Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway

Göran Ståhl[1]
Sören Holm[1]
Timothy G. Gregoire[2]
Terje Gobakken[3]
Erik Næsset[3]
Ross Nelson[4]

[1] Dept. of Forest Resource Management, Swedish University of Agricultural Sciences, 90183 Umeå Sweden. E-mail: Goran.Stahl@srh.slu.se; Soren.Holm@srh.slu.se

[2] School of Forestry and Environmental Studies, 360 Prospect St., Yale University, New Haven CT06511-2104 USA. E-mail: Timothy.Gregoire@yale.edu

[3] Dept. of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5003, NO-1432, Ås Norway. E-mail: Terje.Gobakken@umb.no; Erik.Naesset@umb.no

[4] 614.4/Biospheric Science Branch, NASA-Goddard Space Flight Center, Greenbelt MD 20771 USA. E-mail: Ross.F.Nelson@nasa.gov

Abstract

In forest inventories regression models are often applied to predict quantities such as biomass at the level of sampling units. In this paper we propose a model-based inference framework for combining sampling and model errors in the variance estimation. It was applied to airborne laser (LiDAR) datasets from Hedmark County (Norway) where the model error proportion of the total variance was found to be large for both scanning (ALS; airborne laser scanning) and profiling LiDAR when biomass was estimated. With profiling LiDAR, the model error variance component for the entire county was as large as 71% whereas for ALS it was 43% of the total variance. Partly, this reflects the better accuracy of the pixel-based regression models estimated from scanner data as compared to the models estimated from profiler data. The framework proposed in our study can be applied in all kinds of sample surveys where model-based predictions are made at the level of individual sampling units. Especially, it should be useful in cases where model-assisted inference cannot be applied due to the lack of a probability sample from the target population, or due to problems of correctly matching observations of auxiliary and target variables.

Keywords: forest inventory, model-based inference, regression estimation, scanning laser, profiling laser, variance

Introduction

In forest inventories, quantities at the level of sampling units are often predicted using regression models. Well known examples include volume and biomass models based on measurements of diameter and height on individual trees (e.g. Marklund 1988). Similarly, in inventories based on airborne lasers (LiDAR; light detection and ranging) regression models are applied to predict per-hectare biomass or volume based on measurements derived from the lasers (e.g. Nelson et al. 1988, 2003a; Næsset 1997).

Standard sampling theory assumes that the variables entering the estimators are observed without error (e.g. Gregoire & Valentine 2008). If measurement errors or model -related errors cannot be ignored, uncertainty estimates such as the variance typically are underestimated when standard formulas are applied (e.g. Särndal et al. 1992). Thus, the reported figures are less trustworthy than they appear to be and, as a result, inappropriate decisions may be made based on the information. In some studies (e.g. Gertner & Köhl 1992, Gertner et al. 2002) error budgets are compiled where the contribution of different error sources to the total error (often expressed as variance or mean square error) is assessed. The importance of proper handling of model-related errors also is pointed out by the Intergovernmental Panel on Climate Change (IPCC 2003) in their good practice guidance for greenhouse gas reporting for the land use, land-use change, and forestry sector of the United Nations Climate Change Convention. However, in many forest surveys model related errors are ignored and uncertainty estimates are provided as if predicted quantities were true.

LiDAR based forest surveys have evolved rapidly over the last decade (e.g., Næsset 1997, Means et al. 2000, Holmgren 2004). The strength of the technique in the context of forest inventory is that very detailed information about canopy height and cover can be obtained through measurements of time differences in the returns of laser pulses, emitted from air- or spacecrafts, reflected from the canopy and from the ground. Two different approaches using so-called small-footprint laser data have been developed and demonstrated in operational projects, namely (i) the use of airborne profiling lasers designed for sampling-based inventories (Nelson et al. 2003a, 2004), and (ii) the use of airborne scanning lasers (ALS) for wall-to-wall mapping of forest stands for practical forest planning (Næsset & Bjerknes 2001, Næsset 2002). Today, it is common practice in many countries to apply ALS when stand level information for forest management plans is compiled (e.g. Næsset 2004b, 2007). The profiling system developed at NASA by Nelson et al. (2003b), labeled "Portable Airborne Laser System" (PALS), is an inexpensive and simple device which can be operated at low costs. A profiling system only collects a narrow line of data on the ground, and does not provide data for wall-to-wall mapping. In contrast to this, each flight-line of a scanning system typically has a swath width of, say, 500-1000 m. Thus, scanning systems can provide data for continuous mapping of the forest and are therefore ideal for estimating properties of the forest at stand level. However, ALS can also be used as a strip sampling tool to inventory timber volume and biomass in large areas.

Applications where LiDAR measurements are used in the context of sampling surveys currently are gaining increased interest (e.g. Nelson et al. 2004, Parker & Evans 2004, Andersen & Breidenbach 2007). In such applications, standard sampling estimators and variance estimators are problematic due to the complex structure of the surveys, where long lines or belts may extend over several strata (e.g. Nelson et al. 2008). Further, many different sources of errors are involved, and model-related errors need to be specifically accounted for in the uncertainty analysis or model-assisted estimators (Särndal et al. 1992) be applied. The latter approach requires that there is a sound probability sample of population units, with measured target variables, available from the area of interest, which is not always the case. The reason may be that already existing biomass models are applied in new surveys or that good matching of units from field sampling and remote sensing cannot be achieved.

The objective of this study was to develop and apply a general framework for model-based estimation and error assessment, accounting for both sampling and model errors, in cases where regression models are applied to predict the target variables. Especially, the framework should be useful in cases where model-assisted estimators cannot be applied. The study was based on LiDAR data from Hedmark County, Norway, where both scanning and profiling laser data had been acquired in order to assess biomass resources. While model-based inference may encompass many different approaches (e.g.

Gregoire 1998, McRoberts 2010), in this study we applied regression models at the level of sampling units that had been selected through probability sampling.

Materials and Methods

The Hedmark County survey

The study area was Hedmark County, in southeastern Norway on the Swedish border (Fig. 1). The total area of the county is 27390 km^2. There are 2309 permanent National Forest Inventory (NFI) sample plots available in this area, distributed systematically in a 3x3 km grid; each plot is circular and has a size of 250 m^2. The measurement protocol stipulates that 20% of the permanent plots are re-measured every year; this selection is made according to a Latin square design within a 45x45 km block of plots.

On each sample plot, all trees with diameter at breast height (DBH) ≥5 cm were callipered and tree heights were measured on an average of 10 sample trees per plot. Trees with DBH <5 cm (but taller than 1.3 m) were counted, and their diameters estimated by means of models (Tomter 1998). Total above ground dry biomass of all trees taller than 1.3 m was then estimated according to species-wise allometric models with DBH and height as predictor variables (Marklund 1988). The coordinates of each plot center were determined with an average accuracy <1 m using differential Global Positioning System and Global Navigation Satellite System measurements according to the procedures suggested by Næsset (2001).

