ISPRS Journal of Photogrammetry and Remote Sensing 69 (2012) 167–179

Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment

L. Naidoo [a,*], M.A. Cho [a], R. Mathieu [a], G. Asner [b]

[a] Council for Scientific and Industrial Research (CSIR), Natural Resources and the Environment, P.O. Box 395, Pretoria 0001, South Africa
[b] Carnegie Institution for Science, Stanford, CA, USA

ARTICLE INFO

Article history:
Received 14 July 2011
Received in revised form 13 January 2012
Accepted 13 March 2012

Keywords:
Savanna tree species
Spectral variability
Tree height
Random Forest
Predictor datasets

ABSTRACT

The accurate classification and mapping of individual trees at species level in the savanna ecosystem can provide numerous benefits for the managerial authorities. Such benefits include the mapping of economically useful tree species, which are a key source of food production and fuel wood for the local communities, and of problematic alien invasive and bush encroaching species, which can threaten the integrity of the environment and livelihoods of the local communities. Species level mapping is particularly challenging in African savannas which are complex, heterogeneous, and open environments with high intra-species spectral variability due to differences in geology, topography, rainfall, herbivory and human impacts within relatively short distances. Savanna vegetation are also highly irregular in canopy and crown shape, height and other structural dimensions with a combination of open grassland patches and dense woody thicket – a stark contrast to the more homogeneous forest vegetation. This study classified eight common savanna tree species in the Greater Kruger National Park region, South Africa, using a combination of hyperspectral and Light Detection and Ranging (LiDAR)-derived structural parameters, in the form of seven predictor datasets, in an automated Random Forest modelling approach. The most important predictors, which were found to play an important role in the different classification models and contributed to the success of the hybrid dataset model when combined, were species tree height; NDVI; the chlorophyll *b* wavelength (466 nm) and a selection of raw, continuum removed and Spectral Angle Mapper (SAM) bands. It was also concluded that the hybrid predictor dataset Random Forest model yielded the highest classification accuracy and prediction success for the eight savanna tree species with an overall classification accuracy of 87.68% and KHAT value of 0.843.

1. Introduction

Numerous studies have readily dealt with the classification of plant functional groups, like the mapping of broadleaf and fine-leaf forest trees (Kooistra et al., 2007) or mangrove types (Yingchun et al., 2006), but fewer studies have intimately tackled the classification and mapping of trees at species level (Hestir et al., 2008; Asner et al., 2008; Sobhan, 2007). This is especially the case in African savannas which are complex, heterogeneous, and open environments with high intra-species spectral variability due to differences in geology (e.g. granite and gabbro), topography, rainfall, herbivory and human impacts (e.g. fire, resource harvesting such as fuel wood or foliage browsing) within relatively short distances (Cho et al., 2009, 2010). Unlike more stable boreal and tropical forests, savannas are highly dynamic and are in a constant state of flux in which cyclical successions between the dominance of woody and grassy vegetation are evident (according to patch dynamics theory in Meyer et al. (2007)). The accurate mapping of individual trees at species level in the savanna ecosystem can provide numerous benefits for the managerial authorities, especially for economically useful trees, which are a key source of food production and fuel wood for the local communities, and problematic alien invasive and bush encroaching species, which can threaten the integrity of the environment and livelihoods of the local communities. The Marula Tree (*Sclerocarya birrea*), for example, plays an important role by providing non-timber forest products (NTFPs); for instance, the local communities in the communal rangelands utilise the Marula fruit for beer brewing in cultural and especially trading activities (Shackleton and Shackleton, 2003). Joubert (2007), on the other hand, described the 'plague' of bush encroaching (e.g.

* Corresponding author. Tel.: +27 12 841 2233; fax: +27 12 841 3909.
E-mail addresses: lnaidoo@csir.co.za (L. Naidoo), mcho@csir.co.za (M.A. Cho), rmathieu@csir.co.za (R. Mathieu), gpa@stanford.edu (G. Asner).

0924-2716/$ - see front matter © 2012 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2012.03.005

Dichrostachys cinerea and *Terminalia sericea)* and alien invasive species (water hyacinths – *Eichhornia crassipes)* in the Kruger National Park.

The classification of tree species falls within the realm of possibility for remote sensing but in order to capture the complex inter- and intra-species spectral variability resulting from genetic patrimony and difference in environmental and physical factors (weather, seasonality, geology and edaphic conditions and natural phenological changes such as deciduous versus evergreen species during the savanna dry season – Hestir et al., 2008; Lees and Ritman, 1991; Tong et al., 2004), the spectral resolution of a sensor must be high with numerous, contiguous bands along with a high spatial resolution in relation to the scale of lateral variation of the tree canopy. These requirements are best met by high resolution hyperspectral sensors. Classification studies from Cho et al. (2010) and Cho et al. (2011) have shed some light on the use of spectral band configurations and particular significant bands of hyperspectral imagery in assisting successful savanna tree species classification. Cho et al. (2010) made use of a band redundancy minimisation procedure, known as the Band Add-On procedure, to select and identify the most useful hyperspectral bands for species discrimination using Spectral Angle Mapper (SAM) classifier. They concluded that a total of 31 bands (which occupied a combination of blue, red edge, near-infrared and chemical spectral bands) out of the original 72 bands were found to be the most spectrally significant. Furthermore, Cho et al. (2011) resampled a hyperspectral dataset to simulate the spectral band configuration of the Worldview-2 sensor (traditional spectral regions of red, green, blue and near-infrared plus yellow and red-edge spectral regions) to classify savanna species and achieved higher classification accuracies than the traditional spectral regions (typically available on multispectral sensors such as SPOT, IKONOS). These accuracy results could be attributed to the inclusion of more vegetation sensitive bands (particularly the yellow and red-edge spectral bands) in the overall analytical approaches outlined in Cho et al. (2011).

