Comparison of Randomized Branch Sampling with and without Replacement at the First Stage

Jorge Cancino and Joachim Saborowski

Cancino, J. & Saborowski, J. 2005. Comparison of randomized branch sampling with and without replacement at the first stage. Silva Fennica 39(2): 201–216.

Randomized Branch Sampling (RBS) is a multistage sampling procedure using natural branching in order to select samples for the estimation of tree characteristics. Usually, sampling units are selected with unequal probabilities. Conventional RBS uses sampling with replacement (SWR) for repeated sampling on the first stage, and the sample size equals 1 on all subsequent stages, thus resulting in n so-called sample paths. When the sampling fraction is large multiple selections of first stage units are likely. Sampling without replacement (SWOR) at the first stage is an alternative that is expected to increase efficiency of the procedure. In this case, the second stage sample size m must be larger than 1 to enable unbiased variance estimation. In the present study, a theoretical and an empirical comparison of the conventional RBS and the SWOR variant was accomplished. Requiring a certain precision of the RBS estimation, the conventional RBS method is mostly more time-consuming than the variant with SWOR at the first stage. Only if $m = 1$ is chosen as second stage sample size for the SWOR RBS, this is often more time-consuming. In those cases, conventional RBS is up to 5% cheaper. In general, the larger m is, the more expensive is conventional RBS compared with the variant with SWOR at the first stage. The smaller the ratio of the variance between the primary units to the total variance of the estimate, the larger is the advantage of the SWOR variant. Generally, it can be shown that the gain of efficiency by SWOR is larger in case of weak correlations between auxiliary and target variable.

Keywords Randomized Branch Sampling, sampling without replacement, probability proportional to size, unequal selection probabilities, multistage sampling
Authors´ addresses Cancino: Facultad de Ciencias Forestales, Universidad de Concepción, Chile; Saborowski: Institut für Forstliche Biometrie und Informatik, Georg-August-Universität Göttingen, Germany
E-mail jcancino@udec.cl, jsaboro@gwdg.de
Received 29 July 2004 **Revised** 9 February 2005 **Accepted** 29 March 2005

1 Introduction

RBS uses the natural branching within the crown in order to take samples to estimate one or more parameters of individual trees (see e.g. Jessen 1955, Valentine et al. 1984, Williams 1989, Valentine et al. 1994, Raulier et al. 1999, Good et al. 2001, Mundson et al. 1999, Xiao et al. 2000). It requires the definition of nodes (points at which a branch or part of a branch ramifies into subordinated branches), segments (parts of a branch between two consecutive nodes; see Fig. 1a), and paths (series of successive segments between the first node and a final segment; i.e. a segment without a node at its end).

The selection of a path begins at the first node by drawing one of the adjacent segments. This segment is traced until another node is encountered where the selection procedure is repeated. The path is terminated when a final segment is selected. The target variable such as needle weight or branch biomass is measured whenever it occurs at a path segment. RBS allows the selection of segments with probability proportional to an auxiliary variable. Although, in principle, any segment characteristic can be defined as an auxiliary variable, it is advantageous in terms of precision of the final estimate to select an auxiliary variable closely correlated with the target variable (Jessen 1955, Grosenbaugh 1967, Valentine et al. 1984, Cancino 2003).

Classical RBS uses SWR at the first node (see e.g. Gregoire et al. 1995, Parresol 1999, Good et al. 2001, and Snowdon et al. 2001), which can result in a loss of efficiency. Therefore, Saborowski and Gaffrey (1999) suggested SWOR at the first node. Their approach is based on the well-known fact that, with simple random samples, SWOR is more efficient than SWR (Cochran 1977). The authors developed a sampling strategy which incorporates the method of Sampford (1967) for the selection of sampling units with unequal probabilities and without replacement in the well known multistage sampling estimator (Rao 1975, Saborowski 1990). It combines SWOR with unequal selection probabilities on the first stage and unequal probability SWR on the second and all subsequent stages. On the second stage, sample sizes larger than one are possible and even necessary if unbiased variance estimation is required. On the third and all subsequent stages sample size equals 1. So, sampling beginning with the third stage is actually classical RBS and the SWOR method differs from pure classical RBS only on the first stage and, if the second stage sample size is larger than one, on the second.

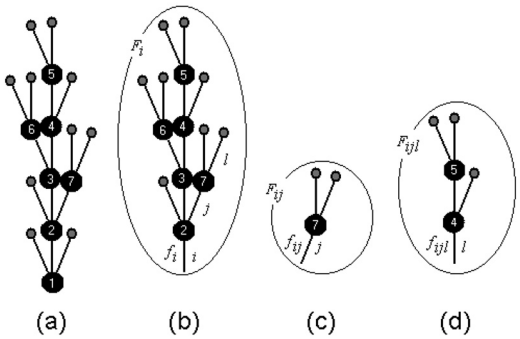

Fig. 1. a) Scheme of a tree with 7 nodes and 16 segments. b)–d) Subordinated crown partitions: a primary (i), a secondary (j), and a tertiary segment (l), the corresponding target variables f_i, f_{ij}, and f_{ijl} at these segments, and the cumulated values F_i, F_{ij}, and F_{ijl}.

In the present paper we compare the classical RBS and the SWOR variant in a case study and by additional theoretical considerations. The comparison is focussed on the precision of estimates and on the time required to achieve a particular level of precision.

2 Classical RBS in the Multistage Sampling Framework

For the comparison of conventional with SWOR RBS it is necessary to embed conventional RBS in the general theoretical framework of multistage sampling. The segments emanating from the first node, usually the lowest node at the stem, can be seen as the primary sampling units of a multistage sampling procedure in a population of N primary units, M_i secondary units in the i^{th} primary unit and K_{ij} tertiary units in secondary unit j of primary unit i and so on. Accordingly, the conditional

selection probabilities of the units in SWR are denoted by q_i, q_{ij} and q_{ijl}.

Following the notation of standard textbooks on sampling techniques (e.g. Cochran 1977), the total of the target variable be $F = \sum_{i=1}^{N} F_i$ where $F_i = \sum_{j=1}^{M_i} F_{ij}$ and $F_{ij} = \sum_{l=1}^{K_{ij}} f_{ijl}$ and, merely for notational convenience, we confine to a three stage design. Note that we use f for the target variable in order to accord with the RBS notation of Gregoire et al. (1995) where f reminds of the amount of foliage but may also represent biomass or other typical characteristics of interest.

Replacing SWOR on the second stage by SWR in Cochran's chapter 11.9 (Cochran 1977) and extending it to three stages, we attain the following unbiased estimators $\hat{F}$, $\hat{F}_i$ (given i) and $\hat{F}_{ij}$ (given ij) for F, F_i and F_{ij} with sample sizes n, m_i and k_{ij}, respectively.

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{F}_i}{q_i}, \quad \hat{F}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{\hat{F}_{ij}}{q_{ij}}, \quad \hat{F}_{ij} = \frac{1}{k_{ij}}\sum_{l=1}^{k_{ij}}\frac{f_{ijl}}{q_{ijl}} \qquad (1)$$

As for the precision of $\hat{F}$, we have

$$Var\,\hat{F} = \frac{1}{n}\sum_{i=1}^{N} q_i\left(\frac{F_i}{q_i} - F\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{Var_2\hat{F}_i}{q_i}$$
$$Var_2\hat{F}_i = \frac{1}{m_i}\sum_{j=1}^{M_i} q_{ij}\left(\frac{F_{ij}}{q_{ij}} - F_i\right)^2 + \frac{1}{m_i}\sum_{j=1}^{M_i}\frac{Var_3\hat{F}_{ij}}{q_{ij}} \qquad (2)$$

with Var_k denoting the conditional variance given the sample selection on stage $k-1$. Finally,

$$V = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{F}_i}{q_i} - \hat{F}\right)^2 \qquad (3)$$

is an unbiased estimator for $Var\,\hat{F}$.