The county was stratified into eight strata based on existing land use maps and Landsat satellite images. The eight strata included four productive forest classes, i.e., (1) high, (2) medium, and (3) low productivity forests and (4) young forest. The remaining four strata were either nonproductive forest or non-forest classes, i.e., (5) nonproductive forest, (6) mountain areas >850 m above sea level, (7) developed areas, e.g., residential areas and infrastructure, and (8) open water. Both profiling and scanning laser data were collected during the summer of 2006. The flight lines were flown east-west and followed the NFI grid (Fig. 2), for practical reasons without any account for the stratification. Observations from NFI sample plots measured in the period from 2005 to 2007 were used as ground training data to construct models of relationships between the airborne laser data and aboveground biomass as determined from field measurements.

Laser profiling

In total, 105 profiling flight lines totaling 9166 km were flown; 763 NFI plots were overflown within 17.8 m of plot center. Profiling laser measurements of canopy height and crown density, similar in form to ALS variables described by Næsset and Gobakken (2008, Section 2.4), were extracted along fixed-length, 17.8 m segments closest to the center of the ground plot and related to total aboveground biomass. The segment dimension, 17.8 m, was defined by the diameter of a 250 m^2 circular ground plot.

The eight-class stratification proved problematic with respect to analysis of the profiling LiDAR data. Due to the variability in data, it was found that only rather approximate predictive biomass models could be developed for each of the four productive forest classes. As a consequence, one generic linear model ($R^2 = 0.59$, RMSE = 39 t ha^{-1}) was developed across all productive forest classes. For the four nonproductive forest/ non-forest classes, the R^2 values ranged between 0.46 and 0.64, with RMSEs between 12.3 and 19.8 t ha^{-1}. Profiling results, then, are reported for productive forest and the four nonproductive/non-forest classes based on estimates made on each of the 17.8 m segments on all flight lines.

In all the models, biomass (t ha^{-1}) was the dependent variable and heights to the 40th, 60th and 90th canopy cover deciles the independent variables together with average canopy height, quadratic mean canopy height, and standard deviation of canopy heights. Standard linear regression was applied. For each model, the independent variables providing the best fit were selected. Typically, two independent variables were included in each model: one variable related to decile height and one to either average height, standard deviation of heights, or quadratic mean height.

Laser scanning

Fifty-three flight lines were flown with the scanning laser with an inter-line distance of 6 km, which means that approximately 50% of the available plots were covered by ALS data (Fig. 2). In total, 4570 km were flown covering 2297 km^2 or 8.4% of the county's area. The average pulse density for the scanner was 2.8 pulses m^{-2}. Only first echoes were used in this study. Non-linear aboveground biomass models were estimated for 7 of the 8 strata. From 30 to 151 sample plots were located in each stratum, except for water where no plots were measured. The estimated models were back-transformed and the calculated RMSE values then ranged between 9.6 and 23.9 t ha^{-1}. The squared Pearson correlation coefficients between observed and estimated biomass after back-transform ranged between 0.79 and 0.92.

All the models included both canopy height metrics and canopy density metrics derived from the ALS data. Altitude was also included in four of the models. The selected models contained from two to four explanatory variables. Typical canopy height metrics were the upper deciles of the canopy height distributions. The canopy density metrics were computed by dividing the respective canopy returns into 10 different vertical layers of equal height. The height of each layer was defined as one tenth of the distance between the 95% percentile and the lowest canopy height (1.3 m). The canopy densities were then computed as the proportion of number of returns above a given layer to total number of returns including those below 1.3 m, see Næsset (2004a) for further details.

Each flight line or strip was divided into regular 250 m^2 grid cells and the cells were allocated to strata. The laser echoes that belonged to each grid cell were used to derive the same canopy height and -density metrics as derived for the NFI plots. The estimated models were used to predict biomass for each cell within 500 m wide belts with the center lines of the ALS scans being the center lines of the belts. The model developed for medium site productivity was used also for predicting biomass for each cell allocated to the stratum water.

Statistical methods

In the following sections, basic estimators, variances, and variance estimators will be derived in steps. First, we address a standard simple random sampling framework where sampling and model errors are combined. Then, this is expanded to account also for cluster sampling and stratification, which were important features of the Hedmark County LiDAR survey.

The following basic setup of an inventory, such as the one in Hedmark, is assumed:

- A first sample, S1, is acquired by simple random sampling. This is the 'application sample' to which we apply the regression models developed based on the S2 sample (see below). Thus, from the S1 sampling units proxy variables are acquired and used as independent variables in a regression model to predict the target variable on each sampling unit. In Hedmark, the proxy variables were the ALS or profiling laser measurements.
- A second sample, S2, is taken where measurements of both target and proxy variables are made. In Hedmark, our S2 sample consists of all ground sample plots and the corresponding metrics derived from the profiling or scanning laser measurements.

We assume that there is only one model step involved, i.e. that biomass is predicted directly based on proxy data at the level of sample plots. Moreover, in the basic set-up we assume that the samples S1 and S2 are independent.

All derivations are made in a model-based context, assuming that a population model $Y(x) = g(x, \alpha, \varepsilon)$ is available. Here, x is the vector of independent variables, α the vector of parameters, and ε the deviation from the true value. The form of the expected value model $E(Y \mid x) = E\ g(x, \alpha, \varepsilon) = g(x, \alpha)$ is assumed to be known. In practice, this model can be fitted using linear or non-linear regression, including back-transformation and correction for transformation bias in case the dependent variable was transformed. Such corrections are available for many common transformations, e.g. Miller (1986, §1.2.3)).

Below, the x-vector variables are denoted x_{i1}, *i*=1,..., *m*, where *m* is the number of sampling units in S1. The sample S2 is assumed to be acquired in a manner appropriate for estimating the parameters in the vector α. According to the assumptions of regression analysis, this can be done in many different ways, ranging from purposive to random sampling [e.g. Royall (1970), Royall & Herson (1973)]. The observations are denoted with index '2' and the model $y_{i2} = g(x_{i2}, \alpha, \varepsilon_{i2})$ is assumed to be valid; the parameters α are estimated as $\hat{\alpha}$. Throughout the derivations, we condition on the sample S2.

True population mean

The true finite population mean can be written

$$(1) \qquad \tilde{\mu}_Y = \frac{1}{M}\sum_{i=1}^{M} \tilde{g}(x_i, \alpha, \varepsilon_i)$$

where *M* is the population size. For large *M,* $\tilde{\mu}_Y$ will differ only negligibly from the population model mean

$$(2) \qquad \mu_Y = \frac{1}{M}\sum_{i=1}^{M} g(x_{i1}, \alpha)$$

obtained by taking expectation with respect to ε . Throughout this study, we adopt μ_Y as definition of the population mean.