Although the use of spectra alone provided acceptable classifications in these studies, it is evident from various structural remote sensing studies (Kim, 2007; Bork and Su, 2007; Geerling et al., 2007; Lucas et al., 2008; Asner et al., 2008) that structural information (especially tree height) plays important roles in assisting or being solely utilised in vegetation cover and tree species level classification and mapping. Bork and Su (2007), for example, integrated LiDAR data in the mapping process by detecting the differences in vegetation height and then implementing vertical height 'thresholds' for the adequate height separation of the different vegetation communities. Geerling et al. (2007) combined image spectroscopy and LiDAR data, by data fusion at the pixel level, to improve the classification of floodplain vegetation types. Hill and Thomson (2005) integrated spaceborne hyperspectral data with airborne LiDAR data in a parcel-based unsupervised classification approach to map homogeneous broadleaf and conifer woodlands. Since savanna vegetation are also highly irregular in canopy and crown shape, height and other structural dimensions with a combination of open grassland patches and dense woody thicket (a stark contrast to the more homogeneous forest vegetation), these structural vegetation parameters should not be ignored. Furthermore, structural variables may help to reduce spectral confusion, for instance when particular tree species possesses spectral properties similar to the underlying grass layer (as was the case for *Acacia nigrescens* in Cho et al., 2011). The potential importance of the simple structural parameter, the tree height, is illustrated in the box plot in Fig. 1, which shows the height statistics of various savanna tree species (obtained from field data). The tree height data show three classes of trees; the short (*Pterocarpus rotundifolius*, *Acacia gerrardii/D. cinerea* and *T. sericea*), medium (*Combretum* species and *Spirostachys africana)* and tall (*A. nigrescens*, *S.* birrea and *Berchemia discolor)* trees. Such distinguishable differences in the height ranges of the different species could enhance species classification. *Combretum* species and *S.* africana however, showed high variability in height, ranging from short to tall trees and this could reduce accuracy of classification of these species.

An integrated approach, which has the ability to combine structural and spectral variables into an automated classification procedure, may help to overcome the high intra-species spectral variability of savanna tree species (Cho et al., 2009, 2010), while taking advantage of known differences in structure between species. Further advantages of integrating different sensor datasets, particularly hyperspectral data, with LiDAR were covered in detail by Lucas et al. (2008). Traditional parametric classification methods, e.g. Maximum Likelihood (MAXLIKE), are affected by the 'Hughes Phenomenon' which arises in high dimensionality data when the training dataset size is not large enough to adequately estimate the covariance matrices (Cortijo and de la Blanca, 1996). In hyperspectral classification studies, acquiring the sufficient number of training data that exceeds the total number of spectral bands, required for the MAXLIKE classifier, is an impractical task especially in highly, spectrally variable environments. For complex savanna tree species classification, the necessary requirements could be met by the implementation of non-parametric classification methods such as a decision tree approach, with the most commonly used approach being the Classification and Regression Trees (CART).

CART is a non-parametric model which constructs important rule sets by iteratively subsetting the target dataset, according to defined thresholds of various important explanatory variables, into smaller homogeneous groups (Ismail et al., 2010; Prasad et al., 2006). This single decision tree approach recursively 'mines' and groups the target data until an end node for classification or a defined class is reached. CART classification approaches have proven successful in species level classification and mapping of tropical forest canopies (Affendi et al., 2009) and invasive aquatic vegetation (Hestir et al., 2008). However, according to Ismail et al. (2010) and Prasad et al. (2006), CART models are sensitive to small changes in the training dataset and have been identified as being occasionally unstable as they are prone to data overfitting. Other non-parametric classifiers such as K-nearest neighbour (KNN), Support Vector Machines (SVM) and artificial neural networks (ANN) were also not considered. ANN and SVM techniques are too computer intensive and time consuming due to the level of complexity and customization that is required, while it is difficult to determine the optimal *K* value for the KNN classifier.

The emergence of the Random Forest (RF) approach was seen as an improvement over the CART approach as concepts such as multiple decision trees, bootstrap aggregation (bagging) and internal cross-validation were introduced which led to improved results, ease of use and overcoming of the issue of over-fitting (Grossmann et al., 2010; Ismail et al., 2010). RF constructs hundreds of decision tree models (hence 'forest') using randomised subsets (hence 'random') of target data and explanatory variables to build each tree (Grossmann et al., 2010). These multiple classification trees are then voted upon by plurality, to ascertain the correct classification (Lawrence et al., 2006; Ismail et al., 2010). The RF approach has been successfully implemented in the mapping of invasive plant species (Lawrence et al., 2006) or forested ecological systems (Grossmann et al., 2010), and the modelling of the potential distribution of pine forest susceptible to wasp infestation (Ismail et al., 2010). In a predictive vegetation mapping study by Prasad et al. (2006), RF outperformed other classification and regression tree techniques such as CART, MARS (Multivariate Adaptive Regression

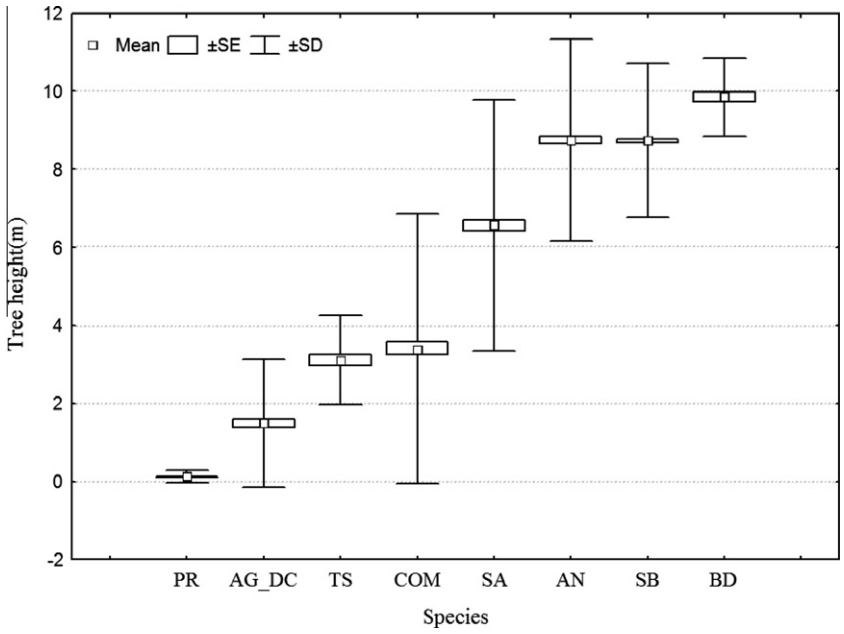

Fig. 1. Box plot of the eight common savanna tree species' height distribution obtained from a field sample collected in the Bushbuckridge and the Sabi Sands Wildtuin regions in the southern part of the Greater Kruger National Park. SE = standard error and SD = standard deviation. AG_DC = *Acacia gerrardii/Dichrostachys cinerea*, AN = *Acacia nigrescens*, BD = *Berchemia discolor*, COM = *Combretum* species, PR = *Pterocarpus rotundifolius*, SA = *Spirostachys africana*, SB = *Sclerocarya birrea* and TS = *Terminalia sericea*.

Splines) and other bagging trees (BT). RF was thus considered as the most applicable approach for the classification of various savanna tree species in such a heterogeneous environment.

