Examining the utility of the random forest ensemble and remotely sensed image data to predict *Pinus patula* forest age in KwaZulu-Natal, South Africa.

By

MICHELLE DYE

203504670

Supervisor: Professor Onisimo Mutanga

Co-supervisor: Dr. Riyad Ismail

Submitted in fulfillment of the academic requirements for the degree of Master of Science in the Discipline of Geography in the
School of Environmental Sciences, Faculty of Science and Agriculture.
University of KwaZulu-Natal, Pietermaritzburg

May 2010

DECLARATION 1

This study was undertaken in fulfillment of a Geography Masters Degree and represents the original work of the author. Any work taken from other authors or organizations is duly acknowledged within the text and references chapter.

……..…

Michelle Dye

……...……

Professor Onisimo Mutanga
Supervisor

DECLARATION 2 – PUBLICATIONS

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part of and/or include research presented in this thesis (includes 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:** Dye, M.[1], Mutanga, O.[2], and Ismail, R.[3] (in preparation). Combining spectral and textural remote sensing variables using the random forest ensemble: Predicting the age of *Pinus patula* forests in KwaZulu-Natal, South Africa.

The work was done by the first author under the guidance and supervision of the second and third authors.

> **Publication 2:** Dye, M.[1], Mutanga, O.[2], and Ismail, R. [3] (in preparation). Examining the utility of the random forest ensemble and hyperspectral image data to predict *Pinus patula* age in KwaZulu-Natal, South Africa.

The work was done by the first author under the guidance and supervision of the second and third authors.

1, 2, 3 Discipline of Geography, School of Environmental Sciences, Faculty of Science and Agriculture, University of KwaZulu-Natal, Pietermaritzburg Campus, Private Bag X01, Scottsville, 3209, South Africa.

Signed: ……..

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The mapping of forest age is important for effective forest inventory as age is indicative of a number of plant physiological processes. Field survey techniques have traditionally been used to collect forest inventory data, but these methods are costly and time-consuming. Remote sensing offers an alternative which is time-effective and cost-effective and can cover large areas. The aim of this research was to assess the capabilities of multispectral and hyperspectral remotely sensed image data and the statistical method, random forest, for *Pinus patula* age prediction. The first section of this study used spatial and spectral data derived from multispectral QuickBird imagery to predict forest age. Five co-occurrence texture measures (variance, contrast, correlation, homogeneity, and dissimilarity) were calculated on QuickBird panchromatic imagery (0.6 m spatial resolution) using 12 moving window sizes. The spectral data was extracted from visible and near infrared (NIR) QuickBird imagery (2.4 m spatial resolution). Using the random forest ensemble, various methods of combining the spectral and texture variables were evaluated. The best model was achieved using backward variable selection which aims to find the fewest number of input bands while maintaining the highest predictive accuracy. Only five of the original 64 variables were used in the final model (R^2 = 0.68).

The second part of this study examined the utility of the random forest ensemble and AISA Eagle hyperspectral image data to predict *P. patula* age. Random forest was used to determine the optimal subset of hyperspectral bands that could predict *P. patula* age. Two sequential variable selection methods were tested: forward and backward variable selection. Although both methods resulted in the same root mean square error (3.097), the backward variable selection method was unable to significantly reduce the large hyperspectral dataset and selected 206 variables for the model. The forward variable selection method successfully reduced the large dataset to only nine optimal bands while maintaining the highest predictive accuracy from the hyperspectral dataset (R^2 = 0.6).

Overall, we concluded that (i) remotely sensed data can produce accurate models for *P. patula* age prediction, (ii) random forest is an effective tool for the combination of spectral and spatial multispectral data, (iii) random forest is an effective tool for variable selection of a high dimensional hyperspectral dataset, and

(iv), although random forest has mainly been used as a classifier, it is also a very effective tool for prediction.

ACKNOWLEDGMENTS

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

- First to my family and friends, thank-you all for your love and support throughout my studies, and for the endless encouragement. A special thanks to my father for his advice and help.

- I would like to thank Professor Onisimo Mutanga, for motivating me to do an MSc, and for his valuable input as my supervisor. Onnie, I am very grateful for your support and guidance.

- Thank-you to everybody in the geography department, to my fellow students for your encouragement, to Mrs. Ruth Howison for always being willing to help with technical issues. To Mr. Brice Gijsbertsen for help with field work and also offering valuable technical support.

- I would like to thank Sappi for providing the field data and imagery used in this study.

- Lastly, to Dr. Riyad Ismail, a very special thank-you. I so appreciated all the time and effort you have put into helping me throughout my studies.

Chapter one

Introduction

1.1. Background

The mapping of forest age is important for effective resource management and scientific research (Buddenbaum *et al.*, 2005). Forest biophysical attributes such as leaf area index (LAI), forest growth rate, canopy cover and net primary production are often related to forest age. Forest age has therefore been used as a surrogate for these variables (Ahern *et al.*, 1991; Danson and Curran, 1993, Niemann, 1995). Inventory data related to forest age has traditionally been collected by means of field surveys. Such surveys are time-consuming, labour intensive and costly (Gower *et al.*, 1999). Alternatively, remote sensing is a time-effective and cost-effective method of collecting forest attributes data for large areas (Cho *et al.*, 2009). Most large forestry companies in KwaZulu-Natal already have extensive databases that include forest age. However, the proposed techniques could be used for estimating age over larger areas, incorporating small growers who perhaps do not have age data for their forest plantations. Another application could be the estimation of forest age for illegal plantations that do not have formal records.

Multispectral remote sensing imagery has proved to be a useful and accurate tool for discriminating forest age (Jensen *et al.*, 1999; Cohen and Spies, 1992; Jakubauskas and Price, 2000; Gerylo *et al.*, 2002; Franklin *et al.*, 2003; Kayitakire *et al.*, 2006; Johansen *et al.*, 2007; Wunderle *et al.*, 2007). A common approach is to examine the tree's spectral response. Research has shown that the spectral response of a tree changes as it gets older due to changes in chlorophyll content and the internal structure of the plant (Jenson *et al.*, 1999). There is generally an inverse relationship between forest age and spectral response in the visible and near infrared (NIR) bands (Cohen and Spies, 1992; Jensen *et al.*, 1999; Gerylo *et al.*, 2002). However, spectral methods work on a per-pixel basis and do not take the spatial arrangements of objects into consideration (Franklin *et al.*, 2000). Structural changes in the stand cause variations in image texture and allow for strong relationships to be developed between forest age and canopy characteristics (Johansen *et al.*, 2007). As a result, several studies have focussed on image texture

for discriminating forest age (Franklin *et al.*, 2003; Johansen *et al.*, 2007; Wunderle *et al.*, 2007, Kayitakire *et al.*, 2006; Gerylo *et al.*, 2002; Jensen *et al.*, 1999).

Multispectral sensors typically collect data using three to six spectral bands from the visible and NIR regions of the spectrum (Govender *et al.*, 2008). With the advancement of hyperspectral sensor technologies, it is now possible to collect data from hundreds of narrow spectral bands, providing a full spectral curve for each pixel (Vane and Goetz, 1993). The contiguous nature of hyperspectral data makes it possible to identify features of interest by their characteristic reflectance signal (Goetz, 2009). Several studies have used hyperspectral remotely sensed data for forest age discrimination (van Aardt and Norris-Rogers, 2008; Buddenbaum *et al.*, 2005; Chan and Paelinckx, 2008). Research has shown that the red edge is very important for the mapping of forest structural variables (van Aardt and Norris Rogers, 2008; Blackburn, 2007; Treitz and Howarth, 1999; Schlerf *et al.*, 2005; Clark *et al.*, 2005). The red edge is the abrupt change in reflectance that occurs approximately between 670 nm and 780 nm in the vegetation spectra (Cho, 2007). The red edge position (REP) is the point of maximum slope in the red edge, and this position varies with forest age. According to Shafri *et al.* (2006) the REP of younger trees shifts towards the longer wavelengths while the REP of mature trees shifts towards the shorter wavelengths.

A common statistical approach for estimating forest biophysical variables (such as forest age) from remotely sensed data is the simple regression (Wulder, 1998). The method generally involves analyzing the empirical relationship between remotely sensed spectral data and stand variables using linear correlation and regression techniques (Gerylo *et al.*, 2002). A disadvantage of linear models is that they require a normal distribution and cannot deal with correlated input variables (Jensen *et al.*, 1999). Ecological data are complex and often involve nonlinear relationships between observed data and remotely sensed data (De'ath and Fabricius, 2000). More robust methods are therefore required. Regression trees have frequently been used for prediction purposes in remote sensing applications (DeFries *et al.*, 1997; Hansen *et al.*, 2002; Lobell *et al.*, 2007; Michaelson *et al.*, 1994). The studies have shown that regression trees overcome the limitations of traditional linear models. For example, when utilizing regression trees the input data (i) can be continuous or discrete and (ii) do not have to be normally distributed. Furthermore, regression trees can deal with non-linear relationships between

predictor variables and observed data (Bel *et al.*, 2009; De'ath and Fabricius, 2000; Prasad *et al.*, 2006). However, regression trees tend to be sensitive to small variations in the training dataset which can affect the overall predictive performance of the model (Breiman, 2001).

The random forest ensemble is an improvement on regression trees (Ismail, 2009). The ensemble grows many trees without pruning and the result is based on the average of all the trees in the ensemble. The trees are built using bootstrap aggregation (bagging) which involves randomly drawing, with replacement, a bootstrap sample of the original training dataset (Breiman, 2001). When a bootstrap sample is drawn, approximately one third of the data is excluded and the tree then makes predictions on the excluded samples to bring the sample to full size (Prasad *et al.*, 2006). The data which are excluded are known as the 'out-of-bag' (OOB) samples; while the replicated data are known as the 'in-bag' samples (Breiman, 2001). The random forest ensemble is easier to use than many other ensemble methods (for example boosting) because it only requires two user-defined parameters: the number of trees to be grown (*ntree*) and the number of variables used at each split (*mtry*) (Lawrence *et al.*, 2006; Ismail, 2009; Liaw and Wiener, 2002; Peters *et al.*, 2007). Sensitivity to these parameters is minimal and the default values are often a good choice (Liaw and Wiener, 2002). A key advantage of random forest is that the ensemble uses the OOB samples to calculate the model's predictive accuracy, making it unnecessary for an independent accuracy assessment (Breiman, 2001; Lawrence *et al.*, 2006). Furthermore, random forest uses the OOB samples to calculate an internal measure of variable importance (Strobl and Zeileis, 2008). Researchers have used this measure of importance to reduce the number of variables in a model to an optimal subset of variables that provide the best predictive accuracies (Ismail, 2009).

1.2. Aims and objectives

The aim of this research is to assess the utility of the random forest ensemble and remotely sensed data to predict *P. patula* forest age. The main objectives are as follows:

- To assess the capability of multispectral remotely sensed data to predict *P. patula* age
- To test the ability of the random forest ensemble to combine spectral and texture variables derived from multispectral imagery
- To assess the capability of hyperspectral remotely sensed data to predict *P. patula* age
- To evaluate the effectiveness of th*e* random forest ensemble to select the optimal subset of hyperspectral bands
- To compare multispectral and hypers*pe*ctral remotely sensed data to predict *P. patula* age

1.3. Outline of thesis

This thesis is presented in four chapters and structured mainly around two core chapters (chapter two and three) that form publishable papers and will be submitted to peer reviewed journals. Since both chapters have major sections dealing with the study area, literature review, and methodology, these sections are not covered in the introductory section of the thesis in order to avoid repetition.

Chapter two will assess the capabilities of multispectral remotely sensed data to predict *P. patula* age. The random forest ensemble will be used to develop a framework for combining spectral and texture variables that were derived from QuickBird imagery. Various combinations, based on window size, texture variables, principal component analysis and a backward variable selection procedure using random forest will be tested

Chapter three will evaluate the effectiveness of hyperspectral data and the random forest ensemble for predicting *P. patula* age. The random forest ensemble will be utilized to test the forward and backward variable selection methods to reduce the large number of AISA Eagle hyperspectral bands (n = 230) to an optimal subset of bands that can accurately predict the age of *P. patula* stands.

Chapter four provides a conclusion to the study. The aims and objectives of the research will be discussed in detail, highlighting important findings from the study. Additionally, the chapter will discuss the optimal remotely sensed data and techniques that are best suited for the predicting *P. patula* age. Finally, the chapter

examines the limitations to the study and makes recommendations for future research.

Chapter two

Combining spectral and textural remote sensing variables using the random forest ensemble: Predicting the age of *Pinus patula* forests in KwaZulu-Natal, South Africa

2.1. Abstract

In this study we examined the utility of the statistical technique, random forest, to combine spectral and texture variables to accurately predict the age *of Pinus patula* stands. Using the QuickBird panchromatic band (0.6 m), five texture variables (variance, contrast, correlation, homogeneity, and dissimilarity) were calculated using 12 moving window sizes. The spectral variables used in this study consisted of the QuickBird visible and near infrared (NIR) bands (2.4 m). Using the random forest ensemble, various methods of combining the spectral and texture variables were evaluated. The texture variables were combined with the spectral variables based on (i) window size, (ii) texture measure, (iii) principal components of the window sizes and (iv) principal components of the texture measures. Additionally, we tested the ability of all the spectral and texture variables (n = 64) to predict the age of *P. patula* stands using (i) the random forest ensemble and (ii) the random forest ensemble with a backward variable selection process. The best model was based on the random forest ensemble with a backward variable selection process. The model used only five variables (NIR, green, variance with a 3 x 3 window, red and blue) of the original 64 variables and obtained the best predictive accuracies (R^2 = 0.68). Overall, results indicated that the random forest ensemble is a flexible and robust method that provides an ideal framework for combining spectral and texture variables derived from high spatial resolution multispectral remotely sensed data.

Keywords: Texture, random forest, backward variable selection, forest age, *Pinus patula*.

2.2. Introduction

Forest age is an important component of forest inventory because it is indicative of a number of forest conditions (Franklin *et al.,* 2003; Jakubauskas and Price, 2000; Jensen *et al.*, 1999). Forest age cannot be directly measured using remotely sensed data but can be inferred through changes in forest structure (height, density, basal area) and biophysical properties such as biomass and leaf area index (Jakubauskas and Price, 2000). Studies have shown that multispectral remote sensing imagery is a very useful and accurate tool for discriminating forest age (table 1). A common approach to determining forest age is to examine the spectral reflectance of the stand (Jenson *et al.*, 1999; Gerylo *et al.*, 2002, Franklin *et al.*, 2003; Johansen *et al.*, 2007; van Aardt and Norris-Rogers, 2008; Gebreslasie *et al.*, 2008; Tomppo *et al.*, 2009). Researchers have shown that the spectral response of a tree changes as it gets older due to changes in chlorophyll content and the internal structure of the plant (Jenson *et al.*, 1999). In general, there is an inverse relationship between spectral reflectance and forest age because as trees grow older, competition for light, water and nutrients causes weaker trees to die off. This results in (i) a decrease in stems per hectare, and (ii) an increase in the size and visibility of shadows (Ahern *et al.*, 1991; Danson and Curran, 1993; Niemann, 1995; Jensen *et al.*, 1999; Jakubauskas and Price, 2000; Gerylo *et al.*, 2002; Franklin *et al.*, 2003). Spectral approaches work on a per-pixel basis and do not take into consideration the spatial arrangements of objects in an image (Franklin *et al.*, 2000). Consequently, alternative remote sensing techniques that incorporate spatial information have been explored (Franklin *et al.*, 2003; Johansen *et al.*, 2007; Wunderle *et al.*, 2007).

Texture analysis is a spatially-based image transformation technique which enhances the spatial relationship between pixels (St-Louis *et al.*, 2006). This technique is useful for predicting forest age because structural changes in the stand cause variations in image texture and allow for strong relationships to be developed between age and canopy characteristics (Johansen *et al.*, 2007). Studies have shown that image texture can successfully discriminate between forest age classes (Kayitakire *et al.*, 2006; Franklin *et al.*, 2001), but more importantly, image texture has also been used in conjunction with spectral reflectance to provide additional discriminatory power. For example, Johansen *et al.* (2007) combined spectral

reflectance and image texture derived from high spatial resolution imagery to discriminate between the different stages of riparian forest growth. The inclusion of image texture improved the classification accuracy of vegetation classes between 2% and 19%. Similarly, Franklin *et al.* (2000) combined image spectral reflectance and texture to classify the composition of forest species in Alberta and New Brunswick. Results showed that the inclusion of image texture increased the overall classification accuracy by 5% in Alberta, and 12% in New Brunswick. Wunderle *et al.* (2007) used pan-sharpened SPOT 5 imagery to estimate forest stand structure. Image texture was included in the study to compliment the spectral response data and increase model accuracy. A stepwise multivariate regression analysis was performed using both the spectral and texture variables and resulted in a high model accuracy ($R^2 = 0.79$).

However, remote sensing studies that have combined spectral and texture variables have tended to use more traditional statistical techniques. Frequently used methods include linear regression models (Jakubauskas and Price, 2000), and linear discriminant analysis (Franklin *et al.*, 2003; Zhang *et al.*, 2004; Moskal and Franklin, 2001). Linear models require normal distribution, linearity, and the absence of collinearity amongst input variables (Jensen *et al.*, 1999). However, ecological data are often very complex and nonlinear interactions may exist between the observed data and the remotely sensed data (De'ath and Fabricius, 2000). For example, in the study by Jensen *et al.* (1999), a nonlinear relationship was shown to exist between the age of loblolly pine and near infrared (NIR) reflectance, with NIR reflectance decreasing at a faster rate as the tree matured. Therefore, more robust methods are required to handle remotely sensed data and their interactions with the response data (i.e. forest age).

