

STATISTICAL PROPERTIES OF MEAN STAND BIOMASS ESTIMATORS IN A LIDAR-BASED DOUBLE SAMPLING FOREST SURVEY DESIGN

H.-E. Andersen^{a,*} and J. Breidenbach^b

^a USDA Forest Service, Pacific Northwest Research Station, Anchorage, Alaska, USA - handersen@fs.fed.us

^b Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg Abteilung Biometrie und Informatik
Wonnhaldestr. 4, 79100 Freiburg, Germany - johannes.breidenbach@forst.bwl.de

KEY WORDS: Forestry, statistics, LIDAR, sampling, inventory, biomass

ABSTRACT:

Airborne laser scanning (lidar) can be a valuable tool in double-sampling forest survey designs. Lidar-derived forest structure metrics are often highly correlated with important forest inventory variables, such as mean stand biomass, and lidar-based *synthetic regression estimators* have the potential to be highly efficient compared to single-stage estimators, which could lead to increased precision for inventory estimates. However, when a limited sample is available to develop the regression model, an estimate based solely on the synthetic regression estimator can yield biased results for stands within a forest area where the regression model was unrepresentative. A number of *modified (approximately) design-unbiased regression estimators* have been proposed that serve to reduce this model-induced bias while also maintaining the efficient, variance-reducing properties of the synthetic regression estimator. In this study, we use a simulation approach to explore the statistical properties of several lidar-based regression estimators of mean stand biomass, using lidar and field plot data collected at a study site in a conifer forest in western Washington State, USA.

1. INTRODUCTION

1.1 Double-sampling in forest inventory

The use of covariate information obtained from remote sensing in a double-sampling design (e.g. with regression estimators) has been a well-established technique in forest survey for decades. A double-sampling design using a combination of remote sensing and field data is particularly cost-effective in the inventory of large, remote forest areas, where the cost of establishing field plots can be considerable, and the number of plots established is necessarily limited. In these cases, the use of remotely-sensed covariate information collected over a larger number of plots can greatly increase the precision and reliability of the inventory estimates for a given area. The use of aerial photos in forest mensuration, and particularly the use of aerial stand volume tables, has been used for many years to decrease forest inventory costs (Paine and Kiser, 2003). Although accurate forest measurements can be acquired from aerial photos through manual interpretation, automated extraction of three-dimensional information from stereo imagery is complex and error-prone, due to the inherently two-dimensional format of photographs, as well as shadows, layover, and the characteristically irregular shapes and surfaces of tree crowns. In addition, tree heights are difficult to measure accurately using aerial photographs, unless accurate terrain models are already available for the area. Because of these issues, the use of aerial photos for acquisition of detailed forest measurements in a double-sampling design has been limited in large-scale forest inventory programs in the United States.

1.2 Lidar for forest inventory applications

Airborne laser scanning (lidar) provides data on the full three-dimensional structure of the forest canopy, at a high resolution, and is readily amenable to automated processing and analysis. Due to the high demand for lidar-derived terrain information in forested areas, high-resolution, discrete-return lidar data is becoming increasingly available to forest managers all over the

world. Therefore, lidar has the potential to be a much more cost-effective sampling tool for operational forest inventory than aerial photography. In fact, the very strong correlations between lidar metrics and plot-level variables suggest that parameters such as stand biomass could be estimated with a high level of precision over a large area using a relatively small number of representative field plots.

The potential of lidar for forest measurement has already been well-established in numerous previous studies. In studies carried out across a wide variety of different forest types in North America, Japan, Europe, and Australia, lidar-derived canopy structure metrics have been shown to be highly correlated with forest inventory variables. Næsset (1997) reported that forest stand volume could be accurately estimated in 36 spruce (*Picea abies*) stands in Norway using a pool of various canopy height and canopy cover density metrics. Means *et al.* (2000) reported that a variety of stand inventory parameters in a Pacific Northwest forest could be accurately estimated using lidar-derived metrics.