This study aimed to classify eight common savanna tree species in the Greater Kruger National Park region, South Africa, using spectral and structural remote sensing information in an automated Random Forest modelling approach. These species were *A.* gerrardii/*D.* cinerea (AG/DC), *A.* nigrescens (AN), *B.* discolor (BD), *Combretum* species (COM), *P.* rotundifolius (PR), *S.* africana (SA), *S.* birrea (SB) and *T.* sericea (TS). Based on the assumption that tree height is an important structural parameter, the use of which might improve the classification of savanna tree species, the objective of this study was to investigate the influence of using tree height data as well as spectral data on accuracy of classifying savanna tree species. The research was made possible by the availability of an integrated airborne hyperspectral and LiDAR sensor dataset collected by the Carnegie Airborne Observatory (CAO). For this investigation, seven predictor datasets – consisting of spectral, structural and a combination of spectral and structural information at the species level – were subjected to Random Forest modelling and compared. The following scientific questions were posed for investigation.

- Which particular explanatory variable (predictor) or suite of explanatory variables, used in the Random Forest model, contributed the most towards the success of the savanna tree species classification?
- Which Random Forest model yielded the highest accuracy results for classifying the eight common savanna tree species when utilising spectral, structural and a combination of spectral and structural predictor datasets in the modelling process?

2. Materials and methodology

2.1. Study area

The study area is located within the broad savanna biome, which occupies over a third of the area of Southern Africa, and is distinguished by the coexistence of a grassy ground layer and a prominent upper layer of woody plants (Rutherford and Westfall, 1986). Regionally, savannas have a long dry winter and a wet summer with an annual precipitation varying between 235 and 1000 mm. This rainfall range, together with grazing pressures and fire, govern the vegetation structure present in this biome. Various vegetation types; particularly Clay Thornbush, Mixed Bushveld and Sweet and Sour Lowveld Bushveld; are supported in this general savanna environment (Rutherford and Westfall, 1986).

The study area under investigation (Fig. 2) is located in the southern portion of the Greater Kruger National Park region in Mpumalanga, South Africa, and consists of two broad study regions or land use types. These are the Sabi Sands Wildtuin, which is a combination of concession and privately owned conserved land, and the Bushbuckridge Municipality District, which includes communal rangelands that are utilised for livestock ranching, harvesting and farming activities of neighbouring informal communities. The Sabi Sands Wildtuin is approximately 54,000 hectares (ha) and is situated at 24°50′S and 31°30′E towards the Western border of the central Kruger National Park (Ben-Shahar, 1991). The entire Bushbuckridge region is approximately 260,000 ha and extends into the southernmost portion of the Limpopo Province. The region supports two broad savanna vegetation types: Lowveld Sour Bushveld (in the wetter western region) and Lowveld Mixed Bushveld (in the drier east) which make up part of the Granite Lowveld Vegetation Unit described in Mucina and Rutherford (2006). The terrain in both study regions is gently undulating with geomorphological catena sequences of crests, slopes and valleys with gabbro intrusions persisting in the Sabi Sands region and granite soil types dominating most of Bushbuckridge. Tall shrubland with few trees to moderately dense low woodland vegetation dominate these crests and slopes with dense thicket to open savannas dominating the valleys (Mucina and Rutherford, 2006). In the west, near the Drakensberg escarpment, the mean annual rainfall is approximately 1200 mm and decreases to 550 mm in the flatter interior to the east (Shackleton, 2000). Most of the rainfall falls in summer between October and April. The mean annual temperature for the region is 22 °C.

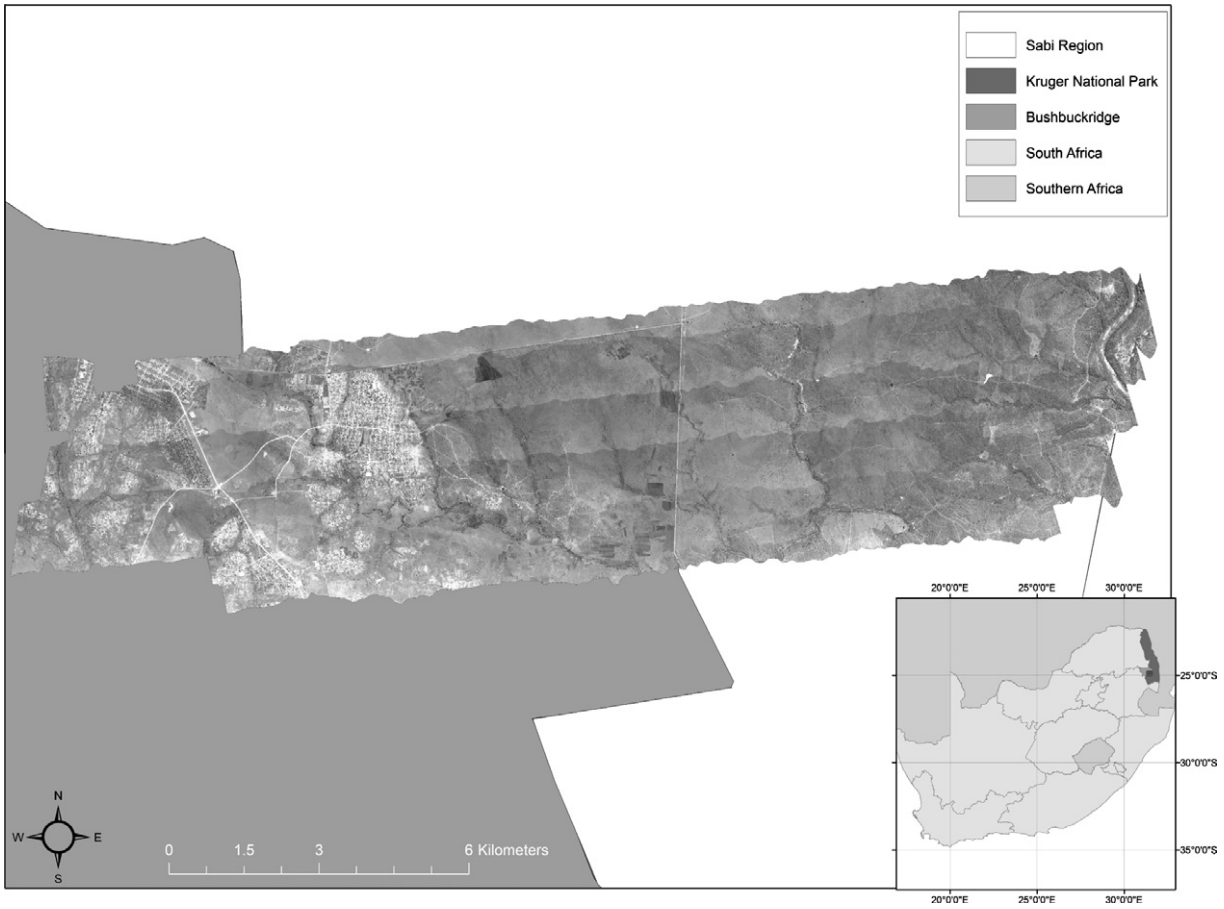

Fig. 2. A study area map of the Greater Kruger National Park showing the CAO hyperspectral dataset acquired over the Bushbuckridge and the Sabi Sands Wildtuin regions.

