Forestry An International Journal of Forest Research

Institute of Chartered Foresters

Forestry 2021; **94**, 311–323, doi:10.1093/forestry/cpaa034
Advance Access publication 3 October 2020

Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock

Diogo N. Cosenza[1,*], Lauri Korhonen[2], Matti Maltamo[2], Petteri Packalen[2], Jacob L. Strunk[3], Erik Næsset[4], Terje Gobakken[4], Paula Soares[1] and Margarida Tomé[1]

[1]*Forest Research Centre, School of Agriculture, University of Lisbon, Tapada da Ajuda, 1349-017 Lisbon, Portugal*
[2]*School of Forest Sciences, University of Eastern Finland P.O. Box 111, 80101 Joensuu, Finland*
[3]*USDA Forest Service, Pacific Northwest Research Station, 3625 93rd Ave SW, Olympia, WA 98512, USA*
[4]*Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway*

*Corresponding author: Tel: +351 213 653 309; E-mail: dncosenza@gmail.com and dncosenza@isa.ulisboa.pt

Received 28 January 2020

In this study, for five sites around the world, we look at the effects of different model types and variable selection approaches on forest yield modelling performances in an area-based approach (ABA). We compared ordinary least squares regression (OLS), k-nearest neighbours (kNN) and random forest (RF). Our objective was to test if there are systematic differences in accuracy between OLS, kNN and RF in ABA predictions of growing stock volume. The analyses are based on a 5-fold cross-validation at five study sites: an eucalyptus plantation, a temperate forest and three different boreal forests. Two completely independent validation datasets were also available for two of the boreal sites. For the kNN, we evaluated multiple measures of distance including Euclidean, Mahalanobis, most similar neighbour (MSN) and an RF-based distance metric. The variable selection approaches we examined included a heuristic approach (for OLS, kNN and RF), exhaustive search among all combinations (OLS only) and all variables together (RF only). Performances varied by model type and variable selection approaches among sites. OLS and RF had similar accuracies and were more efficient than any of the kNN variants. Variable selection did not affect RF performance. Heuristic and exhaustive variable selection performed similarly for OLS. kNN fared the poorest amongst model types, and kNN with RF distance was prone to overfitting when compared with a validation dataset. Additional caution is therefore required when building kNN models for volume prediction though ABA, being preferable instead to opt for models based on OLS with some variable selection, or RF with all variables together.

Introduction

The area-based approach (ABA) is commonly used in applications of forest yield modelling with airborne laser scanning (ALS). In this approach field, measured observations on plots are related to remotely sensed observations for the same plots using a model. Ordinary least squares regression (OLS) was predominately used in early studies (Næsset, 2002, 2004), but machine learning techniques such as k-nearest neighbour (kNN; Dudani, 1976) and random forest (RF; Breiman, 2001) have also been used (Maltamo et al., 2006; Packalén and Maltamo, 2006, 2007). These methods increased in popularity recently (Pascual et al., 2019; Silva et al., 2017; Görgens et al., 2015). An advantage of the kNN and RF techniques is that they require less expertise to implement than linear modelling approaches or other more complex machine learning approaches such as neural networks and support vector regression (see Hastie et al., 2009; Haykin, 2009). While being straightforward to implement for practitioners, they are also efficient in handling high-dimensional data, which is often useful in remote sensing applications, and their requirement for more computational time is less of an issue today due to the increase in computing power.

kNN and RF methods have been used for prediction of both individual tree attributes (Gleason and Im, 2012; Maltamo et al., 2009; Vauhkonen et al., 2010; Yu et al., 2011) and plot-level attributes (Latifi and Koch, 2012; Maltamo et al., 2006; Packalén and Maltamo, 2007; Shataee et al., 2011). The principle of kNN is to predict the target variable based on the closest

observations in the training dataset (Dudani, 1976; Hastie et al., 2009). In kNN, a distance metric is used to determine which k training observations are nearest to the target value in terms of the predictor variables. Parameters that the modeller must decide for kNN include the number of nearest neighbours to impute, how to measure distance and the weighting scheme employed to average imputed values. The RF algorithm is based on decision trees formed from resamples of the input data. Each decision tree uses a randomly selected subset of both the available predictors and observations. After training the model, the average of the 'forest' of decision trees is used to predict new observations.

OLS and machine learning models typically require a number of decision and inspections which we will lump together as 'parameter tuning' steps. Correct specification of tuning parameters for OLS and machine learning methods requires specialized expertise. For regression analysis (OLS), for example, the analyst must select variables, fit the model and perform various diagnostics. kNN models also benefit from careful variable selection (Packalén et al. 2012) and the user must choose the number of neighbours and the distance metric. RF, on the other hand, has been shown to be less sensitive to variable selection (Belgiu and Drăgu, 2016). There are a number of RF parameters than can be tuned, although it is common for analysts to use the defaults.

According to a review of kNN methods by Chirici et al. (2016), $k = 5$ is the most frequent number of neighbours used in remote sensing applications, although values in the range 1–10 are common. Euclidean and Mahalanobis distance are often applied as distance metrics as they perform well relative to other methods and do not depend on response attributes, which makes them simple and fast to compute. Canonical correlation analysis (most similar neighbour–MSN; Moeur and Stage, 1995) and RF-based distance metrics (Lin and Jeon, 2002) are also frequently used. The weighting scheme used for prediction from multiple neighbours ($k > 1$) is usually the inverse of the distances to the k nearest neighbours.

In the case of the RF, Belgiu and Drăgu (2016) stated that the number of trees used to train the models is usually 500 since it is the default value of the most popular RF implementation, the *randomForest* package in R (see Liaw and Wiener, 2002). Furthermore, an exploratory study by Lawrence et al. (2006) showed that the error rates become stable when the number of trees is close to 500. The other parameters are normally set to the default values of the software, which are similar to those proposed by Breiman (2001). Thus, it is a common practice to train the RF model by not allowing to split the trees with less than five nodes and applying one-third of the available predictor variables in the inner bootstrap.

The kNN and RF algorithms have some operational advantages. They are multivariate, non-parametric and quick to implement; however, they also have several limitations. Since kNN and RF predictions are weighted averages of the training data, the algorithms cannot extrapolate beyond the range of the training data. kNN and RF methods also typically require a greater number of training observations than parametric regression to perform well. These methods work best if the training dataset is large (e.g. thousands of observations) such that the full range of population values is represented. The demand for training observations also increases with the number of predictors due to the 'curse of dimensionality' (Bellman, 1961). This issue is very important for kNN, to which increasing the dimensionality can cause the distance to the nearest neighbour to be barely distinguishable from that of a distant neighbour, generating a loss of 'contrast' among the distances (Beyer et al., 1999). This problem can be further accentuated if the kNN model uses 'irrelevant variables', which are redundant or have a low correlation with the response variable (Blum and Langley, 1997). Besides, irrelevant variables can also cause problems in the inversion of positive definite matrixes, a required step in computations of Mahalanobis and MSN distances (McRoberts et al., 2017).

To reduce problems associated with high-dimensional data, studies have used various strategies to select a subset of ALS metrics in RF and kNN models. Many approaches can be applied for this purpose (see Saeys et al., 2007), but it is common to see implementations based on stepwise variable selection (Breidenbach et al., 2010; Hudak et al., 2008; Maltamo et al., 2006), correlation feature selection (García-Gutiérrez et al., 2015), genetic algorithm (Latifi et al., 2010; Latifi and Koch, 2012; McRoberts et al., 2015) or simulated annealing (Packalén et al., 2012). Some researchers have suggested that RF may result in overfitting in cases of irrelevant variables (Segal, 2004), so it is common to see implementations of some form of variable selection for RF in an ABA context (see Genuer et al., 2010). To reduce dimensionality for RF applications, variables are often selected based on their importance (Shi et al., 2018; Silva et al., 2017).

OLS, kNN and RF are common methods for ABA forest yield modelling in the literature, but there has not been a systematic evaluation of the differences in accuracies between these methods, especially when considering differences in tuning strategies. The studies that we are aware of performed comparisons with at most three datasets (Fassnacht et al., 2014; Latifi and Koch, 2012; McRoberts et al., 2017), and most of them focused on boreal or temperate forests (Maltamo et al., 2006; Packalén et al., 2012; Packalén and Maltamo, 2007; Shataeea et al., 2011). Additionally, each study employed its own tuning approach so that comparisons among them are difficult.