Regression trees (Breiman *et al.*, 1984) have been recommended to overcome the limitations of linear based models (Prasad *et al.*, 2006) and have been widely used for prediction purposes in the remote sensing domain (DeFries *et al.*, 1997; Hansen *et al.*, 2002; Lobell *et al.*, 2007; Michaelson *et al.*, 1994). The method is popular amongst remote sensing researchers because (i) the input data can be continuous or discrete, (ii) the input data do not need to be normally distributed, and (iii) non-linear relationships between predictor variables and observed data can be modeled (Bel *et al.*, 2009; De'ath and Fabricius, 2000; Prasad *et al.*, 2006). However, regression trees are sensitive to small variations in the training dataset

(Breiman, 2001). This can lead to instability with regards to variable selection, and can adversely affect the predictive performance of the final model (Elith *et al.*, 2008). Consequently, ensemble methods such as random forest (Breiman, 2001) have been developed to reduce the instability of single regression trees and improve the overall predictive performance (Pal, 2005).

The random forest ensemble grows many regression trees and the final result is based on the average of all the regression trees (Lawrence *et al.*, 2006). Random forests have several advantages over single regression trees. First, the approach is robust to over-fitting (Lawrence *et al.*, 2006). Second, it is also relatively easy to use. No pruning of trees is necessary, and there are only two user defined parameters, the number of trees (*ntree*) and the number of possible splitting variables (*mtry*) (Prasad *et al.*, 2006). Third, random forest provides an internal estimate of the models predictive accuracy using the out-of-bag (OOB) sample data, making it unnecessary to have an independent accuracy assessment dataset (Breiman, 2001; Lawrence *et al.*, 2006). Finally, random forest provides an internal measure of variable importance based on the multivariate interactions between variables (Strobl and Zeileis, 2008). While other machine learning methods such as neural networks have been successfully used to accurately predict forest age (Jensen *et al.*, 1999; Gebreslasie *et al.*, 2008), random forest is valuable as it provides insight regarding the spectral and texture variables that best contribute to the final model (Strobl and Zeileis, 2008).

The aim of this paper is to develop a framework based on the random forest ensemble that will incorporate spectral and texture remote sensing data to predict the age of *P. patula* stands in KwaZulu-Natal. Various combinations of spectral and texture variables will be evaluated based on window size, texture variable, principal component analysis and a backward variable selection procedure using random forest ranked variable importance. The best combination will then be tested using an independent test dataset in order to assess the model's predictive performance.

Table 1. Key studies using multispectral remote sensing imagery for mapping forest age.

Reference	**Sensor (resolution)**	**Age (yrs)**	**Species**	**Method**	**Input variables**	**Overall accuracy (%)**
Jensen *et al.* (1999)	Landsat TM (30m)	4-40	- Loblolly pine (*Pinus taeda*)	Regression analysis using multiple regression and artificial neural networks	Spectral bands	98.2%
Jakubauskas and Price (2000)	Landsat 5 TM (30m)	0-250	- Lodgepole pine (*Pinus contorta*)	Regression analysis using stepwise multiple regression	Spectral bands	90%
Gerylo *et al.* (2002)	Landsat 5 TM (30m)	21-198	- Jack pine (*Pinus banksiana*) - Trembling Aspen (*Populus tremuloides*)	Regression analysis using stepwise multiple regression	Spectral bands	75% (Jack pine) 44% (Trembling Aspen)
Franklin *et al.* (2003)	Landsat 5 TM (30m)	21-198	- White spruce (*Picea glauca*) - Jack pine (*Pinus banksiana*)	Classification using linear discriminant analysis (LDA)	Texture images and spectral bands	92% (Jack pine) 63% (White spruce)
Kayitakire *et al.* (2006)	IKONOS-2 (4m)	27-110	- Norway spruce (*Picea abies*)	Regression analysis using simple linear model	Texture images	81%
Johansen *et al.* (2007)	QuickBird (2.4m)	Young Mature Old	- Western hemlock (*Tsuga heterophylla*) - Western red cedar (*Thuja plicata*) - Amabilis fir (*Abies amabilis*) - Yellow cedar (*Chamaecyparis nootkatensis*) - Sitka spruce (*Picea sitchensis*) - Douglas fir (*Pseudotsuga menziesii*) - Red alder (Alnus *rubra*)	Classification using object-oriented classification algorithm	Texture images and spectral bands	78.9%
Wunderle *et al.* (2007)	SPOT-5 (10m)	0-50	- Lodgepole pine (*Pinus contorta*) - White spruce (*Picea glauca*) - Balsam fir (*Albies balsamea*) - Trembling aspen (*Populus tremuloides*)	Regression analysis using stepwise multiple regression	Texture images and spectral bands	74%

2.3 Materials and methods

2.3.1 Study area

The study area (figure 1) consists of 6391 ha of commercial forestry and forms part of the Hodgsons Sappi plantation which is located near Greytown in KwaZulu-Natal, South Africa (Centroid: Latitude 29º13'42"S Longitude 30 º29'56"E). The study area falls under the midlands mist belt grassland bioregion unit as defined by Mucina and Rutherford (2006). The area experiences summer rainfall and has a mean annual precipitation of 915 mm (range 730-1280 mm). Some of the winter, spring and early summer precipitation are in the form of cold front activity. Frequent and heavy mist provides significant amounts of additional moisture. The mean annual temperature is 15.8ºc. The dominant soils in the study area are apedal and plinthic soil forms derived mostly from the ecca group. The landscape is classified as hilly and rolling (Mucina and Rutherford, 2006) with elevation ranging from 1030 m to 1590 m above sea level. The study site contains same-aged stands (compartments) of *Acacia*, *Eucalyptus* and *Pinus* species. However, the majority of the stands consist of *P. patula* trees that are grown under a pulpwood management regime. The majority of the stands are between 1 and 25 years old, with harvesting typically occurring when the *P. patula* trees are 25 years old (Owen, 2000).

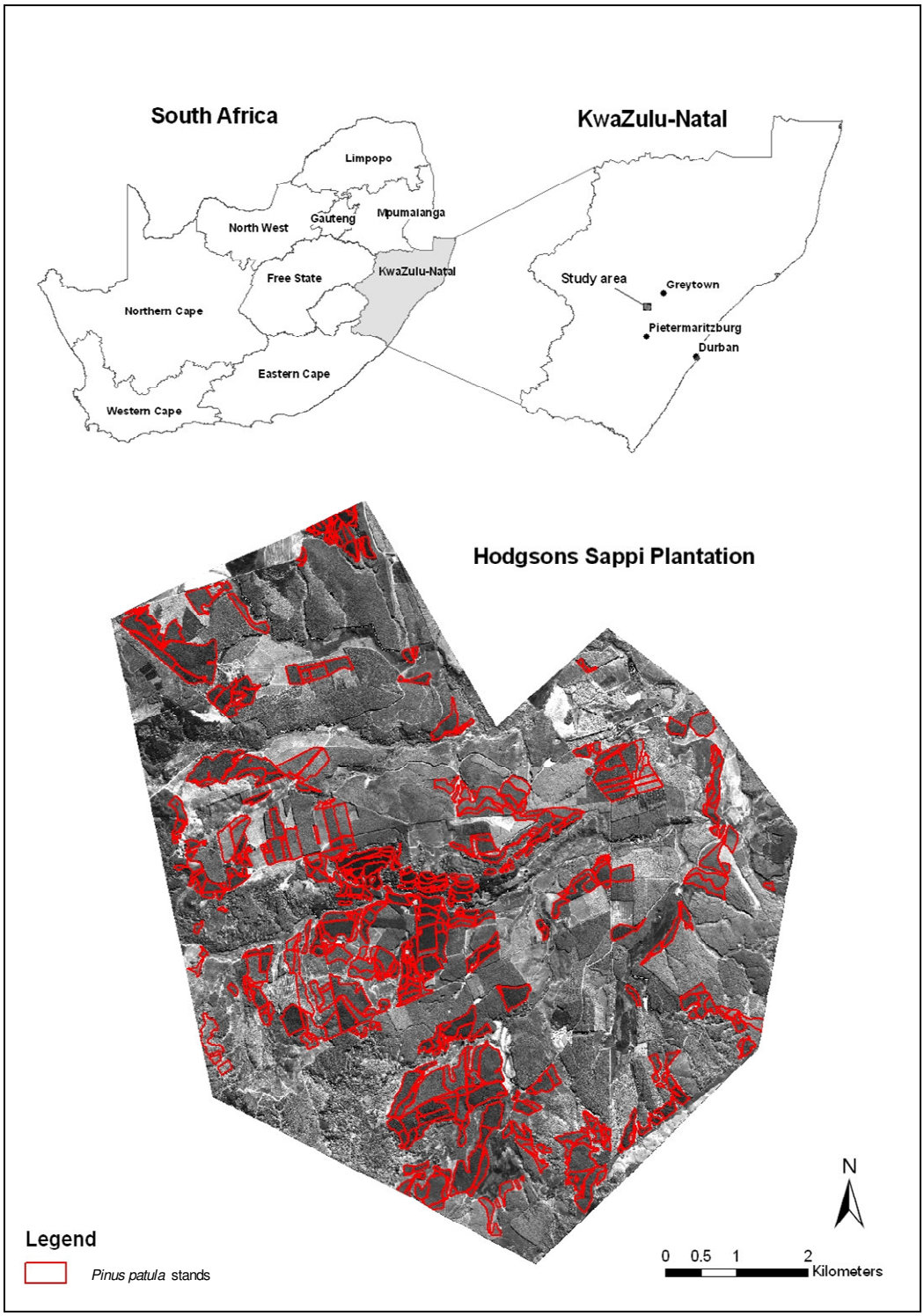


Figure 1. Location of the study area. The *Pinus patula* stands (n= 142) in the study area are highlighted in red.

2.3.2. Data acquisition

QuickBird multispectral (2.4 m resolution) and panchromatic (0.6 m resolution) imagery was acquired on the 10th of September 2008 under cloudless conditions (table 2). The images were georectified using the random polynomial correction (RPC) QuickBird model (Johansen *et al*., 2007). RPC's were calculated from a digital photogrammetry technique that uses a collinearity equation to construct sensor geometry (ENVI, 2006). The RPC model uses sensor information as well as elevation data to orthorectify an image. The digital elevation model (DEM) used in the rectification process was created from 5 m contours of the study area. Following Johansen *et al*. (2007), the image was also atmospherically corrected using pre-launch calibration coefficients as described in the ENVI 4.3 image processing software.

Table 2. Spectral and spatial resolution of the QuickBird imagery used in the study.

Band	Colour	Spectral range (nm)	Spatial resolution (m)
1	Blue	450-520	2.4
2	Green	520-600	2.4
3	Red	630-690	2.4
4	NIR	760-900	2.4
Panchromatic	Pan	450-900	0.6

Source: http://www.digitalglobe.com/index.php/85/QuickBird .

2.3.3. Field data

Field data for the study area was supplied by Sappi, a paper and pulp company. Field data comprising of species composition, age, stems per hectare, mean diameter at breast height and mean tree height were collected by the forestry company using industry standard enumeration techniques (Owen, 2000).

Following Jensen *et al* (1999), the field data served as ground reference data that were used to predict *P. patula* age. Stands that were three years and younger were not used in the study due to their spectral similarity to bare soil and grass (Jensen *et al.*, 1999). After removing outliers such as newly cut stands, and stands that were not consistent with the ground reference data, the final dataset consisted of 142 samples (1214 ha). Age ranged from 4 to 24 years; however the majority of the samples consisted of younger stands (4 to 12 years) (figure 2). Using Hawthorne's analysis tools (www.spatialecology.com/htools/overview.php) in ArcGIS 9.1 (ESRI, 2006), 70% of the samples were randomly selected and used as a training dataset while the remaining 30% of the samples were used to test the final model (Ismail and Mutanga, 2010).

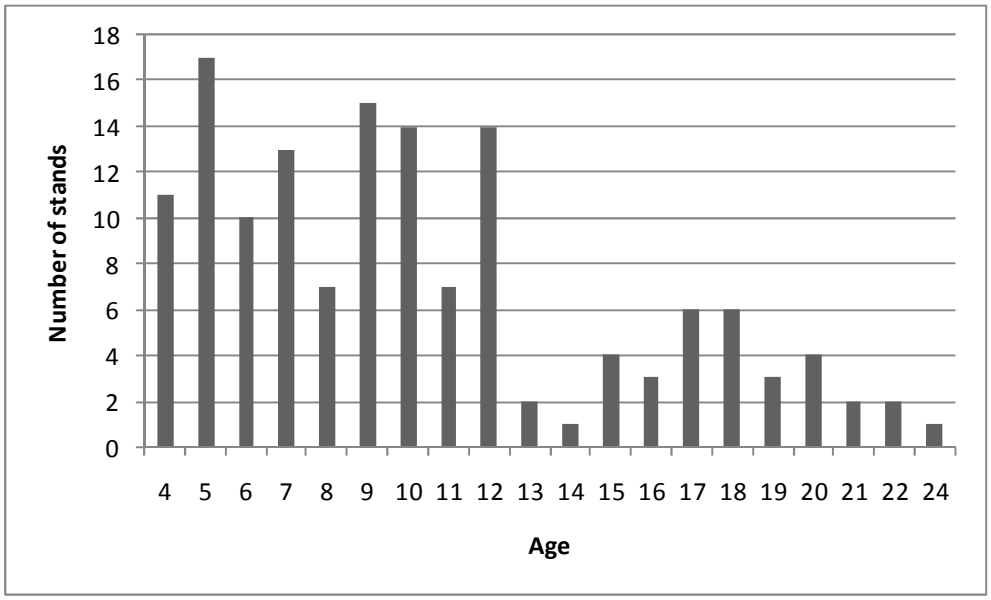


Figure 2. Age distribution of the *Pinus patula* stands (n = 142) located in study area.

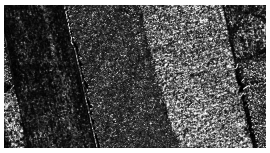
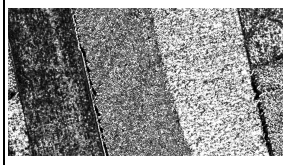
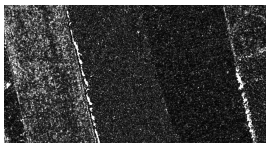
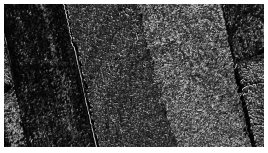
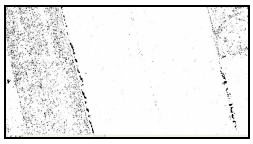
2.3.4. Texture analysis

Haralick *et al*. (1973) defined 14 texture statistics that are derived from the grey level co-occurrence matrix (GLCM). The approach describes the probability of any grey level occurring spatially relative to any other grey level within a moving

window. The probabilities are stored in a GLCM and statistics are applied to the matrix to generate texture features which are assigned to the centre pixel of the window (Jobanputra and Clausi, 2006). Johansen *et al.* (2007) used texture analysis and spectral reflectance data to discriminate between old and young forest in British Columbia, Canada (table 1). Second-order contrast, dissimilarity, and homogeneity provided the most significant discrimination between the age classes. Kayitakire *et al.* (2006) utilized second-order variance, contrast and correlation to discriminate the age of common spruce stands in Belgium. Therefore, based on findings of Johansen *et al.* (2007) and Kayitakire *et al.* (2006), the following five co-occurrence texture measures were calculated from the QuickBird panchromatic image: variance, contrast, correlation, homogeneity, and dissimilarity. Table 3 provides a detailed description of the texture measures used in this study.

Window sizes are an important component of a texture analysis because texture is a multi-scale phenomenon (Moskal and Franklin, 2001). Using small window sizes could result in poorly sampled co-occurring probabilities and an inconsistent estimate of individual texture measures; while focusing on only larger window sizes could result in the eroding of class boundaries (Jobanputra and Clausi, 2006). It is therefore necessary to use a range of small, medium and large window sizes. Johansen *et al.*, (2007) showed that semivariograms can be used to obtain the optimal window size for a texture analysis. However, in this study we followed the recommendation of Moskal and Franklin (2001) by calculating texture using multiple window sizes. The texture variables used in this study were computed using 12 window sizes (3 x 3, 5 x 5, 7 x 7, 9 x 9, 11 x 11, 13 x 13, 15 x 15, 17 x 17, 19 x 19, 21 x 21, 23 x 23, 25 x 25) and the mean value for each sample (n = 142) was extracted using the zonal statistics functionality in ArcGIS 9.1 (ESRI, 2006).

Table 3. Co-occurrence texture measures used in this study (Adapted from Dye *et al.* 2008). Where *i* and *j* are equal, the cell is on the diagonal and (*i-j*) = 0.

Texture measure	**Formula**	**Description**	**Example: (3 x 3 window)**
Contrast	$\sum_{i,j=0}^{N-1} P_{i,j} (i-j)^2$	Contrast is a measure of the overall amount of local variation in a window (i.e., it is proportional to the range of grey levels) (Yuan *et al*, 1991).	
Dissimilarity	$\sum_{i,j=0}^{N-1} P_{i,j} \lvert i-j \rvert$	The dissimilarity measure is similar to the contrast measure. However, where contrast weights increase exponentially (0, 1, 4, 9, etc.) as one moves away from the diagonal, dissimilarity weights increase linearly (0, 1, 2, 3 etc.) (Hall-Beyer, 2010).	
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$	Homogeneity measures the smoothness of image texture. Large changes in spectral values will result in very small homogeneity values, while small changes will result in larger homogeneity values (Tuttle *et al.*, 2006).	
Variance	$\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j} (i-\mu_i)^2$ $\sigma_j^2 = \sum_{i,j=0}^{N-1} P_{i,j} (j-\mu_j)^2$	Accounts for the variability of the spectral response of pixels (Tuttle *et al*, 2006) but considers the pairwise combinations of variability.	
Correlation	$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(i-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$	The correlation texture algorithm measures the grey level linear-dependency within the image (Kayitakire *et al*, 2006).	

