West *New Zealand Journal of Forestry Science* (2017) 47:22
DOI 10.1186/s40490-017-0101-7

New Zealand Journal of Forestry Science

RESEARCH ARTICLE Open Access

Simulation studies to examine bias and precision of some estimators that use auxiliary information in design-based sampling in forest inventory

P. W. West

Abstract

Background: Various double sampling methods using both target and auxiliary variables have been developed over many years for use in natural resource inventory.

Methods: Simulations of inventory were carried out using four different ratio estimators and model-assisted estimation in each of five rather different example forest populations. Estimates of population means and their standard errors from each of these methods were compared with those obtained using simple random sampling.

Results: With all five double sampling estimators, bias in estimates of means and standard errors (the latter estimated analytically or through bootstrapping) was generally small and consistent with theoretical expectations. Their efficiency increased as either the first- or second-phase sample sizes increased. All were more efficient than estimates obtained using simple random sampling as long as there was some positive level of correlation between the target and auxiliary variable. However, none of the double sampling estimators was more efficient than any of the others.

Conclusions: For many forest inventory tasks, users may well be able to use whichever of the estimators is most convenient to their purpose. However, model-assisted estimation has application in a wider range of circumstances than the other methods, which perhaps recommends it for general use.

Keywords: Inventory, Design-based sampling, Ratio estimators, Model-assisted estimation

Background

Modern forest inventory aims to inform both long- and short-term objectives of forest management, objectives that can be as varied as estimation of carbon sequestration, determination of sustainable wood supply, maintenance of biodiversity or estimation of the level of forest destruction or rehabilitation (Corona et al. 2003; Wulder et al. 2004; Köhl et al. 2006; McRoberts and Tomppo 2007; Falkowski et al. 2009). Broadly speaking, forest inventory either maps how some forest characteristic (the target variable) varies across a forested area or estimates the total or mean value of that variable across the area. To assess the level of confidence that one may have in the result, it is usual also to estimate the level of precision of such estimates (Paré et al. 2016), commonly expressed as a confidence limit. Numerous texts describe forest inventory methods (Schreuder et al. 1993; Kangas and Maltamo 2006; Köhl et al. 2006; Gregoire and Valentine 2008; Mandallaz 2008) and several reviews have summarised recent developments in the discipline (McRoberts et al. 2010; Mandallaz 2013; Corona 2016; Ståhl et al. 2016).

Forest target variables can be difficult, time consuming and expensive to measure on the ground. Typical examples include forest biomass or volume of wood products available for harvest. However, it is common to have available one or more auxiliary variables that are correlated, at least to some extent, with the target variable and can be measured relatively easily. Stand basal area might be such a variable for the two examples mentioned above. Such variables can be used very effectively

Correspondence: pwest@nor.com.au
Forest Research Centre, School of Environment, Science and Engineering, Southern Cross University, Lismore, NSW 2480, Australia

to reduce the time and effort involved in undertaking a forest inventory.

Recently, much research effort has revolved around the use of newly available airborne or satellite remote sensing technologies that provide values of one or more auxiliary variables (Köhl et al. 2006; Falkowski et al. 2009); in particular, laser imagery has received much attention for forest inventory. These technologies can provide imagery at scales as small as metres or tens of metres across large parts of entire forest regions or, indeed, entire countries. Using ground-based measurements of the target variable and one or more auxiliary variables obtained from remote imagery of the measurement sites, research then develops a working model system to predict the target variable from the auxiliary variable(s). Using the remote imagery from across the entire region, that model is then applied to estimate and map the target variable across the region and/or provide estimates of the total and mean of the target variable over the region. These processes are often termed 'wall-to-wall' inventory and are used to obtain broad-scale, strategic information; recently, they have been much used to estimate carbon sequestration capability of forests as part of climate change research (e.g. Neigh et al. 2013). There are many examples of the application of this remotely sensed imagery for forest inventory both with estimation of the precision of estimates, in one form or another, (Cohen et al. 2013; Tomppo et al. 2014; Kangas et al. 2016; Ringvall et al. 2016) and without precision estimation (Du et al. 2014; Ometto et al. 2014; Waser et al. 2015; Clerici et al. 2016; Immitzer et al. 2016).

There is also a long history of forest inventory done at other scales, often down to compartment level or smaller, for tactical or operational management purposes or for other reasons (Ahamed et al. 2011; Corona et al. 2014; Mandallaz 2013; Melville et al. 2015; McRoberts et al. 2006, 2016). Such inventory generally uses 'design-based' sampling which involves selection from the forest population under consideration of a random sample. This is done using some sampling design such that the probability of inclusion of any sampling unit from the population is known. Design-based sampling may be aided also through use of one or more auxiliary variables through which the statistical efficiency of the inventory may be improved, that is, the precision of the estimates increased (Gregoire and Valentine 2008, Chap. 6). Of course, if no suitable auxiliary variable(s) can be identified, design-based inventory can only use simple random sampling.

One way to use auxiliary variables with design-based sampling is to undertake double sampling, where the auxiliary variable is measured on a first-phase sample and the target variable is measured on a smaller subsample in a second phase; the sampling design can differ between the first and second phase. This approach aims to minimise the time, effort and cost involved with sampling whilst achieving the level of precision required in the estimate of the population mean or total. Further, it may be anticipated that the higher is the level of correlation between the target and auxiliary variable, the higher will be that precision.

The estimates of the population mean from such double samples have often been determined using one of a number of what are termed 'ratio' estimators. These are used in many areas of research, including economics (Knottnerus 2011), medicine (Al-Omari and Bouza 2015), agriculture (Francis et al. 1979; Reich et al. 1993), land- use assessment (Li et al. 2014) and animal ecology (Stevenson 1979; Neilson et al. 2013). Their principles are described by Cochran (1977, Chap. 6) and theoreticians continue to develop them for particular circumstances (e.g. Oral and Oral 2011; Lin and Chao 2014; Magnussen et al. 2014; Al-Omari and Bouza 2015; Weiskittel et al. 2015; Kumar and Chhaparwal1 2016; Ringvall et al. 2016). Another method that is being used increasingly is 'model-assisted' estimation; Baffetta et al. (2009) offer a formal and generalised description of its approach whilst Kangas et al. (2016) offer a recent example of its use.

Many of the commonly used double sampling estimators use sample data that have a similar form, although somewhat different sampling methods may have been used to obtain them. To the practitioner of forest inventory dealing with any particular forest population, it is often far from obvious as to which of the sampling methods and estimators available should be used for that population to provide the 'best' inventory result. By 'best' here is meant a result that, for the sampling effort that the resources available can afford, produces an unbiased estimate of the population mean and its standard error and which has the minimum possible estimate of the standard error and the corresponding confidence limit.

The present work aims to address this issue by comparing estimates of population means and their standard errors obtained using each of model-assisted estimation and a number of the more commonly used ratio estimators, when these are applied with double sampling in forest inventory. Further, these results are compared with those obtained using simple random sampling so that the advantage gained with the double sampling estimators may be assessed. These comparisons are done through simulation studies using five example forest populations that cover a range of rather different forest types and target variables. As well as comparing sampling methods, also examined are the effects on estimation efficiency both of changing the level of correlation between the target and auxiliary variable and of changing the sizes of the first- or second-phase samples.

Methods

Estimators

Numerous authors have described in formal detail the various design-based sampling techniques and estimators used in forest inventory (Schreuder et al. 1993; Kangas and Maltamo 2006; Gregoire and Valentine 2008; Mandallaz 2008; Ståhl et al. 2016); these texts informed the approach and the estimators considered here.

Following Schreuder et al. (1987), many double sampling estimators assume that, over the population as a whole, a linear regression model relates a target variable of interest (Y) to an auxiliary variable (X) as

$$Y = \alpha + \beta X + \epsilon \quad , \tag{1}$$

where α and β are parameters and the error term (ϵ) has variance $V(\epsilon)$, given by

$$V(\epsilon) = \phi^2 X^g \quad , \tag{2}$$

where ϕ^2 is a variance and g is a constant that takes a value ≥0.

Many commonly used ratio estimators assume that the Y–X relationship in the population passes through the origin, that is, $\alpha = 0$ in Eq. (1). Ratio estimators that make that assumption will be considered in the present work. For such estimators, if the Y–X relationship is not linear, it may be possible to make it so through data transformation. If it does not pass through the origin, it may be possible to transform the data to make it so; that possibility is considered later in the present work.

In the case of model-assisted estimation, other models may be used that employ the full power of linear or non-linear regression analysis with one or more auxiliary variables as appropriate. For the present work, Eq. (1), with the intercept α retained, will be used for model-assisted estimation as it is both simple and commonly used; in a forest inventory context, Mandallaz (2013) employed a rather more complex case of model-assisted estimation where the intercept was retained also.

Often too, it is assumed that the Y–X relationship is homoscedastic, that is, $g = 0$ in Eq. (2); as it is a common circumstance, this will be assumed to be the case in the present work for both ratio estimators and model-assisted estimation. Särndal et al. (1992, Sects. 7.3, 7.4) considered this issue of hetereoscedasticity for one of the ratio estimators used here (the mean of ratio estimator—see later). Certainly, it would require that weighted least-squares regression be applied for model-assisted estimation.

Further work beyond the present would be necessary to examine the effects of variations to any of the various model assumptions made above.

