Forestry *An International Journal of Forest Research*

Institute of Chartered Foresters

Forestry 2017; **90**, 112–124, doi:10.1093/forestry/cpw041
Advance Access publication 19 August 2016

Methods for variable selection in LiDAR-assisted forest inventories

Paolo Moser[1]*, Alexander C. Vibrans[1], Ronald E. McRoberts[2], Erik Næsset[3], Terje Gobakken[3], Gherardo Chirici[4], Matteo Mura[4] and Marco Marchetti[5]

[1]*Universidade Regional de Blumenau, Forest Engineering, 3250 São Paulo Street, Blumenau 89030-000, Santa Catarina, Brazil*
[2]*Northern Research Station, Forest Inventory & Analysis, U.S. Forest Service, 1992 Folwell Ave, Saint Paul, MN 55108, USA*
[3]*Norwegian University of Life Sciences, Ecology and Natural Resource Management, 12 Høgskoleveien, Ås 1430, Norway*
[4]*University of Florence, Gestione dei Sistemi Agrari, Alimentari e Forestali, 13 Via San Bonaventura, Florence 50145, Italy*
[5]*University of Molise, Bioscienze e Territorio, Contrada Fonte Lappone, Molise 86090, Italy*

*Corresponding author. E-mail: paolo.moser@gmail.com

Received 27 April 2016

Estimation of wood volume and biomass is an important assignment of any National Forest Inventory. However, the estimation process is often expensive, laborious and sometimes imprecise because of small sample sizes relative to population variability. Remote sensing techniques are an option to assist in surveying large areas by providing data that can be related to the forest attribute of interest through mathematical models of relationships. Light Detection and Ranging (LiDAR) is a technology that can provide data that are closely related to forest wood volume and biomass. With these data, linear regression is often used to estimate forest attributes. If the relationship provides evidence of nonlinearity, a transformation in the variables can be considered. However, modern computation allows fitting nonlinear regression models without transformations of the variables. Nonlinear least squares (NLS) techniques also give more freedom to assure satisfaction of natural conditions such as non-negativity and/or lower and upper asymptotes. Like any estimation technique, NLS is subject to overfitting when using a large number of predictor variables. Because NLS is more computationally intensive than linear regression, stepwise selection techniques may require considerable programming effort. We compared three methods to select predictor variables for nonlinear models of relationships between forest attributes and LiDAR metrics, two of them based on genetic algorithms (GAs) and one based on random forest (RM). GAs were implemented to optimize a cost function that yields root mean square error or the Akaike Information Criterion (AIC), while RM was based on variable importance in decision trees. A model with the predictor variable most correlated with the response variable was also considered. We compared the results of overall estimation for two datasets using the model-assisted, generalized regression estimator and concluded that the combination of GAs and AIC was the most efficient and stable procedure for selection of variables. We attribute this result to the penalty that AIC applies to models with large numbers of variables, which leads to a more efficient model with a minimum loss of information.

Introduction

Wood volume (or biomass) estimates at local, regional and global levels are fundamental for estimation of carbon stock and for evaluating an ecosystem's response to climatic changes and anthropic influences (Hese *et al.*, 2005, Ni-Meister *et al.*, 2010). Among the many reasons for estimating forest carbon stock, two are motivated by climate change considerations: (1) agreements with the United Nations Framework Convention on Climate Change (UNFCCC) and (2) the carbon credit market (Brown 2002).

For large area estimation, estimates are usually calculated by aggregating the values of volume/biomass for individual trees at the plot level. Plot data are then added or averaged to produce large area estimates (McRoberts and Westfall 2014). Individual tree predictions of wood volume and biomass are commonly based on allometric models that use diameter at breast height (DBH), total height (h) and, sometimes, wood density (d) as predictor variables. If these models are species specific tree species is typically one more predictor variable. For some sampling strategies, also an additional diameter at a certain height percentile is included (Cormier *et al.*, 1992, Chave *et al.*, 2005, Vallet *et al.*, 2006, Ni-Meister *et al.*, 2010, Zianis and Seura 2005). For inventory purposes, the final product is an inference in the form of a confidence interval (CI) for a population parameter such as mean volume or biomass per unit area. The CI can be constructed by adding/subtracting an amount to the sample mean ($\bar{X}$). This amount is based on standard error (SE) –the square root of the ratio between variance and sample size. Assuming a confidence level of 95%, the CI assumes the form of $\bar{X} \pm t_{(\alpha/2;n-1)} \times$ SE, where $t_{(\alpha/2;n-1)}$ is the two tailed percentile of t-distribution, for a given significance level (α) and degrees of freedom ($n-1$). In particular, models for which prediction accuracy is a measure and maps for which accuracy assessments are measures are only intermediate products enroute to this inference.

Remote sensing techniques can be used to obtain estimates for areas that are not sampled. Among these techniques, data based on microwave (synthetic aperture radar — SAR) and optical sensors (i.e. multi- and hyper-spectral) can enhance large scale inferences (McRoberts *et al*., 2010). However, Light Detection and Ranging (LiDAR) data –as such photogrammetric 3D data based on stereo imagery –are considered more informative, because they provide metrics that can be used to predict the vertical structure of the forest/area of interest (Chen *et al*., 2006, Zhao *et al*., 2011, Yao *et al*., 2011, Zhao *et al*., 2012). In practical applications of LiDAR data, georeferenced sample units can be used, in a first stage, to develop empirical models of relationships between field measurements and the derived LiDAR metrics (Næsset 2002). In a second stage, these models are applied for the entire area of interest, predicting the forest attributes based only on LiDAR metrics.

The use of LiDAR for estimation of forest attributes is under development around the world. (Nelson 2013) provided a list of important studies concerning this issue. Of interest, (Nelson *et al*. 1988) constructed linear models to estimate height, volume and biomass and (Næsset 2002) proposed a two stage procedure to predict forest stand characteristics using airborne laser scanning (ALS) and field inventory data. The results of these studies provided evidence that regression models using ALS data can be used to improve estimation of parameters related to forest variables such as volume and biomass.

Two issues are of concern when estimating volume using linear regression. The first is negative or extremely large predictions with no effective biological meaning. As an alternative to linear models, (McRoberts *et al*. 2013a) suggested using a nonlinear asymptotic logistic model, which is constrained by the lower horizontal asymptote of $y = 0$ and by an upper asymptote that can be estimated using the sample data. The second issue is the selection of predictor variables, which may cause overfitting of the model and increases the probability of poor predictions. This phenomenon often occurs with large numbers of highly correlated predictor variables that can cause the model to reflect the noise and peculiarities rather than the general trend in the data and to adversely affect the quality of the predictions when applying the fit model to a new dataset (Kohavi and Sommerfield 1995, Santos *et al*., 2009). For multiple linear regression models, stepwise variable selection procedures are commonly used (Næsset 2002, 2011, Ene *et al*., 2012, Lu *et al*., 2012, Zhao *et al*., 2012, He *et al*., 2013). For nonlinear regression models, the approach is more laborious and difficult to implement. Another issue with stepwise algorithms is related to its inefficiency when predictor variables are strongly correlated, which is the case of ALS height and density metrics (Harrell 2013). McRoberts *et al*. (2013a, 2013b) provided an iterative approach based on the pseudo-R^2 value as an alternative to stepwise procedures.

The advent of machine learning techniques introduced a new paradigm into the data mining process, and these techniques can be used to select subsets of predictor variables that optimize criteria such as the Akaike Information Criteria (AIC) or root mean square error (RMSE). One of these techniques is genetic algorithms (GAs) (Goldberg and Holland 1988, Holland 1992). With this approach, a population of k subsets, each of which includes a random combination of predictor variables, is constructed; the model is then fit for each subset, and an optimization criterion such as RMSE or AIC is calculated. An iterative process based on biological genetics and evolution is then used to construct a new population of subsets. According to (Broadhurst *et al*. 1997), after i iterations, the subset that returns the most optimal value for the criterion is selected as the most suitable subset of predictor variables.

Another technique that can be used is random forest (RF) (Breiman 2001). This algorithm constructs an ensemble of prediction trees and through bootstrapping of the training data, one prediction is assigned to each bootstraped sample. The final prediction is computed as the mean over the predictions of the single trees (Strobl *et al*., 2008). Different than the GA, this technique does not return the optimal subset of predictor variables, but rather returns each variable's 'importance' in the prediction procedure, which can be used to select predictor variables for the regression model.