2.3.5. Random forest for regression applications

Random forest (Breiman, 2001) is an ensemble method that grows many regression trees without statistical pruning, and the result is based on the average of all the regression trees. Individual regression trees in the forest are built using bootstrap aggregation (bagging) which involves randomly drawing, with replacement, a bootstrap sample of the original training dataset (Breiman, 2001). When a bootstrap sample is drawn, approximately one third of the data are excluded and the tree then makes predictions on the excluded samples to bring the sample to full size (Prasad *et al.*, 2006). The excluded one third of the samples is known as the 'out-of-bag' (OOB) samples; while the replicated dataset is known as the 'in-bag' samples (Breiman, 2001). Given that the OOB samples were not used in the training process, calculating the mean square error (MSE) using the OOB samples provides an unbiased assessment on the model's predictive accuracy (Prasad *et al.*, 2006; Lawrence *et al.*, 2006). By aggregating the OOB predictions of all trees in the forest, the MSE can then be estimated (Liaw and Wiener, 2002) as follows:

$$MSE^{OOB} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{OOB}\right)^2}{n} \qquad \text{(Equation 1)}$$

Where n is the number of samples, Y_i is the observed forest age and Y_i^{OOB} is the average of the OOB predictions for the ith observation.

In addition to bagging, each tree split is based on a random subset of the input variables (Breiman, 2001). Therefore, the randomness introduced in the dataset selection (i.e. bagging) and in the variable selection (*mtry*) makes random forests an accurate tool for prediction (Breiman, 2001; Liaw and Wiener, 2002; Peters *et al.*, 2007). Additionally, there are only two tuning parameters required for growing random forests, the number of trees to be grown (*ntree*), and the number of possible splitting variables *(mtry*) which are sampled at each node (Peters *et al.*,

2007). However, researchers have shown that sensitivity of the user defined parameters is minimal and the default values are often a good choice (Lawrence *et al.*, 2006; Liaw and Wiener, 2002; Ismail and Mutanga, 2010). We used the R statistical software (R Development Team, 2008) and the random forest libraries (Liaw and Wiener, 2002) for all statistical analysis.

2.3.6. Random forest: Variable importance

The random forest ensemble is often referred to as a 'black box' (Prasad *et al.*, 2006) because of the limited interpretability of the final model (i.e. the results are based on the average of many trees in the forest). Consequently, the random forest ensemble also produces a measure which ranks the variables according to their importance. Simply stated, the variables associated with the OOB sample are randomly permuted and regression trees are grown on the modified dataset. The importance measure of each variable is then calculated as the difference in the MSE between the original OOB dataset and the modified dataset (Breiman, 2001). It follows that the MSE will decrease substantially if the original variable (i.e. texture or spectral variables) was associated with the response variable (i.e. forest age). Therefore, the difference in MSE before and after permuting the variables can be used to measure the importance of variables used in the final random forest model (Breiman, 2001). A key advantage of the random forest variable importance is that it not only deals with the impact of each variable individually, but also looks at multivariate interactions with other variables (Strobl and Zeileis, 2008).

2.3.7. Random forest as a framework for incorporating texture and spectral data

The framework used to combine texture and spectral variables was adapted from Puissant *et al.* (2005). Figure 3 shows the four methods that Puissant *et al.* (2005) used to combine spectral and texture variables in an effort to improve classification accuracy. The first method involved combining spectral and texture

variables according to window size (for example 3 x 3 contrast, 5 x 5 contrast and so on) while the second method combined spectral and texture variables according to the optimal texture measure (for example 3 x 3 contrast, 3 x 3 variance and so on). In total, 12 random forest models were developed based on the first method (i.e. optimal window size) and five random models were developed based on the second method (i.e. optimal texture measure). Due to the fact that the window sizes and texture measures are highly correlated (St-Louis *et al*., 2006), Puissant *et al.* (2005) transformed the texture variables used in method one and method two using principal component analysis (PCA). PCA is an image enhancement technique which uses a linear transformation of a set of numerical variables in order to create a new set of components that are uncorrelated and ordered in terms of the amount of variance explained in the original data (Eastman and Fulk, 1993). The first three components which explained 98% of the variance from method one and method two were subsequently extracted and combined with the spectral variables.

In the original framework developed by Puissant *et al.* (2005) the coefficient of variation was used to select the optimal model for each method. However, in this study we used the internal measure of error calculated by the random forest ensemble to select the optimal model from the methods proposed by Puissant *et al.* (2005). Generally, random forest uses the MSE to assess the accuracy of a model, however, since MSE is scale dependant, model selection was based on the normalised out-of-bag (NOOB) error (Grimm *et al*., 2008) which is calculated as follows:

$$\text{NOOB error} = \frac{MSE^{OOB}}{VAR(Yk)} \qquad \text{(Equation 2)}$$

Where *VAR(Yk)* is the variance of the response variable (forest age). MSE^{OOB} (Equation 1) refers to the mean square error as determined by the out-of-bag samples.

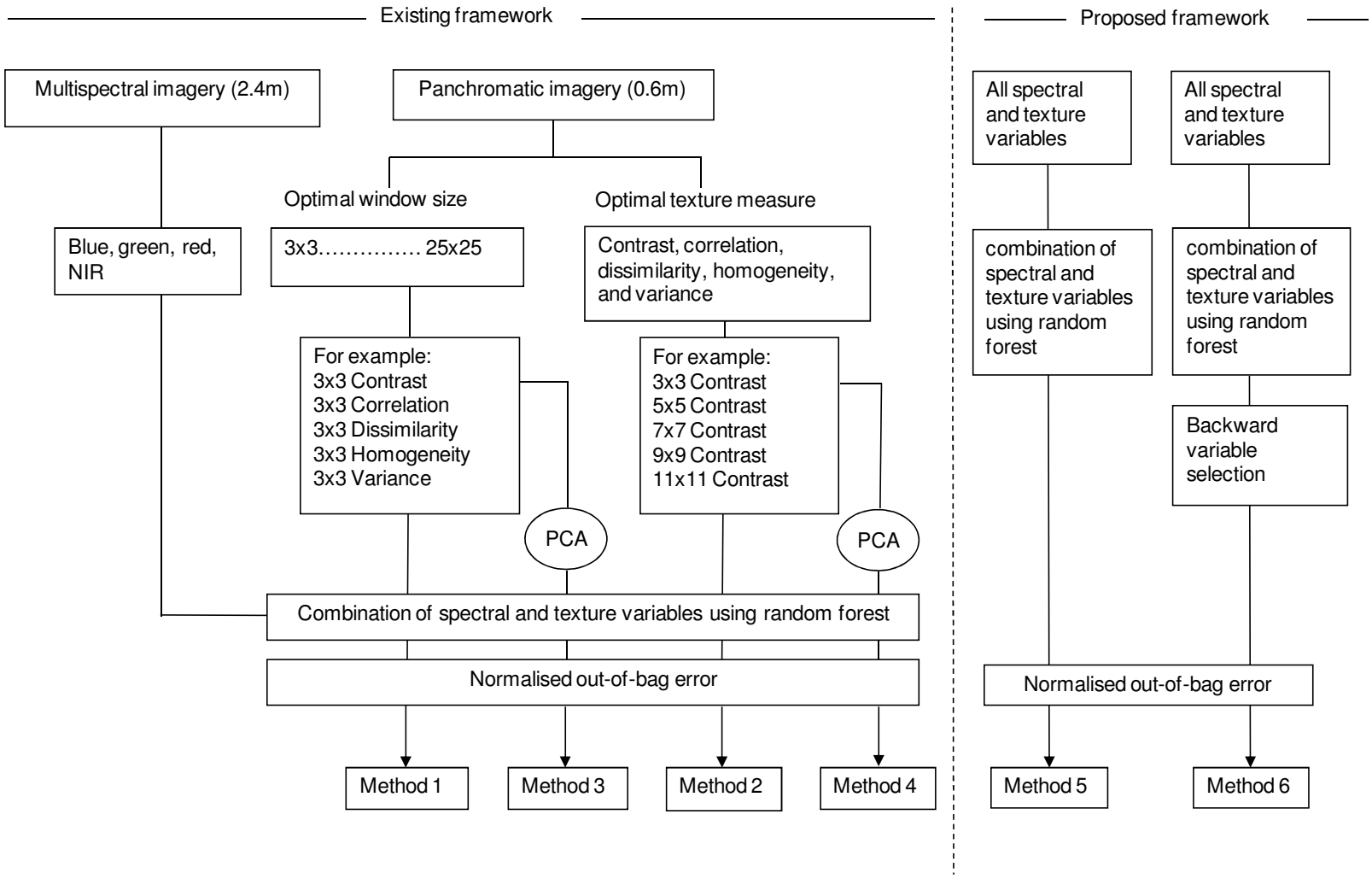


Figure 3. Framework for combining texture and spectral remote sensing data. PCA refers to the principal components analysis.

In addition to the combination framework developed by Puissant *et al.* (2005), we propose two additional methods (method 5 and method 6) for combining texture and spectral variables (figure 3). Method five combines all the texture (window sizes and texture measures) and spectral variables (n = 64) using the random forest ensemble, and examines the final accuracy as determined by the NOOB error. Method six is similar to method five but uses the random forest ensemble with backward variable selection (Strobl and Zeileis, 2008; Kuhn, 2009). Initially, method six builds a random forest using all the variables (n = 64), ranks the importance of each variable and then calculates the model performance using the NOOB error. The model is then run again by dropping the least important variable, and model performance is again calculated. This process continues until there are no variables left. Subsequently, the model with the lowest NOOB error is the best model with the most significant variables (Kuhn, 2009). The rationale of using the backward variable selection with the random forest ensemble is that prediction accuracy will stay relatively constant when unimportant variables are dropped, but will decrease when relevant ones are excluded (Strobl and Zeileis, 2008).

2.4. Results

2.4.1. Combining spectral and texture variables using optimal window size

We tested the utility of the first method (figure 3) by combining the spectral variables with different texture measures that were derived using the same window size. As mentioned earlier, texture measures (n = 5) were calculated using 12 window sizes. The texture measures (e.g. 3 x 3 contrast, 3 x 3 variance, 3 x 3 correlation and so on) were then combined with the spectral variables (n = 4). From the 12 models that were created (figure 4.a), the texture measures calculated using a 3 x 3 window size when combined with the spectral variables yielded the lowest NOOB error (30.45%). Also noticeable from figure 4a

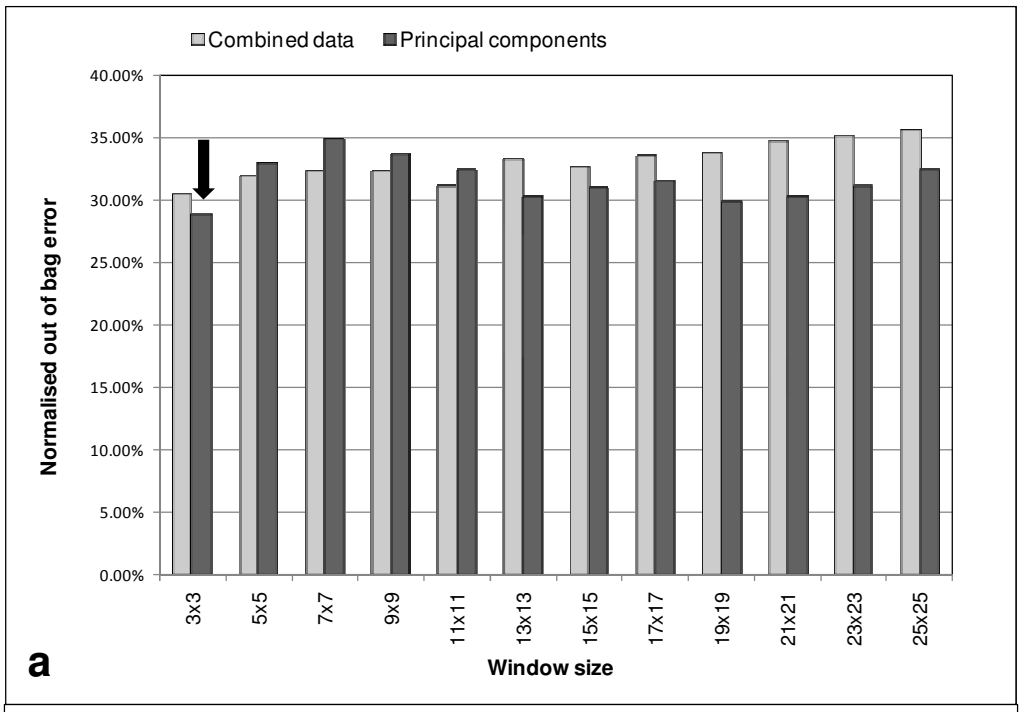
is that the larger window sizes have higher NOOB error rates when compared to the smaller window sizes.

The first three principal components of the texture measures that were derived using the same window size were also combined with the spectral variables (method 3). In the majority of the models, combining the spectral variables with the principal components of the texture variables produced better results than using the original texture variables. Figure 4a shows that eight out of the twelve models that used principal components of the texture variables yielded lower NOOB error rates. Overall, the principal components of the texture measures calculated using a 3 x 3 window when combined with the spectral variables produced the lowest NOOB error (28.78%).

2.4.2. Combining spectral and texture variables using optimal texture measure

We tested the utility of the second method (figure 3) by combining the spectral variables with individual texture measures calculated at different window sizes (e.g. 3 x 3 contrast, 5 x 5 contrast, 7 x 7 contrast and so on). From the five models that were created (figure 4b), the model that used the variance texture measure produced the lowest NOOB error (30.71%). The NOOB error rates for the other texture measures were as follows: contrast (35.67%) and dissimilarity (36.63%), homogeneity (37.64%) and correlation (39.22%).

The first three principal components of individual texture measures calculated at different window sizes were also combined with the spectral variables (method 4). Results show that four out of the five principal components models produced better results than using the original texture variables. Figure 4b shows that the model that used the principal components of the contrast texture measure produced the lowest NOOB error (33.75%) while the principal components of variance texture measure when combined with the spectral variables yielded the highest NOOB error (34.57%).



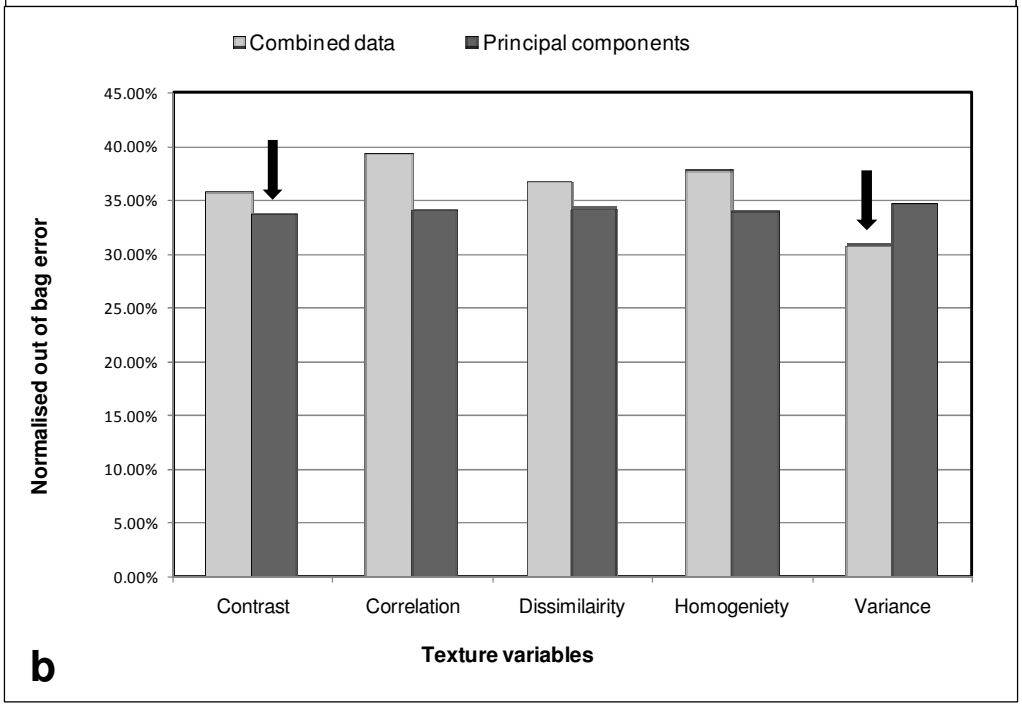


Figure 4a. Combining the spectral variables with various window sizes calculated from the original data and the principal components of the various window sizes. Figure 4b. Combining the spectral variables with various texture measures using the original texture images and the principal components of the texture images. The arrows indicate the lowest error obtained.

2.4.3. Combining all the spectral and texture variables using random forest

Method five of the proposed framework combined all the spectral and texture variables (*n* = 64) using the random forest ensemble. Using the default parameters (*mtry* = 8; *ntree* = 500) the random forest ensemble produced a NOOB error of 34.30%. Figure 5 shows the ranked variable importance as determined by the mean decrease in accuracy. The NIR band was the most important variable with a 15% decrease in accuracy, followed by the green band (11.68%) and the variance texture measure calculated using a 3 x 3 window (11.09%). Excluding the variance texture measure, the spectral bands are more highly ranked than the texture variables used in this study.