2.2. Hyperspectral, LiDAR, and field datasets

At the end of May 2008 an integrated hyperspectral and LiDAR dataset was acquired for 35,000 ha over the study area (Fig. 2) with the Carnegie Airborne Observatory (CAO) Alpha system. The CAO Alpha system consist of three integrated sub-systems (i) a high fidelity Compact Airborne Spectrographic Imager (CASI-1500), (ii) a waveform LiDAR (wLiDAR) capable of operating simultaneously in discrete-return and waveform modes and (iii) a GPS-IMU system allowing for an accurate registration and projection of the hyperspectral and LiDAR data. The dataset included (i) 1.1 m resolution hyperspectral images consisting of 72 bands (from 384.8 to 1054.3 nm, bandwidth) and (ii) raw LiDAR point clouds consisting of up to four ranges or returns per laser shot. The laser beam divergence of the CAO Alpha is custom-designed to precisely match the IFOV of the imaging spectrometer, meaning that there is one laser shot per pixel. For more information on the CAO system specifications, the reader will refer to Asner et al. (2007).

The hyperspectral images were converted from raw digital number (DN) measures to relative surface reflectance measures. Apparent surface reflectance was derived from the radiance data using an automated atmospheric correction model, ACORN 5Li-Batch (Imspec LLC, Palmdale, CA). Inputs to the atmospheric correction algorithm included surface elevation (captured from the LiDAR), aircraft altitude (from the GPS-IMU system), solar and viewing geometry, and estimated visibility (in km). The code used a MODTRAN look-up table to correct for Rayleigh scattering and aerosols. Water vapour was estimated directly from the 940 nm water vapour feature in the radiance data (Asner et al., 2007). For the LiDAR data, the GPS-IMU data were combined with the laser ranging data to determine the three dimensional location of the laser returns. The LiDAR point density per 1.1 m CASI pixel was approximately 1.3 for the image under investigation. From the laser point cloud data, a physical model was used to estimate surface and ground models (Digital Surface Model including the canopy surface and Digital Ground Model). Canopy height models (CHM) were computed by subtracting the DSM from the DEM.

For the field preparation, snap shot images of the hyperspectral imagery were compiled at a resolution in which individual tree canopies were clearly visible. Within these snap shot images, prominent tree canopies were marked with a point shapefile for navigation (via GPS) and identification once in the field. These marked canopies were chosen based on their ease of accessibility and their geographical representation and coverage across the study area. The pre-selected tree canopies were visited during a field visit in May 2010. Other trees and species of interest (e.g. bush encroaching species), which were too small to be clearly visible during the canopy pre-selection process, were encountered during the field visit and demarcated on the image snap shots. This was conducted in order to ascertain an appropriate level of species diversity within the modelling data since some of the pre-selected canopies may over-represent a certain few species. This over-representation was due to the tall tree height (and thus tall tree species) bias in the canopy pre-selection process as larger trees were easily visible and easier to navigate to in the field than smaller trees.

2.3. Data preparation

The pre-selected and field-demarcated tree canopies were processed by overlaying these points over the hyperspectral and LiDAR

Table 1
Attribute information relating to the biology, ecology and anthropogenic importance of the common savanna tree species under analysis.

Scientific name	Common name	Code	Attributes
Acacia gerrardii	Red Thorn	AG	Shrub to medium sized tree. Erect branches and a flattened crown Bark is grey to blackish and rough. Younger branches are reddish and hairy Thorns are in short pairs. Leaves are tiny and clustered on prominent woody cushions Fruit are sickle-shaped, hairy pods. Thorny bush encroaching species. Bark contains constricting tannin chemicals used for medicinal purposes and the inner bark is used to create twine
Acacia nigrescens	Knob Thorn	AN	Medium to large tree up to 30 m. Common in arid bushveld Bark is brown to black and covered with persistent thorn-tipped knobs Thorns are in hooked pairs and almost black. Leaves are twice-compound, leathery and hairless. Fruit are straight, olive to black pods. Timber is very hard and is thus used for making posts and mine props and can be used for flooring material
Berchemia discolor	Brown Ivory	BD	Generally large tree up to 20 m. Usually on river banks and on termitaria. Pale green covered in brown lenticels when young. Bark is dark grey and roughly fissured. Leaves are simple and slightly ovate or elliptic. Side veins form a distinctive herringbone pattern. Not as prevalent as other species. Date-like fruit are harvested as local food produce. An excellent timber species for pole and furniture making
Combretum species	Bushwillow species	COM	Small to medium sized tree. Widespread across savanna. Range from single stemmed to multi-stemmed trees. Bark ranges from pale to dark blackish/brownish grey. Leaves range from oval with rounded apex to oblong and broadly ovate. Leaves are also dull to glossy green and slighter paler below Fruits are 4-winged and have distinct colouring patterns (reddish/brownish) and vary in size from very small to distinctly large. Very common savanna tree species family which is highly abundant in the study region. Good for charcoal production and possesses numerous medicinal properties (treats certain snakebite and dysentery)
Dichrostachys cinerea	Sickle-bush	DC	Shrub or small rounded tree, often encroaching if veld is mismanaged. Branching low down and bark is rough with fissures. Side twigs are modified to form spines Small leaves clustered on spines/side shoots. Fruit occurs distinctively as a curled and twisted mass of brown pods. Hardy and pervasive bush encroaching species which impedes cattle and local movements
Pterocarpus rotundifolius	Round-leaved Bloodwood	PR	Large, rounded, woody shrub or tree. Often forming dense colonies. Usually multi-stemmed with grey young bark. Leaflets are large and rounded with distinguishing parallel side veins. Active bush encroaching species. Good for apiculture due to the rich pollen and nectar sources and plays a role in soil erosion control
Sclerocarya birrea	Marula	SB	Common in SA savannas especially on sandy frost free soils Large and dominant tree (up to 20 m). Protected tree species in SA Leaves are compound, dark green above and paler below Separate male and female trees. Has a large ovoid tasty fruit. Fruit is utilised in local brewery industry for small scale distribution and for cultural purposes
Spirostachys africana	Tamboti	SA	Large erect tree with round canopy and common on brackish flats and along seasonal streams and rivers. Occur in dense stands. Bark is very dark with cracks in rectangular blocks. White latex is present. Leaves are simple and ovate. Have small glands present on top of the petiole at the base. Fruit is a 3-lobed capsule with brown seeds. Prominently utilised in the woodcraft industry for furniture and/or sculptures tailored towards tourism
Terminalia sericea	Silver Cluster-leaf	TS	Small to medium sized tree with rounded crown to characteristically flat-topped Upright stem with reddish-brown to purplish-brown branches. Often bearing small rounded woody galls. Leaves are crowded at the branch ends. Foliage have a distinct blue-grey colour at a distance. Although being a known bush encroaching species, it is primarily utilised as fuel wood to satisfy the energy requirements of local communities