Suppose the objective of an inventory is to estimate the mean, $\overline{Y}$, of the target variable Y in a population that consists of a total of N sampling units and where an auxiliary variable X is available. In the context of forest inventory, the sampling units might be individual trees, or fixed area plots, or points at which variable radius plot sampling (known also as Bitterlich sampling) is done (Gregoire and Valentine 2008 Chaps. 7, 8; West 2016). If the sampling units are individual trees, target and/or auxiliary variable values will be obtained as measurements of some characteristic of the individual trees. If the sampling units are fixed area or variable radius plots, the measurements will be values for whole plots derived from the trees within any one plot.

Suppose a double sample is taken from the population. Assume the first-phase sample is either a simple random sample of size f ($<N$) or it may involve a complete enumeration of the entire population ($f = N$) as in 'wall-to-wall' inventory. Assume the values of the auxiliary variable are measured on the first-phase sampling units and denote them as x_i^f ($i = 1, 2, \ldots, f$). Suppose that from this first-phase sample, a second-phase sub-sample of size n ($<f$) is selected, by methods considered later, on which the target variable is measured and denoted as y_i ($i = 1, 2, \ldots, n$), with the corresponding (measured in the first-phase) auxiliary variable values denoted as x_i ($i = 1, 2, \ldots, n$).

A number of sampling methods and estimators that were chosen for the present work and that might be used under these circumstances to provide estimates of the population mean are listed in Table 1. The first is simple random sampling (Eq. 3.2); note that there is no (Eq. 3.1) for reasons outlined below. Excepting model-assisted estimation (Eqs. 6.1 and 6.2), all the others are ratio estimators. The ratio of means (Eq. 4.1) (Cochran 1977, Sects. 6.2–6.4), mean of ratios (Eq. 5.1) (Hartley and Ross 1954) and probability proportional to size (PPS) sampling (Eq. 7.1) (Cochran 1977, Sect. 9A.2; Särndal et al. 1992, Sect. 3.6; Schreuder et al. 1993, Sect. 3.2.3; Gregoire and Valentine 2008, Sect. 3.3.1) estimators have been in long use in a wide variety of sampling contexts.

The ratio estimator Eq. (8.1) is perhaps less well known outside forest inventory circles. It was developed for particular use in forest inventory by Grosenbaugh (1964, 1965, 1976) and is used to obtain estimates with high precision of amounts of wood available for harvest from smallish forest tracts, usually of only tens of hectares; Gregoire and Valentine (2008, p. 382 et seq.) give a modern summary of the method. Various examples of its use have been reported (Johnson et al. 1967; Johnson and Hartman 1972; Stevenson 1979; Ringvall and Kruys 2005) and various developments to it have been made (Williams and Schreuder 1998; Gregoire and Valentine 1999; Bondesson and Thorburn 2008; Magnussen 2001; Grafström 2010). In its original form, it was known as sampling with probability proportional to prediction,

Table 1 The sampling methods and estimators considered here

Method/estimator	Equation		Equation
One-phase sampling			
Simple random sampling**		$\overline{Y} = (\Sigma_n\ y_i)/n$ $\sigma(\overline{Y}) = \left\{\Sigma_n(y_i-\overline{Y})^2/[n(n-1)]\right\}^{1/2}$	(3.2)
Two-phase (double) sampling			
	Complete enumeration of N sampling units as first-phase	First-phase simple random sample of size = f ($<N$) **	
Simple random sampling in second-phase **			
Ratio of means	$\overline{Y} = \left[(\Sigma_N\ x^f{}_i)/N\right]\overline{R}$ $\overline{R} = (\Sigma_n\ y_i)/(\Sigma_n\ x_i)$ $\sigma(\overline{Y}) = \left\{[1-n/N]\ \left[\Sigma_n(y_i-\overline{R}x_i)^2\right]/[n(n-1)]\right\}^{1/2}$ (Cochran 1977, Eqs. 6.1 and 6.9)	$\overline{Y} = \left[(\Sigma_f x^f{}_i)/f\right]\overline{R}^{\ddagger}$ $\overline{R} = (\Sigma_n\ y_i)/(\Sigma_n\ x_i)$	(4.2)
Mean of ratios	$\overline{Y} = \left[(\Sigma_N x^f{}_i)/N\right]\overline{R} + (N-1)\left[(\Sigma_n\ y_i)-(\Sigma_n\ x_i)\overline{R}\right]/[N(n-1)]$ $\overline{R} = (\Sigma_n\ y_i/x_i)/n$ (Hartley and Ross 1954)	$\overline{Y} = \left[(\Sigma_f x^f{}_i)/f\right]\overline{R} + (f-1)\left[(\Sigma_n\ y_i)-(\Sigma_n\ x_i)\overline{R}\right]/[f(n-1)]$ $\overline{R} = (\Sigma_n\ y_i/x_i)/n$	(5.2)
Model-assisted	$\overline{Y} = \left[(\Sigma_n\ y_i) + (\Sigma_N y^f{}_i) - (\Sigma_n\ y_i)\right]/N$ where $y^f{}_i = \alpha + \beta x^f{}_i,\ y_i = \alpha + \beta x_i$ (Ståhl et al. 2016)	$\overline{Y} = \left[(\Sigma_n\ y_i) + (\Sigma_f y^f{}_i) - (\Sigma_n\ y_i)\right]/f$ where $y^f{}_i = \alpha + \beta x^f{}_i,\ y_i = \alpha + \beta x_i$	(6.2)
Sampling with probability proportional to size in second-phase			
Probability proportional to size (PPS) sampling	$\overline{Y} = \left[(\Sigma_N\ x^f{}_i)/N\right]\overline{R}$ $\overline{R} = (\Sigma_n\ y_i/x_i)/n$ $\sigma(\overline{Y}) = \left[(\Sigma_N x^f{}_i)/n\right]\ \left\{[N-n]\left[\Sigma_{n\ j,k}\left(y_j/x_j-y_k/x_k\right)^2\right]/\left[2N^3(n-1)\right]\right\}^{1/2}$ (Schreuder et al. 1993, Eqs. 3.7, 3.9)		
Quick probability proportional to size (QPPS) sampling	$\overline{Y} = \left[(\Sigma_N\ x^f{}_i)/N\right]\overline{R}^{\dagger}$ $\overline{R} = (\Sigma_n\ y_i/x_i)/n$ (Grosenbaugh 1965, Eq. 3PSEVENTH; Furnival et al. 1987, Eq. 9; West 2011, Eq. 11)	$\overline{Y} = \left[(\Sigma_f\ x^f{}_i)/f\right]\overline{R}^{\dagger\dagger}$ $\overline{R} = (\Sigma_n\ y_i/x_i)/n$ (Adapted from West 2011, Eq. 9; 2017, Eq. 1)	(8.2)

Notation: $\overline{y}$ estimate of the population mean of target variable, $\hat{\sigma}(\overline{y})$) estimate of the standard error of the estimate of the population mean; bootstrapping was used where no analytical estimator is shown, N population size, f first-phase sample size ($\leq N$), x^f_i auxiliary variable value measured in the ith member of a first-phase sample, n second-phase sample size ($<f$), y_i, x_i target and auxiliary variable values, respectively, measured in the ith member of a second-phase sample (the auxiliary variable value will have been measured already as part of the first-phase sample), α, β intercept and slope, respectively, of the ordinary least-squares straight-line fit between the target and auxiliary variable values in a second-phase sample, as defined by Eq. (1)

**All first-phase simple random sampling was done with replacement in this work

‡At their Eq. (6.95), Gregoire and Valentine (2008) show a modified version of this that can be written as $\overline{Y} = \{\overline{R}\left[(\Sigma_f\ x^f_i)-(\Sigma_n\ x_i)\right] + [(\Sigma_n\ y_i)]\}/f$, together with an approximate analytical standard error estimator. The form here was adapted from the commonly used estimator shown by Cochran (1977, Eq. 6.1)

†Originally called sampling with probability proportional to prediction (3P sampling) by its inventor L.R. Grosenbaugh, as discussed in the text

††A method developed as an extension of 3P sampling by West (2011), renamed QPPS sampling by West (2017). In applying this method in the present work it was assumed that the minimum auxiliary variable value in the population was zero, as done originally by Grosenbaugh in 3P sampling. That constraint was not applied by West (2011, 2017), but was found here to yield more reliable estimates of the standard error of the estimate of the population mean

referred to commonly as 3P sampling. West (2011, 2017) developed its approach for more general use and has termed it quick probability proportional to size (QPPS) sampling, the term applied generally to the approach in Table 1 (Eqs. 8.1 and 8.2).

The ratio estimators of Table 1 all assume that the Y–X relationship over the population as a whole passes through the origin. Of course this is not necessarily the case when any individual sample is chosen from the population and it may be that this leads to problems with the estimation of the population mean; the present simulation studies aimed to examine if this was a problem. In the case of model-assisted estimation (Eqs. 6.1 and 6.2), the model used allows that the Y–X relationship over the population may or may not pass through the origin; the implications of this are discussed later.

In the second and third columns of Table 1, the double sampling estimators have been divided into two groups. In the second column (Eqs. 4.1–8.1), it has been assumed that the first-phase 'sample' involved measurement of the auxiliary variable on each and every sampling unit in the entire population. Of course, this is a not a sample as such but is complete enumeration of the auxiliary variable across the population. Theoreticians have often presented these estimators only in this form (Hartley and Ross 1954; Cochran 1977, Eqs. 6.1 and 6.9; Grosenbaugh 1965, Eq. 3PSEVENTH; Schreuder et al. 1993, Eqs. 3.7 and 3.9). Such complete enumeration is appropriate in practice when the population is not too large or 'wall-to-wall' inventory is being undertaken. However, as shown in the third column (Eqs. 3.2–8.2), all the double sampling estimators except PPS (Eq. 7.1) may be formulated assuming that the first-phase sample is a simple random sample that is smaller, usually much smaller, than the whole population; in these cases, the population size (N) is not necessarily known. There can be difficulties with selecting a simple random sample from a large population of unknown size, but methods are available to do so for most forestry circumstances (West 2016).