1.3 Use of lidar in a double-sampling forest inventory

Although the utility of lidar as a predictive tool has been demonstrated in previous studies, the issues that arise in using lidar as sampling tool in an operational inventory sampling design have received less attention. Parker and Evans (2004) presented an approach to using lidar in a double-sampling forest inventory design in southern Idaho. In this study, lidar was collected along a strip of plots, where every 5th plot was measured on the ground. Lidar-derived individual tree-based estimates of height and stem density were used to estimate DBH, basal area, and volume for all plots. Næsset (2002) developed a two-stage lidar-based forest sampling procedure in a conifer forest in Norway. This approach used a pool of lidar-based structural metrics at the plot level, and then used stepwise regression techniques to select the best predictive model for the inventory variables. This study found that lidar-based stand-level estimates for all inventory parameters were more precise

that those obtained from conventional techniques. Although these authors found that stand-level estimates were unbiased in most cases (after correcting for the log-transformations), it is likely that regression models developed using fewer plots (e.g. 25-30 plots instead of 35 – 60 plots) will result in biased estimates for small stands within the coverage area. The models that are developed from lidar tend to draw from a large pool of structural metrics, and are often developed using an automated variable selection technique (such as stepwise regression), and therefore may not be representative of the full range of forest conditions within the entire lidar coverage area, potentially leading to bias in parameter estimates for the smaller stands in a given area.

In most forest surveys, the number of plots available for model development is constrained by accessibility and cost. Although efforts are sometimes made to obtain a representative sample, often the sample can be considered a random sample from the population. Although this is certainly a simplification of reality – managers often have previous knowledge of stand conditions and can use this to increase the sample size in more variable stands – for the purposes of this study we will assume that very little *a priori* information is available, beyond stand boundary information. If this sample is used in a double-sampling design with regression, the simple *synthetic* regression estimators for small domains, or stands, typically have low variance, but can have considerable bias due the use of an unrepresentative regression model. Approaches have been developed to reduce the bias in estimates for small domains within a double-sampling design (Särndal and Hidiroglou, 1989). In particular, these authors introduced a *modified* regression estimator that is (approximately) design-unbiased but with increased variance.

In this paper, we will present an investigation of the statistical properties of several lidar-based regression estimators for mean stand biomass, using simulation to estimate the sampling distribution (variance and mean) of these statistics. In particular, we will discuss the use of a synthetic regression estimator, a modified synthetic regression estimator, a

dampened regression estimator, and the possible effect of transformation bias on mean stand biomass estimates.

2. STUDY AREA

2.1 Capitol Forest study area

The study area for this project was a conifer forest within Capitol State Forest, in western Washington state, USA. This forest is composed primarily of Douglas-fir (*Pseudotsuga menziesii*), western hemlock (*Tsuga heterophylla*), and western redcedar (*Thuja plicata*). This area is the site for an ongoing silvicultural trial resulting in a wide variety of residual stand densities and structures, including patch cuts, group selection, heavy thinning, light thinning, clearcut, and control (see Figure 1). The stands used in this study varied in age from 35 to 70 years.

2.2 Field plot data

The USDA Forest Service and University of Washington have established 98 growth plots in each of these stands, as well some surrounding younger stands (Figure 1), with plot sizes ranging from 0.02 ha to 0.2 ha.

Species and diameter were recorded for each tree with diameter greater than 14.2 cm. Total height was measured using a handheld laser rangefinder on a representative subset of these trees, and regression-based height-diameter models were used to estimate height for all unmeasured trees within the plots. In addition, very accurate locations for the plots were acquired through a closed-traverse survey. More detailed information on the plot measurements can be found in Curtis *et al* (2004). Using the measured tree list data, biomass estimates (kg/ha) for each plot were generated using the BIOPAK software package (Means *et al.*, 1994).