Sources: Schmidt et al. (2007), Shackleton and Shackleton (2003) and Shackleton et al. (2005).

height images to create the tree species spectral and structural libraries that were used in the analysis. Spectral and structural height data were collected from eight common savanna tree species found in the study area. These species were *A.* gerrardii/*D.* cinerea (AG/DC), *A.* nigrescens (AN), *B.* discolor (BD), *Combretum* species (COM), *P.* rotundifolius (PR), *S.* africana (SA), *S.* birrea (SB) and *T.* sericea (TS). Species such as *Combretum apiculatum, Combretum collinum* and *Combretum hereroense* were grouped together in the *Combretum* species class while *A.* gerrardii and *D.* cinerea were also grouped together under a single class because these species share very similar spectral and structural characteristics and traits. The associated ecological and social importance of these species was briefly addressed in Table 1. The spectra, which contained representative pixels (i.e. pixels encompassing complete canopies and which minimized as much of the expected ground spectral contamination as possible) for the eight different tree species, and the structural height parameter, were extracted using the Region of Interest (ROI) tool in ENVI 4.7 remote sensing software. From the hyperspectral imagery, ROIs were created to cover each of the tree species canopies from the field data which were compiled into a general ROI list. These same ROIs were overlaid over the LiDAR imagery to extract the corresponding tree height parameter. The recorded number of canopies sampled and the total number of pixels per species, from which the spectral and structural information were extracted, are summarised in Table 2. In Table 2, it is important to note a particular anomalous value for the mean height of *P.* rotundifolius (0.126 m) which was attributed to the limitation of the LiDAR sensor in detecting short trees or shrubs (<1.5 m). The CAO LiDAR data were validated against field height measurements of approximately 800 trees. There was a strong, positive relationship ($r^2 = 0.93$, $p < 0.001$, standard error = 0.73 m) but the relationship broke down for tree heights below 1.5–1.7 m which were often estimated at below 0.5 m, or were not detected at all by the LiDAR sensor (Wessels et al., 2011). This limitation resulted from a combination of factors including the limited target cross section within 1.1 m pixels of small woody plants, the user-defined thresholds for separating ground and tree returns and the LiDAR reset time which equates to approximately 0.71 m.

2.4. Random Forest predictor datasets

An ensemble of seven sets of predictor data, incorporating different combinations of remotely sensed spectral and/or structural data about the tree species, was investigated by submitting each of the seven predictor datasets in turn into a Random Forest modelling procedure to ascertain which variable(s) drove or enhanced the classification and differentiability of the target tree species on each occasion. These seven predictor datasets, including their descriptions, the band wavelengths used and associated references are summarised in Table 3. The combinations of predictors were

Table 2
Total number of recorded canopies, tree pixels sampled, and tree height statistics (obtained from the LiDAR height image layer) of the tree sample collected from the eight savanna tree species.

Species	# of canopies	Total # of pixels	Mean Ht (m)	Stdev Ht (m)
AG_DC	48	304	1.494	1.633
AN	58	792	8.748	2.580
BD	3	57	9.852	1.002
COM	71	451	3.407	3.453
PR	20	133	0.126	0.163
SA	36	619	6.561	3.222
SB	73	1590	8.732	1.972
TS	22	73	3.118	1.141

chosen for various reasons. Apart from investigating the importance of tree height in this study, two particular predictor datasets (Indices; Nutrient and Leaf Mass) were considered, which made use of particular spectral vegetation indices and spectral bands (Table 3) which are understood to detect differences in the primary and secondary plant chemical compounds found in the savanna vegetation. Since the CAO hyperspectral imagery was taken during a dry rainfall period of May 2008 (the transition between wet and dry seasons), it was thought that these differences could be significant both within and between different tree species. A subset of the CASI bands, selected by using a Spectral Angle Mapper (SAM) approach that had been previously applied by Cho et al. (2010), were also used as one of the predictor datasets. SAM measures the angle between two spectra, which are considered as vectors in *n*-dimensional space, to determine the degree of spectral similarity between those spectra (Cho et al., 2009; Sobhan, 2007; Du et al., 2004; Van der Meer, 2006). The smaller the angle between the two spectra, the more spectrally similar the spectra are while the converse is true if the angle is larger. Cho et al. (2010) utilized a Band Add-On mathematical procedure to select and identify the most appropriate bands for species discrimination. The Band Add-On algorithm selects the bands that maximise inter-species SAM and starts off by selecting the two bands which have the highest average SAM, among all pair-wise combinations. The procedure is then to consecutively add pairs of bands until no bands contribute further to the discriminatory power of the species in question (Cho et al., 2010; Keshava, 2003). For more detailed information regarding this Band Add-On procedure, the reader is recommended to refer to Keshava (2003). The selected bands, displayed in Table 3, were found to improve savanna tree species discrimination, according to Cho et al. (2010), in comparison to the implementation of all available bands in the entire dataset and were thus considered for this study. The raw spectral reflectance data from the CAO hyperspectral imagery were considered as a baseline predictor dataset in which all 72 bands of the collected species' spectral endmembers were fed into the Random Forest model. This 72 band

Table 3
Seven predictor datasets that were modelled in RF including their description and the formulae or wavelengths used to create them including their associated references.