Estimation

A straightforward model-based estimator of the population mean, μ_Y, following simple random sampling of size *m* (the sample size of S1) is

$$(3) \qquad \hat{\mu}_Y = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1}, \hat{\alpha})$$

A total is applied by multiplying with the known population size. Our ambition now is to derive a formula for the variance, and a variance estimator (although omitting finite population correction).

Variance and variance estimation

We assume that our α-estimates are accurate enough so that we can linearize the g-model in the neighborhood of the true value, and use the first and second moments of the linear function as proxies for the true moments.

$$(4)\; g(x,\hat{\alpha}) \approx g(x,\alpha) + (\hat{\alpha}_1 - \alpha_1)\cdot g_1'(x,\alpha) + (\hat{\alpha}_2 - \alpha_2)\cdot g_2'(x,\alpha) + \ldots + (\hat{\alpha}_p - \alpha_p)\cdot g_p'(x,\alpha)$$

where $g_j'(x,\alpha) = \dfrac{\partial g(x,\alpha)}{\partial \alpha_j}$; p is the number of parameters. Expected values, etc., are then taken considering both the distribution of $\hat{\alpha}$-estimates and the S1 sample. The estimator of $\hat{\alpha}$ is independent of all x_{i1}-values, since the sample S2 is assumed taken independently of S1. All $\hat{\alpha}_j$ are further assumed to be unbiased, or approximately so. According to the assumed model, $E(\hat{\mu}_Y) = \mu_Y$. Furthermore,

$$(5)\;\; \hat{\mu}_Y - \mu_Y = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1},\hat{\alpha}) - \mu_Y = \left[\frac{1}{m}\sum_{i=1}^{m} g(x_{i1},\alpha) - \mu_Y\right] + \frac{1}{m}\sum_{i=1}^{m}(g(x_{i1},\hat{\alpha}) - g(x_{i1},\alpha)) =$$

$$= D_1 + D_2$$

where D_1 is the term within brackets and D_2 the second sum

D_1 and D_2 are (at least approximately) uncorrelated, and thus the variance for each term can be derived separately and the variances added. The details are provided in Appendix 1. The resulting variance is

$$(6) \qquad V(\hat{\mu}_Y) = \frac{1}{m}\cdot\sigma_g^2 + \sum_{j}^{p}\sum_{k}^{p} Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot E_{S1}(\bar{g}_j'\bar{g}_k')$$

The first term arises from the sampling error due to S1, while the second term arises from the effects of the uncertainty of the $\hat{\alpha}$ -estimates (which is related to the selection of S2) on D_2 . In equation (6), σ_g^2 is the population variance of the g-values and the $\bar{g}_i^{'}$ -terms are the average values of the first order derivatives of the g-function. Here and elsewhere, the indices $S1$ and $S2$ denote the sample within which the moments are considered. An estimator of this variance is (see Appendix 1):

$$(7) \qquad \hat{V}(\hat{\mu}_Y) = \frac{1}{m} s_{\hat{g}}^2 + \sum_{j=1}^{p} \sum_{k=1}^{p} C\hat{o}v_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot \hat{\bar{g}}_j' \cdot \hat{\bar{g}}_k'$$

In this formula, s_g^2 is the sample based estimate of the population variance of the g-values. The covariance term is estimated from the sample S2.

Cluster sampling

We now expand the set-up to accommodate the important case where the model is applied at the level of population elements selected through cluster sampling (in the S1 sample). This is the case when LiDAR sampling lines (profiler) or strips (ALS) are divided into smaller pieces (profiling 17.8 m segments or ALS 250 m^2 cells in the Hedmark case) for which biomass is predicted using a model (or when models for biomass or volume are applied to trees on sample plots).

We introduce the notation *G* for cluster totals. In the context of LiDAR sampling, a cluster is a flight line, i.e., the total of biomass estimates made on all segments (profiler) or cells (ALS) in a given laser flight line. The 'average of the cluster totals' then can be expressed in a form similar to (3) as

$$(8) \qquad \hat{\mu}_Y = \frac{1}{m} \sum_{i=1}^{m} G_i(\hat{\alpha}) = \frac{1}{m} \sum_{i=1}^{m} \hat{G}_i$$

where $\hat{G}_i = \sum_{t=1}^{T_i} g(x_{it1}, \hat{\alpha})$ is the sum of the $g(x, \hat{\alpha}) -$values for the T_i objects within the *i*th cluster. The average from (8) can easily be converted to total biomass or average biomass per hectare for the region of interest. Note that there is a weighting mechanism implicit in the consideration of cluster totals. Longer flight lines contain more segments or cells, thereby contributing larger numbers of biomass observations to the flight line (or cluster) totals.

Variances can be derived following the same logic as in the former section, and thus the following variance formula is obtained:

$$(9) \qquad V(\hat{\mu}_Y) = \frac{1}{m} \cdot \sigma_G^2 + \sum_{j}^{p} \sum_{k}^{p} Cov_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot E_{S1}(\overline{G}_j' \overline{G}_k')$$

Here σ_G^2 is the variance of cluster totals in the population and $\overline{G}_j' = \frac{1}{m} \sum_{i=1}^{m} G_i'(\alpha) =$

$= \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{T_i} g_j'(x_{it}, \alpha)$. Further, a variance estimator can be derived along the previous lines of derivation, using the notation $G_i = \sum_{i=1}^{m} g(x_{it}, \alpha)$ instead of $g(x_i, \alpha)$, $\hat{G}_i = \sum_{i=1}^{m} g(x_{it}, \hat{\alpha})$ instead of $g(x_i, \hat{\alpha})$, $\overline{G}_j' = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{T_i} g_j'(x_{it}, \alpha)$ instead of $\overline{g}_j'$, and $\hat{\overline{G}}_j' = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{T_i} g_j'(x_{it}, \hat{\alpha})$ instead of $\hat{\overline{g}}_j'$. The resulting variance estimator is:

$$(10) \qquad \hat{V}(\hat{\mu}_Y) = \frac{1}{m} \cdot s_{\hat{G}}^2 + \sum_{j=1}^{p} \sum_{k=1}^{p} C\hat{o}v_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot \hat{\overline{G}}_j' \cdot \hat{\overline{G}}_k'$$

Stratification and post-stratification

In case of ordinary stratification (e.g. Gregoire and Valentine 2008) samples are selected independently in different strata. If this is the case, and if separate regression models are used for predictions in each stratum, stratified sampling can be handled easily by making separate estimates (of population totals and variances) for each stratum and then add the estimates to obtain overall estimates.

However, an important feature in many large-scale forest surveys, such as the LiDAR surveys in Hedmark, is that strata are formed based on available map data but sampling units are distributed independently of this stratification, denoted post-stratification in our study. In this case there will be dependencies between the estimates from different strata, due to the fact that some cluster sampling units extend over several strata.

