elements, and this approximation has been used here. With these inclusion probabilities, $\hat{t}_{\pi\mathrm{SUN2}}$ is indicated by the right-hand-side (rhs) of (1). An approximation to $\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN2}})$ is given by (2), wherein (3) is used to obtain π_{kl} for $l > k > N - n$.

The differences between SUN1 and SUN2 are noteworthy. With SUN1 the joint inclusion probabilities are computed exactly for all pairs, but the selection is not genuine πps because of the introduction of SRS in part. In Sunter's variant 2 the selection is exactly πps, but $\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN2}})$ can only be approximated. We use $\mathrm{VAR}_{\mathrm{SUN2}}$ to denote this approximation.

2.3 RHC Design

A description of the RHC design is straightforward; properties of the RHC estimator are well documented in Rao, Hartley and Cochran (1962), and Rao (1966, 1978). After fixing the sample size n, the universe of size N is randomly divided into n groups of size N_i where $N = \sum_i N_i$ $(i = 1, \ldots, n)$. Let X_{ik} denote auxiliary information for element u_k in group i, $k = 1, \ldots, N_i$, and put $X_{i.} = \sum_{k=1}^{N_i} X_{ik}$. From each group one element is selected with selection probability $p_{ik} = X_{ik}/X_{i.}$. The estimator for the total in group i is given as

$$\hat{t}_{i\pi} = \sum_{k=1}^{N_i}\frac{y_{ik}}{p_{ik}}I_{ik},$$

where I_{ik} is the sample inclusion indicator function for element u_k in group i. The population total is then estimated by

$$\hat{t}_{gr} = \sum_{i=1}^{n}\hat{t}_{i\pi}, \quad (4)$$

with variance

$$\mathrm{Var}(\hat{t}_{gr}) = \frac{1}{N(N-1)}\left(\sum_{i=1}^{n}N_i^2 - N\right)\left(\sum_{k=1}^{N}T_N y_k^2/x_k - t^2\right). \quad (5)$$

Note that (5) depends on the group sizes and is minimized when all are equal. In our application, we determined N_i such that some groups were of size $N_i = [N/n]_{gif}$ where *gif* denotes the greatest integer function and the remainder of size $N_i = [N/n]_{gif} + 1$. The number of groups of each size is chosen so that the sum of the group sizes is N. If N/n is an integer, all groups are of course of equal size. We denote (5) by $\mathrm{VAR}_{\mathrm{RHC}}$ in the sequel.

The RHC design is not an exact πps design, since the subdivision of the population introduces a source of randomness unrelated to the size of the auxiliary variable and (4) is not a Horvitz-Thompson estimator. The inclusion probability depends jointly on the size of X_{ik} and on the probability of an element being assigned to group i. Ordering of the population has no effect on $\mathrm{VAR}_{\mathrm{RHC}}$.

3. TREE POPULATIONS

Table 1 shows the tree populations under consideration and Figure 1 displays the relationship between the various choices for x_k and the target characteristic for the yellow poplar population. We notice almost perfect proportionality between d^2h and volume, the relationship between d and volume is clearly curvilinear, and the relationship between d^2 and volume is intermediate. No noticeable trend between sequence number and volume is apparent in the unordered sampling frame. For the remaining three populations similar patterns hold.

For the four populations and the various combinations of auxiliary variable and sampling intensity, there were no observations for which $nx_k > T_N$, thus no records were measured with certainty.