The conditional selection probabilities of segments usually are proportional to the values of an appropriate auxiliary variable x that is measured or assessed at each segment of a node. Choosing the first node (stage) as an example, the conditional selection probability of the i^{th} segment is given by $q_i = x_i / \sum_{i=1}^{N} x_i$.

Apart from two additional restrictions, conventional RBS can be seen as a special case of this multistage sampling framework with $m_i = 1$ and $k_{ij} = 1$. The restrictions are that with the standard multistage design variables of interest are measured only on the last stage (f_{ijl}) and that all paths of the tree under study must be of equal length. With RBS, units on all stages may principally contribute to the totals F, F_i and F_{ij} and the number of stages encountered may vary among different paths. Therefore, we have to redefine the totals as follows

$$F = f + \sum_{i=1}^{N} F_i, \quad F_i = f_i + \sum_{j=1}^{M_i} F_{ij} \text{ and } F_{ij} = f_{ij} + \sum_{l=1}^{K_{ij}} f_{ijl}$$

where f_{ij} is the biomass of secondary segment j emanating from primary segment i of node 1 and f the biomass of the stem below the first node (see Fig. 1b–d). Unbiased estimators of these totals are now given by

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n}\left(f + \frac{\hat{F}_i}{q_i}\right), \quad \hat{F}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\left(f_i + \frac{\hat{F}_{ij}}{q_{ij}}\right), \quad \hat{F}_{ij} = \frac{1}{k_{ij}}\sum_{l=1}^{k_{ij}}\left(f_{ij} + \frac{f_{ijl}}{q_{ijl}}\right) \qquad (1')$$

and with $m_i = 1$, $k_{ij} = 1$ this yields

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n}\left(f + \frac{\hat{F}_i}{q_i}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(f + \frac{f_i}{q_i} + \frac{\hat{F}_{ij}}{q_i q_{ij}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(f + \frac{f_i}{q_i} + \frac{f_{ij}}{q_i q_{ij}} + \frac{f_{ijl}}{q_i q_{ij} q_{ijl}}\right) \qquad (4)$$

Summation over j and l has vanished because $m_i = k_{ij} = 1$ for all i and j, and $Q_i = q_i$, $Q_{ij} = q_i q_j$ and $Q_{ijl} = q_i q_{ij} q_{ijl}$ are the unconditional selection probabilities of segments i, ij and ijl, respectively. Thus, apart from f, the contribution of the stem, $\hat{F}$ according to Eq. 4 coincides with the classical RBS estimator $\hat{F}$ given in Gregoire et al. (1995). It is based on n paths each of which yields an unbiased estimate

$$f + \frac{f_i}{Q_i} + \frac{f_{ij}}{Q_{ij}} + \frac{f_{ijl}}{Q_{ijl}} = f + \frac{\hat{F}_i}{q_i}$$

of the total biomass F (see $\hat{F}_i$ in Gregoire et al. (1995)).

From Eq. 2, with $m_i = 1$, we obtain the variance

$$Var\,\hat{F} = \frac{1}{n}\sum_{i=1}^{N} q_i\left(f + \frac{F_i}{q_i} - F\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{q_{ij}}{q_i}\left(f_i + \frac{F_{ij}}{q_{ij}} - F_i\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{Var_3\hat{F}_{ij}}{q_i q_{ij}} \qquad (2')$$

and from Eq. 3 the well known unbiased RBS variance estimator

$$V = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(f + \frac{\hat{F}_i}{q_i} - \hat{F} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{F}_i}{q_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{F}_i}{q_i} \right)^2 \quad (3')$$

Although these results are already proven by our derivation from well-known multistage sampling theory, we additionally show the equivalence of Eq. 2' with Jessen's variance formula $V(\hat{X})$ (Jessen 1955) in the appendix.

Whereas formulas (Eq. 1' to Eq. 3' and Eq. 4) can obviously be extended to an arbitrary number of stages, it must be explained how varying path lengths are covered by the multistage approach given above. In a given tree, the number of stages may vary among the n primary units selected (Fig. 2) up to a maximum number. This can easily be handled in the multistage framework if one assumes just one artificial segment on each of the missing stages with conditional selection probability 1 and "foliage" 0. If, e.g. secondary segments are missing for primary segment i, then define $M_i = 1$ with $q_{i1} = 1$, $f_{i1} = 0$ implying $F_i = f_i$. So, the second and subsequent stages neither contribute to Eq. 4 nor to the variance Eq. 2' and all results given above remain valid.

Of course, one can only speculate about why classical RBS sampling as described above has been so popular in the past. Probably, the simplicity of the variance estimator (Eq. 3') and of the calculation of selection probabilities yielding inclusion probabilities $\pi_i = nq_i$ are the most important reasons. The discussion in Cochran (1977) reveals how difficult it is to define selection procedures without replacement and with inclusion probabilities π_i proportional to the size of an auxiliary variable. Different methods have been suggested for the computation of π_i and $\pi_{ii'}$ (see e.g. Grundy 1954, Fellegi 1963, Hartley and Rao 1962). Sampford (1967) states that with most of these methods, in particular the determination of the joint inclusion probabilities $\pi_{ii'}$, necessary for variance estimation, is "almost impossibly severe for sample sizes larger than 2" (Sampford 1967, Stevens 1958; see also Yates and Grundy 1953, Durbin 1967, Brewer and Hanif 1970). He generalized the procedure of Durbin (1967) (see Cochran 1967, p. 261–262) to samples of size $n > 2$ and proposed selection procedures that realize the desired inclusion probabilities π_i. Contemporary hard- and software equipment now allows for convenient application of his computationally extensive methods.

3 RBS with SWOR on the First Stage

Saborowski and Gaffrey (1999) proposed the SWOR variant of the classical RBS method. They recommended to draw n units WOR on stage 1, $m_i \geq 1$ units WR on stage 2 in primary unit i, and 1 unit WR on all following stages. Selection of units on the first stage and calculation of inclusion probabilities π_i and $\pi_{ii'}$ for unit i and both units i and i', respectively, being in the sample is accomplished by the method of Sampford (1967).

With this method, π_i is proportional to q_i, i.e. $\pi_i = n \cdot q_i$. In the general multistage setting, this leads to the unbiased estimator $\hat{F}$ according to Eq. 1 with variance

$$Var\,\hat{F} = \sum_{i=1}^{N} \sum_{i'>i}^{N} (\pi_i \pi_{i'} - \pi_{ii'}) \left(\frac{F_i}{\pi_i} - \frac{F_{i'}}{\pi_{i'}} \right)^2 + \sum_{i=1}^{N} \frac{1}{\pi_i} Var_2 \hat{F}_i \quad (5)$$

and variance estimator

$$V = \sum_{i=1}^{n} \sum_{i'>i}^{n} \frac{\pi_i \pi_{i'} - \pi_{ii'}}{\pi_{ii'}} \left(\frac{\hat{F}_i}{\pi_i} - \frac{\hat{F}_{i'}}{\pi_{i'}} \right)^2 + \sum_{i=1}^{n} \frac{1}{\pi_i} V_i \quad (6)$$

where V_i is an unbiased estimator of $Var_2 \hat{F}_i$ (see Eq. 8).

Similar to Eq. 4 and with $k_{ij} = 1$, we have

$$\hat{F} = \frac{1}{n} \sum_{i=1}^{n} \left(f + \frac{f_i}{q_i} + \frac{1}{m_i} \sum_{j=1}^{m_i} \left(\frac{f_{ij}}{q_i q_{ij}} + \frac{f_{ijl}}{q_i q_{ij} q_{ijl}} \right) \right) \quad (7)$$

with the variance according to Eq. 5 and $Var_2 \hat{F}_i$ as given in Eq. 2 but with $f_i + F_{ij}/q_{ij}$ instead of F_{ij}/q_{ij}. The unbiased variance estimator (Eq. 6) is finally obtained with

$$V_i = \frac{1}{m_i(m_i - 1)} \sum_{j=1}^{m_i} \left(\frac{\hat{F}_{ij}}{q_{ij}} - \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{\hat{F}_{ij}}{q_{ij}} \right)^2 \quad (8)$$

This follows from the fact that the sampling procedure beginning on the second stage is identical to conventional RBS, so Eq. 8 is merely a copy of Eq. 3'.

