International Journal of Digital Earth,
Vol. 5, No. 2, March 2012, 106–132

Modelling stratified forest attributes using optical/LiDAR features in a central European landscape

Hooman Latifi[a]*, Arne Nothdurft[b], Christoph Straub[a] and Barbara Koch[a]

[a]*Department of Remote Sensing and Landscape Information Systems, University of Freiburg, Tennenbacherstraße, Freiburg, Germany;* [b]*Department of Biometry and Informatics, Forest Research Institute Baden Württemberg, Wonnhaldestraße, Freiburg, Germany*

(*Received 29 September 2010; final version received 21 April 2011*)

Improvements in the acquisition of three-dimensional (3D) information from the Airborne Laser Scanner (ALS) increase its applications for studying Earth's surface. The use of ALS data in natural resource inventories is still in an experimental stage in central Europe. Here, a survey was completed in Germany, where plot-level features from LANDSAT Thematic Mapper and ALS data were applied. An automated process was developed for forest stratification using orthoimages. A genetic algorithm was applied for variable screening. Variable subsets of different sizes were employed for simultaneous predictions of structural forest attributes using the 'Random Forest' (RF) method. Performance was assessed by leave-one-out cross-validations on bootstrap resample data. Results indicate that the stratification of forest notably improved the results of predictions. The improvements were more obvious for the strata-related attributes. Accuracy was enhanced as the number of selected variables increased. However, parsimonious models are still essentially required for practical applications. The RF errors were slightly greater than those from least squares regression, as the non-parametric methods do not share the same mix of error components as regression. Through the combination of remote sensing and modelling, we conclude that our results are helpful for bridging the gap between regional earth observation and on-the-ground forest structure.

Keywords: spatial prediction; natural resource inventory; ALS; Digital Earth; data fusion; remote sensing

1. Introduction

The Digital Earth is an integral part of advanced technologies including earth observation (Anon 2009), and can play a sustainable role in addressing challenges such as environmental and natural resource depletion. To this aim, compiling information on biophysical characteristics of natural resources is a vital task. Improving the understanding of interactions between environmental phenomena at different levels, including the local level, will foster the provision of reliable, accurate, and openly accessible information at the relevant scale for practical application. Geosensors are defined as satellite and airborne devices receiving geographically referenced environmental stimuli (Craglia *et al.* 2008). Information derived from high resolution geosensors has received prominent attention as multi-temporal and

*Corresponding author. Email: hooman.latifi@felis.uni-freiburg.de

ISSN 1753-8947 print/ISSN 1753-8955 online

http://dx.doi.org/10.1080/17538947.2011.583992
http://www.tandfonline.com

cost-effective tools for small and regional scale natural resource inventory during the last decade. As previously studied by e.g. Gruen (2008) and Estronell *et al.* (2010), 3D information derived from Airborne Laser Scanning (ALS) remote sensing is practical for representing and analysing Earth in a digital form. Since the late 1990s, ALS data have been extensively used as auxiliary information for predictions of attributes on forest stand and single tree levels where they have been applied for the characterisation of highly variable forest canopy structures (Hudak *et al.* 2008, Hyyppä *et al.* 2009, Koch *et al.* 2009).

In the Federal state of Baden-Württemberg in southwest Germany, the forest inventory is designed based on concentric fixed radius plots in a regular sampling grid. The stratification of forest stands into forest types is accomplished during field surveys for mensurational descriptions. For each of the inventory sample plots, the information on forest type is derived from the corresponding forest stand to which the sample plot belongs. However, the personal expertise of the forestry technician plays a major role in the stratification. Due to the relatively high number of dominant species in forest stands, this process is often a challenging and time-consuming task in ground surveys. Therefore, a straightforward automated method of forest type stratification would provide valuable assistance in the current data processing workflow. The use of remote sensing data for stratification of forest areas has a long tradition e.g. utilising aerial photography or satellite imagery (Koch and Dees 2008), where the aim is to subdivide the heterogeneous forest area into a set of relatively homogeneous regions according to specific forest characteristics. The near infrared spectral information has been determined to be helpful for stratifying forest areas into deciduous and coniferous forest covers (e.g. Stibig *et al.* 2004, Straub *et al.* 2008), as well as for the delineation of various vegetation cover types (e.g. Haapanen *et al.* 2004).

Amongst the preceding studies, Popescu and Wynne (2004) used leaf-off multispectral Airborne Terrestrial Land Applications Scanner (ATLAS) image to differentiate amongst bare earth, deciduous and coniferous trees using a Maximum Likelihood (ML) supervised approach. As reported by an error matrix, they achieved an overall accuracy of 82.87%. Tiede *et al.* (2006) used multispectral line scanner data with 0.5 m spatial resolution in combination with ALS data to segment and classify coniferous, deciduous and dead trees using an object-based segmentation technique. They report an overall accuracy of 75%. Most recently, Kim *et al.* (2011) classified high-resolution IKONOS images and analysed different classification methods (including pixel-based ML and object-based methods) to generate a detailed forest cover map. Seven different forest cover types in addition to grass land and other non-forest areas were differentiated and overall accuracies between 63.9 and 76.8% were consequently achieved featuring better accuracies for segment-based classification.

In this study, we have applied an automated procedure for the stratification of forest cover, which is based on spectral information derived from colour infrared (CIR) orthoimages (Straub *et al.* 2009). Stratification is used to enhance non-parametric models for predictions of important forestry variables on a sample plot level. So far, various remote sensing data sources have been applied to predict stand or plot-scale forest characteristics by parametric and non-parametric models (e.g. Hudak *et al.* 2006, Packalén and Maltamo 2007, Straub *et al.* 2009, McInerney *et al.* 2010). Parametric models generally come with strong assumptions of distributions for the parameters and the variables, which sometimes may not be met by the data. In contrast, non-parametric methods may enable more flexibility for using the unknown regression relationships (Härdle 1990). Yet, they often require larger sample sizes than parametric counterparts, as the underlying data also serves as a model. Nearest Neighbour (NN) imputation methods select units from the reference set to serve as surrogates for members of the target sets using a measure of similarity based on predictor variables (Stage and Crookston 2007). The methods have been used for modelling stand characteristics, and have shown promising results in landscapes of Scandinavia (Packalén and Maltamo 2006, 2007), the United States (Hudak *et al.* 2008), and central Europe (Nothdurft *et al.* 2009, Latifi *et al.* 2010). Koch (2010) reviewed the state of the art in the application of ALS, optical and Synthetic Aperture Radar (SAR) data for forest biomass assessment, in which the importance of combined use of ALS and optical data for such purposes was highlighted. Amongst the most recent studies in the field, McInerney *et al.* (2010) combined airborne ALS and spaceborne Indian Remote Sensing multispectral data to model stand canopy height using the NN method. They reported ALS as the major means of canopy height retrieval, and achieved a relative root mean square error (*RMSE*) between 28 and 31%. In the current study, an NN model based on the Random Forest (RF) approach (Breiman 2001) was used. The RF method consists of regression and classification trees for resampled predictor variable sets, which has been shown by studies such as Hudak *et al.* (2008) and Latifi *et al.* (2010) to produce higher prediction accuracy compared to various other imputation methods including the Most Similar Neighbour (MSN) in which the weight matrix of the predictors is defined using their canonical correlations. Yet, Breidenbach *et al.* (2010) reported an approximately similar performance of RF over and MSN, as their study yielded the RMSE of 32.41% (for MSN) and 32.81% (for RF) when predicting the total standing timber volume by averaging eight NNs. Yu *et al.* (2011) built RF models to predict the timber volume using ALS data on a single-tree scale. They assessed the accuracy based on separate training/test sets and reported relatively poor results of 45.77% RMSE for test sets, which were roughly similar to those obtained from linear regression models. They assessed a significant portion of prediction errors to come from the remarkably less-abundant old stems across the study site.

The RF, like other multivariate models, allows for the simultaneous use of several predictor variables (e.g. multiple remote sensing features). Although adding more features in a regression model could theoretically improve its accuracy, there is always a common interest to build parsimonious models that are valid in other domains beyond the underlying region of parameterisation. This requirement is of particular importance in the presence of a high dimensional predictor set or when the number of samples is relatively low.

A method is used here to reduce high dimensional remote sensing data sets in order to select the optimal candidates for NN modelling. While deterministic methods such as stepwise selection show major shortcomings when applied to highly correlated variable sets, stochastic search algorithms have been reported to be more efficient, especially for complex multivariate models (Barros and Rutledge 1998, Tomppo and Halme 2004). Genetic Algorithm (GA) is an approach that functions by evolving sets of variables (chromosomes) to meet a certain optimisation criterion. Tomppo and Halme (2004) applied the GA method for selecting predictor variables among a set of optical features, in which the aggregation of standard error and bias of kNN predictions was defined as the fitness function to be minimised. Tomppo *et al.* (2009) extended the algorithm to be used for predicting categorical variables such as forest type. In both cases the GA was assessed to be an effective means of optimisation for pruning the weight vector of predictor variables, regardless of the fitness function used. Latifi *et al.* (2010) applied a GA on categorised response variables to select predictors amongst CIR orthoimages, Thematic Mapper (TM) and ALS-derived features. Although the study evaluated the GA to efficiently reduce the relative RMSE of total volume and biomass compared to stepwise selection of variables, the method was reported to produce unstable results in multiple runs. This was attributed to strong correlations amongst the high dimensional predictor variables. Here, by using an appropriate hypothesis test, GA is applied to continuous response variables to produce parsimonious variable subsets.