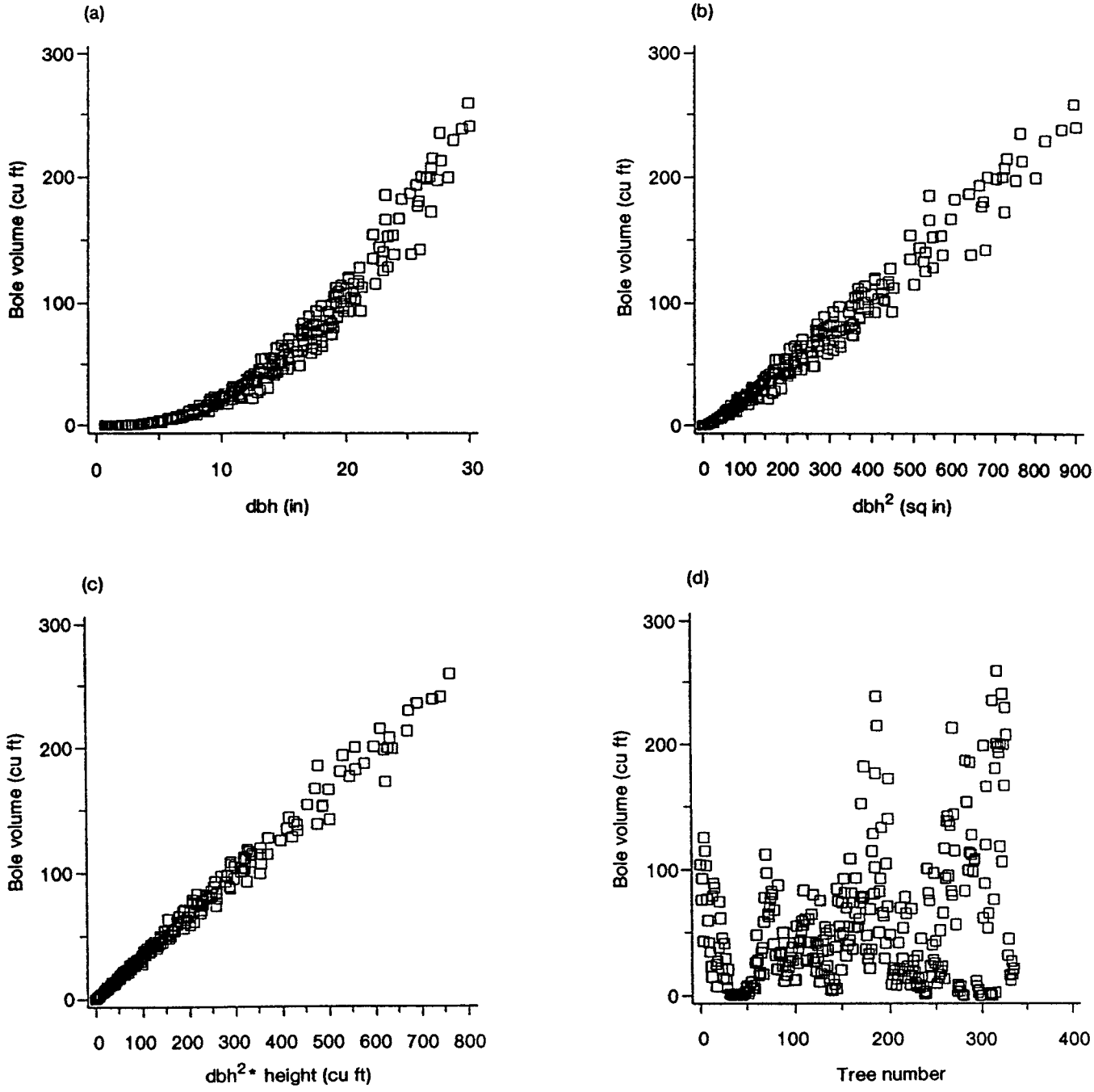

Figure 1. Relation of bole volume to bole dimensions in yellow poplar: (a) diameter at breast height; (b) diameter squared; (c) squared diameter times height; (d) tree sequence number.

Table 1

Tree Populations Examined in an Empirical Comparison of SUN1, SUN2, and RHC
The Last Four Columns Contain Pearson Correlation Coefficients Between x_k and y_k

Species		$N^{(1)}$	$t(ft^3)^{(2)}$	$\rho_{(y;x)}$			
				d	d^2	d^2h	No
Ponderosa pine	*Pinus ponderosa*	140	9,366.6	0.99	0.99	0.99	0.31
Yellow poplar	*Liriodendron tulipifera*	336	18,255.5	0.96	0.96	0.99	−0.07
Loblolly pine	*Pinus taeda*	437	1,835.8	0.96	0.96	0.99	−0.32
Red pine	*Pinus resinosa*	91	4,075.7	0.96	0.96	0.97	−0.05

(1) N is the number of trees in the population.
(2) t is total volume.