Two details are worth mentioning. 1) Since π_i is a probability, it must be assured that $0 < n \cdot q_i \leq 1$. Moreover, Sampford's (1967) procedures require $n \cdot q_i < 1$ for the calculation of the conditional selection probabilities. That means that the maximum sample size for SWOR is determined by the maximum selection probability according to $n < 1/\max(q_i)$. The restriction of this condition can be lessened by selecting segments having extreme selection probabilities with probability 1, what means that they can be deleted from the crown structure and segments emanating from them are attached to the node preceding the deleted segment (Cancino 2003). 2) Unbiased variance estimation requires $m_i > 1$ for the second stage sample size, as shown in Eq. 6 and Eq. 8 in comparison with Eq. 3'.

4 Theoretical Considerations on the Efficiency of RBS Methods

4.1 Comparison of SWR and SWOR RBS

Repeated selection of single units on the first stage, which can occur with SWR more or less frequently, is expected to make SWOR more efficient than SWR. From the following ratio of the variances (Eq. 2' and Eq. 5) ($f = 0$ for convenience and $m_i = 1$ for both methods)

$$\frac{Var_{wr}\hat{F}}{Var_{wor}\hat{F}} = \frac{\frac{1}{n}\sum_{i=1}^{N} q_i \left(\frac{F_i}{q_i} - F\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{Var_2\hat{F}_i}{q_i}}{\sum_{i=1}^{N}\sum_{i'>i}^{N} (\pi_i\pi_{i'} - \pi_{ii'}) \left(\frac{F_i}{\pi_i} - \frac{F_{i'}}{\pi_{i'}}\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{Var_2\hat{F}_i}{q_i}} \qquad (9)$$

it is obvious that the gain in precision of SWOR at the first stage is the smaller the larger the second stage variances $Var_2\hat{F}_i$ are, compared to the variances between first stage units (see also Cochran 1977). From Eq. 9, it follows (Cancino 2003) that SWOR RBS is more efficient if

$$\frac{n-1}{n} - \frac{\sum_{i=1}^{N}\sum_{i'\neq i}^{N} \frac{\pi_{ii'}}{\pi_i\pi_{i'}} F_i F_{i'}}{F^2} > 0 \qquad (10)$$

If exact proportionality exists between the selection probability and the cumulated target variable of a primary unit, both F_i/π_i and $F_{i'}/\pi_{i'}$ are error free estimators of F/n, so that the expression in Eq. 10 with $\sum_{i'\neq i}^{N} \pi_{ii'} = (n-1)\pi_i$ and $\sum_{i=1}^{N} \pi_i = n$ (Cochran 1977) equals zero. I.e., only in case of a weaker relationship between y_i and π_i SWOR RBS will be more efficient than conventional SWR RBS.

4.2 Optimal Allocation of Sampling Units to Stages (SWR)

Without taking account of costs, the most efficient combination of primary (n) and secondary (m_i) sample size for SWR on both stages can easily be deduced from the following analysis. Let λ be a factor that describes the enlargement and the reduction, of n and m_i, respectively, so that an increase of n to λn is compensated by decreasing m_i down to m_i/λ, keeping the total number of measured segments constant (divisibility supposed). Then, in case of a two-stage procedure with unequal selection probabilities and SWR at both stages (see Eq. 2), for the variance ratio of both settings it holds

$$\frac{\frac{1}{\lambda n}\sum_{i=1}^{N} q_i \left(\frac{F_i}{q_i} - F\right)^2 + \frac{1}{\lambda n}\left[\sum_{i=1}^{N} \frac{\lambda}{m_i}\sum_{j=1}^{M_i} \frac{q_{ij}}{q_i}\left(\frac{f_{ij}}{q_{ij}} - \sum_{j=1}^{M_i} f_{ij}\right)^2\right]}{\frac{1}{n}\sum_{i=1}^{N} q_i \left(\frac{F_i}{q_i} - F\right)^2 + \frac{1}{n}\left[\sum_{i=1}^{N} \frac{1}{m_i}\sum_{j=1}^{M_i} \frac{q_{ij}}{q_i}\left(\frac{f_{ij}}{q_{ij}} - \sum_{j=1}^{M_i} f_{ij}\right)^2\right]} < 1$$

because λ reduces the primary variance component in the numerator while the secondary one remains unchanged. Apparently, it is recommendable to include more primary and, correspondingly, fewer secondary units in the sample. As long as costs are not regarded, the most efficient option is the application of the procedure with $m_i = 1$.

For multistage procedures with SWOR on the first and SWR on the second stage and with unequal selection probabilities, the corresponding ratio of the error variances (Eq. 5) with sample sizes λn and m_i/λ for the numerator and n and m_i for the denominator should be discussed. But, we were not able to reveal a similar effect of λ on the efficiency of RBS. Therefore, case studies might

be helpful to understand the effects of a trade-off between n and m_i.

4.3 Optimization of the Multistage SWR Procedure Considering Costs

In order to compare both RBS procedures in case studies with realistic costs or, equivalently, time consumption for field measurements, it will be necessary to determine optimal sample sizes n and m_i for RBS with SWR on all stages. For convenience, we use the abbreviations

$$\sigma_1^2 = \sum_{i=1}^{N} q_i \left(\frac{F_i}{q_i} - \sum_{i=1}^{N} F_i \right)^2$$

and

$$\sigma_{i\mathrm{Rest}}^2 = \sum_{j=1}^{M_i} q_{ij} \left(\frac{F_{ij}}{q_{ij}} - \sum_{j=1}^{M_i} F_{ij} \right)^2 + \sum_{j=1}^{M_i} \frac{Var_3 \hat{F}_{ij}}{q_{ij}}$$

and rewrite variance (Eq. 2') as follows

$$Var\,\hat{F} = \frac{\sigma_1^2}{n} + \frac{1}{nm} \sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i} \qquad (11)$$

The expected total costs (time) C for the selection of multistage samples are assumed to be

$$C = nc_1 + nmc_2 \qquad (12)$$

where c_1 is the expected cost (time) per primary unit and c_2 the expected cost of measuring all remaining segments of a path, taking into account that repeated selection of certain units with SWR may lead to numbers of units actually to be measured which are smaller than the sample sizes n and m (see Eq. 16). Based on Eq. 11 and Eq. 12, optimal n and m can be derived by minimizing

$$G = \frac{\sigma_1^2}{n} + \frac{1}{nm} \sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i} + \mu \left(nc_1 + nmc_2 - C \right)$$

if expected total costs C are pre-defined. By partial differentiation the minimum of G is obtained at

$$m = \sqrt{\frac{c_1}{c_2 \sigma_1^2} \sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i}} = \sqrt{\frac{c_1}{c_2} \left(\frac{\sigma_1^2 + \sum_{i=1}^{N} \sigma_{i\mathrm{Rest}}^2 / q_i}{\sigma_1^2} - 1 \right)} \qquad (13)$$

Eq. 13 minimizes also

$$G = nc_1 + nmc_2 + \mu \left(\frac{\sigma_1^2}{n} + \frac{1}{nm} \sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i} - Var_e \hat{F} \right)$$

when a certain variance of the estimator $\hat{F}$ is to be attained. The sample size n depends on which function, Eq. 11 or Eq. 12, is used as the restriction in G, thus we have

$$n = \frac{K}{c_1 + mc_2} \text{ and } n = \frac{\sigma_1^2 + \frac{1}{m} \sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i}}{Var_e \hat{F}}, \text{ respectively}$$

where $Var_e \hat{F}$ is the pre-determined variance of F.