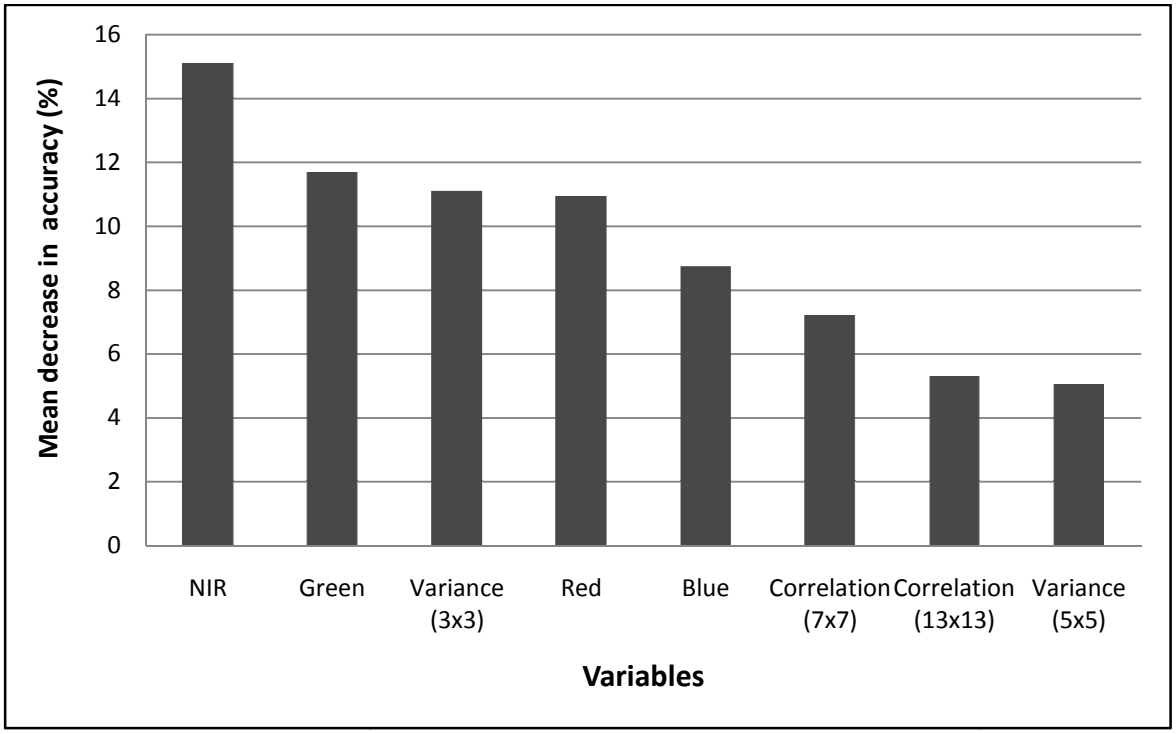


Figure 5. Ranked variable importance as determined by the random forest ensemble. For brevity only variables (n = 8) that had greater than a 5% decrease in accuracy are shown.

2.4.4. Backward variable selection

Method six of the proposed framework combined all the spectral and texture variables (n = 64) using the random forest (*mtry* = 8; *ntree* = 500) ensemble and a backward variable selection process. Using the ranked variable importance (figure 5), the least important variable was dropped and the NOOB error was recalculated. This process was repeated 64 times until there were no more variables to drop (figure 6). Results showed that the best model obtained a NOOB error of 27.72% while using only five variables out of the 64 original variables. The five variables selected by the backward variable selection were the NIR, green, variance (3 x 3 window), red and blue variables. Noticeable from figure 6 is that using all the variables (texture and spectral) does not necessarily improve the model's predictive accuracy; rather there are an optimal number of variables that produce the best results.

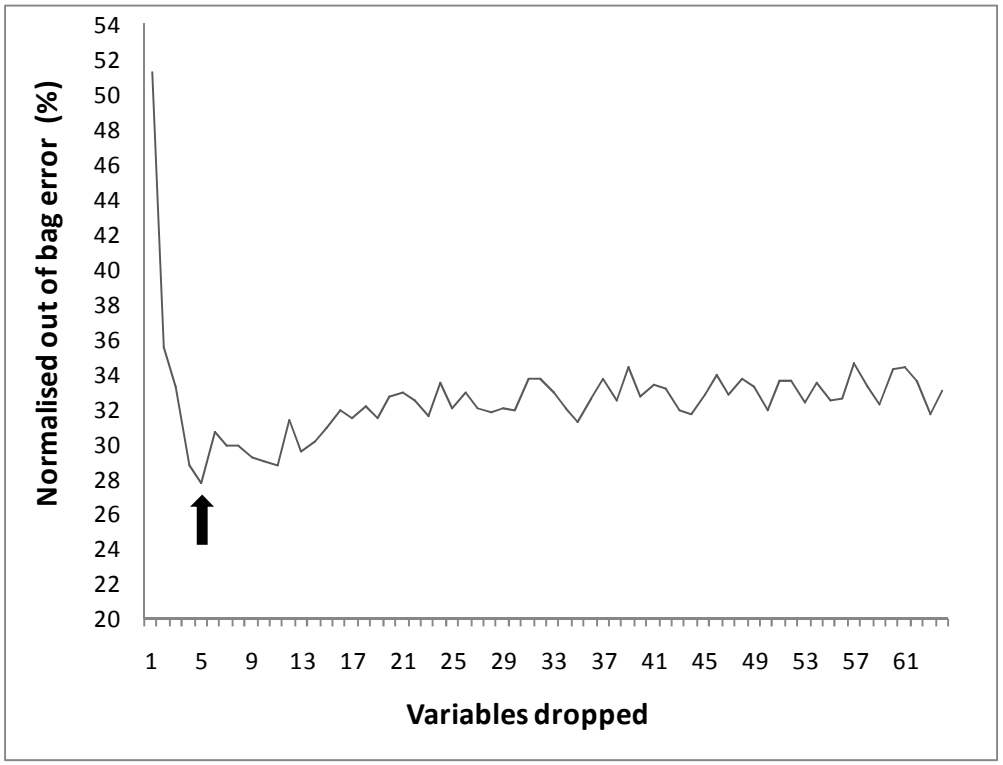


Figure 6. Results of the backward variable selection method and the associated normalised out-of-bag error. The arrow indicates the lowest error obtained using the following five variables: NIR, green, variance (3 x 3 window), red and blue.

2.4.5. Accuracy assessments

Table 4. Summary of the normalised out-of-bag error (NOOB) for the best models for various methods evaluated in this study.

Method	Number of models created	Texture variables used in the best model	NOOB error of the best model
1	12	3 x 3 window	30.45%
3	12	PCA of 3 x3 window	28.78%
2	5	Contrast	33.75%
4	5	PCA of variance	30.71%
5	1	All the texture measures	34.30%
6	64	3 x 3 variance	27.72%

Ideally, model performance is calculated using a large independent test dataset that was not used in the training procedure (Grimm *et al.*, 2008). Therefore, the predictive accuracy (R^2) of the model that produced the lowest NOOB error was tested using 30% of the dataset (n = 43) that was excluded from the modeling process. Overall, the best model with the lowest NOOB error (27.72%) was based on the backward variable selection method (method 6) and yielded a R^2 value of 0.68. For comparative purposes we also calculated the predictive accuracy of the random forest model that used all the texture and spectral variables. It is interesting to note that the backward variable selection method produced a 5% increase in predictive accuracy when compared to the random forest method that used all the texture and spectral variables (R^2 = 0.63).

2.5. Discussion

This study has shown the potential of the random forest ensemble to provide a framework for combining spectral and texture variables to accurately predict the age of *P. patula* stands. The random forest ensemble provides an ideal framework for the integration of spectral and texture data because unlike traditional linear methods which require certain statistical assumptions to be met,

random forest is very robust and can handle complex remotely sensed data. Remote sensing studies utilizing the random forest ensemble have focussed primarily on classification applications (Cutler *et al*., 2007; Lawrence *et al*., 2006; Pal, 2005; Peters *et al*., 2007; Chan and Paelinckx, 2008). However, recent studies have successfully used the ensemble for prediction purposes (Prasad *et al*., 2006; Ismail, 2009; Grimm *et al*., 2008).

2.5.1. Window size versus texture measure

More specifically, an existing framework proposed by Puissant *et al.* (2005) for combining spectral and texture variables was evaluated using the random forest ensemble. The models that combined individual texture measures calculated at various window sizes with the spectral data yielded better results (30.45% NOOB error) than the models that combined various texture measures according to single window size (33.75% NOOB error). The results from this study reiterate the importance of window sizes when implementing texture analysis (Moskal and Franklin, 2001). Smaller window sizes will capture the texture information of individual trees, while the larger window sizes will describe the textural characteristics of forest stands (Moskal and Franklin, 2001; Kayitakire *et al*., 2006). In this study the smaller window sizes proved to be the most suitable for forest age prediction. This is perhaps due to the fact that the trees in the study area are grown for pulpwood and are therefore planted close together. On a typical growth cycle for pulpwood, *P. patula* trees are planted at 1200 stems per hectare (SPH). Even after 25 years the SPH is still high at 950 (Owen, 2000). Smaller window sizes are therefore more appropriate as they provide more detailed textural information of individual trees. Similar results were obtained by Johansen *et al.* (2007) and Wunderle *et al.* (2007). Johansen *et al.* (2007) used semivariograms to assess which window sizes were most appropriate for the separation of vegetation structural stages. Results showed that the 3 x 3 and 11 x 11 window sizes were most appropriate for structural class separation. The 3 x 3 window size was most suitable for discrimination between old and young forest

classes. According to Wunderle *et al.* (2007) there is a relationship between forest structure and textural measures using the larger window size. However the smaller window size (5 x 5 pixels) adds the most information when estimating forest stand structure.

2.5.2. Original texture data versus principal components

It follows that by implementing PCA, the data redundancy inherent in the texture images decreases and the information is compressed into a number of uncorrelated components (Ricotta *et al*., 1999). The ability to deal with correlated variables is especially pertinent in this study since the various texture measures calculated at many window sizes are highly correlated (St-Louis *et al*., 2006). Studies have shown principal components calculated from texture variables can reduce correlation and improve model performance. For example, Johansen *et al.* (2007) calculated principal components of texture variables and found that the principal components produced adequate classification accuracies and were beneficial for compressing image data for classification purposes. Additionally, in the study by Puissant *et al.* (2005), the calculation of principal components from the optimal texture measure resulted in a higher classification accuracy. Results from this study confirm that using the principal components of the various texture variables (window size and texture measures) yielded better results than using the original texture variables with the random forest ensemble. For example, the original texture variables calculated using a 3 x 3 window yielded a higher NOOB error (30.45%) when compared to the PCA of the texture variables which yielded an NOOB error of 28.78%.

2.5.3. Backward variable selection and random forest

Overall, the best model was obtained using backward variable selection with the random forest ensemble (method 6). The proposed method produces a high predictive accuracy (NOOB error = 27.72%, R^2 = 0.68) when compared to using

all the spectral and texture variables with the random forest ensemble (NOOB error = 27.72%; R^2 = 0.63), or any of the existing methods proposed by Puissant *et al.* (2005). Additionally, the method simplifies the modelling process by identifying the smallest number of input variables that offer the best discriminatory power and aid in the empirical interpretation of the final model.

Results from this study show that only five variables (7.8% of original 64 variables) were used in the final model (NIR, green, variance with a 3 x 3 window, red and blue). The spectral variables, especially the NIR and green bands, performed well in the final model. Green plants characteristically absorb visible electromagnetic radiation and strongly scatter in the NIR bands (Curran, 1980). Leaf reflectance is determined by concentrations of plant pigments such as chlorophyll. Young healthy plants will contain higher concentrations of chlorophyll *a* and *b* while older plants will contain less (Schmidt, 2003).

The variance texture measure calculated using a 3 x 3 moving window was the only texture variable utilized in the final model. The variance texture measure is considered to be relevant for forest age discrimination because it accounts for all pair-wise combinations of variability and can therefore detect subtle changes in image texture that occur as a result of plant growth (Kayitakire *et al*., 2002). Younger trees have a more open canopy which results in a higher variation between pixel values due to the alteration of tree crowns, shadows, and gaps. Older stands on the other hand are structurally complex and characterized by a patchy upper canopy with coarse woody debris in all stages of decomposition (Johansen *et al*., 2007).

2.6. Conclusion

To conclude, the random forest ensemble is a useful and robust tool for combining spectral and texture remote sensing data for forest age prediction. Testing various combinations of spectral and texture variables in a random forest environment showed that principal components of the texture datasets produced better results than simply using the original texture images. Furthermore, using

different window sizes to combine spectral and texture variables was more successful than using texture measures. The best model was achieved using the random forest backward variable selection method which reduces error and simplifies the modeling process by selecting only the most important variables to include in the final model. Only five variables were used (NIR, green, variance with a 3 x 3 window, red and blue) and the final model yielded a predictive accuracy of $R^2 = 0.68$. All of the spectral variables (n = 4) were selected for the final model. Since spectral response was significant for *P. patula* age prediction, we wanted to determine whether high spectral remote sensing data could improve the prediction. The next section will assess the capability of hyperspectral remotely sensed data to predict *P. patula* age.

Chapter three

Examining the utility of random forest and AISA Eagle hyperspectral image data to predict *Pinus patula* age in KwaZulu-Natal, South Africa

3.1. Abstract

High data dimensionality is a common problem in hyperspectral data processing. Obtaining reliable predictive accuracies is challenging due to the limited amount of training samples compared to the high number of hyperspectral bands. Consequently, remote sensing techniques that reduce the number of bands are considered essential tools for most hyperspectral applications. The aim of this study was to examine the utility of the random forest ensemble to select the optimal subset of hyperspectral bands that will accurately predict the age of *Pinus patula* stands. AISA Eagle hyperspectral image data (272 bands; 400-970 nm spectral range; 2.4 m spatial resolution; 2.24 nm spectral resolution) was collected over the study area and atmospherically corrected using *in situ* spectroradiometer readings. The random forest ensemble was then used to test whether the forward or backward variable selection methods could identify the optimal subset of AISA Eagle bands to accurately predict the age of *P. patula* stands. Results indicate that both the forward and backward variable selection methods produced high predictive accuracies (RMSE= 3.097). However, the backward variable selection method utilized 206 bands for the final model, while the forward variable selection utilized only a small subset of non-redundant bands (n = 9) while preserving the highest model accuracy (R^2 = 0.6).

Keywords: Hyperspectral remote sensing, random forest, variable selection, forest age, *Pinus patula*.

3.2. Introduction

Hyperspectral remote sensing has demonstrated wide applicability for mapping forest structural variables (Treitz *et al.*, 1999). Unlike multispectral remote sensors which capture a few broad bands, hyperspectral sensors collect data from hundreds of narrow spectral bands, providing a full spectral curve for each pixel (Vane and Goetz, 1993). The contiguous nature of hyperspectral data makes it possible to identify features of interest by their characteristic reflectance signal (Goetz, 2009). Detailed information about how individual objects in a scene reflect or emit electromagnetic energy increases the probability of finding unique characteristics which distinguish the object from others (Govender *et al.*, 2008).

For example, hyperspectral remote sensing has been used to map forest age (van Aardt and Norris Rogers, 2008; Buddenbaum *et al.*, 2005), species composition (Martin *et al.*, 1998; van Aardt and Wynne, 2007; Lawrence *et al.*, 2006; Cho *et al.*, 2008; Cochrane, 2000), and the status and health of forests (Ismail, 2009). Buddenbaum *et al.* (2008) used HyMap imagery (128 bands; 400-2500 nm spectral range; 5m spatial resolution; 10-20 nm spectral resolution) and texture features to classify coniferous tree species and age classes. The overall classification accuracy was 74%. Van Aardt and Norris-Rodgers (2008) used CASI hyperspectral imagery (36 bands; 426–952 nm spectral range; 1m spatial resolution) to discriminate the age of *Eucalyptus* and *Acacia* species in KwaZulu-Natal. The overall classification accuracy was 85% and 71% respectively (van Aardt and Norris-Rogers, 2008).

Although hyperspectral image data provides a wealth of information, the data presents some difficulties, such as increased image costs, data volumes, data redundancy, and data processing costs (Govender *et al.*, 2007; Bajcsy and Groves, 2004; Plaza *et al.*, 2009; Lefsky *et al.*, 2001). A common problem in hyperspectral data processing is related to high data dimensionality (Borges *et al.*, 2007). In most applications utilizing hyperspectral image data, the goal is to classify or discriminate objects in a scene. One would expect that as the number

of hyperspectral bands increases, the accuracy of classification should also increase; however this is not the case in model-based analysis (Bajcsy and Groves, 2004). Generally, the number of training samples (*n*) is limited with respect to the number of hyperspectral bands (*p*). This problem of 'small *n*, large *p*' has been termed the 'curse of dimensionality' (Melgani and Bruzzone, 2004). Due to the limited training set compared to the high number of bands available in hyperspectral remote sensing applications, it is very challenging to get a reliable estimation of statistical class parameters (Melgani and Bruzzone, 2004; Foody and Mathur, 2004; Foody and Arora, 1996). As a result, model accuracy tends to decrease as the number of variables increases (Plaza *et al.*, 2009). This effect is known as the Hughes phenomenon (Hughes, 1968).

Remote sensing techniques that reduce the 'curse of dimensionality' without sacrificing significant information are therefore essential. Consequently, variable selection is considered to be a practical and vital method in hyperspectral data processing (Borges *et al.*, 2007; Ismail, 2009). Variable selection allows the prediction algorithm to focus its attention on relevant variables while ignoring the contribution of irrelevant variables which could be misleading (Dunne *et al.*, 2002). By reducing the number of input variables, the model is faster and less prone to overfitting. It also allows for a better understanding of the underlying processes that generated the data (Sayes *et al.*, 2007).

Variable selection is therefore considered a necessary step to move from research to operational remote sensing applications (van Aardt and Norris-Rogers, 2008). Several researchers have investigated methods that reduce the data dimensionality of hyperspectral data (Bajcsy and Groves, 2004; Vaiphasa et al., 2005; Vaiphasa et al., 2007). These methods can be generalized as variable reduction and variable selection methods (Janecek *et al.*, 2008). Variable reduction refers to methods that create new variables as combinations of the original variables. An example of such a method is the principal component analysis which is commonly used in remote sensing applications (Saeys *et al.*, 2007). Whereas variable reduction methods use the original input variable to

create new variables or components, variable selection methods select a subset of the original input variables (Ismail, 2009). These methods can be classified into filter and wrapper approaches (Guyon and Elisseeff, 2003). Filters are pre-selection methods which are independent of the prediction algorithm (Kohavi and John, 1997). Variables are ranked based on certain statistical criteria and the variables with the highest rankings are selected for further analysis. Examples of filter techniques include: t-test, chi-squared test and Pearson's correlation coefficients (Gheyas and Smith, 2010). Wrappers on the other hand include the prediction algorithm in the variable selection process (Janacek *et al.*, 2008). Although more computationally demanding, wrapper methods generally perform better than the computationally efficient filter methods (Gheyas and Smith, 2010; Ismail, 2009; Chan and Paelinckx, 2008).