Our objective for this study is to identify whether there are modelling and tuning strategies, which are optimal across a broad range of conditions, including sites in North America, South America and Europe, and to identify similarities and differences in OLS, RF and kNN predictions among these datasets. To pursue this objective, we fit models for growing stock volume (V, m^3 ha^{-1}) from lidar data for each of these datasets using OLS, kNN and RF using multiple tuning strategies. We then cross-validated model performances using 5-fold cross-validation. We also evaluated performances using independent validation datasets, which were available for two boreal forest sites in Europe.

Material and methods

Study areas and material

This work was conducted using datasets collected at five locations on three continents: a eucalyptus plantation in Brazil, boreal forests in eastern Finland (Kolari) and northern Finland (Liperi), a boreal forest in southern Norway and a mixed temperate forest in South Carolina in the USA. The sites in Finland are referred to by their municipality, Kolari and Liperi, while the other sites are referred to by their country. Independent validation datasets were available for the two sites Liperi (Finland) and Norway (see section ***Accuracy assessment*** for additional details). A summary of characteristics for each dataset is provided in Table 1. All of the sites have also been used for previous studies; comprehensive descriptions of study areas can be found in the primary articles cited below.

The site in Brazil is a commercial plantation of hybrid eucalyptus located in the state of Bahia (Packalén et al., 2011a, 2011b). At the time of the lidar acquisition, the plantation was composed of several even-aged stands, planted with the same spacing. The tree ages were between 2 and 12 years, the tree heights varied from 17 to 41 m and growing stock varied from 111 to 656 m^3 ha^{-1}.

The Kolari site is located north of the Arctic Circle, where the forests mainly consisted of sparsely spaced Scots pines, although Norway spruce and deciduous trees were also present (Kotivuori et al., 2016). Tree heights ranged from 3 m to only 25 m, and the growing stock were between 4 and 462 m^3 ha^{-1}, the lowest range of volume per hectare of any of the sites.

The other Finnish site, Liperi, was located in eastern Finland, where the forests were denser than Kolari and the trees achieved heights greater than 30 m (Kukkonen et al., 2019). The amount of spruce and pine stands was relatively similar with a minority of deciduous stands. The tree heights ranged from 6 to 41 m, while the growing stock varied from 20 to 797 m^3 ha^{-1} in this dataset.

The Norwegian boreal site is located further south than the Finnish sites and had considerably larger differences in elevation than the two sites in Finland (Gobakken et al., 2013). The forests were dominated by Norway spruce due to the better fertility of the soils. The tree heights ranged from 4 to 37 m, while the growing stock varied from 4 to 692 m^3 ha^{-1}.

The site in the USA was a mix of pine plantations and mixed native hardwoods with a total of 62 different species present on the plots (Strunk et al., 2017). Tree heights for this site ranged from 5 to 38 m and the growing stock in this dataset had the greatest range, from 2 to 1227 m^3 ha^{-1}.

The ALS metrics used for each dataset were prepared for previous studies (referred to in Table 1) using different data processing software and therefore differ between the sites. However, the available ALS metrics were comparable among all sites, representing the vertical return distributions, density metrics, return intensity and return numbers. All of them are described in Appendix 1.

Growing stock modelling

The same variable selection and modelling protocols were used for each of the five datasets (Brazil, Kolari, Liperi, Norway,

Table 1 Summary of the field and ALS data.

Region	Forest type	Dominant species	Number of plots	Plot radius	Point density (pt m^{-2})	Volume* ($m^3 \cdot ha^{-1}$)	Reference
Brazil	Clonal hybrid plantation	*Eucalyptus urophylla* x *Eucalyptus grandis*	195	13 m	1.5	376.1 (132.8)	(Packalén et al., 2011a, 2011b)
Kolari (Finland)	Boreal	*Pinus sylvestris*; *Picea abies*; *Betula pendula*	534	9 m; 12.62 m; 12.65 m	0.6	100.9 (64.2)	(Kotivuori et al., 2016)
Liperi (Finland)	Boreal	*Pinus sylvestris*; *Picea abies*; *Betula pendula*	Train: 578 Test: 444	9 m; 12.6 m	4.8	Train: 186.4 (100.4) Test: 205.6 (94.6)	(Kukkonen et al., 2019)
Norway	Boreal	*Pinus sylvestris*; *Picea abies*; *Betula pendula*	Train: 499 Test: 78	8.92 m 17.84 m	0.7	Train: 211.4 (119.7) Test: 230.7 (119.3)	(Gobakken et al., 2013)
USA	Temperate	62 species dominated by *Pinus taeda* (30.2%) and *Pinus palustris* (15.6%); all other species represent less than 10% of dominant species	194	11.3 m; 22.6 m	6	228.3 (160.7)	(Strunk et al., 2017)

*Plot means growing stock volume with its respective standard deviation (in parenthesis).

USA). For simplification, we use the term 'training' hereafter to refer to both OLS model fitting and the construction of non-parametric models. Each of these approaches was paired with variable selection schemes that were applied every time a model was trained. The same heuristic selection of five predictor variables was used for OLS, kNN and RF (section ***Heuristic variable selection***), while other variable selection approaches differed by model type. The choice of five predictors was a compromise between not overfitting the OLS models and having an adequate number of predictors for the nonparametric methods. The specifics of each variable selection approach and modelling scheme are described in greater detail in the following sections.

OLS modelling

The OLS models were based on Formula (1), in which the square root transformation of the response variable was applied to avoid problems of heteroscedasticity. Because of this transformation, the back-transformed value requires a correction for bias. According to Gregoire et al. (2008), the residual variance of the models was added as described in Formula (2).

$$\sqrt{V_i} = \beta_{i0} + \beta_{i1}x_{i1} + \cdots + \beta_{im}x_{im} + \varepsilon_i \quad (1)$$

$$\hat{V}_i = \left(\sqrt{\hat{V}_i}\right)^2 + \sigma^2 \quad (2)$$

where V_i and $\hat{V}_i$ are the observed and the predicted timber volume (m^3 ha^{-1}) of the plot $i = 1, 2, \ldots, n$; β_{ij} is the model parameter $j = 0, 1, \ldots, m$ of the plot i; x_{ij} is the predictive variable $j = 1, 2, \ldots, m$ for the plot i; ε_i is the random error for plot i; and σ^2 is the residual variance of the model (m^3 ha^{-1}).

Two approaches were used for variable selection for OLS. The first was a machine-learning based **He**uristic approach (OLS_**He**), which means that the derived models are just for prediction purposes since statistical assumptions related to the linear models were ignored. In this approach, we used a five-predictor model ($m = 5$, see Formula (1)) and the predictors were defined by the heuristic variable selection algorithm (detailed in section ***Heuristic variable selection***). The second variable selection approach used for OLS was an all combinations **Exh**austive search algorithm (OLS_**Exh**) implemented in the R statistical environment (R Core Team, 2020). The best model met the following three criteria: smallest relative root mean squared error (RMSE%, Formula (3)), considering the constraints that estimated parameters are significantly different from zero (t-test, $\alpha = 5$ percent), and variance inflation factors (VIF) are less than 10 (Myers, 1989). A three-predictor model ($m = 3$, see Formula (1)) was used in OLS_Exh due to computational limitations. RMSE% was computed as

$$\text{RMSE } (\%) = 100 * \overline{V}^{-1} \sqrt{\frac{\sum_{i=1}^{n}\left(V_i - \hat{V}_i\right)^2}{n}} \quad (3)$$

where V_i and $\hat{V}_i$ are, respectively, the observed and predicted timber volume (m^3 ha^{-1}) for plot $i = 1, 2, \ldots, n$; $\overline{V}$ is the observed mean volume (m^3 ha^{-1}).

kNN modelling

As shown by Packalén et al. (2012) and McRoberts et al. (2015), the prediction accuracy of kNN is sensitive to the choice of distance metric. For this reason, we tested four commonly used distance metrics: **Euc**lidean distance (kNN_**Euc**), **Mah**alanobis distance (kNN_**Mah**), **M**ost **S**imilar **N**eighbour (kNN_**MSN**) and a distance metric based on the **R**andom **F**orest proximity matrix (kNN_**RF**). All the approaches were trained using $k = 5$. The predicted values were obtained as weighted means of the k neighbours where the weights were $1/(1 + d)$, d being the distance between a neighbour and the target. kNN models were trained using the *yaImpute* package (Crookston and Finley, 2015) with the default parameters besides distance metric and k. Given the large number of potential ALS-derived auxiliary variables, we used a heuristic variable selection routine (detailed in section ***Heuristic variable selection***) to choose five 'good' ALS metrics for each kNN model.