Furthermore, we examined whether a remote sensing-supported prior stratification of the sampling units into forest strata may lead to more accurate predictions of six common forestry attributes. Stratification was accomplished by means of an automated procedure using visible/near infrared domains of CIR orthoimages. A broad range of statistical metrics extracted from medium resolution TM imagery and normalised point cloud of ALS data were used for the analysis. The RF model performance was assessed by means of randomly bootstrapped leave-one-out cross-validation. Relative RMSE was used as the diagnostic tool. The unstratified variant was compared to the stratified variant and also to the results of multivariate linear regression models.

2. Material and methods

2.1 Materials

2.1.1. Study area

The study area is located to the north of Karlsruhe in the state of Baden-Württemberg, southwestern Germany (49° 03′ 37″ N and 8° 24′ 09″ E to 49° 01′ 1″N and 8° 25′ 49″ E in WGS84) and covers approximately 900 ha of managed forests (Figure 1). Stands are dominated by Scots Pine (*Pinus sylvestris* L.; 56.3%), European Beech (*Fagus sylvatica* L.; 17.8%), Sessile Oak (*Quercus petraea* Liebl.), and Pedunculate Oak (*Quercus robur* L.; jointly 14.9%). Other species play a minor role. The mean standing timber volume is 264 m^3 ha^{-1} that is less than the average standing volume in Baden-Württemberg (362 m^3 ha^{-1}). The stands have approximately 492 trees ha^{-1} featuring the mean basal area of 25 m^2 ha^{-1} (Table 1).

2.1.2 Forest inventory data

The reference data was collected in the summer of 2006 during a forest inventory based on 297 permanent circular sample plots arranged in a regular 100 × 200 m grid. Each plot consists of four concentric circles featuring 2, 3, 6 and 12 m of radii, on which various attributes were measured. Trees with diameter at breast height (DBH) $<$ 10 cm are measured if their distance to the plot centre is $\leq$ 2 m; for trees smaller than 15 cm DBH and for those smaller than 30 cm DBH, the maximum distances are 3 and 6 m respectively. Trees with DBH $\geq$ 30 cm are measured on the plot with 12 m radius. Two dominant heights of each main tree species and one

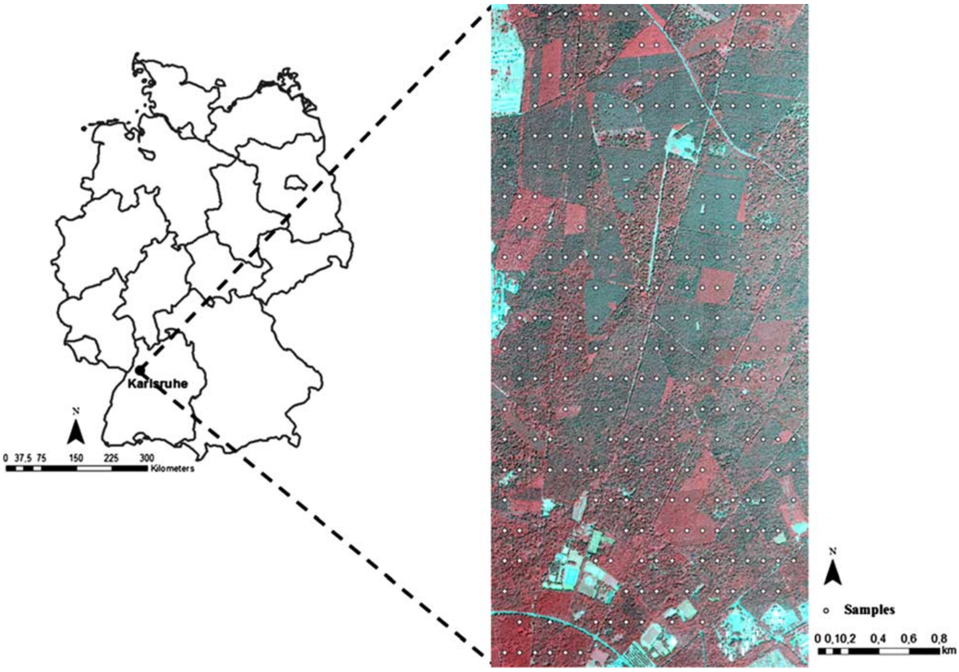

Figure 1. Geographical location of the study site in Baden-Württemberg state-Germany, overlaid by the ground sample plots.

dominant height per other mixed species were measured. According to common forest inventory practice, heights of trees without measurement were predicted by uniform stand height curves. The single tree timber volume was calculated using the taper functions of Kublin (2003), based on integrating tensor-product splines with regression coefficients for form factors. The plot-level total timber volume in m^3 ha^{-1} was derived by summing the single tree volumes weighted by the inverse of the corresponding sample plot area.

The single tree biomass was estimated with Zell (2008) parameters for the allometric equation:

$$B = \beta_0 d^{\beta_1} h^{\beta_2} \tag{1}$$

Table 1. Summary information on forest stands from the study site.

Species group	Deciduous/ coniferous	Timber volume [%]	Timber volume [m^3 ha^{-1}]	Basal area [m^2 ha^{-1}]	Trees ha^{-1}
Pinus sp.	coniferous	56.3	149.0	14.1	214
Fagus sp.	deciduous	17.8	47.2	4.6	110
Quercus sp.	deciduous	14.9	39.5	3.7	78
Other deciduous	deciduous	5.8	15.5	2.0	61
Picea sp./*Abies* sp.	coniferous	4.7	12.3	1.3	26
Larix sp.	coniferous	0.5	1.3	0.1	2
Total		100	264.8	25.7	492

where d is DBH, h is tree height, and β_0, β_1, β_2 are the tree-specific function parameters. The total biomass on each sample plot was derived by summing all single tree biomass estimates. The following attributes were selected to be modelled as response variables: total standing timber volume [m^3 ha^{-1}], volume of coniferous trees [m^3 ha^{-1}], total volume of trees featuring >40 cm DBH [m^3 ha^{-1}], stem number of trees featuring >25 cm DBH, total biomass [tons ha^{-1}], and biomass of coniferous trees [tons ha^{-1}]. It was assumed that the RF models based on the prior stratification would result in more accurate predictions for the forest strata-related target variables such as volume and biomass of the coniferous trees. In case of the stem number, stems with DBH >25 cm were considered as the first-pulse ALS metrics mostly characterise older stems in the overstory. Because information on the portion of mature wood is of major relevance for tree harvest planning, the volume of trees having DBH >40 cm was also modelled as a response variable.

2.1.3 Remote sensing data

Full wave ALS data was acquired in August 2007 by Toposys GmbH using the Harrier56 LiDAR system on a Helicopter and a Riegl LMS-Q560 laser scanner (Table 2). The laser scanning flight was completed twice to enable the collection of a high point density of 16 points m^{-2} over the study site. Furthermore, a large side lap of over 50% was used in scanning (Table 2). RiANALYZE 560 softwareVer. 5.1.10 (RIEGL 2010) was used by the data provider to extract the total number of echoes within each laser beam. Thus, all possible echoes from the laser beams were delivered as pulse-form data featuring a high density point cloud, out of which the first and last echoes were extracted as the most relevant echoes for the analysis.

The high resolution optical data in four spectral bands was recorded in July 2008 by Toposys GmbH using a RGB/NIR line scanner installed on a 'Falcon II' system (Table 2). The data covers an approximate range of 450 nm in the electromagnetic spectrum including visible (RGB) and near infrared (NIR) domains. The uncorrected individual strips were rectified and georeferenced by the data provider using an elevation model derived from ALS data. Finally, true colour (RGB) and CIR orthoimages data featuring 25 cm of Ground Sampling Distance was delivered. This data was used to automatically stratify coniferous and deciduous forest areas based on a two-dimensional feature space using the near infrared and green channels. In order to keep the estimation entirely independent from the stratification process, the CIR data was excluded from the underlying covariate data set of the NN models.

The medium resolution Landsat TM images were utilised to extract spectral input predictor variables to provide features of visible and near/mid-infrared domains. As previously reported by Latifi *et al.* (2010), spectral metrics from TM data could be used as surrogate for high resolution CIR images in plot scale predictions without causing a considerable reduction in prediction accuracy. Thus, one cloud-free satellite scene recorded by TM sensor in July 2006 was acquired for the study area. Neither geometric nor radiometric distortions were found in the data. A linear contrast enhancement was applied on the six spectral bands to expand narrow range of brightness values of the input image over a range of 256 gray values. The thermal band was excluded from the analysis due to the coarse

Table 2. Flight and technical parameters of LiDAR system 'Harrier 56,' RGB/NIR line scanner 'Falcon II', and Landsat-TM satellite scanner.

Harrier 56 LiDAR system	
Measurement rate	100 [kHz]
Field of view	45 [°]
Flying height	700 m above ground level
Flying speed	30 [m s^{-1}]
Point density	16 [points m^2]
Vertical/horizontal accuracy	<0.20 [m]/ <0.50 [m]
Average footprint size	0.35 [m]
Beam divergence	0.5 [mrad]
Beam deflection	Rotating polygon
Sensor dimensions	64 × 30 × 48 [cm]
RGB/NIR Falcon II line scanner	
Flying height	700 [m]
Spectral channels	B: 450–490 [nm] | G: 500–580 [nm] R: 580–660 [nm] NIR: 770–890 [nm]
Viewing angle	21.6 [°]
Line rate	Up to 330 [Hz]
Pixel per line	682
Ground sampling distance	0.25 [m]
TM satellite data	
Spacecraft	Landsat 5
Acquisition date	04.07.2006-Day
Sun elevation	59.42506336
Sun azimuth	140.94471286
Spatial resolution	30 [m]
Spectral channels	Band 1: 0.45–0.52 [nm] Band 2: 0.52–0.60 [nm] Band 3: 0.63–0.69 [nm] Band 4: 0.76–0.90 [nm] Band 5: 1.55–1.75 [nm] Band 7: 2.08–2.35 [nm]

spatial resolution. The technical parameters of the remote sensing data sets are listed in Table 2.