From Eq. 13, we recognize that the classical RBS ($m = 1$) is always optimal when the variability among primary units per cost unit (σ_1^2 / c_1) is equal to the variability per cost unit within the paths beginning at a secondary segment

$$\left(\sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i} / c_2 \right).$$

Classical RBS is inefficient when $\sum_{i=1}^{N} \frac{\sigma_{i\mathrm{Rest}}^2}{q_i} / c_2$ is larger than σ_1^2 / c_1, because then the same precision (higher precision) could be achieved with lower costs (the same costs). The larger the optimal sample size m, the more classical RBS deviates from the optimum, and the more time must be spent for the required precision.

5 Alternative Variance Estimators for SWOR RBS ($m_i = 1$)

In Chapter 4.2 it was shown that $m_i = 1$, i.e. the classical RBS, is the optimal choice among all two-stage sampling procedures with SWR on both stages, if costs are ignored. If we choose $m_i = 1$ as well for the SWOR RBS then there is no unbiased estimator for $Var\hat{F}$. Three alternative ad hoc variance estimators are discussed in this paragraph and will be further examined in the case studies.

The first choice, V_{swr}, could be the variance estimator (Eq. 3'), which is unbiased in case of the classical RBS but biased if SWOR is applied on the first stage instead of SWR. The gain in precision of the SWOR RBS compared with classical RBS depends on the variance components and on the relationship between the cumulated target variable and the selection probability of the primary units (see Chapter 4.1). Particularly,

there is no gain if that relationship is exactly proportional. Therefore, V_{swr} should give reliable, although conservative, estimates of the variance whenever the relationship is close to proportionality. Remarkable over-estimation can be expected to occur for trees with only weakly related F_i and π_i. One might try to reduce this over-estimation by means of a correction factor that plays the role of the finite population correction $(1-n/N)$ in simple random sampling. It is derived from the unconditional selection probabilities of the selected paths as follows in the next paragraph.

If we consider each path as a singular unit, the SWOR of n primary units and $m_i = k_{ij} = \ldots = 1$ units at the subsequent stages approximately corresponds to SWOR of n from a population of N_{Paths} paths with inclusion probability $\pi_p = n \cdot Q_{R_p}$ for path p. Since Q_{R_p} corresponds to $1/N$, the selection probability in simple random sampling, $1-\sum_{p=1}^{n} Q_{R_p}$ is used instead of $(1-n/N)$, hence

$$V_{corr}(\hat{F}) = K \cdot V_{swr}(\hat{F}) \quad \text{with } K = 1 - \sum_{p=1}^{n} Q_{R_p} \qquad (14)$$

The third variance estimator is obtained if we assume each path a sampling unit with inclusion probability $\pi_p = n \cdot Q_{R_p}$ and $\hat{F}_p$, the estimator of the total in that crown compartment which begins with primary segment p, a fixed measurement. The resulting estimator formally coincides with the Sen-Yates-Grundy variance estimator in a one-stage sampling design.

$$V_{1-stage}(\hat{F}) = \sum_{p=1}^{n} \sum_{p'>p}^{n} \left(\frac{\pi_p \pi_{p'} - \pi_{pp'}}{\pi_{pp'}} \right) \left(\frac{\hat{F}_p}{\pi_p} - \frac{\hat{F}_{p'}}{\pi_{p'}} \right)^2 \qquad (15)$$

In order to compute Sampford's (1967) $\pi_{pp'}$ for the selected paths, N_{Paths} as well as the probabilities of all N_{Paths} paths must be known. Therefore, we need estimations of N_{Paths} and the $N_{Paths}-n$ unknown selection probabilities Q_{Rj}, $j = n+1,\ldots,N_{Paths}$.

In sampling with unequal probabilities, $1/q_i$ represents the number of units on the first stage for the estimation of the total by $\hat{F}_i / q_i$, such as $N\hat{Y}_i$ in simple random sampling. Similarly, the reciprocal of the selection probability of each path supplies an estimation of the total number of paths of the tree, i.e. an ad hoc estimator of N_{Paths} could be $\hat{N}_{Paths} = \frac{1}{n} \sum_{p=1}^{n} \frac{1}{Q_{R_p}}$. Further, the unknown probabilities Q_{Rj}, $j = n+1,\ldots,N_{Paths}$, are substituted by the average

$$Q_{R_j} = \left(1 - \sum_{p=1}^{n} Q_{R_p}\right) \Big/ \left(\hat{N}_{Paths} - n\right)$$

derived from $\sum_{p=1}^{N_{Paths}} Q_{R_p} = 1$,

i.e. $\sum_{j=n+1}^{N_{Paths}} Q_{R_j} = 1 - \sum_{p=1}^{n} Q_{R_p}$

6 Database for Case Studies

Data of complete trees of three different species were available for the analysis: spruce (*Picea abies* (L.) Karst.), European mountain ash (*Sorbus aucuparia* L.), and Monterey pine (*Pinus radiata* D. Don) (Table 1, Fig. 2a).

The data for the young spruce trees were collected in the Solling mountains (Lower Saxony, Germany) by the plant-modelling working group of the Institute of Forest Biometry and Informatics of the University of Goettingen. One tree was completely measured and the other trees only sampled. The missing values of the target variable "needle biomass" were estimated by regression. The base diameter of each segment is available.

The eight pine trees come from two pure, even-aged (14 and 29 years old) stands in Cholguán (VIII Región, Chile). For each tree, the position of the branch (height above ground), its length and base diameter, as well as the total weight of each fifth branch were measured. The missing weights were determined by regression, and branches located between two whorls were assigned to the nearest whorl or to an additional node.

Moreover, a time study was carried out during the field measurements of 9 additional European mountain ashes in Bärenfels (Sachsen, Germany). Diameter and leaf biomass were measured for all sampled segments of a tree by a working team of three people. Sampling was done by classical RBS with multiple stages as well as by SWOR RBS ($m_i = 1$) at each tree.

At each node, individual time consumption for two different working units was measured; i.e. the working time at the node (counting of segments, measurement of the auxiliary variable, selection of segments, marking of selected segments) and

Table 1. Tree characteristics.

Tree	Age (years)	DBH (cm)	Height (m)	Biomass	Number of nodes	Number of segments	Number of paths
Norway spruce							
1	14	–	0.4	16.6 [a]	11	598	337
2	16	–	–	41.6	29	623	318
3	12	–	–	99.6	50	901	456
4	11	–	–	11.2	34	233	119
Young Monterey pine							
1	14	25.5	14.4	186.6 [b]	27	164	138
2	14	18.6	14.2	81.8	23	114	92
3	14	14.8	16.4	25.9	7	52	46
4	14	14.2	14.4	31.5	13	84	72
Old Monterey pine							
1	29	51.5	37.6	249.8 [b]	45	184	140
2	29	51.2	33.2	1035.9	56	198	143
3	29	40.7	37.9	146.6	31	147	117
4	29	36.8	40.2	277.7	53	235	183
European mountain ash							
1	16	2.3	4.5	106.9 [c]	23	54	28
2	16	4.0	4.7	351.3	32	156	79
3	26	4.5	6.9	234.8	25	114	58
4	19	7.8	7.8	386.4	32	274	138

[a] Dry weight of needles (g); [b] Fresh branch biomass (kg); [c] Dry weight of leaves (g)

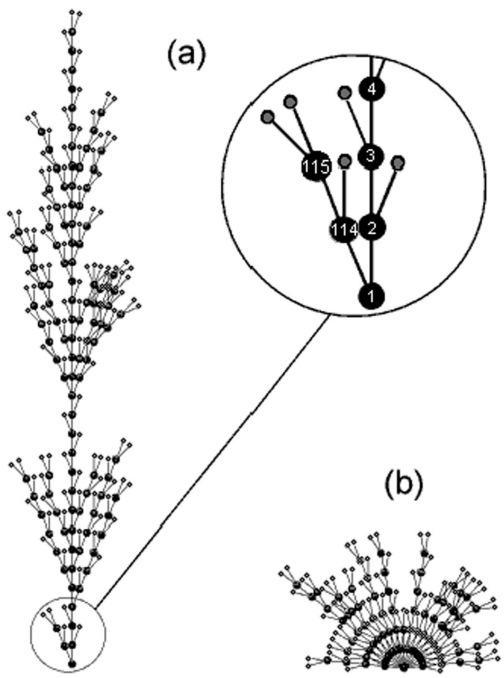

Fig. 2. a) Two-dimensional representation of spruce 4 with and b) without stem.

at the selected segments (gathering leaves). These times were recorded on all stages and later used to compute time consumption (costs) for simulated samples in the completely measured trees (Table 1).

