A hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data

Nicole Reddy[a], Michael Gebreslasie[a], Riyad Ismail[a,]

[a]School of Agriculture, Earth and Environmental Sciences, University of KwaZulu-Natal, Westville campus, Durban, South Africa

DOI: http://dx.doi.org/10.4314/sajg.v6i3.8

Abstract

Up to date forest inventory data has become increasingly essential for sustainable planning and management of a commercial forest plantation. Forest inventory data may be collected in the form of traditional field based approaches or using remote sensing techniques. The aim of this study was to examine the utility of the partial least squares regression (PLSR), random forest (RF) and a PLSR-RF hybrid machine learning approach for the prediction of four forest structural attributes: (basal area, volume, dominant tree height and mean tree height) within a commercial Eucalyptus forest plantation using a combination of spectral and textural information of high spatial resolution (0.15m) remote sensing data. The best model for this study was produced for mature E. dunnii species for dominant tree height using the PLSR-RF hybrid model (R^2 = 0.82 and RMSE = 2.07m). The results of this study highlight the robustness and potential of the PLSR-RF hybrid model for the prediction of forest structural attributes using high resolution imagery within a commercial Eucalyptus forest plantation.

1. Introduction

For sustainable plantation forest management and planning it is crucial and necessary to acquire up to date measurements of forest structural attributes (Dye *et al*. 2012). In general remotely sensed data has demonstrated the potential to map forest structural attributes (Gebreslasie *et al*., 2011; Nichol and Sarkar, 2011; Dye *et al*., 2012; Ismail *et al*., 2015). Satellite based remote sensing data have been used to predict multi-source forest structural attributes (Gebreslasie *et al*., 2010). Their advantages include large geographic coverage and a broad spectral range (Lillesand *et al* 2008). Airborne remote sensing data on the other hand offers high spatial resolution and a narrow spectral range. The nature of the data thus allows for the extraction of texture features in order to predict forest structural attributes (Tuominen and Haakana 2005). This data can also be manipulated by machine learning algorithms for prediction and classification applications as shown in various case studies (Ismail *et al*., 2015). Machine learning algorithms such as the Random Forest (RF) algorithm uses recursive binary partitioning based on the classification and regression tree (CART) ruleset and is an ensemble learning algorithm that benefits from random subspace selection and bagging (Breiman, 2001; Abdel-Rahman, 2013). In general machine learning techniques have been used extensively to estimate forest structural attributes using remote sensing data. Studies such as Shataee *et al*. (2012) compared *k*-nearest neighbour (*k*-NN), support vector machine learning (SVM) and RF regression using ASTER data. These authors concluded that overall SVM and RF produced the lowest root mean square errors (RMSE), however RF proved to be superior to the other methods by producing unbiased results especially for basal area and stems per hectare (RMSE = 18.39 and 20.64, respectively). Subsequently, owing to its promising predictive potential, forest structural estimation studies such as Dye *et al*. (2012) used the RF algorithm with a combination of spectral and textural variables derived from QuickBird imagery to produce an overall model accuracy of $R^2 = 0.68$.

In contrast, certain researchers have favoured the utility of linear machine learning algorithms such as the Partial Least Squares Regression (PLSR) algorithm which uses an iterative process (Wold *et al*., 2001). The algorithm can compress data which allows for the reduction in a large number of variables that are collinear thus allowing for the development of a few non-correlated latent variables also known as factors/components (Vyas and Krishnayya, 2014). Wolter *et al*. (2009) used SPOT-5 sensor data to estimate DBH, tree height, basal area and vertical length of live crown within a forest using a PLSR approach. The outcome of the study showed favourable results for DBH and tree height estimations with R^2 values of 0.82 and 0.69 respectively. A LiDAR based study by Næsset *et al*. (2005) used a combination of ordinary least squares (OLS), seemingly unrelated regression (SUR) and PLSR to estimate forest structural attributes at stand level using laser scanning technology and PLSR produced a best overall R^2 value of 0.94.

Research has shown that the PLSR and RF algorithms have powerful modelling potential and it is with this background that this study proposes a novel approach to predicting forest structural attributes by combining the PLSR and RF algorithms to form a PLSR-RF hybrid algorithm for the prediction of forest structural attributes within a commercial forest plantation. The hybrid approach uses the RF ensemble creating methodology with the addition of the PLSR components instead of using individual remote sensing variables. To the best of our knowledge, no study has assessed a hybrid PLSR and RF (PLSR-RF) machine learning approach to modelling forest structural attributes using multispectral remote sensing imagery within a commercial forest plantation. Therefore, our main objective was to investigate the robustness of these three (PLSR, RF, PLSR-RF) machine learning algorithms in predicting forest structural attributes using spectral and textural remote sensing image characteristics extracted from high spatial resolution (0.15m) imagery.

2. Materials and Methods

2.1 Study site

The study was conducted at the Sappi Riverdale plantation located West of the town of Richmond in the Midlands of KwaZulu-Natal, South Africa, located at 29° 52′ 0″ S, 30° 16′ 0″ E (Figure 1). The total area of the plantation spans 6200ha and is located in the upper catchment area along the Lovu River. The average altitude and temperature is 1190m and 16.1°C respectively. The area receives a mean annual precipitation and runoff of 9-16mm and 143mm respectively. The forested area is characterised by extensive commercial forestry dominated by *Eucalyptus* species such as *Eucalyptus dunnii* and *Eucalyptus grandis*. The *Eucalyptus* species are rapid growing species and are harvested every six to ten years (Owen and Van Der Zel 2000).