2.2 Methods

2.2.1. Automatic forest stratification using CIR data

An unsupervised technique with application of CIR data was developed for stratification of the study site into coniferous, deciduous and mixed strata. The main advantage of an unsupervised method is that no user-defined training data is required. As the final stratified units were offset to 1 m grid cells, the original CIR bands were first resampled into the bands featuring 1 m spatial resolution to speed up the image processing. Coniferous and deciduous trees in the overstory were segmented using the two-dimensional (2D) pixel classification technique described in Straub *et al.* (2009). The NIR and green channels were selected to define the feature space, as these channels revealed the main difference in the spectral reflectance of coniferous and deciduous trees. Clusters were delineated within the feature space using the following procedure.

First, cluster centres were extracted by searching for local maxima. To this aim, the feature space was iteratively smoothed until just two maxima were found. One maximum represented the centre of the 'coniferous cluster' and the other one the centre of the 'deciduous cluster.' The defined cluster centres were used as the initial starting values to delineate the cluster boundaries using a 'pouring algorithm' that considers the feature space as topographical surface (MVTec 2010). The algorithm functions similar to 'watershed segmentation,' a segmentation method based on greyscale mathematical morphology (Soille 1999). Starting from the maxima, regions were grown downwards (similar to water flowing downhill from the maxima in all directions) until 'valley bottoms' were reached. Then each pixel of the image domain was assigned to one of the clusters. Additionally, ground pixels were extracted using the height information from ALS. First, a forest mask was generated with the help of a fully automated procedure for the delineation and classification of forests and trees outside the forest. The method makes use of multiple echoes from ALS data. Thus, ground pixels were masked out. Coniferous and deciduous trees were classified within the forest mask. Details can be found in Straub *et al.* (2008).

Based on this pixel-wise classification, the inventory plots were assigned to one of the following strata: Coniferous forest (*C*), Deciduous forest (*D*) or Mixed forest (*M*). Thus, the proportion of coniferous P_C and deciduous trees P_D was computed within the inventory plot circle (size = 452.4 m^2). Afterwards, the plots were assigned to a stratum using the following definition:

$$\text{Stratum} = \begin{cases} C \text{ if } P_C \geq 70\% \\ D \text{ if } P_D \geq 70\% \\ M \text{ if } P_C \text{ and } P_D < 70\% \end{cases} \quad . \qquad (2)$$

Field measurements from the forest inventory were used to evaluate the accuracy of stratification by computing the percentage of coniferous and deciduous plots based

Table 3. Error matrix showing the comparison of the automatic classification with field measurements.

Field data	Classified data					
	C	*D*	*M*	Total	Producer's accuracy [%]	User's accuracy [%]
C	138	1	12	151	91.39	86.25
D	2	57	14	73	78.08	79.17
M	20	14	39	73	53.42	60.00
Total	160	72	65	297		
					Overall accuracy%: 78.97	

Note: C, Coniferous; D, Deciduous; M, Mixed.

on their proportion of the total timber volume (Table 3). Using this information, the plots were stratified again into coniferous (*C*) deciduous (*D*) and mixed (*M*) forest.

2.2.2. Processing of remote sensing data

The ALS point cloud was normalised using a Digital Terrain Model (DTM) to represent the height above ground. The DTM was created using the filtering algorithm adapted from Kraus and Pfeifer (1998) and implemented by McGaughey (2009), based on linear prediction implemented as an iterative process. After generating a surface (called prior surface) with equal weights for all ALS points, the terrain surface lies between the true ground and the canopy surface. Terrain points are more likely to lie below the surface and vegetation points above the surface. The distance and direction to the surface are used to compute weights for each ALS point using the following weight function:

$$\rho_\imath = \begin{cases} 1 & v_\imath \leq g \\ \frac{1}{1+(a(v_\imath - g)^b)} & gv_\imath \leq g + w \\ 0 & g + wv_\imath \end{cases} \tag{3}$$

where *a* and *b* determine the steepness of the weight function, and *g* determines which points are assigned a maximum weight of 1.0 (assigned to the points below the surface by more than *g*). The above ground offset parameter *w* is used to establish an upper limit of points, which results in assigning a weight of zero to the points above the level ($g+w$). The V_i is the residual value, which is the oriented distance from the prior surface to the measured point. Terrain points are more likely to have negative residuals, whereas the vegetation points are more likely to have small negative or positive residuals. Points with large negative residuals have maximum weights and attract the computed surface, whereas points with medium residuals have smaller weights and less influence on the computed surface. After eliminating the points in which $g+w < V_\text{i}$, the residuals for these measurements Z_i can be computed. If a residual is within the specified range, the corresponding measurement Z_i will be used again in the next iteration step. Following the iteration, the points satisfying the first two conditions of the weight function are considered bare-earth points (McGaughey 2009).

Since the ALS point cloud was acquired during summer time, both first/last pulse data appeared to be similar and thus correlated. The reason was that the majority of forest stands in the study site included a dense understory (mainly composed of deciduous species such as Beech), whereas the conifers occupied a greater proportion in the overstory. Based on results from numerous pre-analysis computations, we only applied the first pulse data to reduce the considerable amount of inter-correlations that existed within the enormous predictor variable set.

Various laser metrics were derived from the data at sample plot level. There was a high density of near-ground laser hits, which were excluded by only including points higher than 2 m (i.e. canopy hits) when calculating canopy metrics (see Packalén and Maltamo 2006). The height metrics consisted of measures of central tendency (i.e. measures locating a sample in the middle, around which the data is distributed), measures of dispersion (important for describing the variation of the data), and height percentiles and proportions that were computed directly from the normalised first pulse point cloud. Intensity values were also used to derive metrics because they have been reported to be correlated with forest attributes under consideration (e.g. Boyd and Hill 2007). Similar statistics (except proportions) were calculated for the ALS intensity values.

The six spectral bands of the TM data were used to derive mean gray values at sample plot level, where the sample plots were the average of cells in the underlying raster image calculated at each sample plot zone. In addition, Normalised Difference Vegetation Index (NDVI) and Infra Red/Red (IR/R) ratio were calculated from visible/near-infrared bands as they were already reported to be correlated with forest structure (Freitas *et al.* 2005) and forest biomass (Gonzalez-Alonso *et al.* 2006). Two main components of Principal Component Analysis (PCA) were derived from separate sets of visible and infrared bands that represented over 98% of the spectral variance of the respective dataset. The PCA has been reported as an appropriate tool for deriving low-dimensional subspaces that capture most of the information of the whole dataset (e.g. Cadima *et al.* 2004). The linear transformation of Tasseled Cap (Crist and Cicone 1984) was carried out on the entire TM data set, returning three components of brightness, greenness and wetness. Altogether 62 features were calculated using the remote sensing data sources.

2.2.3. Variable screening

When dealing with datasets with numerous variables, a frequent aim is to reduce the dimensionality of the dataset (Cadima *et al.* 2004). Though heuristics may be used to deal with highly correlated variable sets, the application of appropriate variable selection methods has also become an important issue. The main objective is to optimise the models efficiency by achieving a certain performance level with maximum degree of freedom. The performance of stepwise selection procedures has been criticised compared to the local search heuristics such as simulated annealing and GA (Barros and Rutledge 1998, Cadima and Jolliffe 2001, Cadima *et al.* 2004).

2.2.3.1 The genetic algorithm (GA). The GA is a variable search procedure based on the principle of evolution by natural selection. The procedure works by evolving sets of variables (called chromosomes) that fit specific criteria from an initial random population via cycles of differential replication, recombination and mutation of the fittest chromosomes. The whole cycle of replication, cross-over and mutation is called a generation. The procedure was introduced by Holland (1975) and is now adaptable to different optimisation scenarios including classification, regression and survival analysis (Trevino and Falciani 2006a, b). Mitchell (1996) is suggested for further reading on the theoretical basis of GA. The currently implemented GA procedure (Cerdeira *et al.* 2009) essentially consists of the following steps (Figure 2):

1. The procedure starts with the creation of a number of random variable sets (initial chromosomes), forming a population of chromosomes (niche). In each iteration, couples are formed as half the size of the population. The father is selected from the chromosomes with the probability of their value of the fitness criterion. The mother is selected from the chromosomes featuring

