Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Does simultaneous variable selection and dimension reduction improve the classification of *Pinus* forest species?

Kabir Yunus Peerbhay
Onisimo Mutanga
Riyad Ismail

SPIE.

Does simultaneous variable selection and dimension reduction improve the classification of *Pinus* forest species?

Kabir Yunus Peerbhay,* Onisimo Mutanga, and Riyad Ismail
University of KwaZulu-Natal, School of Agriculture, Earth and Environmental Sciences, Discipline of Geography, P/Bag X01, Scottsville 3209, Pietermaritzburg, South Africa

Abstract. Tree species information is important for forest inventory management and supports decisions related to the composition and distribution of forest resources. However, traditional methods of obtaining such information involve time consuming and cost intensive ground-based methods. Hyperspectral data offer an alternative source for obtaining information related to forest inventory. Utilizing Airborne Imaging Spectrometer for Applications Eagle hyperspectral data (393 to 994 nm), this study compares the utility of two partial least squares (PLS)-based methods for the classification of three commercial Pinus tree species. Results indicate that the sparse partial least squares discriminant analysis (SPLS-DA) method performed variable selection and dimension reduction successfully to produce an overall accuracy of 80.21%. In comparison, the PLS-DA method and variable importance in the projection (VIP) selected bands produced an overall accuracy of 71.88%. The most effective bands selected by PLS-DA and VIP coincided within the visible region of the spectrum (393 to 700 nm). However, SPLS-DA selected fewer wavebands within the blue (415 to 483 nm), green (515 to 565 nm), and red regions (674 to 694 nm) to confirm the importance of the visible in discriminating tree species. Overall, this study shows the potential of SPLS-DA to perform simultaneous variable selection and dimension reduction of hyperspectral remotely sensed data resulting in improved classification accuracies. [DOI: 10.1117/1.JRS.8.085194]

Keywords: tree species classification; hyperspectral; forestry; sparse partial least squares discriminant analysis; variable importance in the projection.

Paper 14382SS received Jun. 29, 2014; accepted for publication Nov. 25, 2014; published online Dec. 22, 2014.

1 Introduction

Accurate tree species information is a substantial part of any forest inventory and supports forest managers' efforts to conduct sound management decisions.[1] Tree species identification provides valuable spatial data that may benefit operational tasks such as modeling the spread of pest and pathogens, such as *Sirex noctilio*,[2] promoting effective weed control strategies in relation to particular forest species,[3,4] determining optimal bioclimatic site conditions[5] and species level carbon sequestration.[6,7] Additionally, determining the composition and distribution of tree species is valuable for assessing indicators related to the ecological integrity of forest ecosystems and could assist in monitoring ecosystem health and ultimately guide forest management policies.[8,9] However, obtaining information on forest tree species is challenging when using traditional approaches.

While ground-based methods such as field measurements prove to be costly, time consuming and labor intensive, remote sensing provides a reliable alternative for obtaining information for forest inventory.[10] Hyperspectral remotely sensed data have often provided more effective results for mapping tree species over multispectral data, due to the improved spectral resolution that

*Address all correspondence to: Kabir Yunus Peerbhay, E-mail: Peerbhaykabir@gmail.com

samples the electromagnetic spectrum using hundreds of narrow wavebands.[11,12] Mapping forests at species level, however, is challenging since tree species exhibit reflectance that are strongly correlated.[11] The variation present at canopy scale may further hamper tree species discrimination applications due to the effects of tree age, phenology, nonphotosynthetic material, and background effects.[13,14] Additionally, studies have generally expressed difficulties in classifying tree species that are closely related and within the same genus,[15–18] since the variation between subgenera species is less than the variation between species of different genera.

For example, Goodwin et al.[15] showed that the majority of the *Eucalyptus* species considered in their study was individually inseparable compared to other mesic vegetation; however, they obtained an overall accuracy of 94% when merging all of the *Eucalyptus* species into one class. Reference 16 discriminated 11 forest types including mixed species and produced an overall accuracy of 75%, yet the study was unsuccessful in classifying individual deciduous species. Recently studies have applied feature selection methods in the context of tree species classification.[19–22] For instance, Dalponte et al.[19] used support vector machines (SVM) and Airborne Imaging Spectrometer for Applications (AISA) Eagle hyperspectral data to classify 11 Southern Alps tree species and produced the best kappa accuracy of 0.70, with user's and producer's accuracies ranging between 60% and 100%. Hyperspectral data were combined with Lidar data to map five tree species at different scales using SVM and random forest (RF) classifiers.[21] Minimum noise fraction (MNF) transformed bands with an 8-m spatial resolution produced the best accuracy of 86% and a kappa value of 0.83. Using SVM and RF, Fassnacht et al.[22] compared three feature selection methods to classify tree species at three different test sites. SVM classification results in conjunction with MNF input data proved significant in most cases and outperformed results produced by RF when using genetic algorithm (GA), SVM wrapper and sparse generalized PLS selection methods. Finally, using AISA Eagle image data, Peerbhay et al.[20] showed that it was possible to accurately classify six commercial forest species using the PLS discriminant analysis (PLS-DA) algorithm. The study produced an overall accuracy of 80.61%, a kappa value of 0.77 and user's and producer's accuracies ranging from 50% to 100%.

The PLS-DA algorthim is able to suppress background effects, address the spectral similarity between tree species and can effectively deal with the computational and statistical problems associated with hyperspectral datasets.[20] The method is based on the decomposition of explanatory variables (i.e., the hyperspectral wavebands) into PLS latent components that retain the most important information.[23,24] However, the generation of fewer initial components from highly correlated wavelengths are suggested to reduce the chances of model overfitting.[24]

While only a few studies have investigated the utility of PLS for classification in remote sensing,[20,25] PLS-DA has become popular in other research domains. Some of these domains include genetics,[26] biology,[27,28] and chemometrics.[29,30] However, in the analysis of hyperspectral data, it is also of interest to identify the most effective spectral regions that allow for the best discrimination between samples.[31] While PLS alone does not provide insight on the most effective bands that may contribute to the final classification task,[32] the utility of novel variable selection techniques has been advocated. Many studies often adopt preselection approaches for variable selection in order to improve the performance of PLS classifications.[26,33,34] Usually, these approaches are based on some criterion to select high ranking variables which are later included for PLS analysis. For instance, Peerbhay et al.[20] showed that selecting wavebands based on the variable importance in the projection (VIP) score is a robust measure for determining individual waveband importance and for producing the best PLS classification accuracies. In their study, incorporating the optimal subset of VIP selected wavebands ($n = 78$) in the PLS-DA model resulted in an improved overall accuracy of 88.78% and a kappa value of 0.87, with user's and producer's accuracies ranging from 70% to 100%.

Although preselection approaches have been effective, their execution does not involve a complete and computationally efficient way of selecting important variables while performing simultaneous classification. Nonetheless, certain studies have extended the PLS approach to impose sparseness within the technique for the combined purpose of variable selection and dimension reduction.[35] Designed explicitly for optimal group discrimination in high-dimensional settings, SPLS-DA effectively overcomes the problem of being affected by a large number of predictors.[35] This ability makes SPLS-DA well suited for analyzing high-dimensional data and for selecting important variables when classifying features of interest.

It is within this context that this study aims to determine whether simultaneous variable selection and dimension reduction improves the classification of *Pinus* tree species (*Pinus taeda*, *Pinus elliotii*, *Pinus patula*) using SPLS-DA and AISA Eagle hyperspectral imagery. In addition, incorporating wavebands selected by the VIP method with PLS-DA were assessed.