The various methods considered here have been subdivided also in the rows of Table 1 depending on the nature of the sampling involved. Firstly, there is simple random sampling (Eq. 3.2) that involves only a single sample on which the target variable only is measured. All the remaining estimators involve double sampling with either simple random sampling or complete enumeration of the auxiliary variable in the first-phase. The second to fourth rows (Eqs. 4.1–6.2) involve simple random sampling also in the second phase. The fifth and sixth rows (Eqs. 7.1–8.2) involve sampling with probability proportional to size in the second phase, size being determined by the auxiliary variable value of a sampling unit. The case in the fifth row, PPS sampling (Eq. 7.1), is a long recognised and standard form of probability proportional to size sampling (Cochran 1977, Sect. 9A.2). QPPS sampling (Eqs. 8.1, 8.2) uses a form of probability proportional to size sampling based on work of Lahiri (1951). This allows a decision to be made to include or not a sampling unit in the second-phase sample immediately the auxiliary variable value has been measured on that sampling unit; that is how the term 'quick' in the name of these methods was derived. Where field sampling is required for both samples in both phases, Lahiri's method removes the need to revisit sampling units to select the second-phase sample and then measure the target variable on it. However, QPPS sampling does require that a preliminary survey of the population be undertaken to determine the range of values of the auxiliary variable that will be encountered anywhere in the population.

For several of the estimators in Table 1, an analytical estimator of the standard error of the estimate of the population mean, $\hat{\sigma}\left(\overline{Y}\right)$, is well established and these are shown in the table. In all other cases, the standard error was estimated using bootstrapping as described below; research always continues to develop analytical estimators (e.g. Mandallaz 2013), but bootstrapping was applied here for consistency across cases where analytical estimators are less well known or not established.

Simulated populations

Forest inventory simulations were conducted with five example populations each containing 10,000 sampling units and each with a different target variable. All five were based on real forest populations. The first two example target variables were stand basal area and stand stocking density in a primary, closed canopy rainforest growing in a warm temperate climate on Yakushima Island in southern Japan (Kohyama 1986). The target variables in the third and fourth examples were individual tree stem diameters over bark at 1.3 m above ground in each of two 11-year-old plantation forests, one of *Pinus radiata* D.Don in temperate south-eastern South Australia and the other of a mixture of Australian rainforest species in subtropical north-eastern New South Wales, Australia; data collected from those forests were available to the author. The fifth example target variable was stand stem wood volume of sawlog (volume of logs of a size suitable for sawmilling) in native regrowth eucalypt forest in temperate north-eastern Victoria, Australia (Hamilton and Brack 1999). Examples 1, 2 and 5 represent forest populations in which the sampling units are fixed area or variable radius sample plots. Examples 3 and 4 represent populations in which the sampling units are individual trees.

These examples were used also by West (2017) to investigate various properties of QPPS sampling; additional

details about the forests concerned are given there. As shown in his Fig. 1, the frequency distributions of the five target variables differed widely in shape, from quite symmetrical through to having a marked skew to the right or left. The examples were chosen deliberately because of this wide variation in population structure, typical of the variation encountered in different forest populations.

To produce a simulated population data set for any one of the five examples, a set of 10,000 auxiliary variable values were generated, based on the frequency distribution of that example. A set of 10,000 corresponding target variable values were then generated by adding random normal deviates to these auxiliary values using a standard deviation chosen by trial and error to give a particular level of correlation between the target and auxiliary variables. For each of the five examples, nine such target variable data sets were generated with correlation levels chosen to be close to values of 0.1, 0.2, ..., 0.9. Scatter plots of target against auxiliary variables for the first 500 of the 10,000 values generated for one of the examples for each of two correlation levels are shown in Fig. 1. As is evident there, the data were generated so that the target and auxiliary variables bore a straight-line relationship to each other that, over the whole population, passed through the origin. It was assumed also that the target-auxiliary variable relationship was homoscedastic as mentioned in the discussion following Eqs. (1) and (2).

Simulation of inventory

Simulations of inventory were carried out to examine differences arising from all possible combinations of the five examples, the 10 methods and estimators (Table 1), the nine levels of correlation between the target and auxiliary variables and from three different second-phase sample sizes, that is, $5 \times 10 \times 9 \times 3 = 1350$ possibilities. For any one of these 1350 possibilities, 5000 samples were drawn from the example population concerned, using random sampling techniques appropriate to the method or estimator being considered, to yield 5000 estimates of the population mean, $\overline{Y}$, and its standard error, $\hat{\sigma}\left(\overline{Y}\right)$. Thus, the results reported here derive from a total of $1350 \times 5000 = 6¾$ million simulated samples.

Three different second-phase sample sizes were considered because it can be expected that the precision of estimate of the population mean will increase as the sample size increases, that is, its standard error will decrease. Three arbitrarily chosen second-phase sample sizes (n) were tested, 10, 40 and 100. For estimators (4.1–8.1), the first-phase 'sample' included the auxiliary variable values of the entire 10,000 sampling units of a population. For estimators (4.2–8.2), the corresponding first-phase sample sizes (f) were 25, 95 and 240, respectively, for examples 1 and 3–5 and 30, 290, 120 respectively for example 2; experience by the author with these examples had suggested that these might be appropriate first-phase sample sizes. In normal use of QPPS (Eqs. 8.1 and 8.2), the second-phase sampling process employed makes it impossible to ensure a particular second-phase sample size is achieved. However, in these simulations the auxiliary variable values were in fact known in all of the sampling units in the population. This meant it was possible to select a given number of sampling units that conformed to the selection criteria for a second-phase sample for QPPS sampling and a given number that did not conform to those criteria to complete both first- and second-phase samples of the required size; without this constraint, it would have been impossible to compare

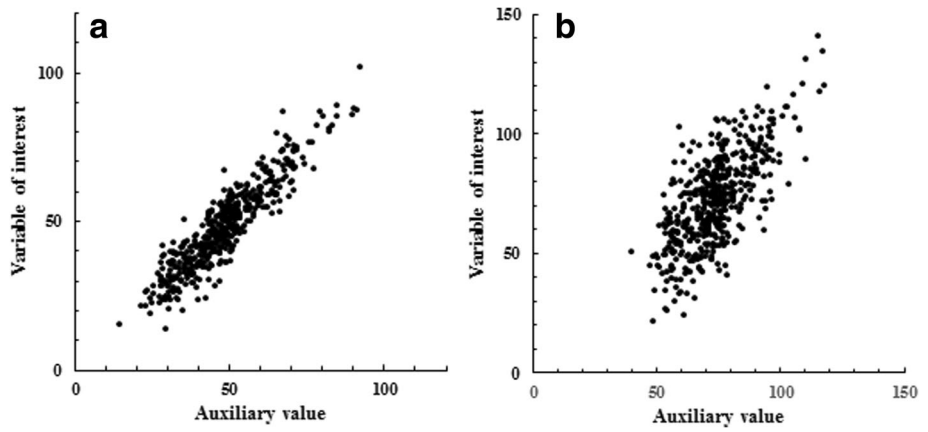

Fig. 1 Example of simulated data. The example is of stand basal area of Japanese rainforest and shows scatter plots of 500 randomly selected values of the 10,000 simulated population values of the variable of interest (target variable) plotted against the corresponding auxiliary variable values when the correlation between the two variables was **a** 0.9 and **b** 0.7. The data were generated so that the straight-line relationship between the two variables for the 10,000 data values passed through the origin

the QPPS estimators properly with the others because their first-phase sample sizes would have differed from that of the others.

For the estimators of Table 1 for which no analytical estimator of the standard error of the estimate of the population mean from any sample was available, the estimate was determined using bootstrapping. Wang and Butar (2006) reported that bootstrapping seems to perform well when applied to double sampling methods. West (2017) found it was the most appropriate estimator of the standard error of the estimate of the population mean for Eq. (8.2) when compared with some analytical estimators that had been proposed by various authors.

When performing bootstrapping with any of the sampling methods used here, a single bootstrap sample was chosen from the original sample by sampling with replacement from each of the first- and second-phase subsamples so that the sizes of both sub-samples in the bootstrap sample was the same as in the original sample. Bootstrapping of any one sample was based on 1000 bootstrap samples a number that was found to give consistent results for the five examples. Given $m = 1000$ population mean estimates from 1000 such bootstrap samples, the estimate of the standard error, $\hat{\sigma}(\overline{Y})$, of the estimate of the population mean from the original sample was determined as $\left\{\left[\Sigma_m\left(\overline{Y}_k - \overline{\overline{Y}}\right)^2\right]/m\right\}^{1/2}$, where $\overline{Y}_k$ was the kth ($k = 1, \ldots, m$) bootstrap estimate of the population mean and $\overline{\overline{Y}} = \left(\Sigma_m \overline{Y}_k\right)/m$.