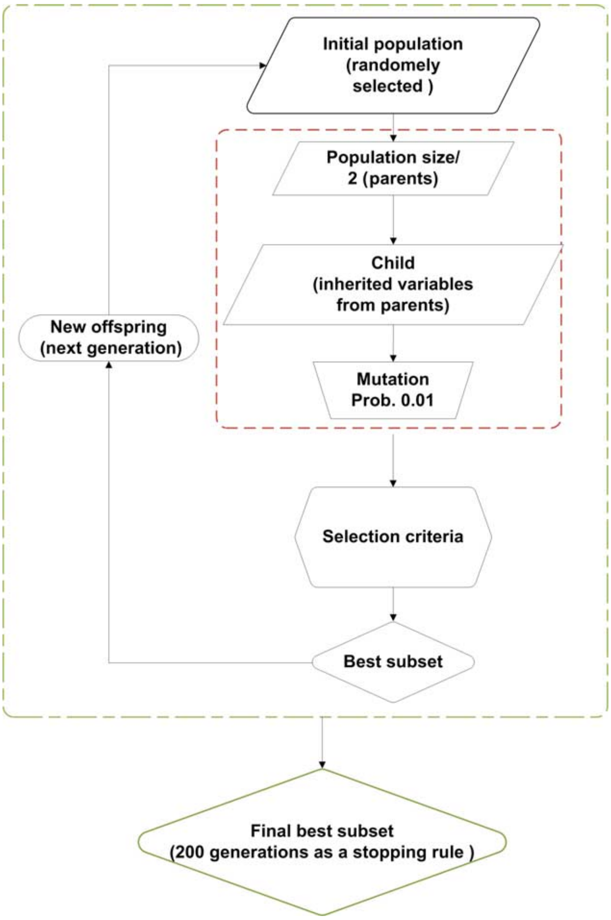
Initial population (randomely selected)
Population size/ 2 (parents)
Child (inherited variables from parents)
Mutation Prob. 0.01
New offspring (next generation)
Selection criteria
Best subset
Final best subset (200 generations as a stopping rule)

Figure 2. GA search method for current implementation.

the same probability but having at least two different variables as those of the father.

2. Those couples are mated to produce new chromosomes (called children), which then inherit the parent's properties. They consist of those variables from both parents, but also of the variables with equal probability from the symmetric difference of their parents. The number of generations (200 generations here) is used as a stopping rule.
3. Cross-over and mutations may occur randomly in the chromosome. Thus, new variables may be created to be used in chromosomes. The probability of each child undergoing a mutation was set to 0.01 here, to prevent the algorithm from replicating the same solution.
4. The procedure is repeated until an accurate chromosome is obtained.

Normally, a large number of clones (replicates) of a small number of variables tend to appear after a few generations. As a result, the algorithm may get stuck in a population in which the size of the symmetric difference of every pair is not greater than two. To prevent this, the algorithm counts the number of existing replicates of the new-born child amongst all the generations. The child is rejected and substituted by an allowable solution selected at random with a uniform distribution if the maximum number of clones surpasses one.

Based on the maximisation of the Tau squared index τ^2 as fitness criterion, the subsets of predictor variables are searched through GA. The maximisation of τ^2 is equivalent to the minimisation of the Wilks Lambda criterion Λ (see Duarte Silva 2001). Assuming the multivariate linear model

$$\begin{aligned}
&Y \quad = \quad X \quad B \quad + \quad E \\
&n \times q \quad \mathrm{n} \times (p+1)(p+1) \times q \;\; n \times q \\
&Y = (y_{.1}, \ldots, y_{.q}), B = (\beta_{.1}, \ldots, \beta_{.q}), E = (\in_{.1}, \ldots, \in_{.q}) \\
&Y = \begin{pmatrix} y_{11} \cdots y_{1q} \\ \vdots \qquad \vdots \\ y_{n1} \ldots y_{nq} \end{pmatrix}, X = \begin{pmatrix} 1 \; x_{11} \; \cdots \; x_{1p} \\ \vdots \qquad\qquad \vdots \\ 1 \; x_{n1} \; \cdots \; x_{np} \end{pmatrix}, \\
&B = \begin{pmatrix} \beta_{01} \cdots \beta_{0q} \\ \vdots \qquad \vdots \\ \beta_{p1} \ldots \beta_{pq} \end{pmatrix}, E = \begin{pmatrix} \in_{11} \; \cdots \; \in_{1q} \\ \vdots \qquad\qquad \vdots \\ 1 \; x_{n1} \; \cdots \; x_{np} \end{pmatrix},
\end{aligned} \tag{4}$$

with $E(\in_{i.}) = 0$, $Cov(\in_{i.}) = \Sigma$ and $Cov(\in_{i.,} \in_{i'.}) = 0$ $\forall \mathrm{i} \neq i'$ for the row $i = 1, \ldots, n$ of E and with $\mathrm{E}(\in.\mathrm{j}) = 0$ and Cov $(\in_{.j}, \in_{.\mathrm{j}'}) = \sigma_{j,j'}$ for the column $j = 1, \ldots, q$ of E, and where $\in_{i.} \sim \mathrm{Nq}(0,)$, the ordinary least square (OLS) estimator for B is

$$\hat{B} = (X'X)^{-1}X'Y. \tag{5}$$

For the variable search applied here, the Matrices Y and X were centred around their corresponding column means to Y_c and X_c, which led to the reduced parameter matrix

$$B_c = \begin{pmatrix} \beta_{11} \cdots \beta_{1q} \\ \vdots \qquad \vdots \\ \beta_{p1} \ldots \beta_{pq} \end{pmatrix} \tag{6}$$

The overall goodness of fit test for the hypotheses H_0: $B_c = 0$ and H_1: $B_c \neq 0$ is then based on a F-test with the test statistic, $F = \frac{1-\Lambda}{\Lambda} \frac{n-p-q}{q} \sim F_{q,N-p-q}$, where Wilks Lambda Λ is calculated by

$$\Lambda = \frac{|Q_e|}{Q_0} = \frac{|Q_e|}{|Q_e + Q|}, \tag{7}$$

with $Q_e = Y_c'Y_c - \hat{B}_c'X_c'X_c\hat{B}_c$ and $Q_0 = Y_c'Y_c$. The hypothesis H_0 is rejected if $F > F_{q,N-p-q}(1-\alpha)$.

The Tau squared index will be

$$\tau^2 = 1 - \Lambda^{\frac{1}{r}}, \tag{8}$$

with *r* being the rank of Q_0, for the matrix of response variables $Y$and any GA selected subset of the full predictor matrix. In addition, a local improvement algorithm was applied to the GA selected optimal variable subset. It follows the sequence of steps:

1. Arrange all variables, which do not belong to the GA-selected subset, in a queue.
2. Select a variable from the queue and swap this variable with each variable in the variable subset.
3. Remove the selected variable from the queue.
4. Measure the optimisation criterion τ for each of the swapping variants.
5. The best swapping variant is then chosen if it leads to a further improvement of the current state. Otherwise the state is reset to the prior state.
6. The variable, which has been exchanged by the variable from the queue now enters the queue, but only if it has not been included in the queue before.
7. The algorithm performs as a loop from Step 2 to Step 6 until the queue is emptied.

A high solution rate (i.e. 1000 bootstraps with a random split into training and testing sets) was set in order to stabilise the selected variable subset in multiple runs. To investigate the effect of the number of the selected variables on the model accuracy, a sequence of different subset sizes ranging from 5 to 18 variables were selected. The performance of the NN model was assessed for each of the optimal variable subsets. The variable selection was carried out irrespective of the remote sensing data sources. That is, none of the data sources were assigned a priori with any preference. The initial population was set to 100 (Cerdeira *et al.* 2009), the probability of mutation was 0.01, maximum number of identical replicates (clones) was 1, and the algorithm was run for 200 generations. The Subselect library (Cerdeira *et al.* 2009) of the R statistical computing environment was used for the computations.

2.2.4. Non-parametric models

The classification and regression tree method of RF (Breiman 2001) was used for simultaneous predictions of the six response variables. The RF contributed an additional layer of randomness to the method of bagging (Breiman 1996). In addition to constructing each tree using a different bootstrap sample of the data, RF alters how the classification or regression trees are constructed. The algorithm works as follows (Liaw and Wiener 2002):

1. Bootstrap samples are drawn from the original data.
2. For each bootstrap sample, an *unpruned* regression tree is grown. The best splits are chosen from the randomly sampled variables at each node.
3. The new predictions are then made by aggregating the predictions of the total number of trees. That is, the mode votes from the total trees will be the predicted value of the respective variable.

Adapted to be used as a NN imputation method, the distance between the target and reference units in RF is calculated as ‘one minus the proportion of terminal nodes from all regression trees where the target observation is in the same terminal node as the specific reference unit' (Crookston and Finley 2008). The method produces more robust results compared to single-tree classifications (Hudak *et al.* 2008). In order to return stable results in repeated runs, RF imputation here was set to generate 300 classification trees per response variable.

The number of NNs k considered for predictions of the target unit can be set to any number equal or greater than the total amount of samples n. Different numbers of k have been applied in several studies (e.g. Holmström and Fransson 2003, Haapanen *et al.* 2004, Maltamo *et al.* 2006, Packalén and Maltamo 2007, Finley and McRoberts 2008). Here, single NN was used by setting $k = 1$. The choice of k often depends on the analyst's expertise, as increasing k leads to a stronger shift of the predictions towards the sample mean. This may return seriously biased results particularly in cases where the distribution of observations is skewed. However, an increase of k reduces the prediction error (Hudak *et al.* 2008).

The RF prediction was applied on both sets of stratified and unstratified samples using the variable subsets of different sizes selected from GA search. Performance of the predictions for all response variables were compared amongst different variable subset sizes. To provide a means of comparison, the predictions of unstratified variant were also compared to the prediction results based on multivariate linear regression models as displayed by Equation (4) using the OLS estimator for parameter estimation shown in Equation (5).