2 Methods and Materials

2.1 *Study Area*

The research was conducted in the 6391 ha Sappi Hodgsons plantation (Centroid: 29° 13′18″ S and 30° 23′13″ E) in KwaZulu-Natal, South Africa (Fig. 1). Evenly aged stands consisting of *P. patula*, *P. elliotii*, *P. taeda* are the dominant commercial softwood tree species occurring in

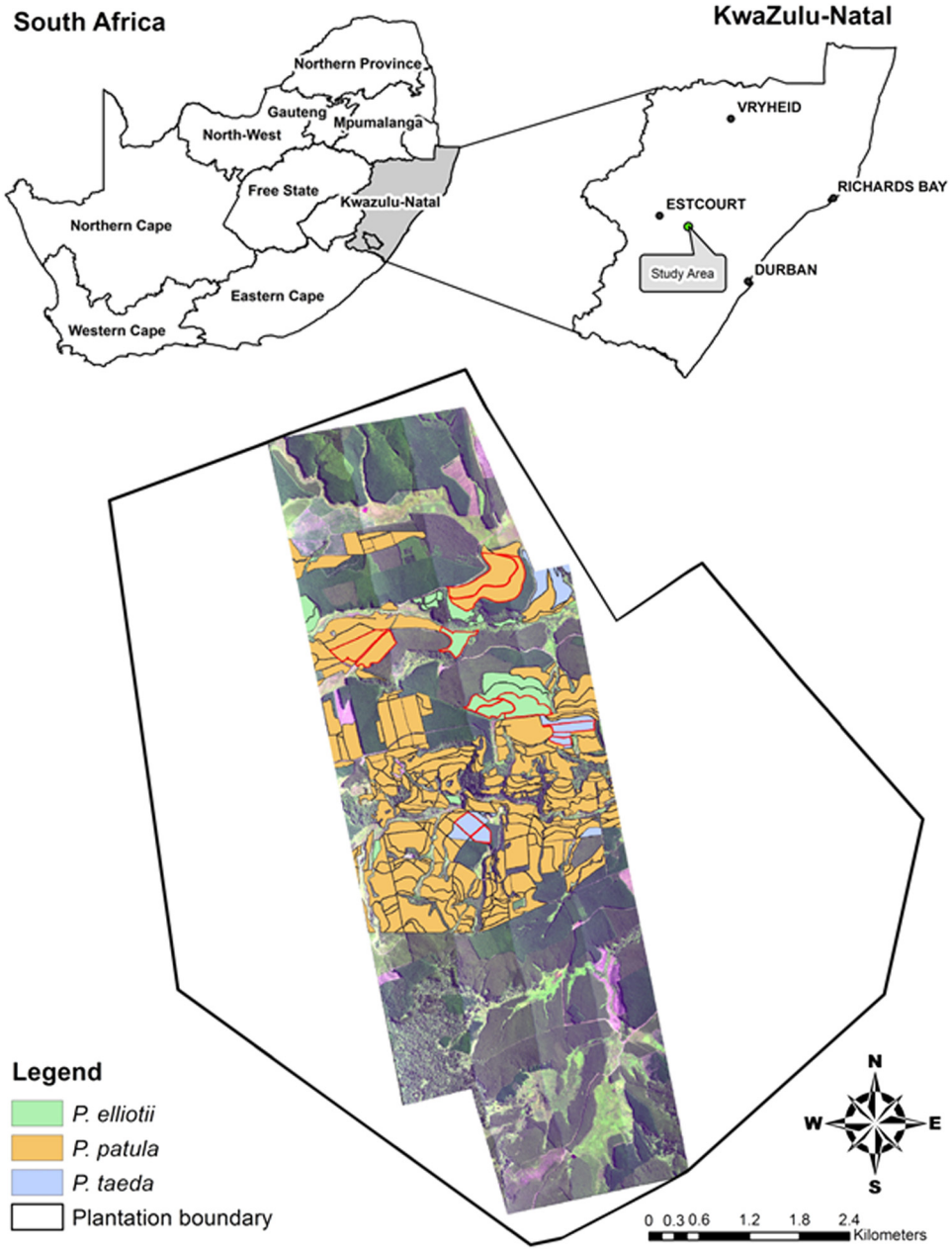

Fig. 1 Location of the study area and the composition of tree species in the Airborne Imaging Spectrometer for Applications (AISA) Eagle hyperspectral scene. Forest stands selected in this study ($n = 12$) are indicated in red.

the study area. The plantation is situated in the mist belt grassland bioregion of the KwaZulu-Natal midlands with average temperatures in the region of 15.9°C. Rainfall ranges between 730 and 1280 mm/annum, with highly variable precipitation occurring during the summer and additional moisture is provided by heavy mist during the winter.[36] The relief of the area is generally hilly and covered by diminutive grasslands with slopes peaking between 1030 and 1590 m above sea level.[37] The establishment of the invasive tree, *Solanum mauritianum* (bugweed), within the study area has not gone unnoticed. Bugweed trees primarily grow in association with the *Pinus* trees in low to high densities. The prolific dispersal of bugweed is particularly evident when extensive occurrences dominate parts of the forest canopy, whereas other *Pinus* stands are richly invaded in the forest understory. Due to the prevalence of bugweed trees occurring within the *Pinus* stands, the invader species was included in this study to provide a more realistic assessment of the classification method.

2.2 *Hyperspectral Image Acquisition and Preprocessing*

During the summer of February 2009, AISA Eagle hyperspectral imagery was obtained under cloudless conditions. Four AISA flight lines with a pixel size of 2.4 m were collected. The applied sensor delivers hyperspectral imagery in 272 bands with a spectral range between 393.23 and 994.09 nm.

A light aircraft was used to collect the hyperspectral imagery at a mean GPS altitude of 2728.42 m and a swath width of 3058 m. The image was atmospherically calibrated using the empirical line method,[38] which is based on the linear relationship between *in situ* measured ground reflectance and the sensor spectral signal. The Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer (350 to 2500 nm) was used for the acquisition of field measurements to calibrate the reflectance data. The image was topographically corrected using a digital elevation model with contours of 5 m created from 1:50 000 topographic maps. The image was referenced to the Universal Transverse Mercator (UTM zone 36S) projection using WGS-84 Geodetic system. Although wavebands after 900 nm showed the presence of spectral noise, these bands were included in this study. ENVI 4.7 image processing software[39] was used for the preprocessing of the AISA Eagle imagery.

2.3 *Training Data*

Field data for *P. taeda*, *P. elliotii*, and *P. patula* consisted of four forest stands per species that were randomly selected from all the forest stands occurring in the study area. A field visit was conducted to assess the condition of the selected *Pinus* species and coincided with the acquisition of the AISA Eagle imagery during February. Each pine stand was further subsampled randomly using field points to collect image spectra from single pixels (Table 1). Additionally, the occurrences of bugweed within the selected *Pinus* stands were recorded in field and used as point samples to collect image spectra. Using the R statistical software package,[40] the number of test and training samples for each species was then statistically balanced. This was implemented to ensure the ideal optimization of the PLS-DA models and classification using hyperspectral data.[41] Figure 2 displays the average spectral reflectance curves in each of the tree species considered in this study.

Table 1 The sample size for the respective tree species considered in the study.

