A statistical approach for modelling forest structural attributes using multispectral remote sensing data within a commercial forest plantation

by

NICOLE REDDY

211532606

Supervisor: Dr Michael Gebreslasie

Co-supervisor: Dr Riyad Ismail

Submitted in fulfilment of the academic requirements for the degree of Master of Science in the Discipline of Geography in the

School of Agriculture, Engineering and Environmental Sciences

University of KwaZulu-Natal

Westville campus

March 2017

DECLARATION 1 - PLAGIARISM

I, ……. declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a. Their words have been re-written but the general information attributed to them has been referenced
   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: ……
Nicole Reddy

Signed: ……
Dr Michael Gebreslasie

Signed: ……
Dr Riyad Ismail

DECLARATION 2 - PUBLICATIONS

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part and/or include research presented in this thesis (include publications in preparation, submitted, *in press* and published and give details of the contributions of each author to the experimental work and writing of each publication)

Publication 1: A hybrid partial least squares and random forest approach for modelling forest structural attributes using multispectral remote sensing data. In preparation

Publication 2: Texture based ratios derived from high spatial resolution multispectral imagery for prediction of forest structural attributes within a commercial forest plantation. In preparation

Signed: …….

Table of contents

List of figures

List of tables

Abstract

Achieving accurate measurements of forest structural attributes within commercial forest plantations is vital for efficient forest management and understanding commercial forest ecosystem dynamics. There are two ways in which forests may be surveyed to achieve such data; namely traditional field based approaches versus remotely sensed data. Although the former is unbiased and accurate, it does provide challenges due to it being time consuming and expensive. The latter method of using remotely sensed data to achieve such measurements has received increasing attention due to its ability to cover vast areas of land in an efficient manner. The aim of this research was to establish whether or not remotely sensed data (consisting of spectral and textural information) could be used in conjunction with machine learning statistical approaches to estimate forest structural attributes within a commercial forest plantation. This study used four band colour infrared high resolution (0.15m) imagery where the first part of this research used a combination of spectral and textural measures (e.g. homogeneity, dissimilarity and angular second moment) with the partial least squares regression (PLSR) algorithm, the random forest (RF) ensemble and a developed methodology of a PLSR-RF hybrid algorithm to predict forest structural attributes such as tree height basal area and volume within a commercial forest plantation. Once the data were analysed it came to light that the PLSR-RF hybrid outperformed the other two machine learning algorithms with predicting dominant tree height with an 82% accuracy. The latter part of this research focused on whether texture feature ratios could be used to determine forest structural attributes within a commercial forest plantation using the RF ensemble and the PLSR-RF hybrid. Ten texture features were extracted using eCognition software through Grey Level Co-Occurrence Matrices and Grey Level Difference Vector Matrices which were later used to create ratios for the prediction of various forest structural attributes such as tree height, basal area and volume within a commercial forest plantation. The result of the second portion of this study revealed that the RF ensemble algorithm produced the best result for dominant tree height ($R^2 = 0.69$). It was concluded that (i) high resolution imagery can be used to accurately predict forest

structural attributes within a commercial forest plantation using statistical techniques (ii) the developed methodology of the PLSR-RF hybrid algorithm is an effective tool for the prediction of forest structural attributes using high resolution imagery (iii) the RF ensemble is a robust tool for predicting forest structural attributes using texture features.

Acknowledgements

I would like to express my gratitude to the following for supporting me in completing this research:

- First to my fiancé, family and friends, thank you all for your support throughout my studies and for the endless encouragement

- I would like to thank Sappi for providing the field data and imagery used in this study and for opening up their offices and Riverdale plantation to us for the duration of this research

- I would like to thank the National Research Foundation for awarding me with the funding to complete my MSc and for the continued financial support throughout my academic endeavours

- Lastly, I would like to sincerely thank Dr Michael Gebreslasie and Dr Riyad Ismail, for motivating me to do an MSc and for their valuable input as my supervisors... Michael and Riyad, I am very grateful for your unwavering support, guidance and compassion throughout this academic journey. I so appreciated all the time and effort you have put into helping me complete this research and for standing by me through all the tough times, your kindness will never be forgotten.

Chapter one

General Introduction

1.1. Background

South Africa has a dynamic and lucrative commercial forest industry which has grown substantially over the past few decades and now supports significant employment and foreign currency making its longevity and sustainability of key concern (Armstrong et al., 1998; Tewari, 2001). Commercial forest activities for plantations include seeding large areas with exotic plant species which are later harvested and further processed into pulp and paper and other derivatives constituting a total area of 1.28 million hectares of land while generating 1.2% of the country's Gross Domestic Profit (GDP) as a whole (Tewari, 2001; DAFF, 2013).

The use of remotely sensed data over a period of time has the ability to shed light on how a forest system has changed or is changing pre and post tree removal and/or tree growth which then further facilitates sustainable management of such resources (Evans et al., 2006). South Africa's pulp and paper industry is a leading contender in international export markets and has promoted much economic development by creating employment allowing for the increase in the country's GDP (Tewari, 2001; DAFF, 2013). Plantation forests have the potential to be key contributors to the social and economic development of the country by indirectly providing raw materials for mining timber, sawmilling and paper and pulp industries as well as having the ability to provide domestic fuel wood to socially vulnerable areas (Gebreslasie, 2008; Tesfamichael, 2009). Forests not only play vital roles in terms of economic sustainability but they also contribute to regulating temperature and humidity while acting as a natural carbon dioxide absorbent as this is stored in the biomass of forest systems in both the above and below ground structures (Gebreslasie, 2008; Tesfamichael, 2009). Forest plantations may also aid in soil stability and reducing soil surface runoff (Tewari, 2001).

In recent years, measurements in the forestry sector of forest biophysical parameters have become increasingly sought after in the most accurate and timely manner covering vast areas for active application in sustainable forest management and understanding the underlying factors of a forest ecosystem (Popescu et al., 2003; Evans et al., 2006; Tesfamichael et al., 2010; Wulder et al., 2012). This need gave rise to an increase in popularity of collecting data from remote locations by means of remote sensing technology due to traditional field based

approaches being impracticable and inopportune (Tesfamichael et al., 2010; Sarkar and Nichol, 2011; Wulder et al., 2012).

The introduction of remote sensing technology creates a platform to assess and address challenges that the forestry sector faces in terms of sustainable forest management. Optical remote sensing imagery has the ability to provide a synoptic view of places that may be inaccessible while providing coherent continuous coverage to be utilised for feasible and efficient forest resource management (Gebreslasie, 2008; Akay et al., 2009; Tesfamichael et al., 2010; getzin et al., 2010). Remotely sensed data have the potential to provide topography characterizations for the mapping of forest stand parameters such as age, height, density, leaf area index and biomass by using multisensor systems (Popescu et al., 2003; Sarkar and Nichol, 2011; Clementel et al., 2012; Sheridan et al., 2014). Specifically, optical satellite data or aerial photography have proved to be a striking source of data owing to wide ranges of spatial and spectral resolutions which can be manipulated to model forest structural attributes (Nichol and Sarkar, 2011; Wood et al., 2011). Digital aerial photography has many advantages such as (i) its powerful nature that is accompanied by the multispectral nature of the image over conventional photographs (ii) the digital structure of the aerial photographs which allows for a number of geometric corrections allowing for this type of imagery to be directly compatible with geographic information systems for easy interpretation and analysis (iii) this type of imagery has the ability to be radiometrically corrected allowing for the compensation of view angles as well as atmospheric and path radiances effects (Gougeon, 1995). However, it is important to bear in mind that forest structural attributes such as height and volume cannot be measured directly from optical satellite data or aerial photographs hence, studies have explored the notion of remotely sensed image texture characteristics as a reliable representation for forest structural attribute modelling (Gebreslasie et al., 2010; Sarkar and Nichol, 2011; Wood et al., 2011).

A broad literature survey highlighted the potential use of texture features in estimating forest structural attributes by producing promising results in their respective areas of interest with the use of statistical techniques during data analysis (Gebreslasie et al., 2011; Nichol and Sarkar, 2011; Gallardo-Cruz et al., 2012). There are some noteworthy limitations with the use of texture measurements such as the notion that texture is a particularly complex property with significant variance regarding the object of interest, window size selection and physiographic condition (Sarkar and Nichol, 2011). Upon processing texture properties great amounts of data may be

generated with no distinction or indication of which variables of texture or combination of variables for that matter hold the greatest potential relating to a particular study (Champion et al., 2008). High spatial resolution data is said to be the most effective when processing texture variables due to the fact that finer structural details may be discriminated (Fuchs et al., 2009; Sarkar and Nichol, 2011).

Remote sensing data differs from traditional field data approaches in the sense that it provides spatial and temporal data that can be manipulated by machine learning algorithms for classification and regression applications in various case studies. There are several reasons why machine learning algorithms have increasingly been used but some of the main reasons are that: (i) it has the ability to learn features to produce the relevant results required for a particular study from datasets that are disordered in nature, (ii) it has the ability for parameter optimisation by employing a gradient based method for optimising a large array of parameters, (iii) such parameters may be large in numbers and when set appropriately, can select optimal features to save time for data analyses, (iv) the whole set of available data is used for both model training and validation which is an advantage when data is scarce and expensive to obtain.

The present study used the partial least squares regression (PLSR) algorithm, the random forest (RF) machine learning algorithm and a PLS-RF hybrid approach. Machine learning algorithms such as PLSR is robust in nature and was developed by Wold (1995). PLSR is able to overcome problems such as over-fitting and multicollinearity (Vyas and Krishnayya, 2014). PLSR is able to compress data which allows for a reduction in a large number of variables. The PLSR method is an iterative process and produces a number of models with the aim of finding a few partial least square latent variables which are able to explain as much of the variation in both the explanatory and response variables as possible (Oumar et al., 2013).

Another powerful machine learning algorithm that has made consequential gains for promising predictive potential is the RF algorithm which uses recursive binary partitioning based on the classification and regression tree (CART) ruleset and is an ensemble learning tool that benefits from random subspace selection and bagging (Breiman, 2001; Adam et al., 2013; Abdel-Rahman, 2013). Machine learning techniques in general have been used extensively to estimate plot-level volume per hectare and basal area per hectare where some authors tend to compare and contrast various machine learning approaches to distinguish which algorithm works best with a certain type of data set. This approach has been done with varying success however, within the context of this study the aim of this research was to determine whether image data

of a spatial and spectral nature can be used to determine forest structural attributes within a commercial forest plantation using high resolution (0.15m) remotely sensed imagery and machine learning statistical techniques.

This research was conducted with the following objectives:

1. To assess the capability of a combination of spectral and textural features extracted from high resolution imagery to predict forest structural attributes within a commercial forest plantation
2. To test the ability of the PLSR, RF and PLSR-RF models in predicting forest structural attributes such as basal area, tree height and volume
3. To assess the capability of using only texture features extracted from multispectral data in predicting forests structural attributes within a commercial forest plantation
4. To offer recommendations for the use of remotely sensed measures of texture characteristics from high resolution imagery for commercial forest plantation management

1.2. Outline of thesis

This thesis is presented in four chapters and is structured around two core chapters namely; chapter two and three. These two chapters are in the form of publishable papers and will be submitted to peer reviewed journals. Both chapters each have major sections that deal with their own study area, literature review and methodology and for that reason theses sections have not been covered in the introductory section of the thesis to avoid repetition.

Chapter two will assess a hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data. The PLSR and RF models will be used to develop a hybrid methodology to be used in conjunction with a combination of spectral and textural features derived from high resolution optical imagery for the prediction of forest structural attributes within a commercial forest plantation.

Chapter three will ascertain whether the ratios of texture features derived from high resolution imagery can determine forest structural attributes within a commercial forest plantation. The RF ensemble and the PLSR-RF hybrid models will be used to compare and contrast whether texture features are in fact a plausible proxy for forest structural attribute estimations.

Chapter four provides a conclusion to the study where the aims and objectives of this research are discussed in detail, highlighting the key findings from this study. Finally, the concluding chapter will also address limitations and challenges experienced throughout this study with useful recommendations for future research endeavours of this nature.

Chapter two

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

Nicole Reddy[1], Riyad Ismail[1&2] and Michael Gebreslasie[1]

[1]School of Agriculture, Engineering and Environmental Sciences, University of KwaZulu-Natal, Westville, Durban

[2]SAPPI Forests (Pty) Ltd, 17 Montrose Park Boulevard, Victory Country Club Estate, Pietermaritzburg

2.1. Abstract

To ensure sustainable planning and management within a commercial forest plantation, forest inventory data that is up to date has become increasingly essential and necessary. Data for these measurements 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 three machine learning algorithms (partial least squares regression (PLSR), random forest (RF) and a PLSR-RF hybrid) 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. When using the combined species data (*E. grandis* and *E. dunnii*) all three machine learning algorithms returned statistically weak accuracies. Thus, owing to the various inter species characteristics, coefficient of determination values further improved when the species were separated into individual *E. grandis* and *E. dunnii* species across all ages with the highest accuracy being reported for *E. dunnii* species using the RF algorithm for volume (R^2 = 0.57 and RMSE = 61.31tons/ha). These individual species data were further partitioned according to young and mature *E. grandis* and *E. dunnii* species due to the intra species characteristics that affect machine learning algorithm predictions. 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 = 1.89m). 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.

Keywords: PLSR, partial least squares regression, random forests, RF, forest attributes, volume, tree height, DBH, texture, age, basal area

2.2. 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). Traditional field based forest inventory approaches are time and labour consuming. Remotely sensed data has demonstrated the potential to map forest structural attributes such as tree age, tree diameter at-breast-height (DBH), tree height, stems per hectare (SPH), tree volume, leaf area index and biomass using passive and active remote sensing systems (Popescu et al., 2003; Gebreslasie, 2008; Gebreslasie et al., 2011; Sarkar and Nichol, 2011; Dye et al., 2012; Clementel et al., 2012; Sheridan et al., 2014; Dube et al., 2014; Ismail et al., 2015).

The age of a forest is an important component of forest inventory missions because it serves as a valuable indicator on a number of forest conditions. Forest structural attributes such as tree height and basal area are often related to the age of a forest and can be used as a surrogate in estimating these attributes when combined with remote sensing data. Remotely sensed data in the form of high resolution aerial imagery reflects landscapes in a two-dimensional space and holds the promise of achieving better estimations with higher resolutions (Getzin et al., 2012). The use of aerial imagery for forestry analysis has been noted as a very cost effective alternative to traditional field based approaches of forest inventory and is used by most forest administrators for the recurrent monitoring of forest resources and yields in commercial forest plantations (Pasher and King, 2010; Gebreslasie et al., 2011 Getzin et al., 2012). Remote sensing data provides spectral and textural information that can be manipulated by machine learning algorithms for prediction and classification applications in various case studies. Machine learning algorithms have increasingly been used for many reasons but the main reasons are that: (i) it has the ability to learn features to produce relevant results required for a particular study from datasets that are disordered in nature and (ii) it has the ability for parameter optimisation by employing a gradient based method for optimising a large array of parameters. Machine learning algorithms such 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; Adam et al., 2013; Abdel-Rahman, 2013; Dube et al., 2014). RF is a powerful algorithm that has made consequential gains for its promising predictive potential. In general machine learning techniques have been used extensively to estimate forest structural attributes using remote sensing data where some authors tend to compare and contrast various machine learning approaches such as Shataee et al. (2011) who 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 authors 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 $R^2 = 0.68$. More recently, Dube et al. (2014) utilized vegetation indices derived from RapidEye imagery with stochastic gradient boosting and the RF algorithm to produce an overall predictive accuracy of $R^2 = 0.80$ and $R^2 = 0.79$ respectively for various *Eucalyptus* species within a commercial forest plantation.

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 1995; Oumar et al., 2013). It is able to compress data which allows for the reduction in a large number of variables that are collinear allowing for the development of a few non-correlated latent variables also known as factors/components (Vyas and Krishnayya, 2014). Wolter et al. (2005) 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 that 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 done by Næsset et al. (2005) used a combination of ordinary least squares (OLS), seemingly unrelated regression (SUR) and PLSR to estimate forest structural attributes (height, mean diameter, volume, basal area etc.) of a forest stand using laser scanning technology and produced a best overall R^2 value of 0.94.

PLSR and RF have demonstrated powerful modelling potentials and with that background 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.

2.3. Materials and Methods

2.3.1. Study site

The study was conducted at the Sappi Riverdale plantation located West of the town of Richmond in the Midlands region of KwaZulu-Natal, South Africa, 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 and drains into the ocean between Park Rynie and Amanzimtoti on the Natal South Coast. This area consists of moist grassveld streams and includes a mix of veld types ranging from Ngongoni veld of the Natal Mistbelt (40%), the Highland Sourveld (30%) and Southern Tall Grassveld (20%). The terrain is low and mountainous with undulating hills consisting of a geology made up of a mix of mudstone, sandstone, tillite, ampholite and basalt with mostly sandy-clay and sandy-clay loams. 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* (43%) and *Eucalyptus grandis* (57%) with an average age of seven years and an average height of 23.5m. The *Eucalyptus* species are rapid growing species that are planted with seedlings or clones and is set to be harvested every six to ten years (Owen and Van Der Zel 2000).