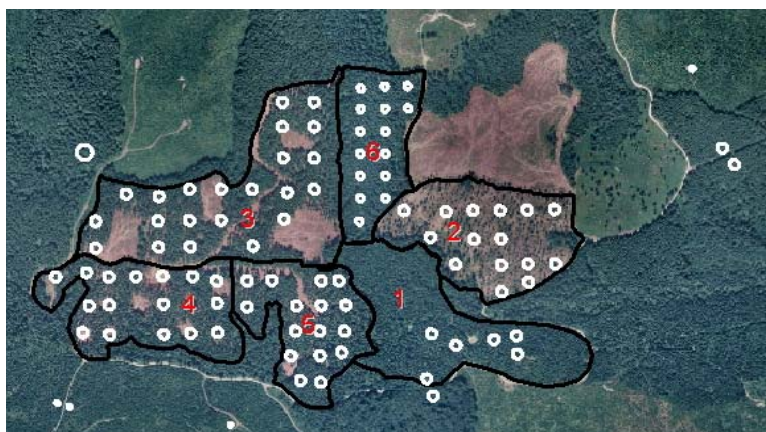


Figure 1. Capitol Forest study area, Washington State, USA. Stand numbers are shown in red, field plots are shown in white.

3. LIDAR DATA

Lidar data were collected over the study area in March, 1999 with a SAAB TopEye system mounted on a helicopter platform.

The details of the lidar acquisition are provided in Andersen *et al.* (2005). The nominal pulse density was 4 returns/m², and the

footprint diameter was approximately 0.4 m. The contractor provided raw lidar point data along with ground returns filtered using a proprietary algorithm.

4. METHODS

4.1 Background

Previous analyses of lidar-based double sampling techniques have used cross-validation (Næsset, 2002) and comparison to the field plot data used in the second stage of the survey (Parker and Evans, 2004) to validate their survey methods. Using a leave-one-out cross-validation procedure, Næsset (2002) assessed the predictive value of the models developed for three different stand types (young, mature(poor site), mature(good site)). This was essentially a test of the predictive quality of the regression models, as opposed to an assessment of the sampling distribution of the regression estimator, since all of the plots (except one) were used to develop the regression models. Parker and Evans (2004) implemented a traditional double-sampling design, where only a limited number of the lidar plots were measured on the ground. The relatively limited number of ground-measured field plots allowed for an assessment of bias, but limited their ability to assess the variance of the regression estimator.

In this study, we used a simulation approach to analyse the sampling distributions of several lidar-based regression estimators of mean stand biomass in the Capitol Forest study area. For the purposes of this study, we assumed that the complete set of 98 plots represented the population, and in each iteration of the simulation, 30 plots were randomly sampled from this population. Using the *R* statistical package, at each iteration a stepwise regression procedure was used to find the (presumed) best fit model relating a suite of lidar-derived, plot-level metrics (mean height (ht), maximum ht, coefficient of variation of heights, 10th percentile ht, 25th percentile ht, 50th percentile ht, 75th percentile ht, 90th percentile ht, and 2-dimensional canopy cover) to the square root of the biomass at the plot (*R*-Development-Core-Team, 2006). Previous analyses had indicated that the square-root transformation was appropriate in the estimation of biomass (Andersen *et al.*, 2006). The predictive model that was selected using the sample of 30 plots was then used to estimate the biomass for all 98 plots in the area. Various estimators of stand biomass (sample mean, synthetic regression estimator (*SY*), modified regression estimator (*MRE*), and dampened regression estimator (*DRE*)) were then generated from these predicted plot-level biomass values. This procedure was repeated for 50000 iterations to develop the sampling distribution of the various estimators. Although all of the plots were available in the model development stage of this study, only stands with multiple plots were used in the stand-wise analysis, giving a total of six stands (Stand 1: 35-yr Douglas-fir, Stand 2: 70-yr Douglas-fir (heavily thinned), Stand 3: 70-yr Douglas-fir (group selection), Stand 4: 70-yr Douglas-fir (patch-cut), Stand 5: 70-yr Douglas-fir (lightly thinned), and Stand 6: 70-yr Douglas-fir (uncut)). The variables selected in each iteration were also observed to assess the stability of the models. Canopy cover was selected as a significant predictor variable in every iteration, while the other selected variables tended to vary among the different height-based metrics (52% of models included 25th percentile ht, 48% of models included mean ht, 40% of models included 50th percentile ht, etc.). Interestingly, the least-used variable was maximum height, possibly due to the generally homogeneous nature of the stands used in this study, where height was much less variable than density, understory density, etc.