Species	Number of tree stands	Point samples	Total sample size
P. patula	4	20	80
P. taeda	4	20	80
P. elliotii	4	20	80
S. mauritianum	80	1	80

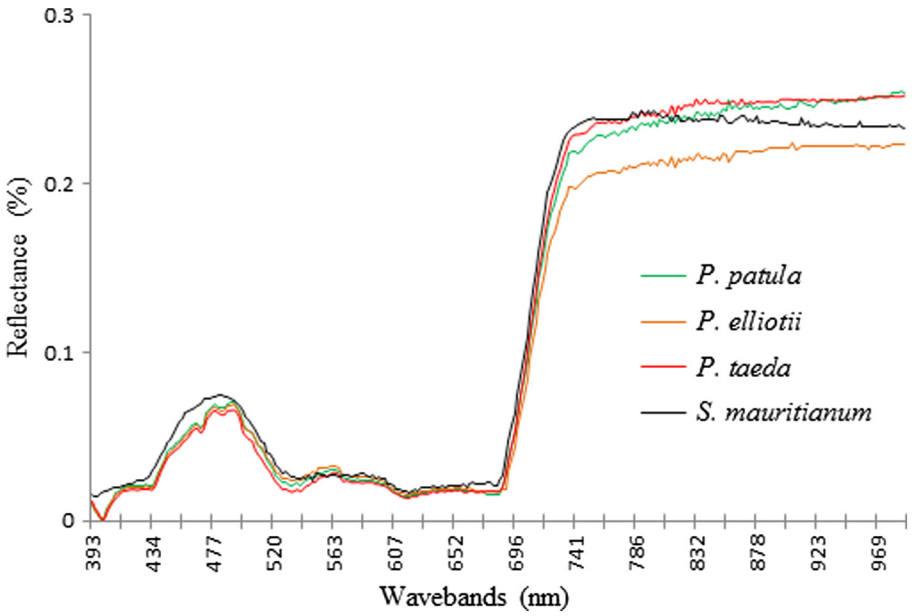

Fig. 2 Average spectral reflectance curves of the three pine tree species and bugweed considered in this study.

2.4 *Statistical Analysis*

2.4.1 *Partial least squares discriminant analysis*

PLS-DA is based upon the classical PLS regression method for constructing predictive models,[42] where dimension reduction and the latent decomposition of the X and Y matrices is principle. PLS projects the X matrix in the K dimension space where each column of X defines one coordinate axis. In an A-dimensional hyperplane, which is represented by one line and one direction per component, the X matrix is projected down onto an orthogonal axis, whereas at the same time, the positions of the projected data are related to the values of the response matrix (Y).[42] Since the latent component matrix (T) produces K linear combinations or scores for X and Y, finding the direction vectors within T is focal to a PLS operation. PLS seeks the columns of which direction vectors relate to X and Y and obtains the most effective variable directions in the X space.[23,43] The method can be statistically described by

$$\mathrm{X} = \mathrm{TP}' + \mathrm{E}, \tag{1}$$

where X represents the matrix of the wavebands ($n = 272$), T is a factor score matrix, P is the X loadings, and E is the residual or a noise term.

$$\mathrm{Y} = \mathrm{TQ}' + \mathrm{F}, \tag{2}$$

where Y is a matrix of the response variable (forest species), T is the scores for Y, Q is the Y loadings and F is the residuals.

2.4.2 *Variable importance in the projection*

While PLS-DA provides no insight regarding the most effective wavelengths that may contribute toward the final classification,[32] studies have demonstrated the benefit of utilizing the VIP score for identifying individual waveband importance[34,42,44] and determining the most effective spectral regions for classification.[26,27] The VIP method[42] computes the importance of each waveband by producing scores that serve as a ranked measure of importance amongst the explanatory variables.[33] Using the VIP scores to preselect important wavebands in a dataset is, therefore, an

essential requirement for a PLS model to achieve good classification performance[26] and is defined as follows:

$$\mathrm{VIP}_k = \sqrt{K \sum_{a=1}^{A} [(q_a^2 t_a^T t_a)(w_{ak}/\|w_k\|^2)] / \sum_{a=1}^{A} (q_a^2 t_a^T t_a)}, \tag{3}$$

where VIP_k is the importance of the k'th waveband based on a PLS-DA model with a components, w_{ak} is the corresponding loading weight of the k'th waveband in the a'th PLS-DA component, t_a, w_a, and q_a are the a'th column vectors, and K is the total number of bands.[45] The important variables of the PLS-DA model were identified by selecting those wavebands that had a VIP score of >1, since the average of squared VIP scores is equal to 1.[33] A new PLS-DA model using the selected VIP bands was developed and then used to classify the test dataset.

2.4.3 *Sparse partial least squares discriminant analysis*

SPLS-DA closely follows the PLS-DA approach whereby the categorical response variables are initially observed as continuous in order to construct latent components. However, SPLS-DA imposes sparseness within the latent components to promote variable selection while performing simultaneous dimension reduction. Irrelevant and noisy variables are scored to zero by imposing L_1 penalty,[46] thus eliminating any contribution toward the models' discrimination power. In addition, the latent components are built to explain the best discrimination among classes by using only the few informative variables (non-zero variables). Class membership of each variable is then assigned by reference cell coding the response matrix (Y) with dummy variables.[35] Y is assumed to be one of the classes $(G+1)$ indicated by $0,1,\ldots,G$. The recoded response matrix is then defined as an $n \times (G+1)$ matrix with:

$$y^*_{i,(g+1)} = I(y_i = g), \tag{4}$$

where $i = 1,\ldots,n$; $g = 0,1,\ldots,G$, and I is an indicator function of event (A). After constructing latent components, the final step required in SPLS-DA is to fit a classifier since the number of latent components (K) is generally smaller than n. For this purpose, linear classifiers such as linear discriminant analysis (LDA) are commonly utilized.[35]

2.4.4 *Optimizing PLS-DA and SPLS-DA*

To determine the number of components for PLS-DA, 10-fold cross validation (CV) was implemented.[42] Each component was systematically added to the PLS-DA model and the cross validated error was then calculated. The process was repeated on the training data until the addition of further components did not improve the significance of the PLS-DA model.[20] In the case of SPLS-DA, there are only two key tuning parameters that require optimization for ideal model performance.[35,46] These include the number of latent components "k" and a sparsity thresholding parameter "eta" that can be optimized using CV. While "k" largely depends on the number of variables and sample size it has been recommended to search for components between 1 and 10 with a thresholding parameter ranging between 0 and 1.[46] The most optimal latent component, therefore, retains the most effective wavebands, whereas other non-important bands would have a probability of zero. The optimized SPLS-DA model developed was then used to classify the test dataset. PLS-DA and SPLS-DA model optimization, VIP calculations and classification was done using the R statistical software package.[40]

2.4.5 *Classification accuracy assessments*

The dataset ($n = 320$) was divided into training (70%; $n = 224$) and validation data (30%; $n = 96$). Confusion matrices were calculated based on classification results conditioned on the validation dataset. The entire process was repeated 100 times to account for the variation in classification accuracy due to differing compositions of training and validation samples.[22,47]

The quantity and allocation disagreement was then used to measure the disagreement within the error matrix as suggested by Pontius and Millones,[48] who criticize the utility of kappa analysis. The quantity disagreement quantifies the amount of tree samples in the training data that differs from the quantity of samples of the same tree species in the test data while the allocation disagreement measures the amount of tree samples of a particular species in the training dataset that were allocated to different locations of the same species in the test dataset. For the purpose of this study, the quantity and allocation disagreement were combined and the total disagreement of the error matrix reported.[48] Additionally, individual class accuracies are reported by the user's and producer's accuracies. The former is calculated by dividing the number of correctly classified species by the total number of species that were classified in that particular class and is represented by the row total in the confusion matrix. Producer's accuracy is computed by dividing the number of correctly classified species in each class by the number of training data used for that particular class and is expressed by the column total in the confusion matrix.[47]

2.5 *Results*

2.5.1 *PLS-DA Optimization*

Figure 3 illustrates a significant decrease in the CV error from the first component (59.63%) to using 10 components which yields a CV error of 17.08%. The lowest error was produced by using five components (11.60%), with the model stabilizing when using nine components to produce a constant CV error (17.08%). The five latent components were used to develop the PLS-DA model and VIP scores for individual bands were then calculated.