The most straightforward search strategies for wrappers are based on the sequential elimination or addition of variables. Backward variable selection starts with a full set of variables (ranked by importance) and progressively eliminates the least promising variables. The goal of backward variable selection is to first consider the contributions of all the variables, then remove the irrelevant variables so that only a few remain, providing a smaller and more predictive subset of variables. Forward variable selection commences with no input variables and then progressively adds one variable at a time. If there is no improvement from adding any further variables or all the variables have been added, the search is terminated (Dunne *et al.*, 2002).

Support vector machines, regression trees, and neural networks are popular methods in remote sensing applications (Pal and Mather, 2004; van Aardt and Norris-Rogers, 2008; Mutanga and Skidmore, 2004). Although these machine learning methods have proven successful, they lack insight regarding the bands that best contribute to the final model (Adam *et al.,* in press). Alternatively, random forest (Breiman, 2001) has been successfully used for both band selection and for prediction purposes (Ismail and Mutanga, 2010). Random forest grows a large number of regression trees by repeatedly taking random subsets of the training data and using random subsets of the input variables to

determine each split of the tree (Lawrence *et al.*, 2006). One of the most important features of the random forest algorithm is the internal measure of variable importance (Ismail, 2009; Ismail and Mutanga, 2010). Ismail and Mutanga (2010) used the random forest ensemble and backward variable selection to select the optimal set of hyperspectral parameters to predict *Sirex noctilio* induced water stress in *Pinus patula* trees. By identifying the minimum number of spectral parameters, the predictive accuracy of the model increased from $R^2 = 0.73$ (using all spectral parameters) to $R^2 = 0.76$ (using the parameters selected by the wrapper). Similarly, Adam *et al.* (in press) used field spectrometry to discriminate between papyrus swamps and its coexistence species. The random forest ensemble and a forward variable selection were used to select key bands and to reduce the number of input variables. The study by Adam *et al.* (in press) selected 10 significant bands located in the visible and short-wave infrared (SWIR) portions of the spectrum, and resulted in a very high accuracy (90.5%).

Adam *et al.* (in press) showed that by utilizing the random forest ensemble with a forward variable selection method produced the best results. In contrast, Ismail and Mutanga (2010) showed that the backward variable selection method when used with the random forest ensemble produced the optimal subset of variables with the highest predictive accuracies. The question remains: Which variable selection method (forward or backward) when used in conjunction with the random forest ensemble produces the best results?

In this study the random forest ensemble will be tested as a wrapper using both forward and backward variable selection methods. Additionally, we intend to examine the utility of the random forest ensemble to accurately predict the age of P*. patula* stands using AISA Eagle image data.

3.3. Materials and methods

3.3.1. Study area

The study area (figure 7) is located in the Hodgsons Sappi plantation in KwaZulu-Natal, South Africa (Centroid: Latitude 29º13'40"S Longitude 30 º29'56"E). The site is hilly with elevation ranging from 1030-1590 m. Soil is predominantly apedal and plinthic from the ecca group, and the area experiences summer rainfall with an annual mean precipitation of 915 mm (range 730-1280 mm). The area falls within the midlands mistbelt grassland bioregion (Mucina and Rutherford, 2006) and frequent mist provides significant amounts of additional moisture. The mean annual temperature is 15.8ºc. The study area contains same-aged stands of *Acacia*, *Eucalyptus* and *Pinus* species. The majority of the stands consist of *P. patula* trees which are grown for pulpwood, with tree age ranging from 1 – 24 years.

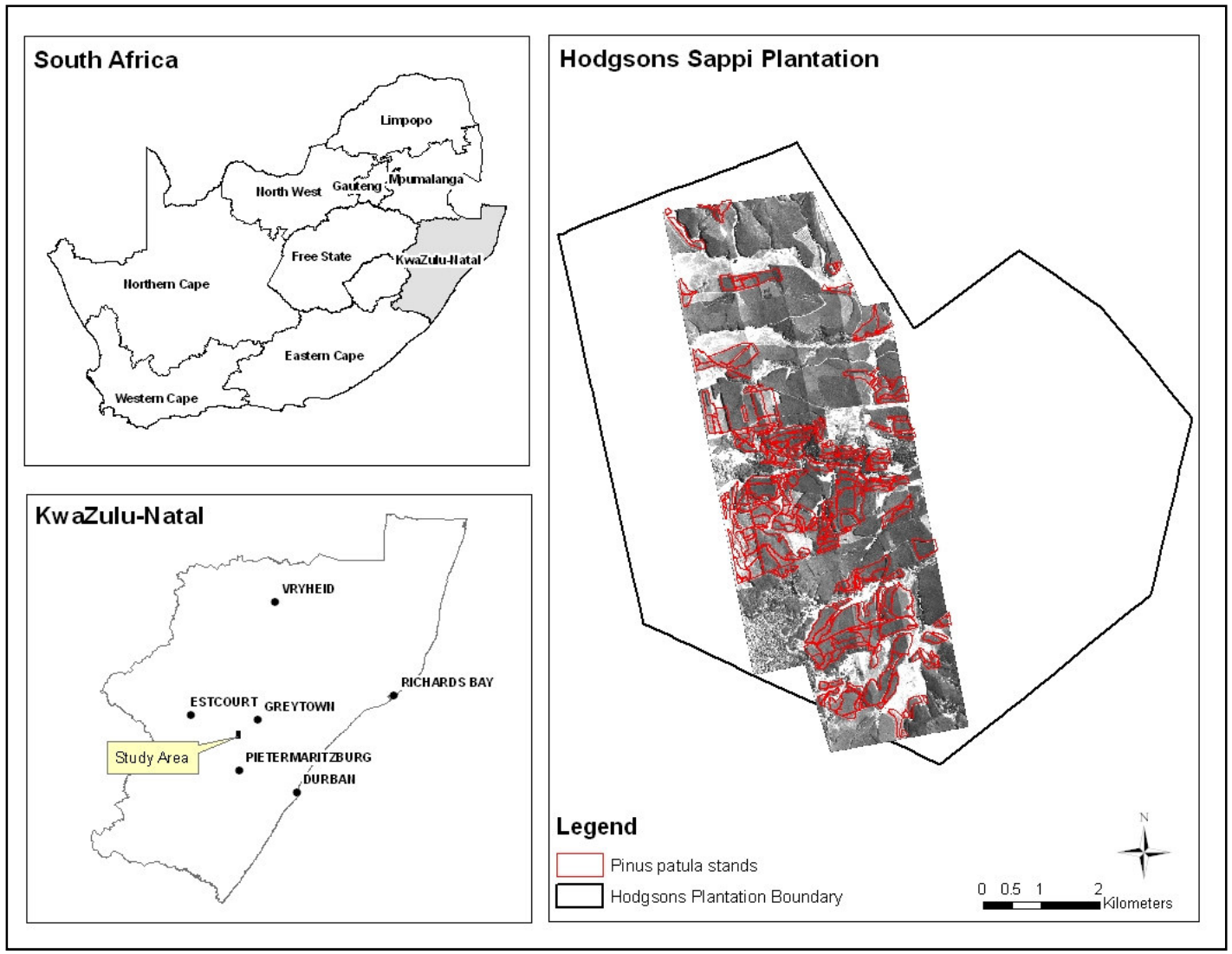


Figure 7. Location of the study area. The *Pinus patula* stands (n= 90) in the study area are highlighted in red.

3.3.2. Data acquisition

AISA Eagle hyperspectral data (272 bands; 400-970 nm spectral range; 2.4 m spatial resolution; 2.24 nm spectral resolution) was acquired during March 2009 under cloudless conditions. The image was atmospherically corrected using the empirical line method which is based on the linear relationship between *in situ* measured ground reflectance and the sensor signal (Roberts *et al.*, 1986). The Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer (350 nm – 2500 nm) was used to collect ground measurements and these measurements were used to calibrate the airborne data as well as to assess the data quality of the AISA Eagle imagery. The image was orthorectified using bilinear resampling and geo-referenced in Universal Transverse Mercator (UTM) with WGS-84 Geodetic datum. Due to the amount of noise in the dataset, bands after 900 nm were excluded from the study (figure 8), and the final dataset consisted of 230 bands.

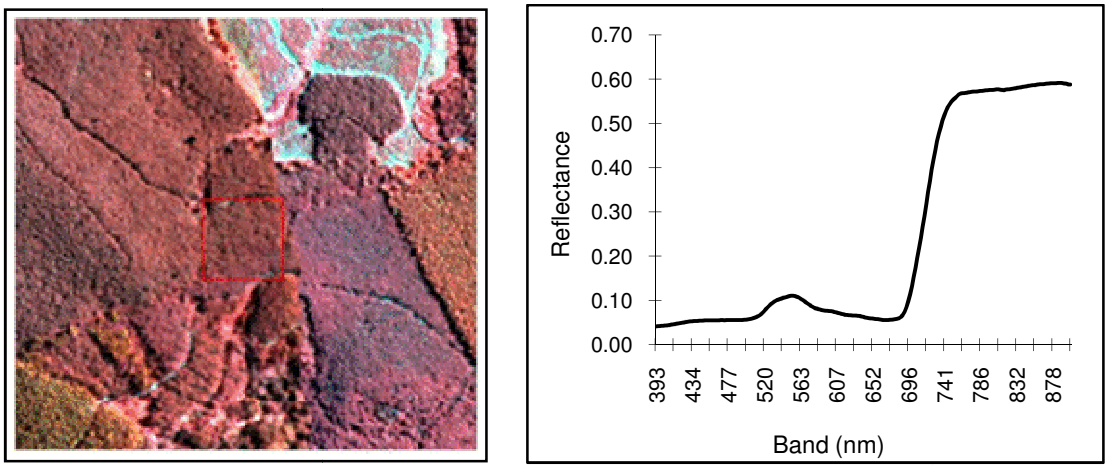


Figure 8. Example of *P. patula* sample plot and the corresponding spectral curve.

3.3.3. Field data

Field data for this study were collected by Sappi, a paper and pulp company, using industry standard enumeration techniques (Owen, 2000). The data consisted of species composition, age, mean stems per hectare, mean diameter at breast height and mean tree height. Following Jensen *et al.* (1999) the field

data were used as ground truthing data to predict forest age. Stands that were not consistent with the ground truthing data such as newly cut stands were excluded from sampling. Stands three years and younger were also excluded due to their spectral similarity to bare soil and grass (Jensen *et al*., 1999). The final sample dataset consisted of 90 *P. patula* stands (651.6 ha) with age ranging from 4 – 24 years. However, most of the samples consisted of younger stands (4-12 years). A 30 x 30 pixel window was used to collect image spectra from each sample stand. The spectra were collected and averaged using ENVI 4.5 (ENVI, 2006). Seventy percent of the samples (n = 63) were used as a training dataset while the remaining 30% of the samples (n = 27) were used to test the final model (Ismail, 2009).

3.3.4. Random forest

Random forest is a relatively new data mining technique that is designed to produce accurate predictions that does not overfit the data (Breiman, 2001). Bootstrap samples are drawn to create many regression trees (generally between 500 and 2000 trees) and the final prediction is the average of all individual tree outputs (Prasad *et al*., 2006). When a bootstrap sample is drawn during the training procedure, approximately one third of the data is excluded; these are called the 'out-of-bag' (OOB) samples. The tree makes predictions on the OOB data to bring the sample to full size; the replicated data are called the 'in-bag' data (Prasad *et al*., 2006). Since the OOB samples were not used in the training process, the prediction error offers an unbiased assessment of the trees predictive accuracy (Breiman, 2001). Random forest is easy to use as it requires only two input parameters: the number of trees (*ntree*) and the number of input variables (*mtry*). Random forest introduces additional randomness by selecting a random subset of the input variables (*mtry*) to determine the best split at each tree node (Lawrence *et al*., 2006). The *mtry* function in random forest is an important component because by growing the regression trees using only a small user defined subset of variables, the 'small *n* large *p*' problem is avoided (Ismail,

2009; Gislason *et al.*, 2006). Additionally, studies have shown that the default *ntree* and *mtry* values provide accurate results (Ismail, 2009; Liaw and Wiener, 2002). For example, Ismail (2009) optimized the *mtry* value by trying all possible values (n = 64), however the default *mtry* produced the lowest error and the study found that the classification accuracy did not increase beyond 500 trees. We implemented the random forest library (Liaw and Wiener, 2002) using the R statistical software (R Development Core Team, 2008).

3.3.5. Random forest variable importance

Random forest offers three measures of variable importance (Breiman, 2001). The first is based on the number of times each candidate variable is selected. The second measure is based on the Gini index (Strobl and Zeileis, 2008). The Gini index was used in the original classification and regression trees method proposed by Breiman (1984). The final measure, which was adopted in this study, utilizes the permutation of variables as an estimate of variable importance (Strobl *et al.*, 2008). This importance measure is calculated from the difference between the OOB error of each regression tree, and the error calculated after permuting a predictor (Breiman, 2001). The change in OOB error for each randomly permutated predictor gives an indication of its importance because the permutation process should have little effect on the OOB error if a predictor is irrelevant (Grimm *et al.*, 2008). The main advantage of the permutation measure is that the method takes into account the impact of individual predictor variables, as well as the multivariate interactions with other variables (Strobl and Zeileis, 2008). The permutation of variables is therefore seen as a better estimate of variable importance than the other two methods (Breiman 2001, Grimm *et al.*, 2008; Ismail, 2009; Adam *et al.*, in press; Chan and Paelinckx, 2008). However, a limiting factor of random forest variable importance is that it does not automatically select the optimal number of variables that have the lowest error (Adam et al, in press). Subsequently, this study implements the forward and backward variable selection methods which allow for the testing of all

combinations of bands in order to select the model with the lowest error and optimal subset of bands.

3.3.6. Variable selection

In this study, the backward and forward variable selection methods were implemented in conjunction with the random forest ensemble (figure 9). Backward variable selection commences with all of the input variables which are ranked by importance and progressively drops the least promising variable. Forward variable selection on the other hand commences with no variables and progressively adds the highest ranking variable. For both selection methods, variable importance was calculated once using all the hyperspectral bands and the subsequent rankings were then used to add or remove variables (Ismail and Mutanga, 2010). In contrast, the recursive approach recalculates variable importance for each model producing a new ranking of variables before any variables are removed or added to the model (Ismail, 2009). However, Svetnik *et al.* (2003) showed the recursive approach performs poorly, since it is much greedier (more prone to getting stuck in a local optimum) than the non-recursive approach. After running the forward and backward variable selection methods, the root mean square error (RMSE) was calculated (equation 3) and used to select the optimal number of bands.

It should be noted that the selection of variables (when $n < p$) is an unstable process and can lead to the selection of very different subsets of variables for each replicate of the study (Gheyas and Smith, 2010; Granitto *et al.*, 2006). Bands that do not appear relevant may become relevant when taken in conjunction with bands. (Gheyas and Smith, 2010). In order to test the stability of the final model, the variable selection process that produced the best results was repeated (replications =100) to determine the frequency that the optimal bands appear in subsequent replicates of the study (Ismail, 2009). For comparison purposes we also examined the utility of all the hyperspectral bands as input variables into the random forest algorithm.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{OOB}\right)^2}{n}} \qquad \text{(Equation 3)}$$

Where Y_i is the observed forest age and Y_i^{OOB} is the predicted forest age using the out-of-bag samples, and n is the number of samples.

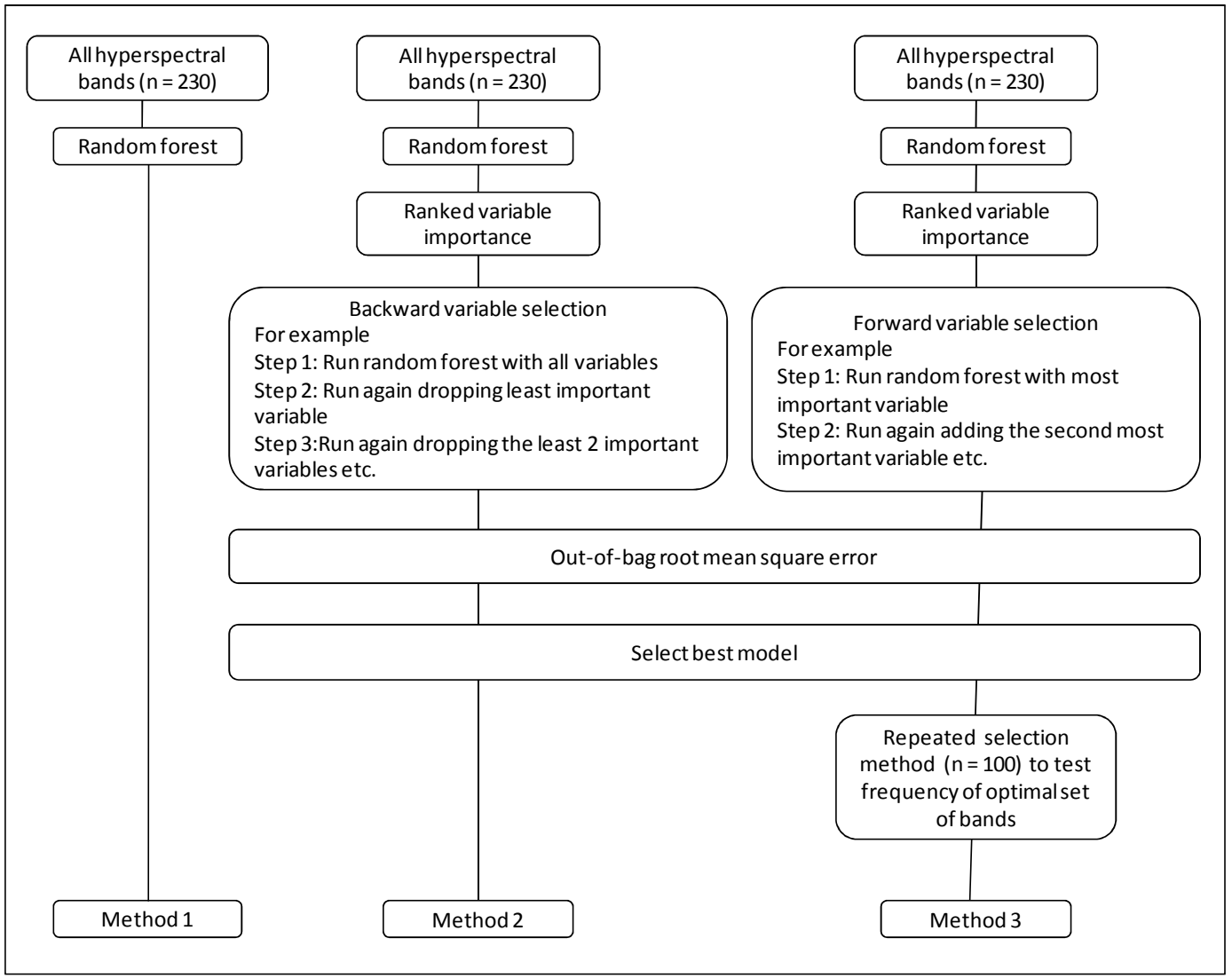


Figure 9. The variable selection methods used in this study.