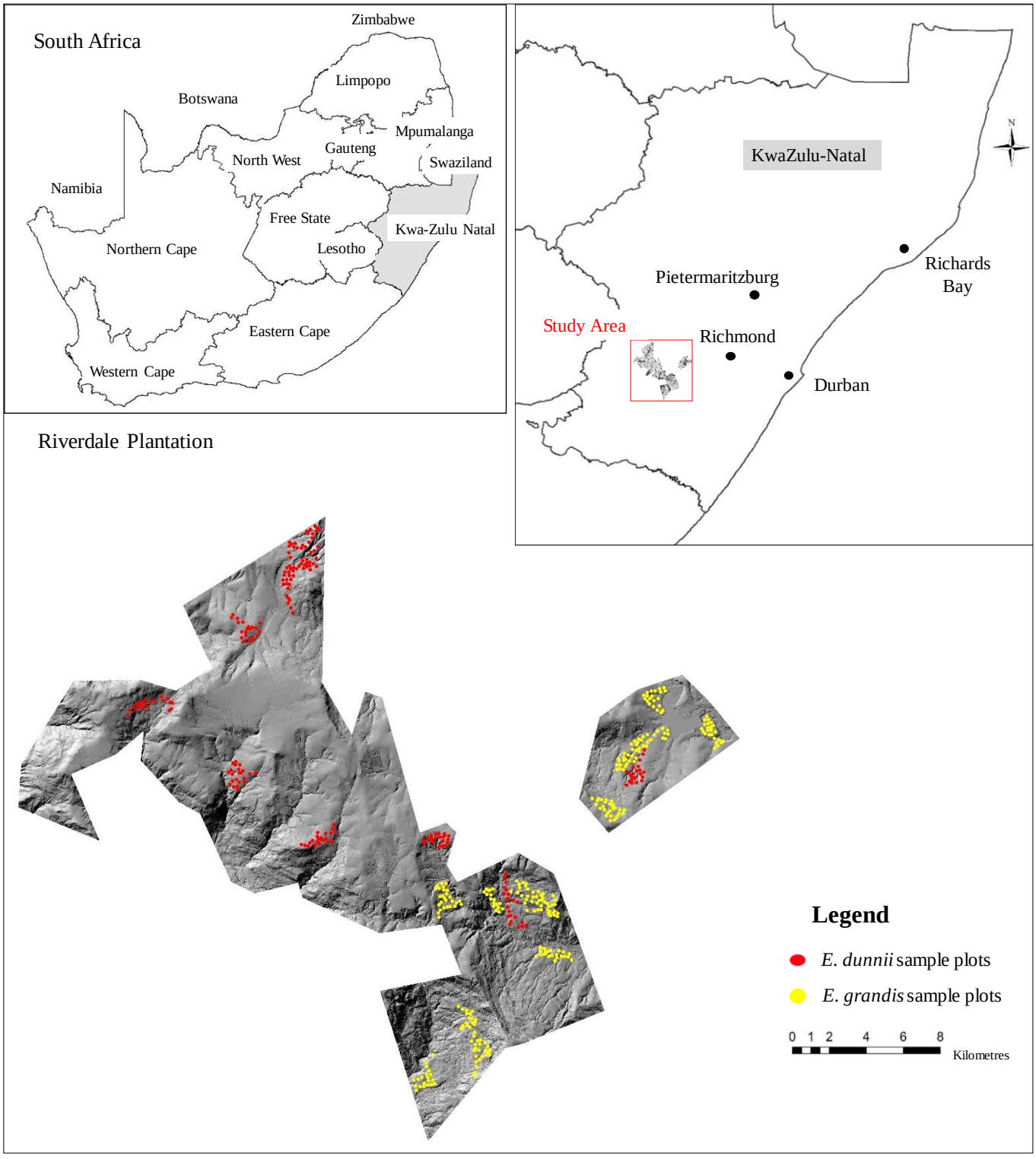

Figure 1: Location of the study area indicated as the Sappi Riverdale plantation located in the KwaZulu-Natal Midlands region (South Africa)

2.3.2. Field data

The field survey campaign was conducted between the 12th of April and the 22nd May 2014 at the Sappi Riverdale plantation near Richmond, KwaZulu-Natal, South Africa, using industry standard enumeration techniques (Owen, 2000). 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 and dominant tree height) and basal area (Baha) were measured for each circular plot. *Eucalyptus dunnii* and *Eucalyptus grandis* species that were between two and ten years (Figure 2) were considered for this study. The tree height and DBH for each plot was measured using the Vertex IV laser instrument and Haglof Digitech Calliper, respectively.

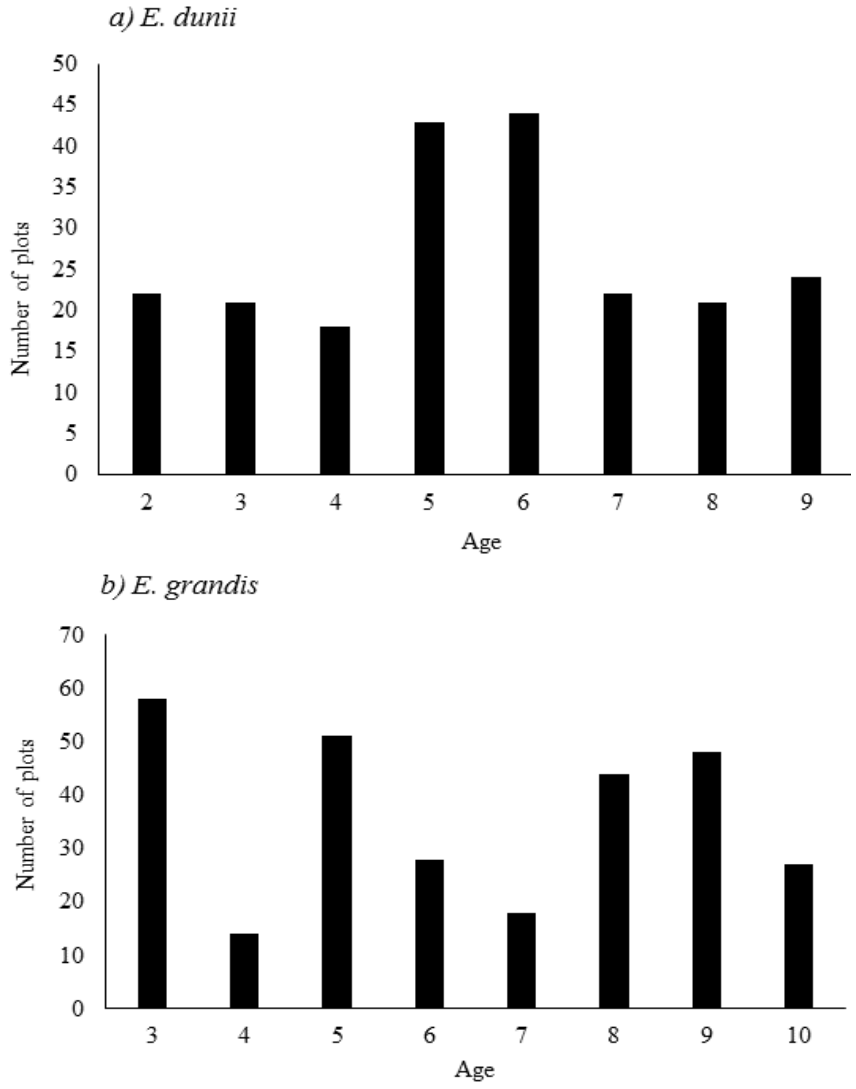

Figure 2: Age distribution of the a) *E. dunnii* species (n = 214) and b) *E. grandis* species (n = 288) located in the study area

2.3.3. 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 processed by the service provider using a full photogrammetric processing procedure with customised in-house photogrammetry suites and Geomatica 10.1 photogrammetry software. Global positioning system data and inertia measurements were incorporated into the photogrammetry process. Additionally, all imagery was 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). The four band colour imagery is multispectral in nature indicating that it was collected from several parts of the electromagnetic spectrum. This image data contains information from the red, green, blue and near infrared bands (Table 1). The four band colour infrared imagery is particularly common with studies involving crop inventory or yield estimates (USDA Forest Services, 2008) such as the present study. This imagery is mostly used when the health of a yield is in question resulting in the NIR band becoming important as well as the green wavelength where chlorophyll in plants are reflected (Dye et al., 2012). This four band imagery has the ability to penetrate atmospheric haze better than most natural colour imagery which results in sharper imagery for analysis (Paine and Kiser, 2003).

Table 1: Showing the four bands of colour infrared with their associated image band configuration

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

2.3.4. Texture feature extraction and multiresolution segmentation

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) (Kayitakire et al., 2006). 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, 1979). The present study used a first principal component to extract texture features in Definiens eCognition software version 9.0 which added a new dimension to this study by performing a multiresolution segmentation (MRS) of the individual stand plots without selecting a designated moving window size, contrary to previous studies that have used texture and an optimum window size (Chan et al., 2003; Kayitakire et al., 2006; Dye et al., 2008; Dye et al., 2012; Blaschke, 2010; Gebreslasie et al., 2011; Nichol and Sarkar, 2011; Mhangara and Odindi., 2013; Dube et al., 2014, Wang et al., 2015). The MRS process relies on the heterogeneity of pixels within and imagery that are adjacent to each other. The main parameters to consider for this particular process are the shape and compactness of the pixels within the image being manipulated (Definiens Developer, 2012). Various criterion combinations were tested and visually inspected to determine which values best represented the study area by delineating the canopy cover of the commercial forest stands within the 25 measured compartments. Through a process of trial and error the ideal criterion values for shape and compactness were set to 0.1 and 0.5 respectively which allowed for the canopy cover to be accurately discriminated for the extraction of the texture features within the individual stand plots. The texture features (Appendix A) 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. Texture features were extracted from four directions (0°, 45°, 90°, 135°) thus to achieve directional invariance on each individual stand plot the four directions were summed together providing results for ‘all directions’ which represented the optimal results pertaining to forest structural attribute prediction for this study.

2.3.5. Partial least squares regression (PLSR)

PLSR is a linear statistical method used to model the relationship between two data sets. It 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; Ramírez et al., 2010; Abdel-Rehman et al., 2014). All partial least square (PLS) methods work on the underlying assumption that

the observed data is generated by a process led by a small number of latent variables. 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 (Sampson et al., 2011; Lopatin et al., 2015; Belgiu and Drăgut, 2016). This PLSR model not only considers the variance of the sample set within the data but it also considers the class labels that are used within the predictors. Each block of observed variables can be summarised by latent variables consisting of linear relationships that exist between them. Unlike PCA, PLSR accounts for the covariates explicitly during the process by calculating the latent variables between the *x* and *y* variables (Sampson et al., 2011; 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).

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm} X_j \qquad (1)$$

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

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \qquad i = 1,\ldots.n \qquad (2)$$

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. Equation 1 and 2 adapted from Sampson et al. (2011) and Lopatin et al. (2015).

PLSR is able to retain the number of factors/components that are essentially necessary to maximise linear relationships with the response variables. Oumar et al. (2013) suggested that the first few latent variables produced by the PLSR are the ones that usually explain most of the correlation in the dataset. In this study the PLSR library was implemented from 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 using the R statistical software.

2.3.6. Random forest (RF)

In recent years the RF algorithm has demonstrated consequential gains in prediction and classification accuracies (Breiman, 2001). The RF ensemble is a vital improvement to the original classification and regression (CART) tree method as well as to the statistical methods for the reduction in predictive variance known as bagging or bootstrap aggregation (Breiman et al., 1982; Breiman, 1996). The basic idea behind RF 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 deals with an ensemble of trees and has evolved into an important non-parametric method that has capabilities to fit interactions that may be highly non-linear. It may also be used to deal with irrelevant predictor variables and robust outliers in the predictor variable list as suggested by Cutler et al. (2009). 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. RF seeks to add randomness into the data by selecting random subsets of input variables to establish the most efficient split at each tree node thus reducing bias and maintaining diversity within the data since no pruning is performed (Breiman, 2001). The RF model uses the remaining 34% of the data known as the out-of-bag (OOB) data for the model prediction. The ensemble predicts the data using the difference in the mean square error of the OOB data and the data that is used to grow the homogeneous regression trees. The RF regression algorithm was implemented using a package suggested by Kuhn et al. (2015) called ‘caret’ and the ‘randomForest’ package (Liaw ad Wiener, 2002) in the R statistical software version 3.2.2 (R Development Core Team, 2014). One of the main advantages of the RF algorithm is that it only has two main tuning parameters, i.e. the number of input variables (*mtry*) and the number of trees (*ntree*) therefore the algorithm is not that difficult to implement. In this study a default *ntree* value of 500 was used because most of the errors reach stability before this number of regression trees is reached (Lawrence et al., 2006). The RF ensemble method based on decision trees is more robust than single decision tree models and is able to

produce promising results compared to traditional methods of regression (Ramírez et al., 2010). During model development resampling was done to ensure the best predictive performance of the model thus the model was tuned with a 10-fold cross validation technique using R statistical software.

2.3.7. Partial least squares-random forest (PLSR-RF) hybrid

The PLSR-RF hybrid model involves the random forest non-parametric methodology with the addition of linear PLSR approach. Once this is done, the process is repeated and optimised. The PLSR part of the hybrid creates factors from the explanatory *(x)* variables that are the most relevant for the response *(y)* variables. The *x* and *y* variables are decomposed into latent structures in an iterative process where both structures of the *x* and *y* variables are considered. The PLSR model extracts the scores of the vectors which serve as new predictor representations and regresses the response variables on these new predictors thus creating components. Following this process, the RF method works with 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. RF is considered to be an ensemble algorithm that is robust in nature and does not over fit the data due to the random selection of explanatory variables that are selected from a large dataset (Bassa et al., 2016). Therefore, the PLSR-RF hybrid uses the PLSR methodology of latent variables and converts them into components which is then put through the RF binary recursive partitioning method to sift out the optimal components for prediction. During this process the bagging decision tree principle is applied to the selected components from the PLSR method. During model development the PLSR-RF hybrid selects 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 (Figure 2). Each tree node is randomly selected from the subset of components for splitting at each of these nodes further producing the result of an optimal component and variable for predictive purposes with hyperparameters listed as *ncomp, mtry* and *ntree*. 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.

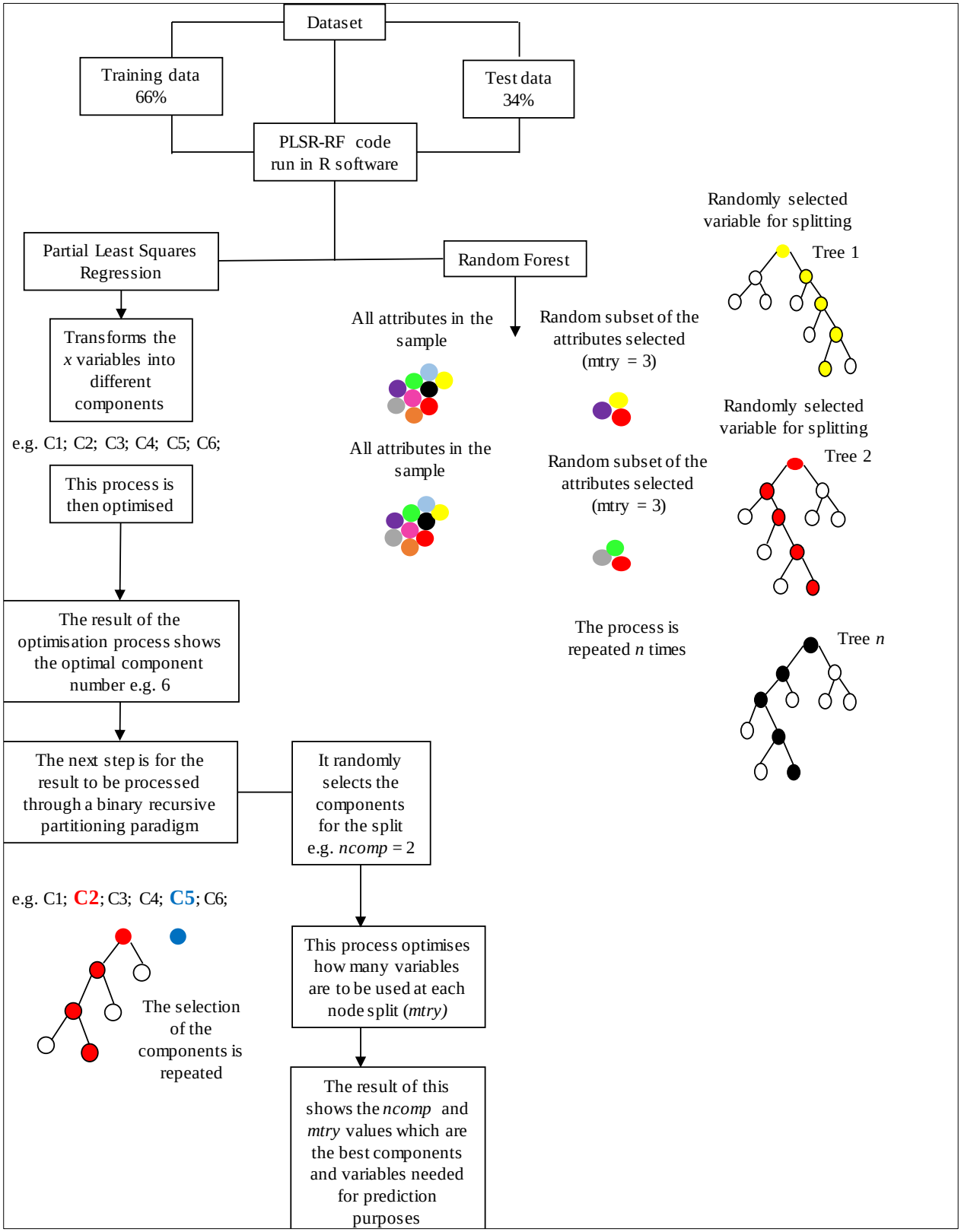

Figure 3: A graphical representation of how the PLSR-RF hybrid model works

2.3.8. 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). The unbiased validation was achieved by performing a uniform stratified 10-fold cross validation (CV) which was applied to all models to ensure the models' true predictive performance was recorded. Kohavi (1995) recommended this type of CV as the best for ensuring fair model prediction. The "stratified" description in this CV refers to the folds that were created based on the full range of values of the response variables and was done in this manner to ensure that each fold was a good representation of the whole dataset. Additionally, the independent validation was calibrated with parameter tuning being done on a training dataset so that the test data remained "unseen" in each of the 10 CV folds to ensure no bias and error came to the final model development results.

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).

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

$X_{measured}$ 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.

2.4. Results

This study used a combination of spectral and textural features with three machine learning models; PLSR, RF and a PLSR-RF hybrid to predict *Eucalyptus* forest structural attributes within a commercial forest plantation. A combination of *Eucalyptus* species across 25 compartments within the commercial forest plantation were used as inputs dataset resulting in the three statistical algorithms returning statistically weak predictive accuracies with R^2 values ranging between 0.08 and 0.54. The RF method produced the best results for all forest structural attributes followed by the PLSR-RF hybrid model (Figure 4).