An estimator of the mean value μ_{Yh} in stratum h is:

(11)
$$\hat{\mu}_{Yh} = \frac{\frac{1}{m}\sum_{i=1}^{m} G_{ih}(\hat{\alpha}_h)}{\frac{1}{m}\sum_{i=1}^{m} A_{ih}}$$

In this formula, the summation extends over all clusters just like in the previous cases (e.g. Eq. 8) but the cluster totals are computed only based on the sub-units belonging to stratum h (if a stratum is not present in a cluster this quantity is zero). The variable A_{ih} is the area (ALS) or length (profiler) of stratum h within cluster i. The $\hat{\alpha}_h$ define the regression model used in stratum h.

The corresponding mean across all H strata, which can then be multiplied with known total area to obtain an overall total, is:

(12)
$$\hat{\mu}_Y = \sum_{h=1}^{H} W_h \, \hat{\mu}_{Yh}$$

Here, W_h is the known area proportion of stratum h from, for instance, the digital map used to stratify the entire study area. The derivations of a variance formula and a variance estimator of (12) are provided in Appendix 1. The variance is, approximately:

$$(13) \quad V(\hat{\mu}_Y) = \frac{1}{m}\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{Cov(G_h(\alpha_h) - \mu_{Yh}A_h, G_k(\alpha_k) - \mu_{Yk}A_k)}{E(A_h)E(A_k)}W_h W_k +$$

$$+\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{W_h W_k}{E(A_h)E(A_k)}\sum_{j_1}^{p_h}\sum_{j_2}^{p_k}Cov_{S2}(\hat{\alpha}_{j_1h}, \hat{\alpha}_{j_2k})E_{S1}(\overline{G}'_{j_1h}\overline{G}'_{j_2k})$$

The first part corresponds to the sampling error and the second part to the errors introduced due to the uncertainties of the parameter estimates of the regression model. An estimator of this variance is:

$$(14) \quad \hat{V}(\hat{\mu}) = \frac{1}{m}\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{W_h W_k}{\overline{A}_n \overline{A}_k}\frac{\sum_{i=1}^{m}(G_{ih}(\hat{\alpha}_h) - \hat{\mu}_{Yh}A_{ih})(G_{ik}(\hat{\alpha}_k) - \hat{\mu}_{Yk}A_{ik})}{m-1} +$$

$$+\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{W_h W_k}{\overline{A}_h \overline{A}_k}\sum_{j_1}^{p_h}\sum_{j_2}^{p_k}C\hat{o}v_{S2}(\hat{\alpha}_{j_1h}, \hat{\alpha}_{j_2k})\,\hat{\overline{G}}'_{j_1h}\,\hat{\overline{G}}'_{j_2k}$$

The covariances of the parameter estimates are important. If separate models have been derived for each stratum based on independent datasets, then all cross-stratum covariances are zero. If the same model is applied in several (or all) strata, then cross-stratum covariances should be included.

For an individual stratum, the variance estimator is:

$$(15) \quad \hat{V}(\hat{\mu}_{Yh}) \quad = \quad \frac{1}{\overline{A}_h^2}\frac{\sum_{i=1}^{m}(G_{ih}(\hat{\alpha}_h) - \hat{\mu}_{Yh}A_{ih})^2}{m(m-1)} \quad + \frac{1}{\overline{A}_h^2}\sum_{j_1}^{p_h}\sum_{j_2}^{p_h}C\hat{o}v_{S2}(\hat{\alpha}_{j_1h}, \hat{\alpha}_{j_2h})\,\hat{\overline{G}}'_{j_1h}\,\hat{\overline{G}}'_{j_2h}$$

Results

The estimation framework described above was applied to data from the Hedmark County survey, using both ALS and PALS data. The results are summarized in Table 1 where both overall and stratum-level results are provided. Specifically, the total variance was separated into model and sampling error components, in order to illustrate the magnitude of the different sources of variability.

Laser Profiling

We note the following based on the results presented in Table 1:

1. The estimates based on profiler data are about 10% smaller than the corresponding ground-based estimates.

2. In all five strata where separate models were developed, more than 50% of the profiler variance is due to variability associated with model parameter estimation. The proportion of model variance is large; in three of the five profiling strata, model variance accounts for more than 90% of the total variance.

3. In all five strata where profiler versus ALS model variance proportions can be compared, the ALS model variance component is consistently smaller than the profiler's.

4. Considering the productive forest class, the profiler standard error is larger than the ground-based standard error.

Point 1 speaks to accuracy, and points 2-4 speak to the precision of the profiling LiDAR estimates. The larger proportion of model error variance for the profiler reflects the combined effects of (1) the better laser pulse geolocation accuracy of the scanning system, (2) the fact that the scanner acquires ranging measurements across an entire ground plot whereas the profiler measures only a linear slice in the proximity of a given ground plot, and (3) the fact that the profiler flew twice as many flight lines as the scanner. Factors (1) and (2) result in smaller ALS model errors; factor (3) reduces the profiler sampling error.

The Table 1 results quantitatively describe the limitations of the PALS profiling LiDAR and suggest two items that should be addressed in order to improve profiling results. First, the geolocation accuracy of the individual LiDAR pulses must be improved (see Gobakken and Næsset (2009) for description of effects of ALS-ground misregistration). Second, alternative model forms should be considered, e.g., ln-ln models (e.g., Næsset 2002) or the square root models utilized by Andersen and Breidenbach (2007) and Boudreau et al. (2008), see also Gregoire et al. (2008) for an appropriate back-transformation correction. Model error is simply overwhelming the profiling error term, and addressing the two points above may decrease the model error term and the total variance.

Laser Scanning

We note the following based on the ALS-results presented in Table 1:

1. The ALS based estimates are slightly larger than the corresponding ground-based estimates. The differences range between 0 and 10% in the different productive forest strata; in average, the difference is about 5%. However, larger differences were found in nonproductive forests and mountain areas, where the ALS based estimates were considerably smaller than the ground-based estimates, respectively. For the entire county the ALS based biomass estimate is very close to the corresponding ground-based estimate.

2. In total 43% of the ALS variance is due to variability associated with model parameter estimation. For the productive forest, the model variance accounts for 42% and it varies between 58% and 85% for the four productive forest strata.

3. Even if the model variance component is consistently smaller than the profiler's, the proportion of model variance is high.

4. In general the ALS standard errors are smaller than the ground-based standard errors, however, the differences are not very large.

The LiDAR-based biomass estimates in mountain areas and in developed areas were considerably smaller than the corresponding ground-based estimates. We argue that the lack of correspondence for these strata was due to the fact that the NFI has been focusing on productive forest and only measured sample plots located close to the productive forest areas. Thus, no representative ground-based samples were available for these areas.