Thus, the objective of the study was to compare estimates of means and variances as the defining components of CIs for mean wood volume and biomass per unit area for the variable selection techniques for two ALS datasets, one from Norway and one from Italy. Models were constructed by fitting nonlinear regression models with subsets of five predictor variables obtained using three approaches: (1) GA with RMSE as the criterion, (2) GA with AIC as the criterion, (3) RF to select variables with the greatest importance. Also, to provide comparison, (4) models were constructed using all predictor variables and (5) only the predictor variable most highly correlated to the response variable.

Given the multidimensionality of the predictor variable space used for ALS-assisted estimation of volume, this study focused on a consistency analysis of the selected subsets of predictor variables and their impact on large scale inferences in the form of CIs for population parameters such as mean volume or biomass per unit area. We were motivated to conduct this study because no comparison among methods for variable selection for LiDAR-assisted forest inventories regarding CIs is known to have been reported, especially when using methods that were developed in a different field (i.e. machine learning) for which statistical inference is not the primary interest. In addition, we proposed an original iterative procedure that addresses the stability of the methods, providing evidence of reliability under different datasets. Finally, we assessed issues of technical efficiency of the methods, keeping in mind implementation aspects like convergence and randomness. To accomplish this task, we used an exhaustive 'brute-force' search for the optimal subset of variables. This subset was used as basis to evaluate the quality of our proposed selection algorithm. The novel features of the study are the comparison of the approaches to select variables with respect to the CI that they produce rather than an intermediate product such as a measure of prediction accuracy, the assessment of stability and efficiency by means of an iterative approach and the investigation of RF to select variables for regression with variables of the same type varying in many levels.

Data

Norwegian data

Data were acquired in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway as part of an operational forest inventory (Figure 1) (McRoberts *et al*., 2013a).

A PA31 Piper Navajo aircraft carried the Optech ALTM 3100 laser scanning system (Optech, Canada) used in the study. The ALS data were acquired between 15 July 2006 and 12 September 2006 from a height above ground of ~1700 m with average aircraft speed of 75 m/s. The pulse repetition frequency was 50 kHz, the scan frequency was 31 Hz, the maximum scan angle was 16, which corresponded to an average swath width of ~975 m, the mean footprint diameter was ~50 cm, and the average point density was 0.7 pulses/m^2. Only echoes with heights greater than 2 m were considered. To match the 250 m^2 size of the field plots, the study area was tessellated into square 250 m^2 cells that served as population units. For each plot and population unit, heights corresponding to the 10th, 20th, …, and 100th percentiles of the distributions were calculated and denoted h_1, h_2, …, and h_{10}, respectively. Densities were calculated as the proportions of echoes with heights greater than 0%, 10%, …, and 90% of the range between 2 m above ground and the 95th height percentile and were denoted d_0, d_1, …, and d_9, respectively.

Norwegian National Forest Inventory circular field plots of 250 m^2 located at the intersections of a 3 km × 3 km grid were used to acquire field measurement data (Tomter *et al.*, 2010). On each plot, all trees with DBH (1.3 m) of at least 5 cm were callipered. An average of 10 sample trees per plot was selected with probability proportional to stem basal area, to provide measured heights. Heights for the remaining trees were predicted using height-DBH models (Vestjordet 1967, Fitje and Vestjordet 1977).

The volume of each sample tree was estimated using species-specific volume models with DBH and either measured height or predicted height as predictor variables (Braastad 1966, Brantseg 1967, Vestjordet 1967). The effects of uncertainty in these model predictions have been demonstrated to be negligible for studies in Sweden and Finland (Ståhl *et al.*, 2014), in Austria (Berger *et al.*, 2014), in Norway (Breidenbach *et al.*, 2014), in Brazil (McRoberts *et al.*, 2015) and in Oregon (Shettles *et al.*, 2015).

The total plot volume (VOL) was estimated as the sum of volume estimates for individual trees. A variogram analysis indicated no meaningful spatial correlation among plot VOL observations. To minimize the effects of forest change between the plot observation dates and the 2006 date of the ALS acquisition, only the 145 plots measured between 2005 and 2007 were used for this study. The study area includes 1259 km^2 and features altitudinal variations ranging from 204 to 1134 m above sea level (asl) with a mean of 570 m asl. The dominant tree species are Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.) (McRoberts *et al.*, 2013b).

Italian data

The Italian study area is located in the south-west part of Molise Region, Italy (Figure 2).

ALS data were acquired for scientific purposes related to ITALID project — Use of LiDAR data to study Italian forests — in June 2010 (Scrinzi *et al.*, 2013). The survey, carried out by Partenavia P68 aircraft equipped with a Optech Gemini sensor, covered 36 380 ha. The maximum scan angle was of 15°, with a frequency of 70 kHz. The average point density is of 3.5 points/m^2 and it varies from 9 points/m^2 in forested and multilayered

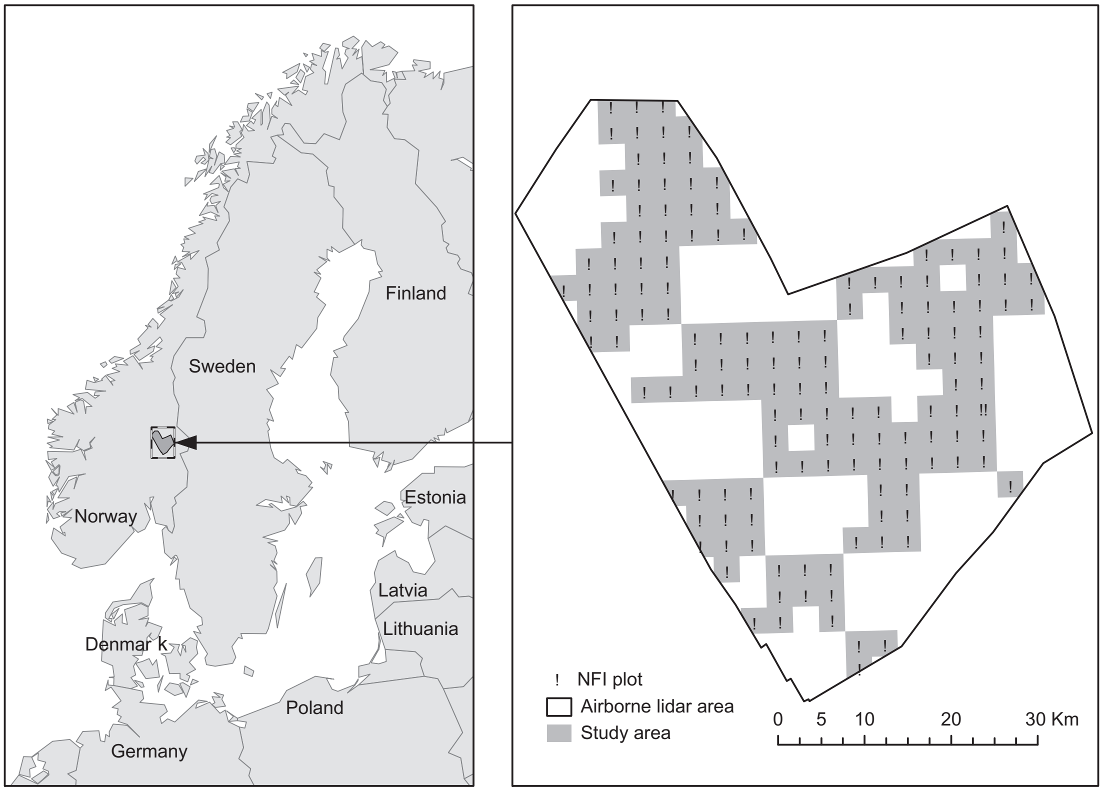

Figure 1 Norwegian study area and sampling design.

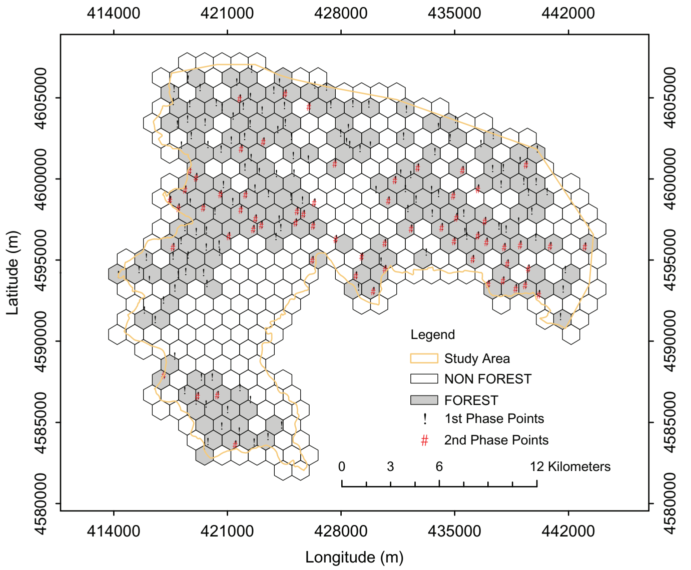

Figure 2 Italian study area and sampling design.

areas, to less than 1 point/m^2 in the bare ground or flat terrain. Twenty-two ALS height and density metrics were calculated. Canopy density metrics were the proportions of all returns above 1.3 m, the proportion and the count of returns between 1.3 and 10 m. Canopy cover was the proportion of first returns above 1.3 m to all first returns. Canopy height metrics were the same as those described for the Norwegian data. Height summary statistics such as minimum, maximum, average, standard deviation, coefficient of variability, kurtosis, skewness and canopy relief ratio (Parker and Russ 2004) were also calculated. All ALS metrics were calculated for 23 m × 23 m cells of approximately 531 m^2 and that served as population units. As for Norway, this configuration was adopted to match the size of the field plots.