4. RESULTS

4.1 Comparison of Variances

The variance of the estimators of t corresponding to the SUN1, SUN2, and RHC design, expressed as a proportion of the MSE under the ROM strategy are compared in Table 2 for the yellow poplar population for each of the sampling intensities investigated and Table 3 depicts pertinent results for the remaining populations. For the SUN1 strategy, the populations were ordered by decreasing size of X, as recommended by Sunter (1977a, 1977b). We focus initially on the results for the yellow poplar population in Table 2.

Table 2

Relative Performances of SUN1, SUN2 and RHC Design
for the Yellow Poplar Population where
Ratio-of-means Estimation (ROM) Serves as a Benchmark

$n/N\%$	X	n	$\frac{\text{VAR}_{\text{SUN2}}}{\text{MSE}_{\text{ROM}}}$	$\frac{\text{VAR}_{\text{SUN1}}}{\text{MSE}_{\text{ROM}}}$	$\frac{\text{VAR}_{\text{RHC}}}{\text{MSE}_{\text{ROM}}}$	k^*[1]
1	No	4	4.8120	3.3136	4.7767	
1	d	4	0.6735	0.6684	0.6731	332
1	d^2	4	0.4605	0.4596	0.4613	333
1	d^2h	4	0.3361	0.3378	0.3402	330
2	No	7	5.1327	2.6346	5.0568	
2	d	7	0.7090	0.6982	0.7081	325
2	d^2	7	0.5731	0.5694	0.5751	318
2	d^2h	7	0.4263	0.4542	0.4369	316
5	No	17	5.4938	1.6643	5.2793	
5	d	17	0.7305	0.7808	0.7283	309
5	d^2	17	0.6541	0.6992	0.6608	291
5	d^2h	17	0.4603	1.2638	0.4935	285
10	No	34	5.8326	1.0985	5.3594	
10	d	34	0.7385	0.7083	0.7339	247
10	d^2	34	0.6712	0.9687	0.6864	260
10	d^2h	34	0.4298	3.0140	0.5037	250

[1] k^* is the observation in the ordered sampling frame at which the SUN1 design switches from πps to SRS sampling.

For a given sampling intensity the precision of all designs relative to ROM increases in the order $X \equiv No$, d, d^2, d^2h; *i.e.*, with increasing proportionality between auxiliary variable and tree bole volume. Given that the approximation of the variance of SUN2 performs well, VAR_{SUN2} can be regarded as measuring the closeness of the RHC and SUN1 designs to matching the efficiency of a genuine πps selection. At low sampling intensities and with meaningful auxiliary information the two designs do not deviate much from SUN2. The performance of both RHC and SUN1 appears to deteriorate at higher sampling intensities relative to SUN2 depending on the choice of size measure. For $X \equiv d^2h$, in which case $\rho_{(y;x)} \cong 0.99$ (see Table 1), RHC is still .85 (.4298/.5037) as efficient as SUN2 but SUN1 is only .14 (.4298/3.014) as efficient, when $n/N\% = 10$. The performance of RHC and SUN1 relative to SUN2 improves for other choices of X which are less well correlated with Y. Indeed, when $X = No$, SUN1 is much more efficient than SUN2.

A puzzling aspect of these results is the indication that SUN2 is less efficient than either RHC or SUN1 for some choices of auxiliary variable and sampling intensity. We speculate that it may be an artifact of the approximation of some second-order inclusion probabilities incorporated into VAR_{SUN2}. It also may depend on the particular ordering used in SUN1 or the group sizes used in RHC sampling, respectively. It is feasible to calculate the exact $\text{Var}(\hat{t}_{\pi\text{SUN2}})$ for $n = 2$. We did so for the ponderosa pine and the red pine populations. The results indicate that VAR_{SUN2} approximates the precision of the SUN2 design very well, but is slightly conservative. The ratios $\text{Var}(\hat{t}_{\pi\text{SUN2}})/\text{VAR}_{\text{SUN2}}$ took on values between 0.975 and 0.999. For larger sample sizes there is no feasible way to determine how well the approximation VAR_{SUN2} performs.