Results

Bias in estimators

The bias in estimates of the population mean based on the means of the simulation estimates for each example and for the three second-phase sample sizes are shown in Table 2. There was no indication for any of the five examples or any of the methods or estimators that the level of bias was related to the target-auxiliary variable correlation level. Thus, each result in Table 2 is the mean of the 45,000 simulations done over all nine correlation levels for each example. There was little evidence of any appreciable bias with any of the methods or estimators or with any second-phase sample size. The

Table 2 Bias (%) in estimates of the population mean from simulations of inventory in five example populations

Example	Methods with complete enumeration of *N* sampling units as first phase					Methods with first-phase sample size = *f*				
	Ratio of means	Mean of ratios	Model-assisted	PPS sampling	QPPS sampling	Simple random sampling	Ratio of means	Mean of ratios	Model-assisted	QPPS sampling
Second-phase sample size = 100										
1	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	− 0.61
2	− 0.01	− 0.01	− 0.02	0.02	0.03	− 0.01	− 0.02	−0.02	− 0.02	0.13
3	− 0.01	− 0.01	− 0.01	0.00	0.01	− 0.01	− 0.01	− 0.01	− 0.01	− 0.47
4	− 0.02	− 0.02	− 0.02	− 0.03	− 0.01	− 0.02	− 0.02	− 0.02	− 0.02	− 0.72
5	0.01	0.00	0.01	0.00	− 0.01	0.02	0.01	0.00	0.01	− 0.65
Second-phase sample size = 40										
1	0.02	0.03	0.02	0.02	0.04	0.02	0.02	0.03	0.02	− 0.55
2	0.00	0.00	0.01	0.04	0.01	− 0.02	0.00	0.00	0.01	0.03
3	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00	0.00	− 0.47
4	0.02	0.02	0.02	− 0.02	− 0.01	0.03	0.01	0.01	0.01	− 0.70
5	0.04	0.04	0.05	0.01	0.03	0.05	0.04	0.03	0.04	− 0.57
Second-phase sample size = 10										
1	− 0.02	− 0.01	− 0.02	0.06	− 0.07	0.01	0.00	0.01	0.00	− 0.76
2	− 0.03	− 0.04	− 0.03	− 0.01	− 0.01	− 0.08	− 0.02	− 0.03	− 0.02	− 0.03
3	0.01	0.01	− 0.02	0.00	0.00	− 0.03	− 0.02	− 0.02	− 0.04	− 0.55
4	0.03	0.03	0.02	− 0.02	− 0.03	0.02	0.04	0.04	0.03	− 0.84
5	0.02	0.02	0.01	− 0.02	0.00	0.01	0.02	0.01	0.00	− 0.75

Bias was determined as the difference between the mean of the simulation estimates of the population mean and the population true mean, expressed as a proportion of the true mean. Results are shown for each of the methods applied in this work both when a complete enumeration of the auxiliary variable was done in the first-phase for the entire population or when the first-phase sample was smaller, of size *f* (<*N*). Each value shown in the table is the mean of 45,000 simulations of each estimator, those being made up of 5000 simulations for each of nine levels of target-auxiliary variable correlation. Results are given also when three different second-phase sample sizes were used. The examples are numbered as 1 – Basal area of Japanese rainforest, 2 – Stocking density of Japanese rainforest, 3 – Tree stem diameters, *P. radiata*, South Australia, 4 – Tree stem diameters, rainforest planting, New South Wales, 5 – Stand sawlog volume, Victorian eucalypt forest. Values shown as zero were actually <0.005%

QPPS estimator (Eq. 8.2) showed a slight tendency to under-estimate the population mean, but even then, the under-estimate was always less than 1%.

Several tests were made of the estimates of the standard error of the estimates of the population mean. Firstly, consideration was given to simple random sampling (Eq. 3.2). Consider a population of size N in which the target variable of the ith sampling unit takes the value y_i; these target variable values were known in each of the five population examples used here. Following Sokal and Rohlf (1995, Eq. 7.2a), the expected value of the standard error $\left[\sigma\left(\overline{Y}\right)\right]$ of the set of sample means of size n determined using simple random sampling from the population is

$$\sigma\left(\overline{Y}\right) \;=\; \sigma/n^{1/2} \quad , \tag{9}$$

where $\sigma = \{[\Sigma_N (y_i - \mu)^2]/N\}^{1/2}$ and $\mu = (\Sigma_N y_i)/N$. For each of the five example populations, the average deviation from this expected value of the actual standard error of the 45,000 simulation estimates of the population mean obtained using simple random sampling is shown in the second column of Table 3. The deviations were small, ranging over − 0.3 to + 0.7%, and showed no apparent pattern over the five examples or with differing sample sizes. These results are consistent with the theoretical expectation of Eq. (9) and confirm the efficacy of the simulation process employed here.

Table 3 Bias (%) in estimates of standard errors of estimates of the population mean in inventory using simple random sampling

Example	Bias in relation to expected values		Bias in relation to actual standard error of simulation means of mean of simulation estimates of standard error
	Of actual standard error of simulation means	Of mean of simulation estimates of standard error	
Sample size (n) = 100			
1	− 0.04	− 0.23	− 0.19
2	0.53	− 0.37	− 0.87
3	0.48	− 0.26	− 0.73
4	0.01	− 0.21	− 0.21
5	− 0.32	− 0.27	0.07
Sample size (n) = 40			
1	0.46	− 0.65	− 1.10
2	0.28	− 0.89	− 1.15
3	0.73	− 0.79	− 1.49
4	0.24	− 0.62	− 0.85
5	0.70	− 0.68	− 1.36
Sample size (n) = 10			
1	− 0.17	− 2.78	− 2.60
2	− 0.09	− 3.46	− 3.37
3	0.60	− 3.27	− 3.84
4	0.17	− 2.63	− 2.79
5	− 0.23	− 3.05	− 2.81

However, whilst the square of the estimate of the standard error of the estimate of the population mean obtained using a simple random sample, that is $\hat{\sigma}\left(\overline{Y}\right)^2$ from Eq. (3.2), is an unbiased estimator of the square of the standard error of the population true mean (its variance), that is σ^2 in Eq. (9), its square root is not. That is, $\hat{\sigma}\left(\overline{Y}\right)$ is not an unbiased estimator of σ (Gurland and Tripathi 1971; Sokal and Rohlf 1995, p. 53). The bias is generally small for larger sample sizes, but may become appreciable for small samples. Of course, it is the estimate of the standard error that is of prime interest to the user of inventory because it, rather than the variance, is used in determining the confidence limit of the estimate. Thus, bias in the estimate of standard error will introduce bias in the confidence limit and it is of interest to the user to know how large that bias might be. The deviations, from their expected value (Eq. 9), of the mean of the 45,000 estimates of the standard error of the estimate of the population means from the simulations using simple random sampling are shown in the third column of Table 3. There was consistent under-estimation with all five examples and the bias increased as sample size decreased, consistent with theoretical expectations. However, with a sample size of n = 100, the under-estimation was small over all five examples, never being more than 0.4%.

Analytical formulae such as Eq. (9) are not necessarily available to determine the expected value of estimates of the standard error of estimates of the population mean for the other forms of sampling being considered here (Table 1). Accordingly, the level of bias in such estimates was determined as the deviation of their mean over many simulations with the actual standard error of the corresponding estimates of the population means from the simulations. This has been done in the fourth column of Table 3 for results obtained using simple random sampling. The results there are little different from those in the third column which were deviations from the expected values. This confirms the efficacy of using this approach to investigate bias in estimates of the standard error of estimates of the population mean.

These biases were then determined for all other sampling methods considered here, in each case the bias being considered in relation to the actual standard error of simulation means. Results are shown in Table 4. As for bias in parameter estimates (Table 2), there was no evidence of differing levels of target-auxiliary variable correlation on the level of bias, so data in Table 4 were

Table 4 Bias (%) in estimates of standard errors of estimates of the population mean for the various methods considered here

Example	Methods with complete enumeration of *N* sampling units as first phase					Methods with first-phase sample size = *f*				
	Ratio of means	Mean of ratios	Model-assisted	PPS sampling	QPPS sampling	Simple random sampling	Ratio of means	Mean of ratios	Model-assisted	QPPS sampling
Second-phase sample size = 100										
1	− 0.7	− 0.8	− 0.9	− 0.7	− 1.7	− 0.2	− 0.8	− 0.9	− 0.9	− 1.3
2	− 0.7	− 0.9	− 0.7	− 0.7	− 0.5	− 0.9	− 1.0	− 0.9	− 1.1	− 0.9
3	− 1.1	− 2.1	− 0.7	− 1.1	− 0.8	− 0.7	− 1.1	− 1.2	− 1.3	− 0.9
4	− 1.2	− 1.3	− 1.2	− 1.0	− 0.9	− 0.2	− 0.7	− 1.0	− 0.8	− 0.8
5	− 0.6	− 0.3	− 0.3	− 0.9	− 0.5	0.1	− 0.2	− 0.2	− 0.3	− 0.4
Second-phase sample size = 40										
1	− 1.5	− 2.7	− 2.2	− 0.7	− 1.8	− 1.1	− 2.0	− 2.1	− 1.9	− 1.6
2	− 1.2	− 2.3	− 1.4	− 1.4	− 2.2	− 1.2	− 1.9	− 2.2	− 1.5	− 2.3
3	− 1.5	− 2.4	− 1.9	− 1.1	− 2.0	− 1.5	− 2.3	− 2.4	− 2.0	− 1.9
4	− 0.9	− 2.1	− 1.9	− 1.4	− 1.5	− 0.8	− 1.6	− 1.8	− 1.6	− 1.6
5	− 1.1	− 2.3	− 1.9	− 0.7	− 1.8	− 1.4	− 2.2	− 2.4	− 2.1	− 1.6
Second-phase sample size = 10										
1	− 2.7	− 8.1	− 1.0	− 2.6	− 8.2	− 2.6	− 6.8	− 7.3	− 2.3	− 7.4
2	− 3.5	− 8.3	35.4	− 3.6	− 8.5	− 3.4	− 7.0	− 7.5	21.4	− 7.4
3	− 3.4	− 8.4	− 1.2	− 3.0	− 8.2	− 3.8	− 7.7	− 7.9	− 2.6	− 7.5
4	− 3.5	− 8.3	4.0	− 3.0	− 8.0	− 2.8	− 6.9	− 7.3	− 1.2	− 7.1
5	− 3.4	− 8.3	8.6	− 3.1	− 7.8	− 2.8	− 7.2	− 7.6	1.5	− 7.3