4.2 Estimators of mean stand biomass

4.2.1 Single-stage estimator

The single stage estimator of mean stand biomass is the arithmetic mean of plot-level biomass measurements from a given stand, or the sample mean. This estimator is unbiased, but can have a high variance, depending upon the number of plots sampled and the variability of a given stand. Following Särndal and Hidiroglou (1989), U denotes the population of plots $U = \{1, \dots, k, \dots, N\}$ that is divided into D domains (or stands), $U_1, \dots, U_d, \dots, U_D$. If the biomass for a given plot is denoted as y_k , U_d are the plots in U that fall in stand d , and N_d is the size of U_d , then we want to estimate the mean stand biomass

$$t_d = \sum_{k \in U_d} y_k / N_d \quad (1)$$

If s denotes a sample of size n that is drawn randomly from U with inclusion probabilities π_k , then s_d denotes the part of U that falls in stand d . The estimated mean biomass for stand d is then given by: $\hat{t}_d = \sum_{s_d} y_k / n_d$. The sampling distributions for the

single-stage estimate of mean stand biomass for each stand is show in Figure 2.

4.2.2 Lidar-based two-stage regression estimators

The use of auxiliary covariate information obtained over a larger number of plots, or in this case, every element within the population, has the potential to greatly increase the efficiency of an estimator. For example, a vector of lidar-based metrics generated at the plot level can be used to increase the precision of estimates of mean stand biomass. In the case of double-sampling with regression, and again following Särndal and Hidiroglou (1989), a linear regression model is used to relate the variables of interest, y , to x , a vector of correlated variables. If the coefficients of the population linear model of y on x can be denoted as B , then the estimated coefficients are \hat{B} . The predicted values are $\hat{y}_k = x'_k \hat{B}$, and the $e_k = y_k - \hat{y}_k$ are the residuals. The so-called *synthetic regression estimator (SY)* of the mean stand biomass is then given by: $\hat{t}_{dSY} = \sum_{U_d} \hat{y}_k / N_d$. In

cases where the regression model is not representative of the entire population, the synthetic regression estimator can yield estimates for small areas that are significantly biased. In order to reduce this bias, Särndal (1981, 1984) developed the (approximately) design-unbiased estimator:

$$\hat{t}_{dRE} = \frac{\left(\sum_{U_d} \hat{y}_k + \sum_{s_d} e_k / \pi_k \right)}{N_d} \quad (2)$$

This estimator consists of the synthetic regression estimator (the left term in the numerator) and an adjustment term (right term in the numerator) that will correct for bias due to use of an unrepresentative model. However, the variance of the design-unbiased estimator is typically higher than the synthetic estimator, because the adjustment term is, in effect, inflated by the expansion factor π_k . Hidiroglou and Särndal (1985) went on to develop a modified design-unbiased estimator:

$$\hat{t}_{dRE} = \frac{\left(\sum_{U_d} \hat{y}_k + N_d \frac{\sum_{s_d} e_k / \pi_k}{\hat{N}_d} \right)}{N_d} \quad (3)$$

where $\hat{N}_d = \sum_{s_d} 1/\pi_k$. As Särndal and Hidiroglou (1989) point out,

this estimator tends to have smaller variance than the unmodified version because the ratio term will give heavier weight to the adjustment term in cases where the model fit in a particular domain is poor. Unlike the unmodified version, the modified estimator has the additional property that, in the case of simple random sampling, it is consistent as the size of the sample approaches the size of the population, or $\hat{t}_d = t_d$ when $s_d = U_d$. However, these authors also note that in cases where the sample size for a domain is particularly small (e.g. $n_d < 5$), and the model fit is therefore particularly poor in this domain, the modified regression estimator can yield unacceptable results due to the heavy weight given to the adjustment term (for example, negative estimates in cases where the residuals are overwhelmingly negative). Särndal and Hidiroglou (1989) therefore suggest using a *dampened* version of the modified regression estimator:

$$\hat{t}_{dRE} = \frac{\left(\sum_{U_d} \hat{y}_k + \left(\hat{N}_d / N_d \right)^{H-1} \sum_{s_d} e_k / \pi_k \right)}{N_d} \quad (4)$$

where:

$$H = 0 \quad \text{if} \quad \hat{N}_d \geq N_d$$

$$= h \quad \text{if} \quad \hat{N}_d < N_d$$

Previous studies have found that using $h = 2$ provided a reasonable level of dampening. This has the effect of inverting the ratio term when a sample is disproportionately undersampled, giving less weight to the correction term.