We focus now on the comparison of RHC to SUN1, again with reference to Table 2. At low sampling intensities, VAR_{SUN1} and VAR_{RHC} are essentially equivalent when $X \equiv d^2h$. But using this auxiliary variable at higher intensities led to a substantially better performance of $\hat{t}_{gr}$ in some cases. The most noteworthy case is $n/N\% = 10$ where $\hat{t}_{gr}$ is nearly 6 times more precise than $\hat{t}_{\pi\text{SUN1}}$.

We surmise from these results that the better $x_k \propto y_k$ holds, the better is the precision of $\hat{t}_{gr}$ relative to $\hat{t}_{\pi SUN}$ owing chiefly to the effect of k^* on VAR_{SUN}. Small values of k^* indicate an early switch to a SRS selection and coincide with small values of VAR_{SUN2}/VAR_{SUN1}. Large values of k^* on the other hand correspond to variance ratios close to 1. For yellow poplar, $n/N\% = 10$ and $X \equiv d^2h$ the SUN1 design selects only three-fourths of the population according to a πps design; we conjecture that the early transition to SRS serves also as an explanation for its poor performance compared to the RHC design. When X = tree sequence number, SUN1 is much more precise than RHC, and its relative precision increases as n increases.

The sharp improvement in efficiency when using an auxiliary variable other than tree sequence number provides an indication of the effectiveness of the strategies discussed here when X is positively correlated to Y, and to the liability of sampling with probability proportional to an auxiliary variable when it is unrelated to Y.

The pattern evident in the results for yellow poplar are generally seen, also, in the results for the other species. Some of them are summarized in Table 3. For ponderosa pine SUN1 relative to RHC is always less precise when $X \equiv d^2h$ regardless of the sampling intensity and SUN2 performs always best when this variable is used. For all species the combination $n/N\% = 10$, $X \equiv d^2h$ leads to low precision of SUN1 compared to the other designs and with the exception of the loblolly pine population, SUN1 performs poorer than ratio-of-means estimation. For all populations, the order of magnitude better precision of ROM over the genuine πps, non-IPPS or approximate πps design when X = tree sequence number is remarkable.

From Figure 1 it can be seen that the ordering of volume by tree numbers is haphazard, *i.e.*, the sequence number carries no information about bole volume. And, there is a price to pay if one uses this uninformative auxiliary information to determine inclusion probabilities. The inefficiency of unequal probability sampling in presence of uninformative auxiliary information is an important limitation for the simultaneous estimation of multiple population attributes, where some may be closely related to the auxiliary design variable but others might be uncorrelated with it. Rao (1966) discusses this point in detail and he proposes alternative estimators based on the unbiased estimators in equal probability sampling and the estimator $\hat{t}_{gr(alt)} = N\sum_i y_i \xi_i$, where $\xi_i = \sum_k^{N_i} p_{ik}$ in the RHC design. Applying this estimator in the case of unequal probability sampling leads to bias, but to better mean-square error performance. For the RHC design with X = tree sequence number, the alternative estimator proposed by Rao (1966) improved the ratio $MSE_{RHC(alt)}/MSE_{ROM}$ remarkably. For the yellow poplar population for example, these ratios were between 1.34 ($n = 4$) and 2.58 ($n = 34$), corresponding to a mean square error of the alternative estimator of only 28% to 48% ($n = 34$) of the RHC estimator (5). Similar patterns hold for the other tree species.

Since the alternative estimator is inconsistent, its bias does not depend on n, the larger ratios within the range for each species appear for larger sample sizes. It thus seems reasonable to limit the use of this estimator to smaller sample sizes. When n gets larger, another alternative is to use a ratio estimator, *e.g.*, Hajek's estimator $N\{(\sum y_i/\pi_i)/(\sum 1/\pi_i)\}$ under a genuine πps design.