Discussion

The proposed model-based framework for biomass estimation based on LiDAR data is only one out of several possible estimation approaches. One alternative would be a model-assisted framework (Särndal et al. 1992) where the models would be used to provide proxy values to which adjustments based on actual measurements are applied to ensure unbiased estimates; this approach was adopted by Andersen and Breidenbach (2007) and further by Gregoire et al. (2010). The model-assisted approach has the advantage of staying within the realm of design-based estimation, but it relies on the availability of a sound (probability-based) subsample of ground plots within the target area and good geographical matching between ground and LiDAR data.

Some features of the Hedmark study complicated usage of model-assisted estimation, e.g. that a random sample of field plots was not available in all regions and that location errors in the PALS data sometimes made it difficult to match field data and LiDAR data. In previous studies with profiling lasers (Nelson et al. 2003a; 2004; 2008) the S1 and the S2 samples have been independent; in such cases a model-based approach would be the only straightforward estimation framework.

The empirical findings illustrate that profiling LiDAR may be less adequate for large area sampling than ALS unless predictive models associated with the profiler can be significantly improved. ALS systems seem to be more efficient with respect to collecting LiDAR measurements over existing ground plots as the wide swath of a scanning system in most cases will ensure overlapping measurement of the plots from the air and on the ground. However, we consider the further development of sampling designs and applications that allow regional estimates based on profiling LiDAR observations to be of great importance because, for the next decade, profiling LiDARs will be the only space LiDARs available capable of providing continental perspectives; the NASA ICESat II and DESDynI are the only space LiDAR missions currently under consideration for launch - both are currently configured as profilers. The presented estimation framework is applicable for profiling as well as scanning LiDAR applications and thus should allow for timely and rapid biomass assessments at several geographical scales, from regional to continental and even global levels.

However, there are several possibilities to further improve the proposed methodological framework. While the current approach accounts for the parameter estimation uncertainty in one model step, cluster sampling, and stratification/post-stratification, it could be further developed to cover also prediction errors for individual units, systematic sampling, and additional model steps. Inclusion of prediction errors for individual units would be particularly relevant for estimates within smaller regions. They would be straightforward to include when random sampling of single population units is applied, although with cluster sampling the dependencies between elements within clusters would lead to complications. Systematic sampling only would affect the sampling error part of the variance, and methods such as successive differencing (e.g. Wolter 1984) probably could be applied to accommodate this effect. Additional model steps also could be included along similar lines as the first model step.

There are several forest inventory cases where the proposed framework would be straightforward to apply; one example is field-based inventories where volume or biomass models are used for tree-level predictions on sample plots. These models are normally developed independently of the application sample. Also, the formulas allow for investigations into the trade-offs between sample sizes in applications and sample sizes for developing the regression models. Clearly, poor models applied to large samples will lead to small sampling error components and large model error components.

It should be noted that the model-based estimators derived require that the complete (estimated) variance-covariance matrix of the estimators of the regression parameters is known. This requirement may be difficult to meet in cases where existing regression models are applied, and where there is no access to data so that the covariance matrices can be computed (if they are not reported).

Although the results of the estimation framework appeared correct and logical when applied to LiDAR data, a small additional simulation study was carried out in order to check for the correctness of some of the basic formulas. The variance estimator (7) was evaluated based on the model: $y_i = \exp(\alpha + \beta \cdot x_i + \varepsilon_i)$, with $\alpha = \ln(0.1)$, $\beta = 2.5$ and $\sigma_\varepsilon = 0.5$, and different distributions for *x* where $5 \le d = \exp(x) \le 45$. Also, two different selection methods regarding the sample S2 (simple random sampling and PPS sampling, using *d* as the 'size' variable) were applied. This model can be seen as a fair approximation of volume or biomass (*y*) as a function of diameter (*d*). The parameter estimates were obtained following a linearization through logarithmic transformation. In each repetition of the simulation new S1 and S2 samples were selected. The distributions for *d* were (i) rectangular and (ii) half-triangular with *d*=5 nine times as frequent as *d*=45. In all cases evaluated the mean of the variance estimator corresponded closely to the simulated (true) variance, indicating a solid performance of the proposed variance estimator.

We conclude that the proposed model-based framework should be very useful in inventory programs where regression models are used to predict the quantities of interest at the level of individual sampling units. The application described in this paper – LiDAR-based estimation of biomass – is an important example, which demonstrated that the model error contribution to the total variance may often be substantial. Especially, the proposed framework should be useful in cases where model-assisted inference cannot be applied; the reasons may be either that models have been developed from an independent dataset or that matching of the samples for model development and application cannot be fully achieved, as was the case when profiling LiDAR was applied in Hedmark.

References

Andersen, H.-E. and Breidenbach, J. 2007. Statistical properties of mean stand biomass estimators in a LIDAR-based double sampling forest survey design. *Proceedings of the ISPRS Worskhop Laser Scanning, 12-14 September 2007, Espoo, Finland*. IAPRS, Volume XXXVI, Part 3 / W52, 2007, pp. 8-13.

Boudreau, J., Nelson, R.F., Margolis, H.A., Beaudoin, A., Guindon, L., and Kimes, D.S. 2008. Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. Remote Sens. Environ. 112: 3876-3890.

Gertner, G.Z. and Köhl, M., 1992. An assessment of some nonsampling errors in a national survey using an error budget. For. Sci. 38: 525–538.

Gertner, G.Z., Wang, G., Fang, S., and Anderson, A. 2002. Error budget assessment of the effect of DEM spatial resolution in predicting topographical factor for soil loss estimation. *Journal of Soil and Water Conservation* 57: 164-174.

Gobakken, T. and Næsset, E. 2009. Assessing effects of positioning errors and sample plot size on biophysical stand properties derived from airborne Laser scanner data. Can. J. For. Res. 39: 1036–1052

Gregoire, T. G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28: 1429-1447.

Gregoire, T.G., Lin, Q.F., Boudreau, J., and Nelson, R. 2008. Regression Estimation Following the Square-Root Transformation of the Response. For. Sci. 54: 597-606.

Gregoire, T.G. and Valentine, H.T. 2008. Sampling strategies for natural resources and the environment. Chapman & Hall, Boca Raton. 474 pp.

Gregoire, T.G., Ståhl, G., Naesset, E., Gobakken, T., Nelson, R., and Holm, S. 2010. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Can J. For. Res. (in press.)

Holmgren, J. 2004. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. Scand. J. For. Res. 19: 543-553.

IPCC, 2003. Good practice guidance for land use, land-use change, and forestry. ISBN 4-88788-003-0.

Marklund, L.G. 1988. Biomass functions for pine, spruce and birch in Sweden. Swedish University of Agricultural Sciences, Department of Forest Survey.Umeå. (In Swedish.).

McRoberts, R.E. 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment* 114: 1017-1025.

Means, J. E., Acker, S. A., Brandon, J. F., Renslow, M., Emerson, L. & Hendrix, C. J. 2000. Predicting forest stand characteristics with airborne scanning lidar. Photogrammetric Engineering and Remote Sensing 66: 1367-1371.

Miller, R. G. 1986. *Beyond ANOVA: Basis of Applied Statistics*. Wiley: New York.