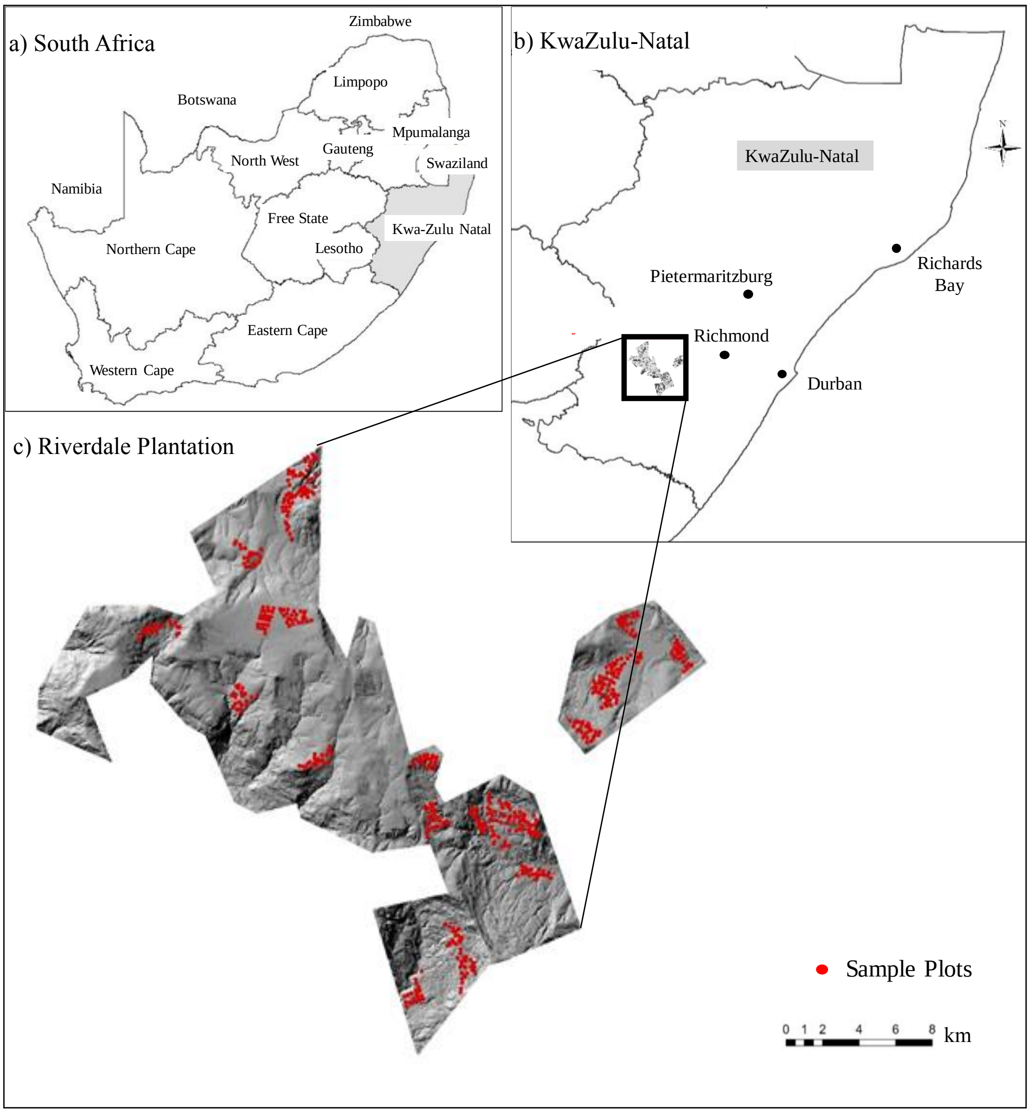

Figure 1: Location of the study area; a) South Africa; b) KwaZulu-Natal; c) Riverdale Plantation

2.2 Field data

The field survey campaign was conducted between the 12th of April and the 22nd May 2014 using industry standard enumeration techniques. A total of 502 georeferenced 10m radius circular plots across 25 compartments were developed based on a systematic grid sampling technique. Selected tree structural attributes such as volume (Volha), tree height: (mean tree height (Htm) and dominant tree height (HtD)) and basal area (Baha) were measured for each circular plot. *Eucalyptus dunnii* and *Eucalyptus grandis* species that were between two and ten years old were considered for this study. The data was partitioned according to the individual *Eucalyptus* species. The *E. grandis PCA*(n = 288) and *E. dunnii* (n = 214) were processed as separate input data. The tree height and DBH for each plot was measured using the Vertex IV laser instrument and Haglof Digitech Calliper, respectively.

2.3 Remote sensing data

Multispectral airborne image data was collected on the 12th April 2014 by Land Resource International under cloudless conditions. The image data was geometrically and radiometrically corrected and supplied as Geo-TIFF files. The data had an 8-bit radiometric resolution with a 0.15m spatial resolution and four spectral bands (Table 1).

Table 1: Spectral characteristics of the multispectral airborne imagery

Band number	Colour	Band configuration
Band 1	Red	650 to 680 nm
Band 2	Green	550 to 580 nm
Band 3	Blue	450 to 480 nm
Band 4	Near Infrared	720 to 750 nm

2.4 Texture feature extraction

The texture features used in the present study were proposed by Haralick *et al*. (1973) who suggested that texture measures depend heavily on the spatial resolution, spectral domain and the object characteristics within the image (shape and dimension). Nichol and Sarkar (2011) have suggested that image texture may be considered as a plausible proxy for forest structural attribute modelling and may be extracted by means of a Grey Level Co-occurrence Matrix (GLCM) and a Grey Level Difference Vector (GLDV). GLCM describes the texture features by the stochastic properties in the image relating to the spatial distribution of the grey levels in an image (Haralick, 1979). GLDV refers to the sum of the diagonals of the GLCM and makes reference to a pixel and its neighbour by counting the occurrence of the absolute difference between them (Haralick, *et al*. 1973). The texture features that were extracted using GLCM method were; entropy, dissimilarity, contrast, second angle moment, homogeneity, mean, standard deviation and correlation. The features extracted using GLDV were; second angle moment, mean, contrast and entropy. Using the first component from a principal component analysis of the multispectral image data, texture features were extracted from four directions (0°, 45°, 90°, 135°) thus to achieve directional invariance on each individual stand plot.