3.4. Results

3.4.1. Random forest using entire hyperspectral dataset

Initially, the random forest algorithm (*ntree* = 500, *mtry* = 76) was run using all the AISA Eagle bands (*n* = 230) and the resulting RMSE was 3.297. Ranked variable importance showed that bands relating to the red-edge were important in the final model (figure 10). Out of the ten bands selected in the model, eight were from the red-edge portion of the spectrum (749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 687.44 nm, 743.21 nm, 740.98 nm, 720.89 nm), and the remaining two were just outside the red-edge in the red band (660.76nm, 658.54nm). The model's predictive accuracy was tested on the independent test dataset and resulted in a poor prediction accuracy (R^2 = 0.53).

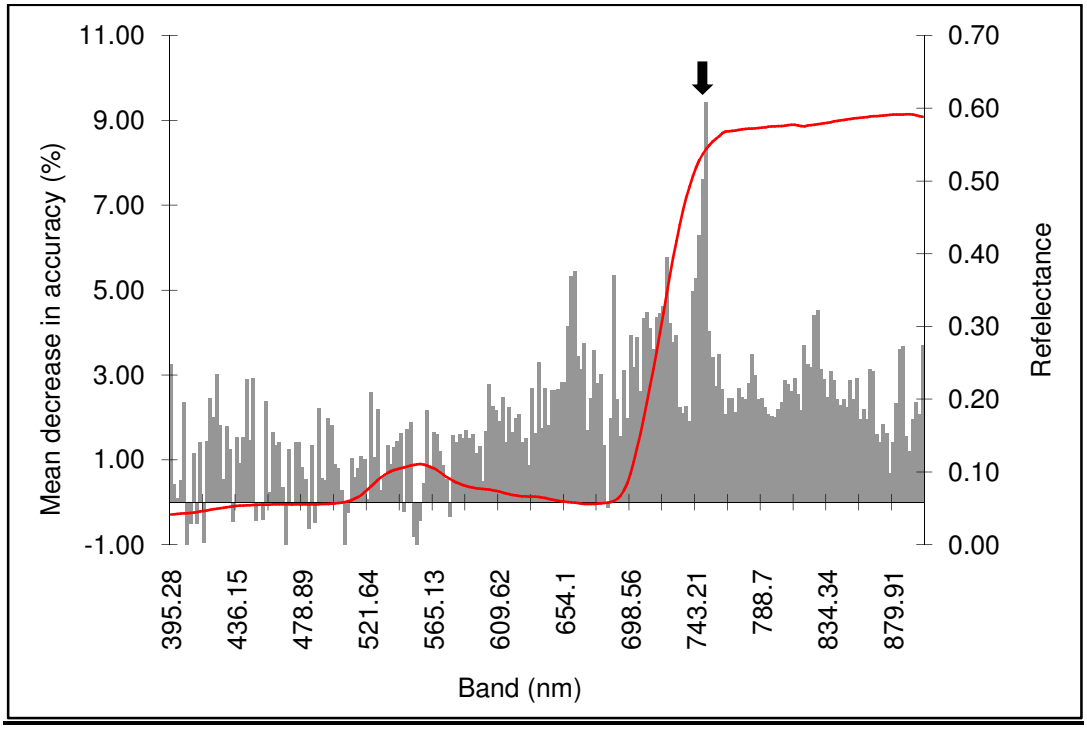


Figure 10. Variable importance of all the AISA Eagle bands expressed as the mean decrease in accuracy. The red line represents a typical vegetation reflectance signature of a *Pinus patula* tree. The black arrow shows the location of band with the largest mean decrease in accuracy.

3.4.2. Random forest variable selection

A backward variable selection was implemented (method 2) and figure 11 shows that that the best model with the lowest RMSE (3.097) consists of 206 bands.

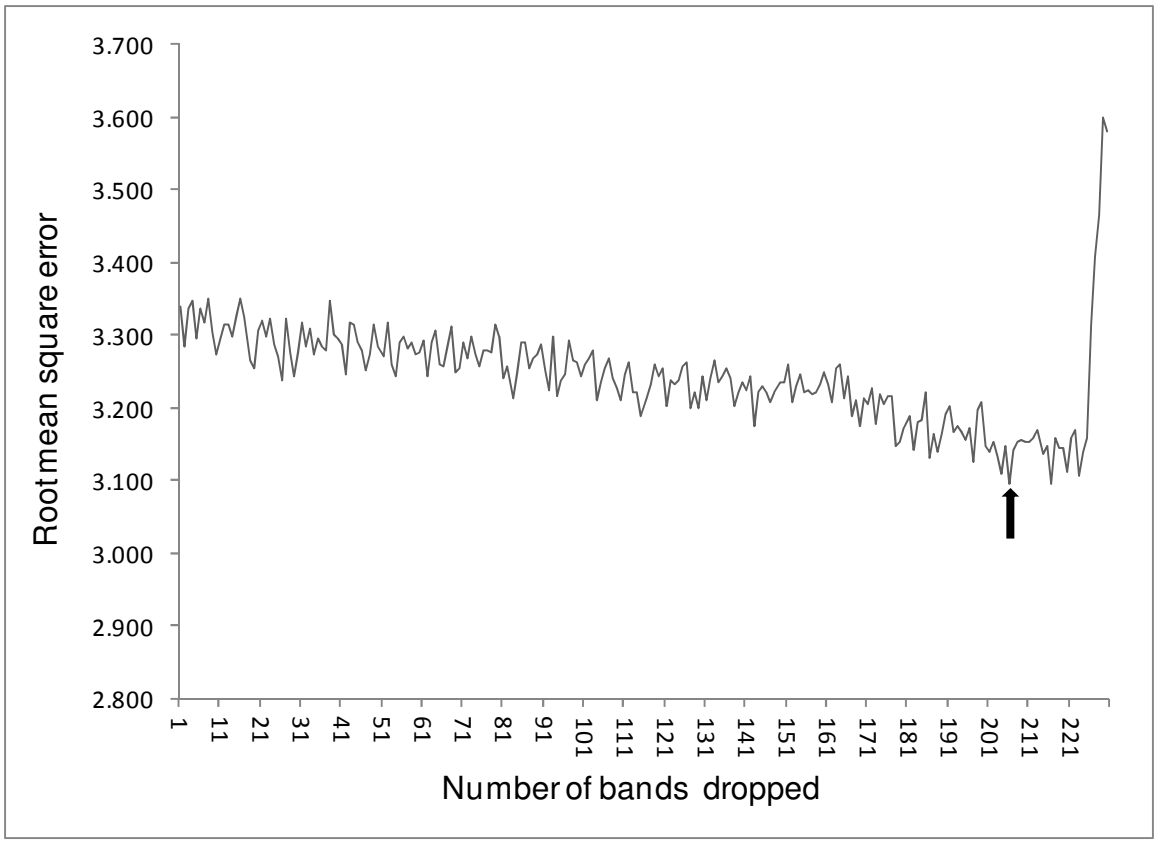


Figure 11. Results of the backward variable selection method. The black arrow indicates the iteration that produced the optimal subset of bands with the lowest RMSE.

Subsequently, a forward variable selection (method 3) was carried out and the final model selected only nine variables (749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 660.76 nm, 687.44 nm, 658.54 nm, 743.21 nm, 740.98 nm) and produced a RMSE of 3.097 (figure 12). Although both variable selection methods selected the best model using a RMSE of 3.097, the backward variable selection method failed to sufficiently reduce the number of bands. The forward variable selection method selected only nine bands compared to the 206 bands selected

by the backward variable selection method. Since the forward variable selection method produced the best results, subsequent analysis will concentrate on the bands selected by the forward variable selection method.

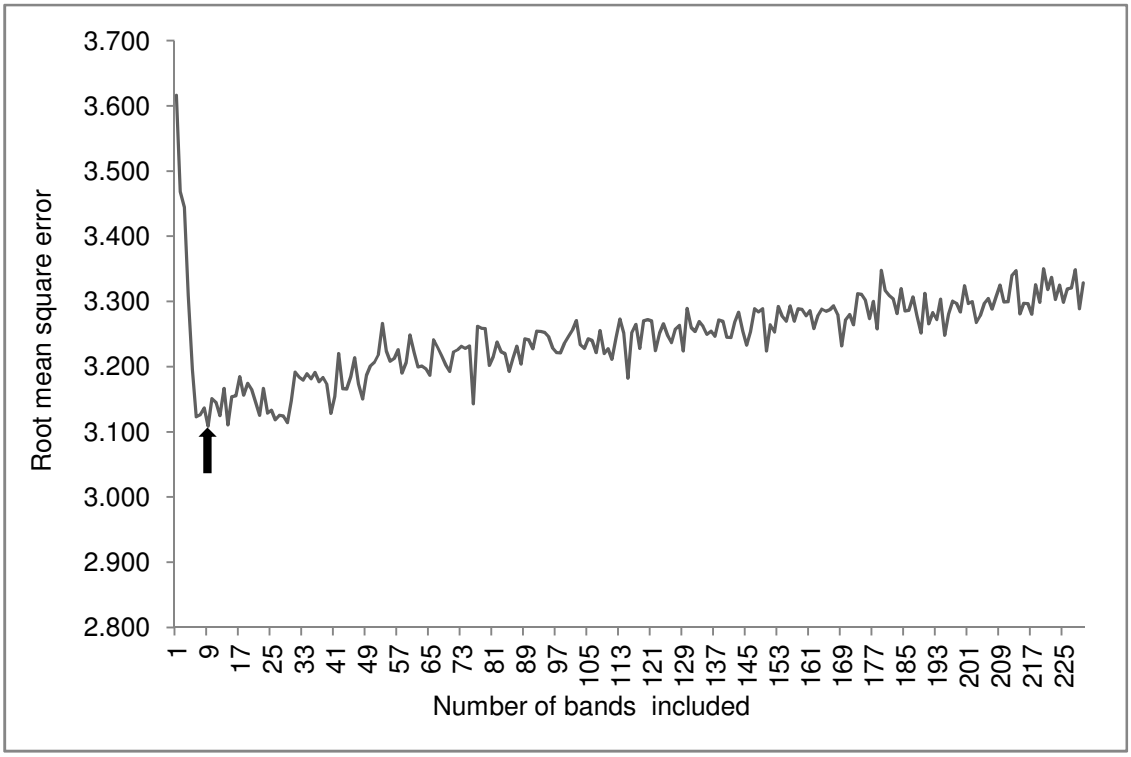


Figure 12. Results of the forward variable selection method. The black arrow indicates the iteration that produced the optimal subset of bands with the lowest RMSE.

Figure 13 shows the location of the optimal bands as determined by the forward variable selection method. The optimal subset of bands (n = 9) are primarily located in the red-edge portion of the spectral curve. Out of the nine bands selected by the forward variable selection method, seven are located within the red-edge portion. The model's predictive accuracy was tested on the independent test dataset and produced a R^2 value of 0.6.

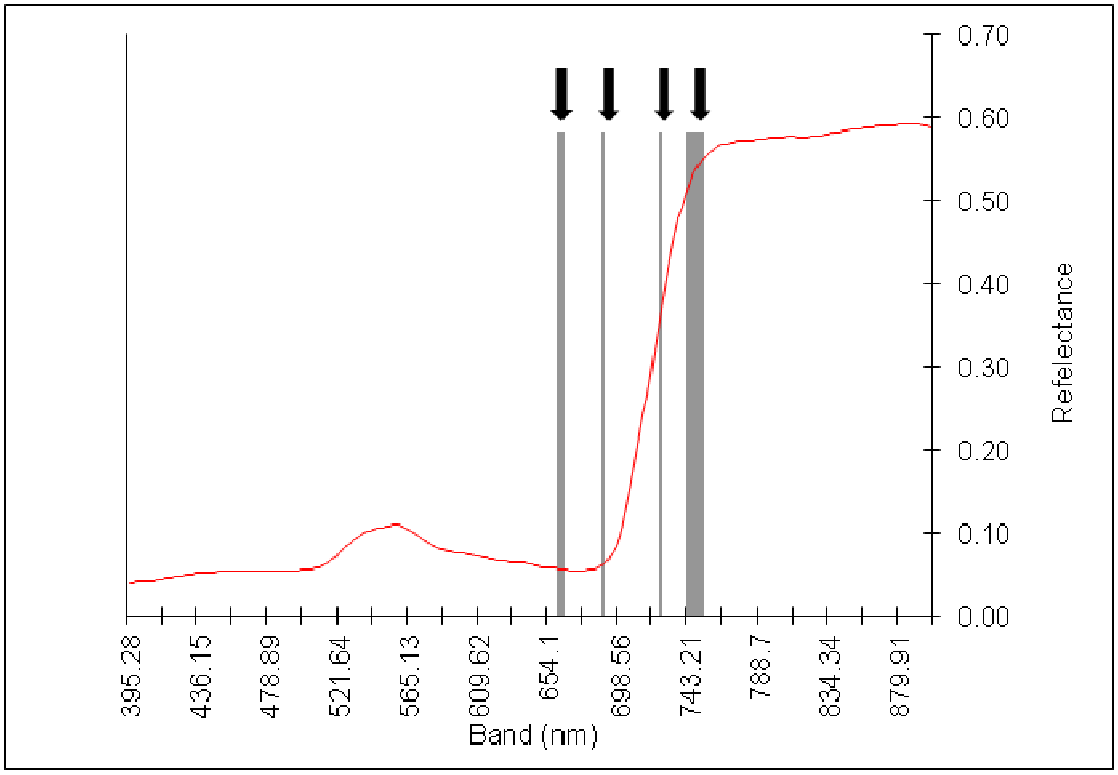


Figure 13. Location of the optimal bands selected by the forward variable selection method.

In order to test the stability of the bands selected by forward variable selection method, the selection process was repeated 100 times to determine the frequency that the selected bands (749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 660.76 nm, 687.44 nm, 658.54 nm, 743.21 nm, 740.98 nm) appear in subsequent replicates of the study. Results indicate that all the bands selected using the forward variable selection method have a very high selection probability (i.e. greater than 80%). More specifically, figure 14 shows the frequency of the selected bands: 749.91 nm (frequency = 100%), 747.68 nm (frequency = 100%), 745.44 nm (frequency = 100%), 723.12 nm (frequency = 95%), 660.76 nm (frequency = 82%), 687.44 nm (frequency = 91%), 658.54 nm (frequency = 93%), 743.21 nm (frequency = 83%) and 740.98 nm (frequency = 88%). Bands 749.91 nm, 747.68 nm and 745.44 nm are selected in all replicates of the study.

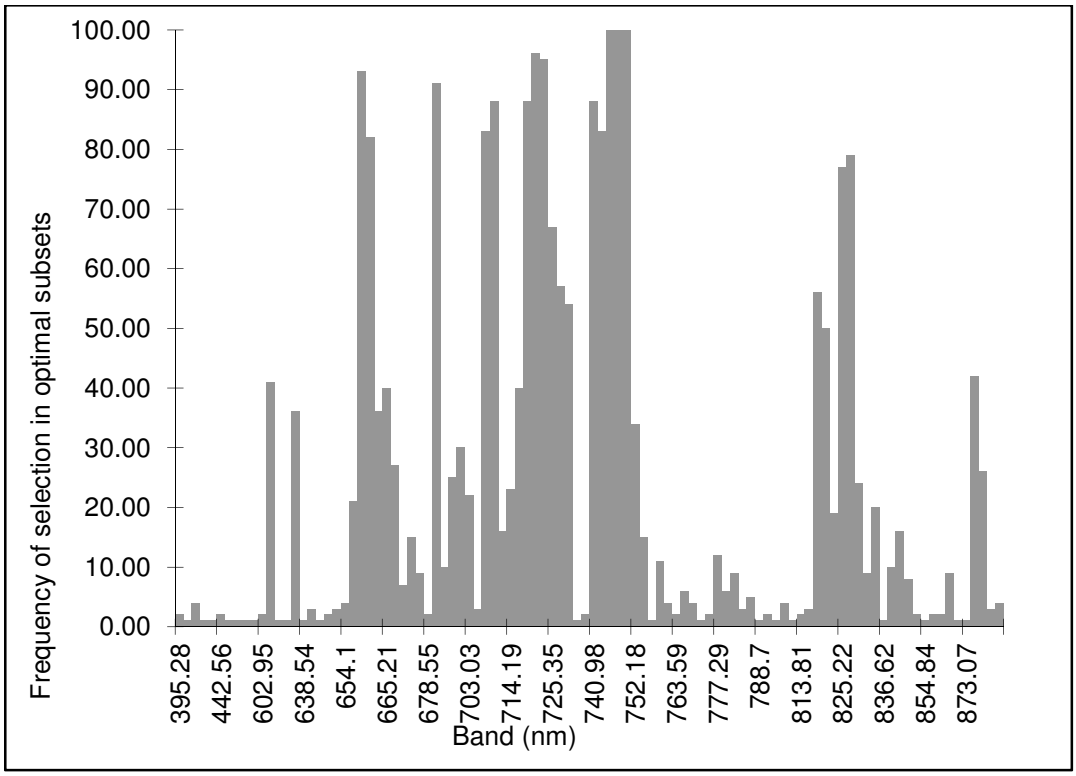


Figure 14. The frequency and the location of bands that are selected by the forward variable selection method during each replicate (n = 100) of the study.