Bias was determined as the mean of differences between estimated values of standard errors of estimates of the population mean from simulations and the actual standard error of the simulation estimates of the population mean, expressed as a proportion of the actual value. The structure and content of the table is otherwise as that of Table 2. The column headed 'Simple random sampling' contains the same values as in the fourth column of Table 3

obtained by pooling results for the 45,000 simulations in each case over the nine levels of correlation. The results obtained using simple random sampling that were shown in the fourth column of Table 3 have been repeated in Table 4; they were determined on the same basis as the remainder of the results in Table 4 and so may be compared directly with them. The under-estimation of standard errors that was evident and discussed for results obtained using simple random sampling in Table 3 is evident consistently across all sampling methods in Table 4. It became larger as the second-phase sample size decreased. For a second-phase sample size of 100, under-estimation was generally less than 1% for all five example populations. It generally tended to be slightly less for results obtained using simple random sampling than for the other methods.

At the smallest second-phase sample size of 10 in Table 4, the mean of ratios and QPPS sampling estimators appear to display somewhat greater under-estimation of standard errors than the other methods. And in particular for this sample size, model-assisted sampling in example 2 (the stocking density of Japanese rainforest) stands out from the other results by displaying gross over-estimation of estimates of standard error. The reason for this is apparent from information in Fig. 2 which shows the actual second-phase sample target and auxiliary variable data values from two example simulations from that population. The example in Fig. 2a was a case where, by chance, the sample selected showed a non-significant relationship between the target and auxiliary variable. Even though over the whole population from which this example sample was chosen, this correlation was 0.7, it is, of course, possible to find an individual sample where the correlation is poor. In this example sample, the model-assisted estimate of the standard error of the estimate of the population mean from the sample was 7775, much greater than the actual standard error over the 5000 simulations of 1769; such an over-estimation is inevitable when there is no substantial relationship between the target and auxiliary variables in the sample. By contrast, in the case of Fig. 2b, there was an appreciable and significant target-auxiliary variable correlation of 0.77 and an estimate of the standard error of the estimate of the population mean from the sample of 927, much lower than the actual standard error. In the example 2 simulations with the smallest second-phase sample size, it seems that cases such as that shown in Fig. 2a must have occurred sufficiently

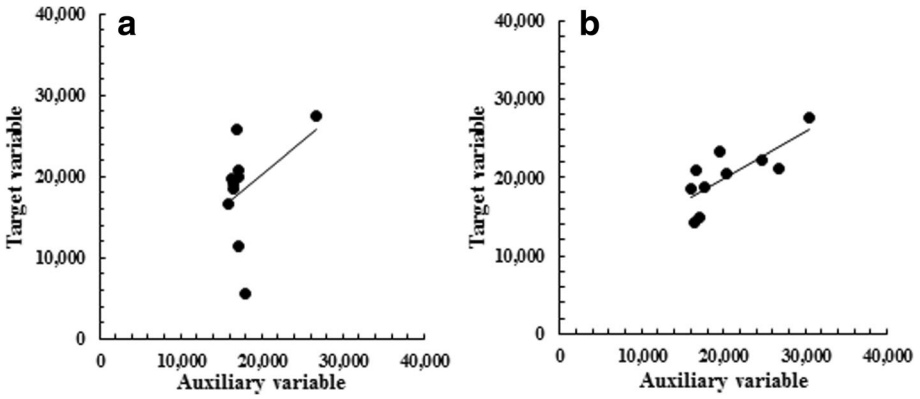

Fig. 2 Examples illustrating problems with model-assisted estimation. Two examples are shown of second-phase simple random samples, of size 10, chosen from amongst the 5000 simulations done with population example 2, the stocking density of Japanese rainforest, with a first-phase sample size of 30 and where the target-auxiliary variable correlation over the population had been set at $r = 0.7$. In both cases, the solid line shows the ordinary least-squares straight-line fit to the target-auxiliary variable data, that is, the model used to obtain model-assisted estimates of the population mean using Eq. (6.2). In the case of **a** the estimates of the parameter values were $\alpha = 3533$ and $\beta = 0.083$, but the correlation level between the target and auxiliary variable data was $r = 0.51$, which was not significantly different from zero (at $p = 0.05$). In the case of **b**, $\alpha = 7800$, $\beta = 0.604$ and the correlation was $r = 0.77$, which was significantly greater than zero (at $p < 0.01$ at least)

often that model-assisted sampling was unable to offer reasonable estimates of the standard error of the estimate of the population mean with the smallest sample size.

Relative efficiencies of double sampling estimators

For each of the five examples and for the different second-phase sample sizes, the efficiency of all the double sampling estimators relative to results obtained using simple random sampling are shown in Table 5 when the level of correlation between the target and auxiliary variable values in the whole population was 0.9. These values were determined as the difference between the mean of the simulation standard error estimates of the population mean for each double sampling estimator and the mean of the simulation standard error estimates obtained using simple random sampling, expressed as a proportion of the mean obtained using simple random sampling.

As might be expected, all the double sampling estimators were much more efficient than estimates obtained using simple random sampling. They were about 50–65% more efficient when the first-phase sampling involved complete enumeration of the N sampling units in the population but less, 25–35%, when the first-phase sample was much smaller (methods with first-phase sample size = f in Table 5). This higher efficiency in the former case is inevitable because much more information is being used from the population because of the complete enumeration of the auxiliary variable.

However, there seemed to be no consistent or substantial differences between the efficiencies of any of the double sampling estimators. The only exception was for model-assisted sampling in example 2 with the smallest second-phase sample size, which was clearly much less efficient than any of the other estimators, although still appreciably more efficient than results obtained using simple random sampling; the reasons for that were discussed earlier and explained through Fig. 2.

The results in Table 5 were obtained from populations in which the target-auxiliary variable correlation level had been set at 0.9. The only difference that was found in the results for the populations with other levels of target-auxiliary variable correlation was that the improvement in efficiency of the double sampling estimators, when compared with results from simple random sampling, declined progressively as the correlation level declined. However, the improvement still remained even at the lowest correlation level used, 0.1. Further, there was no evidence of any difference in efficiency of any of the double sampling estimators at any particular level of correlation.

As an example, Fig. 3 shows for the stand basal area of Japanese rainforest (population example 1) how the sampling efficiency of each of the double sampling estimators changed, relative to that achieved using simple random sampling, as the level of the target-auxiliary variable correlation changed. These results were for the case of a second-phase sample size of 100. The advantage of the greater first-phase sample size is obvious and the progressive improvement in efficiency with increasing correlation is apparent. Similar diagrams were drawn for all the example populations; in each case, all of the five parts of those

Table 5 Sampling efficiency (%) of the various methods considered here relative to simple random sampling

Example	Methods with complete enumeration of *N* sampling units as first phase					Methods with first-phase sample size = *f*			
	Ratio of means	Mean of ratios	Model-assisted	PPS sampling	QPPS sampling	Ratio of means	Mean of ratios	Model-assisted	QPPS sampling
Second-phase sample size = 100									
1	− 59	− 53	− 58	− 58	− 57	− 28	− 25	− 28	− 29
2	− 55	− 47	− 55	− 63	− 52	− 31	− 26	− 31	− 27
3	− 57	− 55	− 56	− 58	− 56	− 28	− 27	− 28	− 29
4	− 59	− 52	− 59	− 62	− 57	− 28	− 24	− 28	− 29
5	− 61	− 52	− 61	− 65	− 58	− 29	− 24	− 29	− 30
Second-phase sample size = 40									
1	− 59	− 54	− 58	− 58	− 58	− 28	− 26	− 28	− 29
2	− 54	− 47	− 54	− 63	− 52	− 31	− 26	− 31	− 28
3	− 57	− 56	− 56	− 58	− 57	− 28	− 27	− 28	− 29
4	− 59	− 53	− 59	− 62	− 57	− 29	− 25	− 28	− 29
5	− 61	− 53	− 60	− 65	− 58	− 29	− 24	− 29	− 29
Second-phase sample size = 10									
1	− 58	− 56	− 55	− 58	− 59	− 31	− 28	− 29	− 33
2	− 53	− 49	− 22	− 62	− 53	− 31	− 27	− 14	− 30
3	− 56	− 57	− 52	− 57	− 58	− 30	− 29	− 26	− 31
4	− 59	− 55	− 52	− 62	− 59	− 31	− 28	− 28	− 33
5	− 60	− 55	− 51	− 65	− 61	− 31	− 27	− 27	− 33

Each value in the table was determined as the difference between the mean of the simulation estimates of the standard error of estimates of the population mean for the estimator concerned and the mean when simple random sampling was used, as a proportion of the mean in the simple random sampling case. Each value in the table was the mean of 5000 simulations when the target-auxiliary variable correlation level was 0.9. Otherwise, the structure of the table is similar to that of Table 2. The results in the first row of the table are repeated in Fig. 3

diagrams were virtually identical overlays of each other, showing that all the double sampling estimators were behaving with the same efficiency of estimation in any of the examples.