2.5 Partial least squares regression (PLSR)

PLSR is a linear statistical method that combines and uses the theory of principal component analysis (PCA) with the theory of multiple linear regression (MLR) and is considered to be a very effective modelling tool for feature extraction and dimension reduction (Abdi, 2007). The PLSR linear multivariate model is useful for analysing datasets with many high dimensional and collinear predictors (Wold *et al.*, 2001). The PLSR model creates orthogonal (uncorrelated) weight vectors by maximising the covariance between the explanatory and response variables while reducing the dimensionality of these *x* variables by sifting out the factors that explain the most information between all the *x* and *y* variables (Lopatin *et al.*, 2015). The PLSR operates by transforming the original predictors $X_1, X_2,\ldots,X_p$ into uncorrelated latent variables such as $Z_1, Z_1,\ldots Z_M$ where $M < p$ and Z is the weighted linear combinations of the original predictors (p) (Equation 1). New variables that are created are denoted by Z_m (m = 1, 2,….M). The X scores are estimated as linear combinations of the original variables X_i with the coefficient weights $\emptyset_{jm}$ (m = 1, 2,….M).

$$Z_m = \sum_{i=j}^{p} \emptyset_{jm} X_j \text{...Equation (1)}$$

The linear regression model is then fit to the latent variables known as the PLS factors in an orthogonal space (M) (Equation 2).

$$Y_i = \theta_0 + \sum_{m=1}^{m} \theta_m Z_{im} + \epsilon_i \text{...Equation (2)}$$

$I = 1,\ldots n,$ θ_0 is the regression intercept and θ_m is the regression coefficients for each of the PLS factor z across n observations where y_i is the response variables and ϵ_i the residuals.

In this study the PLSR was implemented using the R statistical software version 3.2.2 (R Development Core Team, 2014). During model development a 10-fold cross validation was done to obtain the optimum number of PLSR factors.

2.6 Random forest (RF)

The basic idea behind the random forest (RF) algorithm is to achieve an improved predictive accuracy by growing a large number of decorrelated trees. This is done to obtain a prediction accuracy by averaging the prediction values from all the trees in the ensemble for each observation. RF is thus, especially beneficial for data sets with a large number of predictors that may be correlated (Breiman, 2001). The RF method is a bagging method and uses recursive partitioning to form regression trees. Each regression tree that is created is then independently grown until its maximum size is reached based on the training data set, known as the bootstrap sample consisting of 66% of the total population. The RF model uses the remaining 34% of the data known as the out-of-bag (OOB) data for the model prediction (Breiman, 2001). The RF regression algorithm was implemented using a package developed by Kuhn and Johnson (2013) called ‘caret’ and the ‘randomForest’ package (Liaw ad Wiener, 2002) in the R statistical software version 3.2.2 (R Development Core Team, 2014).

2.7 Partial least squares-random forest (PLSR-RF) hybrid

The PLSR-RF hybrid model (Figure 3) improves the random forest non-parametric methodology with the addition of the linear PLSR approach. The PLSR part of the hybrid algorithm creates latent variables from the explanatory *(x)* variables that are the most relevant for the response *(y)* variables. These latent variables now serve as new predictor variables that can be used by the RF algorithm. Subsequently, the RF algorithm (i) creates an ensemble of trees and each tree is grown from a sample that is randomly selected from the bootstrap sample of the training data with replacement and (ii) randomly selects a component from the subset of components for splitting at each node of the trees. The hybrid model combines the benefit of the linear regression model of the PLSR algorithm with the non-linear RF ensemble method. During model calibration a 10-fold cross validation approach was applied to ensure that the prediction accuracies were unbiased and accurate.

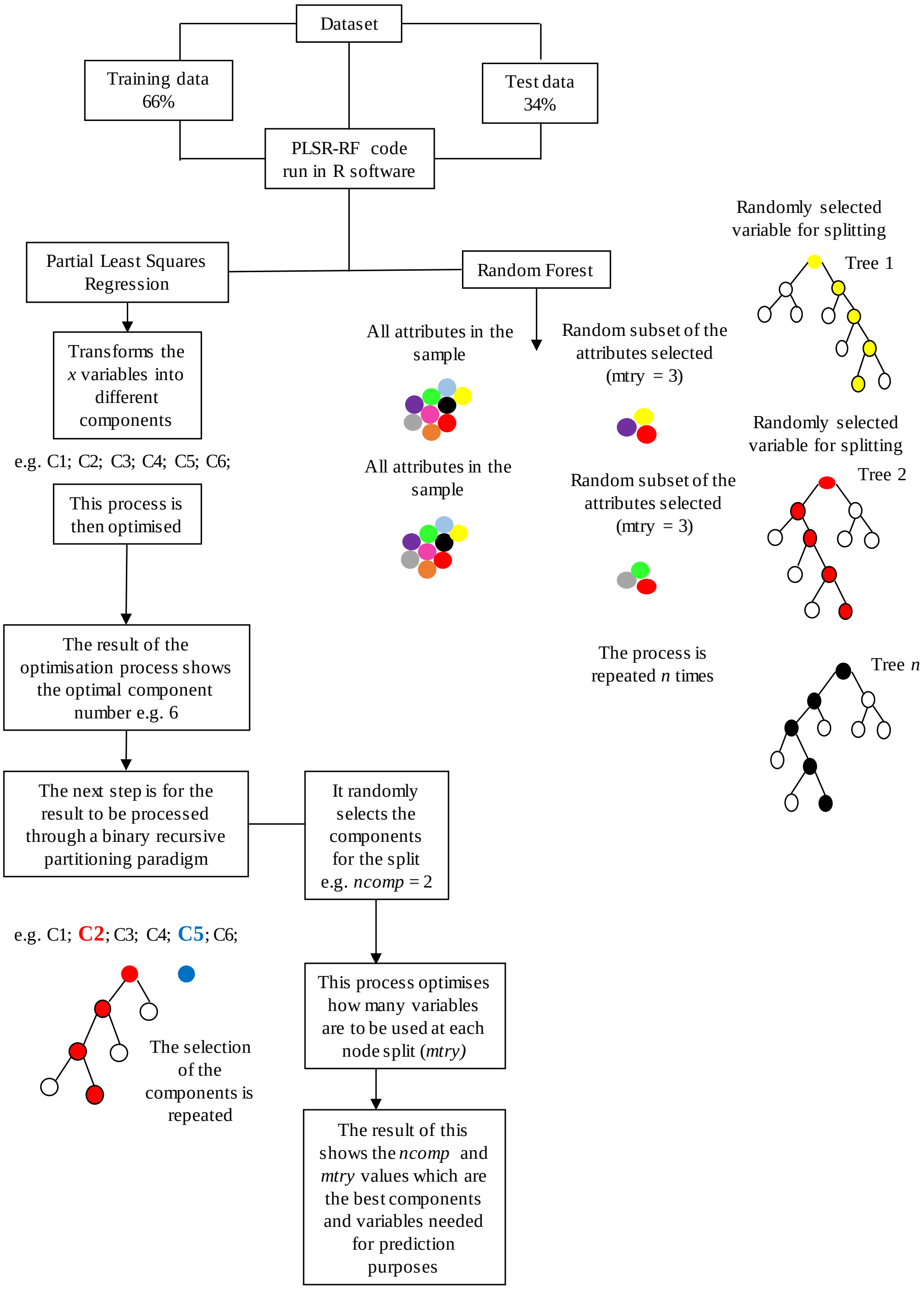
Dataset
Training data 66%
Test data 34%
PLSR-RF code run in R software
Partial Least Squares Regression
Random Forest
Transforms the x variables into different components
e.g. C1; C2; C3; C4; C5; C6;
This process is then optimised
The result of the optimisation process shows the optimal component number e.g. 6
The next step is for the result to be processed through a binary recursive partitioning paradigm
It randomly selects the components for the split e.g. ncomp = 2
e.g. C1; C2; C3; C4; C5; C6;
The selection of the components is repeated
This process optimises how many variables are to be used at each node split (mtry)
The result of this shows the ncomp and mtry values which are the best components and variables needed for prediction purposes
All attributes in the sample
Random subset of the attributes selected (mtry = 3)
All attributes in the sample
Random subset of the attributes selected (mtry = 3)
The process is repeated n times
Randomly selected variable for splitting
Tree 1
Randomly selected variable for splitting
Tree 2
Tree n