4.3 Transformation bias

Typically in a double-sampling framework it is desirable to obtain estimates in the units of the original data. However, simply applying the reverse-transformation of the square-root, or logarithmic, transformation, can result in biased estimates (Næsset, 2002). In the case of the square-root transformation, it has been shown that adding the residual variance (σ^2) to the predicted values can correct for much of this bias (Miller, 1984).

5. RESULTS AND DISCUSSION

The summary statistics (mean, variance) of the simulated sampling distributions for each estimator, and for each stand, are shown in Table 1. It should be noted that the mean coefficient of determination (R^2) values for the 50000 regression models for *sqrt*(biomass) was 0.88, and the standard

deviation of the R^2 values was 0.04. The simulated sampling distributions for the various estimators, and the true mean stand biomass values, are shown in Figures 2-5. The possible influence of transformation bias in converting back to original data units (tons/ha for biomass) is shown in Table 2.

In general, the variance of the single stage estimator is quite high, especially in highly heterogeneous stands (e.g. 3, 4, and 6) (Figure 2). In contrast, in homogeneous stands (e.g. 1 and 2) the sampling error is quite low and even small samples can precisely characterize the population parameter. However, it should be noted that the single-stage mean stand biomass estimates shown here are based only on cases where at least one plot was available in the sample from a given stand, and therefore underestimates the variance of the single-stage estimator, especially in stands with few plots, such as Stand 1 (which was likely unsampled in many of the iterations). As expected, in general the application of the synthetic regression estimator dramatically reduces the variance of the estimator, especially in the more heterogeneous stands (Figure 3). For example, in stand 4, the variance decreased from 3494.8 to 192.7, and in stand 6, the variance decreased from 4569.6 to 892.2. However, as expected, the synthetic estimator's complete reliance on the sometimes ill-fitting regression model led to significant bias for most of the stands (Table 1). This is particularly striking in the case of stand 3, where use of the synthetic estimator led to an 82% reduction in the variance but also introduced a significant 5% bias. Application of the modified design unbiased regression estimator served to dramatically reduce this bias in almost all stands (Figure 4). However, the price of this reduction in bias was a consistent increase in the variance. In general, the variance was still well below that of the single stage estimator. For example, in stand 3, the bias was reduced to 0.5 %, while the variance was reduced to 65% of the variance of the single stage estimator. The form of the dampened estimator appears to moderate both the bias-inducing influence of the synthetic regression term and the variance-inflating effect of the adjustment term (Figure 5). The application of these modified regression estimators may be particularly useful in situations where unbiased estimates are desired for smaller stands within a lidar coverage area.

The results indicate that applying the reverse square-root transformation to recover the original data units does generally lead to a slight negative bias, as we would expect from the explanation in Miller (1984) (Table 2). In all but one stand, application of the bias correction as proposed by Miller (1984) does remove a portion, but not all, of this bias.

	Stand					
	1	2	3	4	5	6
Population mean stand biomass	583.9	311.3	625.7	562.6	620.9	668.2
Single-stage estimator	583.9 (149.8)	311.2 (648.6)	625.7 (894.8)	562.8 (3494.8)	620.7 (418.5)	667.6 (4569.6)
Synthetic regression estimator (SY)	575.0 (965.7)	334.5 (592.3)	594.4 (162.8)	561.9 (192.7)	628.5 (222.2)	692.3 (892.2)
Modified design unbiased regression estimator (MRE)	580.3 (755.5)	315.6 (680.9)	622.3 (315.9)	561.4 (759.5)	622.3 (340.8)	671.4 (2240.1)
Dampened design unbiased regression estimator (DRE)	578.6 (760.2)	321.1 (650.0)	616.4 (298.8)	561.3 (458.3)	623.8 (249.8)	677.1 (1636.9)

Table 1. Statistical properties of (square-root transformed) mean stand biomass estimators (mean (above) and variance (below) of simulated sampling distribution). The stand biomass for the population is shown in the top row.