2.5.2 *PLS-DA variable importance using VIP*

Figure 4 shows the waveband importance as determined by the VIP method. The VIP method placed importance on bands located within the visible (393 to 700 nm) region of the electromagnetic spectrum. A total of 80 bands obtained VIP scores of >1 and were located within the blue (393 to 500 nm), green (521 to 560 nm), and red (676 to 700 nm) regions. More specifically, 49 bands were considered important in the blue region, 19 in the green, and 12 in the red portion of the spectrum.

Results indicate that utilizing the VIP bands ($n = 80$) produced an overall accuracy of 71.88% and a total disagreement of 28. Accuracies for individual species user's and producer's

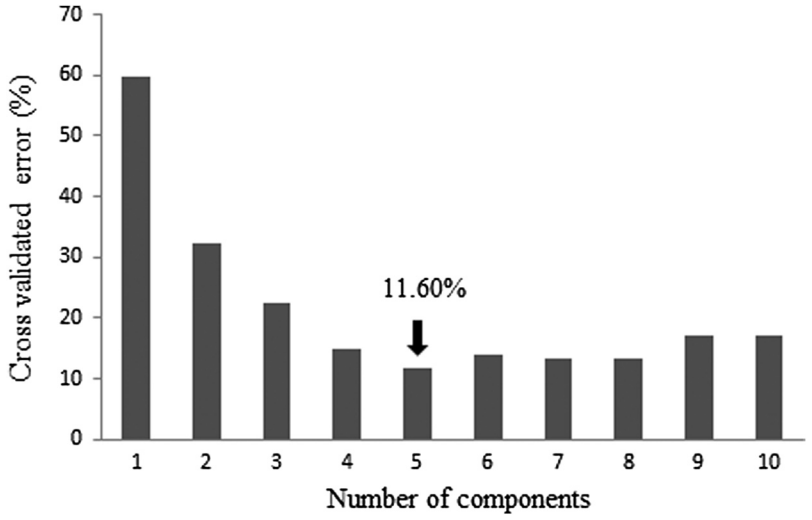

Fig. 3 Assessing the discriminatory power of partial least squares discriminant analysis (PLS-DA) components using all AISA Eagle bands ($n = 272$). 10-fold cross validation was used to determine the lowest error rate conditioned on the training dataset. The optimal component with the lowest error is indicated by the black arrow.

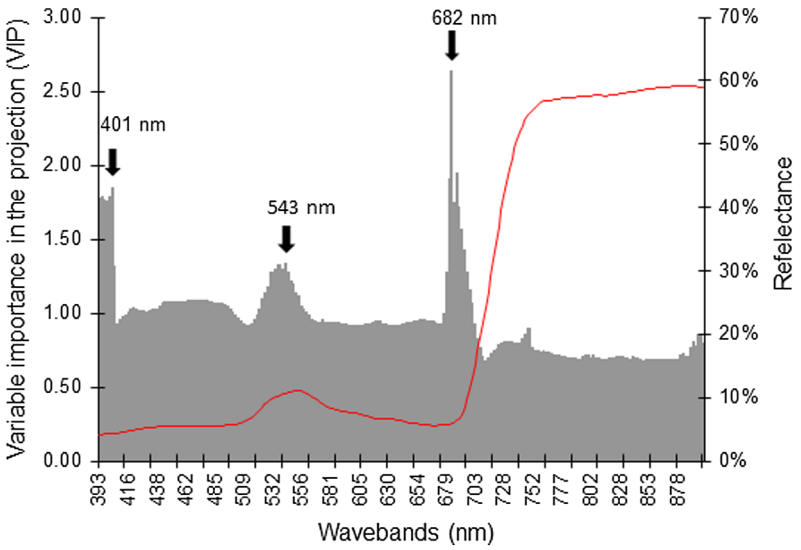

Fig. 4 Waveband importance for all AISA Eagle bands measured by the variable importance in the projection (VIP) method. The red line represents a typical vegetation reflectance curve. The important wavebands are those with scores greater than one and are indicated by the black arrows.

accuracies ranged from 58% to 83% (Table 2). In comparison, using all the AISA Eagle bands ($n = 272$) produced a lower classification accuracy of 68.75% with user's and producer's accuracies ranging between 50% and 79%. For comparison purposes, LDA was used to classify the AISA dataset using the VIP bands. The LDA results revealed an overall classification accuracy of 66.42% with user's and producer's accuracies ranging between 50% and 77%.

2.5.3 *SPLS-DA model optimization*

Figure 5 indicates the significance of each SPLS-DA latent component. The first component yielded a CV error of 40.05% which was later reduced to 13.33% by using 10 latent components. The most significant component (k), however, was achieved by using eight latent components with an "eta" of 0.9 and produced the lowest CV error rate of 10.36%. The model eventually

Table 2 Summed confusion matrix based on the PLS-DA classification algorithm and wavebands selected by the VIP (wavebands = 80). The values in bold indicate the number of correctly classified samples.

	P. elliotii	*P. patula*	*P. taeda*	*S. mauritianum*	Row total	User's Accuracy (%)
P. elliotii	**1800**	100	200	100	2200	82
P. patula	200	**2000**	100	300	2600	77
P. taeda	100	0	**1700**	600	2400	71
S. mauritianum	300	300	400	**1400**	2400	58
Column total	2400	2400	2400	2400	**9600**	
Producer's accuracy (%)	75	83	71	58		
Overall accuracy = 71.88%						
Allocation disagreement = 26						
Quantity disagreement = 2						

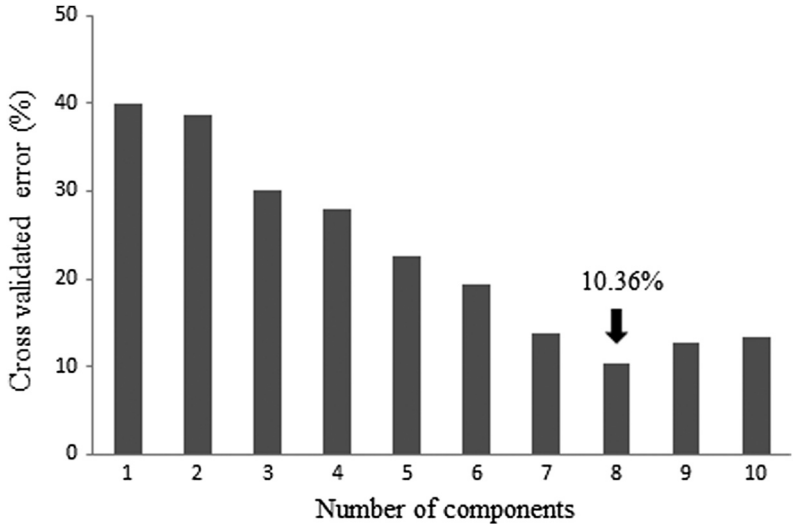

Fig. 5 Assessing the discriminatory power of SPLS-DA components using all AISA Eagle hyperspectral bands ($n = 272$). 10-fold cross validation was used to determine the most significant component conditioned on the training dataset. The optimal component is indicated by the black arrow.