Discussion

The results suggested that all the double sampling estimators considered here (Table 1) performed equally well with any of the five example populations with which they were tested. They all displayed negligible bias as estimators of the population mean (Table 2). This is perhaps somewhat unexpected because it is well known that ratio estimators are subject to bias (Hartley and Ross 1954; Cochran 1977, Sect. 6.3). In the case of the mean of ratio estimator (Eqs. 5.1 and 5.2), the bias correction factor of Hartley and Ross (1954) was incorporated into the estimator. Otherwise, no correction factors were used in the other estimators and bias was still negligible, even when the second-phase sample size was small ($n = 10$, a sampling intensity of 0.1%).

All the estimators were slightly biased towards underestimation of the standard error of the estimate of the population mean (Table 4) whether or not an analytical estimator (Eqs. 3.2, 4.1, 7.1) or bootstrapping (all other cases) was used. This bias was evident even when simple random sampling was used and, as discussed in the previous section, this was to be expected theoretically. There was no hint that the analytical estimators were superior to bootstrapping in this context. The bias increased as the second-phase sample size decreased, but with the largest second-phase sample size used here ($n = 100$, a sampling intensity of 1%) the under-estimation was generally less than 1% over all five example populations, a degree of bias that would often be considered acceptable for many inventory purposes.

All the double sampling estimators with their various sampling designs were more efficient in determining the standard error of the estimate of the population mean when compared with results obtained using simple random sampling. Their efficiency increased substantially as the level of correlation between the target and auxiliary variables increased (Fig. 3) and as the first-phase sample size increased (Table 5, Fig. 3). In the case of the ratio of means estimator with complete enumeration of the auxiliary variable over the whole population (Eq. 4.1),

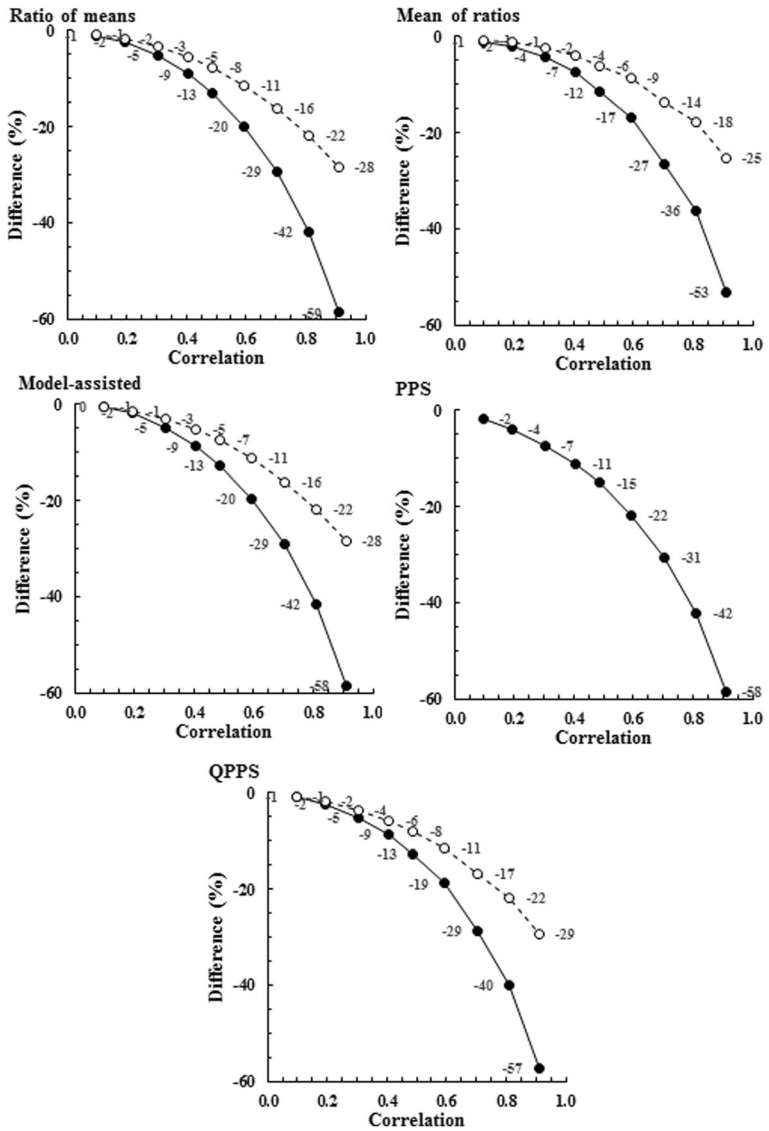

Fig. 3 Effects of target-auxiliary variable level of correlation on efficiency of the various double sampling estimators for the stand basal area of Japanese rainforest example. Scatter plots are drawn, against the target-auxiliary variable level of correlation, of the difference (%) between the mean of the simulation estimates of the standard error of estimates of the population mean for the estimator concerned and the mean of the standard error of estimates obtained when simple random sampling was used, as a proportion of the mean from the simple random sampling case. The results are for the case that the second-phase sample size was 100. Each data point is the mean of 5000 standard error estimates. Results are shown for the case that values of the auxiliary variable were available from a complete enumeration of the population (·—·) or for a first-phase sample size of f = 240 (O- - -O). The annotation on each data point shows the actual percentage difference for that point; values for correlation level of 0.9 there are shown also as the first line of Table 5

Cochran (1977, Sect. 6.6) showed that, for large sample sizes, it will be a more efficient estimator than results obtained using simple random sampling when

$$\rho \; > \; \left(S_X \overline{Y}_Y\right) / \left(2 S_Y \overline{Y}_X\right) \qquad , \tag{10}$$

where ρ is the level of correlation between the target and auxiliary variable in the sample being considered, $\overline{Y}_Y$ and $\overline{Y}_X$ are the population means of the target and auxiliary variables, respectively, and S_Y and S_X are their respective standard deviations. In the present simulations, this condition was found to be satisfied in 80% or more of all the samples selected from all five example populations, the proportion rising rapidly to 100% as the target-auxiliary variable correlation level increased. Thus, it is not surprising that the double sampling methods used here were consistently more efficient than results obtained using simple random sampling. Of course, the very reason

for using auxiliary information with double sampling is to gain this advantage.

All of the double sampling estimators seemed to be equally efficient (Table 5). There was one exception, where model-assisted estimation, with a small second-phase sample size in population example 2, was less efficient than the other estimators; as explained through Fig. 2, this was a consequence of some samples having, by chance, a non-significant relationship between the target and auxiliary variables in the second-phase sample. Even though such sampling cases will occur when using the ratio estimators, there was no hint that this led to any estimation problems. In effect, the problem is avoided in those cases because the Y–X relationship for a sample is always forced through the origin with the ratio estimators.

Of course, different theoreticians have developed all these different double sampling estimators over the years hoping to find estimators that are more efficient than those devised previously. In particular, the PPS and QPPS estimators (Eqs. 7.1, 8.1, 8.2) involve sampling with probability proportional to size in selecting the second-phase sample. No doubt the developers of those estimators hoped that this would render them more efficient than the other estimators tried here (Eqs. 4.1–6.2), all of which used simple random sampling at the second phase. However, the present results showed no such gains.

Consideration of the way in which the estimators used here operate offers an explanation as to why they display similar efficiencies of estimation. Each of the ratio estimators uses a method to estimate a ratio (shown as $\bar{R}$ in Table 1) from the second-phase sample. That ratio is then used in subsequent computations. It is an estimator of the slope (β) of the model assumed here as defined by Eq. (1) with $\alpha = 0$; Särndal et al. (1992, Sects. 7.3, 7.4) discussed this specifically in relation to the mean of ratio estimator. In the case of model-assisted estimation, this slope is estimated also (it is shown as β in Eqs. 6.1 and 6.2) using least-squares regression with Eq. (1). For any one sample, the distribution of the target and auxiliary variable values along that straight-line relationship line will tend to be uniform with simple random sampling, whilst larger values will tend to be more common when sampling with probability proportional to size. In regression analysis, those distributions may affect the precision of the estimate of the slope, although in ways that are not easily predictable (Box and Draper 1959), and, by inference, the precision of the estimates here of $\bar{R}$ and ultimately the estimates of the standard error of the estimate of the population mean. The present results suggest that neither the differences in the methods used by the different estimators to estimate this slope nor the differences in target and auxiliary variable distributions from different forms of sampling in the second phase (an issue recognised also by Gregoire and Valentine 2008, p. 165) have led to any appreciable differences in efficiency of the various double sampling estimators tested here. Särndal et al. (1992, p. 274) suggested that model-assisted estimation should generally be more efficient than ratio estimators, especially for larger sample sizes; the present results suggest that this advantage was negligible in any of the five examples.

As considered in the discussion following Eqs. (1) and (2), it was an implicit assumption of the ratio estimators used here that the target-auxiliary variable relationship can be represented as a straight line passing through the origin (Cochran 1977, Sect. 6.7; Särndal et al. 1992, Sect. 7.3; Gregoire and Valentine, 2008, p. 167). This is not a problem with model-assisted estimation because the intercept of the relationship on the target variable axis (the parameter α in Eqs. 1, 6.1 and 6.2) is determined and used directly as part of the estimation process. When the population example data used here were adjusted by addition of a constant to the target variable values, so that the target-auxiliary variable relationship was no longer through the origin, additional simulation studies found that the efficiency of all the ratio estimators declined and rapidly became much less than results obtained using simple random sampling if the constant used was large enough (results not reported here). However, it was found also that if the target-auxiliary variable relationship for any sample did not pass through the origin, it could easily be made to do so by fitting a straight line regression to the data and then transforming the sample data by subtracting the regression constant from all the target variable values in the sample. The modified data would then yield an estimate of the population mean (to which the constant that had been subtracted was added back) and would also yield the correct estimate of the standard error of the estimate of population mean because the variance of the sample data was unaffected by the transformation (these results also not reported here). Thus, even if the target-auxiliary variable relationship does not pass through the origin, it is quite simple to transform sample data so that any of the ratio estimators used here may be applied satisfactorily.