The field data were acquired from 62 field plots spatially distributed using two-phase unaligned systematic sampling (Chirici *et al.*, 2016). The study area was tessellated into 437 hexagons, each with area of 1 km^2 and a point was randomly selected in each hexagon serving as centre point for the plot (Figure 2). Each plot consists of two concentric plots of 4 and 13 m radius. The DBH of all the trees with DBH $\geq$ 2.5 cm within the 4 m radius area, and trees with DBH $\geq$ 9.5 cm in the 13 m radius area were collected. The height of at most 10 trees was measured in a subsample of trees in the plot. The sample trees were selected according to the three DBH-largest trees, the five trees nearest to the plot centre and the two trees whose species or DBHs were less frequently observed. The heights of the remaining trees were estimated for the main species using the DBH-Height models fit with the gathered data. For each tree, VOL was estimated by national double-entry tables (Castellani *et al.*, 1984). As for the Norwegian data, the uncertainty in these predictions can be considered negligible. The value of the biomass was estimated by:

$$B = \mathrm{VOL} \times \mathrm{BEF} \times \mathrm{WBD} \tag{1}$$

where B is the biomass (t/ha), BEF is the biomass expansion factor and WBD is the wood basic density (t/m^3). The values of BEF and WBD were extracted from (Federici *et al.* 2008).

Forests were found in 205.18 km^2, covering about the 64% of the area. The forested area is dominated by Turkey oak (*Quercus cerris*) at 61.17 km^2 (29.81% of the forested area), Downy oak (*Quercus pubescens*) for 58.86 km^2 (28.69%), Hop Hornbeam (*Ostrya carpinifolia*) for 36.32 km^2 (17.70%), Beech (*Fagus sylvatica*) for 18.54 km^2 (9.04%) and Holm oak (*Quercus ilex*) 14.12 km^2 (6.88%). Hygrophilus forests, plantations, pioneer deciduous vegetation, shrublands, synanthropic forests and Chestnut forests complete the forest landscape.

Methods

The analyses are based on three statistical assumptions: (1) there is a finite population of N units (cells) in the form of squares of size 250 m^2 for Norway and 531 m^2 for Italy; (2) there is an equal probability sample of n cells and (3) ALS metrics are available for all plots and cells.

Because the final product of a forest inventory is an inference in the form of a CI for a population parameter, comparisons among the proposed variable selection techniques (described below) were made with respect to this CI ($\bar{X} \pm t_{(\alpha/2;n-1)} \times \mathrm{SE}$), rather than an intermediate product such as a measure of prediction accuracy. Nevertheless, prediction accuracy is used to select variables and is therefore reported for informational purposes.

Variable selection methods

Genetic algorithm

GAs are stochastic optimization techniques that are conceptually based on biological genetics and evolution (Goldberg and Holland 1988, Holland 1992). This approach searches for the subset of predictor variables that optimizes a cost function defined by the user. Although there is no guarantee of finding the optimal predictor variable subset (Garey and Johnson 1979), locally optimal solutions can be found in a feasible computational time.

Let X be the ordinated set of all the predictor variables and $f(x_i, \beta)$ be a function that returns a value to be optimized (cost function). The basic building steps for all GAs can be summarized as follows (Broadhurst *et al.*, 1997):

(1) A population of k subsets is constructed, each one containing a random combination of predictor variables (genes).
(2) Each subset is considered a binary string [0,1] with 1's meaning that the variable x_i is 'selected' in the subset and 0's meaning 'not selected'. This is called a 'chromosome'.
(3) A weighted random selection is applied to the population of chromosomes, selecting two of them (parents). The weights are proportional to the cost function response, meaning that chromosomes that yield near optimal (or locally optimal) responses have greater probabilities to be selected.
(4) The parent chromosomes are partitioned and recombined, creating a 'child' that carries a mix of parents' characteristics.
(5) A probability function is assigned to each child, adding the possibility of mutation (changes between 0's and 1's).
(6) Steps 3–5 are repeated j times and the cost function is evaluated for each new chromosome.
(7) The whole process is replicated until a criterion is satisfied or until a pre-defined number of iterations is completed.

Selection of predictor variables using GA was reported as successful by other studies related to different areas like medicine, economy and chemistry (Broadhurst *et al.*, 1997, Vinterbo and Ohno-Machado 1999, D'heygere *et al.*, 2002, Paterlini and Minerva 2010, Cateni *et al.*, 2011). Regarding forestry, the studies of (Haapanen and Tuominen 2008), (Latifi *et al.* 2010), (Tuominen *et al.* 2013) and (Garcia-Gutierrez *et al.* 2014) corroborate these results, especially when using remotely sensed data. However, we did not find any study that uses the iterative approach proposed here, so no information about stability and efficiency among a large number of datasets were reported in the aforementioned papers.

In this study, the GA approach was used to optimize two cost functions, one based on minimization of AIC and denoted GA_{AIC} and the other based on minimization of RMSE and denoted GA_{RMSE}.

Random forests

RF is an algorithm that belongs to the family of ensemble methods used for both classification and regression problems, based on model aggregation ideas (Genuer *et al.*, 2010). The basic mechanics of RF consist of combining many binary decision trees constructed using several bootstrapped samples where the final prediction/classification is the average over this 'forest' (Breiman 2001). Because this study is focused on applied regression to predict volume values, only RF for prediction is considered.

Let X be the set of all the predictor variables and $f(X, \beta)$ a linear multiple regression model that relates the response variable, VOL/biomass, with the predictors variables (ALS Metrics). In this framework, the basic routine for a RF algorithm follows (Hastie *et al.*, 2009):

(1) A population of t bootstrapped samples (or subsamples) is created from the training dataset.
(2) For each sample obtained in (1), a decision tree $T_i (1 \leq i \leq t)$ is created.
(3) Each T_i is grown by recursively repeating these steps in each node of the tree:
 (a) Random selection of m predictor variables, among p available.
 (b) Pick the most suitable variables/subset of predictor variables.
 (c) Split the node into two daughter nodes.
(4) Step (3) is repeated until a minimum node size defined by the user is reached.
(5) Output the ensemble of trees $\{T_i\}_1^t$.
(6) Predictions are the average of the predictions all over the ensemble, which means:

$$\hat{y}_{RF} = \frac{1}{t} \sum_{i=1}^{t} T_i(X) \qquad (2)$$

In the step (3.b) of RF algorithm, variables are selected using a random permutation process. Basically, when the ith tree is grown, a sample (that was not used when growing the tree) is passed down the tree, and the prediction accuracy is recorded. Then the values for the mth variable are randomly permuted in this sample, and the accuracy is again computed. Decrease in accuracy provides evidence that the mth variable is significant. In fact, according (Hastie *et al.*, 2009), the randomization effectively avoids the effect of a variable, much like setting a coefficient to zero in a linear model.

Besides the prediction process, RF returns predictor variable importance which, as the GA, give insights into which predictor variables should be included in the regression model. The 'importance' of a given variable is computed based on the increase in mean square error for a tree in the forest when the observed values of this variable are randomly permuted in the samples (Genuer *et al.*, 2010). In this context, an advantage of this approach is that it covers the impact of each predictor variable in two ways: individually and interacting with other variables. However, in a multidimensional space where the variables are highly correlated, the true variable importance can only be assessed by a conditional approach (Strobl *et al.*, 2008). This procedure avoids the permutation test to give more weight to highly correlated predictor variables that are not obviously independent. Although computationally intensive, this was the procedure adopted in this study, because a principal component analysis showed that the height percentiles are correlated with each other and the density percentiles are correlated with each other. However, height percentiles are not correlated with the density percentiles, which splits the set of predictor variables in two well-defined subsets.