Table 3 Summed confusion matrix based on the SPLS-DA classification algorithm and the Airborne Imaging Spectrometer for Applications (AISA) Eagle hyperspectral dataset. The values in bold indicate the number of correctly classified samples.

	P. elliotii	*P. patula*	*P. taeda*	*S. mauritianum*	Row Total	User's Accuracy (%)
P. elliotii	**2100**	100	100	100	2400	88
P. patula	0	**2200**	100	300	2600	85
P. taeda	0	0	**1800**	400	2200	82
S. mauritianum	300	100	400	**1600**	2400	67
Column total	2400	2400	2400	2400	**9600**	
Producer's accuracy (%)	88	92	75	67		
Overall accuracy = 80.21%						
Allocation disagreement = 18						
Quantity disagreement = 2						

stabilized at a constant value of 13.33%. The eight latent components were then used to develop the SPLS-DA model.

Test dataset results indicate that using the AISA Eagle hyperspectral bands with eight SPLS-DA components produced an overall accuracy of 80.21% and a total disagreement of 20. User's and producer's accuracies for each species ranged from 67% to 92% (Table 3).

Figure 6 displays the variation in classification accuracy produced by SPLS-DA when using 100 iterations for splitting the training and validation dataset. Classification means were found to be >80% with a standard deviation of 2.87.

Figure 7 shows the most effective wavebands selected by the SPLS-DA algorithm and that were automatically used in the classification process. The method placed importance on bands located within the visible (415 to 694 nm) region of the electromagnetic spectrum. The SPLS-DA model used a total of 55 bands which best explained the discrimination among the tree species and were located in intervals within the blue (415 to 436 nm; 457 to 483 nm), green (515 to 521 nm; 530 to 565 nm), and red regions (674 to 694 nm), respectively.

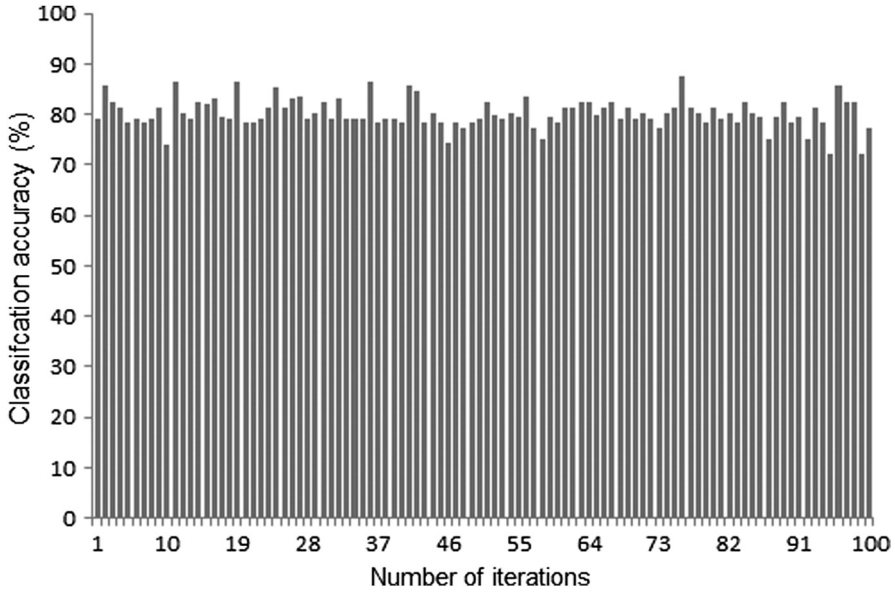

Fig. 6 The variation in classification accuracy produced by SPLS-DA when using 100 iterations for splitting the training and validation dataset.

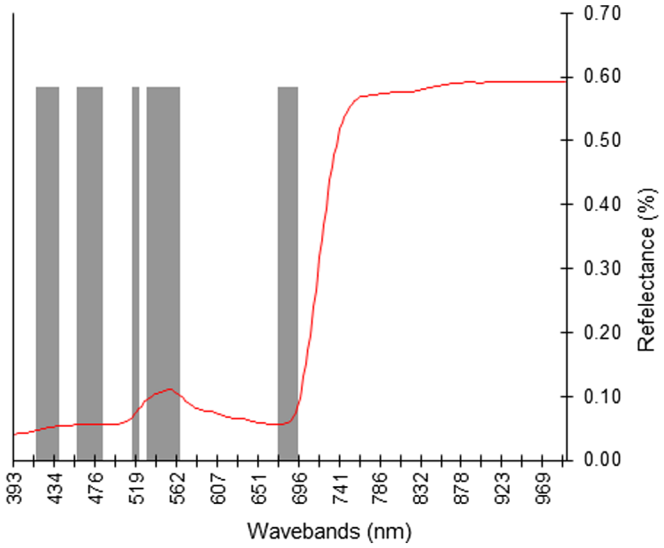

Fig. 7 Location of the most effective wavebands used in the classification of the tree species using the SPLS-DA algorithm. The important wavebands are those with scores of greater than zero.

In total, 24 bands were considered important in the blue, 21 in the green, and 10 in the red portion of the spectrum.

In comparison to the SPLS-DA results, utilizing these bands in LDA revealed an overall accuracy of 72.9% with user's and producer's accuracies between 56 and 80%. Since the SPLS-DA classification produced the best results, a classified tree species map was produced using a subset of the AISA Eagle imagery (Fig. 8). The map is comparable to that of the AISA Eagle airborne hyperspectral image, with *P. patula* being the dominant tree species. *P. taeda* and *S. mauratanum* have the most confusion with each other and are the least correctly mapped species, respectively.

2.6 *Discussion*

One of the most prominent challenges in discriminating forest species using remotely sensed data is to use the subtle spectral variations between species to classify them correctly. This study