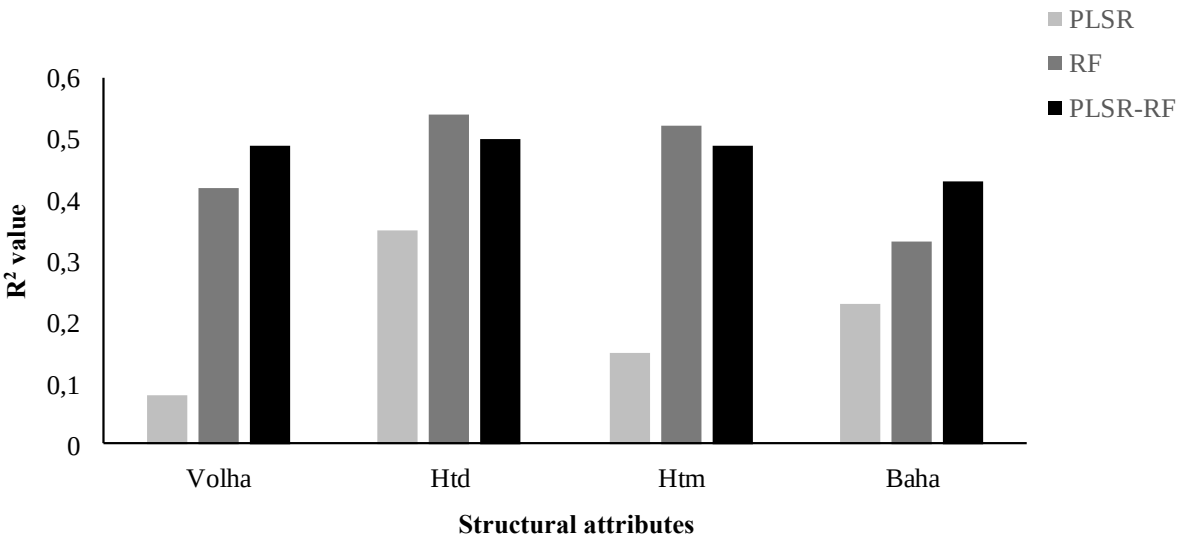

Figure 4: Model predictions using a combination of *E. grandis* and *E. dunnii* species across three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features within a commercial forest plantation for volume (volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

The hyperparameter optimization results for PLSR shows that the optimal number of components was found between one and 10 (Table 3). The RF model suggests that all models (Volha, HtD and Htm) used less than 10 variables for node splitting (*mtry*) and the PLSR-RF results showed that all the models used more than 10 variables for node splitting. The size of all the ensembles (*ntree*) for RF and PLSR-RF was consistent at 500 trees. Models with higher values of the *mtry* and *ntree* parameters did not increase training dataset accuracies across all forest attributes.

Table 2: Optimised *ncomp, mtry* and *ntree* obtained when using various machine learning algorithms and a combination of *E. grandis* and *E. dunnii* species for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

	PLSR	**RF**		**PLSR-RF**		
Forest attribute	*ncomp*	*mtry*	*ntree*	*ncomp*	*mtry*	*ntree*
Volha	1	7	500	10	7	500
HtD	9	12	500	13	9	500
Htm	2	6	500	13	8	500
Baha	10	7	500	10	5	500

Following the poor performance of the combined species (n =502) in predicting forest structural attributes, the data was further partitioned according to the individual *Eucalyptus* species. The *E. grandis* (n = 288) and *E. dunnii* (n = 214) were processed as separate input data. The PLSR model performance for *E. grandis* was poor but improved when predicting *E. dunnii* forest structural attributes. For the *E. grandis* species the RF model produced promising results for dominant tree height (R^2 = 0.47 and RMSE = 3.52m) however the PLSR-RF algorithm produced the best models for volume (R^2 = 0.49 and RMSE = 38.58tons/ha) and dominant tree height (R^2 = 0.50 and RMSE = 2.37m) (Figure 5a). Using the *E. dunnii* species as the input data the RF model produced comparative results when predicting volume (R^2 = 0.55 and RMSE = 62.61tons/ha). However, the best overall model results for *E. dunnii* species were produced using the PLSR-RF model for mean tree height (R^2 = 0.57 and RMSE = 1.93m) and volume (R^2 = 0.57 and RMSE = 39.55tons/ha) (Figure 5b).

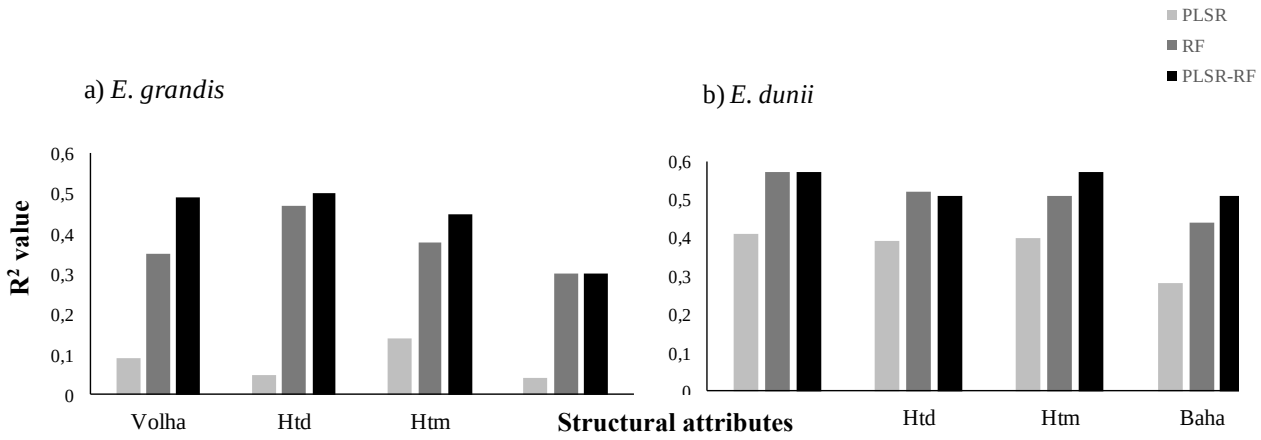

Figure 5: Model predictions using individual tree species (a) *E. grandis* and (b) *E. dunnii* across three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha).

The hyperparameter optimization results when using the individual *Eucalyptus* species shows that for the *E. grandis* species all models except for basal area used seven as the optimal component number for the PLSR method (Table 3). For the RF methodology, volume, dominant and mean tree height used less than 10 variables for node splitting (*mtry*). The PLSR-RF hybrid algorithm shows that the optimal components (*ncomp)* selected for this study were greater than 7 with less than 10 variables being used for node splitting. Table 3 shows that three of the models (Volha, HtD and Htm) used for *E. dunnii* species selected the optimal component number to be six when using the PLSR methodology. RF models used a total of less than 10 variables for node splitting (*mtry*). The PLSR-RF methodology selected the optimal components that were greater than 10 accept for volume. The models used for node splitting were all less than nine for the structural attributes (Volha, Baha, Htd and Htm).

Table 3: Optimised *ncomp, mtry* and *ntree* obtained when using the various machine learning algorithms and the *E. grandis* and *E. dunnii* species individually for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha). A default *ntree* value (*ntree* = 500) was used for all models.

	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
DBHq	6	3	500	12	8	500

This study further looked at mature and young forests of the two species individually. Young *E. grandis* species were grouped at ages three to six years (Figure 6a) and mature *E. grandis* species were grouped at seven to 10 years (Figure 6b). 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 6a). 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.

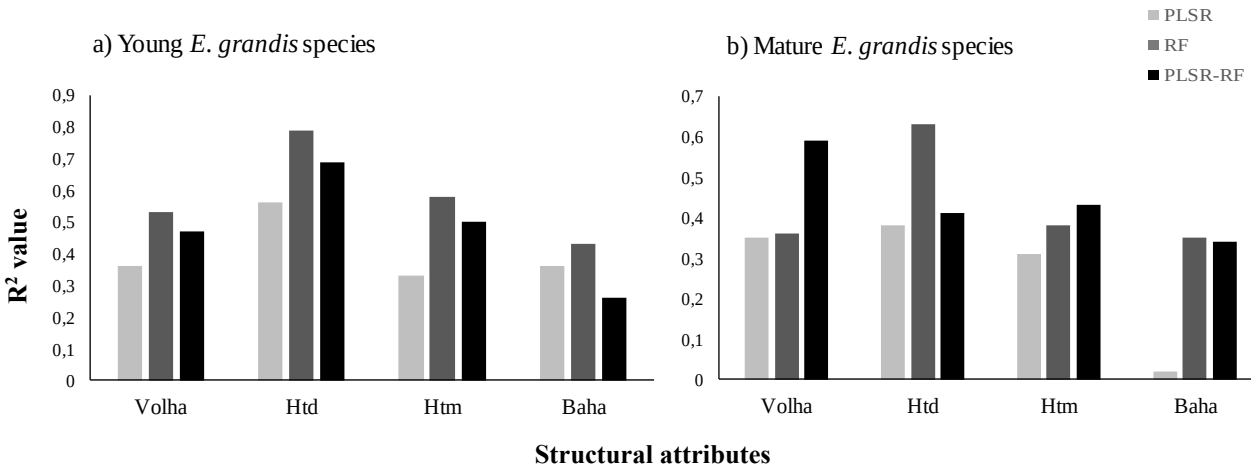

Figure 6: Model predictions using only *E. grandis* species that are (a) young and (b) mature across three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal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

The hyperparameter optimization results when using the young *E. grandis* species (Table 4) shows that all models selected five as the optimal component number except for volume. Using the RF methodology the variables used for node splitting (*mtry*) was below six for all models (Volha, HtD, Htm and Baha). The PLSR-RF hybrid algorithm shows the optimal components to be less than 10. Table 4 also shows that all models used for *E. grandis* mature species selected the optimal components greater than 10 with the exception of basal area when using the PLSR methodology. RF models used a total of less than 7 variables for node splitting (*mtry*) with the exception of mean tree height. The PLSR-RF methodology selected the optimal components that were greater than seven with all the models using less than 10 variables for node splitting (Volha, HtD, Htm and Baha).

Table 4: Optimised *ncomp, mtry* and *ntree* obtained when using the various machine learning algorithms and young and mature *E. grandis* species for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha). A default *ntree* value (*ntree* = 500) was used for all models.

	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	1	2	500	10	4	500

When using an age partition for *E. dunnii* species accuracies continued to improve across all forest attributes. Unfortunately for young (Figure 7a) *E. dunnii* species PLSR could not explain more than 39% of the variation. However, when using the RF model for 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). The PLSR-RF model produced the highest accuracies for the young *E. dunnii* species for basal area (R^2 = 0.55 and RMSE = 2.93ha) and dominant tree height (R^2 = 0.65 and RMSE = 1.85m). For mature *E. dunnii* species the best model accuracy was reported for dominant tree height across all three machine learning algorithms; 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). Basal area did not produce accuracies above 31% using mature *E. dunnii* species across all three machine learning algorithms.

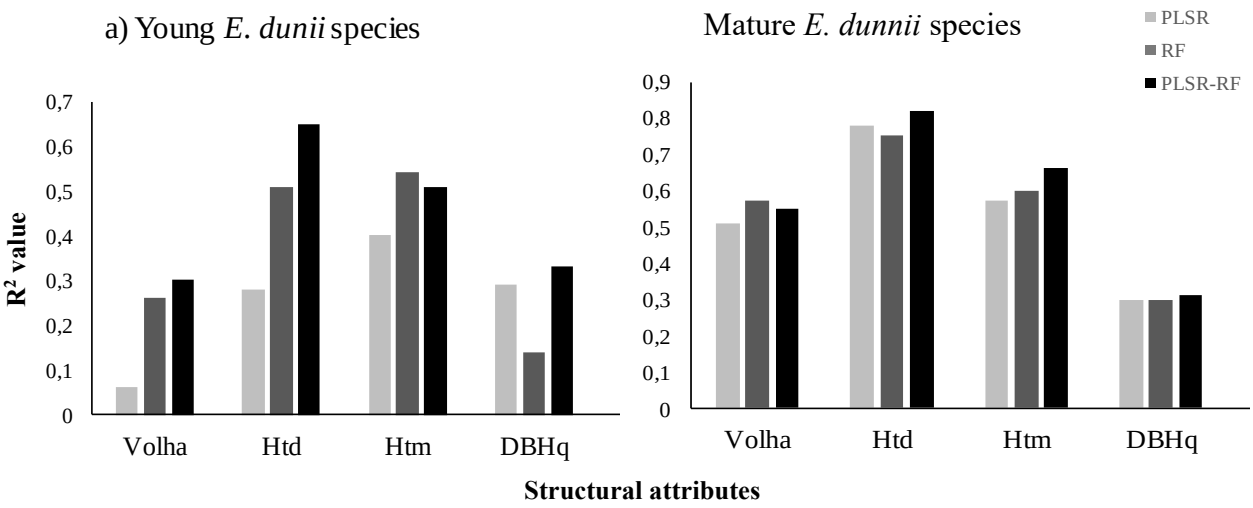

Figure 7: Model predictions using only *E. dunnii* species that are (a) young and (b) mature across three machine learning techniques i.e. PLSR, RF and PLSR-RF hybrid using a combination of spectral and textural features for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

Hyperparameter optimization results when using the young *E. dunnii* species (Table 5) shows that all models selected optimal component numbers less than eight when using the PLSR methodology. Using the RF methodology the variables used for node splitting (*mtry*) were below six with the exception of basal area. The PLSR-RF hybrid algorithm shows the optimal components to be greater than seven with less than 10 variables being used for node splitting. Table 5 also shows that the models used for mature *E. dunnii* species selected the optimal component number six with the exception of volume and dominant tree height when using the PLSR methodology. RF models used variables between eight and 16 for node splitting (*mtry*) with the exception of mean tree height. The PLSR-RF methodology selected the optimal components that were between three and 11 with all the models using less than 10 variables for node splitting except volume.

Table 5: Optimised *ncomp, mtry* and *ntree* obtained when using the various machine learning algorithms and young and mature *E. grandis* species for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha). A default *ntree* value (*ntree* = 500) was used for all models.

	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

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 8a and b) for dominant tree height followed by the RF algorithm producing the best result for the young and mature *E. grandis* species (Figure 8c and d) for dominant tree height. The prediction value of 0.82 suggests that the PLSR-RF regression equation could significantly predict dominant tree height when using mature *E. dunnii* species as displayed in the predicted vs observed graphs in Figure 8.

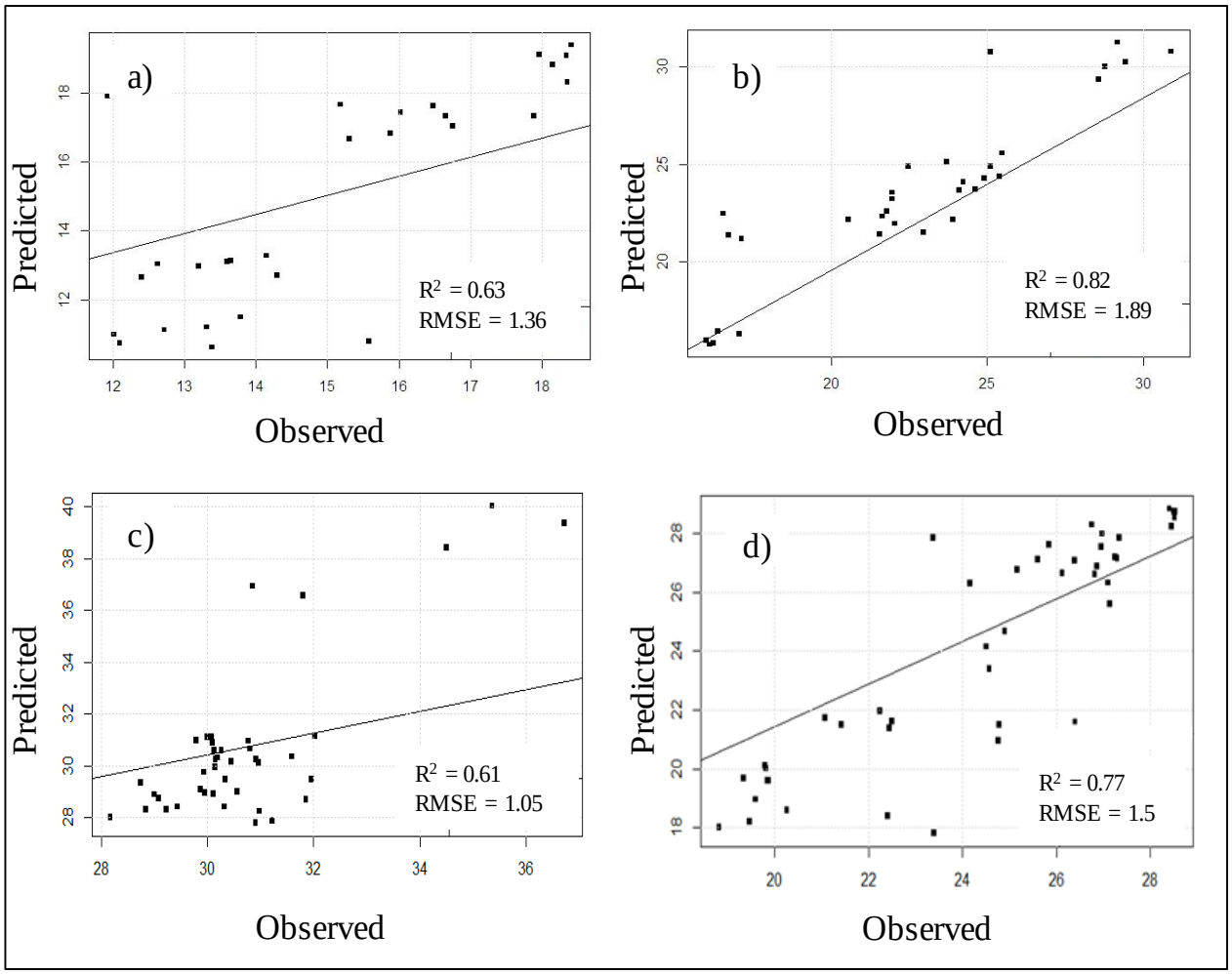

Figure 8: Observed vs predicted graphs for the best models produced in this study. All models produced the highest accuracies for dominant tree height (HtD) a) young *E. dunnii* b) mature *E. dunnii* c) young *E. grandis* d) mature *E. grandis*

2.5. Discussion

2.5.1. Combination of spectral and texture variables

We built our predictive models using a combination of spectral and texture features across three machine learning algorithms: Partial Least Squares Regression (PLSR), Random Forest (RF) regression and a PLSR-RF hybrid approach. Our study used high spatial resolution imagery with the main advantage being its ability to discriminate forest features. Despite this advantage, high resolution imagery tends to have limited spectral capabilities (Momeni et al., 2016). The image used for this study had four (visible and near infrared) spectral bands. Spectral information only has one dimension of diversity whereas the use of texture measures has many dimensions. In our study the 502 candidate plots used to build the predictive models varied in structural characteristics and species composition. A texture analysis is often used to introduce the underlying spatial information/variation within an image (Puissant et al., 2005; Gallardo-

Cruz et al., 2012). In the case of this study texture features measure the forest structural attributes using the distribution of grey levels among the neighbouring pixels in the multispectral image (Haralick et al. 1973). Gallardo-Cruz et al. (2012) highlighted the importance of introducing texture feature information as the spatial resolution increases. This could be due to the fact that texture features have an enhanced spatial relationship of the pixels within the high resolution image (Sarkar and Nichol, 2011). Our study combined the benefits of spectral and textural features to predict forest structural attributes.