Predictor dataset	Description	Formulae/wavelengths used (nm)	References
Height	Tree height of individual tree species (recorded in metres)		
Indices	Four main Vegetation Spectral Indices were selected:		
	Carotenoid Reflectance Index (CRI)	$\lambda 800(1/\lambda 520 - 1/\lambda 550)$	Gitelson et al. (2002)
	Photochemical Reflectance Index (PRI)	$(\lambda 531 - \lambda 570)/(\lambda 531 + \lambda 570)$	Gamon et al. (1992)
	Normalized Difference Vegetation Index (NDVI)	$(\lambda 800 - \lambda 678)/(\lambda 800.5 + \lambda 678)$	Rouse et al. (1973)
	Red Edge NDVI (RE)	$(\lambda 750 - \lambda 705)/(\lambda 750 + \lambda 705)$	Gitelson and Merzlyak (1994)
Height + indices	Tree species' height data and Vegetation Spectral Indices (CRI, PRI, NDVI and RE) combined in a single dataset		
Raw bands	Spectral reflectance data of the 72 raw bands of the CAO hyperspectral sensor	384.8; 394.3; 403.7; 413.1; 422.6; 432; 441.4; 450.9; 460.3; 469.7; 479.2; 488.6; 498.1; 507.5; 517; 526.4; 535.9; 545.3; 554.8; 564.2; 573.7; 583.1; 592.6; 602; 611.5; 620.9; 630.4; 639.9; 649.3; 658.8; 668.2; 677.7; 687.1; 696.6; 706; 715.5; 724.9; 734.4; 743.8; 753.3; 762.7; 772.1; 781.6; 791; 800.5; 809.9; 819.3; 828.8; 838.2; 847.6; 857; 866.5; 875.9; 885.3; 894.7; 904.1; 913.5; 922.9; 932.3; 941.7; 951.1; 960.5; 969.9; 979.3; 988.7; 998.1; 1007.4; 1016.8; 1026.2; 1035.6; 1044.9; 1054.3	
Continuum removed transformed (CRT) bands	Spectral reflectance data in the continuum removed transformed format (72 transformed bands) *Utilized the built-in function in the spectral profile viewer in ENVI 4.7*	$S_{cr} = (S/C)$ where S_{cr} = Continuum-removed spectra, S = original spectrum (λ), C = continuum curve (λ)	Mutanga and Skidmore (2003)
Spectral Angle Mapper (SAM) selected bands	Spectrally significant bands (31 bands) selected from mathematical Band Add-On procedure It selects bands which have highest average SAM among all pairwise comparisons and keeps adding on the next consecutive bands until none are left	706; 762.7; 696.6; 668.2; 677.7; 687.1; 715.5; 724.9; 734.4; 743.8; 753.3; 384.8; 394.3; 403.7; 413.1; 422.6; 913.5; 819.3; 828.8; 838.2; 847.6; 857; 866.5; 875.9; 885.3; 894.7; 904.1; 1016.8; 922.9; 932.3; 941.7	Cho et al. (2010)
Nutrient and Leaf Mass (N + LM) bands	Selected bands representing leaf nutrients (e.g. chlorophyll) and leaf mass (e.g. LAI) Associated with green biomass	466 (Chlorophyll *b*) 695 (Total chlorophyll) 725 (Total chlorophyll, leaf mass) 740 (Leaf mass and LAI) 786 (Leaf mass) 846 (Leaf mass, LAI, chlorophyll)	Cho et al. (2007)

raw dataset was then subjected to a continuum removed transformation to create a new predictor dataset. This transformation was done to enhance the absorption features of the mean reference spectral values evident in the spectral profiles and to minimize the differences caused by the variability of solar illumination at each pixel-crown position (Odagawa and Okada, 2009). This transformation could also contribute to minimizing any effects arising from any possible Bi-directional Reflectance Distribution Function (BRDF) effect in the imagery.

Finally, the predictors which contributed most to classification success were identified from each of the seven initial datasets of predictors. These were then combined to create a further so-called 'hybrid' dataset, which was then submitted to the RF in an attempt to improve overall classification accuracy.

2.5. Random Forest model background, methods and validation

Random Forest, developed by Leo Breiman and Adele Cutler, is a type of data mining technology which combines information from a collection of virtually grown decision trees (Salford Systems, 2004). This collection or 'forest' of decision trees are grown from user-defined target and eligible predictor data via bootstrap sampling, where only randomly iterated two third's of the original training data is used for each tree, and the random selection of splitting variables, used to split the nodes in the tree construction. The 'forest' of decision trees is then grown out to its maximum possible size (defined by the user) and is left unpruned (Salford Systems, 2004). These individual trees are then combined through a weighted voting process to determine the most effective model. Similarly to other decision tree techniques, such as CART, Random Forest automatically selects the most significant predictors from a suite of eligible candidates and are insensitive to missing data values but unlike other decision tree and data mining methods, it is not prone to model over-fitting (as each tree is grown independently) and possesses built-in self validation via the implementation of an 'Out-of-Bag' dataset (to be elaborated upon later) (Salford Systems, 2004).

The Random Forest modelling was performed in the Random Forest integrated module of the Salford Predictive Modeller Builder 6.6 software package (Salford Systems, 2004). The different datasets of the predictor types were inputted separately into the Random Forest dialogue and the various model settings were adjusted accordingly. The class weights were 'balanced' for all instances which meant that the small classes were 'up-weighted' to equal the size of the largest target class. Species classes such as *A. nigrescens* and *S. birrea* contain much larger sample sizes than for instances *B. discolor* so a balancing of classes is required to reduce possible bias. According to Ismail et al. (2010) and Prasad et al. (2006), there are two main tuning parameters required in a Random Forest – the number of trees to be built in the 'forest' and the number of possible splitting variables/predictors considered for each node in the trees. For this study, the number of trees to be built was kept at the default number of 500 trees while a standard rule of thumb, the squared root of the total number of predictors, was implemented to determine the appropriate number of possible predictors considered for each node. Researchers have reported that these default values and the rule of thumb often produce acceptable results (Liaw and Wiener, 2002 cited in Ismail et al., 2010; Salford Systems, 2004; Dahinden, 2011).

Since Random Forest makes use of an internal Out-of-Bag (OOB) sampling procedure, which calculates an unbiased and reliable error rate, an independent validation dataset was not necessary for this study (Lawrence et al., 2006; Prasad et al., 2006). During this OOB sampling procedure, approximately a third of the sample, selected at random and excluded from the bootstrapped sample going into the construction of a tree in the Random Forest, would be reserved as an internal test dataset for the validation of that tree in the Random Forest model (Ismail et al., 2010). This process is iterated for every tree in the 'forest'. Consequently each decision tree is grown and validated independently from the other trees in the Random Forest. The reliability of using this OOB dataset and its resulting estimates of accuracy was supported by the accuracy assessment comparisons of a separate test and OOB datasets in the Lawrence et al. (2006) study and was also successfully documented in other studies (Prasad et al., 2006; Furlanello et al., 2003; Grossmann et al., 2010). Once the Random Forest models have been executed, various results per predictor dataset were available but only the most informative results are presented.