Figure 2: A graphical representation of the PLSR-RF hybrid model.

2.8 Model Validation

To ensure that the validation results were unbiased a training-test set (66%-34% split respectively) partition was done as suggested by Kuhn and Johnson (2013).

To gauge the predictive accuracy of the PLSR, RF and PLSR-RF hybrid models in predicting forest attributes within a commercial forest plantation the coefficient of determination (R^2) and root mean square error (RMSE) for the validation sample data were computed (Equation 3).

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(X_{measured} - X_{predicted})2}}{2} \qquad \text{Equation (3)}$$

$X_{measured}$ represents the measured forest attributes, $X_{predicted}$ represents the predicted values from the validation data and *i* represents the explanatory variables included in the summation process.

3. Results

The hyper-parameter optimization results for all the *E. grandis and E.dunnii* models developed in this study are shown in Table 2. For the PLSR models, between 6 and 14 latent components (*ncomp*) were selected using 10 fold cross validation. For the RF algorithm, the values selected for the *mtry* hyperparameter were between 3 and 12, while the *ntree* parameter was consistently set at 500. For the PLSR-RF algorithm (i) between 6 and 12 latent components were selected (ii) the selected *mtry* hyperparameters were between 2 and 9 and (iii) similar to the RF algorithm a *ntree* value of 500 produced the best results

Table 2: Optimal hyper-parameters (*ncomp, mtry* and *ntree*) for the *E. grandis and E.dunnii* models. Models were optimized for estimating: volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

	PLSR	**RF**		**PLSR-RF**		
Forest attribute	***ncomp***	***mtry***	***ntree***	***ncomp***	***mtry***	***ntree***
		***E. grandis* species**				
Volha	7	7	500	7	7	500
HtD	7	12	500	10	9	500
Htm	7	8	500	11	8	500
Baha	14	4	500	11	5	500
		***E. dunnii* species**				
Volha	6	6	500	6	2	500
HtD	11	5	500	11	5	500
Htm	6	6	500	10	4	500
Baha	6	3	500	12	8	500

For the *E. grandis* species (Figure 4a), the PLSR-RF algorithm produced the best models for volume (R^2 = 0.49 and RMSE = 38.58tons/ha), mean tree height (R^2 = 0.47 and RMSE = 3.52m) and dominant tree height (R^2 = 0.50 and RMSE = 2.37m). For the *E. dunnii* species (Figure 4b), the PLSR-RF model produced the best result when predicting volume (R^2 = 0.57 and RMSE = 61.31tons/ha) and mean tree height (R^2 = 0.57 and RMSE = 1.93m).

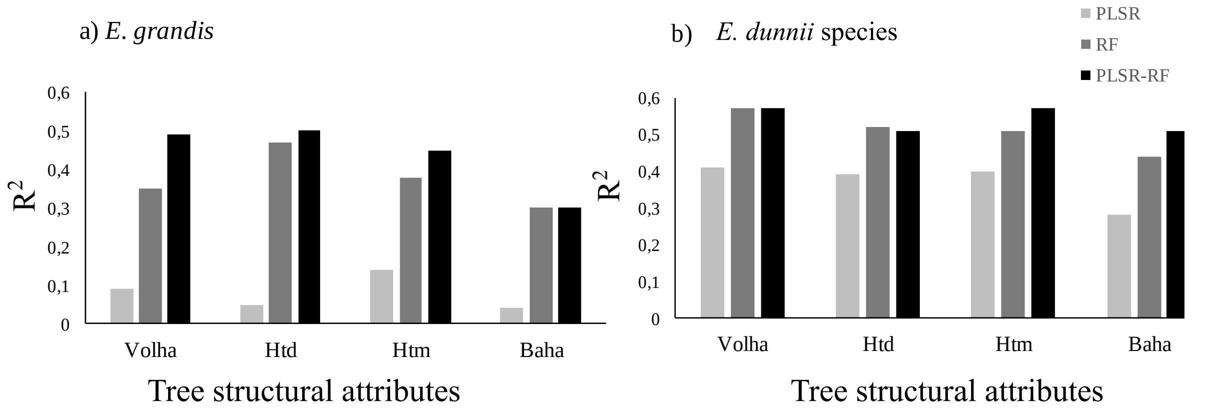

Figure 4: Model predictions using individual tree species (a) *E. grandis* and (b) *E. dunnii.* The results are shown for the three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features. Models accuracies are shown for volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

This study further looked at predicting the forest structural attributes for the young and mature forests.