Studies have demonstrated that the inclusion of textural features to spectral features could improve predictions of forest structural attributes (Wulder et al., 1998). From the results of our study it is suggested that height was the most accurate variable to be predicted using the combined spectral band information and texture features as suggested by Kayitakire et al. (2006) who reported a similar trend. Wang et al. (2015) and Johansen et al. (2007) noted a similar trend in their study where the inclusion of texture feature information improved model accuracies for Johansen et al. (2007) by 19%. Nichol and Sarkar (2011) further suggested the importance of texture and spectral information as input variables for model prediction and showed great potential for biomass prediction with the best R^2 of 0.91 being achieved using the average texture and spectral band information of two remote sensing sensors.

2.5.2. Model development and accuracies for inter and intra species characteristics

Our study consisted of two *Eucalyptus* species with different age ranges. The age of a forest is an important component of forest inventory missions because it serves as a valuable indicator on a number of forest conditions (Dye et al. 2012). Forest age is considered to have a useful effect on image texture because structural changes in the forest stands cause a variation in the texture of the image. Using a combination of the *Eucalyptus* species (*E. grandis and E. dunnii*) all three machine learning algorithms returned statistically weak accuracies due to the inter-species characteristics. The PLSR method did not perform well during model development for prediction purposes of various forest structural attributes. The highest reported model accuracy for PLSR was for dominant tree height (R^2 = 0.33 and RMSE = 5.54m) followed by basal area (R^2 = 0.23 and RMSE = 0.23ha). Following the poor performance of the PLSR algorithm, models were developed using the RF algorithm with *E. grandis* and *E. dunnii* combined. Accuracies showed promising improvement when compared to the PLSR with the highest model accuracy being reported for dominant tree height (R^2 = 0.54 and RMSE = 4.74m). The

poor performance of the RF model may be attributed to high input variable inconsistency and the presence of mixed species of *Eucalyptus* species planted in homogeneous compartments within the commercial forest plantation and not necessarily due to algorithm strength (Dye et al., 2012; Dube et al., 2014; Wang et al., 2015). In addition to the results obtained for the PLSR and RF algorithms individually our study demonstrated the strength and applicability of a hybrid algorithm that has been 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. The PLSR-RF hybrid produced comparable results for dominant tree height (R^2 = 0.50 and RMSE = 4.89m) but did not outperform RF in this case as hoped.

In an attempt to further improve model accuracies across all three machine learning algorithms, the data set was split into individual species of all ages ranging from two to 10 years. When using the PLSR method for *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 accuracies improved by up to 40% with the highest being 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 a volume 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 there was a notable improvement when the species were processed separately compared to the combined species results which 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. High R^2 values for young and mature *E. grandis* species was produced using the RF methodology because a large number of decision trees are grown to ensure that the data is not over fitted. Therefore, bias is low because of the random predictor selection (Prasad et al 2006) which explains why the RF model performs better in some cases.

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., 2012). 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 taxonomic differences as well as the occurrence of dense forest canopy cover (Dube et al.,

2014). It is important to use spectral and spatial 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. This study highlights that spatial resolution plays a crucial part in prediction studies hence more information is needed to establish the optimum pixel size for texture extraction and machine learning statistical techniques. 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 WorldView-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 touched on the difference between the traditional RF approach 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.

2.6. 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 high resolution (0.15m) remotely sensing image data for texture extraction using a PCA to reduce data redundancy.

Chapter three

Texture based ratios derived from high spatial resolution multispectral imagery for the prediction of forest structural attributes within a commercial forest plantation

Nicole Reddy[1]

1. School of Agriculture, Engineering and Environmental Sciences, University of KwaZulu-Natal, Westville, Durban

3.1. Abstract

Accurate forest measurements collected in a timely fashion have become increasingly desirable across large areas such as commercial forest plantations for sustainable forest management and understanding ecosystem dynamics within this type of environment. Data may be collected via field based approaches and remote sensing techniques. Although the field based approaches are typically unbiased it is very time consuming and expensive hence the appeal of remote sensing technology. This study aims to examine if texture based ratios derived from high resolution imagery (of various *Eucalyptus* species) can be used to predict forest structural attributes (basal area, volume, dominant tree height and mean tree height) within a commercial forest plantation using two machine learning algorithms; random forest (RF) and a partial least squares regression-random forest (PLSR-RF) hybrid approach. The commercial forest plantation contained two species of *Eucalyptus* known as *E. dunnii* and *E. grandis* where the data was later portioned into young and mature forests. When using young *E dunnii* species the predictive accuracies of the forest structural attribute estimations did not exceed 39% however, these model accuracies significantly improved when the mature *E. dunnii* species were used. The PLSR-RF algorithm for mature *E. dunnii* species produced the best result for dominant tree height (R^2 = 0.60 and RMSE = 2.78). However, for the mature *E. grandis* species the PLSR-RF did produce the best result for basal area (R^2 = 0.42 and RMSE = 3.82m). The best

overall model for this study was derived from the young *E. grandis* species using the RF model (R^2 = 0.69 and RMSE = 2.18m) for dominant tree height. These results demonstrate the great robustness and potential for using the RF and PLSR-RF algorithms for the prediction of forest structural attributes within a commercial forest plantation using texture based ratios as the input data set and brings attention to various inter and intra species characteristics that may affect predictive accuracies within a commercial forest plantation.

Key words: random forests, RF, PLSR, partial least squares regression, forest attributes, volume, tree height, DBH, texture, age, basal area, regression

3.2. Introduction

Measurements of forest stand parameters have become increasingly pursued in the most efficient and operational manner covering vast areas for active submission in sustainable forest management (Popescu et al., 2003; Evans et al., 2006; Wulder et al., 2012). The need for these measurements, advances the need for data collection by means of remote sensing technology due to traditional field based approaches being unfeasible and untimely at the required frequency and coverage (Sarkar and Nichol, 2011; Ismail et al., 2015).

Digital aerial photography has been greatly explored in forest structural attribute modelling as an attractive source of data owing to widespread varieties of spatial and spectral resolutions producing higher coefficient of determination values than that of satellite datasets (Hyyppa et al., 2000; Kim et al., 2009; Nichol and Sarkar, 2011; Wood et al., 2011). However, forest structural attributes such as height and volume cannot be measured directly from optical satellite data hence, recent studies have explored the notion of remotely sensed image texture features as a probable alternative for forest structural attribute modelling (Kayitakire et al., 2006; Dye et al., 2012; Dube et al., 2014; Dube and Mutanga, 2015a; Ismail et al., 2015). Haralick et al. (1973) proposed the extraction of texture features by means of a Grey Level Co-Occurrence Matrix (GLCM) which is one of the most frequently used algorithms for texture derivation and has been explored on the premise of texture segmentation or classification but seldom for the estimation of forest structural attributes (Franklin et al., 2000; Rao et al., 2002; Kayitakire et al., 2006; Gallardo-Cruz et al., 2012). Haralick (1979) proposed another algorithm known as a Grey Level Difference Vector Matrix (GLDV) and operates on the distance and angular spatial relationships over a specific image region and is considered to be diagonal variations of GLCMs. Literature has further suggested that the most effective way to use pixel-based texture with these algorithms is to determine an appropriate moving window

size for optimal predictive potential which may also depend on the skills of the interpreter and the desired output of the study.

In recent years, high spatial resolution imagery has been used by a number of authors such as, Gebreslasie et al. (2011) who used IKONOS imagery a machine learning approach known as artificial neural network statistics and image texture to produce a R^2 of 0.86 for volume. Dye et al. (2012) used QuickBird imagery, five texture variables and the random forest machine learning technique with a combination of spectral and textural variables to produce an overall model accuracy of $R^2 = 0.68$. Dube et al. (2014) also used very high resolution RapidEye imagery and derived vegetation indices coupled with stochastic gradient boosting and the random forest technique to produce an overall predictive accuracy of $R^2 = 0.80$ for the *E. grandis* species within a commercial forest plantation. This literature suggests that high resolution image texture has been used extensively to map forest structural attributes however; this process has been limited by a fixed window size of various image processing techniques, which doesn't take into account the edge effect. The present study will be using image texture at a plot level where a principal component analysis (PCA) will be used to calculate the average of the pixels thus the notion of the edge effect within the pixels and the need for an optimum window size is rendered moot in this situation.

In a forested landscape texture characteristics are dependent on high resolution imagery and the size and spacing of the tree crowns (Nichol and Sarkar, 2011). If a pixel falls on a tree its neighbouring pixel may also fall on the same tree resulting in a low local variance hence Nichol and Sarkar (2011) used this analogy to investigate texture parameters of two high resolution optical sensors (AVNIR-2 and SPOT-5) using multiple regression models developed from image parameters extracted from image processing. The results from this investigation improved as the texture parameters were averaged over both sensors ($r^2 = 0.911$ and RMSE = 30.10) however, further enhancements in the results were observed using a simple ratio of the texture measures of AVNIR-2 ($r^2 = 0.99$) and SPOT-5 ($r^2 = 0.916$) with the most hopeful result being obtained from the ratio of the texture measures from both sensors ($r^2 = 0.939$ and RMSE = 24.77).

Sarkar and Nichol (2011) continued to explore the concept of texture variable ratios and conducted an investigation using high resolution optical data, image processing methods such as texture parameters and ratios of texture parameters in simple linear and stepwise regression models for the estimation of forest parameters/biomass. An encouraging result from this

investigation suggested that the use of texture parameters of the spectral bands yielded operative biomass estimations (r^2 = 0.76 and RMSE = 46 t/ha) which were further improved using a simple ratio method of the texture features (r^2 = 0.88 and RMSE = 32t/ha). The concluding remarks of Sarkar and Nichol (2011) suggested that forest parameter/biomass estimations improved significantly using texture parameters of high resolution optical data.

Gallardo-Cruz et al. (2012) conducted research on carbon stock estimation and biodiversity monitoring by investigating whether texture from very high resolution satellite imagery could be used to estimate structure and diversity of secondary vegetation communities and their changes over time. This research involved the use of 40 texture variables across 6 structural attribute characteristics such as basal area, stand age and canopy height producing R^2 values of 0.93, 0.85 and 0.89 respectively. A simple method of a single image with linear models was adopted and produced reliable basal area and age estimates further allowing the application of image texture to be used in carbon sequestration monitoring and biodiversity loss in a more practicable and conclusive manner.

Commercial satellites are expected to produce very high resolution images holding the ability to reduce errors in forest structural attribute estimation for operational uses hence research conducted by Kayitakire et al. (2006) assessed the capability of very high resolution IKONOS-2 imagery to estimate various forest attributes such as age, stand density and basal area on the basis of texture measurements derived from second order statistical texture methods (GLCM). The R^2 values ranged from 0.76 to 0.82 for the 4 forest attributes in question (top height, circumference, stand density and age) except basal area which demonstrated a low correlation to the texture variables (R^2 = 0.35) which may warrant further inquiry in future research if need be.

Previous studies have demonstrated that the ability to quantify image texture using optical remote sensing and statistical approaches (GLOM and GLCM) has created the opportunity to estimate forest structural attributes without the need for traditional field based methodologies. This research endeavours to determine whether ratios of image texture features can be used to predict forest structural attributes in a commercial forest plantation in the Midland region of the KwaZulu-Natal Province, South Africa, without a selected moving window size but rather at a stand plot level based on the pixels from within the boundary of the image object. This will be done using high resolution (0.15m) four band colour infrared imagery focusing on the entire plot within the commercial forest plantation by means of machine leaning statistical analyses.

Hence, the objectives of this research are therefore to:

1. Perform a multiresolution segmentation on delineated plot boundaries to extract texture features after Haralick from a high resolution imagery using a principal component analysis, GLCM and GLDV in eCognition software
2. Conduct a correlation analysis to distinguish which of the extracted texture features have the most influence on forest structural attributes followed by a texture feature ratio analysis to predict forest structural attributes (such as volume, basal area dominant and mean tree height) using various machine learning algorithms (RF and PLSR-RF hybrid)

3.3. Materials and Methods

3.3.1. Study site

This study was conducted in a commercial forest plantation located west of Richmond, a town found in the KwaZulu-Natal Midlands region, South Africa, at 29° 52′ 0″ S, 30° 16′ 0″ E (Figure 9). The Sappi Riverdale plantation has a total planted area of 6200ha. This area is located along the Lovu River with its catchment draining into the ocean near Amanzimtoti and Park Rynie. The main veld types found in this region are Highland Sourveld (30%), Southern Tall Grassveld (20%) and the Ngongoni veld of the Natal Mistbelt (40%). Riverdale terrain is characterised by undulating hills with a low and mountainous geology. The geology of this region consists of a mix of mudstone, tillite, sandstone and basalt. The average altitude is 1190m with an accompanied average temperature of 16.1°C. This catchment area receives an annual precipitation of between 9-16mm. The forested area within the Riverdale plantation consists of commercial forestry of *Eucalyptus* species such as *Eucalyptus dunnii* (43%) and *Eucalyptus grandis* (57%) with an average age of seven years and an average height of 23.5m.

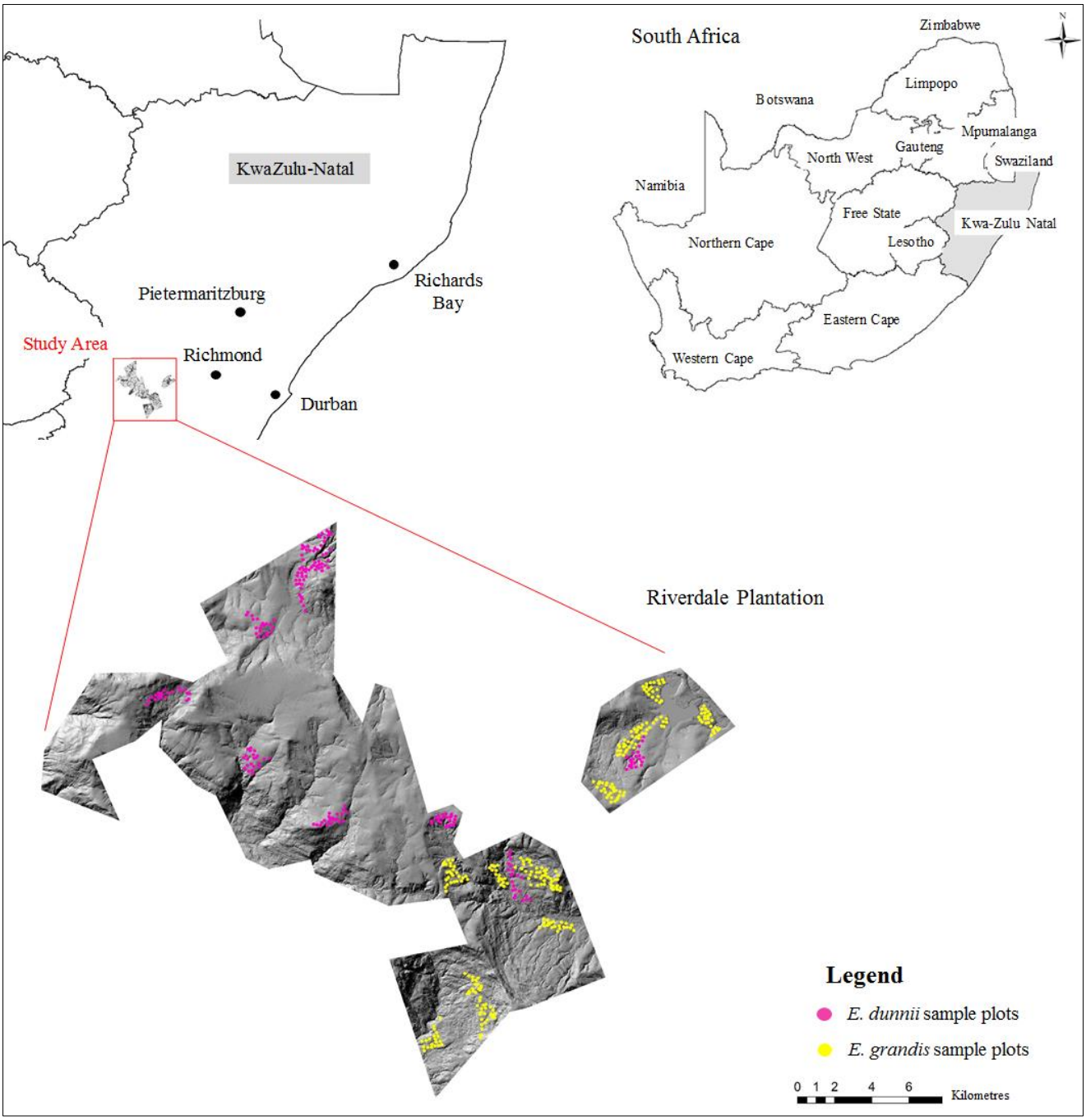

Figure 9: Location of the study area indicated as the Sappi Riverdale plantation the Midlands region of KwaZulu-Natal (South Africa)

3.3.2. Field data

Ground truth data was collected at the Sappi Riverdale plantation, Richmond, KwaZulu-Natal, South Africa, between the 12th of April and the 22nd May 2014 using standard enumeration techniques as suggested by Owen (2000). The ground truth data was collected in 10m radius circular plots where each plot was georeferenced across 25 compartments with a total of 502

sample plots. A systematic grid sampling technique was used where forest structural attributes such as basal area (Baha), volume (Volha), mean (Htm) and dominant tree height (HtD) were recorded for each circular plot during the field campaign. *Eucalyptus dunnii* and *Eucalyptus grandis* species that were between two and ten years (Figure 10) were considered for this study. The tree height and DBH for each plot was measured using the Vertex IV laser instrument and Haglof Digitech Calliper, respectively.