Field measurements, simulation of the RBS procedures and calculation of time consumption (costs) for simulated samples were accomplished with the program BRANCH (Cancino et al. 2002), the latter according to Eq. 12. BRANCH computes C based on all sampled units of a tree, the average time for each path by $c_0 = C/(nm)$ and c_2 by

$$c_2 = c_0 - \frac{c_1}{m} \tag{16}$$

7 Experimental Results and Discussion

All experimental comparisons of SWR and SWOR RBS are based on multistage sampling, the number of stages encountered depending on the tree under study. Both methods are identical (pps with sample size 1) on the third and all subsequent stages, differing only in the sample selection method for the primary units and in the second stage sample size m of the SWOR RBS ($m=1$ for SWR RBS). With both methods paths are sampled sequentially from node 1 onwards, but only with SWOR RBS repeated sampling of the same primary unit is avoided. As for the sampling frame they equally require measurement of the auxiliary variable x at all segments emanating from a node.

For the simulations, the stem was always excluded from the crown structure as exemplified in Fig. 2b, resulting in a large number N of primary segments and therefore allowing for a larger range of sample sizes n in the study. With the stem included, N and the related efforts to measure the auxiliary variable would be equally reduced for both RBS procedures. Additionally, the conditions would be even less favourable for conventional RBS because the probability of repeated sampling of the same primary unit would increase considerably.

In practice, stems are usually excluded when stratified RBS is applied to crown sections (strata) where e.g. whirls or groups of adjacent whirls are used as nodes. If the stem were used as a segment in such cases, the correlation between auxiliary and target variable would remarkably be reduced because of the usually extreme dimensions of the stem among the segments of the node and the relatively small total of the target variable in the paths above the stem segment. Only those branches above the stem segment which belong to the same stratum contribute to the total, instead of the entire crown partition above a stem segment in unstratified RBS. If unstratified RBS is used, inclusion of the stem leads to low precision of the estimator $\hat{F}$.

Of course, this is only a case study supposed to show that there are tree species and crown structures where SWOR RBS is superiour to conventional RBS, due to the observed variance components of the multistage setting.

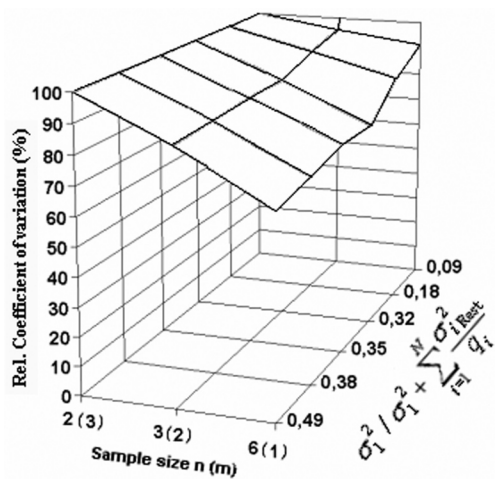

Fig. 3. Interpolated relative coefficients of variation of the biomass estimate from SWOR for different combinations of sample sizes ($nm=6$), and for 6 trees without stem and with different variance ratios (0.09–0.49). The value for $n=2$ and $m=3$ was considered 100% for each tree.

7.1 Optimum Allocation of Sample Sizes for SWOR RBS, if Time Is Not Considered

For 6 trees (Table 1, Norway spruces 1, 3, 4, and European mountain ashes 2, 3, 4) with increasing variance ratios

$$\sigma_1^2 \Big/ \left(\sigma_1^2 + \sum_{i=1}^{N} \frac{\sigma_{i\text{Rest}}^2}{q_i} \right) \tag{17}$$

from 0.09 to 0.49, the coefficients of variation were calculated based on Eq. 6 and with the stem removed from the crown structure, so that all segments emanating from the stem are attached to a virtual node 1, similar to Fig. 2b). The total number of 6 primary and secondary sampling units was realized in three different combinations of n and m. The empirical results, linearly interpolated, reveal the same tendency as it was found in 4.2 for SWR on both stages, as it could have been conjectured but was not generally proved for SWOR (Fig. 3). The precision of the SWOR RBS increases with decreasing m and is highest

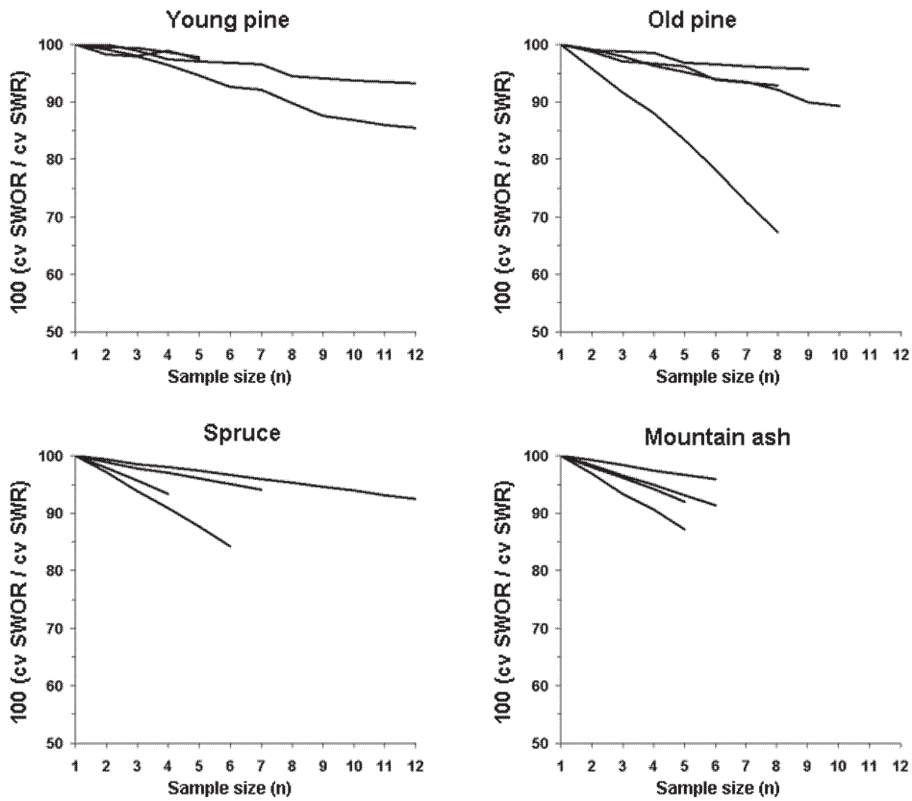

Fig. 4. Ratio of the coefficients of variation of SWOR (numerator) and classical RBS (denominator) with $m = 1$ and different primary sample sizes n; trees without stem, each line represents a tree (auxiliary variable: cross section).

for $m = 1$. Finally, the reduction of precision by larger second stage sample sizes is highest for the largest variance ratio. The coefficient of variation for $n = 6$ and $m = 1$ is below 70% of that for $n = 2$ and $m = 3$.

7.2 Relative Precision of SWOR and Classical RBS, if Time Is Not Considered

With trees without the stem and $m = 1$ for both procedures, the SWOR estimates were always more precise than the classical RBS estimates (Fig. 4). The larger the sample size, the larger is the gain in precision, independent of tree and species. But only for 5 out of 16 trees and larger sample sizes, it exceeds 10%.

Within the species, greater gains were achieved for the trees with weaker relationships between target variable and selection probabilities of the primary units, in accordance with the results of 4.1. The larger the variance ratio (Eq. 17) the greater is the increase of precision.