On the other hand, if the target-auxiliary variable relationship in the population is not a straight line or if more than one auxiliary variable is available, the ratio estimators can no longer be used. As mentioned earlier, model-assisted estimation would then become appropriate and the full power of linear, non-linear and multiple regression may be employed in the place of the simple straight-line model shown in Eqs. (1), (6.1) and (6.2). When the data are heteroscedastic ($g > 0$ in Eq. 2), weighted least-squares regression may be used to fit the chosen model.

The present results may be compared with those from some other simulation studies. In two example forest populations, Schreuder et al. (1987) tested several of the estimators that were used here, the ratio of means estimator, model-assisted estimation and PPS sampling (denoted by them as $\hat{Y}_{rm}$, $\hat{Y}_{lr}$ and $\hat{Y}_{HT}$, respectively) in circumstances that assumed the auxiliary variable value had been measured on all sampling units in the population. As in the present work, they found that bias in estimates of the population mean was small (less than ± 0.5%) with any of the estimators. However, they found that estimates of the standard error of estimates of the population mean were often appreciably smaller (as much as 85%) with model-assisted estimation. They did not state specifically if their target-auxiliary variable relationships passed through the origin and the advantages they found with model-assisted estimation may reflect that issue.

Reich et al. (1993) compared use of the ratio of means estimator and model-assisted estimation to estimate the mean unit area biomass of a perennial prairie grass, native to north America, across a 1500-m^2 sample area in Colorado; they used a process equivalent to having the auxiliary variable measured only on a first-phase sample rather than the entire population. The relationship between their target and auxiliary variable (the auxiliary variable was an estimate by eye of grass biomass on any sampling unit) was a straight line passing through the origin; they did not specify the target-auxiliary variable correlation level, but the nature of their sampling procedure suggests it would have been quite high, perhaps greater than 0.5. They found (their Tables 2 and 3) that both estimators tended to under-estimate the true mean of their population by 2–4%, the bias tending to decline with increasing sample size. Consistent with present results, they found that both estimators were equally efficient (based on their results from bootstrap estimation of standard errors) and both were appreciably more efficient than results obtained using simple random sampling by at least 24%, rising to as much as 61% as sample size increased. Francis et al. (1979) also studied biomass sampling of prairie grass and concluded that model-assisted estimation was preferable to the ratio of means estimator. However, as Reich et al. noted (their p. 90), the target-auxiliary variable relationships of Francis et al. did not necessarily pass through the origin.

The present results appear to be at variance with those of West (2017). He used the same simulation examples as here, but considered only QPPS sampling when the auxiliary variable was not measured on all sampling units (Eq. 8.2). In contrast to the present work, he used simulation data sets where the target-auxiliary variable relationship did not pass through the origin. This led to cases where the use of simple random sampling appeared to give more efficient estimates than QPPS sampling at lower levels of target-auxiliary variable correlation. Correction of that issue in the present work ensured that, at any positive correlation level, all the estimators tested here gave more efficient estimates than those obtained using simple random sampling.

Conclusions

The present work compared results obtained when sampling from five simulated forest populations that had rather differently shaped frequency distributions of their target variables. Real forest population circumstances were used to derive the simulated populations; these then reflected a range of circumstances that might reasonably be encountered in actual forest inventories.

Simulations of sampling from these five populations were compared with the use of simple random sampling and double sampling that used model-assisted estimation or each of four ratio estimators (Table 1). Bias in estimates of the population mean was negligible for all the methods used, even though their formulations did not always preclude this possibility. Whilst all displayed some bias in the estimation of the standard error of estimates of the population mean, the bias was small and consistent with theoretical expectations. In some cases, analytical estimators were used to estimate standard errors and in others bootstrapping was used; there was no indication that bootstrapping was unsatisfactory. Efficiency of all the double sampling estimators increased as either the first or second-phase sample sizes increased. All gave more efficient results than the use of simple random sampling as long as there was some positive level of correlation between the target variable of the inventory and the auxiliary variable being used; their efficiency increased progressively as the level of correlation increased.

However, none of the double sampling estimators was more efficient than any of the others. This applied even when sampling with probability proportional to size was used in the second-phase of sampling rather than simple random sampling. This seemed surprising because development and use of these double sampling estimators have continued over many years, presumably because it has been anticipated that one or other of them is preferable to the others. A few other studies have compared one or more of these estimators from which it may be concluded also that there is little difference between the results obtained with any of these estimators.

With the exception of model-assisted sampling, it is essential to the double sampling procedures that there be no more than one auxiliary variable used and that the

relationship between it and the target variable in the population being considered can be represented adequately as a straight line passing through the origin. If the relationship does not pass through the origin, it was found in additional work that was not reported in detail here that the data could be transformed very simply to ensure that it does so upon which the sample data may be used readily to make the requisite estimates of population variables. If more than one auxiliary variable is available or the relationship between them and the target variables is other than a straight line, then model-assisted estimation must be used and may employ the full power and options of regression analysis.

Two things must be borne in mind when considering the ramifications of the present results. Firstly, whilst a positive level of correlation between the auxiliary and target variables always led to increased efficiency of the double sampling estimators over that achieved using simple random sampling, there may be more work involved in double sampling than with simple random sampling alone because of the need to obtain the data for the auxiliary variables. On the other hand, to achieve the same efficiency of estimates, it may be necessary to use a larger sample size if simple random sampling with measurement of the target variable only is to be done. It will depend on the balance between the cost of these alternatives that will determine which of them is preferable.

Secondly, the present results were obtained using just five example populations. Whilst these were based on real forest populations that had rather differently shaped frequency distributions, there may be other populations and target variables for which different conclusions might be reached. However, the present results suggest strongly that none of the double sampling estimators considered here offers better results than any of the others. That is, for many forest inventory tasks, users may well be able to use whichever of the estimators is most convenient to their purpose. Given that model-assisted estimation has application in a wider range of circumstances than the ratio estimators, perhaps it can be recommended for more general use than the others.

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Author's information
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

Received: 3 July 2017 Accepted: 13 September 2017
Published online: 16 October 2017

References

Ahamed, T., Tian, L., Zhang, Y., & Ting, K. C. (2011). A review of remote sensing methods for biomass feedstock production. *Biomass and Bioenergy, 35*, 2455–2469.

Al-Omari, A. I., & Bouza, C. N. (2015). Ratio estimators of the population mean with missing values using ranked set sampling. *Environmetrics, 26*, 67–76.

Baffetta, F., Fattorini, L., Franceschi, S., & Corona, P. (2009). Design-based approach to *k*-nearest neighbours techniques for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment, 113*, 463–475.

Bondesson, L., & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics, 35*, 466–483.

Box, G. E. P., & Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association, 54*, 622–654.

Clerici, N., Rubiano, K., Abd-Elrahman, A., Hoestettler, J. M. P., & Escobedo, F. J. (2016). Estimating aboveground biomass and carbon stocks in Periurban Andean secondary forests using very high resolution imagery. *Forests, 7*, 138.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cohen, R., Kaino, J., Okello, J. A., Bosire, J. O., Kairo, J. G., Huxham, M., & Mencuccini, M. (2013). Propagating uncertainty to estimates of above-ground biomass for Kenyan mangroves: A scaling procedure from tree to landscape level. *Forest Ecology and Management, 310*, 968–982.

Corona, P. (2016). Consolidating new paradigms in large-scale monitoring and assessment of forest ecosystems. *Environmental Research, 144*, 8–14.

Corona, P., Fattorini, L., Franceschi, S., Scrinzi, G., & Torresan, C. (2014). Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: Model-based, design-based, and hybrid perspectives. *Canadian Journal of Forest Research, 44*, 1303–1311.

Corona, P., Kohl, M., & Marchetti, M. (2003). *Advances in Forest Inventory for Sustainable Forest Management and Biodiversity Monitoring*. Dordrecht, The Netherlands: Kluwer.

Du, L., Zhou, T., Zou, Z. H., Zhao, X., Huang, K. C., & Wu, H. (2014). Mapping forest biomass using remote sensing and national forest inventory in China. *Forests, 5*, 1267–1283.

Falkowski, M. J., Wulder, M. A., White, J. C., & Gillis, M. D. (2009). Supporting large-area, sample-based forest inventories with very high spatial resolution satellite imagery. *Progress in Physical Geography, 33*, 403–423.

Francis, R. C., Van Dyne, G. M., & Williams, B. K. (1979). An evaluation of weight estimation double sampling as a method of botanical analysis. *Journal of Environmental Management, 8*, 55–72.

Furnival, G. M., Gregoire, T. G., & Grosenbaugh, L. R. (1987). Adjusted inclusion probabilities with 3P sampling. *Forest Science, 33*, 617–633.

Grafström, A. (2010). On a generalization of Poisson sampling. *Journal of Statistical Planning and Inference, 140*, 982–991.

Gregoire, T. G., & Valentine, H. T. (1999). Composite and calibration estimation following 3P sampling. *Forest Science, 45*, 179–185.