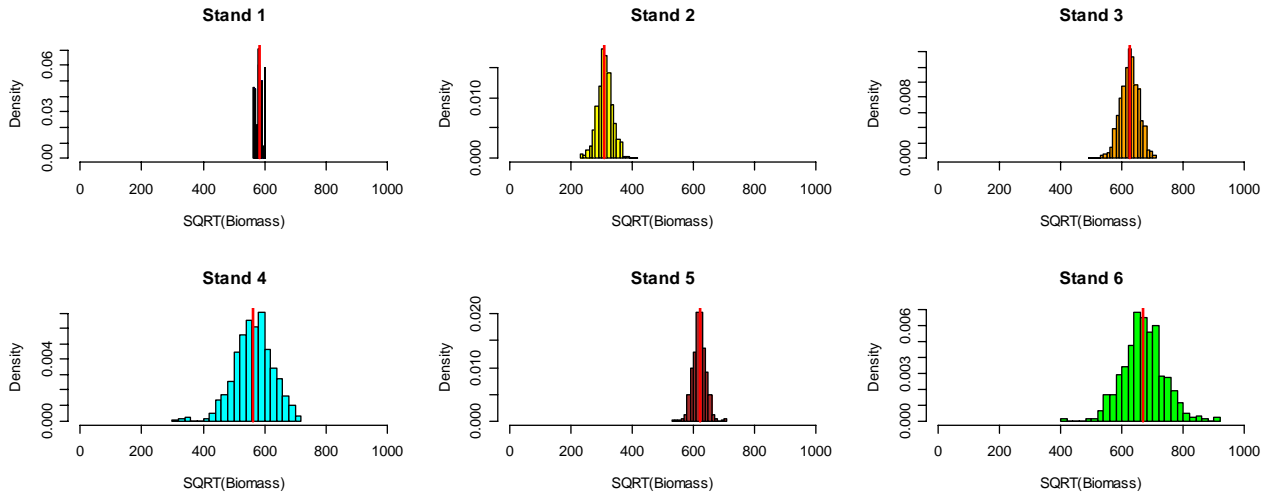


Figure 2. Simulated sampling distributions for the *single-stage estimator* for mean stand biomass. Vertical red line indicates the true mean stand biomass within the population.