3.5. Discussion

Hyperspectral remote sensing has proved to be successful in the mapping of forest structural variables (Treitz *et al.*, 1999). One of the main problems with hyperspectral data is the large number of bands compared to the number of training samples (Melgani and Bruzzone, 2004). Techniques that reduce the number of input bands without decreasing model accuracy are therefore vital in hyperspectral data processing. The aim of this study was to utilize the random forest ensemble as a wrapper to reduce the large number of hyperspectral bands to an optimal subset for the accurate prediction of *P. patula* age. Additionally, the forward and backward variable selection methods were compared to determine which wrapper produces the best results when implemented with the random forest ensemble.

3.5.1. Random forest as a wrapper for variable selection

Results showed that the random forest ensemble successfully reduced the large hyperspectral dataset (n = 230) to a more manageable subset of bands (n = 9), while preserving the highest model accuracy (R^2 = 0.6). These results are comparable to other studies that used random forest for band selection (Lawrence *et al*., 2006; Ismail, 2009; Ismail and Mutanga, 2010; Adam *et al*., in press). More specifically, the findings show that by concentrating only on significant variables, the predictive accuracy increased by 7% (from R^2 = 0.53 using all the bands to R^2 = 0.6 when using only nine bands). These results indicate that it is beneficial to remove irrelevant bands and focus only on significant bands. An exhaustive search of all possible subsets of bands will guarantee that the optimal subset of bands is found (Gheyas and Smith, 2010). While the variable selection process increases computational time, studies have shown that the random forest ensemble is faster than any of the competing machine learning algorithms (Ismail, 2009).

3.5.2. Forward variable selection versus backward variable selection

Although the forward and backward variable selection methods selected optimal models based on the same RMSE (3.097), the forward variable selection was able to significantly reduce the number of bands. The forward variable selection methods selected nine bands compared with the backward selection method that selected 206 bands. One of the problems with the backward variable selection method is that the method assumes that the prediction accuracy does not decrease as the number of bands increases. However, in reality the predictive ability of the wrapper may decrease as the variable subspace increases (Gheyas and Smith, 2010). As a result, when dealing with a high dimensional datasets (for example hyperspectral image data), the backward variable selection method often finds it difficult to identify the effect of individual variables (AISA Eagle bands) on the target variable (*P. patula* age). Therefore good predictors could be

removed early on by the wrapper. In the forward variable selection method, once a band is added it is never removed. The forward variable selection is therefore considered to be a more robust method (Gheyas and Smith, 2010).

3.5.3. Significance of the red edge

Only nine bands (749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 660.76 nm, 687.44 nm, 658.54 nm, 743.21 nm, 740.98 nm) were selected by the forward variable selection method. The majority of the bands were related to the red-edge portion of the spectrum which is located between 670 nm and 780 nm (Cho, 2007). The position of the red edge moves as a tree matures; this is due to changes in chlorophyll content (Shafri *et al.*, 2006). Younger trees contain a larger quantity of chlorophyll which absorbs more visible light, moving the red edge towards the longer wavelengths. As the tree matures there is less chlorophyll present and therefore less visible light is being absorbed. This shifts the red edge position towards the shorter wavelengths (Shafri *et al.*, 2006). The strong relationship between red edge reflectance and forest age makes the red edge significant for age discrimination and prediction.

3.6. Conclusion

Overall, the results of this study showed that the random forest ensemble is a robust and accurate method for (i) band selection and (ii) predicting the age of *P. patula* stands using AISA Eagle imagery. While the best models for the forward and backward variable selection methods resulted in a RMSE of 3.097, the backward variable selection method was unable to significantly reduce the number of AISA Eagle bands. The forward variable selection produced the best model (R^2 = 0.6) which included only nine of the original 230 hyperspectral bands. Most of these bands were located in the red-edge portion of the spectrum, suggesting that this region is highly significant for predicting the age of *P. patula* stands.

Chapter four

Conclusion

4.1. Introduction

The mapping of forest age is important for effective forest inventory since age is indicative of a number of forest structural variables. Traditional field survey techniques have commonly been used for the collection of forest attributes. Although this method produces accurate results, it is costly and time-consuming. Remote sensing offers a non-destructive and spatially complete alternative that is time-effective and cost-effective for large scale forest mapping (Cho *et al*., 2009). The aim of this research was to assess the utility of the random forest ensemble and multispectral and hyperspectral remotely sensed data to predict *P. patula* forest age. The main objectives were (i) to assess the capability of multispectral remotely sensed data to predict *P. patula* age, (ii) to test the ability of the random forest ensemble to combine spectral and texture variables derived from multispectral imagery, (iii) to assess the capability of hyperspectral remotely sensed data to predict *P. patula* age, (iv) to evaluate the effectiveness of th*e* random forest ensemble to select the optimal subset of hyperspectral bands, and (v) to compare multispectral and hyper*spe*ctral remotely sensed data to predict *P. patula* age. The section below will discuss each objective.

4.2. Assessing the capability of multispectral remotely sensed data to predict *P. patula* age

The results from the study confirm the potential of multispectral remotely sensed (QuickBird) data to accurately predict forest age. Similar results were obtained by Jensen *et al.* (1999), Jakubauskas and Price (2000), Gerylo *et al.* (2002), Kayitakire *et al.* (2006), and Wunderle *et al*. (2007). The multispectral dataset

produced high accuracies (R^2 = 0.68) by using only five variables. The NIR band was the most significant in the model, followed by the green band, variance texture measure with a 3 x 3 window, the red band, and the blue band. The significance of the NIR bands can be explained by the presence of chlorophyll in green vegetation, which strongly scatters reflectance in the NIR band (Curran, 1980). The amount of chlorophyll varies greatly with age and this will affect leaf reflectance properties; young healthy plants contain higher concentrations of chlorophyll and reflect highly in the NIR, while older plants will contain less chlorophyll and reflect less in the NIR band (Schmidt, 2003).

The variance texture measure calculated using a 3 x 3 moving window was the only texture variable that was selected by the backward variable selection. The variance texture measure performed better than the other texture measures because it accounts for all pair-wise combinations of variability and can therefore detect subtle changes in image texture that occur as a result of plant growth. Johansen *et al*. (2007) demonstrated that there is a structural difference between trees of different ages. Whereas younger trees have a more open canopy and therefore show a higher variance between pixel values, older trees are characterized by a complex and patchy canopy. Texture is a multi-scale phenomenon and window sizes are therefore an important component of a texture analysis (Moskal and Franklin, 2001). The smallest window size (3 x 3) was significant because the trees in the study area are grown for pulpwood and are therefore planted close together. The methods used in this study would be useful for larger scale operation forestry, as remote sensing imagery offers a synoptic and spatially complete coverage over large areas of forest.

4.3. Testing the ability of the random forest ensemble to combine spectral and texture variables derived from multispectral imagery.

The random forest ensemble has shown great potential for combining spectral and texture variables to accurately predict the age of *P. patula* stands. Traditional linear methods require normal distribution, linearity, and the absence of

collinearity amongst input variables (Jensen *et al.*, 1999). Remote sensing data are often complex and involve nonlinear relationships between observed data and remotely sensed data. In Jenson *et al.* (1999), a non-linear relationship was found between forest age and NIR reflectance. Reflectance in the NIR decreased at a larger rate as the trees matured. The random forest ensemble was successful in the integration of texture and spectral remote sensing data, confirming that the ensemble can deal with complex non-linear data and correlated predictor variables.

Results from the study also showed that (i) utilizing principal components of the texture datasets produced better results than simply using the original texture images, (ii) combining spectral and texture variables according to window size (for example 3 x 3 contrast, 5 x 5 contrast and so on) was more effective than combing the variables according the optimal texture measure (for example 3 x 3 contrast, 3 x 3 variance and so on), and (iii) utilizing a backward variable selection incorporated with the random forest ensemble produced the best results. The backward variable selection method simplified the modelling process by selecting only five significant bands from the original 230 bands. Using all the texture and spectral variables produced a predictive accuracy of R^2 = 0.63, whereas using the variables selected by the backward variable selection resulted in a higher accuracy of R^2 = 0.68 (5% increase). These results illustrate the benefit of utilizing the proposed method to identify the best texture and spectral variables. The method developed in this study can also be potentially applied to other remote sensing applications that combine spectral and texture variables.

4.4. Assessing the capability of hyperspectral remotely sensed data to predict *P. patula* age.

Results from this study showed that utilizing all the AISA Eagle bands with the random forest ensemble produced a R^2 value of 0.53. However, the random forest ensemble when utilized with the forward variable selection method selected an optimal subset of bands (n = 9) and improved the final predictive

accuracy (R^2 = 0.6). The following bands were selected by the forward variable selection method: 749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 660.76 nm, 687.44 nm, 658.54 nm, 743.21 nm, 740.98 nm. Seven out of the nine bands used in the final model were located in the red edge portion of the spectrum. These results correspond with other studies that have found the red edge to be a region of spectral significance for mapping forest structural attributes (van Aardt and Norris Rogers, 2008; Blackburn, 2007; Treitz and Howarth, 1999; Schlerf *et al.*, 2005; Clark *et al.*, 2005). Research has shown that the position of the red edge changes with age. In younger trees the red edge shifts towards the longer wavelengths while for mature trees it shifts towards the shorter wavelengths (Shafri *et al.*, 2006). Hyperspectral remote sensing and red edge position analysis therefore holds great potential for predicting forest age.

4.5. Evaluating the effectiveness of the random forest ensemble to select the optimal subset of hyperspectral bands.

An important feature of random forest is the internal measure of variable importance. The ranked variable importance allows the user to determine which variables were the most significant in the final model (Prasad *et al.*, 2006). However, variable importance calculated by the random forest ensemble does not automatically select the optimal number of variables that have the lowest error (Ismail, 2009). Results from this study indicate that variable selection is required to test all the possible combinations and select the model with the lowest error. More specifically, forward and backward variable selection methods were implemented using the random forest ensemble. Both variable selection methods produced a RMSE of 3.097, but the backward variable selection method was unable to significantly reduce the number of bands. One of the disadvantages of the backward variable selection is that the method assumes that the prediction accuracy does not decrease as the number of variables increases, therefore a good predictor could be removed early in the selection process (Gheyas and Smith, 2010). In a forward variable selection method, once

a variable is added it is not removed. The forward variable selection produced the best result and reduced the large number of hyperspectral bands (n = 230) to a few significant bands (n = 9). By reducing the number of variables the predictive accuracy of the final model was improved by 7%. It can therefore be concluded that the random forest ensemble can be an accurate and robust tool for band selection in a hyperspectral applications.

4.6. A comparison between multispectral and hyperspectral remotely sensed data to predict *P. patula* age.

Overall, the multispectral dataset resulted in the best predictive model (R^2 = 0.68), surprisingly outperforming the hyperspectral dataset (R^2 = 0.6). There are several possible reasons for the lack of performance from the hyperspectral data, (i) the hyperspectral analysis did not include a texture analysis, (ii) the spatial resolution of the hyperspectral image (2.4 m) was not as fine as the multispectral panchromatic image (0.6 m), (iii) the variables used in the final model were highly correlated, and (iv) there was large amount of noise present in the AISA Eagle imagery. The section below discusses these limitations in detail.

4.6.1. Texture analysis

The multispectral dataset (spectral and texture variables) resulted in a better predictive model than the hyperspectral dataset (only spectral variables). This indicates that it is not sufficient to work on a per pixel basis for discriminating forest age, but that the spatial arrangement of objects in the image need to be considered. Forest canopy structure changes with age which results in variations in image texture. This allows for relationships to be made between forest age and image texture (Johansen *et al.*, 2007). Previous research has shown that image texture can be effective for the discrimination of forest age (Franklin *et al.*, 2003; Johansen *et al.*, 2007; Wunderle *et al.*, 2007; Kayitakire *et al.*, 2006; Gerylo *et*

al., 2002; Jensen *et al.*, 1999). So, it follows that both spatial and spectral variables are needed to predict forest age.

4.6.2. Spatial resolution

The texture analysis on the multispectral dataset was carried out using 0.6 m pixel resolution. In the case of plantation forestry where the trees are planted close to one another and the range of ages are relatively small (1-24 years), pixel resolution is an important factor in predicting forest age. Due to the fact that a tree's structure changes with age, fine spatial resolution should improve age discrimination through the separation of small-scale features, such as foliage, gaps, and shadows that would be mixed in a moderate resolution pixel (Lefsky *et al.*, 2001). For example, mature forest canopies might be more accurately discriminated by measuring the quantity and spatial organization of shadow in their canopies (Cohen and Spies, 1992). The poor performance of the hyperspectral dataset could be due to the larger spatial resolution (2.4 m) of the hyperspectral image, compared to the multispectral panchromatic image (0.6 m).

4.6.3. Correlated variables

Results from the hyperspectral study predominately concentrate on bands in the red edge which are highly correlated (749.91 nm, 747.68 nm, 745.44 nm, 723.12 nm, 660.76 nm, 687.44 nm, 658.54 nm, 743.21 nm, 740.98 nm). Previous studies have shown that random forest is not affected by correlated variables (Archer and Kimes, 2008). However, research in the field of bioinformatics (Strobl *et al.,* 2008) has suggested that correlations between predictor variables can affect the random forest variable importance measures. A new conditional permutation method was suggested for calculating of variable importance that does not completely eliminate the preference of correlated variables, but provides a fair means of comparison that allows for the identification of relevant predictor variables. It is recommended that the proposed method be tested on

remote sensing data to establish not only whether it increases prediction accuracy, but also whether it selects non-correlated variables.

4.6.4. Image noise

The poor performance of the hyperspectral dataset could be explained by the amount of noise present in the hyperspectral dataset. Bands above 900 nm contained so much noise that they were excluded from the study. Additionally, the AISA Eagle image data was atmospherically corrected using the empirical line method which is based on the linear relationship between *in situ* measured ground reflectance and the sensor signal (Roberts *et al.*, 1986). The method assumes that the effects of the atmosphere are uniform across the entire image. In reality this is often not the case as there are differences in illumination across the image, for example, shadows caused from topography (Smith and Milton, 1999). The noise present in the hyperspectral image may have affected the overall result.

4.7. Recommendations for future research

Several recommendations for future research are discussed below:

- It is recommended that future research focuses on the effect of texture analysis on hyperspectral imagery for forest age prediction. Bunting *et al.* (2009) assessed the ability of textural information calculated from 2.6 m HyMap hyperspectral data for the classification of broad forest types. Texture measures included first and second order derivatives at different scales. A multiple stepwise discriminant analysis was performed on the resulting datasets. A classification accuracy of 60% was achieved using the combined reflectance and texture data, compared to the classification accuracies of 55% and 43% using only the reflectance and texture datasets. This result shows that a texture analysis using a hyperspectral

dataset can increase classification accuracy. It is also recommended that future research looks at hyperspectral data to estimate chlorophyll content. Texture analysis could be conducted on chlorophyll spectral variables and related to forest age.

- This research has highlighted the importance of spatial resolution for forest age discrimination. This aspect needs to be explored further to establish which pixel size best discriminates forest age in commercial forestry, where *P. patula* age ranges from 1 to 24. In order for remote sensing methods to become operational for mapping forest variables, the spatial resolution of the image needs to be suitable for the specific application (Treitz and Howarth, 2000). It is recommended that future research focuses on methods that consider objects at their optimal spatial resolution (Marceau *et al.*, 1994). By finding the optimal spatial resolution for specific features of interest, the information content per pixel is increased (Atkinson, 1997). Previous studies have shown that the optimal spatial resolution and analysis technique can be derived by examining the spatial variation between objects in an image (Atkinson, 1993; Atkinson; 1997; Atkinson and Aplin, 2004; Woodcock and Strahler, 1987). Future studies should determine on what basis the researcher should select the optimal spatial resolution for mapping forest age.

- The performance of the random forest ensemble in regression applications using a dataset where the number of samples exceeds the variables (small *n* large *p*) is not fully understood. Current research has focused on classification applications (Ismail, 2009; Adam *et al.,* in press). The few studies that have used random forest for prediction in regression applications (Ismail and Mutanga, 2010; Grimm, 2009) were not $n < p$ applications. Additionally, no single machine learning algorithm is superior in all applications (Kohavi *et al.*, 1996) and it is therefore recommended that future studies compare the robustness of the random forest ensemble

against other tree based ensembles (e.g. bagging, boosting). Additionally, the random forest ensemble should also be tested against other methods such as artificial neural networks which have proved successful in forest age prediction (Jensen *et al.*, 1999).

- The effect of noise on the random forest ensemble should be examined in more detail. Previous research has shown that random forest is robust to noise because the ensemble exploits noise in the dataset in order to create a more diverse classifier or predictor (Hamza and Larocque, 2005). However, Ismail (2009) has recently shown that classification accuracies decline substantially due to noise, especially in applications when there are limited samples compared to variables (i.e. $n < p$). The effect of noise on the random forest ensemble for regression applications needs to be fully examined.

4.8. Conclusion

The aim of this thesis was to investigate the potential of remote sensing technologies and the random forest ensemble to predict *P. patula* age. The research carried out in this study showed that it is possible to map forest age using a combination of spatial and spectral remote sensing technologies. The final conclusion was based on the following observations in this thesis:

1. Multispectral remotely sensed data has the capability to accurately predict forest age. The combination of NIR, green, variance with a 3 x 3 window, red and blue resulted in the best predictive model ($R^2 = 0.68$).
2. The random forest ensemble has shown great potential for combining spectral and texture variables to accurately predict the age of *P. patula* stands. More specifically, (i) calculating principal components of the texture datasets reduced data redundancy and produced better results

than simply using the original texture images, ii) combining spectral and texture variables according to window size (for example 3 x 3 contrast, 5 x 5 contrast and so on) was more effective than combing the variables according the optimal texture measure (for example 3 x 3 contrast, 3 x 3 variance and so on), and (iii) utilizing a backward variable selection incorporated with the random forest ensemble was successful in reducing the large number of bands, and resulted in the best predictive model.