Table 3

Pertinent Results About the Relative Performances of SUN1, SUN2 and RHC Design for the Remaining Populations where Ratio-of-means Estimation (ROM) Serves as a Benchmark

$n/N\%$	X	n	$\frac{VAR_{SUN2}}{MSE_{ROM}}$	$\frac{VAR_{SUN1}}{MSE_{ROM}}$	$\frac{VAR_{RHC}}{MSE_{ROM}}$	k^*
			Ponderosa Pine			
1	*No*	2	1.9608	1.9794	1.9507	
1	d^2h	2	0.1050	0.1096	0.1077	137
2	*No*	3	2.2976	1.9264	2.2275	
2	d^2h	3	0.1768	0.1919	0.1859	135
5	*No*	7	2.8717	2.0681	2.7819	
5	d^2h	7	0.3113	0.3890	0.3670	129
10	*No*	14	3.2528	2.2745	3.0294	
10	d^2h	14	0.2928	1.3724	0.4488	97
			Red Pine[1]			
2	*No*	2	2.0210	1.9485	2.0029	
2	d^2h	2	0.9076	0.9026	0.9104	90
5	*No*	5	2.9295	2.3141	2.8236	
5	d^2h	5	0.8874	1.3456	0.8991	87
10	*No*	9	3.5548	2.0124	3.2958	
10	d^2h	9	0.8699	1.3192	0.8942	81
			Loblolly Pine			
1	*No*	5	4.8011	3.7104	4.7625	
1	d^2h	5	0.4043	0.4161	0.4174	431
2	*No*	9	5.5940	3.7441	5.5044	
2	d^2h	9	0.5129	0.5510	0.5476	419
5	*No*	22	6.5290	3.3082	6.5253	
5	d^2h	22	0.5035	0.6385	0.6085	406
10	*No*	44	7.7977	2.6635	6.5708	
10	d^2h	44	0.3854	0.7214	0.6146	375

[1] The sampling intensity 1% was omitted since it would have resulted in $n = 1$.

4.2 The Effect of Ordering on The Precision of Sunter's Variant 1

Sunter and others have noted that the precision of the SUN1 design depends on the ordering of the population. The recommendation to sort the sampling frame by decreasing size of x_k's is rooted in the assumption that larger x_k are more likely to be proportional to y_k than smaller ones. The goal is to apply the πps part of the SUN1 design not only to as big a portion of the population as possible but also to those elements for which $x_k \propto y_k$ holds best. Under this assumption it was thus advised to put the elements with large x_k values at the top of the frame. However, it is clear that this is only a rough rule of thumb, since the assumption of greater proportionality with increasing size may not hold.

To investigate the effect of ordering the ponderosa pine and red pine populations were first sorted by increasing x_k and then grouped into 10 groups of approximately equal size. The Pearson correlation coefficient between x_k and y_k was computed within each group and the populations were then sorted by

(a) groups of decreasing correlation and increasing size of x_k within each group,

(b) groups of decreasing correlation and decreasing size of x_k in each group

and SUN1 sampling was repeated for the combinations of x_k's and sampling intensity 10%. Table 4 shows the results.

Table 4

$\text{Var}_{\text{SUN1}}/\text{MSE}_{\text{ROM}}$ for Ponderosa Pine and Red Pine and Different Ways of Ordering the Population

	Ponderosa Pine Ordered by			Red Pine Ordered by		
X	decr.x_k	decr.ρ incr.x_k	decr.ρ decr.x_k	decr.x_k	decr.ρ incr.x_k	decr.ρ decr.x_k
d	0.5614	0.6165	0.6043	1.0307	1.0236	0.6454
d^2	0.3478	0.6562	0.5869	1.2077	0.9373	0.6948
d^2h	1.3724	60.861	0.4459	1.3192	0.8674	0.7461

The results are rather surprising. For red pine the order by decreasing correlation improved all measures of precision. Sorting by increasing x_k within each group now made VAR_{SUN1} very close to VAR_{RHC}, and with $x = d^2h$, $\text{VAR}_{\text{SUN1}} < \text{VAR}_{\text{RHC}}$. Sorting by decreasing x_k within each group achieved an even greater improvement. In contrast to these results, sorting the ponderosa pine population by decreasing ρ and increasing x_k made things worse. The very high value of 60.861 is caused by a premature switch to SRS, since in this setting k^* is only 28, corresponding to only 20% of the population being sampled πps. Moreover, using order of decreasing ρ and decreasing x_k improved VAR_{SUN1} only for $x = d^2h$.