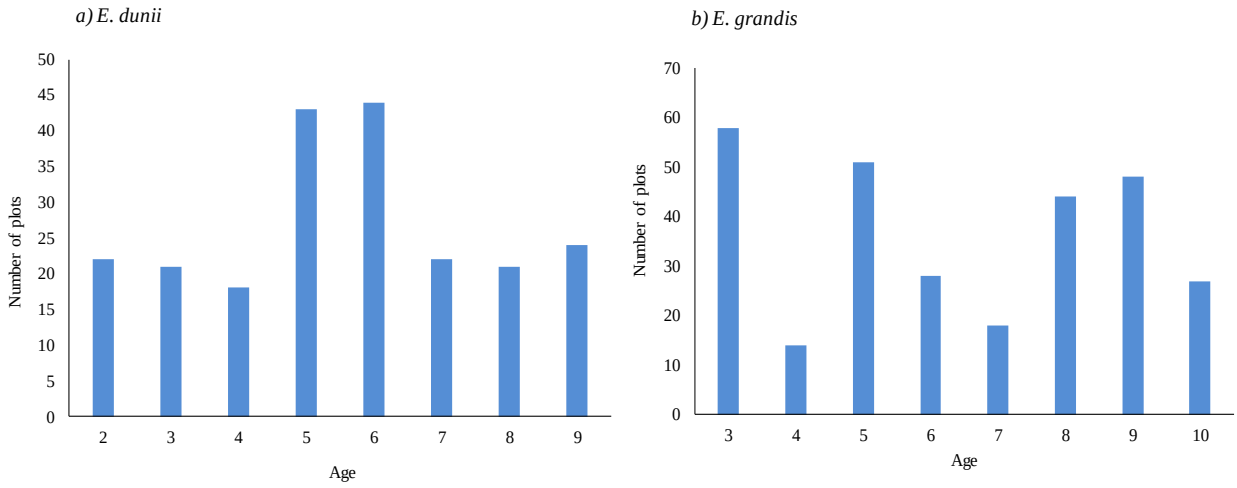

Figure 10: Age distribution of the a) *E. dunnii* species (n = 214) and b) *E. grandis* species (n = 288) located in the study area for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

3.3.3. Remote sensing data

The multispectral airborne imagery was sourced from Land Resource International (LRI), an independent service provider, under cloudless conditions on the 12th April 2014. LRI processed the image data using full in-house photogrammetry suites and Geomatica 10.1 photogrammetry software where Global Positioning System data was incorporated into this process. The imagery was supplied as Geo-TIFF files that were radiometrically corrected. The data had an 8-bit radiometric resolution with a 0.15m spatial resolution and four spectral bands (Table 6).

Table 6: Showing the four bands of colour infrared with their associated image band configuration

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

3.3.4. Segmentation

Segmentation algorithms are necessary when new image objects are needed based on the image layer information but are also crucial when refining existing image objects by subdividing them into smaller segments for a more detailed and focused analysis (Definien's Developer, 2007). The segmentation process, shape and size of a desired object is defined by the heterogeneity between pixels that are adjacent to each other where the scale parameter is considered to be the main input parameter for any segmentation process. The rationale for the segmentation of the individual sample plots (n = 502) is to generate objects that closely mirror the canopy and shadow structure within each plot. To obtain such results several multiresolution segmentation tests were executed using the multiresolution segmentation algorithm in Definien's eCognition Developer to segment the aerial photograph such that the canopy and shadow areas within the sample plots were delineated. The multiresolution segmentation yields the best abstraction and shaping in any application area. This particular algorithm consecutively merges the pixels and is considered to be a bottom-up segmentation algorithm based on a pair wise region merging technique (Definien's Developer, 2012). This segmentation process involves the identification of single image objects from one pixel and the subsequent merging of these objects with its neighbours based on a relative homogeneity criterion. According to Definien's Developer (2012) the homogeneity criterion is based on the combination of shape and colour criteria of both initial and the resulting image objects of the intended merging. Within the multiresolution segmentation algorithm, a higher scale parameter value results in larger image objects and smaller values result in smaller ones. The multiresolution segmentation uses an optimization route that minimizes the average heterogeneity of image objects and maximizes their respective

homogeneity for a given resolution. Key parameters that were tested in this study were the layer weights as well as the shape and compactness criterion. The image layer weights for all the four bands- red, green, blue and NIR were one, five, one and ten respectively. Settings for the composition of homogeneity criterion were assigned as 0.1 for shape and 0.5 for compactness. The testing process involved a variety of segmentation settings applied to the eCognition Developer where the trials indicated that a scale parameter of 40 (Figure 11) was the optimum value for forest canopy segmentation based on visual interpretation of the sample plots.

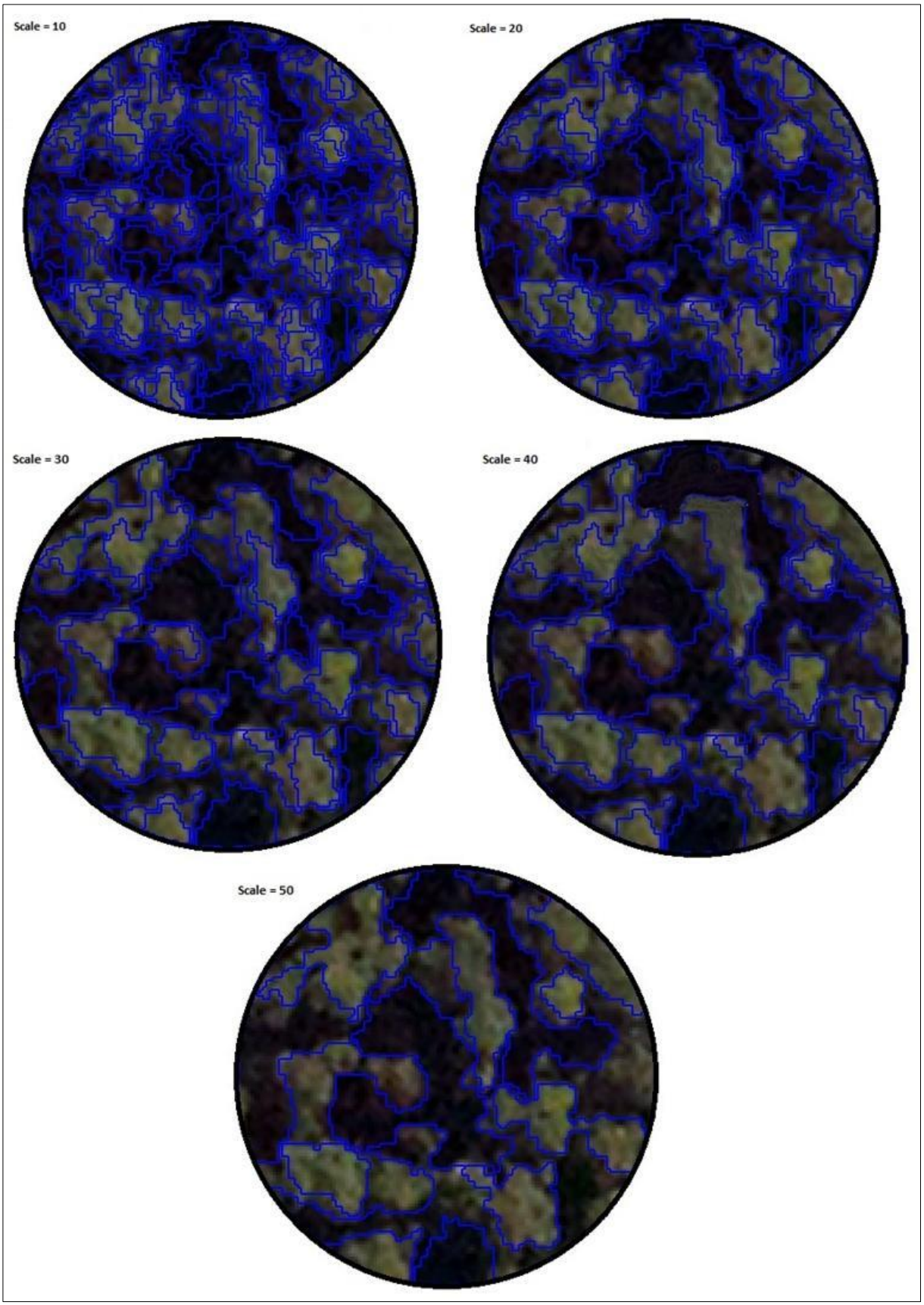

Figure 11: Visual representation of the multiresolution segmentation tests Scale = 10, 20, 30, 40, 50

3.3.5. Principal component analysis

PCA is one of the oldest and most popular multivariate statistical techniques used for feature reduction of redundancy in remote sensing data (Tsai et al., 2007; Sampson et al., 2011). PCA is straight forward, simple and easy to use and has been used in most available remote sensing image processing and analysis packages such as Definien's eCognition Developer. PCA is a linear transformation that projects data onto new orthogonal feature spaces in such a way that the first few components will represent the most variance in the original dataset (Jia and Richards, 1999; Tsai et al., 2007). PCA operates on several dependent variables which may be inter-correlated resulting in the main objective being to extract the most important information from these variables to express the data as new variables called principal components obtained as linear components of the original variables (Sampson et al., 2011). PCA aims to compress the data to keep only the most important information by simplifying the description of the data to allow for analysis of the observations and variables (Abdi and Williams, 2010). The first principal component is considered to have the largest possible variance allowing this component to extract and explain the most important data. The second component is computed under constraint of the first component and the other components are computed just the same. As the other components are computed the data retains less and less information. According to Tsai et al. (2007) when PCA is applied to multispectral imagery with a few discrete bands it is the most effective in extracting satisfactory outcomes and useful information. If PCA is applied to an entire data set the variance among the vegetation and non-vegetation covered pixels will dominate the Eigen analysis (Sampson et al., 2011). Only the most important information is to be extracted from the data however deciding the number of components needed from the PCA can be a challenge. For this study a PCA was used as part of the fundamental framework so that an appropriate feature extraction system was developed to extract the information most important to the variable analysis. Tsai et al. (2007) illustrated that the first principal component band accentuates the distinction between vegetation and non-vegetation clusters. On the basis of this framework texture was extracted by PCA of the first component as it is retained the most useful information from all four bands within the image.

3.3.6. Texture feature extraction

Texture can be defined as the spatial relationships within the image tone that may result in different values for neighbouring pixels for the same target type (Sarkar and Nichol, 2011). It can be said that the texture within high resolution imagery is the most important source of

information and not necessarily the intensity of the image (Podest and Saatchi, 2002). Haralick et al. (1973) proposed a number of image texture features derived from a Grey Level Co-Occurrence Matrix (GLCM) and a Grey Level Difference Vector Matrix (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, 1979). The data extracted by these methods greatly depends on the spatial resolution, spectral domain and sensed object characteristics (shape and dimension) (Kayitakire et al., 2006). An object-oriented classification was done before the extraction of the texture features without selecting a designated moving window size, contrary to previous studies that have used texture and an optimum window size (Blaschke, 2010; Mhangara and Odindi., 2013; Wang et al., 2015). Texture was then extracted after Haralick using a first principle component analysis in Definien's eCognition Developer version 9.0 software. Eight GLCM texture features were used (entropy, dissimilarity, contrast, second angle moment, standard deviation, correlation, homogeneity and correlation) and four GLDV features (second angle moment, mean, contrast and entropy) were computed. A correlation analysis was performed in the R statistical software to distinguish which texture features have the most effect on forest structural attribute prediction and those texture features were later used as a basis for the texture based ratios. Texture features were extracted from four directions (0°, 45°, 90°, 135°) thus to achieve directional invariance on each individual stand plot the four directions were summed together providing results for 'all directions' which represented the optimal results pertaining to forest structural attribute prediction for this study.

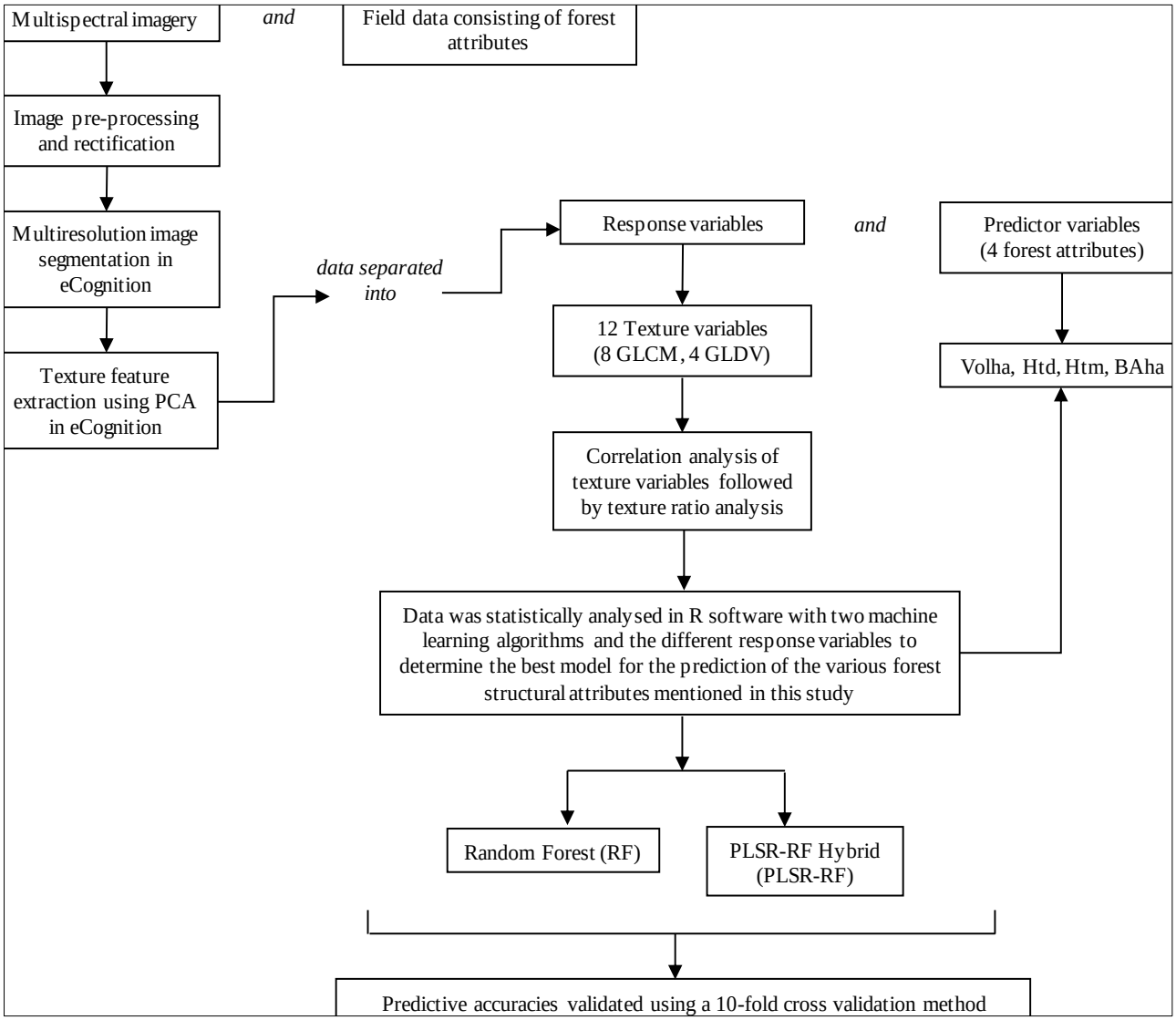

Figure 12: Flow chart of overall methodology

3.3.7. Random forest (RF)

The RF method is known as a bagging method and uses recursive partitioning to form regression trees (*ntree*) which are smaller homogeneous subdivisions of the data. When these regression trees are created the overall result of the trees are averaged (Breiman, 2001). 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. This process of bootstrapping is done with replacement and without the input variable selection stopping at each regression tree node (Adam et al., 2013). RF seeks to add randomness into the data by selecting random subsets of input variables (*mtry*) to establish the most efficient split at each tree node thus reducing bias and maintaining diversity within the data since no pruning is performed (Breiman, 2001; Lawrence et al., 2006; Adam et al., 2013). The RF model uses the remaining 34% of the data known as the out-of-bag (OOB) data for the

model prediction. The ensemble predicts the data using the difference in the mean square error of the OOB data and the data that is used to grow the homogeneous regression trees. The RF library (Liaw ad Wiener, 2002) in the R statistical software version 3.2.2 was used for this study (R Development Core Team, 2014). The two main parameters that are active in this algorithm are the number of input variables (*mtry*) and the number of trees (*ntree*), which was left at its default value of 500 because most of the errors reach stability before this number of regression trees are reached (Lawrence et al., 2006). During model validation a 10-fold cross validation method was used for this study.