RF modelling

RF modelling was performed using the *randomForest* package (Liaw and Wiener, 2002) with all training parameters set to their default values, since they are well established in the literature (Belgiu and Drăgu, 2016; Breiman, 2001). The default number of random trees was 500, the default number of predictors randomly selected to grow a tree from each node was one-third of the available predictors, and the default minimum number of nodes was five for each tree. Latifi and Koch (2012) reported that having a large number of predictor variables is not an issue in RF. To confirm this finding, RF was used both with a **He**uristic variable selection with five variables (RF_**He**; detailed in section ***Heuristic variable selection***) and using **all** the available predictor variables (RF_**All**).

Heuristic variable selection

Heuristic variable selection was performed with the simulated annealing metaheuristic (Kirkpatrick et al., 1983) adapted by Packalén et al. (2012) for kNN modelling. We also use the same variable selection algorithm (Algorithm 1) for kNN, OLS and RF. The idea behind a metaheuristic strategy is to find a 'good' solution for a given 'cost function' iteratively, by making changes in a given initial solution (see Talbi, 2009). In our case, a solution (S) is any set of five variables used to train a model, and the cost function (f_{cost}) is the RMSE% obtained by computing the residuals of the trained model respective to a given solution (Formula (3)).

Algorithm 1. Variable selection by simulated annealing, where n is the number of iterations, i is the number of the current iteration ($i = 1, 2, \ldots, n$), T_0 is the initial temperature, T is the current temperature, S_0 is the initial random solution, S is a temporary solution, S' is a neighbour solution of S, S_{best} is the best solution found until the iteration i, $f_{neighbour}$ is the neighbour function, f_{cost} is the cost function, c is the associated cost of S, c' is the cost associate of S', c_{best} is the associated cost of S_{best}.

1. $n \leftarrow 500$
   $T_0 \leftarrow 0.2$
   $T \leftarrow T_0$
   $S \leftarrow S_0$
   $c \leftarrow f_{cost}(S)$
   $c_{best} \leftarrow c$
   $S_{best} \leftarrow S$
   $i \leftarrow 0$
2. **While** $i < n$ **do:**
   $S' \leftarrow f_{neighbor}(S, \frac{i}{n})$
   $c' \leftarrow f_{cost}(S')$
   $\Delta \leftarrow (c' - c) + 0.001$
   #Evaluation
   If $exp(-\Delta/T) >$ random(from:0, until:1), **then:**
   $c \leftarrow c'$; $S \leftarrow S'$
   #Recording the best solution found
   If $c' < c_{best}$, **then:**
   $c_{best} \leftarrow c'$; $S_{best} \leftarrow S'$
   #Cooling temperature
   $T \leftarrow max\left[0.001,\ T_0\left(1 - \frac{i}{0.8\,n}\right)\right]$
   $i \leftarrow i + 1$
3. **Return:** S_{best}

The changes to the solution are induced by the neighbourhood function ($f_{neighbour}$), which produces an alternative solution ('neighbour solution', S') based on the current solution and the proportion of iterations (1) that were performed (i/n, being $i = 1, 2, \ldots, n$ and n the total number of iterations). Initially, the algorithm randomly replaces two-thirds of the variables in the current solution with different variables. The number of variables replaced in each iteration is reduced over the duration of the variable selection until only one variable at a time is replaced after achieving 80 percent of the iterations. The parameter temperature (T) controls the probability of accepting non-improving solutions, where this probability decreases with the temperature. The initial temperature is high to allow exploring different solutions, and then decreases throughout the iterations in a process known as 'cooling'. In our adaptation, it decreased linearly until achieving 80 percent of the iterations, where the temperature was set to a very small positive value. With this configuration, the solution with the lowest error found is exploited in the last 20 percent of the iterations to produce a better neighbour solution so that temporary solutions with higher errors are no longer allowed. The initial temperature (T_0) and the number of iterations (n) were set to 0.2 and 500, respectively. The routine was implemented in the R environment.

Accuracy assessment

The combinations of modelling type and variable selection approaches resulted in a total of eight approaches including OLS_Exh, OLS_He, kNN_Euc, kNN_Mah, kNN_MSN, kNN_RF, RF_All and RF_He. The performances of these approaches were assessed for each study area using cross-validation and for two sites using an independent validation. The relative RMSE (RMSE%, Formula (3)) and squared Pearson's correlation computed from validation predictions were used to evaluate performances.

Cross-validation was implemented using a 5-fold approach in which a given dataset is randomly split into five evenly sized disjoint sets ('folds'). In each iteration, one of these folds was omitted and a new model was trained with the remaining data to predict the volume of the omitted fold. The fivefolds were combined after the five iterations, and the RMSE% was computed. The entire 5-fold cross-validation was performed 100 times for each site, and the mean RMSE% was used as a final measure to obtain a better view of the real accuracy, as the randomness involved in the variable selection process is decreased. The intervals containing 95 percent of the RMSE% values were also used for comparison.

Independent test datasets that were available for two sites (Liperi and Norway) were used for additional validation. For each site, a new model was first trained using the full dataset previously used in their respective cross-validations, and the resultant models were used to predict volume to the testing datasets. Finally, RMSE% was computed both for training and testing datasets. This process was repeated 100 times. The mean RMSE% and the intervals containing 95 percent of the errors were also used for accuracy assessment.

As a final assessment for each approach, we used the training dataset one more time to train a showcase model to predict volumes to the testing dataset so that we could provide figures with observed versus predicted values and compute squared Pearson correlation (r^2). For the approaches where the heuristic variable selection was used, we also provide the progress of the cost function during the optimization, where the temporary solutions were applied to predict volume to the training and testing dataset and compute their respective RMSE% in each iteration.

Results

Overall, all eight modelling approaches demonstrated relatively similar accuracies for most of the datasets in the 5-fold cross-validation (Figure 1). We saw increasing Mean RMSE% values for forest conditions with greater heterogeneity in forest composition (see section **Study areas and material**). For example, the greatest relative errors were observed for the temperate forest conditions in the USA dataset (RMSE = 34–46 percent), which also had the greatest observed variation in structure and species. The Brazil site is very homogeneous, especially in contrast to the USA, and had the smallest RMSE% values, ranging from 9 to 14 percent. The complexity of the boreal forests (Kolari, Liperi and Norway) fell between Brazil and the USA; they had RMSE% values ranging from 19 to 30 percent. Additionally, the 95 percent intervals for RMSE% for all prediction methods overlapped for the USA dataset. This means that inferences about the relative performances of the tested prediction approaches are less certain for this site, although the trends in performances generally supported the inferences we made from other datasets.

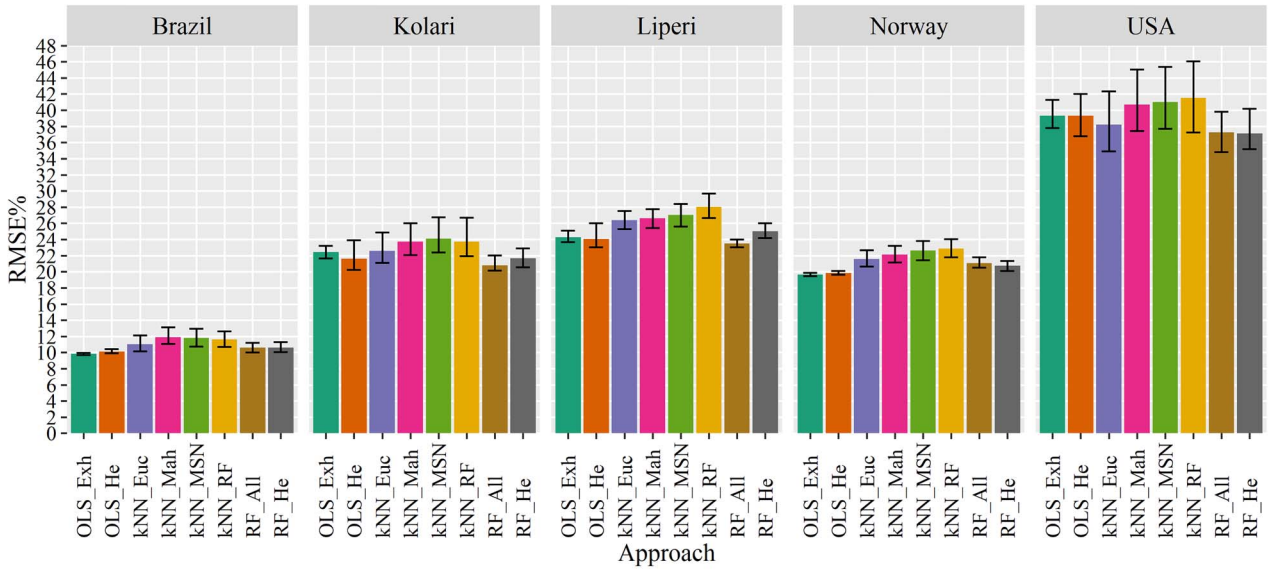

Figure 1 Mean accuracy of the modelling approaches by forest sites assessed through 5-fold cross-validations repeated 100 times. Each bar represents the mean of the relative RMSE% values obtained at the end of the 5-fold cross-validations for a given modelling approach. The whiskers represent intervals containing 95 percent of the RMSE%.