Studies using RF to select predictor variables for regression are not so common. Most of them are related to the use of RF for classification in remote sensing approaches. Investigations regarding regression issues were conducted by (Genuer *et al.* 2010) and (Hapfelmeier and Ulm 2013). These studies successfully used RF to select predictor variables, although no analysis was conducted to assess the efficiency of the method in comparison with other methodology.

Inference methods

The simple random sampling estimator

The most widely used estimator of a population parameter is the simple random sampling (SRS) estimator (Hansen *et al.*, 1983). This statistic is characterized as probability-based (or design-based) because it is derived from the probabilities of selection of population units into the sample. Probability-based estimators rely on three assumptions: (1) the sample is constructed using a probability-based randomization scheme; (2) each population unit has a positive and known probability of being selected and (3) the observation of the response variable for each population unit is a fixed value (McRoberts *et al.*, 2013b). This estimator can be calculated with or without replacement. For our study, we used the SRS estimator as a special case of the Horvitz-Thompson (HT) estimator with equal probability sampling with replacement.

The SRS estimators for means and their variances are

$$\hat{\mu}_{SRS} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (3)$$

and

$$\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{SRS}}) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_{\mathrm{SRS}}\right)^2}{n(n-1)} \tag{4}$$

where n is the sample size and y_i is the observation of VOL/biomass in the ith sample unit.

The SRS estimator has some advantages:

(1) It is easy to calculate, because it only uses the sampled data, with no need to fit a model or some other statistical procedure.
(2) It is intuitive, in the sense that it only uses a common arithmetic mean, and its variance is well stabilized by the Central Limit Theorem.
(3) It is unbiased under any probability sampling design, which means that $E[\mu_{\mathrm{SRS}}] = \mu$ and $E[\mathrm{VAR}(\mu_{\mathrm{SRS}})] = \mathrm{VAR}(\mu)$.

The disadvantage of the SRS estimators is that the variances can be large, mainly when the sample size is small and/or the population is highly variable (McRoberts *et al*., 2013a). However, because it is unbiased, the SRS estimator was used in this study for comparison with the model-assisted estimators used with the different subsets of predictor variables.

When applying the estimator for SRS variance under a systematic design, the variance of the SRS mean estimator is usually over-estimated (Särdnal *et al*., 1992), but the estimator of the mean is still unbiased — this is the primary interest here.

The generalized regression estimator

The generalized regression estimator (GREG) used in this study to estimate the population mean using all the data available for this population is also classified as probability-based, because it relies on the same assumptions as the SRS estimators.

This particular estimator is often called ‘model-assisted’ because it uses a model based on auxiliary data to improve estimation. (Särdnal *et al*. 1992) provided the following model-assisted estimators for the mean and the variance of a population parameter:

$$\hat{\mu}_{\mathrm{GREG}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{5}$$

and

$$\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}) \tag{6}$$

where N is the population size, $\hat{y}_i$ is the prediction of VOL/biomass for each population unit, using equation (1), n is the sample size, y_i is the observation of volume in the ith sample unit and $\epsilon_i = \hat{y}_i - y_i$. The term $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$ in equation (4) is the aforementioned correction for estimated bias.

Despite its label, GREG can be used with any modelling approach that produces reliable predictions, including non-parametric techniques (Zheng and Little 2003, Lehtonen *et al*., 2005, Breidt and Opsomer 2009).

Nonlinear logistic regression model

The nonlinear logistic regression model used to describe the relationship between plot-level ground VOL/biomass values and the ALS metrics has the mathematical structure (McRoberts *et al*., 2013a):

$$y_i = f(x_i, \beta) = \frac{\alpha}{1 + e^{\left(\beta_0 + \sum_{j=1}^{J}\beta_j . x_{ij}\right)}} + \epsilon_i \tag{7}$$

where i indexes plots or population units, y_i is the observed VOL/biomass, αs and βs are parameters to be estimated, x_{ij} is the jth ALS metric and ϵ_i is the residual error. This asymptotic model was selected for this study because of two primary advantages: (1) it does not produce negative values, because the predictions are limited by the asymptote $y = 0$; (2) large values are constrained by the parameter α that can be initially estimated as the maximum value in the sample data (McRoberts *et al*., 2013b). These two assumptions are valuable for purposes of retaining the biological relevance of the predictions.

To assess the quality of fitness of this model, the classical R^2 is not entirely appropriate, because the assumptions underlying R^2 are not fully satisfied when using nonlinear models (Anderson-Sprecher 1994). Keeping that in mind, an efficiency measure R^{2*} (often called pseudo-R^2) was used, calculated as (Vanclay and Skovsgaard 1997), described by

$$R^{2*} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{8}$$

where $R^{2*} = 1$ for a ‘perfect’ fit; $R^{2*} = 0$ indicates that the model is no better than a simple average and $R^{2*} < 0$ reveal a poor quality of fit.

Analysis

For both datasets, we performed the following analyses.

As an initial procedure, the SRS estimates were calculated, for comparison with the GREG estimates and to assess the value of the auxiliary data for reducing the variance of the population estimates.

To start the process of variable selection, a principal component analysis was conducted to investigate the degree of multivariate correlation among the predictor variables.

To proceed with the variable selection, the analysis was conducted according to the following steps:

(1) Using SRS with no replacement, the plot-level dataset was split into two disjoint subsets labelled as calibration dataset and validation dataset, for further reference. These subsets contained, respectively, 70% and 30% of the total of sampled plots (Zhang 2005).
(2) Using the calibration dataset, the methods for variable selection aforementioned were applied.
(3) Steps (1) and (2) were repeated 100 times, always changing the seed that starts the random processes involved. The number of iterations was selected following a critical analysis of the stability of the results with no effective changes found beyond 100 replications.

(4) The five predictor variables with largest relative frequency for $\mathrm{GA}_{\mathrm{AIC}}$ and $\mathrm{GA}_{\mathrm{RMSE}}$ and with the largest mean importance value for RF were chosen as the 'most suitable subset' for each method.
(5) The predictor variable most highly correlated with the response variable was chosen to compose the 'single variable subset'.
(6) All the variables were chosen to compose the 'all variables subset'.

Using the outputs of the variable selection procedures, the parameters of the nonlinear logistic regression model provided by equation (6) were estimated using the nonlinear least-squares (NLS) algorithm and the calibration dataset.

The validation dataset was used to assess the accuracy of the model predictions, through the computation of R^{2*} and RMSE. The underlying regression assumptions were verified through visual inspection of $Q-Q$ and residuals versus predictions graphs. Lack of fit was assessed visually through a graph of VOL/biomass observations versus VOL/biomass predictions. Because lack of fit in the model can lead to false positive results in the analysis of the performance of the selected predictor variables, a statistical test of hypothesis was conducted. A simple linear model was fit to VOL/biomass observations as the response variable and VOL/biomass predictions as the predictor variable. In the absence of lack of fit, the points in the graph of this model should lie along the 1:1 line. An F-test for comparing estimates of the intercepts and slopes jointly to (0,1) was conducted, providing approximate results because no account is made for the uncertainty in the predictions that serve as the predictor variable (Vanclay and Skovsgaard 1997, McRoberts *et al*., 2013a).

Once the model presented no evidence of lack of fit, the GREG estimates were calculated using the models with the different sets of predictor variables. The results were compared with the SRS estimates with emphasis on the absolute values of the estimated means ($\hat{\mu}_{\mathrm{GREG}}$) and their respective standard errors ($\mathrm{SE}(\hat{\mu}_{\mathrm{GREG}}) = \sqrt{\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}})}$). Also, the standard deviations of these statistics over the 100 replications can be used as a measure of the stability of model predictions. In addition, to assess the efficiency of the GREG estimator when compared with SRS estimator, the relative efficiency coefficient (RE) was calculated as,

$$\mathrm{RE} = \frac{\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{SRS}})}{\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}})} \qquad (9)$$

Because RE is the ratio between the variances of $\hat{\mu}_{\mathrm{SRS}}$ and $\hat{\mu}_{\mathrm{GREG}}$, values greater than 1 are evidence of greater precision in the estimates.

Finally, because $\mathrm{GA}_{\mathrm{AIC}}$ presented the most efficient and accurate results, the behaviour of this algorithm among iterations was explored. A 'brute-force' procedure was implemented to find the most suitable variable subset among the 524 287 possible combinations for the Italian dataset. Having the fitting statistics for this subset of variables, the GA procedure was evaluated with respect to number of iterations, stability and convergence to near-optimal results.

Results

Model accuracy

The results of variable selection are shown in Table 1. Because the selection procedure was repeated 100 times, the mean values of R^{2*} and RMSE are reported with their respective standard deviations. These values were obtained by applying the fit models to the 30% of the data that were not used to estimate the parameters of the models, in each iteration. This iterative procedure was used to evaluate the robustness and stability of the variable selection algorithms. Robustness was assessed by verifying the presence or absence of outliers in the R^{2*} and RMSE distributions, while stability was assessed using the standard deviation of these statistics.