These results indicate that there may exist an order that minimizes VAR_{SUN1} and may yield higher precision than a simple ordering by decreasing value of X. But this order will usually differ depending upon the auxiliary information, and even an ordering that is reasonable on intuitive grounds may give unanticipated results. It is not known if any ordering is optimal in the sense of minimizing $\text{Var}(\hat{t}_{\pi\text{SUN1}})$ for the approximate πps design used in this study. According to our present knowledge no optimal strategy has been described.

5. DISCUSSION AND CONCLUSION

Employing some meaningful auxiliary information leads to a considerable gain in precision in the unequal probability designs compared to a ratio-of-means estimation.

A choice between the two Sunter designs can be made on grounds of the relationship between size measure and target characteristic. When $X \propto Y$ is strong, SUN2 offers advantage over SUN1, and SUN1 appears preferable when the relationship is weak. Based on our results, the approximate πps strategy, SUN1 and the non-IPPS design RHC appear to come fairly close to the efficiency offered by genuine πps selection. With increasing sampling intensity, however, the highest precision is obtained with the SUN2 design. But the quality of the approximation VAR_{SUN2} in this case is unclear.

If one's aim is to use an approximate πps or a non-IPPS strategy then the RHC design with estimator $\hat{t}_{gr}$ appears to offer advantages over the Sunter design with $\hat{t}_{\pi\text{SUN}}$, at least for the tree populations studied here with the objective of estimating total bole volume. At reasonably low sampling intensities, both estimators appear to be equally precise.

An advantage of the RHC design is its simplicity. An operational advantage is that it can be applied to every population because it is impervious to its ordering and provides an unbiased estimation within each group. While the first criterion is also met by Sunter's variant 1, the ordering there clearly affects the precision of the estimator $\hat{t}_{\pi\text{SUN1}}$. Variant 2 can only be used if some ordering of the population meets the conditions given in Section 2.2. Otherwise the selection algorithm does not produce a sample of exactly size n.

The precision of the RHC method, however, depends on the group sizes employed. The algorithm given in Section 2.3 is optimal.

While a particular ordering may improve the precision of $\hat{t}_{\pi\text{SUN1}}$, it is unclear at present how to discern an optimal ordering and a fixed sample size. Moreover an optimal

ordering of one choice of auxiliary variable or attribute of interest may be deleterious when implemented with a different auxiliary variable or attribute.

All strategies can be disastrous with uninformative auxiliary information.

Finally and to the extent that computational burden is a meaningful criterion, RHC is arguably less burdensome than variant 1 of Sunter's design.

ACKNOWLEDGMENT

We gratefully acknowledge the comments and suggestions by J.N.K. Rao, C.-E. Särndal, and A. Sunter who reviewed earlier versions of the manuscript as well as the helpful comments of the referees whose contribution helped to improve the paper substantially.

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā* A, 28, 47-60.

RAO, J.N.K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. *Contributions to Survey Sampling and Applied Statistics* (H.A. David, Ed.), New York: Academic Press, 69-86.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society* B, 24, 482-491.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SÄRNDAL, C.-E., SWENSSON B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SUKHATME, P.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications (3rd Ed.)*. Iowa State University Press.

SUNTER, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.

SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.

SUNTER, A. (1989). Updating size measures in a PPSWOR design. *Survey Methodology*, 15, 253-260.

SCHREUDER, H.T., LI, H.G., and SADOOGHI-ALVANDI, S.M. (1990). Sunter's pps Without Replacement Sampling as an Alternative to Poisson Sampling. USDA Forest Service Research Paper RM-290.