The patterns in performance relative to modelling approach varied by dataset, but there were consistent patterns observed across datasets. Our results suggest that mean RMSE% varied little amongst modelling approaches examined. The greatest variation in performance was amongst methods for the USA where kNN_RF and RF_He had a difference of approximately 5 percent in mean RMSE%. This was consistent with other sites in that kNN models almost exclusively performed poorer (had larger RMSE% values) than OLS and RF, and that OLS and RF generally performed similarly. There was only one exception where kNN performed better than other methods: for the USA, kNN_Euc performed better than OLS methods, while other kNN models performed worse than OLS for the USA. kNN_Euc consistently had the best performances amongst kNN distance metrics across all sites. Other kNN distance metrics did not demonstrate obvious trends, and their relative performances amongst kNN methods varied by site.

OLS and RF generally performed similarly across all sites (RMSE%). OLS performed better for Brazil and Norway, RF performed better for Kolari and the USA, and the result for Liperi was mixed. However, differences amongst OLS and RF were very small, with a maximum of 3 percentage points difference (USA), and most differences were less than 2 percentage points. The difference within methods (OLS_Exh versus OLS_He and RF_ALL versus RF_He) were generally also very small (<2 percentage points).

The validation at Liperi's and Norway's external datasets had RMSE% values ranging from 24 to 34 percent (Figure 2). These errors were greater (3 to 5 percentage points) than the cross-validation errors. In Liperi, validation on independent data demonstrated a similar pattern in performance to cross-validation (see Figure 1), where the best RF and OLS methods had nearly identical performances, kNN had the largest errors, and

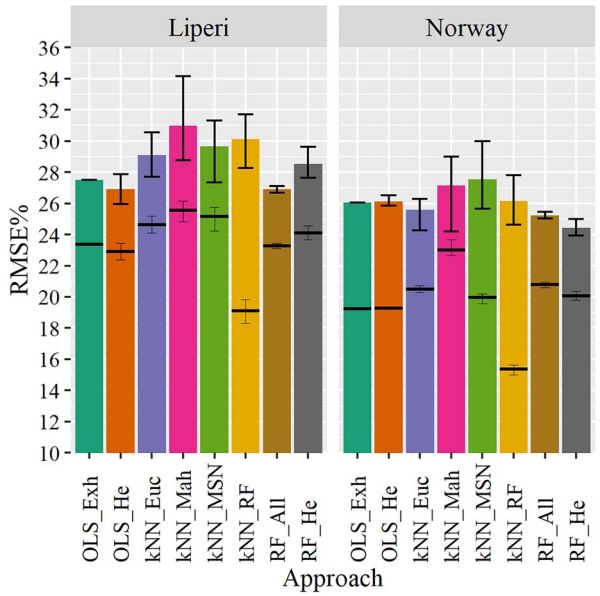

Figure 2 Mean accuracy of the modelling approaches by forest sites assessed through validations repeated 100 times. Each bold line and bar represents, respectively, the mean of the relative RMSE% values obtained at the end of the validation with the application of a trained model to the training and testing dataset. The whiskers represent intervals containing 95 percent of the RMSE%.

kNN_Euc performed best amongst kNN distances. Independent validation for Norway differed slightly in that kNN_Euc and kNN_RF performed similarly to OLS and RF models. Interestingly, the relative performance of RF_He clearly differed between the

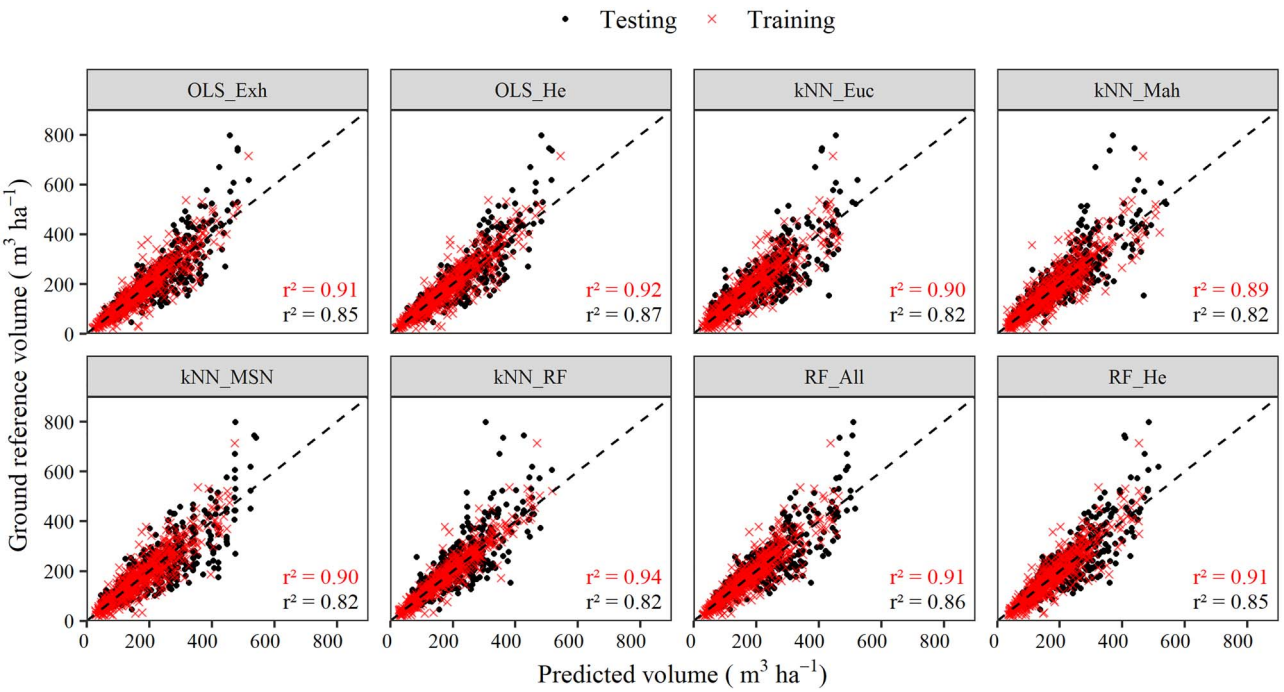

Figure 3 Observed and predicted values for testing and training datasets from Liperi.

two sites, performing better than all other methods in Norway, and performing worst amongst OLS and RF methods for Liperi. Furthermore, most of the testing errors followed a similar pattern as the training ones. The greatest difference was observed for kNN_RF, which had notably smaller RMSE% values for the training than for the testing datasets, showing that RF distance metric in kNN had a clear tendency for overfitting.

The relative agreement between observed and predicted values of the showcase models can be see for training and validation sets in Figure 3 (Liperi) and Figure 4 (Norway). As might be expected, the agreement between predicted (*x*-axis) and observed (*y*-axis) values was better for the training dataset than for the validation dataset. The r^2 values for the training datasets ranged from 0.89 to 0.96 and for independent validation datasets ranged from 0.82 to 0.89. As was observed with RMSE%, OLS and RF methods generally perform the best, and there is evidence that kNN_RF has a tendency to overfit the training dataset. For Liperi, kNN_RF performed the best for the training dataset ($r^2 = 0.94$), but performed among the worst with respect to the validation dataset ($r^2 = 0.82$). In Norway, a similar decline for kNN_RF was observed where the agreement (r^2) between predicted and observed values was greater than for any other methods at 0.96 in the training dataset, but dropped to 0.88 for the validation dataset, which is in the middle amongst other methods for the Norway validation dataset. According to the testing correlation, the best approaches for the Liperi dataset were the OLS and RF models with OLS_He and RF_All reaching $r^2 = 0.87$ and 0.86, respectively. Most approaches had comparable performance for the Norway dataset where the best performances were achieved for RF_All and RF_He, both having $r^2 = 0.89$.