The hyper-parameter optimization results for the young and mature *E. grandis* models are shown in Table 3. For the PLSR models, between 5 and 14 latent components (*ncomp*) were selected. For the RF algorithm, the values selected for the *mtry* hyperparameter were between 2 and 15, while the *ntree* parameter was consistently set at 500. For the PLSR-RF algorithm (i) between 3 and 11 latent components were selected (ii) the selected *mtry* hyperparameters were between 1 and 8 and (iii) a *ntree* value of 500 produced the best results

Table 3: Optimal hyper-parameters (*ncomp, mtry* and *ntree*) for the young and mature *E. grandis* models. Models were optimized for estimating: volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

	PLSR	**RF**		**PLSR-RF**		
Forest attribute	***ncomp***	***mtry***	***ntree***	***ncomp***	***mtry***	***ntree***
			Young *E. grandis* species			
Volha	5	2	500	3	1	500
HtD	7	5	500	6	3	500
Htm	8	5	500	8	2	500
Baha	5	5	500	7	2	500
			Mature *E. grandis* species			
Volha	12	6	500	9	7	500
HtD	12	2	500	11	8	500
Htm	14	15	500	7	4	500
Baha	10	2	500	10	4	500

Young *E. grandis* species were grouped at ages three to six years (Figure 5a) and mature *E. grandis* species were grouped at seven to 10 years (Figure 5b). The best model for young *E. grandis* species were developed for dominant tree height using the RF model (R^2 = 0.79 and RMSE = 1.76m) followed by a 10% decrease when using the PLSR-RF hybrid (R^2 = 0.69 and RMSE = 2.10m) (Figure 5a). The RF algorithm continued to produce promising results when applied to the mature *E. grandis* species with the best model being produced for dominant tree height (R^2 = 0.63 and RMSE = 2.05m) but could not explain more than 35% of the variation for the other forest structural attributes. The PLSR-RF model produced the best model for volume (R^2 = 0.59 and RMSE = 51.02tons/ha) for mature *E. grandis* species. The PLSR algorithm could not explain more than 40% of variation across all forest structural attributes for both young and mature *E. grandis* species.

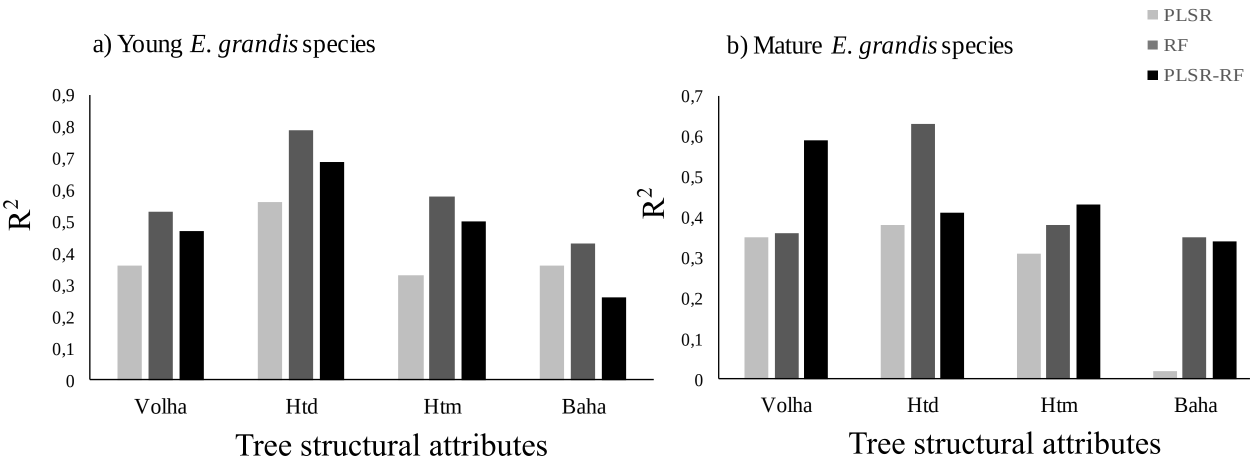

Figure 5: Model predictions shown for (a) young and (b) mature *E. grandis* species. The results are shown for the three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features. Models accuracies are shown for volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

The hyper-parameter optimization results for the young and mature *E. dunnii* models are shown in Table 4. For the PLSR models, between 5 and 12 latent components (*ncomp*) were selected. For the RF algorithm, the values selected for the *mtry* hyperparameter were between 3 and 16, while the *ntree* parameter was consistently set at 500. For the PLSR-RF algorithm (i) between 3 and 10 latent components were selected (ii) the selected *mtry* hyperparameters were between 3 and 10 and (iii) a *ntree* value of 500 produced the best results

Table 4: Optimal hyper-parameters (*ncomp*, *mtry* and *ntree*) for the young and mature *E. dunnii* models. Models were optimized for estimating: volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

	PLSR	**RF**		**PLSR-RF**		
Forest attribute	*ncomp*	*mtry*	*ntree*	*ncomp*	*mtry*	*ntree*
			Young *E. dunnii* species			
Volha	6	4	500	9	7	500
HtD	6	5	500	11	8	500
Htm	7	3	500	7	4	500
Baha	5	16	500	10	4	500
			Mature *E. dunnii* species			
Volha	12	16	500	10	10	500
HtD	5	9	500	5	3	500
Htm	6	3	500	10	6	500
Baha	6	16	500	3	3	500

Young *E. dunnii* species were grouped at ages three to six years (Figure 6a) and mature *E. dunnii* species were grouped at seven to 10 years (Figure 6b). For the young *E. dunnii* species PLSR could not explain more than 39% of the variation for all the forest attributes considered in this study. When using the RF model for the young *E. dunnii* species, the highest accuracy was reported for mean tree height (R^2 = 0.54 and RMSE = 1.83m) and dominant tree height (R^2 = 0.51 and RMSE = 2.23m). When considering the young *E. dunnii* species, the PLSR-RF model produced the highest accuracies for basal area (R^2 = 0.55 and RMSE = 2.93ha) and dominant tree height (R^2 = 0.65 and RMSE = 1.85m). For the mature *E. dunnii* species, the reported accuracies for dominant tree height across all three machine learning algorithms were as follows: PLSR (R^2 = 0.78 and RMSE = 2.38m), RF (R^2 = 0.75 and RMSE = 2.34m) and PLSR-RF (R^2 = 0.82 and RMSE = 2.07m).