According to Table 1, regarding the Norwegian dataset, the mean R^{2*} s were larger for the subsets of variables (min = 0.64; max = 0.74) than for all variables (0.56). Among the selection methods, RF produced the smallest R^{2*}, providing evidence that this technique is not the most reliable approach for selecting variables. In the other hand, the largest R^{2*} was produced by $\mathrm{GA}_{\mathrm{AIC}}$ (0.74). This value is ~15% largest than the value produced by RF. Also, while the mean RMSE for all variables was the largest (56.6) and for RF, it was the second largest (52.35), mean RMSE for $\mathrm{GA}_{\mathrm{AIC}}$ was the smallest (43.68). This results indicates that the overall estimates for the nonlinear model fit with the five variables selected by $\mathrm{GA}_{\mathrm{AIC}}$ are more accurate and more precise than the estimates using all the variables. This also suggests the presence of overfitting when no variable selection is done.

Regarding Italian dataset, $\mathrm{GA}_{\mathrm{AIC}}$ also increased the performance of the regression model, corroborating the findings for Norwegian dataset. Although R^{2*}, RMSE and $\mathrm{SD}_{R^{2*}}$ for $\mathrm{GA}_{\mathrm{AIC}}$ subset were quite similar to values for the single variable subset, the standard deviation for RMSE was ~32% smaller for $\mathrm{GA}_{\mathrm{AIC}}$, providing evidence of stability in this procedure through different calibration and validation datasets. In addition, the results of variable selection in the Italian dataset were less consistent than the results for Norwegian dataset. Firstly, special attention should be paid to the results of the regression with all variables for which R^{2*} was negative (−0.31) and was accompanied by a large RMSE (85.26). Standard deviations for these statistics were large, also. These results provide strong evidence of a greater

Table 1 Results of variable subset selection using the field measurements in Norwegian and Italian datasets

Selection	Norway				Italy			
	R^{2*}	RMSE	$\mathrm{SD}_{R^{2*}}$	$\mathrm{SD}_{\mathrm{RMSE}}$	R^{2*}	RMSE	$\mathrm{SD}_{R^{2*}}$	$\mathrm{SD}_{\mathrm{RMSE}}$
None	0.56	56.60	0.24	14.73	−0.31	85.26	1.42	36.26
$\mathrm{GA}_{\mathrm{RMSE}}$	0.69	48.56	0.14	9.89	0.40	54.26	0.38	14.14
$\mathrm{GA}_{\mathrm{AIC}}$	0.74	43.68	0.14	10.69	0.55	46.47	0.30	8.19
RF	0.64	52.35	0.15	11.20	0.43	55.83	0.48	12.46
Single var.	0.67	50.57	0.10	8.14	0.54	46.49	0.25	11.99

†R^{2*} and RMSE are means; $\mathrm{SD}_{R^{2*}}$ and $\mathrm{SD}_{\mathrm{RMSE}}$ are the standard deviations of the respective statistics over 100 iterations. Values obtained using the validation dataset (30% of the data).

degree of overfitting in this dataset than in the Norwegian dataset, probably because of the greater biological variability in the Italian forest. Also, overall RF results were the less efficient again when compared with the methods based on GA, specially GA_{AIC}. R^{2*} for GA_{AIC} (0.55) was 27% largest than R^{2*} for RF (0.43). Regarding RMSE, the value was 20% smaller for GA_{AIC}. Similar proportions were found for the stability measurements. The results of both analyses corroborate the hypothesis that, among the methods used, GA_{AIC} yielded the most reliable results concerning accuracy.

Figure 3 shows the VOL observations versus VOL predictions (Italy) and biomass observations versus biomass predictions (Norway) for the nonlinear logistic model calibrated with the subset of variables selected by GA_{AIC} and with all variables. The dashed line represents the theoretical 1:1 observed versus predicted line for the full model. In the absence of lack of fit, the points in these graphs should lie along this straight line. The *F*-test for comparing estimates of the intercepts and slopes jointly to (0,1) indicated no significant differences for the models with selected subsets of variables by GA_{AIC}, which indicates no

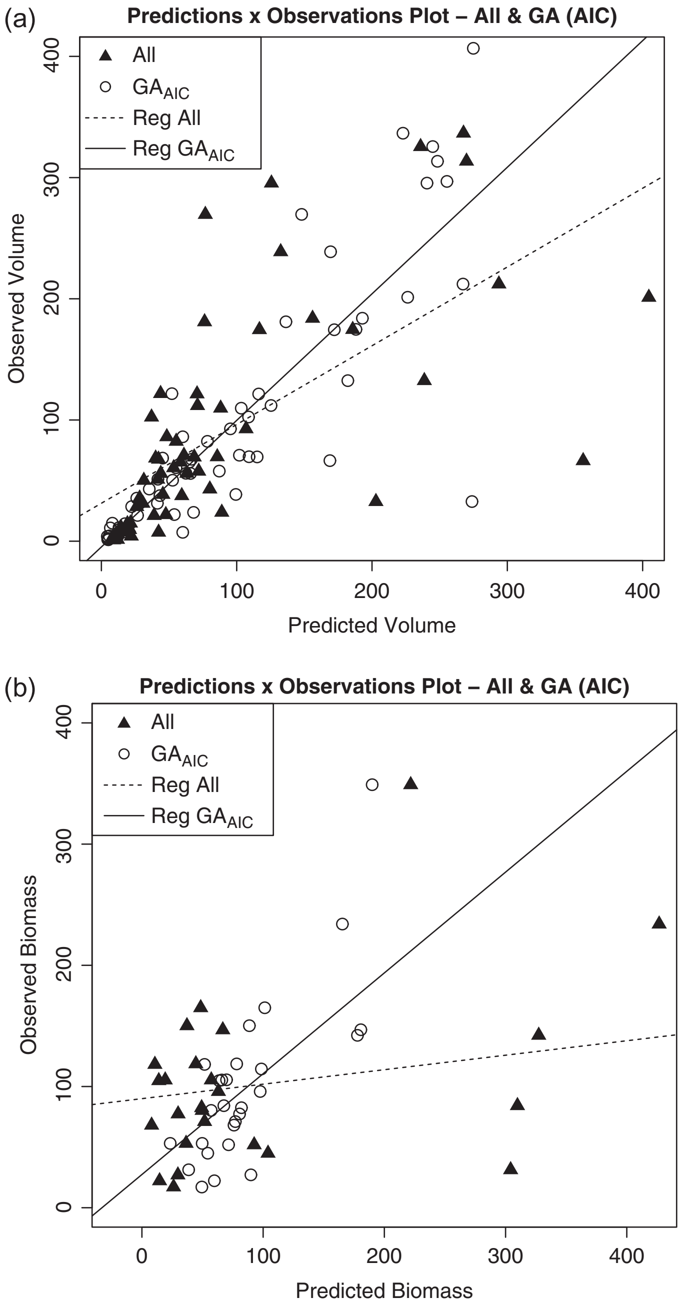

Figure 3 (a) VOL observations versus nonlinear logistic model VOL predictions for Norwegian dataset; (b) biomass observations versus nonlinear logistic model biomass predictions for Italian dataset.

important lack of fit of the models to the data. Also, the accuracies obtained using our models are comparable to accuracies achieved by other LiDAR/biomass/volume studies (Næsset 2002, Latifi *et al*., 2010, Lu *et al*., 2012, He *et al*., 2013, McRoberts *et al*., 2013a, 2013b, McRoberts and Westfall 2014, McRoberts *et al*., 2014). The most important result was that the full model was inefficient for predicting biomass based on R^{2*}. Although visually explicit, conduction of the *F*-test led to the rejection of null hypothesis, providing more evidence that the estimates of this model are less reliable than estimates for model with fewer predictor variables. Graphs for other subsets were quite similar.

Population parameters inference (SRS estimator)

The SRS estimator yielded a value of $\hat{\mu}_{SRS} = 88.99\ m^3/ha$ with a standard error of $SE(\hat{\mu}_{SRS}) = 8.30\ m^3/ha$ for the entire study area in Norway. These values were used for comparison for the results obtained with the GREG estimator with different subsets of predictor variables, especially for comparison of variances.

For the entire Italian study area, we obtained $\hat{\mu}_{SRS} = 108.15$ t/ha with a standard error of $SE(\hat{\mu}_{SRS}) = 11.20$ t/ha. Before the analyses, four plots were considered outliers and excluded from the analysis. The outlier status is likely because of forest harvest that occurred between the ALS acquisition and the field survey.