Figures 5 and 6 show the RMSE% values during the heuristic variable selection progresses for the Liperi and Norway datasets, respectively. As a general pattern, the error values for the training datasets (black) decrease monotonically with the number of iterations. However, RMSE% values for the testing datasets (grey) are erratic and do not necessarily improve with number of iterations, which suggests that heuristic variable selection may not be suitable for some modelling approaches.

Despite the kNN_RF having the smallest final RMSE% for the training dataset for both Liperi and Norway, it had the largest differences between testing and training errors. The pattern of kNN performances for test datasets was highly erratic relative to their testing error progress and did not show the desired systematic improvement. OLS_He and RF_He showed systematic improvements in testing error as a result of each iteration of this variable selection and had a less erratic progress than kNN-based models. These two approaches also had the smallest final RMSE% values for the testing datasets. The differences between the initial and final RMSE% values were larger for the Liperi dataset than for the Norway dataset. The latter also had fewer steep decays. The exact patterns in Figures 5 and 6 vary each time that models are trained because the heuristic variable selection approach is stochastic in nature.

Discussion

This work demonstrates similarities and differences in ABA volume prediction performances between OLS, kNN and RF methods for datasets collected around the world (northern Europe, North America and South America). While the results differ among

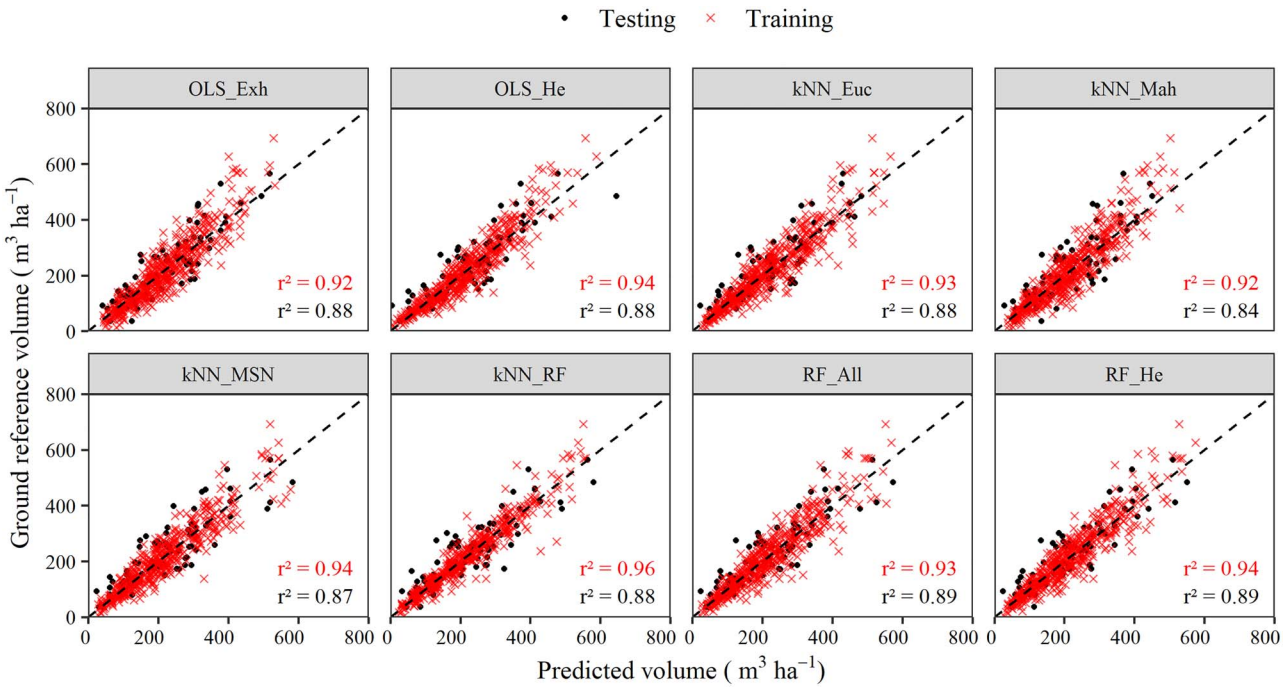

Figure 4 Observed and predicted values for testing and training datasets from Norway.

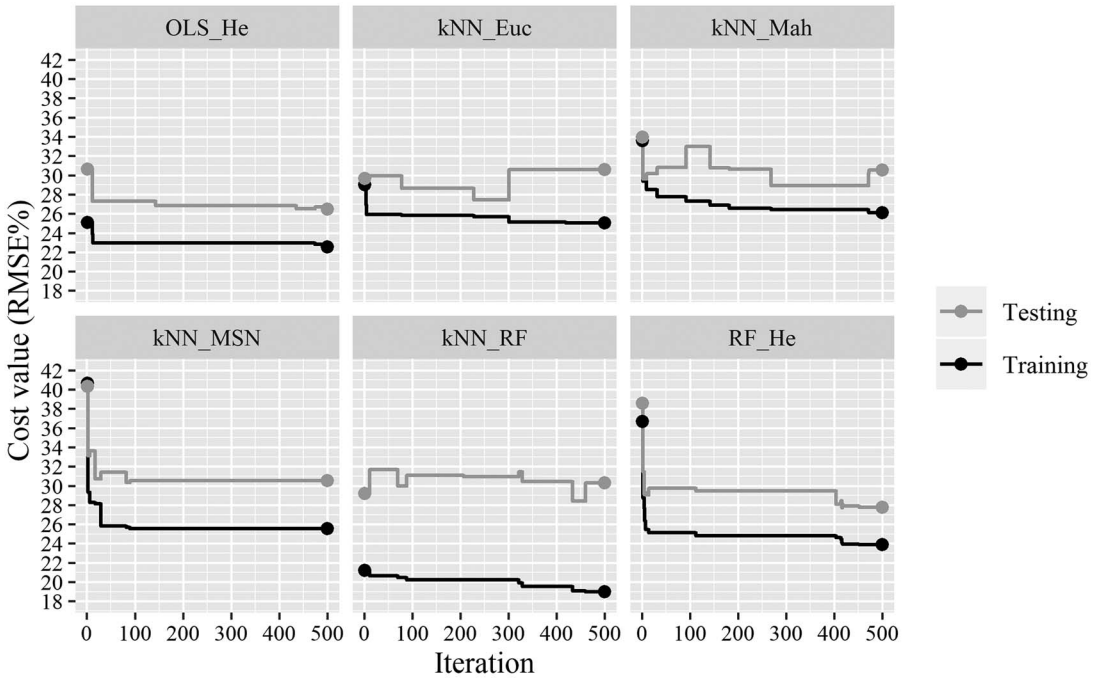

Figure 5 Optimization progress for the variable selection during the model training with the Liperi dataset. The errors were computed with the temporary solutions applied to the training and testing datasets.

datasets, it was possible to distinguish some patterns, which should prove helpful to researchers and practitioners.