Gregoire, T. G., & Valentine, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. Boca Raton, FL, USA.: Chapman & Hall/CRC.

Grosenbaugh, L. R. (1964). *Some suggestions for better sample-tree measurement.* In *Proceedings of the Society of American Foresters Meeting, 20–23 October 1963* (pp. 36–42). Boston, MA, USA.

Grosenbaugh, L. R. (1965). *Three-pee Sampling Theory and Program 'THRP' for Computer Generation of Selection Criteria*. Berkley: Research Paper PSW-21, Res. Paper PSW-21. Berkley, CA, USA: USDA Forest Service, Pacific Southwest Forest and Range Experiment Station.

Grosenbaugh, L. R. (1976). Approximate sampling variance of adjusted 3P estimates. *Forest Science, 22*, 173–176.

Gurland, J., & Tripathi, R. C. (1971). A simple approximation for unbiased estimation of the standard deviation. *American Statistician, 25*, 30–32.

Hamilton, F., & Brack, C. (1999). Stand volume estimates from modelling inventory data. *Australian Forestry, 62*, 360–367.

Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. *Nature, 174*, 270–271.

Immitzer, M., Stepper, C., Böck, S., Straub, C., & Atzberger, C. (2016). Use of WorldView-2 stereo imagery and National Forest Inventory data for wall-to-wall mapping of growing stock. *Forest Ecology and Management, 359*, 232–246.

Johnson, F. A., & Hartman, G. B. (1972). Fall, buck, and scale cruising. *Journal of Forestry, 65*, 722–726.

Johnson, F. A., Dahms, W. G., & Hightree, P. E. (1967). A field test of 3P cruising. *Journal of Forestry, 70*, 566–568.

Kangas, A., & Maltamo, M. (Eds.). (2006). *Forest inventory methodology and applications*. Dordrecht, The Netherlands: Springer.

Kangas, A., Myllymäki, M., Gobakken, T., & Næsset, E. (2016). Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. *Canadian Journal of Forest Research, 46*, 855–868.

Knottnerus, P. (2011). On the efficiency of randomized probability proportional to size sampling. *Survey Methodology, 37*, 95–102.

Köhl, M., Magnussen, S. S., & Marchetti, M. (2006). *Sampling Methods, Remote Sensing, and GIS Multiresource Forest Inventory*. Berlin: Springer.

Kohyama, T. (1986). Tree size structure of stands and each species in primary warm-temperate rain forests of southern Japan. *Botanical Magazine of Tokyo, 99*, 267–279. https://doi.org/10.1007/BF02489543.

Kumar, S., & Chhaparwal, P. (2016). A robust unbiased dual product estimator for population mean through modified maximum likelihood in simple random sampling. *Cogent Mathematics, 3*, 1168070.

Lahiri, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin de l'Institut International de Statistique, 33*, 133–140.

Li, Y. Z., Zhu, X. F., Pan, Y. Z., Gu, J. Y., Zhao, A. Z., & Liu, X. F. (2014). A comparison of model-assisted estimators to infer land cover/use class area using satellite imagery. *Remote Sensing, 6*, 8904–8922.

Lin, F.-M., & Chao, C.-T. (2014). Variances and variance estimators of the improved ratio estimators under adaptive cluster sampling. *Environmental and Ecological Statistics, 21*, 285–311.

Magnussen, S. (2001). Saddlepoint approximations for statistical inference of PPP sample estimates. *Scandinavian Journal of Forest Research, 16*, 180–192.

Magnussen, S., Næsset, E., & Gobakken, T. (2014). An estimator of variance for two-stage ratio regression estimators. *Forest Science, 60*, 663–676.

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Boca Raton, FL, USA.: Chapman & Hall/CRC.

Mandallaz, D. (2013). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research, 43*, 441–449.

McRoberts, R. E., & Tomppo, E. O. (2007). Remote sensing support for national forests inventories. *Remote Sensing of Environment, 110*, 412–419.

McRoberts, R. E., Chen, Q., Domke, G. M., Ståhl, G., Saarela, S., & Westfall, J. A. (2016). Hybrid estimators for aboveground carbon per unit area. *Forest Ecology and Management, 378*, 44–56.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C., & Gormanson, D. D. (2006). Using satellite imagery as ancilliary data for increasing the precision of estimates for the Forest inventory and analysis program of the USDA Forest Service. *Canadian Journal of Forest Research, 36*, 2968–2980.

McRoberts, R. E., Tomppo, E. O., & Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research, 25*, 368–381.

Melville, G., Stone, C., & Turner, R. (2015). Application of LiDAR data to maximise the efficiency of inventory plots in softwood plantations. *New Zealand Journal of Forestry Science, 45:* 9.

Neigh, C. S. R., Nelson, R. F., Ranson, K. J., Margolis, H. A., Montesano, P. M., Sun, G., Kharuk, V., Næsset, E., Wulder, M. A., & Andersen, H.-E. (2013). Taking stock of circumboreal forest carbon with ground measurements, airborne and spaceborne LiDAR. *Remote Sensing of Environment, 137*, 274–287.

Nielson, R. M., Evans, T. J., & Stahl, M. B. (2013). Investigating the potential use of aerial line transect surveys for estimating polar bear abundance in sea ice habitats: A case study for the Chukchi Sea. *Marine Mammal Science, 29*, 389–406.

Ometto, J. P., Aguiar, A. P., Assis, T., Soler, L., Valle, P., Tejada, G., Lapola, D. M., & Meir, P. (2014). Amazon forest biomass density maps: Tackling the uncertainty in carbon emission estimates. *Climate Change, 124*, 545–560.

Oral, E., & Oral, E. (2011). A robust alternative to the ratio estimator under non-normality. *Statistical Probability Letters, 81*, 930–936.

Paré, D., Gertner, G.Z., Bernier, P.Y., & Yanai, R.D. (2016). Quantifying uncertainty in forest measurements and models: approaches and applications. *Canadian Journal of Forest Research, 46(3), v*. https://doi.org/10.1139/cjfr-2016-0029.

Reich, R. M., Bonham, C. D., & Remington, K. K. (1993). Double sampling revisited. *Journal of Range Management, 46*, 88–90.

Ringvall, A., & Kruys, N. (2005). Sampling of sparse species with probability proportional to prediction. *Environmental Monitoring and Assessment, 104*, 131–146.

Ringvall, A. H., Ståhl, G., Ene, L. T., Næsset, E., Gobakken, T., & Gregoire, T. G. (2016). A poststratified ratio estimator for model-assisted biomass estimation in sample-based airborne laser scanning surveys. *Canadian Journal of Forest Research, 40*, 1386–1395.

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

Schreuder, H. T., Gregoire, T. G., & Wood, G. B. (1993). *Sampling Methods for Multiresource Forest Inventory*. New York: Wiley.

Schreuder, H. T., Li, H. G., & Hazard, J. W. (1987). PPS and random sampling estimation using some regression and ratio estimators for underlying linear and curvilinear models. *Forest Science, 33*, 997–1009.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry* (3rd ed.). New York: W.H. Freeman & Co..

Ståhl, G., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., Patterson, P. L., Magnussen, S., Næsset, E., McRoberts, R. E., & Gregoire, T. G. (2016). Use of models in large-area forest surveys: Comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems, 3:* 5. doi:10.1186/s40663-016-0064-9.

Stevenson, S. K. (1979). *Effects of selective logging on arboreal lichens used by Selkirk caribou. Fish and Wildlife Rep. No. R-2*. Victoria, B.C.: British Columbia Ministry of Forests Available from http://www.env.gov.bc.ca/wld/documents/techpub/r2.pdf . Accessed 19 Sep 2017.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H. M., Chamuya, N., Zahabu, E., & Otieno, J. (2014). A sampling design for a large area forest inventory: Case Tanzania. *Canadian Journal of Forest Research, 44*, 931–948.

Wang, J., & Butar, F.B. (2006). An empirical study of the bootstrap and the jackknife methods applied in double sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*. Accessible from http://www.amstat.org/sections/srms/Proceedings/allyearsf.html . Accessed 19 Sep 2017.

Waser, L. T., Fischer, C., Wang, Z. Y., & Ginzler, C. (2015). Wall-to-wall forest mapping based on digital surface models from image-based point clouds and a NFI forest definition. *Forests, 6*, 4510–4528.

Weiskittel, A. R., MacFarlane, D. W., Radtke, P. J., Affleck, D. L. R., Temesgen, H., Woodall, C. W., Westfall, J. A., & Coulston, J. W. (2015). A call to improve methods for estimating tree biomass for regional and national assessments. *Journal of Forestry, 113*, 414–424.

West, P. W. (2011). Potential for wider application of 3P sampling in forest inventory. *Canadian Journal of Forest Research, 41*, 1500–1508.

West, P. W. (2016). Simple random sampling of individual items in the absence of a sampling frame that lists the individuals. *New Zealand Journal of Forestry Science, 46:* 15. https://doi.org/10.1186/s40490-016-0071-1 .

West, P. W. (2017). Population structure and correlation between auxiliary and target variables may affect precision of estimates in forest inventory. *Communications in Statistics – Simulation and Computation, 46*, 4951–4965. https://doi.org/10.1080/03610918.2016.1139128.

Williams, M. S., & Schreuder, H. T. (1998). Outlier-resistant estimators for Poisson sampling: a note. *Canadian Journal of Forest Research, 28*, 794–797.

Wulder, M. A., Hall, R. J., Coops, N. C., & Franklin, S. E. (2004). High spatial resolution remotely sensed data for ecosystem characterization. *Bioscience, 54*, 511–521.