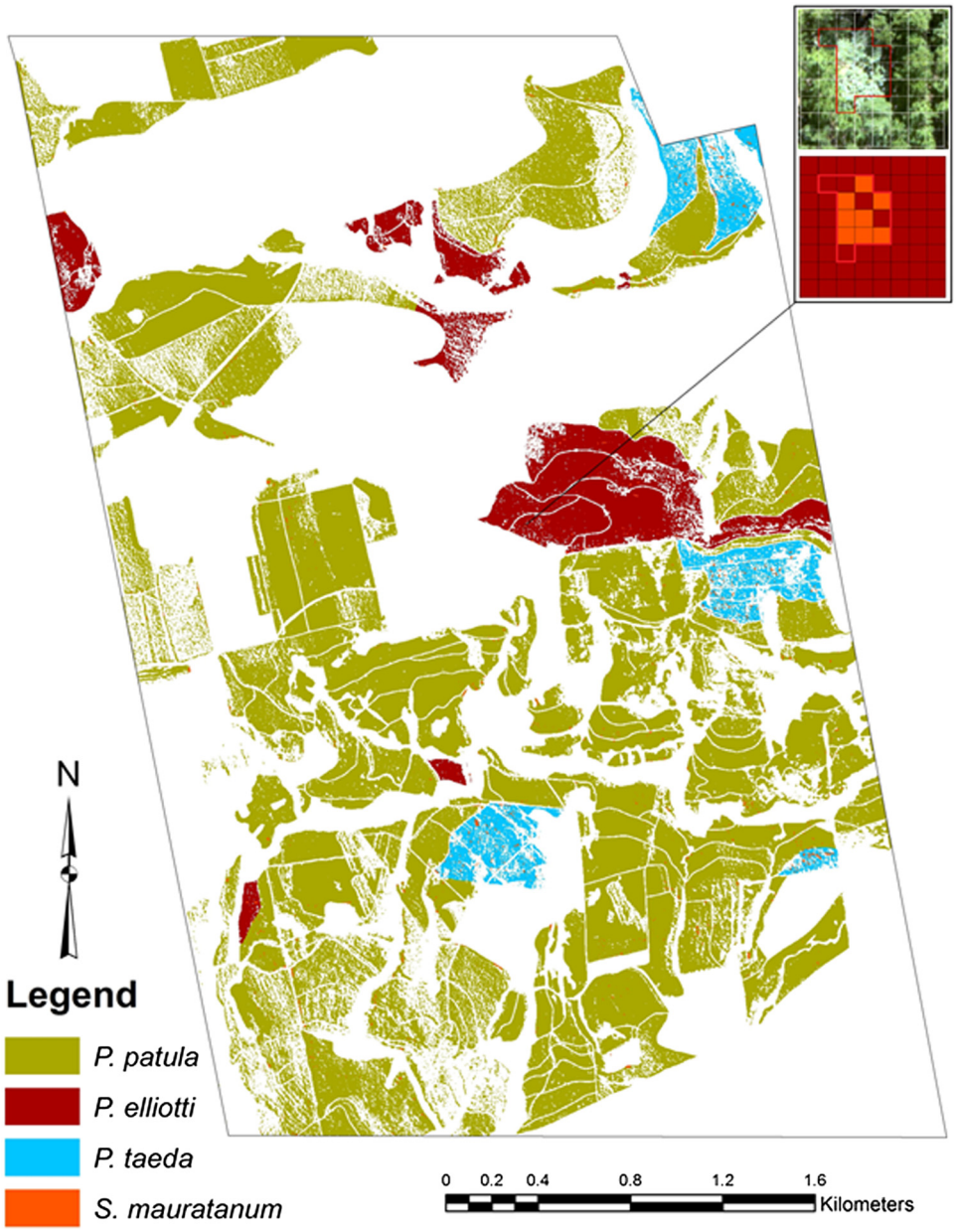

Fig. 8 Tree species classification map produced by the SPLS-DA algorithm using the AISA Eagle hyperspectral image.

presents valuable evidence for the application of utilizing hyperspectral remote sensing to classify commercial tree species in KwaZulu-Natal, South Africa. Results show the capability of the AISA Eagle image data in effectively dealing with the spectral similarity existing between the closely related *Pinus* species considered in this study. In addition, the utility of the SPLS-DA algorithm proved more effective compared to PLS-DA and VIP while providing an accurate framework for executing simultaneous variable selection and dimension reduction of high-dimensional datasets, which is necessary if we are to fully exploit hyperspectral image data in classifying commercial forest tree species.

2.6.1 *PLS-DA and SPLS-DA classification using AISA Eagle hyperspectral data*

The generation of fewer initial components within a PLS-DA model is critical in reducing the risk of overfitting and removing the low order components which do not contribute toward the models' performance.[23,24] Subsequently, the results indicate that the systematic addition of latent

components to the PLS-DA models significantly improves model performance based on the CV error. Using five optimal latent components in PLS-DA in conjunction with VIP selected bands produced an overall classification accuracy of 71.88%. When utilizing eight optimal components, SPLS-DA produced an 8.33% improvement in the overall classification accuracies. This classification result is comparable to that of previous forest species discrimination studies using hyperspectral datasets. [6,11,13–16,49,50] However, this classification result has been achieved using a low number of species (i.e., four species) when compared to the number of species considered in Refs. 9–11, 49, and 50. Alternatively, other feature selection and extraction techniques have been applied for the classification of tree species using hyperspectral data. These include stepwise LDA,[8,51,52] out-of-bag and best-first search method,[53] MNF transformations,[21,54] sequential forward floating selection,[19,55] GA, SVM wrapper, and sparse generalized PLS selection.[22]

When observing the individual classification accuracies of each tree species considered in this study, SPLS-DA produced higher individual class accuracies (67 to 92%) compared to the accuracies produced by PLS-DA and the VIP selected bands (58 to 83%). Furthermore, there was an improvement in the user's and producer's accuracies for *P. elliotii* and *P. patula* when compared to the accuracies obtained in a previous study[20] that discriminated *Pinus*, *Eucalyptus* and *Acacia* tree species. This result confirms the findings of Wolter et al.[24] and Wolter et al.,[43] who suggest that separate PLS models could be constructed to improve individual class accuracies. As a result, individual PLS models use the spectral information to explain the variance for species within a genus (for example, *P. elliotii* and *P. patula*) such as in this study as opposed to species from different genera (for example, *E. grandis* and *P. patula*). However, most of the confusion occurred with *Pinus* trees and bugweed (*S. mauritianum*). The results show that bugweed were the least correctly classified class and that the majority of the confusion occurred between bugweed and *P. taeda*. Nonetheless, the classification accuracies obtained in this study for each tree species may be influenced by a variety of other factors linked to the spectral variability within the canopy of each forest stand. For example, the variation in reflectance within forest species primarily occurs as a result of canopy shadowing, differences in light absorption, and spectral scattering of wavelengths.[14] Additionally, researchers have noted that the classification of tree species may also be affected by the overall structure of the forest canopy, sensor optical properties, and the effects of the nonphotosynthetic material.[13]

2.6.2 *PLS-DA and SPLS-DA variable importance*

While both models (PLS-DA and SPLS-DA) performed classification successfully, the exclusive variable selection approaches provided valuable insight on the most effective wavebands when classifying the tree species. The VIP method successfully reduced the large number of hyperspectral bands to 80 important wavebands to produce a reasonable level of accuracy (71.88%) compared to when all the bands ($n = 272$) were utilized (68.75%). SPLS-DA, however, executed variable selection automatically to include only important variables within the PLS classification and successfully reduced the hyperspectral bands to 55 relevant wavebands to produce the best classification accuracies. Nonetheless, given the spectral range of the AISA Eagle sensor, 80 and 55 bands are still a high number when compared to other forest species classification studies and could be a potential drawback of the methodology. For example, Clark et al.[13] applied 30 bands at crown level and obtained a high accuracy of 86%. Using 30 AISA Bands, Dalponte et al.[19] obtained a kappa accuracy of 0.70. Similarly, Jones et al.[8] applied 40 AISA bands and mapped most tree species with accuracies ranging from >60% to 90%. Liu et al.[54] used 26 spectral bands and obtained 80.67% classification accuracy for mapping temperate forest species. However, their results were based on a MNF transformation. Additionally, Jones et al.[8] and Clark et al.[13] investigated larger spectral ranges beyond the visible and near infrared regions that were used in this study.