The YaImpute library (Crookston and Finley 2008) of the statistical environment R was used for the RF computations.

2.2.5. *Assessment of performance*

Predictions were made for each of the field sample plots. Performance of variable predictions was assessed by means of leave-one-out cross-validation. For the prediction of one specific sample plot, the data from that plot was excluded from the regression space, thus the observation on any specific sample plot could not serve as prediction for itself. The results of cross-validation were reported as diagnostics including relative *RMSE* and performance enhancement $\mathrm{Enh}_{RMSE\%}$. If the absolute *RMSE* of the unstratified samples is

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{9}$$

Where n is the total number of sample plots, y_i is the observed response variable on plot i, and is the prediction of the response variable on corresponding plot, the relative *RMSE*% is obtained by $RMSE\% = RMSE/(1/n\sum_{i=1}^{n} y_i)100\%$.

For the stratified predictions, the RMSE of the corresponding stratum is $RMSE^{(\mathrm{str})} = \sqrt{\frac{1}{n}\sum_{c=1}^{3} n_c \times RMSE_c^2}$, where n is total number of samples, n_c is the number of sub-samples for each forest stratum, and $RMSE_c$ is absolute *RMSE* of the respective stratum. Then the percentage of $RMSE^{(str)}$ can be calculated as:

$$RMSE^{(\mathrm{str})}\% = \frac{RMSE^{(str)}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \times 100\% \tag{10}$$

Finally, the performance enhancement was derived as:

$$\mathrm{Enh}_{RMSE\%} = \frac{mean(RMSE^{(str)} - RMSE)}{mean(RMSE)} \times 100. \tag{11}$$

The *RMSE*s were used to compare the performance of the RF approach applied to the unstratified and the stratified variants. In addition, *RMSE* was used to compare GA-selected best subsets of different sizes. The above-mentioned statistics were calculating for 1000 randomly bootstrapped sub-samples, and the mean values of all runs were reported as final rates. This was done to prevent possible instabilities occurring because of the random element of RF, or when the variable prediction became caught in possible local optima.

3. Results

3.1. Selection of variable subsets

The GA variable selection resulted in optimal subsets consisting of 5 to 18 variables. Generally, the GA search generated variable subsets featuring more ALS variables than spectral features. The predictor variables selected by the GA from the initial 62 features are shown in Table 4. While some of the variables are repeated as the cardinality increases, some are replaced by the variables that were often highly correlated with them.

3.2. RF predictions

As reported in Figure 3, the stratification of the entire regression space led to a considerable performance enhancement of the RF predictions for all of the six attributes. Two examples are illustrated in Figure 3, according to which the stratified data improved the results regardless of the subset size. As one would assume, the improvement is more obvious in two strata-specific response variables including coniferous standing volume and coniferous biomass.

The accuracy of the RF models constantly increased for all the response variables as the number of selected predictor variables increased. Whereas both total timber volume and total biomass achieved the lowest *RMSE*%, the standing volume of trees >40 cm DBH yielded the highest *RMSE*% compared to other variables being predicted (Figure 4). The predictions showed roughly equal performance for timber volume and biomass in both total and coniferous rates. Performance was notably enhanced by the prior stratification of forest type. Figure 5 and Table 5 display a comparison of the performance enhancement rates for all the response variables across the predictor subset sizes. While the improvement was the lowest for the total stem number of trees with DBH >25 cm, the enhancements were apparently higher (up to greater than 10%) for the predictions of those response variables that benefited directly from the stratification (i.e. coniferous volume and coniferous biomass).

Furthermore, the performance enhancement for most of the response variables (except total volume in trees >40 cm DBH) reduced as the size of the subset exceeded 12 variables. This suggests that 12 predictor variables is the optimal subset size for the predictions.

Table 4. GA-selected predictor variables from the total 62 remote sensing features. The superscripts show the size of variable subsets in which the corresponding metric is included.

Data source	Selected predictors	Description
Normalized LiDAR points – First pulse	Height. Mean ^{5,16}	Height mean
	Height. StdDev ^{5,18}	Height Standard Deviation
	Height. InterquartileDistance ^{12,16,18}	Height interquartile distance
	Height. Skewness ^{7,11,13,14,15,18}	Height skewness
	Height.P10 ^{8,13}	Height 10th percentile
	Height.P20 ^{6,12,16,18}	Height 20th percentile
	Height.P30 ^{18}	Height 30th percentile
	Height.P60 ^{10,12,16}	Height 60th percentile
	Height.P80 ^{9,12,13}	Height 80th percentile
	Height.P90 ^{10}	Height 90th percentile
	Height.P95 ^{6,8}	Height 95th percentile
	Percentage.FR.2–5 m ^{9,10,11,12,15,16,18}	Percentage of first pulses >2 and <5 m
	Percentage.FR.10–12.5 m ^{7,11,13,14,15}	Percentage of first pulses >10 and <12.5 m
	Percentage.FR.12.5–15 m ^{9,14,16,18}	Percentage of first pulses >12.5 and <15 m
	Percentage.FR.15–17.5 m ^{16}	Percentage of first pulses >15 and <17.5 m
	Percentage.FR.17.5–20 m ^{14,18}	Percentage of first pulses >17.5 and <20 m
	Percentage.FR.20–25 m ^{6,7,8,9,11,12,13,14,15,16,18}	Percentage of first pulses >20 and <25 m
	Percentage.FR.25–30 m ^{5,6,7,8,9,10,11,12,13,14,15,16}	Percentage of first pulses >25 and <30 m
	Int. Mean ^{8,18}	Intensity mean
	Int. Mode ^{10,11,13,14,15,16,18}	Intensity mean
	Int. Median ^{7,15}	Intensity median
	Int. StdDev ^{18}	Intensity Standard Deviation
	Int. InterquartileDistance ^{9}	Intensity interquartile distance
	Int. Skewness ^{6}	Intensity skewness
	Int.P10 ^{14,15,16,18}	Intensity 10th percentile
	Int.P20 ^{11,15,16}	Intensity 20th percentile
	Int.P30 ^{7,10,11,14,18}	Intensity 30th percentile
	Int.P40 ^{10,11,12,14,16,18}	Intensity 40th percentile
	Int.P60 ^{5,12}	Intensity 60th percentile
	Int.P70 ^{8,13}	Intensity 70th percentile
	Int.P95 ^{9,10,12,13,14,15,16}	Intensity 95th percentile
TM data	TM2 ^{10,14,15,18}	TM band 2
	TM4 ^{13}	TM band 4
	TM5 ^{6,13,14}	TM band 5

Table 4 (*Continued*)

Data source	Selected predictors	Description
	TM7 [8,9,11]	TM band 7
	TM NDVI [13,15]	NDVI
	TM IR/R [12,13,18]	IR/R band Ratio
	TM_B457_PC1 [8,9,10,11,12,15]	2nd principal components of infrared bands
	TM_B457_PC2 [7]	2nd principal components of infrared bands
	TASSEL.BRIGHTNESS [15]	Tasseled Cap transformation- Brightness
	TASSEL.GREENNESS [5,15,16]	Tasseled Cap transformation- Greenness
	TASSEL.WETNESS [14,16,18]	Tasseled Cap transformation- Wetness

The *RMSE*% of the multivariate linear model predictions was slightly smaller than those of RF in most cases (except for coniferous volume and coniferous biomass) when applying unstratified samples (Table 6).

4. Discussion and conclusion

Continuous advancements in the acquisition of laser scanner data are leading to enhancements in the range of its applications to the study of Earth's surface (Estronell *et al.* 2010). Information integration (combining multi-source and multi-disciplinary information) and the development of specific, well-defined case studies of natural phenomena have been stated as two of the essential research requirements to achieve the vision of the next generation of Digital Earth (Craglia *et al.* 2008). According to Ehlers (2008), geoinformatics encompasses activities such as the acquisition and storage of geospatial data, the development of algorithms, modelling and presentation of spatial information. This study, based mainly on using open source computing facilities for data analysis, aimed at linking two and three dimensional remote sensing data to build reliable models and describe forest structure in a local level.

The significance of ALS data for studying forest structure is confirmed by previous investigations that have shown its importance compared to other remote sensing data sources. The application of various statistical metrics from multispectral and ALS data resulted in a high dimensional predictor variable set, in which many variables were strongly correlated, particularly those from ALS metrics. Thus, an efficient variable pruning scheme was essential. The traditional practice of variable selection by stepwise procedures has been criticised on several grounds by many authors (e.g. Barros and Rutledge 1998, Cadima and Jolliffe 2001), though it has been previously used to extract spectral features for NN modelling (Chirici *et al.* 2008). One serious shortcoming of these approaches is their failure to recognise that the importance of each variable for a given analysis is influenced by the final set of