Næsset, E. 1997. Estimating timber volume of forest stands using airborne laser scanner data. Remote Sens. Environ. 61: 246-253.

Næsset, E. 2001. Effects of differential single- and dual-frequency GPS and GLONASS observations on point accuracy under forest canopies. Photogramm. Eng. Remote Sensing 67: 1021-1026.

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

Næsset, E. 2004a. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scand. J. For. Res.19: 164-179.

Næsset, E. 2004b. Accuracy of forest inventory using airborne laser-scanning: evaluating the first Nordic full-scale operational project. Scand. J. For. Res. 19: 554-557.

Næsset, E. 2007. Airborne laser scanning as a method in operational forest inventory: Status of accuracy assessments accomplished in Scandinavia. Scand. J. For. Res. 22: 433 - 442.

Næsset, E. & Bjerknes, K.-O. 2001. Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. Remote Sens. Environ. 78: 328-340.

Næsset, E. & Gobakken, T. 2008. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sens. Environ. 112: 3079-3090.

Nelson, R., Krabill, W., and Tonelli, J. 1988. Estimating forest biomass and volume using airborne laser data. Remote Sens. Environ. 24: 247-267.

Nelson, R., Valenti, M.A., Short, A. and Keller, C. 2003a. A multiple resource inventory of Delaware using airborne laser data. BioScience 53: 981-992.

Nelson, R., Parker, G. & Hom, M. 2003b. A portable airborne laser system for forest inventory. Photogramm. Eng. Remote Sensing 69: 267-273.

Nelson, R., Short, A., and Valenti, M. 2004. Measuring biomass and carbon in Delaware using an airborne profiling LIDAR. Scand. J. For. Res. 19: 500-511. [Erratum. 2005, 20: 283-284.]

Nelson, R.F., Næsset, E., Gobakken, T., Ståhl, G., and Gregoire, T.G. 2008. Regional forest inventory using an airborne profiling LiDAR. Jour. Forest Planning 13: 287-294.

Parker, R.C. and Evans, D.L. 2004. An application of LiDAR in a double-sampling forest inventory. West. J. Appl. For., 19: 95-101.

Royall, R. M. 1970. On finite population sampling theory under certain linear regression models. Biometrika 57: 377-387.

Royall, R. M. and Herson, J. 1973. Robust estimation in finite populations. J. Am. Statist. Assoc. 68: 880-889

Särndal, C.-E., Swensson, B. and Wretman, J. 1992. Model assisted survey sampling. Springer, New York. 694 pp.

Tomter, S.M. 1998. Beregning av volum de første år etter bestandsetablering. [Volume estimation during the first years after stand establishment]. Unpublished note. Norwegian Institute of Land Inventory. 2 p. (In Norwegian)

Wolter, K.M. 1984. An investigation of some estimators of variance for systematic sampling, Journal of the American Statistical Association 79: 781-790.

Table 1. Mean, standard error, standard error in percent of mean and percentage model error of total variance for total above ground dry biomass using profiling (PALS) and scanning (ALS) LiDAR. The corresponding ground-based Norwegian National Forest Inventory estimates are reported in the two rightmost columns.

	PALS					ALS					Ground plots		
Land cover	No of plots	Mean (t ha^{-1})	SE (t ha^{-1})	SE %[a]	% model error	No of plots	Mean (t ha^{-1})	SE (t ha^{-1})	SE %[a]	% model error	No of plots	Mean (t ha^{-1})	SE (t ha^{-1})
Productive forest:													
High						46	133.8	6.07	4.5	85.1	92	121.3	8.21
Medium						105	97.8	3.43	3.5	57.8	243	94.5	3.77
Low						138	47.4	2.19	4.6	64.0	306	46.6	2.29
Young						151	44.6	3.59	8.0	63.3	334	40.3	2.77
All prod. forest	554	57.6	2.21	3.8	56.5	440	67.7	2.16	3.2	41.8	975	64.0	1.73
Non prod. Forest/nonforest:													
Nonproductive forest	109	29.7	2.03	6.9	86.8	107	27.4	2.40	8.7	70.9	167	22.5	2.13
Mountain areas	78	6.37	3.06	48.1	98.1	85	6.0	0.69	11.5	29.6	182	6.5	0.85
Developed areas	22	10.9	3.00	27.6	97.7	30	5.8	0.89	15.2	72.6		--- no plots ---	
Water	0	1.63	4.78	294.0	99.4	0	3.2	0.30	9.5	14.6	77	0.6	0.58
Total	763	35.5	1.49	4.2	71.5	662	40.3	1.18	2.9	42.7	1401	39.4	0.99

[a] SE %: standard error in percent of mean above ground dry biomass.

Figure legends

Figure 1. The study area: Hedmark County, Norway.

Figure 2. Fifty three airborne laser scanning flight lines (spacing of 6 km) and 1401 National Forest Inventory ground plots (black dots). Profiling laser was flown at every 3 km, i.e., twice as many flight lines as with the scanning laser.

Fig.1.

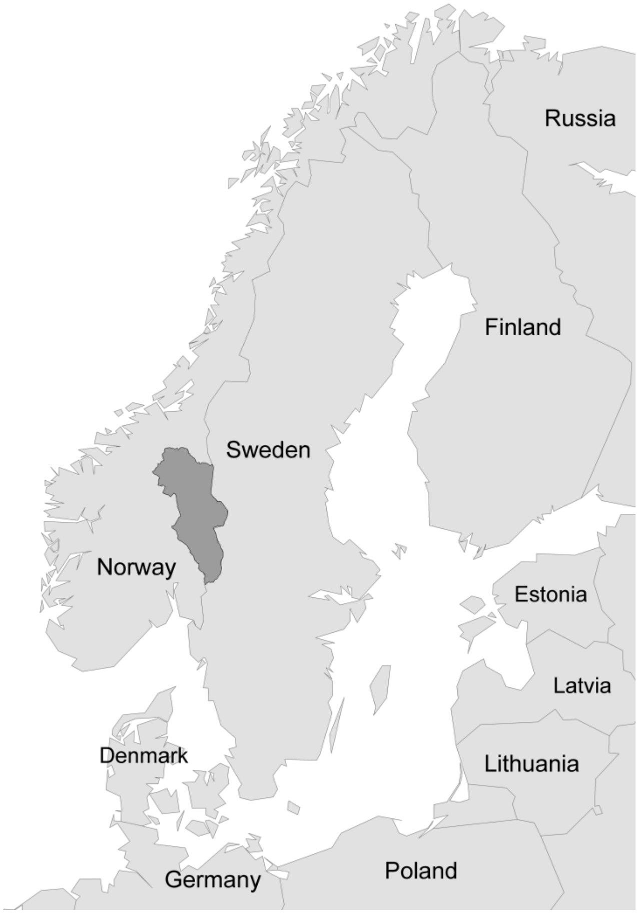
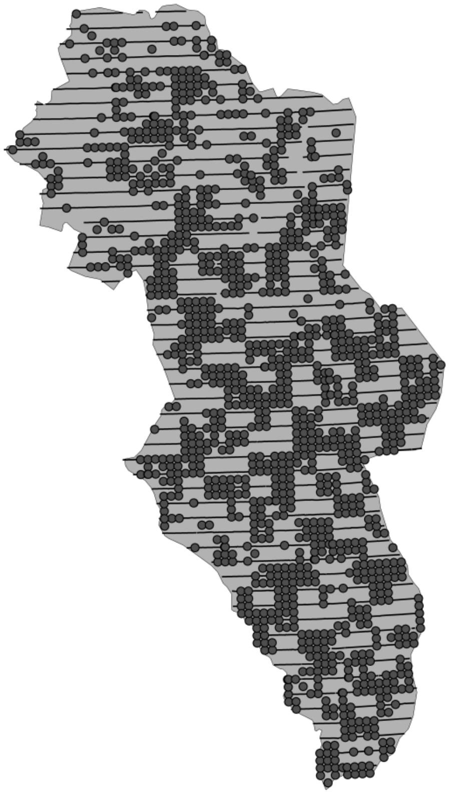