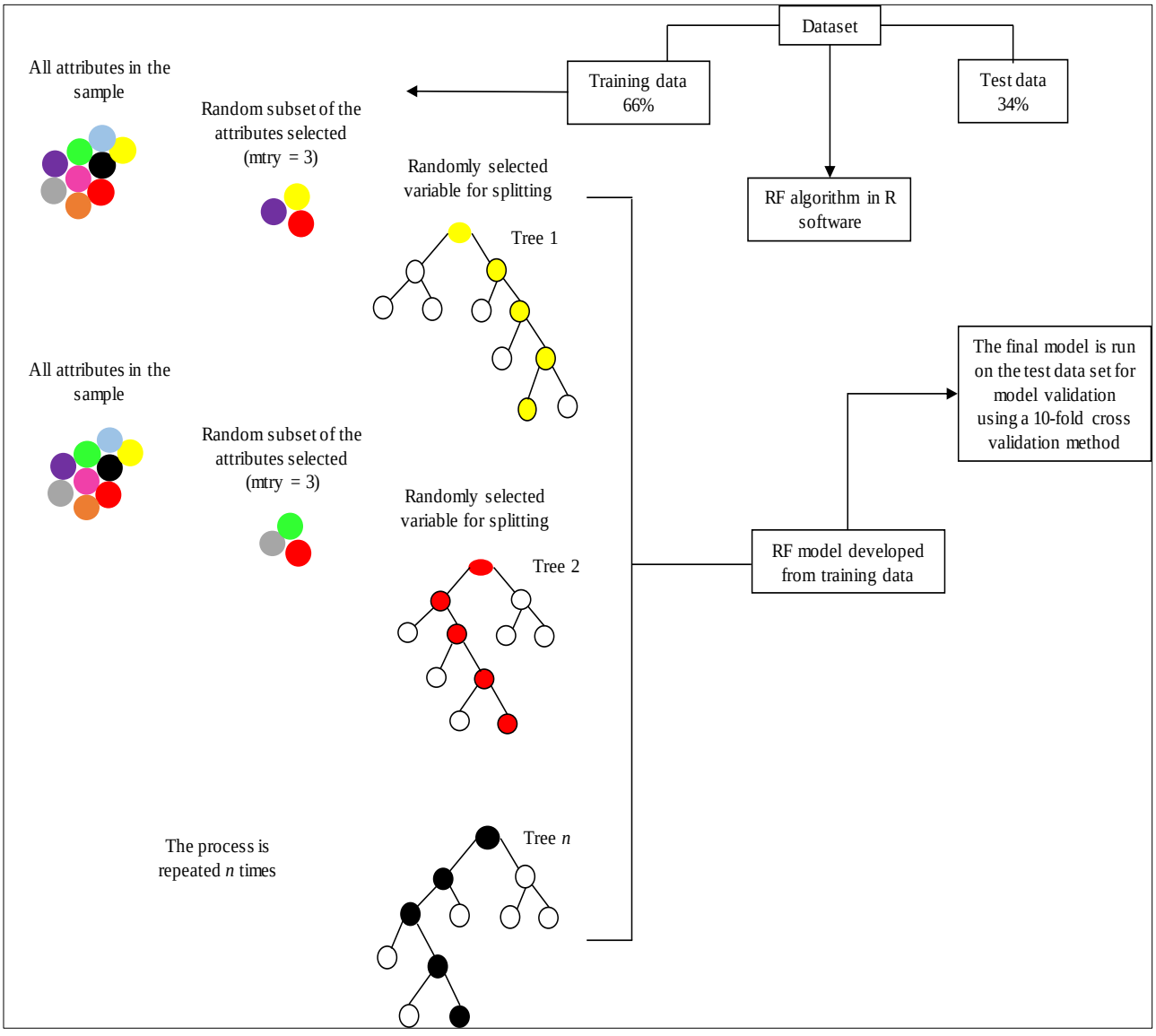

Figure 13: A graphical representation of how the RF algorithm works

3.3.8. Partial least squares-random forest (PLSR-RF) hybrid

PLSR is a linear statistical method used to model the relationship between two data sets. It 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; Ramírez et al., 2010; Abdel-Rehman et al., 2014). 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 (Sampson et al., 2011; Lopatin et al., 2015; Belgiu and Drăgut, 2016). The PLSR-RF hybrid model involves the random forest non-parametric methodology with the addition of linear PLSR approach. The hybrid method uses the explanatory variables and transforms them using the PLSR methodology to yield different components. Once this is done, the process is repeated and optimised. The PLSR part of the hybrid creates factors from the explanatory *(x)* variables that are the most relevant for the response *(y)* variables. The *x* and *y* variables are decomposed into latent structures in an iterative process where both structures of the *x* and *y* variables are considered. The PLSR model extracts the scores of the vectors which serve as new predictor representations and regresses the response variables on these new predictors thus creating components. Following this process, the RF method works with 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. RF is considered to be an ensemble algorithm that is robust in nature and does not over fit the data due to the random selection of explanatory variables that are selected from a large dataset (Bassa et al., 2016). Therefore, the PLSR-RF hybrid uses the PLSR methodology of latent variables and converts them into components which is then put through the RF binary recursive partitioning method to sift out the optimal components for prediction. During this process the bagging decision tree principle is applied to the selected components from the PLSR method. During model development the PLSR-RF hybrid selects 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 (Figure 12). Each tree node is randomly selected from the subset of components for splitting at each of these nodes further producing the result of an optimal component and variable for predictive purposes with hyperparameters listed as *ncomp, mtry* and *ntree*. 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.

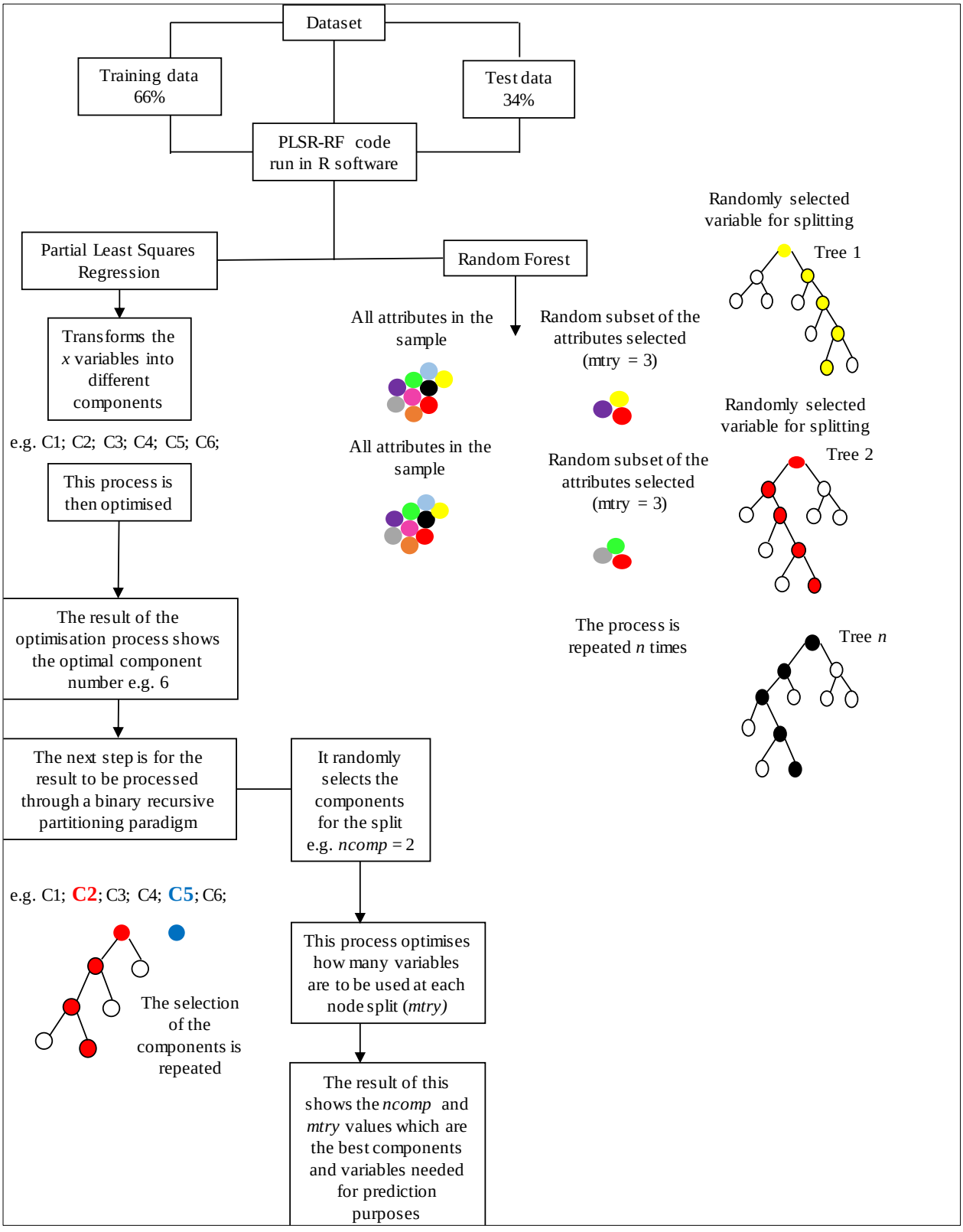

Figure 14: A graphical representation of how the PLSR-RF hybrid model works

3.3.9. Model Validation

In order to ensure the results obtained in this study remained unbiased, the data was randomly split into a training (66%) and a test (34%) dataset (Kuhn and Johnson, 2013). The model calibration and validation was done using a uniform stratified 10-fold cross validation (CV) method. Kohavi (1995) recommended this type of CV as the best for ensuring fair model prediction. The folds that were created in this process were "stratified" which meant that a full range of values from the response variables was used to ensure that there was a good representation of the entire dataset. The independent model validations were done using the test data which was untouched and completely random from the training data ensuring that when the 10 CV folds were created the data was free from error and bias.

To gauge the predictive accuracy of the RF and PLSR-RF hybrid models in predicting forest structural 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 4).

$$RMSE = \sqrt{\frac{\sum_{i-1}^{n}\left(x_{measured} - x_{predicted}\right)^2}{n}} \quad (4)$$

$X_{measured}$ 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.

3.4. Results

This study used texture features extracted through a first principal component analysis with two machine learning algorithms; Random Forest (RF) and a Partial Least Squares Regression-Random Forest Hybrid (PLSR-RF). *Eucalyptus* species within a commercial forest plantation over 25 compartments were used to predict various forest structural attributes using only the texture features extracted through eCognition software.

Texture features extracted through the Grey Level Co-occurrence Matrix (GLCM) method were the most popular texture features extracted after Haralick in eCognition software and appear to have the most influence on the forest structural attributes (Table 7 and 8).

Table 7: Texture feature results from the correlation analysis showing the texture features with the most influence in predicting forest structural attributes for young and mature *E. dunnii* and *E. grandis* species

Species	Age group	
E. grandis	**Young**	**Mature**
	GLCM Mean	GLCM Entropy
	GLCM Correlation	GLDV Entropy
	GLCM Standard Deviation	GLCM Homogeneity
	GLCM Dissimilarity	GLCM Second Angle Moment
E. dunii	**Young**	**Mature**
	GLCM Standard Deviation	GLCM Entropy
	GLDV Contrast	GLCM Contrast
	GLDV Entropy	GLCM Homogeneity
	GLDV Second Angle Moment	GLDV Second Angle Moment

Table 8: Results from the texture ratio analysis used for the young and mature *E. dunnii* and *E. grandis* species used for this study

Species	Age group	
E. dunii	**Young**	**Mature**
	$\frac{GLCM\ Standard\ Deviation}{GLDV\ Contrast}$	$\frac{GLCM\ Entropy}{GLCM\ Contrast}$
	$\frac{GLCM\ Standard\ Deviation}{GLDV\ Entropy}$	$\frac{GLCM\ Entropy}{GLCM\ Homogeneity}$
	$\frac{GLCM\ Standard\ Deviation}{GLDV\ Second\ Angle\ Moment}$	$\frac{GLCM\ Entropy}{GLDV\ Second\ Angle\ Momen}$
	$\frac{GLDV\ Contrast}{GLDV\ Entropy}$	$\frac{GLCM\ Contrast}{GLCM\ Homogeneity}$
	$\frac{GLDV\ Contrast}{GLDV\ Second\ Angle\ Moment}$	$\frac{GLCM\ Contrast}{GLDV\ Second\ Angle\ Momen}$
	$\frac{GLDV\ Entropy}{GLDV\ Second\ Angle\ Moment}$	$\frac{GLCM\ Homogeneity}{GLDV\ Second\ Angle\ Momen}$
E. grandis	**Young**	**Mature**
	$\frac{GLCM\ Mean}{GLCM\ Correlation}$	$\frac{GLCM\ Entropy}{GLDV\ Entropy}$
	$\frac{GLCM\ Mean}{GLCM\ Standard\ Deviation}$	$\frac{GLCM\ Entropy}{GLCM\ Homogeneity}$
	$\frac{GLCM\ Mean}{GLCM\ Dissimilarity}$	$\frac{GLCM\ Entropy}{GLCM\ Second\ Angle\ Moment}$
	$\frac{GLCM\ Correlation}{GLCM\ Standard\ Deviation}$	$\frac{GLDV\ Entropy}{GLCM\ Homogeneity}$
	$\frac{GLCM\ Correlation}{GLCM\ Dissimilarity}$	$\frac{GLDV\ Entropy}{GLCM\ Second\ Angle\ Moment}$
	$\frac{GLCM\ Standard\ Deviation}{GLCM\ Dissimilarity}$	$\frac{GLCM\ Homogeneity}{GLCM\ Second\ Angle\ Moment}$

Using two machine learning algorithms the young *E. dunnii* species returned statistically weak predictions for all forest structural attributes. The highest predictive accuracy was recorded for dominant tree height using the RF algorithm (R^2 = 0.38 and RMSE = 2.50m). Volume predictions using both the algorithms could not explain more than 10% of the variability in the

young *E. dunnii* forests. The RF model performed better than the PLSR-RF hybrid model across all forest structural attributes for young *E. dunnii* species (Figure 14a). Accuracies improved as the mature forests were used against the machine learning algorithms with the PLSR-RF hybrid algorithm producing the highest accuracy for dominant tree height ($R^2 = 0.60$ and RMSE = 2.78m). Using the mature *E. dunnii* forest the predictions results improved by up to 40% compared to the young *E. dunnii* species across both machine learning algorithms. Using the mature *E. dunnii* species the PLSR-RF performed better than the RF algorithm across majority of forest structural attributes (Figure 14b).

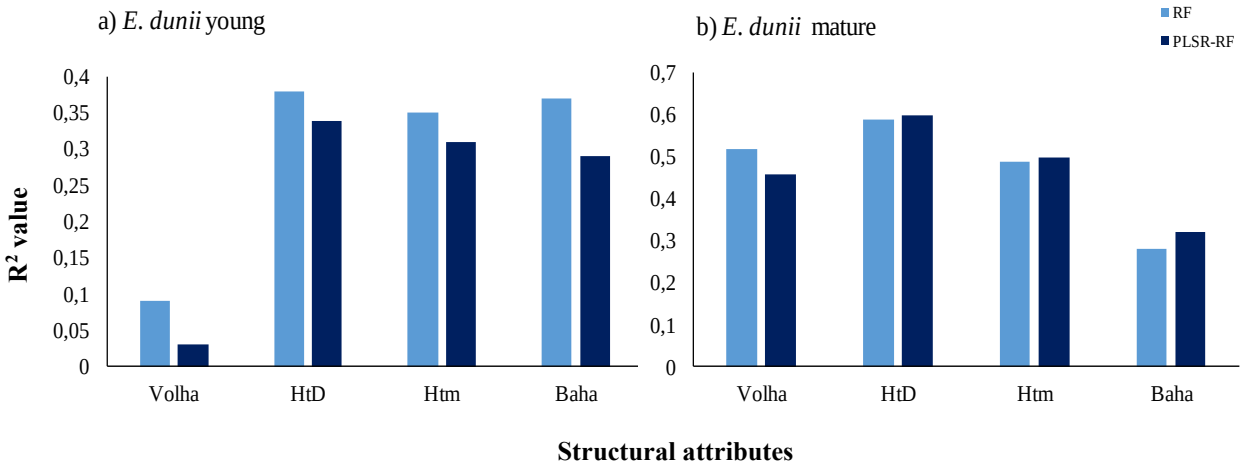

Figure 15: Results for a) young *E. dunnii* species and b) mature *E. dunnii* species using the RF and PLSR-RF algorithms for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

The hyperparameter optimization results when using only the young *E. dunnii* species suggests that all models used 2 variables for node splitting (*mtry*) when the RF model was trained and validated. The PLSR-RF model showed that the optimal number of components used for model development in young *E. dunnii* species was less than and equal to five accompanied by less than three variables used for node splitting. When the mature *E. dunnii* species was applied to the RF model the optimal number of variables used for node splitting was two with the exception of dominant tree height (Table 9). The PLSR-RF model highlighted that the optimal components used was three with the exception of dominant tree height and basal area. The PLSR-RF model also suggested that the most common number of variables used to for node splitting was one and three across all the forest structural attributes. The size of all the

ensembles *(ntree)* using both young and mature *E. dunnii* species was consistent at 500 trees when using the RF and PLSR-RF models because higher values of *ntree* parameters did not increase model accuracies (Table 9).

Table 9: Optimal *mtry, ncomp* and *ntree* hyperparameters obtained when using the various machine learning algorithms with young and mature *E. dunnii* species for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha). A default *ntree* value (*ntree* = 500) was used for all models.

	RF		PLSR-RF		
Forest attribute	*mtry*	*ntree*	*ncomp*	*mtry*	*ntree*
			Young *E. dunii* species		
Volha	2	500	5	2	500
HtD	2	500	1	1	500
Htm	2	500	2	2	500
Baha	2	500	5	1	500
			Mature *E. dunii* species		
Volha	2	500	3	3	500
HtD	6	500	4	1	500
Htm	2	500	3	3	500
Baha	2	500	5	1	500

This study looked at the young and mature *E. grandis* species (in comparison to the *E dunnii* species) across two machine learning algorithms to predict various forest structural attributes. Using only the young *E. grandis* forests the RF model performed the best for dominant tree height (R^2 = 0.69 and RMSE = 2.18m) which is a 31% improvement to the previously recorded young *E. dunnii* species. The RF model could not explain more than 20% of the variation for basal area prediction as indicated in figure 15a. However, the PLSR-RF model did show accuracy improvements for young *E. grandis* forests for volume and basal area when compared to using the RF model (Figure 15a). The mature *E. grandis* forests showed improved results when compared to the young *E. grandis* forests where basal area accuracies improved by up to 14% using the PLSR-RF model (R^2 = 0.42 and RMSE = 3.82m) (Figure 15b). The RF and

PLSR-RF models could not explain more than 9% of model variation for mean tree height when predicting the mature *E. grandis* forests.

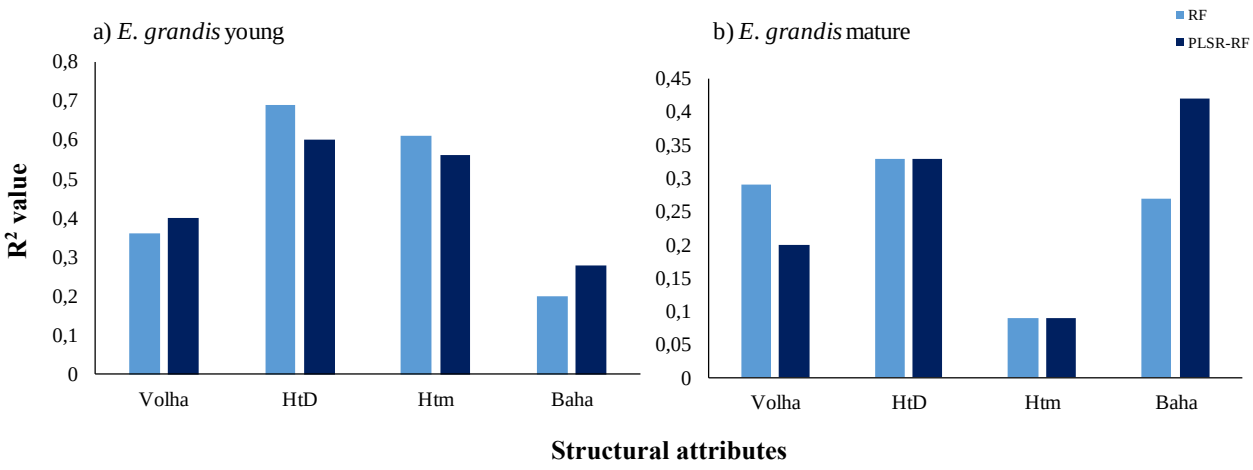

Figure 16: Results for a) young and b) mature *E. grandis* species using the RF and PLSR-RF algorithms for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha)

The hyperparameter optimisation results for young *E. grandis* species suggests that the majority of the models used less than six variables for node splitting using the RF model (Table 10). Using the PLSR-RF hybrid model the optimal components used was less than six with the majority of the forest structural attributes using one as the optimal variable for node splitting with the exception of dominant tree height and basal area. During analysis of the mature *E. grandis* forests the variables used for node splitting with the RF model was less than seven. The PLSR-RF model used the optimal components that were less than six (HtD, Htm and Baha) and one as the preferred number for node splitting accept for volume. The size of all the ensembles (*ntree*) in the RF and PLSR-RF methodology was consistent at 500 trees. Models with higher values of the *ntree* parameters did not increase training dataset accuracies.