When comparing the important bands selected by PLS-DA using VIP and those inherently selected by SPLS-DA, results show that bands in the visible region of the spectrum (393 to 700 nm) were most effective in the classification. More specifically, PLS-DA and VIP placed importance on 49 bands in the blue (393 to 500 nm), 19 bands within the green (421 to 560 nm), and 12 bands in the red (676 to 700 nm). In comparison, SPLS-DA selected fewer bands, also

within the visible portion and along narrower wavelength intervals than the band ranges of VIP. For instance, SPLS-DA placed importance on 24 blue wavebands located between 415 to 436 nm and 457 to 483 nm, 21 green wavebands between 515 to 521 nm and 530 to 565 nm and 10 bands within the red at 674 to 694 nm. While wavebands within the blue portion of the spectrum are recognized for classifying tree species, those located within the green region confirm the importance of the green reflectance peak around 550 nm.[20,56] The significance of the red region is also recognized for the discrimination of tree species[20] and is a result of the red portion being sensitive to plant pigment concentrations within the leaf tissue.[57–59] Overall, the importance of visible wavebands selected in this study for the classification of tree species is comparable to that of other studies who also recognize the significance of wavebands in the visible for the classification of tree species using hyperspectral data.[13–15,20,60]

The operational limitation of this study is, however, highlighted by the procurement of relatively homogenous pixels of each tree species to exploit the subtle variations existing between them. Nonetheless, the proposed methodology of this study should be tested in areas that have a heterogeneous composition of the selected tree species and could be expanded to species that are native to South Africa. This would require some variation in the methodology due to the denser spatial configuration of native trees. Future studies should also consider the application of stability measures or iterative bootstrap classification approaches.[21,22] Such approaches would capture the variation created by changing the composition of training and validation datasets to improve the reliability and quality of classification results. The robustness of the waveband regions selected by the SPLS-DA technique should also be investigated using other commercially available sensors for classifying tree species. For example, spectral regions of importance included narrow band ranges in the blue (415–483 nm), green (515–565 nm), and red (674–694 nm) portions of the spectrum. This provides an opportunity to exploit the new generation of multispectral sensors (such as WorldView-2), with fine spatial resolution and spectral resolutions, to discriminate among tree species in South Africa.

2.7 *Conclusion*

This study has shown the capability of utilizing SPLS-DA for the combined purpose of variable selection and dimension reduction of high-dimensional data for the classification of commercial tree species. SPLS-DA produced an overall accuracy of 80.21% and a total disagreement value of 20. Accuracies for the individual tree species ranged between 67% and 92% with the most effective wavebands located in the visible portion (415 to 694 nm) of the spectrum. Overall, the utility of SPLS-DA provided an accurate and computationally efficient methodology for selecting important variables within the PLS framework, while performing simultaneous classification for the successful discrimination of commercial tree species.

Acknowledgments

The authors would like to acknowledge the support from the Applied Centre for Climate and Earth Systems Science (ACCESS) and Sappi forest-SA for the successful completion of this paper.

References

1. M. Wulder, "Optical remote-sensing techniques for the assessment of forest inventory and biophysical parameters," *Prog. Phys. Geogr.* **22**(4), 449–476 (1998).
2. R. Ismail, O. Mutanga, and L. Kumar, "Modeling the potential distribution of pine forests susceptible to sirex noctilio infestations in Mpumalanga, South Africa," *Trans. GIS* **14**(5), 709–726 (2010).
3. K. Little, J. Kritzinger, and M. Maxfield, *Some Principles of Vegetation Management Explained*, University of KwaZulu-Natal, Pietermaritzburg (1997).
4. J. T. Atkinson, R. Ismail, and M. Robertson, "Mapping bugweed (solanum mauritianum) infestations in pinus patula plantations using hyperspectral imagery and support vector machines," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(1), 17–28 (2014).

5. A. Morris and R. Pallett, *Pines*, in D. L. Owen, ed., South African Institute of Forestry, Pretoria (2000).
6. J. A. N. v. Aardt and R. H. Wynne, "Examining pine spectral separability using hyperspectral data from an airborne sensor: an extension of field-based results," *Int. J. Remote Sens.* **28**(2), 431–6 (2007).
7. T. Dube et al., "Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution spaceborne multispectral rapideye sensor and advanced machine learning algorithms," *Sensors* **14**(8), 15348–15370 (2014).
8. T. G. Jones, N. C. Coops, and T. Sharma, "Assessing the utility of airborne hyperspectral and lidar data for species distribution mapping in the coastal Pacific Northwest, Canada," *Remote Sens. Environ.* **114**(12), 2841–2852 (2010).
9. S. E. Sesnie et al., "The multispectral separability of costa rican rainforest types with support vector machines and random forest decision trees," *Int. J. Remote Sens.* **31**(11), 2885–2909 (2010).
10. M. Vohland et al., "Remote sensing techniques for forest parameter assessment: multispectral classification and linear spectral mixture analysis," *Silva Fennica* **41**(3), 441–456 (2007).
11. M. Cochrane, "Using vegetation reflectance variability for species level classification of hyperspectral data," *Int. J. Remote Sens.* **21**(10), 2075–2087 (2000).
12. S. L. Ustin et al., "Using imaging spectroscopy to study ecosystem processes and properties," *BioScience* **54**(6), 523–534 (2004).
13. M. L. Clark, D. A. Roberts, and D. B. Clark, "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales," *Remote Sens. Environ.* **96**(3), 375–398 (2005).
14. R. Lucas et al., "Classification of australian forest communities using aerial photography, casi and hymap data," *Remote Sens. Environ.* **112**(5),2088–103 (2008).
15. N. Goodwin, R. Turner, and R. Merton, "Classifying eucalyptus forests with high spatial and spectral resolution imagery: an investigation of individual species and vegetation communities," *Aust. J. Bot.* **53**(4), 337–345 (2005).
16. M. E. Martin et al., "Determining forest species composition using high spectral resolution remote sensing data," *Remote Sens. Environ.* **65**(3), 249–254 (1998).
17. J. van Aardt and M. Norris-Rogers, "Spectral-age interactions in managed, even-aged eucalyptus plantations: application of discriminant analysis and classification and regression trees approaches to hyperspectral data," *Int. J. Remote Sens.* **29**(6), 1841–1845 (2008).
18. K. N. Youngentob et al., "Mapping two eucalyptus subgenera using multiple endmember spectral mixture analysis and continuum-removed imaging spectrometry data," *Remote Sens. Environ.* **115**(5), 1115–1128 (2011).
19. M. Dalponte, L. Bruzzone, and D. Gianelle, eds., "Tree species classification in the southern Alps with very high geometrical resolution multispectral and hyperspectral data," in *3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, IEEE, Lisbon (2011).
20. K. Y. Peerbhay, O. Mutanga, and R. Ismail, "Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (pls-da) in Kwazulu–Natal, South Africa," *ISPRS J. Photogramm. Remote Sens.* **79**, 19–28 (2013).
21. A. Ghosh et al., "A framework for mapping tree species combining hyperspectral and lidar data: role of selected classifiers and sensor across three spatial scales," *Int. J. Appl. Earth Obs. Geoinf.* **26**, 49–63 (2014).
22. F. Fassnacht et al., "Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(6), 2547–2561 (2014).
23. L. Li, S. L. Ustin, and D. Riano, "Retrieval of fresh leaf fuel moisture content using genetic algorithm partial least squares (GA-PLS) modeling," *IEEE Geosci. Remote Sens. Lett.* **4**(2), 216–220 (2007).
24. P. T. Wolter, P. A. Townsend, and B. R. Sturtevant, "Estimation of forest structural parameters using 5 and 10 meter spot-5 satellite data," *Remote Sens. Environ* **113**(9), 2019–2036 (2009).