3. The random forest ensemble can successfully reduce a high dimensional hyperspectral dataset to an optimal set of bands and improve model accuracy. Utilizing all the hyperspectral bands (n = 230) resulted in a relatively poor predictive accuracy (R^2 = 0.53). However, the forward variable selection method selected nine significant bands and resulted in a higher predictive accuracy (R^2 = 0.6).
4. The forward variable selection method was more effective in reducing the large number of hyperspectral bands than the backward variable selection method. Although both variable selection methods resulted in the same RMSE (3.097), the backward variable selection method selected 206 bands, while the forward variable selection method selected only nine bands.
5. The multispectral dataset (texture and spectral variables) was more successful in predicting *P. patula* age than the hyperspectral dataset (spectral). Several reasons for this have been suggested and recommendations were made to improve the performance of a hyperspectral dataset for predicting the age of *P. patula* stands.

REFERENCES

Adam, E.M., Mutanga, O., Rugege, D., and Ismail, R. In press. Discriminating the papyrus vegetation (*Cyperus papyrus L.*) and its co-existent species using random forest and hyperspectral data resampled to HYMAP. *International Journal of Remote Sensing*.

Ahern, F.J., Erdle, T., Maclean, D.A., and Kneppeck, I.D. 1991. A quantitative relationship between forest growth rates and thematic mapper reflectance measurements. *International Journal of Remote Sensing*, 12: 387-400.

Archer, K.J. and Kimes, R.V. 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, 52: 2249-2260.

Atkinson, P.M. 1993. The effect of spatial resolution on the experimental variogram of airborne MSS imagery. *International Journal of Remote Sensing*, 14: 1005-1011.

Atkinson, P.M., 1997. Selecting the spatial resolution of airborne MSS imagery for small scale agricultural mapping. *International Journal of Remote Sensing*, 18: 1903-1917.

Atkinson, P.M. and Aplin, P. 2004. Spatial variation in landcover and choice of spatial resolution for remote sensing. *International Journal of Remote Sensing*, 25(2): 3687-3702.

Bajcsy, P. and Groves, P. 2004. Methodology for hyperspectral band selection. *Photogrammetric Engineering and Remote Sensing Journal*, 70: 793-802.

Bel, L., Allard, D., Laurent, J.M., Chaddadi, R., and Bar-Hen, A. 2009. CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics and Data Analysis*, 53: 3082-3093.

Blackburn, G.A. 2007. Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany*, 58(4): 855-867.

Borges, J.S., Marcal, A.R.S., and Dias, J.M.B. 2007. Evaluation of feature extraction and reduction methods for hyperspectral images. In *Proceedings of the 26th EARSeL Symposium, Warsaw, Poland, 29 May-2 June 2006*, *Bochenek,* Z. (Eds.): New Developments and Challenges in Remote Sensing: 225-264.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. Classification and Regression Trees. Wadsworth International Group. Belmont.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45: 5-32.

Buddenbaum, H., Schlerf, M., and Hill, J. 2005. Classification of coniferous tree species and age classes using hyperspectral data and geostatistical methods. *International Journal of Remote Sensing*, 26(24): 5453-5465.

Bunting, P., He, W., Zwiggelaar, R., and Lucas, R. 2009. Combining Texture and Hyperspectral Information for the Classification of Tree Species in Australian Savanna Woodlands. In: Jones, S. And Reinke, K. (eds.). *Innovation in Remote Sensing and Photogrammetry*. Springer. Berlin.

Chan, J.C-W. and Paelinckx, D. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112: 2999–3011.

Cho, M. 2007. Hyperspectral remote sensing of biochemical and biophysical parameters: The derivative red edge “double-peak feature”, a nuisance or an opportunity? PhD dissertation, C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PEandRC), Wageningen University, the Netherlands.

Cho, M.A., Skidmore, A.K., and Sobhan, I. 2009. Mapping beech (*Fagus sylvatica L.*) forest structure with airborne hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 11: 201-211.

Clark, M.L., Roberts, D.A., and Clark, D.B. 2005. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, 96: 375-398.

Cochrane, M.A. 2000. Using vegetation reflectance variability for species level classification of hyperspectral data. *International Journal of Remote Sensing*, 21(10): 2075-2087.

Cohen, W. B. and Spies, T. A. 1992. Estimating structural attributes of Douglas-Fir/Western Hemlock forest stands from LANDSAT and SPOT imagery. *Remote Sensing of Environment*, 41: 1-17.

Curran, P. 1980. Multispectral remote sensing of vegetation amount. *Progress in Physical Geography*, 4: 315-41.

Cutler, R.D., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., and Lawler, J.J. 2007. Random forests for classification in ecology. *Ecology*, 88(11): 2783-2792.

Danson, F.M., and Curran, P.J. 1993. Factors affecting the remotely sensed response of coniferous forest plantations. *Remote sensing of Environment*, 43: 55-65.

De'ath, G. and Fabricius, K. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11): 3178-3192.

DeFries, R.S., Hansen, M., Steininger, M., Dubayah, R., Sohlberg, R. and Townshend, J. 1997. Subpixel forest cover in Central Africa from multisensor, multitemporal data. *Remote Sensing of Environment*, 60: 228-246.

DigitalGlobe 2008. http://www.digitalglobe.com/index.php/85/QuickBird (accessed November 2009).

Dunne, K., Cunningham, P., and Azuaje, F. 2002. Solutions to instability problems with sequential wrapper-based approaches to feature selection (Technical Report TCD-2002-28). Dept. of Computer Science, Trinity College, Dublin, Ireland.

Dye, M., Mutanga, O., and Ismail, R. 2008. Detecting the severity of woodwasp, *Sirex noctilio,* 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.

Eastman, J.R., and Fulk, M., 1993. Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing*, 59 (6): 991-996.

Elith, J., Leathwick, J.R. and Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: 802-813.

ENVI. 2006. *ENVI Version 4.3*. ITT Industries, Inc. Boulder. Colorado.

ESRI. 2006. *ArcGIS Version 9.1*. ERSI, California.

Foody, G.M., and Arora, M. 1996. Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications. *Pattern Recognition Letters*, 17: 1389-1398.

Foody, G.M., and Mathur, A. 2004. Toward intelligent training of supervised image classifications: Directing training data acquisition for SMV classification, *Remote Sensing of Environment*, 93: 107-117.

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.

Franklin, S.E., Wulder, M.A., and Gerylo, G.R. 2001. Texture analysis of IKONOS panchromatic data for Douglas-fir forest age class separability in British Columbia. *International Journal of Remote Sensing*, 22(13): 2627-2632.

Franklin, S.E., Hall, R.J., Smith, L., and Gerylo, G.R. 2003. Discrimination of conifer height, age and crown closure classes using Landsat-5 imagery in the Canadian Northwest Territories. *International Journal of Remote Sensing*, 24(9): 1823-1834.

Gebreslasie, M.T., Ahmed, F.B., and van Aardt, J. 2008. Estimating plot-level forest structural attributes using high spectral resolution ASTER satellite data in even-aged *Eucalyptus* plantations in southern KwaZulu-Natal, South Africa. *Southern Forests*, 70(3): 227-236.

Gerylo, G.R., Hall, R.J., Franklin, S.E., and Smith, L. 2002. Empirical relations between Landsat TM spectral response and forest stands near Fort Simpson, Northwest Territories, Canada. *Canadian Journal of Remote Sensing*, 28(1): 68–79.

Gheyas, A. and Smith, L.S. 2010. Feature subset selection in large dimensionality domains *Pattern Recognition*, 43: 5-13.

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R., 2006. Random Forests for land cover classification. *Pattern Recognition Letters*, 27: 294-300.

Goetz, A.F.H. 2009. Three decades of hyperspectral remote sensing of the Earth: A personal view. *Remote Sensing of Environment*, 113: 5-16.

Govender, M., Chetty, K., and Bulcock, H. 2007. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water South Africa*, 33: 145-152.

Govender, M., Chetty, K., Naiken, V. and Bulcock, H. 2008. A comparison of satellite hyperspectral and multispectral remote sensing imagery for improved classification and mapping of vegetation. *Water South Africa*. 34:147-154.

Gower, S.T., Kucharik, C.J., and Norman, J.M., 1999. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote Sensing of Environment* 70 (1): 29–51.

Granitto, P.M., Furlanello, C., Biasioli, F. and Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83: 83-90.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – Digital soil mapping using Random Forests analysis. *Geoderma*, 146: 102-113.

Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3: 1157-1182.

Hall-Beyer, M. 2010. The GLCM Tutorial Home Page. University of Calgary. Canada. Online at: http://www.ucalgary.ca (accessed March 2010).

Hamza, M. and Larocque, D., 2005. An empirical comparison of ensemble methods based on classification trees. *Journal of Computation and Simulation*, 75(8): 629-643.

Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., Dimiceli, C., and Carroll, M. 2002. Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data. *Remote Sensing of Environment*, 83: 303-319.

Haralick, R.M., Shanmugan, K., and Dinstein, I. 1973. Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6): 610–621.

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14: 55–63.

Ismail, R. 2009. Remote sensing of forest health: The detection and mapping of *Pinus patula* trees infested by *Sirex noctilio*. PhD dissertation, School of Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

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

Jakubauskas, M.E. and Price, K.P. 2000. Regression-based estimation of lodgepole pine forest age from Landsat Thematic Mapper Data. *Geocarto International*, 15(1): 1-6.

Janecek, A.G.K., Gansterer, W.N., Demel, M.A., and Ecker, G.F. 2008. On the relationship between feature selection and classification accuracy. *JMLR: Workshop and Conference Proceedings*, 4: 90-105.

Jenson, J.R., Qui, F., and Ji, Minhe. 1999. Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. *International Journal of Remote Sensing*, 20 (14): 2805-2822.

Jobanputra, R., and Clausi, D.A. 2006. Preserving boundaries for image texture segmentation using grey level co-occurring probabilities. *Pattern Recognition*, 9: 234-245.

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: 29-44.

Kayitakire, F., Giot, P., and Defourny, P. (2002). Automated delineation of the forest stands using digital color orthophotos: Case study in Belgium. *Canadian Journal of Remote Sensing*, 28(5): 629–640.

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

Kocev, D., Dzeroski, S., White, M., Newell, G.R., Griffioen, P. 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modeling*, 220(8): 1159-1168.

Kohavi, R. and John, G.H. 1997. Wrappers for feature subset selection. *Artifical Intelligence Journal*, 97(1-2): 273-324.

Kohavi, R., Sommerfield, D., and Dougherty, J. 1996. Data Mining using MLC++: A Machine Learning Library in C++, Tools with Artifical Intelligence. *IEEE Computer Society Press*: 234-245.

Kuhn, K. 2009. Variable selection using the caret package. http://ftp.udc.es/public/CRAN/web/packages/caret/vignettes/caretSelection.pdf (accessed November 2009)

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

Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. *R News*, 2/3: 18-22.

Lefsky, M.A., Cohen, W.B., and Spies, T.A. 2001. An evaluation of alternate remote sensing products for forest inventory, monitoring, and mapping of Douglas-fir forests in western Oregon. *Canadian Journal of Forest Research*, 31: 78–87.

Lobell, D.B., Oritiz-Monasterio, J.I., Asner, G.P., Naylor, R. and Falcon, W. 2007. Combining field surveys, remote sensing and regression trees to understand yield variations in an irrigated wheat landscape. *Agronomy Journal*, 97: 241-249.

Marceau, D.J., Gratton, D.J., Fournier, R.A., and Fortin, J., 1994. Remote sensing and the measurement of geographical entities in a forested environment. 2. The optimal spatial resolution. *Remote Sensing of Environment*, 49: 105-117.

Martin, M.E., Newman, S.D., Aber, J.D., and Congalton, R.G. 1998. Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing of Environment*, 65: 249-254.

Melgani, F., and Bruzzone, L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions of Geoscience and Remote Sensing*, 42(8): 1778-1790.

Michaelson, J., Schimel, D.S., Friedl, M.A., Davis, F.W. and Dubayah, R.O. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science*, 5: 673-696.

Moskal, L.M. and Franklin, S.E. 2001. *Classifying multilayer forest structure and composition using high resolution, compact airborne spectrographic imager image texture*. In American Society of Remote Sensing and Photogrammetry Annual Conference. St. Louis.

Mucina, L., and Rutherford, M.C. (eds.). 2006. *The Vegetation of South Africa, Lesotho and Swaziland*. Strelitzia 19. South African National Biodiversity Institute. Pretoria.

Mutanga, O., and Skidmore, A.K. 2004. Integrating imaging spectroscopy and neural networks to map tropical grass quality in the Kruger National Park, South Africa. *Remote Sensing of Environment*, 90: 104-115.

Nieman, K.O. 1995. Remote sensing of forest stand age using airborne spectrometer data. *Photographic Engineering and Remote Sensing*, 61: 1119-1127.

Owen, D.L. (ed.) 2000. *Southern African Forestry Handbook*. Volume 1. The Southern African Institute of Forestry.

Pal, M., and Mather, M. 2004. Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generations Computer Systems*, 20: 1215-1225.

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

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P., and Huybrechts, W. 2007. Random forests as a tool for ecohydrological distribution modeling. *Ecological modeling*, 207: 304–318.

Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., and Trianni, G. 2009. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113: 110–122.

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.

R Development Core Team 2008. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ricotta, C., Avena, G.C., and Volpe, F. 1999. The influence of principal component analysis on the spatial structure of a multispectral dataset. *International Journal of Remote Sensing*, 20 (17): 3367-3376.

Roberts, D.A., Yamaguchi, Y. and Lyon, R.J.P. 1986. *Comparison of various techniques for calibration of AIS data*. In Proceedings of the Second Airborne Imaging Spectrometer Data Analysis workshop. JPL Publication 86–35, Pasadena CA: Jet Propulsion Laboratory: 21–30.

Saeys, Y., Inza, I., and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics Advance Access. Published August 24.

Schlerf, M., Atzberger, C., and Hill, J. 2005. Remote sensing of biophysical variables using HyMap imaging spectrometer data. *Remote Sensing of Environment*, 95: 177-194.

Schmidt, K.S. 2003. *Hyperspectral remote sensing of vegetation species distribution in Saltmarsh.* International Institute for Geo-Information Science and Earth Observation. Enchede, Netherlands.

Shafri, H.Z.M., Salleh, M.A.M., and Ghiyamat, A. 2006. Hyperspectral Remote Sensing of Vegetation Using Red Edge Position Techniques. *American Journal of Applied Sciences*, 3(6): 1864-1871.

Smith, G. M. and Milton, E. J. 1999. The use of the empirical line method to calibrate remotely sensed data to reflectance. *International Journal of Remote Sensing*, 20: 2653–2662.

Spatial Ecology. 2010. Hawth's Analysis Tools for ArcGIS. www.spatialecology.com/htools/overview.php (accessed April 2010).

St-Louis, V, Pidgeon, A.M., Radeloff, V.C., Hawbaker, T.J., and Clayton, M.K. 2006. High resolution image texture as a predictor of bird species richness. *Remote Sensing of Environment*, 105: 299-312.

Strobl, C., Boulesteix, A-L., Kneib, T., Augustin, T., and Zeileis, A. 2008. Conditional Variable Importance for Random Forests. Technical Report Number 23. Department of Statistics. University of Munich.

Strobl, C. and Zeileis, A. 2008. Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance. 2008. *Proceedings in Computational Statistics*, 2: 59-66.

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R. and Feuston, B., 2003. Random Forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Science*, 43(6): 1947-1958.

Tomppo, E.O., Gagliano, C., De Natale, F., Katila, M., and McRoberts, R.E. 2009. Predicting categorical forest variables using an improved k-Nearest

Neighbour estimator and Landsat imagery. *Remote Sensing of Environment*, 113: 500–517.

Tuttle, E.M., Jensen, R.R., Formica, V.A., Gonser, R.A. 2006. Using remote sensing image texture to study habitat use patterns: a case study using the polymorphic white-throated sparrow (Zonotrichia albicollis). *Global Ecology and Biogeography*, 15: 349–357.

Treitz, P.M. and Howarth, P.J. 1999. Hyperspectral remote sensing for estimating biophysical parameters of forest ecosystems. *Progress in Physical Geography*, 23(3): 359–390.

Treitz, P.M. and Howarth, P.J., 2000. High spatial resolution remote sensing data for forest ecosystem classification: an examination of spatial scale. *Remote Sensing of Environment*, 72: 268-289.

Vaiphasa, C., Ongsomwang, S., Vaiphasa, T. and Skidmore, A.K., 2005. Tropical mangrove species discrimination using hyperspectral data: a laboratory study. *Estuarine, Coastal, and Shelf Science*, 65: 371-379.

Vaiphasa, C., Skidmore, A., de Boer, W. and Vaiphasa, T., 2007. A hyperspectral band selector for plant species discrimination. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62: 225-235.

Van Aardt, J.A.N. and Norris-Rogers, M. 2008. Spectral-age interactions in managed, even-aged *Eucalyptus* plantations: application of discriminant analysis and classification and regression trees approaches to hyperspectral data. *International Journal of Remote Sensing*. 29: 1841-1845.

Van Aardt, J.A.N. and Wynne, H. 2007. Examining pine seperability using hyperspectral data from an airborne sensor: An extension of field-based results. *International Journal of Remote Sensing*, 28(1-2): 431-436.

Vane, G. and Goetz, A. 1993: Terrestrial imaging spectrometry: current status, future trends. *Remote Sensing of Environment*, 44: 117-26.

Woodcock, C.E. and Strahler, A.H. 1987. The factor of scale in remote sensing. *Remote Sensing of Environment*, 21: 311-332.

Wulder, M. 1998. Optical remote-sensing techniques for the assessment of forest inventory and biophysical parameters. *Progress in Physical Geography*, 22 (4): 449-476.

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

Yuan, X., King, D., and Vleck, J. 1991. Sugar Maple Decline Assessment Based on Spectral and Textural Analysis of Multispectral Aerial Videography. *Remote Sensing of Environment*, 37: 47-54.

Zhang, C., Franklin, S.E., and Wulder, M.A. 2004. Geostatistical and texture analysis of airborne-acquired images used in forest classification. *International Journal of Remote Sensing*, 4: 859–865.