7.3 SWOR vs. Classical RBS Considering Time Consumption

The results presented in Fig. 5 were derived as mean time consumptions of 10 000 simulations of samples for each combination of n and m ($n = 1\ldots,6$; $m = 1,\ldots6$), based on the time study of the 9 European mountain ashes (see Chapter 6). According to Eq. 13 and that time study, the optimal second stage sample size is $m_0 = 1$ for the European mountain ash which is studied in

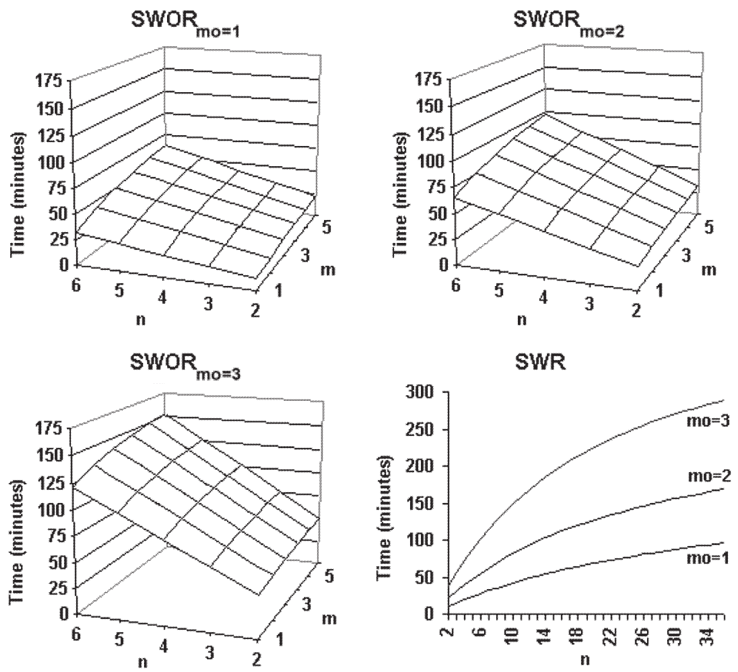

Fig. 5. Time consumption for the SWOR and the classical RBS with different sample sizes n and m and optimal secondary sample sizes $m_0 = 1,2,3$ for European mountain ash 2 without stem, according to Table 1.

Fig. 5. In this case, with $n \leq 6$ time consumption is slightly smaller for SWR than for SWOR RBS (compare lower-right and upper-left graph of Fig. 5). But this changes with increasing m. With $m = 2$ and $n = 6$ time consumption is clearly below that of SWR RBS with $n = 12$. Please, note that this is a purely time-related comparison of both methods for equal numbers of paths $n \cdot m$ (SWOR) and n (SWR) without considering the sampling errors achieved.

Additionally, the analysis was done assuming $m_0 = 2$ and $m_0 = 3$ yielding even higher advantages of SWOR RBS compared to the according time consumption for classical RBS. A larger m_0 is either the result of higher relative costs for first stage units, c_1, or of a smaller variance ratio (Eq. 17) in Eq. 13. Although it is quite logical that higher values of m_0 imply larger times (costs) for the primary units, Fig. 5 reveals that the increase of time consumption due to the enlargement of m_0 is larger with classical than with SWOR RBS. With increasing m_0 time consumption for classical RBS using $n = 36$ increases by a factor 3 from about 95 to about 285, whereas for SWOR RBS with $n = m = 6$ this factor is about 2.4. Classical RBS becomes less effective with higher stage 1 costs and with smaller variability among first stage units.

Total time consumption of SWOR RBS is an increasing linear function of the primary sample size n. This linearity is the result of SWOR, which ensures the repetition rate 1. SWR on the second and all following stages tolerates repeated selection of units causing a decreasing slope of time consumption with increasing secondary sample size m. The larger m, the larger is the repetition rate and the smaller is the contribution of each additional unit in the (nominal) sample size m to total time consumption (Fig. 5). With classical RBS, this behavior can be observed also for increasing first stage sample size n (Fig. 5).

Another effect of the enlargement of m_0, observed in Fig. 5 with SWOR RBS, is a change in the dependence between n and m for a given expenditure of time (e.g. 25 minutes for the European mountain ash). When m_0 becomes larger,

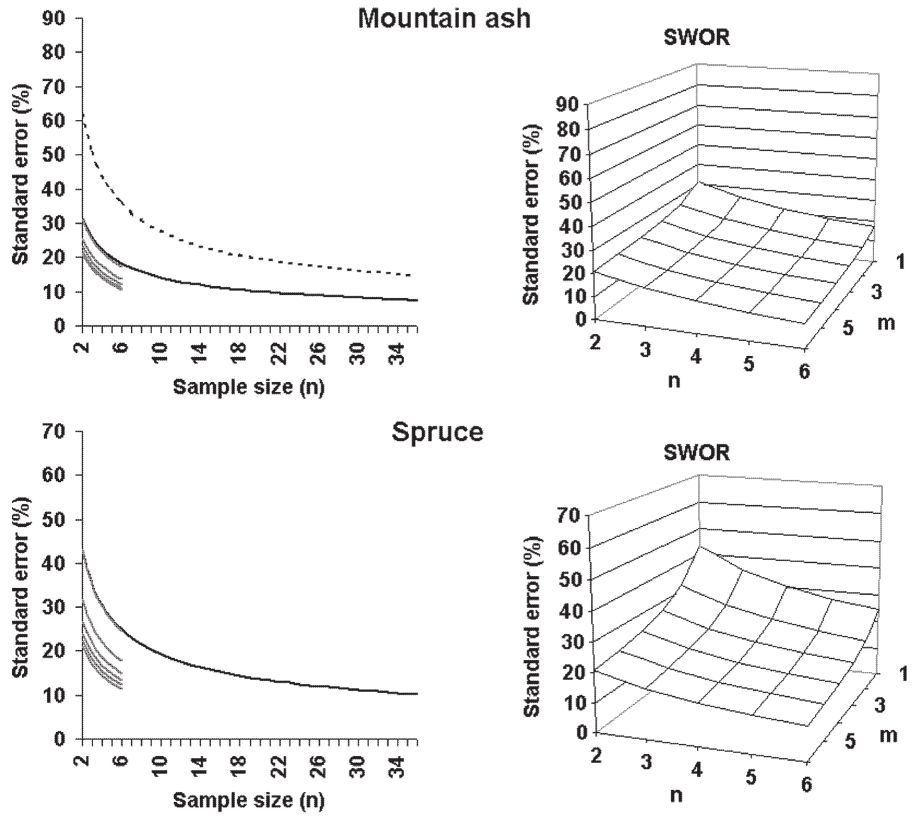

Fig. 6. Standard errors for SWOR and classical RBS with different sample sizes for trees without stem (right). The black lines on the left represent classical RBS without (continuous) and with stem (broken), the gray lines different secondary sample sizes for SWOR RBS ($m=1$: highest; $m=6$: lowest) and the trees without stem: European mountain ash 2, Norway spruce 3; auxiliary variable: cross section.

an increasing number of secondary units have the same effect on the expenditure of time as one primary unit.

A result which SWOR RBS could benefit from is depicted in Fig. 6. The decrease of the standard error achieved by sample size $m=2$ instead of $m=1$ is rather large. This holds for all Norway spruces and European mountain ashes. E. g. with $n=6$, that enlargement of the secondary sample size yielded a decrease of the standard error by 16.4% (spruce 4) up to 27.2% (spruce 3). For the European mountain ashes, the decrease varied between 20.9% (tree 2) and 23.4% (tree 3). The smaller the ratio of the variance between the primary units, the larger is the effect of m on the standard error.