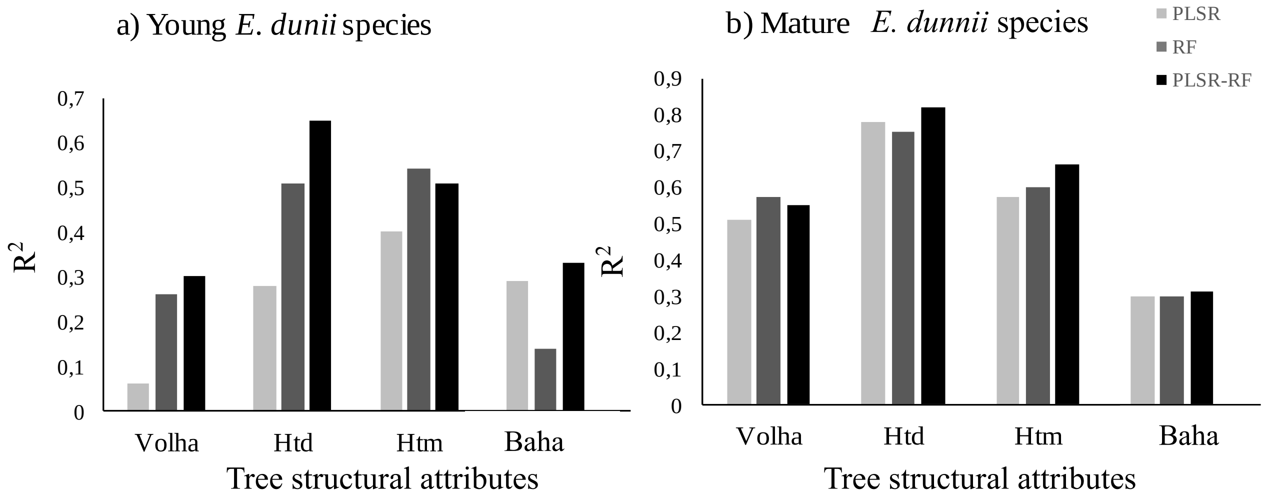

Figure 6: Model predictions using only *E. dunnii* species that are (a) young and (b) mature across three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features. Models accuracies are shown for volume (Volha), mean tree height (Htm), dominant tree height (HtD) and basal area (Baha).

The results of this study suggest that dominant tree height was the forest structural attribute that was the most accurately predicted using a combination of spectral and textural features. The PLSR-RF hybrid algorithm produced the highest model accuracies for the young and mature *E. dunnii* species (Figure 7a and b). The RF algorithm produced the highest model accuracies for the young and mature *E. grandis* species (Figure 7c and d). Overall, the best prediction model was obtained by the PLSR-RF algorithm (R^2 = 0.82) and could be potentially used to predict the dominant height for mature *E. dunnii* species.