25. B. Panneton et al., "Discrimination of corn from monocotyledonous weeds with ultraviolet (UV) induced fluorescence," *Appl. Spectrosc.* **65**(1), 10–9 (2011).
26. M. Pérez-Enciso and M. Tenenhaus, "Prediction of clinical outcome with microarray data: A partial least squares discriminant analysis (PLS-DA) approach," *Hum. Genet.* **112**(5–6), 581–592 (2003).
27. K. A. Lê Cao and C. Le Gall, "Integration and variable selection of 'omics' data sets with PLS: A survey," *Journal de la Société Française de Statistique* **152**(2), 77–96 (2011).
28. A. J. Hobro et al., "Differentiation of walnut wood species and steam treatment using ATR-FTIR and partial least squares discriminant analysis (PLS-DA)," *Anal. Bioanal. Chem.* **398** (6), 2713–2722 (2010).
29. R. Castillo et al., "Multivariate strategies for classification of eucalyptus globulus genotypes using carbohydrates content and nir spectra for evaluation of their cold resistance," *J. Chemom.* **22**(3–4), 268–280 (2008).
30. O. Galtier et al., "Comparison of PLS1-DA, pls2-da and simca for classification by origin of crude petroleum oils by mir and virgin olive oils by NIR for different spectral regions," *Vib. Spectrosc.* **55**(1), 132–140 (2011).
31. M. Kalacska et al., "Hyperspectral discrimination of tropical dry forest lianas and trees: comparative data reduction approaches at the leaf and canopy levels," *Remote Sens. Environ.* **109**(4), 406–415 (2007).
32. B. H. Menze et al., "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinf.* **10**(1), 213–229 (2009).
33. I. G. Chong and C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemom. Intell. Lab. Syst.* **78**(1), 103–112 (2005).
34. L. Cécillon et al., "Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts," *Soil Biol. Biochem.* **40**(7), 1975–1979 (2008).
35. D. Chung and S. Keles, "Sparse partial least squares classification for high dimensional data," *Stat. Appl. Genetics Mol. Biol.* **9**(1), 1554–6115 (2010).
36. M. Dye, O. Mutanga, and R. Ismail, "Examining the utility of random forest and AISA Eagle hyperspectral image data to predict pinus patula age in Kwazulu-Natal, South Africa," *Geocarto Int.* **26**(4), 275–289 (2011).
37. L. Mucina and M. C. Rutherford, *The Vegetation of South Africa*, South African National Biodiversity Institute, Lesotho and Swaziland (2006).
38. D. Roberts, Y. Yamaguchi, and R. Lyon, "Comparison of various techniques for calibration of AIS data," NASA STI/Recon Technical Report No. 8712970 (1986).
39. ENVI, ENVI 4.7: environment for visualizing images. Exelis Visual Information Solutions. ITT Industries, Colorado (2009).
40. R Development Core Team, "R: A language and environment for statistical computing," 2012, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project (1 November 2012).
41. S. W. Lindström et al., "The importance of balanced data sets for partial least squares discriminant analysis: classification problems using hyperspectral imaging data," *J. Near Infrared Spectrosc.* **19**(4), 233–241 (2011).
42. S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemom. Intell. Lab. Syst.* **58**(2), 109–130 (2001).
43. P. T. Wolter et al., "Remote sensing of the distribution and abundance of host species for spruce budworm in northern Minnesota and Ontario," *Remote Sens. Environ.* **112**(10), 3971–3982 (2008).
44. J. Read et al., "Narrow-waveband reflectance ratios for remote estimation of nitrogen status in cotton," *J. Environ. Qual.* **31**(5), 1442–1452 (2002).
45. C. Gomez, P. Lagacherie, and G. Coulouma, "Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements," *Geoderma* **148**(2), 141–148 (2008).
46. H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. R Stat. Soc. B* **72**(1), 3–25 (2010).

47. R. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publisher, Boca Raton, Florida (1999).
48. R. G. Pontius, Jr. and M. Millones, "Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.* **32**(15), 4407–4429 (2011).
49. P. Gong and B. Yu, "Conifer species recognition: effects of data transformation," *Int. J. Remote Sens.* **22**(17), 3471–3481 (2001).
50. K. L. Castro-Esau, G. Sánchez-Azofeifa, and T. Caelli, "Discrimination of lianas and trees with leaf-level hyperspectral data," *Remote Sens. Environ.* **90**(3), 353–372 (2004).
51. B. Datt, Ed., "Recognition of eucalyptus forest species using hyperspectral reflectance data," in *Geoscience and Remote Sensing Symposium, 2000 Proc. (IGARSS 2000)*, IEEE, Honolulu (2000).
52. T. Fung et al., "Band selection using hyperspectral data of subtropical tree species," *Geocarto Int.* **18**(4), 3–11 (2003).
53. J. C.-W. Chan and D. Paelinckx, "Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sens. Environ.* **112**(6), 2999–3011 (2008).
54. L. Liu et al., edis., "Fusion of airborne hyperspectral and lidar data for tree species classification in the temperate forest of northeast China. Geoinformatics," in *19th International Conf. IEEE*, IEEE, Shanghai (2011).
55. M. Dalponte et al., "The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas," *Remote Sens. Environ.* **113**(11), 2345–2355 (2009).
56. N. C. Coops et al., "Chlorophyll content in eucalypt vegetation at the leaf and canopy scales as derived from high resolution spectral data," *Tree Physiol.* **23**(1), 23–31 (2003).
57. G. A. Blackburn, "Hyperspectral remote sensing of plant pigments," *J. Exp. Bot.* **58**(4), 855–867 (2007).
58. J. Gamon and J. Surfus, "Assessing leaf pigment content and activity with a reflectometer," *New Phytologist* **143**(1), 105–117 (1999).
59. A. A. Gitelson, M. N. Merzlyak, and O. B. Chivkunova, "Optical properties and nondestructive estimation of anthocyanin content in plant leaves," *Photochem. Photobiol.* **74**(1), 38–45 (2001).
60. N. C. Coops et al., "Assessment of crown condition in eucalypt vegetation by remotely sensed optical indices," *J. Environ. Qual.* **33**(3), 956–964 (2004).

Kabir Yunus Peerbhay is a PhD candidate at the University of KwaZulu-Natal specializing in remote sensing. He received the MSc degree (cum laude) in applied environmental science in 2011. His research is focused on using machine learning algorithms to maximize the benefit of remotely sensed data for forest inventory practices and weed detection applications.

Onisimo Mutanga is a professor and academic leader in research at the University of KwaZulu-Natal, South Africa. His research is focused on ecological assessment and monitoring with special emphasis on vegetation pattern analysis using GIS and remote sensing. He is currently expanding this domain into mapping vegetation species, wetland mapping, disease detection in plantation forests and agricultural crops as well as quantifying forest fragmentation and its impact on biodiversity and ecosystem condition.

Riyad Ismail received the MSc degree in GIS (cum laude) and PhD degree in remote sensing from the University of KwaZulu-Natal, South Africa. He has over 15 years of experience in implementing spatial technologies (GIS, GPS and remote sensing) at commercial, academic and research institutions. He was recently appointed as a senior research associate at the University of KwaZulu-Natal and is currently employed as a principal research officer at Sappi forests.