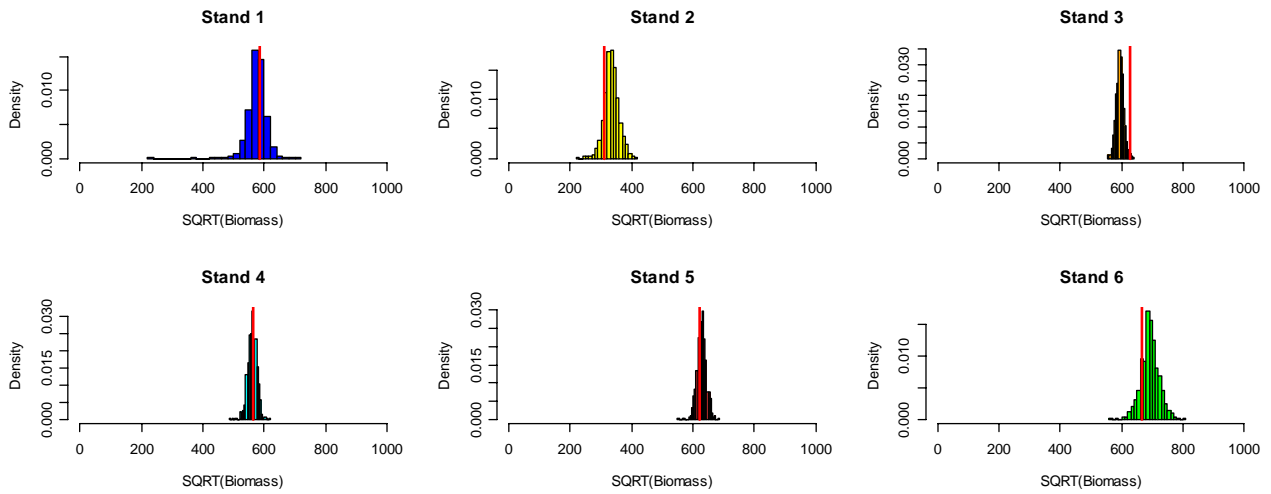


Figure 3. Simulated sampling distributions for the *synthetic regression estimator* for mean stand biomass. Vertical red line indicates the true mean stand biomass within the population.

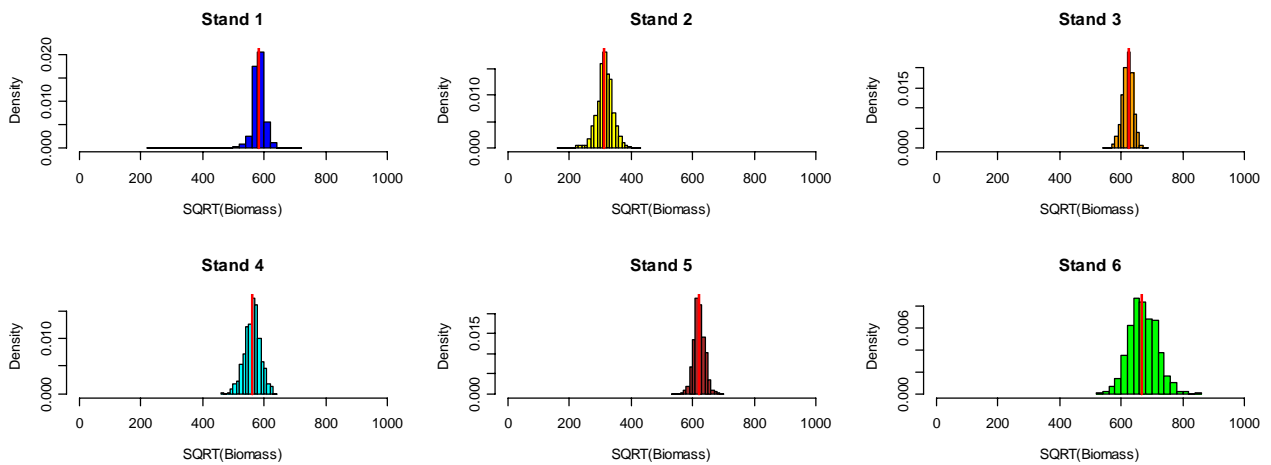


Figure 4. Simulated sampling distributions for the *modified design-unbiased regression estimator* for mean stand biomass. Vertical red line indicates the true mean stand biomass within the population.