Among the summary reports created by the Random Forest software, variable importance, misclassification and prediction success were chosen for presentation in this study. Variable importance is evaluated based on the degradation of the prediction if the data for the particular predictors were interchanged randomly (Prasad et al., 2006). This is important for ascertaining which predictor(s) are driving the differences in accuracy between the different classifications. Hence, it helps in improving the understanding of which predictor(s) are most suitable for modelling by identifying the smallest number of predictors that possess the best discriminatory potential (Ismail et al., 2010). The Gini Index was considered to ascertain the most important predictors (i.e. the scores greater than 80). In the Gini Index the most important predictor(s) receive a score of 100 while the remaining less significant predictor(s) receive a decreasing score (Salford Systems, 2004). The Gini Index is a type of splitting criterion which functions in a manner in which at every split or branching of a tree in the 'forest' created by a particular predictor, there is a resultant decrease in the Gini Index. The summation of all the decreases attributed by that particular predictor in the 'forest', normalised by the number of trees in the 'forest', is termed the Gini Index for that predictor (Breiman, 2003). The scores obtained by the Gini Index have no unit. The Gini Index score of 80 and greater was chosen as the authors' conservative interpretation of which predictors were considered valuable and qualified for incorporation into the hybrid dataset classification model. Misclassification and prediction success both indicate the overall effectiveness of the Random Forest model in terms of classification accuracy assessment. Confusion matrices were created while overall, and species specific user's and producer's accuracies were computed. The producer's accuracy indicates the probability of a reference pixel being correctly classified (i.e. the total number of correctly classified pixels in a class divided by the total number of reference pixels of that class – a measure of omission error). The user's accuracy, on the other hand, indicates the probability that a pixel classified on the image actually represents that class on the ground (i.e. the total number of correctly classified pixels in a class divided by the total number of pixels that were classified in that particular class – a measure of commission error) (Congalton, 1991). The confusion matrices were created by comparing the modelled data against the internal test OOB sample data in each predictor dataset model.

A Kappa statistic (KHAT) was also calculated, complementing the overall classification accuracy, to ascertain the most accurate Random Forest model while the Gini Index variable importance values were reviewed to determine the most significant predictor(s). The Kappa statistic evaluates the pairwise agreement among a set of classes while correcting for expected chance agreement (Carletta, 1996; Prasad et al., 2006). The values range from −1, which indicates complete disagreement between classes, to +1, which indicates a perfect agreement (Prasad et al., 2006). This statistic is a powerful technique in its capacity to compare the results from multiple confusion matrices (Congalton, 1991). The formula for KHAT (Formula (1)) and accompanying explanation is included below:

$$\mathrm{KHAT} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+} \times x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+} \times x_{+i})} \quad (1)$$

where r is the number of rows in the confusion matrix, x_{ii} is number of observations in row i and column i, x_{i+} and x_{+i} are the totals of row i and column i, respectively and N is the total number of observations (Congalton, 1991).

A hybrid dataset of predictors was then created by obtaining the most important predictors (i.e. Gini Index score of greater than 80) from the seven modelled predictor datasets in order to attempt to achieve a superior Random Forest model and classification accuracy assessment than the results of the different predictor datasets separately. These important predictors which created the hybrid dataset are displayed in Table 4 in the results section. Lastly, a flow diagram (Fig. 3) was included, in order to summarise all the necessary steps taken in the methodology.

3. Results

3.1. Predictor importance

From the results in Table 4 of the seven separate RF models, 23 predictors (excluding the duplicated results between the predictor datasets) had a Gini score greater than 80. Among these, tree height; NDVI; chlorophyll *b* wavelength and selected raw (mostly in the blue region around chlorophyll *b*), continuum removed (mostly in the red region) and SAM (mostly blue and red) wavelengths appeared to contribute the most to the classification success. When the datasets with the highest Gini scores were input to a further RF model as a hybrid dataset, the tree species height predictor was by far the most valuable predictor (Gini Index of 100) contributing to prediction success. The second most significant predictor was the continuum removed transformed band 30 (658.8 nm) which only achieved a Gini Index score of 65.84.

3.2. Modelled prediction success

The summary results in Fig. 4 illustrate the classification accuracies for the Random Forest models of the different predictor datasets and for the hybrid dataset. Fig. 5 illustrates the total number of misclassified pixels for the seven predictor datasets and the hybrid dataset. Amongst the seven initial predictor dataset results, the Random Forest model combining the predictors tree height and vegetation spectral indices (Ht + Indices) yielded the highest overall classification accuracy of 82.38%, KHAT of 0.776 and the least number of misclassified pixels (708) compared to the other models. The use of the tree height variable only in the Random Forest modelling yielded by far the lowest classification accuracy (overall accuracy of 31.90% and KHAT of 0.186) while the use of raw bands produced the highest accuracy amongst the strictly spectral datasets (overall accuracy of 80.29% and KHAT of 0.755). However, the hybrid dataset yielded the highest overall classification accuracy of 87.68%, a KHAT of 0.843 and only 495 misclassified pixels.

The confusion matrix resulting from the raw band RF model, which achieved the highest overall classification accuracy for a spectra-only predictor dataset (according to Fig. 4), is presented in Table 5. In Table 5, the producer's accuracies were very high for most of the eight tree species with the exception of *S.* birrea which achieved an accuracy of 67.23%. The user's accuracies achieved, however, were highly variable with some species achieving high accuracies of upper 80% and 90% (e.g. *S. birrea*, *A.* nigrescens and *S. africana*) with others achieving poorer accuracies of less than 50% (e.g. *T.* sericea with 36.70% and *B.* discolor with 29.73%). Majority of the spectral confusion in the model was evident between *S.* birrea and some of the shorter tree species (particularly *A. gerrardii*/*D.* cinerea and *T.* sericea with 160 and 111 classified pixels being incorrectly predicted on the ground, respectively). Almost all producer's and user's accuracies clearly improved in the confusion matrix results of the hybrid model, displayed in Table 6. In Table 6, all eight tree species classes were classified at a very high producer's accuracies with the lowest being 78.27% for *Combretum* species. *T.* sericea yielded the highest producer's accuracy (97.26%) within the sample population. The user's accuracy, on the other hand, complemented most of the species with high performing producer's accuracy with only a few exceptions. *B. discolor*, similarly in Table 5, was the most problematic species in the dataset with the lowest user's accuracy of 35.29% (confusion with *S.* africana and mostly *S. birrea*) which starkly contrasted with its 94.74% producer's accuracy. *S.* birrea achieved an improved producer's accuracy of 82.33% in comparison to the modelled results in Table 5. The

Table 4
The Gini Index score summary table and the most significant predictors in each predictor dataset which scored greater than 80.* For clarity of the hybrid dataset result, the Gini Index scores of all modelled predictors were included.