Note that the dependent variable for the Italian dataset is biomass rather than volume, but the results corroborated the findings for the Norwegian dataset. In fact, overfitting in the models for the Italian dataset is much more evident and leads to less meaningful predictions when considering all the variables. The analyses were conducted exactly in the same way as for the Norwegian dataset. The difference in the dependent variables occurred because the two datasets where collected for different studies. This difference is negligible with respect to the overall results for two reasons: (1) volume is closely related to biomass because it is the basis for biomass assessment and (2) the similarity between the results for the two study area suggests that the proposed methodology is equally efficient, regardless the response variable that is being estimated.

Discussion

Regarding the Norwegian dataset, the results of RF variable selection and the model with one predictor variable were the less accurate and efficient, concerning SE and relative efficiency (Table 2). The results for RF corroborate the hypothesis of (Strobl *et al*. 2007) that RF variable importance is not reliable when variables of the same type vary in many levels in the present sample. These authors used a series of simulations to show that RF algorithm may artificially prefer suboptimal predictor variables in two scenarios: when predictor variables vary in their scale of measurement or when variables of the same type vary in their number of categories. The former does not occur here. The latter is exactly the case of height and density returns, that are stratified in percentiles, creating many 'levels' for the same variable, which leads to a suboptimal selection of variables by RF technique.

Concerning the model with one predictor variable, the one most highly correlated to VOL (d9) was an efficient predictor when interacting with others, but it did not have enough information to be used alone in the model. Under this view GA_{RMSE} is a good choice. Again, the model with no variable selection produced the largest standard deviations, almost twice the standard deviations for the models with selected variables. This is a strong evidence that overfitting occurred, affecting the robustness of the model over different calibration and estimation datasets and leading to poor predictions. Given this result, extreme caution should be exercised when using a model with all variables included.

Table 2 Results of VOL estimation for Norwegian dataset using the nonlinear logistic regression model with predictor variables selected by different methods

Selection	$\hat{\mu}_{GREG}$	$\widehat{VAR}(\hat{\mu}_{GREG})$	RE	$SD(\widehat{VAR}(\hat{\mu}_{GREG}))$	SD_{RE}
None	81.84	3.67	5.29	0.41	1.01
GA_{RMSE}	81.94	3.91	4.51	0.11	0.25
GA_{AIC}	80.10	3.61	5.31	0.09	0.25
RF	76.54	4.42	3.54	0.14	0.19
Single var.	79.10	4.63	3.22	0.03	0.05

†RE is the relative efficiency (related to SRS estimator) and $SD(\widehat{VAR}(\hat{\mu}_{GREG}))$ and SD_{RE} are the standard deviations of the respective statistics over 100 iterations.

The GREG estimates for the Norwegian area were smaller than the SRS estimates for all the combinations of variables (min = 76.54 m^3/ha; max = 81.94 m^3/ha). Also, the means and SEs were not statistically different among 100 iterations. This result is expected when using regression models, because parameter estimates for predictor variables that are unrelated to the response variable are usually not statistically significantly different from zero and, therefore, have little effect on the model predictions.

Relative efficiency increases when using the variables selected by GA_{AIC} but not for GA_{RMSE}. This may occur because RMSE is more closely related to R^{2*} than is AIC.

The standard deviations of $\widehat{VAR}(\hat{\mu}_{GREG})$ and RE, analysed together with the parameters estimates, are the main result of this study. GA_{AIC} yielded the second smallest value for $SD(\widehat{VAR}(\hat{\mu}_{GREG}))$ and the third smallest value for SD_{RE}. This indicates that the model with variables selected by GA_{AIC} is ~65% more stable with respect to estimates of SEs and ~75% more stable with respect to RE than the full model. This result is more evident when considering that GA_{AIC} produced the largest RE and the second smallest $\widehat{VAR}(\hat{\mu}_{GREG})$. The single variable subset was the most stable (~95% more stable with respect to the full model), but both $\widehat{VAR}(\hat{\mu}_{GREG})$ and RE were the largest. This result may be explained by the fact that because this subset has only one variable (*d*9), it leads to more similar results over the 100 iterations, but more imprecise predictions at each iteration.

Given these results, GA_{AIC} is the more reliable method for selecting variables with respect to stability, reducing overfitting in an stable way, increasing the efficiency of the estimates and producing small SEs. This result may occur because of the penalty term in the AIC algorithm; while RMSE always leads to fewer parameters, AIC tends to select the most efficient and informative model based on the assumption that the model is a good reflection of reality (Sileshi 2014).

Regarding the Italian study area (Table 3), the smallest value for the GREG estimator was obtained from the full model ($\hat{\mu}_{\mathrm{GREG}}$ = 104.92 t/ha). As for Norway, means and SEs were not statistically different among the 100 iterations.

Results of this analysis were similar to the results for the Norwegian analysis, with $\mathrm{GA}_{\mathrm{AIC}}$ producing larger RE and smaller $\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}})$. Also, $\mathrm{GA}_{\mathrm{AIC}}$ was more stable over the 100 iterations. Regarding stability, $\mathrm{SD}(\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}}))$ for $\mathrm{GA}_{\mathrm{AIC}}$ was ~60% smaller than the second smallest one ($\mathrm{GA}_{\mathrm{RMSE}}$) and for $\mathrm{SD}_{\mathrm{RE}}$, $\mathrm{GA}_{\mathrm{AIC}}$ yielded a value ~30% smaller than the single variable subset. For all the statistics, the full model presented the less reliable values, probably a direct consequence of strong overfitting.

GA implementation and efficiency

Based on the statistics aforementioned, $\mathrm{GA}_{\mathrm{AIC}}$ produced the most suitable subset of variables among the assessed methods. Our analyses included a search for the most suitable subset over all possible subsets. This procedure is computationally intensive because of the extremely large number of combinations and should be avoided in operational use. The analysis was conducted only with the Italian dataset, because we found more convergence problems in these data than in the Norwegian dataset.

Table 4 shows the fitting statistics when applying the nonlinear model to different variable subsets with different sizes. These statistics were calculated using all the data to fit and test model, so we are not assessing overfitting based on this approach. Note that the same combination, for each number of variables, produces the largest R^{2*}, the smallest RMSE and the smallest AIC. Based on that, Table 4 shows the optimal results for each size of predictor variables subsets. As expected R^{2*} increased and both RMSE and AIC initially decreased with the inclusion of more variables. A different behaviour happened when fitting the model with 18 and 19 variables. This fact may occur for two reasons: (1) the number of variables is large when compared with the sample size and (2) the last variable included (P99) is highly correlated ($r > 0.95$) with seven other variables (37%), which inhibits convergence of the NLS routine. So, excluding these two last cases, Table 4 indicates that R^{2*} has an asymptote of ~0.83, which is reached with eight variables. Our main goal here is to show that there is a threshold for the maximum number of predictor variables with little to be gained beyond the threshold.

Table 3 Results of biomass estimation for whole area of Italian dataset using the nonlinear logistic regression model with predictor variables selected by different methods

Selection	$\hat{\mu}_{\mathrm{GREG}}$	$\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}})$	RE	$\mathrm{SD}(\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}}))$	$\mathrm{SD}_{\mathrm{RE}}$
None	104.92	8.88	2.89	7.88	1.50
$\mathrm{GA}_{\mathrm{RMSE}}$	108.51	5.70	3.95	0.53	0.56
$\mathrm{GA}_{\mathrm{AIC}}$	108.54	5.34	4.43	0.22	0.32
RF	108.05	5.53	4.21	0.58	0.69
Single var.	108.80	5.66	4.02	0.80	0.45

†RE is the relative efficiency concerning SRS estimator and SD ($\widehat{\mathrm{VAR}}(\hat{\mu}_{\mathrm{GREG}})$) and $\mathrm{SD}_{\mathrm{RE}}$ are the standard deviations of the respective statistics over 100 iterations.

Given this result, to assess the efficiency of the $\mathrm{GA}_{\mathrm{AIC}}$ algorithm, we evaluate the number of iterations necessary to find the near optimal solution. Two aspects must be mentioned here: (1) the GA procedure is based on a random starting value so the number of iterations is usually different among runs and

Table 4 Goodness-of-fit statistics over all possible subsets of independent variables

Variables	Largest R^{2*}	Smallest RMSE	Smallest AIC
1	0.77	40.83	601.89
2	0.79	39.82	599.96
3	0.79	39.78	600.78
4	0.80	39.42	600.67
5	0.81	38.99	600.30
6	0.82	38.10	598.52
7	0.82	37.82	598.55
8	0.83	37.87	599.55
9	0.83	38.21	601.42
10	0.83	38.45	602.95
11	0.83	38.70	604.47
12	0.83	39.10	606.43
13	0.83	39.46	608.21
14	0.83	40.11	610.80
15	0.83	40.58	612.82
16	0.83	41.32	615.55
17	0.83	41.99	618.03
18	0.82	42.66	620.43
19	0.72	54.15	648.63

†The number of variables means that the optimal particular combination of that number of predictor variables produces the values in the other columns.
†RMSE values were calculated concerning the number of variables included in the nonlinear model.