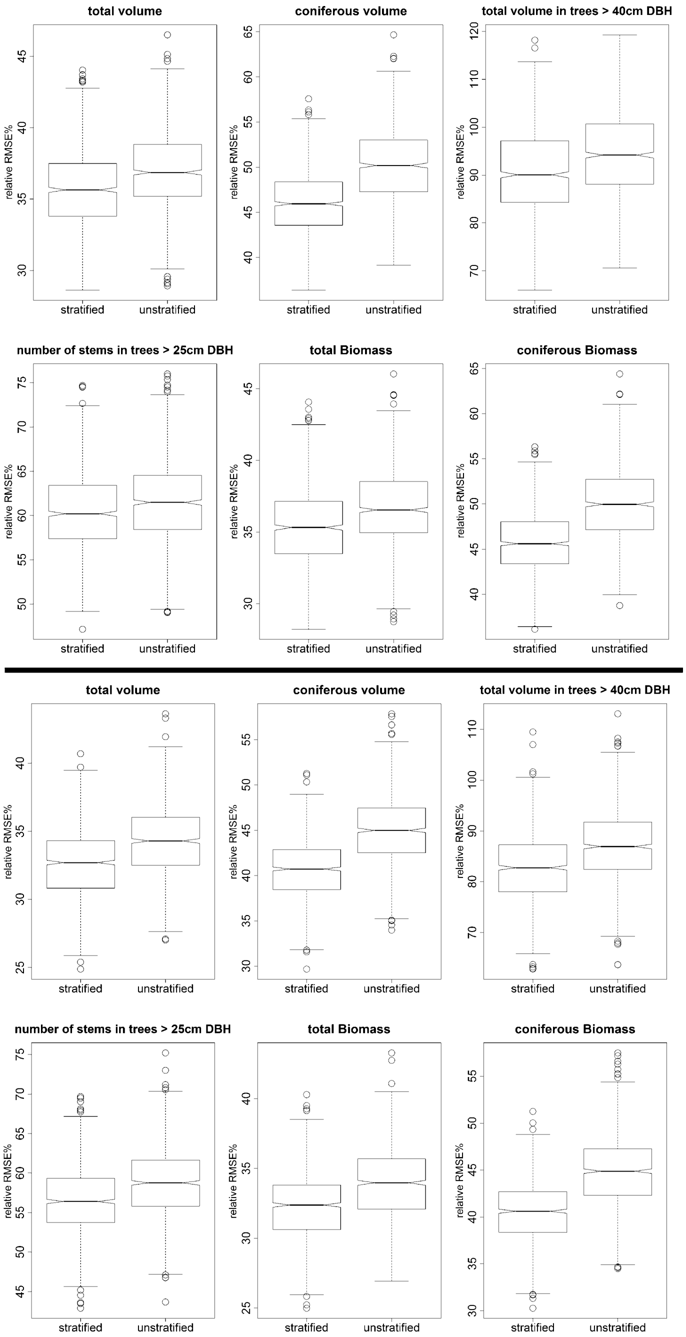

Figure 3. Box-plots for the relative *RMSE%* of RF predictions of the six response variables for the stratified and the unstratified variant from 1000 bootstrap resample for two examples of variable subsets featuring 6 (top) and 14 (bottom) variables. Non-overlapping notches between box-plots provide strong evidence of significant difference of medians (Chambers *et al.* 1983).

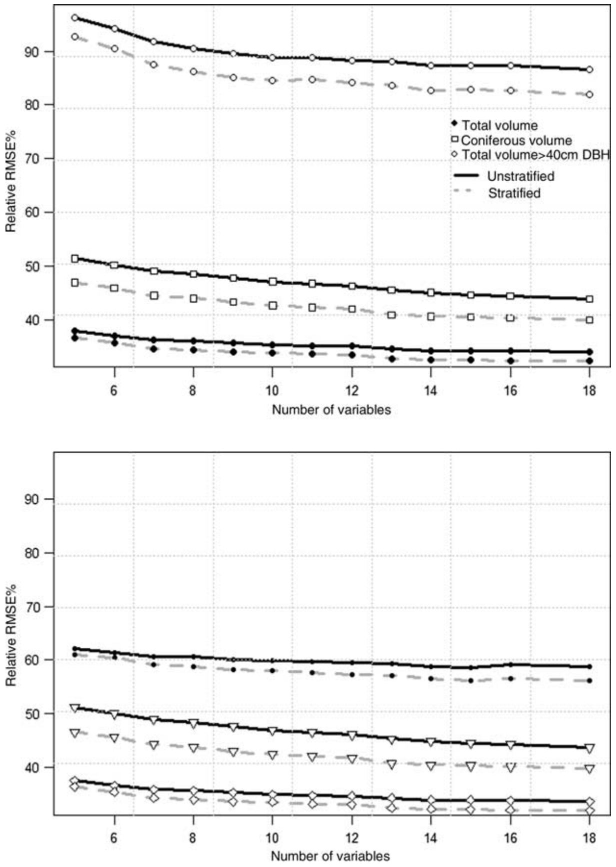

Figure 4. Relative *RMSE%* of RF predictions over the cardinality of the optimal variable subset for the unstratified (solid lines) and the stratified variant (dashed lines).

selected variables. This problem can be avoided by procedures such as the GA. As previously mentioned, Tomppo and Halme (2004) and later Tomppo *et al.* (2009) approved the plausibility of optimising a fitness function by GA to be used for tailoring the correlated predictor subsets derived from remote sensing data. In the present study, the τ^2 optimisation criterion based on Wilks Lambda was applied to obtain a tentative solution, which could then be fed to a local improvement algorithm for refinement. The additional application of a local improvement algorithm enables further enhancement of the selected subset, as also reported by Cadima *et al.* (2004) and Cerdeira *et al.* (2009). As an improvement to the previously reported shortcomings by Latifi *et al.* (2010), our results of variable screening showed more stable and consistent pruned predictor subsets for numerous repeated runs. This is appreciably advantageous over previous GA outputs, in that it optimises a fitness function for continuous multivariate responses (as opposed to Latifi *et al.* (2010)), uses an additional local improvement, and avoids repeating the similar solutions during its runtime, which in turn reduces the divergence time.

The prior stratification of forest type using CIR orthoimages yielded successful results. The method does not require any user-defined training regions and enables

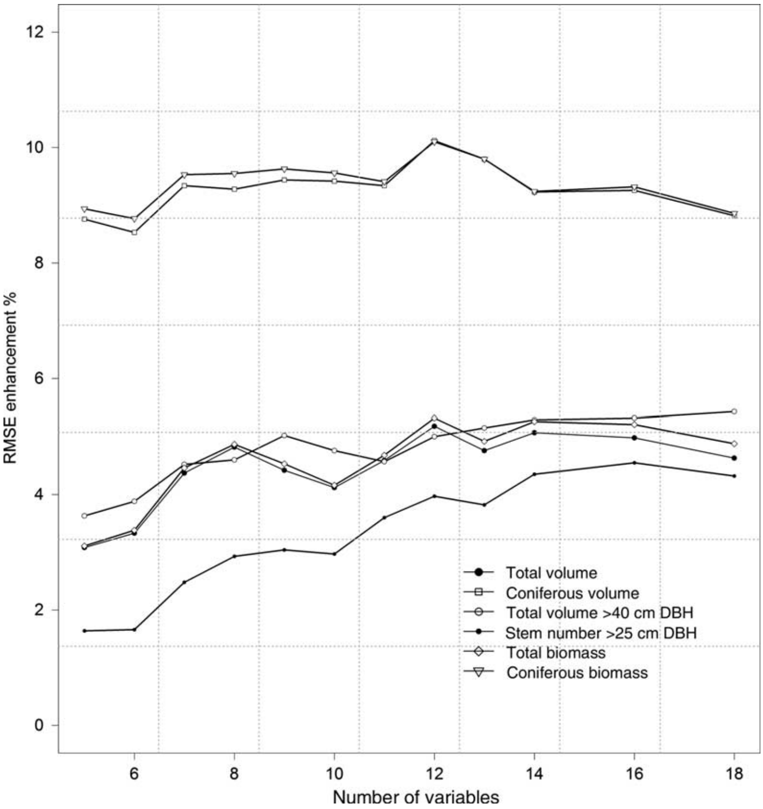

Figure 5. Performance enhancement rates in terms of *RMSE* (in percentage) induced by the prior forest type stratification for RF predictions of six response variables in different GA-selected subset sizes.

stratification of coniferous, deciduous and mixed forests. It is easy to apply to study sites outside of the region described here. However, further study is required to test the possibility of a more detailed categorisation of tree species composition. Our achieved overall accuracy (78.97%) of the classification into coniferous, deciduous and mixed forest is entirely in line with the results of the earlier studies such as those by Popescu and Wynne (2004), Tiede *et al.* (2006) and recently by Kim *et al.* (2011), though it is worth noting that in contrast to those mentioned above, an unsupervised method was applied here that is considered as a fast and cost-effective approach for retrieval of explicit information. However, one may consider that a direct comparison of results is sometimes hardly conceivable due to differences in affecting factors e.g. study sites, classification and validation techniques.

Results showed an increase in prediction accuracy along with increasing predictor subset sizes (Figure 5). The improvement was maximised for the majority of response variables when 12 predictors were selected. It then slightly declined as more predictors were appended. The majority of the selected variables consisted of first pulse height metrics extracted from ALS data (Figure 3). This reveals the dominance of topographic LiDAR data over the other 2D remote sensing sources when predicting forest structural attributes. Nevertheless, few optical metrics were

Table 5. Performance enhancement rates in terms of *RMSE* (in percentage) induced by the prior forest type stratification for RF predictions of six-fold response variables in different GA-selected subset sizes.