Finally, we discuss the ratio of the times needed for the attainment of a certain precision (15% to 25%) by classical RBS ($m=1$) and by SWOR RBS (Fig. 7) with different sample sizes m. The according first stage sample sizes n are not reported but limited above by the total number of primary segments at the first node. The primary sample size n depends on the required precision for SWR RBS and additionally on m for SWOR RBS. Again, all results were obtained by 10 000 simulated samples for each realizable combination of n and m. The resulting standard errors for each m were interpolated by cubic splines over a range of n's in order to fix the sample sizes n and m necessary for the actually required precisions 15%, 20% and 25%. Thus n may be frac-

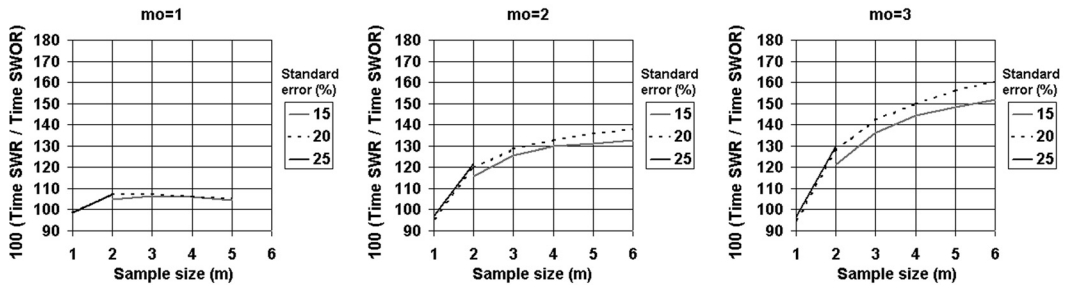

Fig. 7. Ratio of time consumption of the classical ($m=1$) to SWOR RBS with different sample sizes and required precisions (15%–25%) for European mountain ash 2 without stem, auxiliary variable: cross section.

tional what is unrealistic in practice but serves to avoid uncontinuities in the functional relationship between time ratio and m. The results for a tree with $m_0=1$, the most advantageous situation for classical RBS, are depicted in Fig. 7. The curves for different precisions have different ranges. A precision of 15% can not be achieved with $m<2$ in case of the SWOR RBS, because the primary sample size n is limited above by N. The time ratio for a required precision of 25% can be given only for $m\leq 2$, for larger m, precision would be higher than 25% with SWOR RBS.

With $m_0=1$, the optimal second stage sample size for SWR RBS in case of the tree under study, classical RBS is less time consuming than SWOR RBS only if the latter is also carried out with $m=1$. In this case, achievement of a predefined precision using classical RBS is not more than 2% cheaper. But with $m\leq 2$ SWOR RBS outperforms classical RBS by up to 8%.

The advantage is even more pronounced when the optimal secondary sample size m_0 is assumed to be larger, a situation occurring whenever costs per primary unit are larger or variance among primary units is relatively smaller (see Eq. 13). Thus, m_0 could be considered to be an indicator of the lack of efficacy of classical RBS. With $m_0=3$, for example, classical RBS ($m=1$) for the European mountain ashes was up to 60% more time consuming than SWOR RBS. The smaller the proportion of the variance among primary units, the greater is the advantage of SWOR.

The large decrease in the standard error caused by sample size $m=2$, as mentioned above, compared with $m=1$, in conjunction with the small amount of time taken for the inclusion of an additional secondary unit into the sample, implies a large advantage of the SWOR over classical RBS regarding costs. The value $m=2$ proved to be optimal for two of the trees in the analysis (European mountain ashes 2 and 3).

Working with larger trees than studied in this paper may require cutting off and ordering the branches (primary units) before the measurement can be done. This additional expenditure of time at the primary units would lead to an even larger m_0. Moreover, long distances between first stage units along the stem would also enlarge c_1 and thus m_0. Therefore, particularly for older trees, classical can certainly be expected to be more costly than SWOR RBS.

The efficiency of classical RBS is further reduced if trees are sampled with stem. Then the repetition rate is remarkably increased by the lower number of primary units at the first node, as well as by the usually extremely large selection probability of the stem segment compared with the adjacent branch segments.

8 Estimating the Variance of the SWOR RBS Estimator with $m = 1$

With second stage sample size $m=1$ there is no unbiased variance estimator, because the variance between the secondary units cannot be estimated. The three variance estimators proposed in Chapter 5 generally have a positive bias (Fig. 8), and the larger the sample size, the larger is that bias. The only exception is Norway spruce 3.

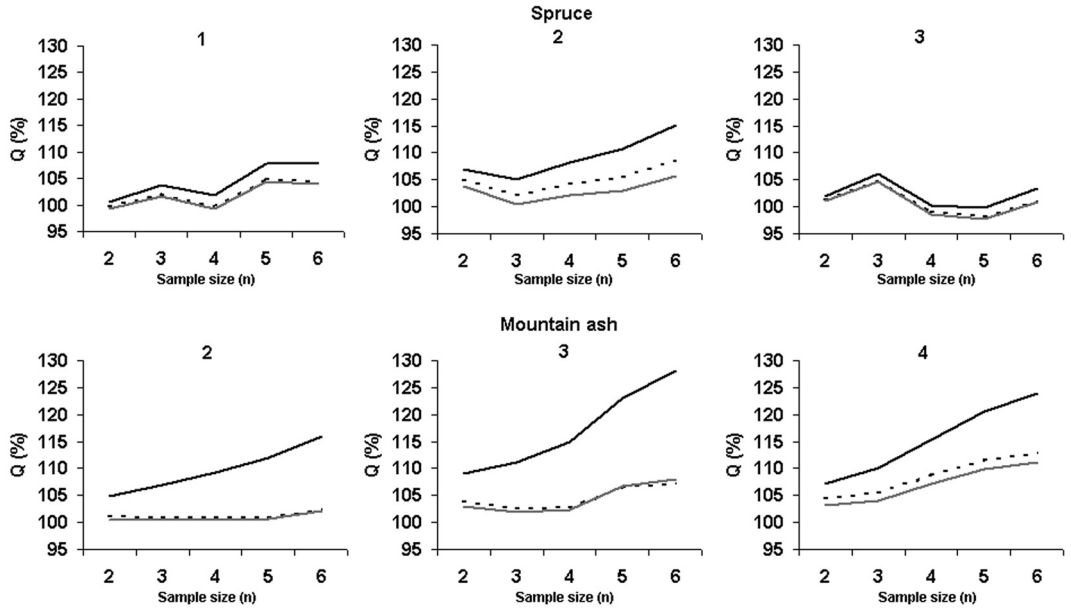

Fig. 8. Ratio of the average values of the variance estimates to the true variance of the estimate for SWOR on the first stage and $m = 1$ on stage 2; trees without stem. Continuous, dark line: classical RBS variance estimator; broken, dark line: classical RBS variance estimator with correction (Eq. 14); gray line: Sen-Yates-Grundy-like estimator (Eq. 15); auxiliary variable: cross section.

The variance estimated by the formula of the classical RBS was most biased and exceeded the true variance by almost 30% (see European mountain ash 3, Fig. 8). For this tree, however, SWOR at the first stage yielded clearly more precise estimates of the total of the target variable, with a variance 10% less than that of classical RBS. The increment in bias with increasing sample size is a result that could have been expected, because efficiency of SWOR in comparison with SWR also increases with the sample size, the same tendency as it can be observed for the bias of the classical RBS variance estimator.

The corrected classical variance estimator is much less biased, the largest observed bias being below 15%. However, some neglectable under-estimations arose.

The Sen-Yates-Grundy-like estimator achieved the best results amongst the three alternative variance estimators. In the worst case, the bias hardly exceeds 10% (see European mountain ash 4, Fig. 8) and usually it is below 5%. As with the corrected estimator, some little under-estimations are to be mentioned.

The three variance estimators yielded the worst results when both the relationship between the target variable and the unconditional selection probabilities of the segments is weak and the ratio of the variance between the primary units to the entire variance is large; i.e. the bias of the alternative variance estimators was largest in those cases in which the strongest gain in precision of the SWOR can be observed.

Acknowledgement

The authors wish to thank the referees, in particular Timothy G. Gregoire, for many helpful comments to the original manuscript.

References

Brewer, K.R.W. & Hanif, M. 1970. Durbin's new multistage variance estimator. Journal of the Royal Statistical Society, Series B, 32: 302–311.