Our first inference is that kNN methods were generally systematically less precise than OLS and RF. Better performance for RF relative to kNN has also been demonstrated in previous studies (Fassnacht et al., 2014; Latifi and Koch, 2012; Tompalski et al., 2019). Possible explanations are that kNN has many parameters, which need tuning, that kNN is more sensitive to correct tuning,

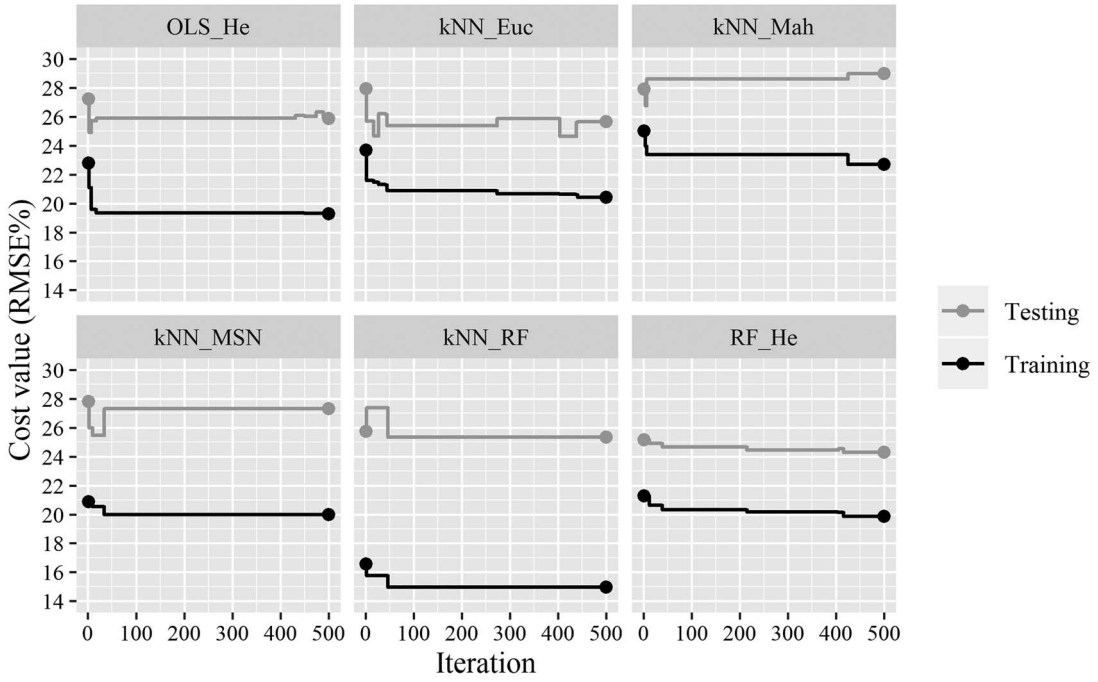

Figure 6 Optimization progress for the variable selection during the model training with the Norway dataset. The errors were computed with the temporary solutions applied to the training and testing datasets.

and that the default kNN tuning parameters are less generic than those for RF. RF default parameters, in contrast, seem to work well across a wide variety of datasets. For example, Shataee (2013) demonstrated that exhaustive tuning of kNN parameters led to smaller errors for a Euclidean-based kNN than for an RF model. McRoberts et al. (2017) showed that kNN performance improves greatly through parameter tuning, which also reduced the variation in performances among kNN distance metrics. In contrast, studies which tune RF parameters are less common in the ALS literature (Shataee, 2013; Shataee et al., 2011; Shi et al., 2018), and the results with respect to specific tuning parameters and their effects on performances are still unclear. Therefore, it is common to use the RF with the parameters set to default, since they appear to produce sufficiently robust models (Belgiu and Drăgu, 2016). However, it is important to highlight that the parameter tuning by itself cannot guarantee a reasonable performance since many factors influence a model's accuracy, such as the inherent properties of the dataset (Fassnacht et al., 2014). Besides, the full parameter tuning is time-consuming and considerably increases the computational complexity of the analysis as well as the required technical capacity of the analyst.

Establishing a comparison between our results and the ones available in existing literature was not a straightforward task since all of the studies had procedural differences in their implementation. Our results concerning the performance of kNN distance metrics disagree with some other studies. We found that kNN_Euc performed best under cross-validation, while kNN_RF was among the worst and with a clear indication of overfitting. Other studies found that kNN_Euc was outperformed by other techniques like kNN_MSN and, especially, kNN_RF (Hudak et al., 2008; Latifi et al., 2010; McRoberts et al., 2017). Both of them were considered the most robust distance metrics in some studies (Packalén et al., 2012; Shataee et al., 2011; Vauhkonen et al., 2010).

The first reason for this difference in findings may be related to differences in the model performance assessment. Earlier studies often applied leave-one-out cross-validation (LOOCV) (Latifi et al., 2010; Packalén et al., 2012; McRoberts et al., 2017), which is a specific type of cross-validation in which a single observation is omitted each time. The use of a 5-fold cross-validation, in contrast, may provide a more robust set of validation data for detection of model overfitting. A related factor is that more complex models have a greater tendency to overfit the data (Hastie et al., 2009). Since kNN_MSN and kNN_RF are more complex than the kNN_Euc, it may be that studies using LOOCV reported performances resulting from overfit models. Lastly, kNN_MSN models are usually applied for simultaneous prediction of several response variables, such as species-specific volume prediction (Packalén and Maltamo, 2007; Vauhkonen et al., 2010). In such scenario, the kNN approaches would probably perform better, as the predictions are guaranteed to remain logical.

Another inference from our results relates to practical considerations of heuristic and exhaustive variable selection procedures for OLS. Our results showed that OLS models for both heuristic and exhaustive variable selection have similar performances, but they differ in some respects. A heuristic approach to variable selection can handle a much greater number of explanatory variables, and is more flexible with respect to model formulations including handling non-linear forms, and can be easily optimized

for a wide range of loss functions. On the other hand, the stochastic process involved in the simulated annealing caused higher variation in the OLS_He errors. The exhaustive variable selection, although having smaller error ranges, quickly becomes computationally infeasible for large numbers of explanatory variables. The computational cost of training OLS_Exh increases factorially with the number of explanatory variables and with the number of variables to search for, and may hence be infeasible for high dimensional datasets.

Among RF methods, RF_He had comparable results to RF_All in most situations. This confirms that, to a degree, RF is robust to overfit when dealing with high dimensional data (Belgiu and Drăgu, 2016). Our findings suggest that variable selection may not be needed for some applications of RF.

The cross-validation and validation with independent data are traditionally focused on the performance of a specific model, where a model with the same predictors is trained and used to predict values to the holdout folds or testing dataset. Differently, our assessment was focused on the modelling approach itself wherein the variable selection was also repeated before predicting the next holdout fold or testing dataset. The fact that these processes were repeated 100 times increased the robustness of inferences regarding the comparisons of modelling approaches and their efficiency. This methodology is therefore strongly recommended for future works that intend to test or compare modelling approaches, even if it is more computationally intensive.

This study demonstrates the importance of validation, especially independent validation. Even with a fairly intensive cross-validation scheme that uses 20 percent of the data for holdout validation, the results suggest that our approach still overestimates model performances. Part of this difference may be the result of differences in the conditions as the training and validation sites, but the divergence in reported performances deserves further investigation. While the use of an independent validation dataset is the ideal way to assess model performance, it is rarely available in practice (Araujo et al., 2005). Despite the divergence between cross-validation and independent validation, our study suggests that the cross-validation can generally still provide information on the relative performances amongst methods, although we were not able to identify overfit by the kNN_RF method. Other assessment approaches which can provide reliable errors when validation datasets are not available include block cross-validation (see Roberts et al., 2017) or leaving out spatial clumps of observations, instead of choosing holdout data at random from across the study areas (Kotivuori et al., 2016, 2018). Despite that, Roberts et al. (2017) showed that a cross-validation with random folds can provide fair error estimates when a model is developed only to predict to a dataset with the same structure, which is a situation that fits to our assessment.

Our findings agree with the results of Fassnacht et al. (2014), who found that the accuracy of ABA-based predictions is influenced by the modelling approach. Therefore, our results are of great importance to remote sensing experts that need to construct accurate ABA maps of the growing stock volume or aboveground biomass. The main results were consistent regardless of variations in ALS metrics computed for the different study sites, as the metrics still represented similar variations in vertical and horizontal canopy structure despite differences in computational methods. The differences between methods are considerably smaller than the differences between datasets, but we showed that they are systematic, and some methods are slightly better in ABA applications than others. For example, the RF_All with default parameters is robust across multiple forest types and can be reliably applied to high-dimensional datasets without time-consuming variable selection processes. OLS methods can provide similar accuracy and might be preferred over RF_all in situations where it is important to interpret how individual predictors influence the predicted values. We did not test the effects of number of training observations in this study, but it is known that OLS tends to be more reliable for small data sets (e.g. $<$100 observations). Finally, we showed that kNN methods are suboptimal in situations where there is only a single response variable.

Conclusion

This study compared OLS, kNN and RF modelling strategies in combination with multiple variable selection strategies for the prediction of growing stock volume in five forests around the globe. RF and OLS models performed better than kNN models, although the maximum difference was about 4 percentage points, and generally less. These results suggest that RF or OLS are preferable for estimating volume from ALS data. In addition, kNN based on RF distance was especially problematic due to its strong tendency to overfit, which was not detected with a 5-fold cross-validation, but became clear during the validation with an independent test data. kNN based on Euclidean metrics outperformed other kNN approaches across all study areas.