	Performance enhancement rates [%]					
Number of Variables	Total volume	Coniferous volume	Total volume in trees >40 cm DBH	Number of stems in tress >25 cm DBH	Total biomass	Coniferous biomass
5	3.08	8.76	3.63	1.64	3.11	8.94
6	3.33	8.53	3.88	1.66	3.38	8.77
7	4.37	9.34	4.52	2.48	4.46	9.53
8	4.82	9.28	4.60	2.93	4.87	9.55
9	4.42	9.44	5.02	3.04	4.53	9.63
10	4.12	9.42	4.76	2.97	4.16	9.56
11	4.59	9.34	4.57	3.60	4.68	9.41
12	5.18	10.12	5.00	3.97	5.32	10.10
13	4.76	9.80	5.15	3.82	4.92	9.80
14	5.07	9.23	5.29	4.35	5.26	9.24
16	4.98	9.26	5.32	4.55	5.21	9.32
18	4.63	8.82	5.43	4.32	4.88	8.86

included in the variable subsets. This generally supports Maltamo *et al.* (2006) who reported that the spectral components, when combined with the limited number of ALS metrics, may be helpful for forest attribute prediction at plot/stand level. While this view was also supported by some other area-based studies such as Koch (2010) and McInerney *et al.* (2010), Hudak *et al.* (2008) reported the inability of spectral predictors to explain variation beyond the variation explained by ALS metrics. The latter study used physiographic features (e.g. slope and aspect) as predictors, whereas we used no physiographic features in the current study as they were of minor relevance in the topographically-gentle landscape surveyed here. Amongst the GA-selected predictor subsets, the intensity metrics from ALS were also repeatedly present. In this regard, our results support the previous reports e.g. by Koch (2010) that the altimetric height information, when combined with physical values derived from ALS intensity, is appropriate for modelling forest structure.

The RF simultaneously takes both predictor and response variables into account and works based on large ensembles of regression trees. The method has previously been reported to produce higher accuracies for continuous forest attributes compared to other NN methods including MSN (Hudak *et al.* 2008, Latifi *et al.* 2010). Although increasing the number of regression trees built for each response variable may stabilise the estimates (Liaw and Wiener 2002, Hudak *et al.* 2008), the RF predictions may still diverge slightly from run to run. Therefore, the predictions were carried out numerous times as a means to stabilise the RF prediction results. The leave-one-out cross-validation was completed using 1000 random bootstraps. The results of predictions using different subset sizes showed a stable increment of accuracy, i.e. the predictions are consequently protected from being trapped in possible local optima. Concerning the response variables, the volume of trees >40 cm DBH was predicted least accurately (i.e. over 80% *RMSE*), which can be mainly attributed to the low abundance of old stems across the study area. This clearly

Table 6. Comparison of RF and OLS across GA-selected variable subsets in unstratified variant.

	RF – RMSE [%]						OLS- RMSE [%]					
No. of Variables	Total volume	Coniferous volume	Total volume of trees >40 cm_DBH	Stem number in trees >25 cm DBH	Total biomass	Coniferous biomass	Total volume	Coniferous volume	Total volume of trees >40 cm_DBH	Stem number in trees >25 cm DBH	Total biomass	Coniferous biomass
5	37.841	51.426	96.178	62.044	37.589	51.209	31.483	58.660	79.172	54.440	30.899	59.040
6	36.944	50.205	94.185	61.412	36.630	50.059	30.913	49.011	81.414	51.770	30.546	49.187
7	36.200	49.110	91.735	60.662	35.892	48.995	32.566	47.559	83.832	53.455	31.972	47.060
8	36.095	48.476	90.405	60.652	35.749	48.386	30.876	47.919	80.588	51.575	30.442	48.295
9	35.62	47.797	89.596	60.111	35.297	47.669	30.876	47.919	80.588	51.575	30.442	48.295
10	35.350	47.102	88.784	59.885	35.005	46.983	32.970	52.486	79.862	54.691	32.251	52.775
11	35.122	46.730	88.885	59.654	34.775	46.587	32.970	52.486	79.862	54.691	32.251	52.775
12	35.060	46.323	88.173	59.475	34.710	46.176	31.018	49.386	77.434	52.443	30.579	49.976
13	34.609	45.515	88.004	59.437	34.284	45.340	30.886	44.798	78.054	51.598	30.457	44.486
14	34.250	45.048	87.251	58.805	33.919	44.906	31.399	45.858	78.065	50.948	30.895	45.325
15	34.228	44.645	87.409	58.650	33.933	44.519	32.113	46.256	77.708	51.979	31.413	45.654
16	34.135	44.462	87.288	59.179	33.825	44.341	30.354	44.986	78.637	52.119	29.850	44.578
18	33.960	43.853	86.664	58.723	33.672	43.741	31.087	44.650	77.472	51.839	30.511	44.377

supports those recent reports from Yu *et al.* (2011) who used the RF method of prediction. Both volume and biomass showed the best prediction results and preformed similarly. This is evidently because the biomass is estimated by allometric equations with the standing volume as descriptor variable. Stem number predictions were not as accurate as standing volume predictions. This follows the results already obtained by Packalén and Maltamo (2007) and Maltamo *et al.* (2009a), both of which report that predictions of stem number were generally poorer than those of standing volume. Though the direct comparison of accuracy amongst different studies is difficult (as the studies are conducted under varying forest conditions, scanning systems and data characteristics), our models for stratified attributes yielded notably improved results compared to those from recent area-based studies in relatively similar landscapes. Breidenbach *et al.* (2010) can be used as a particular case of comparison, who reported a cross-validated *RMSE*% of 32.81% and 44.40% for RF models of total volume and coniferous volume in temperate forests, respectively. Whereas an average of eight NNs was used to yield the above-mentioned results, our stratified predictions using a single NN yielded nearly similar values.

Coniferous volume and coniferous biomass achieved greater performance enhancements compared to other attributes, with coniferous biomass being slightly better than coniferous volume. This supports the improvement from prior stratification, which led to more precise predictions by means of homogeneous strata. Twelve predictors were found to have the maximum relative improvement due to stratification, which, in this respect, is comparable to Hudak *et al.* (2008) and Maltamo *et al.* (2009b) who reported 12 predictors as optimum subset size for NN predictions of forest attributes using both RF and MSN methods.

The prediction errors of multivariate OLS models were slightly lower than RF errors in most cases (except for coniferous volume and coniferous biomass). This also follows Yu *et al.* (2011) who achieved roughly similar performance of RF and OLS in their study on a single tree scale. This is obvious, since non-parametric imputations do not share the same mix of error components as regression predictions. Imputation errors are often greater than regression errors because the errors do not result from a least-squares minimisation, but from selection of a most similar element in a pool of neighbouring observations. In this regard, our results support the previous statements of Stage and Crookston (2007). However, a single NN prediction ($k = 1$) produced predictions with similar variance structure to that of the observations (Moeur and Stage 1995). This is reported by e.g. Hudak *et al.* (2008) to be advantageous over the higher accuracies achievable by the use of OLS. Furthermore, one may note the ability of non-parametric NN models to simultaneously predict multiple responses, which is often missing when using OLS methods.

The ALS has proved its plausibility for reliable biomass estimations (Koch 2010). Due to the expected future technical innovations of these data for biomass assessments, it is assumed that ALS will play a prominent role in biomass modelling tasks. This study yielded promising results in terms of using non-parametric methods for providing area-based predictions of stratified forest structural attributes in the study area. Moreover, it illustrated how an evolutionary GA search can be applied to a high dimensional remote sensing dataset to produce parsimonious variable subsets. Owing to the results, the methodology is expected to produce promising results across similar areas where the problems of small sample size and multiple variables present challenges for representing local-scale forest structure in an accurate digital form.

Acknowledgements

This research was funded through a doctoral scholarship awarded by the German Academic Exchange Service (DAAD).The authors are grateful to the Forest Service of Baden-Württemberg (Landesforstverwaltung) for providing the required forest inventory data. We thank Nicholas L. Crookston (Rocky Mountain Research Station, USDA Forest Service) for revision of a draft of the manuscript and the anonymous reviewers who greatly contributed to enhancing the quality of this paper. Nathan A. Briggs and Mia Fischer are appreciated for editing the English language.

Notes on contributors

Hooman Latifi received a BSc degree in Forestry from the University of Guilan (Iran) in 2003, a MSc degree in Forestry from the University of Mazandaran (Iran) in 2005, and currently works toward a PhD degree at the department of remote sensing and landscape information systems (FeLis) in Albert-Ludwigs University of Freiburg (Germany). Since 2008, he has been a doctoral scholarship holder from the German Academic Exchange Service (DAAD) and a member of the Graduate School of Environment, Society and Global Change (ESGC) of the University of Freiburg. His main research interests are remote sensing-supported forest assessment, spatial statistics, airborne remote sensing data analysis, data mining, and model optimisation.

Dr. Arne Nothdurft received his doctoral degree of Forest Sciences from Georg-August University of Göttingen in 2007. He currently works at the department of biometry and informatics in forest research institute (FVA) Baden-Württemberg in Freiburg. His research interests are in forest inventory, spatial statistics and growth modelling.

Dr. Christoph Straub received a Dipl. Ing in Forest Sciences with specialisation in geographic information systems and landscape management from the University of Applied Forest Sciences in Rottenburg (Germany) in 2004, an MSc in photogrammetry and geoinformatics from Stuttgart University of Applied Sciences (Germany) in 2006, and a doctoral degree from the Faculty of Forest and Environmental Sciences at the Albert-Ludwigs University of Freiburg (Germany) in 2010. He is currently working as a research associate at the department of information technology at the Bavarian State Institute of Forestry (LWF) in Freising (Germany). His main interests are digital photogrammetry, airborne laser scanning, surface modelling and digital image processing.