Russia
Finland
Sweden
Norway
Estonia
Latvia
Denmark
Lithuania
Germany
Poland

Fig 2.



Appendix 1

The starting point for the derivations is the decomposition

$$\hat{\mu}_Y - \mu_Y = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1},\hat{\alpha}) - \mu_Y = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1},\alpha) - \mu_Y + \frac{1}{m}\sum_{i=1}^{m}(g(x_{i1},\hat{\alpha}) - g(x_{i1},\alpha))$$

We seek to derive the variance of $\hat{\mu}_Y - \mu_Y$, namely $V(\hat{\mu}_Y - \mu_Y)$. The notations D_1 and D_2 are introduced by $D_1 = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1},\alpha) - \mu_Y$ and $D_2 = \frac{1}{m}\sum_{i=1}^{m}(g(x_{i1},\hat{\alpha}) - g(x_{i1},\alpha))$. The stochastic nature of D_1 is determined by the sample $S1$ and that of D_2 (the estimator $\hat{\alpha}$) by the sample $S2$. This is indicated by the indices $S1$ and $S2$ below. Now D_1 is the deviation between the average S1 sample value, if the true model *g* is known and its expected value and this deviation is independent of ε; thus $V(D_1) = \frac{1}{m}\cdot\sigma_g^2$ where σ_g^2 is the population variance of the g -values (population from which $S1$ is taken).

Regarding D_2 , the Taylor approximation leads to

$$D_2 = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{p}(\hat{\alpha}_j - \alpha_j)\cdot g'_j(x_{i1},\alpha) = \sum_{j=1}^{p}(\hat{\alpha}_j - \alpha_j)\left[\frac{1}{m}\sum_{i=1}^{m} g'_j(x_{i1},\alpha)\right]$$

The variance of D_2 is obtained by $V(D_2) = V_{S1}\left[E_{S2}(D_2 \mid S1)\right] + E_{S1}\left[V_{S2}(D_2 \mid S1)\right]$.
Conditioned on $S1$ we have

$$V(D_2 \mid S1) = \sum_{j}^{p}\sum_{k}^{p} Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot \bar{g}'_j \bar{g}'_k$$

where $\bar{g}'_j = \frac{1}{m}\sum_{i=1}^{m} g'_j(x_{i1},\alpha)$. This leads to the following expression for the unconditional variance

$$V(D_2)=\sum_{j}^{p}\sum_{k}^{p}Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot E_{S1}(\bar{g}'_j\bar{g}'_k)=$$

$$=\sum_{j}^{p}\sum_{k}^{p}Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot Cov_{S1}(\bar{g}'_j,\bar{g}'_k)+$$

$$+\sum_{j}^{p}\sum_{k}^{p}Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot E_{S1}(\bar{g}'_j)\cdot E_{S1}(\bar{g}'_k)$$

Note that $Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)$ depends only on S2 (not on S1).

Further, conditioned on S1, the bracketed term in D_2 is a constant. Hence, $E_{S2}(D_2\mid S1)=0$, or nearly so and thus also $E_{S2}(D_1\cdot D_2\mid S1)=0$. This implies that both $E_{S1}Cov_{S2}(D_1,D_2)=0$ $Cov_{S1}(E_{S2}(D_1),E_{S2}(D_2))=0$ and thus D_1 and D_2 are at least approximately uncorrelated. Therefore we can simply add the variances of D_1 and D_2 to get the variance of $\hat{\mu}_Y$

$$V(\hat{\mu}_Y)=\frac{1}{m}\cdot\sigma_g^2+\sum_{j}^{p}\sum_{k}^{p}Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot E_{S1}(\bar{g}'_j\bar{g}'_k)$$

The first term arises from the sampling error due to S1, while the second term arises from the uncertainty of the $\hat{\alpha}$-estimates (which is related to the selection of S2).

Variance estimation

A variance estimator will now be derived. To start with, we address the estimation of σ_g^2 from the S1-sample, through substituting $g(x_{i1},\alpha)$ by $g(x_{i1},\hat{\alpha})$ and applying the sample variance s^2 as an estimator. Conditional on S1 we obtain

$$E_{S2}\sum_{i=1}^{m} g^2(x_{i1}.\hat{\alpha}) = E_{S2}\sum_{i=1}^{m}\left[g(x_{i1},\alpha)+(\hat{\alpha}_1-\alpha_1)g'_1(x_{i1},\alpha)+\ldots+(\hat{\alpha}_p-\alpha_p)g'_p(x_{i1},\alpha)\right]^2 =$$

$$=\sum_{i=1}^{m} g^2(x_{i1},\alpha)+\sum_{j=1}^{p}\sum_{k=1}^{p} Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\sum_{i=1}^{m} g'_j(x_{i1},\alpha_j)g'_k(x_{i1},\alpha_k)$$

and

$$E_{S2}\left(\sum_{i=1}^{m} g(x_{i1},\hat{\alpha})\right)^2 = E_{S2}\left(\sum_{i=1}^{m}\left[g(x_{i1},\alpha)+(\hat{\alpha}_1-\alpha_1)g'_1(x_{i1},\alpha)+\ldots+(\hat{\alpha}_p-\alpha_p)g'_p(x_{i1},\alpha)\right]\right)^2 =$$

$$= E_{S2}\left(\sum_{i=1}^{m} g(x_{i1},\alpha)+(\hat{\alpha}_1-\alpha_1)\sum_{i=1}^{m} g'_1(x_{i1},\alpha)+\ldots+(\hat{\alpha}_p-\alpha_p)\sum_{i=1}^{m} g'_p(x_{i1},\alpha)\right)^2 =$$

$$=\left(\sum_{i=1}^{m} g(x_{i1},\alpha)\right)^2+\sum_{j=1}^{p}\sum_{k=1}^{p} Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot\sum_{i=1}^{m} g'_j(x_{i1},\alpha)\sum_{i=1}^{m} g'_k(x_{i1},\alpha)$$