Cancino, J. 2003. Analyse und praktische Umsetzung unterschiedlicher Methoden des Randomized Branch Sampling. Diss. Fakultät für Forstwissenschaften und Waldökologie der Georg-August-Universität Göttingen. 191 p. Available from: http://webdoc.sub.gwdg.de/diss/2003/cancino/index.html.

— , Gockel, S. & Saborowski, J. 2002. Randomized Branch Sampling – Varianten, Programm Branch und erste Analysen. Deutscher Verband Forstlicher Forschungsanstalten, Sektion Forstliche Biometrie und Informatik, 14. Tagung, Tharandt, 3.–5. April 2002. p. 76–87.

Cochran, W.G. 1977. Sampling techniques. Wiley, New York.

Durbin, J. 1953. Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society, Series B, 15: 262–269.

— 1967. Design of multi-stage surveys for the estimation of sampling errors. Applied Statistics 16: 152–164.

Fellegi, I.P. 1963. Sampling with varying probabilities without replacement. Rotating and non-rotating samples. Journal of the American Statistical Association 58: 183–201.

Gaffrey, D. & Saborowski, J. 1999. RBS, ein mehrstufiges Inventurverfahren zur Schätzung von Baummerkmalen. I. Schätzung von Nadel- und Asttrockenmassen bei 66-jährigen Douglasien. Allgemeine Forst- und Jagdzeitung 170: 177–183.

Good, M., Paterson, M., Brack, C. & Mengersen, K. 2001. Estimating tree component biomass using variable probability sampling methods. Journal of Agricultural, Biological and Environmental Statistics 6: 258–267.

Gregoire, T.G., Valentine, H.T. & Furnival, G.M. 1995. Sampling methods to estimate foliage and other characteristics of individual trees. Ecology 76: 1181–1194.

Grosenbaugh, L.R. 1967. The gains from sample-tree selection with unequal probabilities. Journal of Forestry 65: 202–206.

Grundy, P.M. 1954. A method of sampling with probability exactly proportional to size. Journal of the Royal Statistical Society, Series B, 16: 236–238.

Hartley, H.O. & Rao, J.N.K. 1962. Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33: 350–374.

Jessen, R.J. 1955. Determining the fruit count on a tree by randomized branch sampling. Biometrics 11: 99–109.

Mundson, A., Smith, K., Horvath, R., Ruel, J.-C., Ung, C.-H. & Bernier, P. 1999. Does harvesting (CPRS) mimic fire? Verifying for black spruce forests in central Québec. Project Report 1999-11. 16 p.

Parresol, B.R. 1999. Assessing tree and stand biomass. A review with examples and critical comparisons Forest Science 45: 573–593.

Rao, J.N.K. 1975. Unbiased variance estimation for multistage designs. Sankhya C 37: 133–139.

Raulier, F., Bernier, P. & Ung, C.-H. 1999. Canopy photosynthesis of sugar maple (Acer saccharum) comparing big-leaf and multilayer extrapolations of leaf-level measurements. Tree Physiology 19: 407–420.

Saborowski, J. 1990. Schätzung von Varianzen und Konfidenzintervallen aus mehrstufigen Stichproben. Schriften aus der Forstlichen Fakultät der Universität Göttingen und der Niedersächsischen Forstlichen Versuchsanstalt 99.

— & Gaffrey, D. 1999. RBS, ein mehrstufiges Inventurverfahren zur Schätzung von Baummerkmalen. II. Modifizierte RBS-Verfahren. Allgemeine Forst- und Jagdzeitung 170: 223–227.

Sampford, M.R. 1967. On sampling without replacement with unequal probabilities of selection. Biometrika 54: 499–513.

Snowdon, P., Raison, J., Keith, H., Montagu, K., Bi, H., Ritson, P., Grierson, P., Adams, M., Burrows, W. & Eamus, D. 2001. Protocol for sampling tree and stand biomass. National Carbon Accounting System Technical Report 31. 114 p.

Stevens, W.L. 1958. Sampling without replacement with probability proportional to size. Journal of the Royal Statistical Society, Series B, 20: 393–397.

Valentine, H.T., Baldwin, Jr. V.C., Gregoire, T.G. & Burkhart, H.E. 1994. Surrogates for foliar dry matter in loblolly pine. Forest Science 40: 576–585.

— , Tritton, L.M. & Furnival, G.M. 1984. Subsampling trees for biomass, volume, or mineral content. Forest Science 30: 673–681.

Williams, R.A. 1989. Use of randomized branch and

importance sampling to estimate loblolly pine biomass. Southern Journal of Applied Forestry 13: 181–184.

Xiao, Q., McPherson, G., Ustin, S., Grismer, M. & Simpson, J. 2000. Winter rainfall interception by two mature open-grown trees in Davis, California. Hydrological Processes 14: 763–784.

Yates, F. & Grundy, P.M. 1953. Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society, Series B, 15: 253–261.

Total of 28 references

Appendix

It is sufficient to prove the equivalence of formula (Eq. 2') and Jessen's variance formula (Jessen 1955), which is given there for one path ($n=1$), for a two stage RBS procedure. In our notation and with $F_0 := F - f$, his formula turns out to be

$$Var\ \hat{F} = \sum_{i=1}^{N}\sum_{j=1}^{M_i} Q_{ij}\left(f + \hat{F}_i - F\right)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i} Q_{ij}\left(\frac{f_i}{q_i} + \frac{f_{ij}}{q_i q_{ij}} - F_0\right)^2$$

where the summation is over all $\sum_{i=1}^{N} M_i$ paths. Calculating the square and considering $Q_{ij} = q_i q_{ij}$, one obtains

$$Var\ \hat{F} = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\frac{f_i^2}{q_i} q_{ij} + \frac{f_{ij}^2}{q_i q_{ij}} + 2\frac{f_i}{q_i} f_{ij} - 2F_0\left(f_i q_{ij} + f_{ij}\right) + q_i q_{ij} F_0^2\right)$$

Here we use $\sum_{i=1}^{N} q_i = \sum_{j=1}^{M_i} q_{ij} = 1$ and $F_0 = f_\bullet + f_{\bullet\bullet}$, with the abbreviations $f_\bullet = \sum_{i=1}^{N} f_i$ and $f_{\bullet\bullet} = \sum_{i=1}^{N} f_{i\bullet} = \sum_{i=1}^{N}\sum_{j=1}^{M_i} f_{ij}$, so that

$$Var\ \hat{F} = \sum_{i=1}^{N}\frac{f_i^2}{q_i} + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{f_{ij}^2}{q_i q_{ij}} + 2\sum_{i=1}^{N}\frac{f_i f_{i\bullet}}{q_i} - F_0^2$$

The two stage version of the variance (Eq. 2'), with $F_{ij} = f_{ij}$ and $F_i = f_i + f_{i\bullet}$, can be expanded according to

$$\begin{aligned} Var\ \hat{F} &= \sum_{i=1}^{N} q_i\left(\frac{F_i}{q_i} - F_0\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{q_{ij}}{q_i}\left(\frac{f_{ij}}{q_{ij}} - f_{i\bullet}\right)^2 \\ &= \sum_{i=1}^{N}\left(\frac{F_i^2}{q_i} - 2F_0 F_i + q_i F_0^2\right) + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\frac{f_{ij}^2}{q_i q_{ij}} - 2\frac{f_{i\bullet}}{q_i} f_{ij} + \frac{q_{ij}}{q_i} f_{i\bullet}^2\right) \\ &\quad \left(\text{use} \quad \sum_{i=1}^{N} F_i = F_0 \quad \text{and} \quad \sum_{i=1}^{N} q_i = \sum_{j=1}^{M_i} q_{ij} = 1\right) \\ &= \sum_{i=1}^{N}\frac{f_i^2 + 2 f_i f_{i\bullet}}{q_i} - F_0^2 + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{f_{ij}^2}{q_i q_{ij}} \end{aligned}$$

This completes the proof of the equivalence.