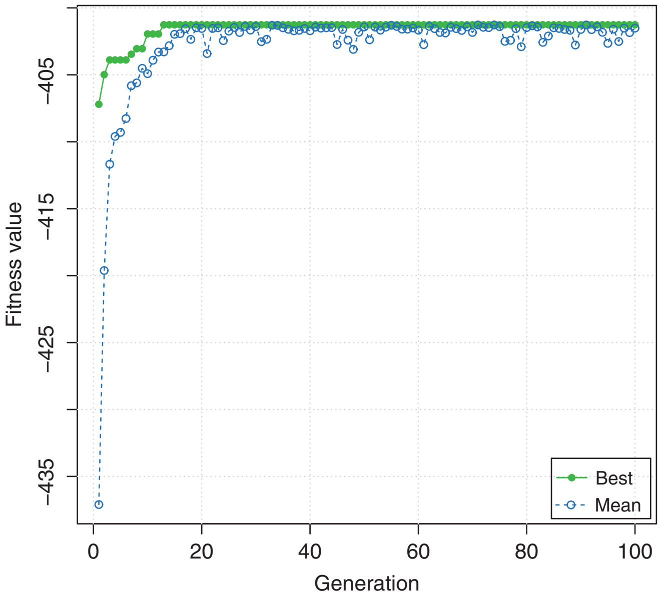

Figure 4 Results of GA procedure for select variables optimizing AIC.

(2) the optimal solution may not be found, although near optimal solutions are expected.

After 100 iterations with different seeds, the mean number of iterations necessary for the GA procedure to converge was ~15 iterations, yielding an AIC value of 607.5, with four variables retained. The results of this experiment are shown in Figure 4. Green dots represent the smallest AIC value for each generation of subsets, while blue dots represent the mean of the subsets. A visual analysis of the means shows that the GA procedure has a high degree of stability, corroborating the analytical results found previously and reported using the standard deviation of the results over all iterations with different training and test datasets.

Conclusions

Five conclusions may be drawn from these analysis. First, appropriate selection of predictor variables contributes to forest inventory by shortening CIs. In particular, if appropriate selection of the predictor variables reduces CI width, then that the sample size could be reduced considerably without affecting the CI. Alternatively speaking, appropriate selection of the predictor variables is equivalent to increasing the sample size by a comparable amount, noting that sample size is proportional to variance, not SE. In this study, GA_{AIC} shortened the CI ~15% and 10% for Norwegian and Italian datasets, respectively, compared with SRS estimator. In the NFI context, this can represent cost efficiency regarding the sampling effort. Second, overfitting is an inherent phenomenon when fitting models with a large number of variables and leads to poor predictions when applied to an independent estimation dataset. Thus, procedures for selecting the most suitable subset of variables are necessary, especially when using ALS data that produce a large number of correlated metrics. Third, models with fewer predictor variables tend to be more stable over different calibration and validation datasets, even if the overall estimated means did not change too much. These insignificant differences in means probably are related to the property of regression models that assign parameter estimates that are not significantly different from zero to variables that are unrelated with the response variable. In addition, decreases in the standard deviation of mean estimates over the replications are obtained when using models with fewer predictor variables. Fourth, RF variable importance is not reliable when variables of the same type vary in many levels in the present sample. Because this is likely to occur with LiDAR data, extra attention is required when using this technique for selection of variables. Fifth, the GA with a cost function that minimizes the Akaike Information Criteria was the most efficient and most stable method for selecting the subsets of variables among the tested algorithms. We attribute this result to the penalty that this criterion applies when adding parameters, a feature that leads to a more parsimonious model with a minimum loss of information. That is not the case for RMSE, which always tends to decrease with the addition of new variables and leads to inclusion of predictor variables that are not significantly correlated with the response variable.

Funding

We thank U.S. Forest Service and CNPq for a research grant received by the two first authors in July 2014. The first author also thanks to Programa UNIEDU Pos-Graduacao (FUMDES) for a scholarship received since 2015.

References

Anderson-Sprecher, R. 1994 Model comparisons and R^2. *Am. Stat.* **48** (2), 113–117.

Berger, A., Gschwantner, T., McRoberts, R.E. and Schadauer, K. 2014 Effects of measurement errors on individual tree stem volume estimates for the austrian national forest inventory. *For. Sci.* **60** (1), 14–24.

Braastad, H. 1966 Volume tables for birch. *Meddelelser Norske Skogforsksvesen* **21** (1), 265–365.

Brantseg, A. 1967 Volume functions and tables for scots pine. South Norway. *Meddelelser Norske Skogforsksvesen* **22**, 689–739.

Breidenbach, J., Antón-Fernández, C., Petersson, H., McRoberts, R.E. and Astrup, R. 2014 Quantifying the model-related variability of biomass stock and change estimates in the norwegian national forest inventory. *For. Sci.* **60** (1), 25–33.

Breidt, F.J., Opsomer, J.D. 2009 Nonparametric and semiparametric estimation in complex surveys. In *Handbook of Statistics - Sample Surveys: Inference and Analisys* **29**. Pfeffermann D. and Rao C. R. (eds). Elsevier, pp. 103–120.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Broadhurst, D., Goodacre, R., Jones, A., Rowland, J.J. and Kell, D.B. 1997 Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta.* **348** (1), 71–86.

Brown, S. 2002 Measuring carbon in forests: current status and future challenges. *Environ. Pollution* **116** (3), 363–372.

Castellani, C., Scrinzi, G., Tabacchi, G. and Tosi, V. 1984 *Inventario forestale nazionale italiano (IFNI): tavole di cubatura a doppia entrata.* Ministero dell'Agricoltura e delle Foreste.

Cateni, S., Colla, V. and Vannucci, M. 2011 A genetic algorithm-based approach for selecting input variables and setting relevant network parameters of a som-based classifier. *In International Journal of Simulation Systems, Science & Technology. UKSim 4th European Modelling Symposium on Mathematical modelling and computer simulation*, Vol. 12.

Chave, J., Andalo, C., Brown, S., Cairns, M., Chambers, J., Eamus, D., *et al.* 2005 Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia.* **145** (1), 87–99.

Chen, Q., Baldocchi, D., Gong, P. and Kelly, M. 2006 Isolating individual trees in a savanna woodland using small footprint LiDAR data. *Photogrammetric Engg. & Remote Sens.* **72** (8), 923–932.

Chirici, G., McRoberts, R.E., Fattorini, L., Mura, M. and Marchetti, M. 2016 Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sens. Environ.* **174**, 1–9.

Cormier, K.L., Reich, R.M., Czaplewski, R.L. and Bechtold, W.A. 1992 Evaluation of weighted regression and sample size in developing a taper model for loblolly pine. *For. Ecology Manage.* **53** (1), 65–76.

D'heygere, T., Goethals, P. and De Pauw, N. 2002 Use of genetic algorithms to select input variables in artificial neural network models for the prediction of benthic macroinvertebrates. In *Integrated Assessment and Decision Support proceedings of the 1st biennial meeting of the International Environmental Modelling and Software Society*, Vol. 2, pp. 136–141.

Ene, L.T., Næsset, E., Gobakken, T., Gregoire, T.G., Ståhl, G. and Nelson, R. 2012 Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. *Remote Sens. Environ.* **123**, 579–592.

Federici, S., Vitullo, M., Tulipano, S., De_Lauretis, R. and Seufert, G. 2008 An approach to estimate carbon stocks change in forest carbon pools under the unfccc: the italian case. *iForest-Biogeosciences and Forestry* **1** (2), 86–95.

Fitje, A. and Vestjordet, E. 1977 Stand height curves and new tariff tables for Norway spruce. *Meddelelser Norske Skogforsksvesen* **34** (2), 23–62.

Garcia-Gutierrez, J., Gonzalez-Ferreiro, E., Riquelme-Santos, J.C., Miranda, D., Dieguez-Aranda, U. and Navarro-Cerrillo, R.M. 2014 Evolutionary feature selection to estimate forest stand variables using lidar. *Int. J. Appl. Earth Observation and Geoinformation* **26**, 119–131.