Barbara Koch is a Professor of Remote Sensing and GIS at the Albert-Ludwigs University of Freiburg (Germany). She received an MSc in forest sciences in 1982, and later in 1988 a PhD degree in Remote Sensing from Ludwig-Maximilians University of Munich. She leads the Department of Remote Sensing and Landscape Information Systems (FeLis) of the University of Freiburg since 1994. Her main research interests include the application of remote sensing in forest management and planning, photogrammetry, digital image processing and laser scanner applications.

References

Anonymous, 2009. 2009 Beijing declaration on digital earth. *International Journal of Digital Earth*, 2 (4), 397–399.

Barros, A.S. and Rutledge, D.N., 1998. Genetic algorithm applied to the selection of principal components. *Chemometrics and Intelligent Laboratory Systems*, 40, 65–81.

Boyd, D.S. and Hill, R.A., 2007. Validation of airborne LiDAR intensity values from a forested landscape using HYMAP data: preliminary analysis. *Proceedings of the ISPRS*

Workshop 'Laser Scanning 2007 and SilviLaser 2007' Part 3/W52, 12–14 September, Espoo-Finland. Helsinki: Helsinki University of Technology, 71–76.

Breidenbach, J., Nothdurft, A., and Kändler, G., 2010. Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. *European Journal of Forest Research*, 129 (5), 833–846.

Breiman, L., 1996. Bagging predictions. *Machine Learning*, 24 (2), 123–140.

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

Cadima, J., Cerdeira, J., and Orestes Minhoto, M., 2004. Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47, 225–236.

Cadima, J. and Jolliffe, I.T., 2001. Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics*, 6, 62–79.

Cerdeira, J.O., *et al.*, 2009. *Package 'subselect' userguide.* R Development Core Team. Available from: http://cran.r-project.org/web/packages/subselect/index.html [Accessed 25 August 2010].

Chambers, J.M., *et al.*, 1983. *Graphical methods for data analysis.* Belmont, CA: Wadsworth.

Chirici, G., *et al.*, 2008. Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sensing of Environment*, 112, 2686–2700.

Craglia, M., *et al.*, 2008. Next-generation digital earth. *International Journal of Spatial Data Infrastructures Research*, 3, 146–167.

Crist, E.P. and Cicone, R.C., 1984. A physically-based transformation of thematic mapper data – the TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing*, 22 (3), 256–263.

Crookston, N.L. and Finley, A.O., 2008. yaImpute: an R package for KNN imputation. *Journal of Statistical Software*, 23 (10), 1–16.

Duarte Silva, A. P., 2001. Efficient variable screening for multivariate analysis. *Journal of Multivariate Analysis*, 76, 35–62.

Ehlers, M., 2008. Geoinformatics and digital earth initiatives: a German perspective. *International Journal of Digital Earth*, 1 (1), 17–30.

Estronell, J., *et al.*, 2010. Analysis of the factors affecting LiDAR DTM accuracy in a steep shrub area. *International Journal of Digital Earth*. DOI: 10.1080/17538947.2010.533201

Finley, A.O. and McRoberts, R.E., 2008. Efficient k-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment*, 112, 2203–2211.

Freitas, S.R., Mello, M.C.S., and Cruz, C.B.M., 2005. Relationships between forest structure and vegetation indices in Atlantic rainforest. *Forest Ecology and Management*, 218 (1–3), 353–362.

Gonzalez-Alonso, F., *et al.*, 2006. Forest biomass estimation through NDVI composites. The role of remotely sensed data to assess Spanish forests as carbon sinks. *International Journal of Remote Sensing*, 27 (24), 5409–5415.

Gruen, A., 2008. Reality-based generation of virtual environments for digital earth. *International Journal of Digital Earth*, 1 (1), 88–106.

Haapanen, R., *et al.*, 2004. Delineation of forest/nonforest land use classes using nearest neighbor methods. *Remote Sensing of Environment*, 89, 265–271.

Härdle, W., 1990. *Applied nonparametric regression. Econometric society monographs series.* Cambridge: Cambridge University Press.

Holland, J.H., 1975. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* Ann Arbor, MI: University of Michigan Press.

Holmström, H. and Fransson, E.S., 2003. Combining remotely sensed optical and radar data in kNN estimation of forest variables. *Forest Science*, 49 (3), 409–418.

Hudak, A.T., *et al.*, 2006. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return LiDAR and multispectral satellite data. *Canadian Journal of Remote Sensing*, 32, 126–138.

Hudak, A., *et al.*, 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, 112, 2232–2245.

Hyyppä, J., *et al.*, 2009. Forest inventory using small-footprint airborne LiDAR. *In*: J. Shan and C.K. Toth, eds. *Topographic laser ranging and scanning, principles and processing*. Boca Raton, FL: CRC Press, 335–370.

Kim, S.R., *et al.*, 2011. Forest cover classification by optimal segmentation of high resolution satellite imagery. *Sensors*, 11 (2), 1943–1958.

Koch, B., 2010. Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 581–590.

Koch, B. and Dees, M., 2008. Forestry applications. *In*: Z. Li, J. Chen, and E. Baltsavias, eds. *Advances in photogrammetry, remote sensing and spatial information sciences.* Balkema: CRC Press, 439–459.

Koch, B., *et al.*, 2009. Airborne laser data for stand delineation and information extraction. *International Journal of Remote Sensing*, 30 (4), 935–963.

Kraus, K. and Pfeifer, N., 1998. Determination of terrain models in wooded areas with airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 53, 193–203.

Kublin, E., 2003. Einheitliche Beschreibung der Schaftform – Methoden und Programme – BDATPro. *Forstwissenschaftliches Centralblatt*, 123, 183–200 (in German with English summary).

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.

Liaw, A. and Wiener, M., 2002. Classification and regression by random forest. *R News*, 2, 18–22.

Maltamo, M., *et al.*, 2006. Non-parametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*, 36, 426–436.

Maltamo, M., *et al.*, 2009a. Combining ALS and NFI training data for forest management planning: a case study in Kuortane, western Finland. *European Journal of Forest Research*, 128, 305–317.

Maltamo, M., *et al.*, 2009b. Non-parametric prediction of diameter distribution using airborne laser scanner data. *Scandinavian Journal of Forest Research*, 24, 541–553.

McGaughey, R.J., 2009. *FUSION/LDV: software for LIDAR data analysis and visualization (Ver. 2.70)*. Washington, DC: USDA Forest service-Pacific Northwest Research Station, 1–127.

McInerney, D.O., *et al.*, 2010. Forest canopy height retrieval using LiDAR data, medium resolution satellite imagery and kNN estimation in Aberfoyle, Scotland. *Forestry*, 83 (2), 195–206.

Mitchell, M., 1996. *An introduction to genetic algorithms.* Cambridge, MA: MIT Press.

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

MVTec, 2010. *HALCON/.NET reference manual.* Available from: http://halcon.cn/download/documentation/reference-8.0/dotnet/pouring.html [Accessed 19 January 2010]

Nothdurft, A., Soborowski, J., and Breidenbach, J., 2009. Spatial prediction of forest stand variables. *European Journal of Forest Research*, 128 (3), 241–251.

Packalén, P. and Maltamo, M., 2006. Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. *Forest Science*, 52 (6), 611–622.

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 Sensing of Environment*, 109, 328–341.

Popescu, S.C. and Wynne, R.H., 2004. Seeing the trees in the forest: using Lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing*, 70 (5), 589–604.

RIEGL, 2010. *RiANALYZE airborne software.* Available from: http://www.rieglusa.com/software-guide/rianalyze.shtml [Accessed 20 January 2011].

Soille, P., 1999. *Morphological image analysis.* New York: Springer.

Stage, A.R. and Crookston, N.L., 2007. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. *Forest Science*, 53 (1), 62–72.

Stibig, H.J., Achard, F., and Fritz, S., 2004. A new forest cover map of continental southeast Asia derived from SPOT-VEGETATION satellite imagery. *Applied Vegetation Science*, 7, 153–162.

Straub, C., *et al.*, 2009. Using airborne laser scanner data and CIR orthophotos to estimate the stem volume of forest stands. *Photogrammetric Fernerkun*, 3, 277–287.

Straub, C., Weinacker, H., and Koch, B., 2008. A fully automated procedure for delineation and classification of forest and non-forest vegetation based on full waveform laser scanner data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37 (B8), 1013–1019.

Tiede, D., Lang, S., and Hoffmann, C., 2006. Supervised and forest type-specific multi-scale segmentation for one-level-representation of single trees. *Proceedings of 1st International Conference on Object-based Image Analysis (OBIA 2006)*, 4–5 July, 2006, Salzburg. Austria: ISPRS.

Tomppo, E., *et al.*, 2009. Predicting categorical forest variables using an improved k-nearest neighbour estimator and Landsat imagery. *Remote Sensing of Environment*, 113 (3), 500–517.

Tomppo, E. and Halme, M., 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. *Remote Sensing of Environment*, 92, 1–20.

Trevino, V. and Falciani, F., 2006a. *GALGO: An R package for genetic algorithm searches.* Available from: www.bip.bham.ac.uk/vivo/galgo/Tutorial.pdf [Accessed 5 April 2011].

Trevino, V. and Falciani, F., 2006b. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, 22 (9), 1154–1156.

Yu, X., *et al.*, 2011. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (1), 28–37.

Zell, J., 2008. Methoden für die Ermittlung, Modellierung und Prognose der Kohlenstoffspeicherung in Wäldern auf Grundlage permanenter Großrauminventuren. Thesis (PhD). Faculty of Forest and Environmental Studies, University of Freiburg (in German with English summary).