In the case of the OLS methods, the heuristic variable selection was shown to perform as well as the more time-consuming exhaustive search, which is especially advantageous in the case of high-dimensional data where the exhaustive search algorithm may not be feasible. We did not observe an improvement to RF from using variable selection, which suggest that RF is a robust approach for practical applications.

Acknowledgements

We thank the Portuguese Science Foundation (*Fundação para Ciência e Tecnologia I.P.*–FCT) for funding the research activities of Diogo N. Cosenza; the University of Eastern Finland for supporting Diogo N. Cosenza during the development of this project; the USDA Forest Service–Savannah River for sharing field measurements and ALS data for the Savannah River Site in South Carolina, USA funded through the United States Department of Energy–Savannah River Operations Office under Interagency Agreement DE-EM0003622 and the USDA Forest Service Pacific Northwest Research Station. We thank Eetu Kotivuori and Mikko Kukkonen for providing the Kolari and Liperi datasets, and the anonymous reviewers that helped improving this text.

Conflict of interest statement

None declared.

Funding

This research was funded by the Forest Research Centre, a research unit funded by Fundação para a Ciência e a Tecnologia I.P. (FCT), Portugal (grant number UIDB/00239/2020). The research activities of Diogo N. Cosenza were supported by FCT (grant number PD/BD/128489/2017).

References

Araujo, M., Pearson, R., Thuiller, W. and Erhard, M. 2005 Validation of species-climate impact models under climate change. *Glob. Chang. Biol.* **11**, 1504–1513. doi: 10.1111/j.1365-2486.2005.001000.x.

Belgiu, M. and Drăgu, L. 2016 Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31. doi: 10.1016/j.isprsjprs.2016.01.011.

Bellman, R. 1961 *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. 1999 When is "nearest neighbor" meaningful? In *Lecture Notes in Computer Science*. G., Goos, J., Hartmanis, J., van Leeuwen (eds.). Springer-Verlag, pp. 217–235.

Blum, A.L. and Langley, P. 1997 Selection of relevant features and examples in machine learning. *Artif. Intell.* **97**, 245–271. doi: 10.1016/S0004-3702(97)00063-5.

Breidenbach, J., Næsset, E., Lien, V., Gobakken, T. and Solberg, S. 2010 Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. *Remote Sens. Environ.* **114**, 911–924. doi: 10.1016/j.rse.2009.12.004.

Breiman, L. 2001 Random forests. *Mach. Learn.* **45**, 5–32.

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T. *et al.* 2016 A meta-analysis and review of the literature on the k-nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sens. Environ.* **176**, 282–294. doi: 10.1016/j.rse.2016.02.001.

Crookston, N.L. and Finley, A.O. 2015 yaImpute: an R package for kNN imputation. *J. Stat. Softw.* **23**, 16. doi: 10.18637/jss.v023.i10.

Dudani, S.A. 1976 The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **4**, 325–327. doi: 10.1109/TSMC.1976.5408784.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P. *et al.* 2014 Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens. Environ.* **154**, 102–114. doi: 10.1016/j.rse.2014.07.028.

García-Gutiérrez, J., Martínez-Álvarez, F., Troncoso, A. and Riquelme, J.C. 2015 A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. *Neurocomputing* **167**, 24–31. doi: 10.1016/j.neucom.2014.09.091.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. 2010 Variable selection using random forests. *Pattern Recognit. Lett.* **31**, 2225–2236. doi: 10.1016/j.patrec.2010.03.014.

Gleason, C.J. and Im, J. 2012 Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sens. Environ.* **125**, 80–91. doi: 10.1016/j.rse.2012.07.006.

Gobakken, T., Korhonen, L. and Næsset, E. 2013 Laser-assisted selection of field plots for an area-based forest inventory. *Silva Fenn.* **47**, 1–20. doi: 10.14214/sf.943.

Görgens, E.B., Montaghi, A. and Rodriguez, L.C.E. 2015 A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. *Comput. Electron. Agric.* **116**, 221–227. doi: 10.1016/j.compag.2015.07.004.

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. doi: 10.1093/forestscience/54.6.597.

Hastie, T., Tibshirani, R. and Friedman, J. 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer.

Haykin, S.O. 2009 *Neural Networks and Learning Machines*. 3rd edn. Pearson education, Inc.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E. and Falkowski, M.J. 2008 Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sens. Environ.* **112**, 2232–2245. doi: 10.1016/j.rse.2007.10.009.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. 1983 Optimization by simulated annealing. *Science (80-.).* **220**, 671–680. doi: 10.1126/science.220.4598.671.

Kotivuori, E., Korhonen, L. and Packalen, P. 2016 Nationwide airborne laser scanning based models for volume, biomass and dominant height in Finland. *Silva Fenn.* **50**, 1–28. doi: 10.14214/sf.1567.

Kotivuori, E., Maltamo, M., Korhonen, L. and Packalen, P. 2018 Calibration of nationwide airborne laser scanning based stem volume models. *Remote Sens. Environ.* **210**, 179–192. doi: 10.1016/j.rse.2018.02.069.

Kukkonen, M., Maltamo, M., Korhonen, L. and Packalen, P. 2019 Multispectral airborne LiDAR data in the prediction of boreal tree species composition. *IEEE Trans. Geosci. Remote Sens.* **57**, 3462–3471. doi: 10.1109/TGRS.2018.2885057.

Latifi, H. and Koch, B. 2012 Evaluation of most similar neighbour and random forest methods for imputing forest inventory variables using data from target and auxiliary stands. *Int. J. Remote Sens.* **33**, 6668–6694. doi: 10.1080/01431161.2012.693969.

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**, 395–407. doi: 10.1093/forestry/cpq022.

Lawrence, R.L., Wood, S.D. and Sheley, R.L. 2006 Mapping invasive plants using hyperspectral imagery and Breiman cutler classifications (random Forest). *Remote Sens. Environ.* **100**, 356–362. doi: 10.1016/j.rse.2005.10.014.

Liaw, A. and Wiener, M. 2002 Classification and regression by random Forest. *R news* **2**, 18–22. doi: 10.1177/154405910408300516.

Lin, Y. and Jeon, Y. 2002 *Random Forests and Adaptive Nearest Neighbors, Technical Report No. 1055*. Madison.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A. and Kangas, J. 2006 Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Can. J. For. Res.* **36**, 426–436. doi: 10.1139/x05-246.

Maltamo, M., Peuhkurinen, J., Malinen, J., Vauhkonen, J., Packalén, P. and Tokola, T. 2009 Predicting tree attributes and quality characteristics of scots pine using airborne laser scanning data. *Silva Fenn.* **43**, 507–521. doi: 10.14214/sf.203.

McRoberts, R.E., Chen, Q., Domke, G.M., Næsset, E., Gobakken, T., Chirici, G. *et al.* 2017 Optimizing nearest neighbour configurations for airborne laser scanning-assisted estimation of forest volume and biomass. *Forestry* **90**, 99–111. doi: 10.1093/forestry/cpw035.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2015 Optimizing the k-nearest neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. *Remote Sens. Environ.* **163**, 13–22. doi: 10.1016/j.rse.2015.02.026.

Moeur, M. and Stage, A.R. 1995 Most similar neighbor: An improved sampling inference procedure for natural resource planning. *For. Sci.* **41**, 337–359.

Myers, R.H. 1989 *Classical and Modern Regression With Applications*. Duxbury Press.

Næsset, E. 2004 Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. *Scand. J. For. Res.* **19**, 164–179. doi: 10.1080/02827580310019257.

Næsset, E. 2002 Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* **80**, 88–99. doi: 10.1016/S0034-4257(01)00290-5.

Packalén, P., Heinonen, T., Pukkala, T., Vauhkonen, J. and Maltamo, M. 2011a Dynamic treatment units in eucalyptus plantation. *For. Sci.* **57**, 416–426.

Packalén, P. and Maltamo, M. 2007 The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sens. Environ.* **109**, 328–341. doi: 10.1016/j.rse.2007.01.005.

Packalén, P. and Maltamo, M. 2006 Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. *For. Sci.* **52**, 611–622. doi: 10.109352.6.611.