Garey, M.R. and Johnson, D.S. 1979 Computers and intractability: an introduction to the theory of NP-completeness. W. H. Freeman.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. 2010 Variable selection using random forests. *Pattern Recognit. Lett.* **31** (14), 2225–2236.

Goldberg, D. E. and Holland, J. H. 1988 Genetic algorithms and machine learning. *Machine Learning* **3** (2), 95–99.

Haapanen, R. and Tuominen, S. 2008 Data combination and feature selection for multi-source forest inventory. *Photogrammetric Engg. & Remote Sens.* **74** (7), 869–880.

Hansen, M.H., Madow, W.G. and Tepping, B.J. 1983 An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J. Am. Stat. Assoc.* **78** (384), 776–793.

Hapfelmeier, A. and Ulm, K. 2013 A new variable selection approach using random forests. *Comput. Stat. Data. Anal.* **60**, 50–69.

Harrell, F.E. 2013 *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media.

Hastie, T., Tibshirani, R. and Friedman, J. 2009 *The elements of statistical learning*. Springer.

He, Q., Chen, E., An, R. and Li, Y. 2013 Above-ground biomass and biomass components estimation using LiDAR data in a coniferous forest. *Forests* **4**, 984–1002.

Hese, S., Lucht, W., Schmullius, C., Barnsley, M., Dubayah, R., Knorr, D., *et al.* 2005 Global biomass mapping for an improved understanding of the CO^2 balance - the Earth observation mission Carbon-3D. *Remote Sens. Environ.* **94** (1), 94–104.

Holland, J. 1992 *Adaptation in natural and artificial systems*. MIT Press, .

Kohavi, R. and Sommerfield, D. 1995 Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In *KDD*, pp. 192–197.

Latifi, H., Nothdurft, A. and Koch, B. 2010 Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/lidar-derived predictors. *Forestry* **83** (4), 395–407.

Lehtonen, R., Särndal, C. and Veijanen, A. 2005 Does the model matter? comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statist. Transit.* **7** (3), 649–673.

Lu, D., Chen, Q., Wang, G., Moran, E., Batistella, M., Zhang, M., *et al.* 2012 Aboveground forest biomass estimation with Landsat and LiDAR data and uncertainty analysis of the estimates. *Int. J. For. Res.* **2012**.

McRoberts, R.E., Cohen, W.B., Naesset, E., Stehman, S.V. and Tomppo, E. O. 2010 Using remotely sensed data to construct and assess forest attribute maps and related spatial products. *Scandinavian J. For. Res.* **25** (4), 340–367.

McRoberts, R.E., Moser, P., Zimermann Oliveira, L. and Vibrans, A.C. 2015 A general method for assessing the effects of uncertainty in individual-tree volume model predictions on large-area volume estimates with a subtropical forest illustration. *Can. J. For. Res.* **45** (1), 44–51.

McRoberts, R.E., Naesset, E. and Gobakken, T. 2013a Accuracy and precision for remote sensing applications of nonlinear model-based inference. *IEEE J. Select. Topics Appl. Earth Observat. Remote Sens.,* **6** (1), 27–34.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2013b Inference for LiDAR-assisted estimation of forest growing stock volume. *Remote Sens. Environ.* **128**, 268–275.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2014 Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information. *Remote Sens. Environ.* **154**, 226–233.

McRoberts, R.E. and Westfall, J.A. 2014 Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. *For. Sci.* **60** (1), 34–42.

Næsset, E. 2002 Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* **80** (1), 88–99.

Næsset, E. 2011 Estimating above-ground biomass in young forests with airborne laser scanning. *Int. J. Remote. Sens.* **32** (2), 473–501.

Nelson, R. 2013 How did we get here? An early history of forestry LiDAR. *Can. J. Remote Sens.* **39**, 6–17.

Nelson, R., Krabill, W.B. and Tonelli, J. 1988 Estimating forest biomass and volume using airborne laser data. *Remote Sens. Environ.* **24** (2), 247–267.

Ni-Meister, W., Lee, S., Strahler, A.H., Woodcock, C.E., Schaaf, C., Yao, T., *et al.* 2010 Assessing general relationships between aboveground biomass and vegetation structure parameters for improved carbon estimate from LiDAR remote sensing. *J. Geophys. Res.* **115**.

Parker, G.G. and Russ, M.E 2004 The canopy surface and stand development: assessing forest canopy structure and complexity with near-surface altimetry. *For. Ecol. Manage.* **189** (1), 307–315.

Paterlini, S. and Minerva, T. 2010 Regression model selection using genetic algorithms. In *Proceedings of the 11th WSEAS International Conference on Neural Networks and Evolutionary Computing and Fuzzy Systems*, pp. 19–27.

Santos, E.M.D., Sabourin, R. and Maupin, P. 2009 Overfitting cautious selection of classifier ensembles with genetic algorithms. *Information Fusion* **10** (2), 150–162.

Särdnal, C., Swensson, B. and Wretman, J. 1992 *Model assisted survey sampling*. Springer-Verlag.

Scrinzi, G., Clementel, F., Colle, G., Corona, P., Floris, A., Maistrelli, F., *et al.* (2013). Impiego di dati lidar di pubblica disponibilità per il monitoraggio forestale a grande e piccola scala: il progetto italid. In *Proceedings of the 9th SISEF National Congress 'MultifunzionalitÃ degli Ecosistemi Forestali: Sfide e OpportunitÃ per la Ricerca e lo Sviluppo'*, pp. 16–19.

Shettles, M., Temesgen, H., Gray, A.N. and Hilker, T. 2015 Comparison of uncertainty in per unit area estimates of aboveground biomass for two selected model sets. *For. Ecol.Manage.* **354**, 18–25.

Sileshi, G. W. 2014 A critical review of forest biomass estimation models, common mistakes and corrective measures. *For. Ecol. Manage.* **329**, 237–254.

Ståhl, G., Heikkinen, J., Petersson, H., Repola, J. and Holm, S. 2014 Sample-based estimation of greenhouse gas emissions from forestsa new approach to account for both sampling and model errors. *For. Sci.* **60** (1), 3–13.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. 2008 Conditional variable importance for random forests. *BMC. Bioinformatics.* **9** (1), 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. 2007 Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC. Bioinformatics.* **8** (1), 25.

Tomter, S.M., Hylen, G., Nilsen, J.-E. 2010 Development of Norway's national forest inventory. In *National Forest Inventories. Pathways for common reporting*. Tomppo E., Gschwantner T., Lawrence M. and McRoberts R.E. (eds). Springer, pp. 411–424.

Tuominen, S., Haapanen, R., *et al*. 2013 Estimation of forest biomass by means of genetic algorithm-based optimization of airborne laser scanning and digital aerial photograph features. *Silva Fennica* **47** (1), 902.

Vallet, P., Dhôte, J.-F., Moguédec, G.L., Ravart, M. and Pignard, G. 2006 Development of total aboveground volume equations for seven important forest tree species in France. *For. Ecol. Manage.* **229** (1), 98–110.

Vanclay, J.K. and Skovsgaard, J.P. 1997 Evaluating forest growth models. *Ecol. Modell.* **98** (1), 1–12.

Vestjordet, E. 1967 Functions and tables for volume of standing trees. Norway spruce. *Meddelelser Norske Skogforsksvesen* **22**, 539–574.

Vinterbo, S. and Ohno-Machado, L. 1999 A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. In *Proceedings of the AMIA Symposium*, American Medical Informatics Association. p. 984.

Yao, T., Yang, X., Zhao, F., Wang, Z., Zhang, Q., Jupp, D., *et al*. 2011 Measuring forest structure and biomass in New England forest stands using Echidna ground-based LiDAR. *Remote Sens. Environ.* **115** (11), 2965–2974.

Zhang, G.P. 2005 Neural networks for data mining. In *Data mining and knowledge discovery handbook* **2**. Maimon O. and Rokach L. (eds). Springer, pp. 419–444.

Zhao, F., Guo, Q. and Kelly, M. 2012 Allometric equation choice impacts LiDAR-based forest biomass estimates: a case study from the Sierra National Forest, CA. *Agricult. For. Meteorol.* **165**, 64–72.

Zhao, F., Yang, X., Schull, M.A., Román-Colón, M.O., Yao, T., Wang, Z., *et al*. 2011 Measuring effective leaf area index, foliage profile, and stand height in New England forest stands using a full-waveform ground-based LiDAR. *Remote Sen. Environ.* **115** (11), 2954–2964.

Zheng, H. and Little, R.J. 2003 Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *J. Off. Stat.* **19** (2), 99.

Zianis, D. and Seura, S.M. 2005 *Biomass and stem volume equations for tree species in Europe* **Vol. 4**. Finnish Society of Forest Science, Finnish Forest Research Institute.