Table 10: Optimal *mtry*, *ncomp* and *ntree* obtained when using the various machine learning algorithms and *E. grandis* for volume (Volha), dominant tree height (HtD), mean tree height (Htm) and basal area (Baha). A default *ntree* value (*ntree* = 500) was used for all models.

	RF		PLSR-RF		
Forest attribute	*mtry*	*ntree*	*ncomp*	*mtry*	*ntree*
	Young *E. grandis* species				
Volha	2	500	4	1	500
HtD	3	500	3	2	500
Htm	3	500	2	1	500
Baha	5	500	5	4	500
	Mature *E. grandis* species				
Volha	2	500	5	2	500
HtD	2	500	3	1	500
Htm	6	500	4	1	500
Baha	4	500	5	1	500

The overall results of this study suggest that dominate tree was the forest structural attribute that was the most accurately predicted using only the extracted texture features and ratios (Appendix A and Table 8) for this study. The PLSR-RF algorithm produced the best result for the mature *E. dunnii* species and the RF model produced the best results using the young *E. grandis* species (Figure 16).

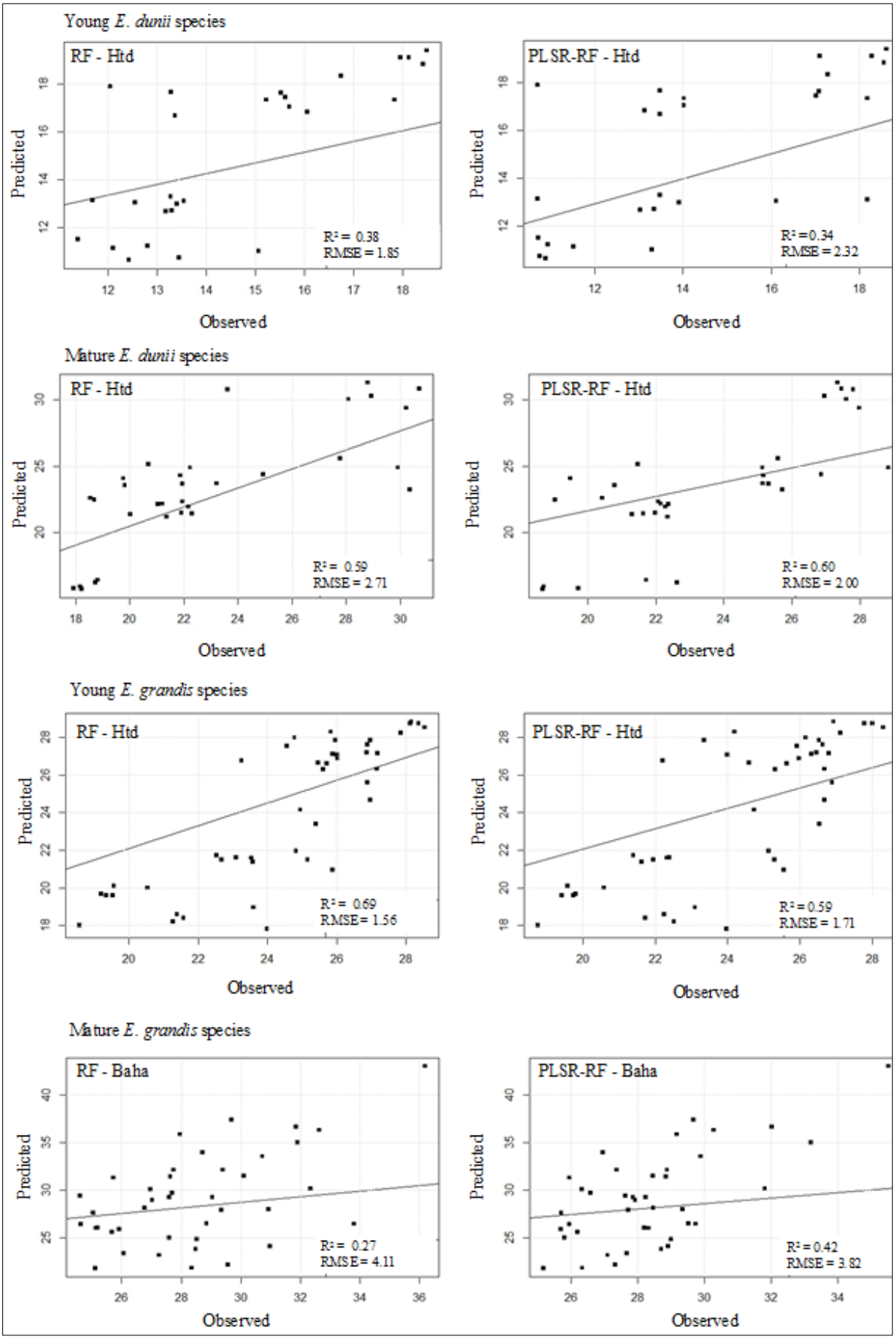

Figure 17: Observed vs predicted graphs for the best models obtained in this study using the RF and PLSR-RF machine learning algorithms for dominant tree height (HtD)

3.5. Discussion

3.5.1. Texture feature extraction and texture ratios

The predictive models for this study were built using image texture features across two machine learning algorithms; Random Forest (RF) regression and a Partial Least Squares Regression (PLSR)-RF hybrid approach. The imagery for this study consisted of a high spatial resolution owing to its main advantage of forest feature discrimination. The texture within this image consisted of many dimensions where texture as suggested by Haralick et al (1973) refers to the information concerning the spatial distribution of tonal variations of the band/s within an image and is considered to be an innate property of practically all surfaces. The key factor surrounding the use of texture variables is that it is more competent in characterizing forest structure with high resolution imagery (Sarkar and Nichol, 2011).

In recent years attempts to extract texture features have been somewhat limited due to algorithm development for the extraction of image properties of a specific nature due to the presence of edges. Hence, this study extracted texture features using a first principle component analysis in eCognition software with a multiresolution image classification of the entire stand plot thus eliminating the edge effect of the moving window size. Dye et al. (2008; 2012) suggested that by using principal components of the data in conjunction with the RF ensemble, more promising results were produced when compared to using just the original data. Nichol and Sarkar (2011) also reported observing this trend, hence our study used the first principle component in a PCA as many studies have shown that using texture to calculate principal components can improve model performance by producing high classification accuracies while reducing the amount of image data when processing very high resolution remotely sensed data (Puissant et al., 2005; Johansen et al., 2007; Nichol and Sarkar, 2011). Literature has demonstrated that texture properties are a more important source of data than reflectance or intensity values and this is especially true for high resolution imagery where better-quality detail of structural attributes can be discriminated. The importance of texture features increases as the spatial resolution increases due to the underlying spatial variation within the image Nichol and Sarkar, 2011; Gallardo-Cruz et al., 2012). Nichol and Sarkar (2011) used this analogy to investigate texture parameters of two high resolution optical sensors (AVNIR-2 and SPOT-5) using multiple regression models developed from image parameters extracted from image processing. The results from that investigation improved as the texture parameters were averaged over both sensors (R^2 = 0.91 and RMSE = 30.10) however, further enhancements in

the results were observed using a simple ratio of the texture measures of AVNIR-2 (r^2 = 0.99) and SPOT-5 (r^2 = 0.916) were used with the most hopeful result being obtained from the ratio of the texture measures from both sensors (r^2 = 0.939 and RMSE = 24.77). The present study adopted a similar methodology to determine whether or not forest structural attributes can be determined from the ratios of the texture features within the high resolution image. The present study suggests that dominant tree height was the most accurately predicted forest structural attribute using only texture ratios of the extracted texture features as input variables with an accuracy level of just less than 70% being obtained. Sarkar and Nichol (2011) continued to explore the concept of texture variable ratios and conducted an investigation using high resolution optical data, image processing methods such as texture parameters and ratios of texture parameters in simple linear and stepwise regression models for the estimation of forest parameters/biomass. An encouraging result from that investigation suggested that the use of texture parameters of the spectral bands yielded operative biomass estimations (r^2 = 0.76 and RMSE = 46 t/ha) which were further improved using a simple ratio method of the texture features (r^2 = 0.88 and RMSE = 32t/ha). The concluding remarks of Sarkar and Nichol (2011) suggested that forest parameter/biomass estimations improved significantly using texture parameters of high resolution optical data which supports the positive accuracies obtained in the present study.

Literature has suggested caveats attached to the use of texture measurements such as the perception that texture is an extremely complicated property with significant variance regarding the object of interest, window size selection and physiographic conditions (Franklin, 2000; Chen et al., 2004; Sarkar and Nichol, 2011). In order to process texture variables in the most effective manner with fine structural detail of forest attributes, optical imagery with high spatial resolutions (<1m pixels) are expected to hold the greatest predictive potential (Fuchs et al., 2009; Sarkar and Nichol, 2011). Contrary to such caveats texture characteristics of high resolution images have the potential to provide important and useful information about the spectral and structural characteristics of a region of interest or object once these features are extracted and statistically analysed (Gebreslasie et al., 2011; Dube et al., 2014; Dube and Mutanga, 2015a; Dube and Mutanga, 2015b; Ismail et al., 2015).

3.5.2. Model development for individual forest species

This study used 502 measured stand plots with two species present; *E. dunnii* and *E. grandis* within the commercial forest plantation consisting of differing age structures (young and mature). A forests age is an important factor for determining its ultimate yield and is valuable in ascertaining certain forest conditions (Dye et al., 2012). Texture is greatly affected by the age of a forest due to various structural changes associated with a developing/maturing forest. Texture analysis by means of a statistical approach considers the distribution and disparity of spectral variability in the region of interest. The statistical methods can be divided into two categories depending on the pixels that define the area of interest. These categories are first order and second order statistical textural measurements (Haralick et al, 1973). Gebreslasie et al. (2011) suggested that there is great potential for using high resolution imagery for the estimation of forest structural attributes by means of statistical techniques and produced a R^2 of 0.86 using IKONOS imagery and the artificial neural networks statistical technique. The RF ensemble method has been used extensively in recent years for classification and prediction in various study areas. Hence, RF coupled with the novelty of the PLSR-RF hybrid approach model development proved to be promising for forest structural attribute modelling in this particular study. Using only the young *E. dunnii* species the RF and PLSR-RF algorithms could not explain more than 40% of the variation which could be attributed to the homogeneous nature of the forest stands and not necessarily due to algorithm strength (Dye et al., 2012). In an attempt to further improve those results the mature *E. dunnii* species were used to develop models using the RF and PLSR-RF algorithms which resulted in up to a 20% improvement on predictive accuracies with the PLSR-RF model producing the best result for dominate tree height. When the *E. grandis* species was used as input variable the young forests produced the highest results for dominant tree height using the RF model. The mature *E. grandis* forests could not explain more than 42% of variation using both the machine learning algorithm across all forests structural attributes. The variation in these findings suggest that the age of the individual species plays a significant role on the prediction of forest structural attributes when using machine learning algorithms. Each individual species has its own structural and taxonomic characteristics (Dube et al., 2014). Within a commercial forest plantation as the trees begin to mature some of the other individual trees may die off due to intra-species competition for resources. This could explain why the young *E. dunnii* and mature *E. grandis* regression models performed poorly.

The RF model demonstrates an overestimation when the data has less variability and is of a small size and an underestimation of the data when it is variable and of a large origin (Ismail, 2009; Dube et al., 2014; Mutanga et al., 2012). However, the RF machine learning regression technique is a robust and useful tool in forest studies when using spectral band information, texture features and a combination of the two when extracted from remote sensing data (Dye et al., 2012; Wang et al., 2015). Results show that dominant tree height was the most accurate variable to be predicted using texture feature ratios with similar findings being reported by Kayitakire et al. (2006). The RF model performed the best with the ratios of the texture features because the texture features have an enhanced spatial relationship of the pixels within the very high resolution image.

Further research is required into the RF and PLSR-RF models and how the amount of noise in the RF ensemble affects the predictive accuracy of both the model. This study highlights that spatial resolution plays a crucial part in the prediction studies hence more information is needed to establish the optimum pixel size for texture extraction and whether or not the inclusion of spectral information to the texture ratios would have any effect on predictive accuracies.

3.6. Conclusion

This paper investigated: (i) whether texture based ratios were able to predict forest structural attributes within a commercial forest plantation using the performance and strength of two machine learning algorithms (RF and PLSR-RF hybrid).

Our results have demonstrated that: (i) the RF model is more robust in predicting dominant tree height in various *E. grandis* forests when derived from the young tree species within the plantation (ii) there is great potential for using high resolution (0.15m) remotely sensing image data for texture extraction using a PCA to reduce data redundancy.

Chapter four

4.1. Conclusion

Accurate measurements of forest structural attributes are vital for ascertaining commercial forest health directly related to the potential yield of commercial forest plantations for various industrial needs. The measurements within these commercial forest plantations are needed over vast distances in the most accurate and efficient manner. Traditional field based approaches

have been employed extensively in the past when budgets and man power allowed for it and although this method is unbiased it is not feasible over such great distances such as commercial forest plantations. This shortcoming gave rise to the inception of more industries using remotely sensed data as an alternative method coupled with statistical approaches for the derivation of forests structural attributes from optical imagery over these commercial forest plantations. The aim of this research with those thoughts in mind endeavoured to determine whether image data of a spectral and textural nature could be used to determine forest structural attributes within a commercial forest plantation using high resolution remotely sensed imagery and machine learning statistical techniques.

The main objectives were (i) to assess the capability of a combination of spectral and textural features extracted from high resolution imagery to predict forest structural attributes within a commercial forest plantation (ii) to test the ability of the PLSR, RF and PLSR-RF models in predicting forest structural attributes such as basal area, tree height and volume (iii) to assess the capability of using only texture features extracted from multispectral data in predicting forests structural attributes within a commercial forest plantation (iv) to assess the capability of the PLSR-RF hybrid model for predicting forest structural attributes (v) to offer recommendations for the use of remotely sensed measures of texture characteristics from high resolution imagery for commercial forest plantation management.

4.2. Assessing the capability of a combination of spectral and textural features for the prediction of forest structural attributes

The results from this study confirm the potential of using a combination of spectral and textural remotely sensed optical imagery to accurately predict forest structural attributes such as tree height, basal area and volume. Similar results were obtained by Wunderle et al. (2007) Dye et al. (2012), Dube et al. (2014), Wang et al. (2015). The combination of spectral and texture data produced high accuracies for dominant tree height ($R^2 = 0.82$).

4.3. Testing the ability of the PLSR, RF and PLSR-RF models in predicting forest structural attributes

Machine learning algorithms have shown great potential for combining spectral and textural variables to accurately predict forest structural attributes within a commercial forest plantation. Traditional linear algorithms such as the PLSR method requires linearity, a normal distribution

and the absence of collinearity amongst the input variables (Jensen et al., 1999). Data collected through remotely sensed techniques are often non-linear demonstrating that the RF and PLSR-RF algorithms were successful in the integration of spectral and texture variables of the remotely sensed data thus confirming that these algorithms do have the capability to handle complex non-linear and correlated predictor variables. The results from this study showed that (i) by using a principal component analysis of the texture variable datasets the results were promising with a similar trend being reported by Dye et al. (2012) (ii) the RF ensemble showed great potential for predicting forests structural attributes with the best result being reported for dominant tree height (R^2 = 0.77) (iii) the developed methodology for this study of the PLSR-RF hybrid algorithm showed how robust the algorithm was in handling a combination of spatial and textural data which resulted in the highest model accuracy being produced using the PLSR-RF algorithm for dominant tree height (R^2 = 0.82). These results illustrate the benefit of utilizing the proposed method for predicting forest structural attributes within a commercial forest plantation where the method developed in this study could potentially be applied to other remote sensing applications that combine spectral and texture variables.

4.4. Assess the capability of using only texture features extracted from multispectral data in predicting forests structural attributes within a commercial forest plantation

Texture within an image is considered to be a multi-scale entity with window sizes playing an important role in texture analysis (Moskal and Franklin, 2011). The smallest window size (3 x 3) is sufficient for commercial forest trees that are grown for pulpwood and are therefore grown close together (Dye et al., 2012) however, this study did not use a selected moving window size but rather used the entire stand plot for the extraction of the texture measures eliminating the edge effects within the pixel. It is important to note that image texture varies with age thus different forest ages within the commercial forest plantation produced different results based on an age range which allows for the relationship between age and image texture to play a role in predictive accuracies (Johansen et al., 2007). Using the ratios of the texture features the results of the forest structural attribute predictions showed great potential for predicting tree height with the best result being recorded for young *E. grandis* species (R^2 = 0.69) using only texture features with the RF algorithm. Similar trends were reported by Sarkar and Nichol. (2012) regarding texture ratios and machine learning algorithms.

4.5. Recommendations for the use of remotely sensed measures of texture characteristics from high resolution imagery for commercial forest plantation management

- This research has highlighted the importance of spatial resolution for the prediction of forest structural attributes within a commercial forest plantation thus future research could endeavour to distinguish which pixel size best discriminates forest structural attributes within this type of commercial forest plantation using similar species as this research has mentioned each species has its own intra and inter specific characteristics which interact differently with the machine learning algorithms. Thus using sensors with enhanced spatial capabilities (such as WorldView-3 imagery) could improve predictive accuracies for future studies.