Packalén, P., Mehtätalo, L. and Maltamo, M. 2011b ALS-based estimation of plot volume and site index in a eucalyptus plantation with a nonlinear mixed-effect model that accounts for the clone effect. *Ann. For. Sci.* **68**, 1085–1092. doi: 10.1007/s13595-011-0124-9.

Packalén, P., Temesgen, H. and Maltamo, M. 2012 Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Can. J. Remote Sens.* **38**, 557–569. doi: 10.5589/m12-046.

Pascual, A., Bravo, F. and Ordoñez, C. 2019 Assessing the robustness of variable selection methods when accounting for co-registration errors in the estimation of forest biophysical and ecological attributes. *Ecol. Model.* **403**, 11–19. doi: 10.1016/j.ecolmodel.2019.04.018.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. [WWW Document]. https://www.r-project.org/ (accessed on 1 October, 20).

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G. *et al*. 2017 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography (Cop.)*. **40**, 913–929. doi: 10.1111/ecog.02881.

Saeys, Y., Inza, I. and Larrañaga, P. 2007 A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517. doi: 10.1093/bioinformatics/btm344.

Segal, M.R. 2004 *Machine Learning Benchmarks and Random Forest Regression*. Center for Bioinformatics & Molecular Biostatistic.

Shataee, S. 2013 Forest attributes estimation using aerial laser scanner and TM data. *For. Syst.* **22**, 484. doi: 10.5424/fs/2013223-03874.

Shataee, S., Weinaker, H. and Babanejad, M. 2011 Plot-level forest volume estimation using airborne laser scanner and TM data, comparison of boosting and random forest tree regression algorithms. *Procedia Environ. Sci.* **7**, 68–73. doi: 10.1016/j.proenv.2011.07.013.

Shi, Y., Wang, T., Skidmore, A.K. and Heurich, M. 2018 Important LiDAR metrics for discriminating forest tree species in Central Europe. *ISPRS J. Photogramm. Remote Sens.* **137**, 163–174. doi: 10.1016/j.isprsjprs.2018.02.002.

Silva, C., Klauberg, C., Hudak, A., Vierling, L., Jaafar, W., Mohan, M. *et al*. 2017 Predicting stem total and assortment volumes in an industrial Pinus taeda L. forest plantation using airborne laser scanning data and random forest. *Forests* **8**, 254. doi: 10.3390/f8070254.

Strunk, J., Gould, P., Packalen, P., Poudel, K., Andersen, H.-E. and Temesgen, H. 2017 An examination of diameter density prediction with k-NN and airborne lidar. *Forests* **8**, 444. doi: 10.3390/f8110444.

Talbi, E.-G. 2009 *Metaheuristics: From Design to Implementation*. John Wiley & Sons.

Tompalski, P., White, J.C., Coops, N.C. and Wulder, M.A. 2019 Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. *Remote Sens. Environ.* **227**, 110–124. doi: 10.1016/j.rse.2019.04.006.

Vauhkonen, J., Korpela, I., Maltamo, M. and Tokola, T. 2010 Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. *Remote Sens. Environ.* **114**, 1263–1276. doi: 10.1016/j.rse.2010.01.016.

Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M. and Viitala, R. 2011 Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS J. Photogramm. Remote Sens.* **66**, 28–37. doi: 10.1016/j.isprsjprs.2010.08.003.

A. Appendix 1

A list of the ALS metrics by data set. The abbreviations consist of a prefix that specifies the metric type, a statistic indicator and a suffix showing the return type. The prefixes are the height (*H*), intensity (*I*), percentage of below fractional height bin (*B*) and return percentage above fixed height (*D*). The statistic indicators for height and intensity metrics are the minimum (*min*), mean (*mean*), maximum (*max*), standard deviation (*std*), median (*med*), skewness (*skew*), kurtosis (*kurt*), variance (*var*), mode (*mode*), coefficient of variation (*cv*), interquartile distance (*iqd*), average absolute deviation (*aad*), the n-th L-moments (mom) and the number *x* related to the *x*-th height percentile. The subscripted suffixes indicate that the metric is computed from first and single return (*F*), or last and single returns (*L*), while no suffix refers to all returns.

Dataset	No. of metrics	Metrics
Brazil	18	$H10_F$ $H30_F$ $H50_F$ $H70_F$ $H90_F$ $H95_F$ $Hmean_F$ $Hstd_F$ $D10_F$ $D30_F$ $H10_L$ $H30_L$ $H50_L$ $H70_L$ $H90_L$ $H95_L$ $Hmean_L$ $Hstd_L$
Kolari	84	$D1_F$ $D2_F$ $D3_F$ $D4_F$ $D5_F$ $D6_F$ $D7_F$ $D8_F$ $D9_F$ $D10_F$ $D11_F$ $D12_F$ $D13_F$ $D14_F$ $D15_F$ $D16_F$ $D17_F$ $D18_F$ $D19_F$ $D20_F$ $D21_F$ $D22_F$ $D23_F$ $D24_F$ $H5_F$ $H10_F$ $H15_F$ $H20_F$ $H25_F$ $H30_F$ $H35_F$ $H40_F$ $H45_F$ $H50_F$ $H55_F$ $H60_F$ $H65_F$ $H70_F$ $H75_F$ $H80_F$ $H85_F$ $H90_F$ $H95_F$ $H99_F$ $Hmax_F$ $Hmean_F$ $Hstd_F$ $D1_L$ $D2_L$ $D3_L$ $D4_L$ $D5_L$ $D6_L$ $D7_L$ $D8_L$ $D9_L$ $D10_L$ $D11_L$ $D12_L$ $D13_L$ $D14_L$ $D15_L$ $D16_L$ $D17_L$ $D18_L$ $D19_L$ $D20_L$ $D21_L$ $D22_L$ $D23_L$ $H50_L$ $H55_L$ $H60_L$ $H65_L$ $H70_L$ $H75_L$ $H80_L$ $H85_L$ $H90_L$ $H95_L$ $H99_L$ $Hmax_L$ $Hmean_L$ $Hstd_L$
Liperi	58	$Hmax_F$ $Hmin_F$ $Hstd_F$ $Hmed_F$ $Hmean_F$ $Hskew_F$ $Hkurt_F$ $Imax_F$ $Istd_F$ $Imed_F$ $Imean_F$ $Iskew_F$ $Ikurt_F$ $H10_F$ $H30_F$ $H50_F$ $H70_F$ $H90_F$ $I10_F$ $I30_F$ $I50_F$ $I70_F$ $I90_F$ $D0.5_F$ $D2_F$ $D5_F$ $D10_F$ $D15_F$ $D20_F$ $Hmax_L$ $Hmin_L$ $Hstd_L$ $Hmed_L$ $Hmean_L$ $Hskew_L$ $Hkurt_L$ $Imax_L$ $Imin_L$ $Istd_L$ $Imed_L$ $Imean_L$ $Iskew_L$ $Ikurt_L$ $H10_L$ $H30_L$ $H50_L$ $H70_L$ $H90_L$ $I10_L$ $I30_L$ $I50_L$ $I70_L$ $I90_L$ $D0.5_L$ $D2_L$ $D5_L$ $D10_L$ $D15_L$
Norway	23	Hcv_F $Hmax_F$ $Hmean_F$ $H0_F$ $H10_F$ $H20_F$ $H30_F$ $H40_F$ $H50_F$ $H60_F$ $H70_F$ $H80_F$ $H90_F$ $B0_F$ $B10_F$ $B20_F$ $B30_F$ $B40_F$ $B50_F$ $B60_F$ $B70_F$ $B80_F$ $B90_F$
USA	53	*Hmax Hmean Hmode Hstd Hvar Hcv Hiqd Hsquew Hkurt Haad Hmom1 Hmom2 Hmom3 Hmom4 H1 H5 H10 H20 H25 H30 H40 H50 H60 H70 H75 H80 H90 H95 H99 B1 B5 B10 B20 B25 B30 B40 B50 B60 B70* Others: *Percentage of first returns above 1.5 m* *Percentage of all returns above 1.5 m* *Ration between all returns above 1.5 m and the total of first returns* *Number of first returns above 1.5 m* *Number of all returns above 1.5 m* *Percentage of first return above Hmean* *Percentage of first return above Hmode* *Percentage of all return above Hmean* *Percentage of all return above Hmode* *Ratio between all returns above Hmean and the total of first returns* *Ratio between all returns above Hmode and the total of first returns* *Ratio between first returns and all returns* *Ratio between first and all returns* *Ratio between second and all returns*