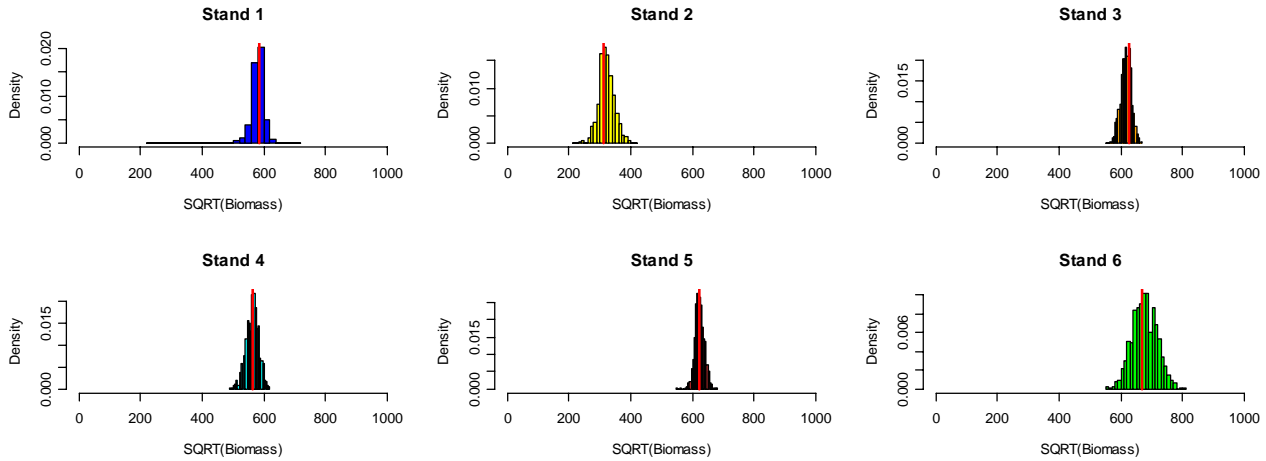


Figure 5. Simulated sampling distributions for the *dampened design-unbiased regression estimator* for mean stand biomass. Vertical red line indicates the true mean stand biomass within the population

	Stand					
	1	2	3	4	5	6
Population mean	341	100	399	335	388	465
Estimate w/o bias correction	336	104	380	316	389	460
Estimate with bias correction	340	107	384	320	392	463

Table 2. Effect of applying reverse square-root transformation to recover original data biomass units (tons/ha).

6. CONCLUSIONS

This investigation confirm the results of previous studies that use of lidar-based regression estimators can significantly increase the precision of estimates for important forest inventory variables, such as mean stand biomass. These results also indicate that use of simple synthetic regression estimators can lead to biased stand-level estimates. The application of a modified regression estimator can reduce the bias at the stand level and will incorporate both the variance-reducing properties of the synthetic regression term and the bias-reducing properties of the correction term.

7. REFERENCES

Andersen, H., R. McGaughey, and S. Reutebuch. 2005. Estimating forest canopy fuel parameters using LIDAR data. *Remote Sensing of Environment* 94, pp. 441-449.

Andersen, H., S. Reutebuch, and R. McGaughey. 2006. Active remote sensing. Chapter 3 in: Shao, G., and K. Reynolds, eds., *Computer Applications in Sustainable Forest Management*. Springer, Dordrecht.

Curtis, R., D. Marshall, D. DeBell, eds. 2004. *Silvicultural options for young-growth Douglas-fir forests: the Capitol Forest study—establishment and first results*. PNW-GTR-598. USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Hidiroglou, M. and C. Särndal. 1985. An empirical study of some regression estimators for small domains. *Survey Methodology* 11, pp. 65-77.

Means, J. H. Hansen, G. Koerper, P. Alaback, M. Klopsch. 1994. *Software for computing plant biomass--BIOPAK users guide*. PNW-GTR-340. USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Means, J. S. Acker, J. Fitt, M. Renslow, L. Emerson, and C. Hendrix. 2000. Predicting forest stand characteristics with airborne scanning lidar. *PE&RS* 66, pp. 1367-1371.

Miller, D. 1984. Reducing transformation bias in curve fitting. *The American Statistician* 38(2), pp. 124-126.

Næsset, E. 1997. Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment* 61, pp. 246-253.

Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning lidar using a practical two-stage procedure and field data. *Remote Sensing of Environment* 80, pp. 88-99.

Paine D. and J. Kiser, 2003. *Aerial photography and image interpretation. 2nd edition*. Wiley, Hoboken.

Parker, R. and D. Evans. 2004. An application of LiDAR in a double-sample forest inventory. *Western J. of Applied Forestry* 19(2), pp. 95-101.

R-Development-Core-Team. 2006. *R: A language and environment for statistical computing*. Technical Report, R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C. 1981. Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse *Bull.of the Int. Stat. Institute* 49, pp. 494-513.

Särndal, C. 1984. Design-consistent versus model-dependent estimators for small domains. *JASA*, 79, pp. 624-631.

Särndal, C., and M. Hidiroglou. 1989. Small domain estimation: A conditional analysis. *JASA*, 84(405), pp. 266-275.