- The performance of using the RF ensemble in various regression studies where the small number of samples versus the large number of variables is not fully understood. Much research has focused on RF for classification purposes but little attention has been given to the regression aspect. It is also important to bear in mind that no single machine learning algorithm is superior in all aspects (Kohavi et al., 1996) thus it is recommended that the RF and PLSR-RF algorithms be compared to other machine learning algorithms in various case studies to establish its robustness and capabilities with new datasets.

- The RF algorithm is affected by noise in the ensemble and this should be explored in more detail in terms of how the noise in the RF algorithm affects the hybrid algorithm in terms of predictive potential.

- Future research should endeavour to explore the differences between oblique RF and traditional RF in creating another hybrid algorithm for more accurate regression analyses within commercial forest plantations.

In conclusion the aim of this research endeavoured to determine whether image data of a spatial and spectral nature can be used to determine forest structural attributes within a commercial forest plantation using high resolution remotely sensed imagery and machine learning

statistical techniques. The final conclusion was based on the following observations in this thesis:

- Spectral and textural image data has the capability to accurately predict forest structural attributes within a commercial forest plantation. The combination of spectral and textural features produced the best results for dominant tree height using the developed PLSR-RF hybrid methodology ($R^2 = 0.82$).

- The comparison of the three machine learning algorithms has suggested that the developed methodology for the PLSR-RF algorithm showed great potential for predicting forest structural attributes and more specifically the use of a principal component analysis of the texture features reduced the data redundancy within the study and contributed to producing the positive results that were obtained.

- Using a texture analysis, it was found that the ratios of the texture features produced the best results for the prediction of forest structural attributes within the commercial forest plantation using the RF algorithm for dominant tree height ($R^2 = 0.69$).

- The use of machine learning algorithms for the prediction of forest structural attributes within a commercial forest plantation was done with acceptable success using high resolution imagery and could potentially do well for other case studies.

4.6. References

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

Abdel-Rahman, E. M., Way, M., Ahmed, F., Ismail, R., and Adam, E. (2013). Estimation of thrips (Fulmekiola serrata Kobus) density in sugarcane using leaf-level hyperspectral data. South African Journal of Plant and Soil, 30 (2), 91-96.

Adam, E., Mutanga, O., and Ismail, R. (2013). Determining the susceptibility of Eucalyptus nitens forests to Coryphodema tristis (cossid moth) occurrence in Mpumalanga, South Africa, International Journal of Geographical Information Science, 27:10, 1924-1938, DOI: 10.1080/13658816.2013.772183.

Bassa, Z., Bob, U., Szantoi, Z., and 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.

Blaschke, T. (2010). Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry, 65 (1), 2-16.

Belgiu, M., Drăgut, L., and Strobl, J. (2014). Quantitative evaluation of variations in rulebased classifications of land cover in urban neighbourhoods using WorldView-2 imagery. ISPRS Journal of Photogrammetry. Remote Sensing. 87, 205–215.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1982) Classification and regression trees. CRC Press.

Breiman, L., (1996). Bagging predictors. Machine Learning, 24 (2), 123-140.

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

Carrascal, L. M., and Gordo. I. G. O. (2009). Partial Least Squares Regression as an Alternative to Current Regression Methods Used in Ecology. Oikos 118: 681-690.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., and Torresan, C. (2012). Estimating forest timber volume by means of “low-cost” LiDAR data. Italian Journal of Remote Sensing, 44 (1), 125-140.

Chan, J. W., Laporte, N., and Defries, R. S. (2003). Texture classification of logged forests in tropical Africa using machine-learning algorithms. International Journal of Remote Sensing, 24 (6), 1401-1407.

Champion, I., Dubois-Fernandez, P., Guyon, D., and Cottrel, M. (2008). Radar image texture as a function of forest stand age. International Journal of Remote Sensing, 29, 1795-1800.

Chen, D., Stow, D. A., and Gong, P. (2004). Examining the effect of spatial resolution and texture window size on classification accuracy: An urban environment case. International Journal of Remote Sensing, 25, 2177-2192.

Cutler, A., Cutler, R.D., and Stevens, J.R. (2009). Tree-based methods. In X. Li and R. Xu (Eds), High-Dimensional Data Analysis in Cancer Research (pg 83-101). Springer Science and Business Media LLC.

Definiens Developer. (2012). Definiens Developer XD 2.0.4- User Guide, Published by Definiens AG, Bernhard-Wicki-Straße 5, 80636, Munchen, Germany. http://www.imperial.ac.uk/media/imperial-college/medicine/facilities/film/Definiens-Developer-User-Guide-XD-2.0.4.pdf

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.

Dube, T., Mutanga, O., Elhadi, A., and Ismail, R. (2014). Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution space borne multispectral rapideye sensor and advanced machine learning algorithms. Sensors, 14 (8), 15348-15370.

Dube, T., and Mutanga, O. (2015a). Investigating the robustness of the new Landsat-8 Operational Land Imager derived texture metrics in estimating plantation forest aboveground biomass in resource constrained areas. ISPRS Journal of Photogrammetry and remote sensing, 108, 12-32.

Dube, T., and Mutanga, O. (2015b). Evaluating the utility of the medium-spatial resolution Landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa. ISPRS Journal of Photogrammetry and Remote Sensing, 101, 36-46.

Dye, M., Mutanga, O., and Ismail, R. (2008). Detecting the severity of *Woodwasp, Sirexnoctilio*, infestation in a pine plantation in KwaZulu-Natal, South Africa, using texture measures calculated from high spatial resolution imagery. African Entomology, 16 (2) 263-275.

Dye, M., Mutanga, O., and Ismail, R. (2011). Examining the utility of random forest and AISA Eagle hyperspectral image data to predict *Pinus patula* age in KwaZulu-Natal, South Africa. Geocarto International, 26 (4), 275-289, DOI: 10.1080/10106049.2011.562308.

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.

Evans, D. L., Roberts, S. D., and Parker, R. C. (2006). LiDAR A new tool for forest measurements. The Forestry Chronicle, 82 (2), 211-218.

Franklin, S. E., Hall, R. J., Moskal, L. M., Maudie, A. J., and Lavigne, M. B. (2000). Incorporating texture into classification of forest species composition from airborne multispectral images. International Journal of Remote Sensing, 21 (1), 61-79.

Fuchs, H., Magdon, P., Kleinn, C., and Flessa, H. (2009). Estimating aboveground carbon in a catchment of the Siberian forest tundra: Combining satellite imagery and field inventory. Remote Sensing of Environment, 113, 518−531.

Gallardo-Cruz, J. A., Meave, J. A., González, E. J., Lebrija-Trejos, E. E., Romero-Romero, M. A., Pérez-García, E. A., Gallardo-Cruz. R., Hernández-Stefanoni. J. L., and Martorell, C. (2012). Predicting tropical dry forest successional attributes from space: is the key hidden in image texture? PloS one, 7 (2), e30506.

Gebreslasie, M. T. (2008). The estimation of eucalyptus plantation forest structural attributes using medium and high spatial resolution satellite imagery. Faculty of Science and Agriculture, University of KwaZulu-Natal Pietermaritzburg, South Africa. Doctor of PhilosophyThesis. http://researchspace.ukzn.ac.za/xmlui/bitstream/handle/10413/354/MT-Gebreslasie.pdf?sequence=1

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.

Getzin, S., Wiegand, K., and Schöning, I. (2012). Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. Methods in Ecology and Evolution, 3 (2), 397-404.

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.

Hyyppa, J., Hyyppa, H., Inkinen, M., Engdahl, M., Linko, S., and Zhu, Y. H. (2000). Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. Forest Ecology and Management, 128 (1-2), 109-120.

Ismail, R., and Mutanga, O. (2010). A comparison of regression tree ensembles: predicating *Sirexnoctilio* induced water stress in *Pinus patula* forest of KwaZulu-Natal, South Africa. International Journal of Applied Earth Observation and Geoinformation, 12, 45-51.

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.

Johansen, K., Coops, N. C., Gergel, S. E., and Stange, Y., (2007). Application of high spatial resolution satellite imagery for riparian and forest ecosystem classification, Remote Sensing of Environment, 110, p. 29–44.

Kayitakire, F., Hamel, C., and Defourny, P. (2006). Retrieving forest structure variables based on image texture analysis and IKONOS-2 imagery. Remote Sensing of Environment, 102(3), 390-401.

Kim, M., Madden, M., and Warner, T. A. (2009). Forest Type Mapping using Object-specific Texture Measures from Multispectral IKONOS Imagery. Photogrammetric Engineering & Remote Sensing, 75(7), 819-829.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* 14 (2), 1137-1145.

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

Kuhn, M. (2015). Caret: classification and regression training. Astrophysics Source Code Library, 1, 05003.

Lawrence, R. L., Wood, S. D., and Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (Random forest). Remote Sensing of Environment, 100, 356-362.

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

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.

Mhangara, P., and Odindi, J. (2013). Potential of texture-based classification in urban landscapes using multispectral aerial photos. South African Journal of Science, 109 (3/4), Article 1273.

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. Remote Sensing, 8 (2), 88.

Mutanga, O., Adam, E., and Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. International journal of Applied Earth Observation and Geoinformatics, 18, 399-406.

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

Oumar, Z., Mutanga, O., and Ismail, R. (2013). Predicting Thaumastocoris peregrinus damage using narrow band normalized indices and hyperspectral indices using field spectra resampled to the Hyperion sensor. International Journal of Applied Earth Observation and Geoinformation, 21, 113-121.

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.

Paine, D. P., and Kiser, J. D. (2003). Aerial Photography and Image Interpretation. 2nd ed. Hoboken, New Jersey. John Wiley and Sons, Incorporated.

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26 (1), 217-222.

Pasher, J., and King, D. J. (2010) Multivariate forest structure modelling and mapping using high resolution airborne imagery and topographic information. Remote Sensing of Environment, 114, 1718–1732.

Popescu, S. C., Wynne, R. H., and Nelson, R. F. (2003). Measuring individual tree crown diameter with LiDAR and assessing its influence on estimating forest volume and biomass. Canadian journal of remote sensing, 29 (5), 564-577.

Prasad, A.M., Iverson, L.R., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9, 181-199.

Puissant, A., Hirsch, J., and Weber, C. (2005). The utility of texture analysis to improve per-pixel classification for high to very high spatial resolution imagery. International Journal of Remote Sensing, 26 (4), 733-745.

Rao, P. V. N., Sai, M. V. R. S., Sreenivas, K., Rao, M. V. K., Rao, B. R. M., Dwivedi, R. S., and Venkataratnam, L. (2002). Textural analysis of IRS-1D panchromatic data for land cover classification. International Journal of Remote Sensing, 23 (17), 3327-3345.

Ramírez, J., Górriz, J. M., Segovia, F., Chaves, R., Salas-Gonzalez, D., López, M., and Padilla, P. (2010). Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. Neuroscience letters, 472 (2), 99-103.

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/.

Sampson, D. L., Parker, T. J., Upton, Z., and Hurst, C. P. (2011). A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. PloS one, 6 (9), e24973.

Sarker, L. R., and Nichol, J. E. (2011). Improved forest biomass estimates using ALOS AVNIR-2 texture indices. Remote Sensing of Environment, 115 (4), 968-977.

Sarker, M. L. R., Nichol, J., Iz, H. B., Ahmad, B. B., and Rahman, A. A. (2013). Forest biomass estimation using texture measurements of high-resolution dual-polarization C-band SAR data. IEEE Transactions on Geoscience and Remote Sensing, 51 (6), 3371-3384.

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.

Sheridan, R. D., Popescu, S. C., Gatziolis, D., Morgan, C. L., and Ku, N. W. (2014). Modelling forest aboveground biomass and volume using airborne LiDAR metrics and forest inventory and analysis data in the Pacific Northwest. Remote Sensing, 7 (1), 229-255.

Tesfamichael, S. G., Van Aardt, J. A. N., and Ahmed, F. (2010). Estimating plot-level tree height and volume of Eucalyptus grandis plantations using small-footprint, discrete return LiDAR data. Progress in Physical Geography, 34 (4), 515-540.

Tewari, D. D. (2001). Is commercial forestry sustainable in South Africa? The changing institutional and policy needs. Forest Policy and Economics, 2 (3), 333-353.

United States Department of Agriculture Forest Service (USDA) (2005). Aerial Photo Guide to New England Forest Cover Types. http://www.nrs.fs.fed.us/pubs/gtr/gtr_nrs195.pdf

Wood, E. M., Pidgeon, A. M., Radeloff, V. C., and Keuler, N. S. (2012). Image texture as a remotely sensed measure of vegetation structure. Remote Sensing of Environment, 121, 516-526.

Wold, S. (1995). PLSR for multivariate linear modelling. In: van de Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design. VCH, Weinheim, Gemany, pp. 195–218.

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 Environment, 113 (9), 2019-2036.

Wulder, M. A., LeDrew, E. F., Franklin, S. E., and Lavigne, M. B. (1998). Aerial image texture information in the estimation of northern deciduous and mixed wood forest leaf area index (LAI). Remote Sensing of the Environment. 64, 64–76.

Wulder, M. A., White, J. C., Nelson, R. F., Næsset, E., Ørka, H. O., Coops, N. C., and Gobakken, T. (2012). LiDAR sampling for large-area forest characterization: A review. Remote Sensing of the Environment, 121, 196-209.

Wunderle, A. L., Franklin, S.E., and Guo, X.G. (2007). Regenerating boreal forest structure estimation using SPOT-5 pan sharpened imagery. International Journal of Remote Sensing, 28, (19), 4351-4364.

APPENDIX A

P (i, j) is the normalised co-occurrence matrix such that SUM (i, j = 0, N-1) (P (i, j)) = 1. V(k) is the normalised grey level difference vector V(k) = SUM (i, j = 0, N-1 and |i-j| = k) P (i, j).

Texture Variable	Formula	Explanation	Reference
GLCM Mean	$Mean(ME) = \sum_{i,j=0}^{N-1} i\ P_{i,j}$	The GLCM mean refers to the average of the grey level pixels in the local window.	Guo et al. (2004) Yu et al. (2006)
GLCM Entropy	$Entropy(EN) = \sum_{i,j=0}^{N-1} i\ P_{i,j}\ (-\ln P_{i,j})$	This variables deals with the size of the image and how much of the information present in the images needs to be compressed. It may also give an indication of how much of the information is lost in the image. If the GLCM values are high the resultant entropy values will be relatively equal.	Yu et al. (2006) Albregsten. (2008) Mohanaiah et al. (2013)
GLCM Standard Deviation	$Standard\ deviation(Std) = \sqrt{VA}$ $where\ VA == \sum_{i,j=0}^{N-1} i\ P_{i,j}\ (i - ME)^2$	GLCM standard deviation refers to the grey level standard deviation in the local window. The GLCM is considered to be high when the local region has a large grey level standard deviation.	Guo et al. (2004) Yu et al. (2006)
GLCM Angular Second Moment	$Angular\ Second\ Moment\ (ASM) = \sum_{i,j=0}^{N-1} i\ P_{i,j}^2$	The ASM is considered to be high when the pixels in the selected image are very similar in nature and is known as a measure of local homogeneity. The ASM is considered to be the opposite of Entropy regarding image texture characteristics.	Guo et al. (2004) Albregsten. (2008) Mohanaiah et al. (2013) Sarkar et al. (2013)
GLCM Correlation	$Correlation(CR) = \sum_{i,j=0}^{N-1} i\ P_{i,j} \left[\frac{(i - ME)(j - ME)}{\sqrt{VA_i\ VA_j}}\right]$	If the scale of the local texture of the image is large in nature it will result in the distance of the spatial correlation being high. Correlation is considered to measure the linear dependency of the grey levels of the neighbouring pixels in the image. If the local texture scale is smaller than the spatial scale there will be relatively low correlation between the pixel pairs.	Guo et al. (2004) Kayitakire et al. (2006) Mohanaiah et al. (2013) Sarkar et al. (2013)
GLCM Contrast	$Contrast(CO) = \sum_{i,j=0}^{N-1} i\ P_{i,j}\ (i - j)^2$	The contrast measures the amount of local variation within the selected image. It is considered to be the opposite of Homogeneity and it is high when the local region has a high contrast in the spatial scale.	Guo et al. (2004) Yu et al. (2006) Albregsten. (2008) Sarkar et al. (2013)
GLCM Dissimilarity	$Dissimilarity(DI) = \sum_{i,j=0}^{N-1} i\ P_{i,j}\ \lvert i - j\rvert$	Dissimilarity is very similar to contrast and is considered to have high contrast in the local region.	Yu et al. (2006) Albregsten. (2008) Sarkar et al. (2013)
GLCM Homogeneity	$Homogeneity(HO) = \sum_{i,j=0}^{N-1} i\ \frac{P_{ij}}{1 + (i - j)^2}$	This variable accounts for the smoothness in the texture of the image. Within the image if there are very large changes in the spectral values the resulting homogeneity values will be much smaller. The homogeneity is said to be high when the GLCM is concentrated along the diagonal.	Guo et al. (2004) Yu et al. (2006) Dye et al. (2012) Sarkar et al. (2013)
GLDV Mean	$GLDV\ Mean\ (GME) = \sum_{k=0}^{N-1} k\ V_k$	The GLDV Mean is considered to be mathematically equivalent to the Dissimilarity measure discussed above.	Guo et al. (2004) Yu et al. (2006) Sarkar et al. (2013)
GLDV Entropy	$GLDV\ Entropy\ (GEN) = \sum_{k=0}^{N-1} V_k\ (-\ln V_k)$	GLDV Entropy is considered to be high when all the elements in the image have similar values and is considered to be the opposite of the GLCM ASM discussed above.	Guo et al. (2004) Yu et al. (2006) Sarkar et al. (2013)
GLDV Contrast	$GLDV\ Contrast\ (GCO) = \sum_{k=0}^{N-1} k^2\ V_k$	The GLDV Contrast is considered to be mathematically equivalent to the GLCM Contrast measure discussed above.	Guo et al. (2004) Yu et al. (2006) Sarkar et al. (2013)
GLDV Angular Second Moment	$GLDV\ Angular\ Second\ Moment\ (GASM) = \sum_{k=0}^{N-1} V_k^2$	The GLDV ASM is similar to the GLCM ASM in the sense that it is a measure of the images' local homogeneity. The ASM is large for the GLDV when some of the present image elements are large and the remaining elements are small.	Guo et al. (2004) Yu et al. (2006) Sarkar et al. (2013)