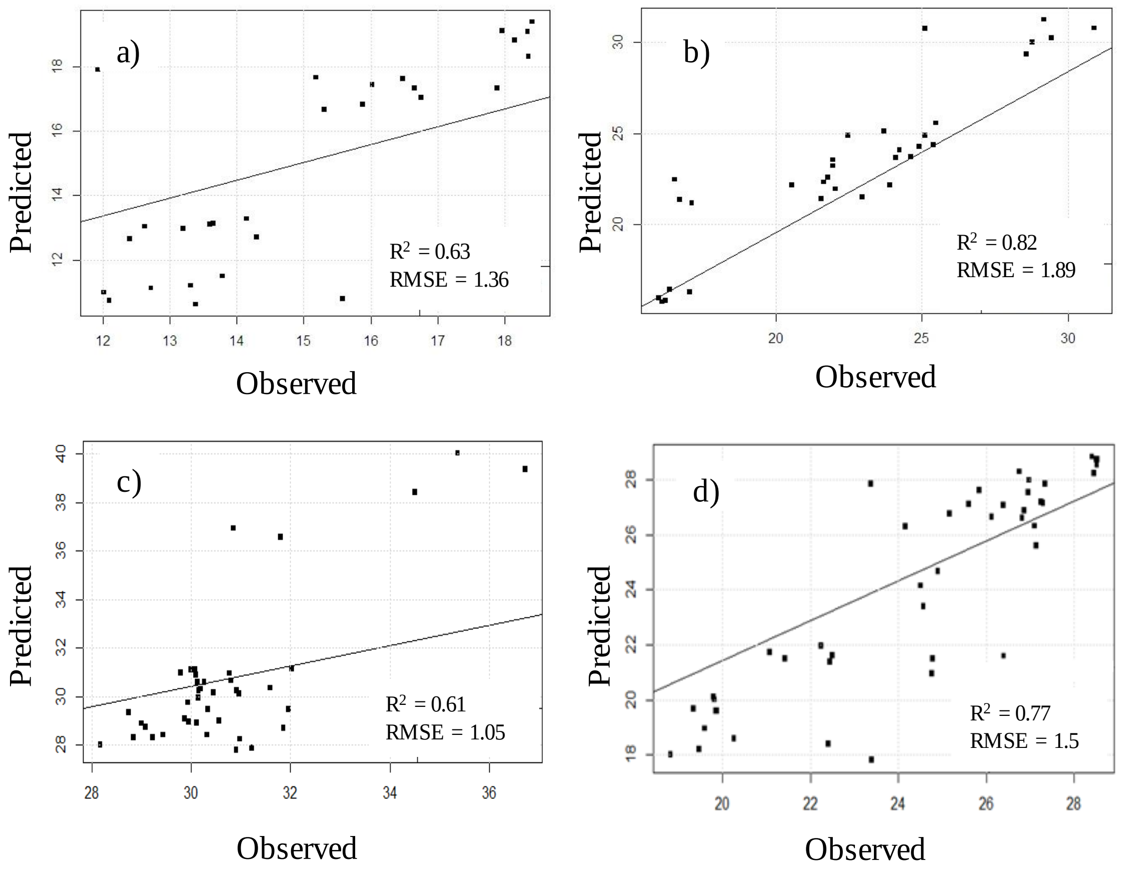

Figure 7: Observed vs predicted graphs for the dominant height models produced in this study. All models produced the highest accuracies for dominant tree height a) young *E. dunnii* b) mature *E. dunnii* c) young *E. grandis* d) mature *E. grandis.*

4. Discussion

This study focussed on predicting forest structural attributes for two *Eucalyptus* species with different age ranges using three machine learning techniques. In addition to the results obtained for the PLSR and RF algorithms, this study has demonstrated the strength and applicability of a hybrid algorithm that was developed using PLSR and RF methodologies for forest structural attribute prediction in a pulpwood *Eucalyptus* forest plantation located in the temperate climatic zone of South Africa.

In an attempt to test the model accuracies across all three machine learning algorithms, the data set was firstly split into individual species of all ages ranging from 2 to 10 years. When using the PLSR method, the *E. grandis* model accuracies still returned statistically weak with coefficient of determination values ranging from 0.09 for volume and reaching a high of 0.14 for mean tree height. When the RF method was applied to the *E. grandis* species dataset, the highest accuracy obtained for dominant tree height (R^2 = 0.47 and RMSE = 3.52m). The results of the *E. grandis* further improved when the PLSR-RF hybrid model was used, with high accuracies being reported for volume (R^2 = 0.49 and RMSE = 68.07tons/ha) and dominant tree height (R^2 = 0.50 and RMSE =

3.34m). The results for *E. grandis* was relatively low when compared to *E. dunnii* where the highest accuracy was reported for volume ($R^2 = 0.57$ and RMSE = 61.31tons/ha) using the RF method followed by the PLSR-RF hybrid model producing an accuracy of $R^2 = 0.55$ and RMSE = 62.61tons/ha. All model accuracies for *E. dunnii* using the RF and PLSR-RF hybrid were above 50%. These results suggest that forest structural attributes could be estimated more accurately for *E. dunnii* and lead to the idea that accuracies could further improve once an age partition was applied to each individual species. Groups of trees within a forest (of all ages) result in diameter-at breast-height and tree height increasing as the tree grows but the crown closure might remain quite small. Hence a commercial forest plantation of certain ages could share similar crown percentages resulting in them potentially having the same spectral reflectance's but differing measurements for the forest structural attributes which could be a reason for poor performance of the models when the ages of the individual species were combined.

Jensen *et al*. (1998) suggested that the phenological cycle of trees have internal structural changes which has a significant effect of spectral responses thus machine learning algorithms may be sensitive to age ranges within the data. Stand age for individual species could have significant effects on forest structural attribute estimation. The results of this study further improved when *E. grandis* was separated into young and mature trees within their individual compartments. Model performance for PLSR reached a high of 56% for dominant tree height for young *E. grandis* species and continued to improve as the PLSR-RF hybrid was applied with a high R^2 value of 0.69 being achieved for dominant tree height. However, for young *E. grandis* species the RF method performed the best with high R^2 values of 0.79 and 0.58 for dominant tree height and mean tree height respectively. When *E. dunnii* species was partitioned according to young and mature trees the highest model accuracies for the young trees were reported for basal area ($R^2 = 0.55$ and RMSE = 2.93ha) and dominant tree height ($R^2 = 0.65$ and RMSE = 1.85m) using the PLSR-RF hybrid. Using the mature *E. dunnii* species model accuracies produced using the PLSR-RF hybrid were the best when being compared to PLSR and RF. This was because the PLSR-RF hybrid used the PLSR methodology of latent variables and converted them into components. These components then underwent a process of RF binary recursive partitioning to select the optimal components for prediction. Hence during the model development process, the PLSR-RF hybrid selected a random subset of the optimal components from the bootstrap sample that were defined using PLSR to allow for an ensemble of trees to be created using the RF methodology. Using this hybrid methodology, the highest model accuracies were reported for dominant tree height ($R^2 = 0.82$ and RMSE = 2.07m) and mean tree height ($R^2 = 0.66$ and RMSE = 1.90m). As the forest develops into mature trees some individual trees begin to die off due to competition for light, water and soil resources. Commercial forests often practices thinning methods which may influence crown closures and resultant canopy gaps (Gebreslasie *et al*., 2011). This could be interpreted as one of the reasons for poor regression models when using only young tree species for both *E. grandis* and *E. dunnii*.

One of the greatest challenges in predicting forest structural attributes is the species structural and the occurrence of dense forest canopy cover (Gebreslasie *et al*., 2011). It is important to use textural data that is capable of overcoming saturation problems in order to produce better forest structural attribute predictions. The hybrid model is useful and robust in the prediction of intra-species predictions using remotely sensed data. The results of this study show that PLSR is less robust in predicting forest structural attributes in a mixed species environment. The promising results of the combination of spectral and textural information with the PLSR-RF hybrid algorithm is owed to the PLSR and RF methodologies. These two algorithms provide the framework for the integration of spectral information and texture features contrary to traditional linear statistical approaches that necessitate specific assumptions to be met, the PLSR and RF frameworks prove to be robust, versatile and capable of handling remotely sensed data that is complex in nature.