Adding the terms we obtain $s_{\hat{g}}^2 = \dfrac{\sum_{i=1}^{m} g^2(x_{i1},\hat{\alpha})-(1/m)\cdot\left(\sum_{i=1}^{m} g(x_{i1},\hat{\alpha})\right)^2}{m-1}$ and (conditional on S1)

$$E_{S2}(s_{\hat{g}}^2) = s_g^2+\sum_{j=1}^{p}\sum_{k=1}^{p} Cov_{S2}(\hat{\alpha}_j,\hat{\alpha}_k)\cdot Cov(g'_j,g'_k)$$

where $Cov(g'_j,g'_k)$ denotes the sample covariance of the variables $g'_j(x_{i1},\alpha)$ and $g'_k(x_{i1},\alpha)$.

The unconditional expectation thus equals

$$E(s_{\hat{g}}^2) = E_{S1}E_{S2}(s_{\hat{g}}^2)) = \sigma_g^2 + \sum_{j=1}^{p}\sum_{k=1}^{p} Cov_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot Cov_{S1}(g'_j, g'_k)$$

Dividing this with *m*, and utilizing that $Cov_{S1}(\overline{g}'_j, \overline{g}'_k) = (1/m) \cdot Cov_{S1}(g'_j, g'_k)$ we obtain

$$V(\hat{\mu}_Y) = E(s_{\hat{g}}^2 / m) + \sum_{j=1}^{p}\sum_{k=1}^{p} Cov_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot E_{S1}(\overline{g}'_j) \cdot E_{S1}(\overline{g}'_k)$$

The term $Cov_{S2}(\hat{\alpha}_j, \hat{\alpha}_k)$ can be estimated from the sample S2, while the product $E_{S1}(\overline{g}'_j) \cdot E_{S1}(\overline{g}'_k)$ probably is fairly well estimated by $\hat{\overline{g}}'_j \cdot \hat{\overline{g}}'_k$, where $\hat{\overline{g}}'_j = \frac{1}{m}\sum_{i=1}^{m} g'_j(x_{i1}, \hat{\alpha})$. This implies a certain bias amounting to $Cov_{S1}(\overline{g}'_j, \overline{g}'_k)$ which would be in the order of $1/m$ (in relation to the estimate as such)

In conclusion, a 'fair' variance estimator can be expressed as

$$\hat{V}(\hat{\mu}_Y) = \frac{1}{m} s_{\hat{g}}^2 + \sum_{j=1}^{p}\sum_{k=1}^{p} C\hat{o}v_{S2}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot \hat{\overline{g}}'_j \cdot \hat{\overline{g}}'_k$$

In matrix notation the double sum can be written as $\nabla\hat{\overline{g}} \cdot C\hat{o}v(\hat{\alpha}) \cdot (\nabla\hat{\overline{g}})^T$ (with the gradient ∇g as a row vector and T as a notation for transposition).

Variances for the post-stratified estimator

The formulas for the variance and its estimator are in principle derived by the same method as above, but the details are more complicated since a ratio estimator is involved.

For the estimator (12)

$$\hat{\mu}_Y = \sum_{h=1}^{H} W_h \hat{\mu}_{Yh}$$

where

$$\hat{\mu}_{Yh} = \frac{\dfrac{1}{m}\sum_{i=1}^{m} G_{ih}(\hat{\alpha}_h)}{\dfrac{1}{m}\sum_{i=1}^{m} A_{ih}}$$

we have the 'generic' formula

$$V(\hat{\mu}_Y) = \sum_{h=1}^{H}\sum_{k=1}^{H} W_k W_h Cov(\hat{\mu}_{Yh}, \hat{\mu}_{Yk})$$

Further, applying the customary expression for ratio estimators,

$$Cov(\hat{\mu}_{Yh}, \hat{\mu}_{Yk}) \approx Cov\left[\frac{\dfrac{1}{m}\sum_{i=1}^{m}(G_{ih}(\hat{\alpha}_h) - R_h A_{ih})}{\dfrac{1}{m}\sum_{i=1}^{m} A_{ih}}, \frac{\dfrac{1}{m}\sum_{i=1}^{m}(G_{ik}(\hat{\alpha}_k) - R_k A_{ik})}{\dfrac{1}{m}\sum_{i=1}^{m} A_{ik}}\right] \approx$$

$$\approx \frac{Cov(\overline{G}_h(\hat{\alpha}) - R_h \overline{A}_h, \overline{G}_k(\hat{\alpha}) - R_k \overline{A}_k)}{E(\overline{A}_h)\cdot E(\overline{A}_k)}$$

Now, by a Taylor expansion

$$\overline{G}_h(\hat{\alpha}_h) - R_h \overline{A}_h \approx \overline{G}_h(\alpha_h) - R_h \overline{A}_h + \sum_{j=1}^{p_h}(\hat{\alpha}_{jh} - \alpha_{jh})\overline{G}'_{jh}(\alpha_h)$$

By inserting this into the numerator covariance above and expanding we obtain

$$Cov(\overline{G}_h(\hat{\alpha}_h) - R_h\overline{A}_h, \overline{G}_k(\hat{\alpha}_k - R_k\overline{A}_k) \approx Cov(\overline{G}_h(\alpha_h) - R_h\overline{A}_h, \overline{G}_k(\alpha_k) - R_k\overline{A}_k) +$$

$$\text{Cov}\left[\sum_{j=1}^{p_h}(\hat{\alpha}_{jh} - \alpha_{jh})\overline{G}'_{jh}(\alpha_h), \sum_{j=1}^{p_k}(\hat{\alpha}_{jk} - \alpha_{jk})\overline{G}'_{jk}(\alpha_k)\right]$$

since two of the four terms are (at least almost) zero, due to the (near) unbiasedness of $\hat{\alpha}$.

The first term equals the covariance estimator with true data and its expectation with respect to S1 is thus equal to

$$\frac{1}{m}Cov(G_h(\alpha_h) - R_hA_h, G_k(\alpha_k) - R_kA_k)$$

The second term equals, by expansion, and taking expectation with respect to S1,

$$\sum_{j_1=1}^{p_1}\sum_{j_2=1}^{p_2}Cov_{S2}(\hat{\alpha}_{j_1h}, \hat{\alpha}_{j_2k})E_{S1}(\overline{G}'_{j_1h}(\alpha_h)\overline{G}'_{j_2k}(\alpha_k))$$

By inserting these two expressions in the ‘generic’ formula we obtain the approximate formula (13) for the variance.

The approximate variance estimator is obtained by inserting sample estimates (and observed values) in the variance formula, in analogy with formula (7).