Further research is required into the PLSR-RF hybrid model and how the amount of noise in the RF ensemble affects the predictive accuracy of the model. More research should be done using this PLSR-RF hybrid algorithm coupled with different remotely sensed data sources such as airborne laser scanning data to improve model predictions of forest structural attributes. In order to improve the limitation posed by high spatial resolution imagery and poor spectral capabilities new research should explore the latest generation of satellite sensors with enhanced spectral capabilities as well as advanced spatial properties (Momeni *et al*., 2016). High resolution imagery from WoorldView-2 and WorldView-3 instruments now acquire imagery with eight spectral bands. These enhanced spectral capabilities may prove useful in discriminating forest structural attributes when modelled with machine learning algorithms. Bassa *et al*. (2016) used WoldView-2 image data to evaluate the potential of the oblique random forest (oRF) algorithm to classify a heterogeneous protected area. These authors examined the difference between the traditional RF approaches and suggested that the oRF has slight improvements compared to the traditional RF algorithm because it builds multivariate trees by learning the optimal split using a supervised model. Future studies should be done to establish whether the hybrid method can be improved using the oRF approach instead of traditional RF for the prediction of forest structural attributes.

5. Conclusion

This paper investigated: (i) the performance and strength of three machine learning algorithms (PLSR, RF and PLSR-RF hybrid) using a combination spectral and texture features for model training and validation in predicting various forest structural attributes within a commercial forest plantation. Our results have demonstrated that: (i) the PLSR-RF hybrid model is more robust in predicting volume and height in various *E. dunnii* species when derived from the mature tree species within the plantation (ii) there is great potential for using the PLSR-RF hybrid algorithm with high resolution remotely sensing image data

REFERENCES

Abdi, H., (2007). Partial least square regression (PLSR regression). In: Salkind, N. (Ed.), *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA, 792–795.

Abdel-Rahman, E. M., Ahmed, F. B. and Ismail, R., (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing*, 34, 712–728.

Bassa, Z., Bob, U., Szantoi, Z., Ismail, R. (2016). Land cover and land use mapping of the iSimangaliso Wetland Park, South Africa: comparison of oblique and orthogonal random forest algorithms, *Journal of Applied Remote Sensing*. 10(1), 015017.

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

Definiens Developer. (2012). Definiens Developer XD 2.0.4- *User Guide*, Published by Definiens AG, Bernhard-Wicki-Straße 5, 80636, Munchen, Germany.

Dye, M., Mutanga, O., and Ismail, R. (2012). Combining spectral and textural remote sensing variables using random forests: predicting the age of *Pinus patula* forests in KwaZulu-Natal, *South Africa. Journal of Spatial Science*, 57(2), 193-211.

Gebreslasie, M. T., Ahmed, F. B., and Van Aardt, J. A. (2010). Predicting forest structural attributes using ancillary data and ASTER satellite data. *International Journal of Applied Earth Observation and Geoinformation* 12s: S23-s26

Gebreslasie, M. T., Ahmed, F. B., and Van Aardt, J. A. (2011). Extracting structural attributes from IKONOS imagery for Eucalyptus plantation forests in KwaZulu-Natal, South Africa, using image texture analysis and artificial neural networks. *International Journal of Remote Sensing*, 32(22), 7677-7701.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions* on, 3, 610-621.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proc IEEE*. 1979, 67, 786-804.

Ismail, R., Kassier, H., Chauke, M., Holecz, F., and Hattingh, N. (2015). Assessing the utility of ALOS PALSAR and SPOT 4 to predict timber volumes in even-aged *Eucalyptus* plantations located in Zululand, South Africa, Southern Forests: *Journal of Forest Science* 1-9.

Kuhn, M., and Johnson, K. (2013). *Applied predictive modelling* (600-603). New York: Springer.

Liaw, A., and Wiener, M. (2002) *Classification and regression by random forest*. R News, (2)3, 18–22.

Lillesand, M. T., Kiefer, R. W., and Chipman, J. W. (2008). *Remote sensing and image interpretation* (6th ed.). Hoboken, NJ: Wiley.

Lopatin, J., Galleguillos, M., Fassnacht, F. E., Ceballos, A., and Hernández, J. (2015). Using a Multistructural Object-Based LiDAR Approach to Estimate Vascular Plant Richness in Mediterranean Forests with Complex Structure. *Geoscience and Remote Sensing Letters, IEEE*, 12(5), 1008-1012.

Momeni, R., Aplin, P., and Boyd, D. S. (2016). Mapping Complex Urban Land Cover from Spaceborne Imagery: The Influence of Spatial Resolution, Spectral Band Set and Classification Approach. *Journal* of *Remote Sensing*, 8(2), 88.

Nichol, J. E., and Sarker, M. R. (2011). Improved biomass estimation using the texture parameters of two high-resolution optical sensors. *Geoscience and Remote Sensing, IEEE Transactions* on, 49(3), 930-948.

Næsset, E., Bollandsås, O.M., and Gobakken, T. (2005). Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using LiDAR Data. *Remote sensing of environment*. 94: 541-553.

Owen, D. L. and Van Der Zel, D. W. (2000). Trees, Forests and Plantations in Southern Africa. Forestry Handbook, *Southern African Institute of Forestry*, Menlo Park. Sec 1: 1-6.

R Development Core Team. (2014) R: *A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna. Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shataee, S., Kalbi, S., Fallah, A. and Pelz, D. (2012). Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms, *International Journal of Remote Sensing*, 33:19, 6254-6280, DOI: 10.1080/01431161.2012.682661.

Tuominen S., Haakana M. (2005). Landsat TM imagery and high altitude aerial photographs in estimation of forest characteristics. *Silva Fennica* 39:4, 573-584. http://dx.doi.org/10.14214/sf.367

Vyas, D and Krishnayya, N. S. R. (2014). Estimating attributes of deciduous forest cover of a sanctuary in India utilizing Hyperion data and PLSR analysis, *International Journal of Remote Sensing*, 35:9, 3197-3218, DOI: 10.1080/01431161.2014.903436.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109-130.

Wolter, P. T., Townsend, P. A., and Sturtevant, B. R. (2009). Estimation of forest structural parameters using 5 and 10-meter SPOT-5 satellite data. *Remote Sensing of the Environment*, 113(9), 2019-2036.