Predictor dataset	Important variables/predictors	Gini Index score
Height	Height	100
Indices	NDVI	100
Height + indices	Height	100
	NDVI	84.06
Raw bands	B8 (450.9 nm)	100
	B35 (706 nm)	97.35
	B9 (460.3 nm)	91.93
	B10 (469.7 nm)	90.82
	B11 (479.2 nm)	89.3
	B7 (441.4 nm)	87.36
	B14 (507.5 nm)	86.54
	B6 (432 nm)	82.1
CRT bands	B30 (658.8 nm)	100
	B32 (677.7 nm)	99.95
	B31 (668.2 nm)	96.66
	B10 (469.7 nm)	94.29
	B33 (687.1 nm)	92.92
	B12 (488.6 nm)	89.07
	B39 (743.8 nm)	88.61
	B29 (649.3 nm)	86.15
	B11 (479.2 m)	82.91
SAM bands	B10 (706 nm)	100
	B4 (413.1 nm)	95.91
	B5 (422.6 nm)	92.15
	B6 (668.2 nm)	87.79
	B7 (677.7 nm)	83.34
N + LM bands	B1 (466 nm)	100
Hybrid	Height	100
	CRT B30 (658.8 nm)	65.84
	CRT B10 (469.7 nm)	65.29
	CRT B32 (677.7 nm)	60.30
	CRT B11 (479.2 m)	57.90
	SAM B10 (706 nm)	51.45
	CRT B12 (488.6 nm)	49.45
	CRT B31 (668.2 nm)	46.49
	CRT B33 (687.1 nm)	42.71
	CRT B29 (649.3 nm)	33.02
	CRT B39 (743.8 nm)	31.87
	NDVI	29.46
	RAW B14 (507.5 nm)	22.14
	RAW B7 (441.4 nm)	18.66
	RAW B6 (432 nm)	15.74
	RAW B8 (450.9 nm)	15.69
	RAW B11 (479.2 nm)	15.42
	RAW B9 (460.3 nm)	13.22
	SAM B4 (413.1 nm)	12.63
	N + LM B1 (466 nm)	12.03
	SAM B5 (422.6 nm)	10.68
	SAM B6 (668.2 nm)	6.32
	SAM B7 (677.7 nm)	5.49

B = bands.
* With exception to the hybrid predictor dataset.

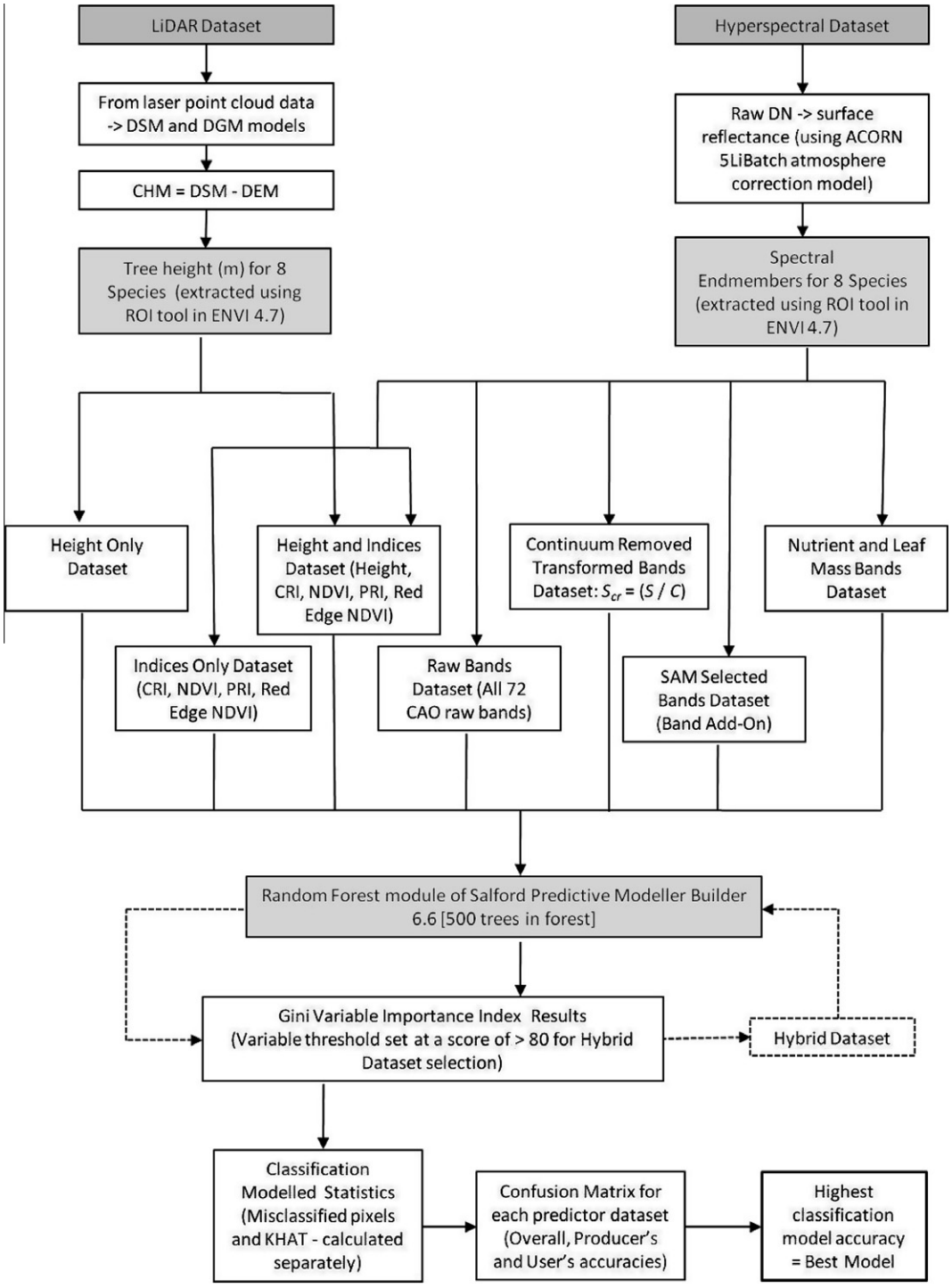

Fig. 3. Methodology schema illustrating all steps required to classify tree species including the LiDAR and hyperspectral data pre-processing, the seven predictor dataset creation, variable importance assessment and misclassification and prediction success. DSM = Digital Surface Model, DGM = Digital Ground Model, CHM = canopy height model, DEM = Digital Elevation Model, DN = digital numbers, ROI = Region of Interest.

remaining species displayed moderate (>60%) to high (>80%) user's accuracies.

4. Discussion

From the Gini Index results for the initial ensemble of datasets used to make RF models (Table 4), the more influential spectral bands (from the raw, CRT and SAM bands) were found to have originated from the visible wavelength spectrum with the available infrared wavelengths playing a lesser role in assisting the Random Forest classification. This observation coincided with the results in the Cho et al. (2010) study which concluded that the most significant bands for savanna tree species discrimination originated from the red-edge and blue region. Due to the limited spectral range of the CAO sensor (384.8–1054.3 nm), the complete infrared region (including shortwave infrared) could not be fully tested and assessed in this study. Amongst the four spectral vegetation indices used in the Indices predictor dataset, it was interesting that NDVI

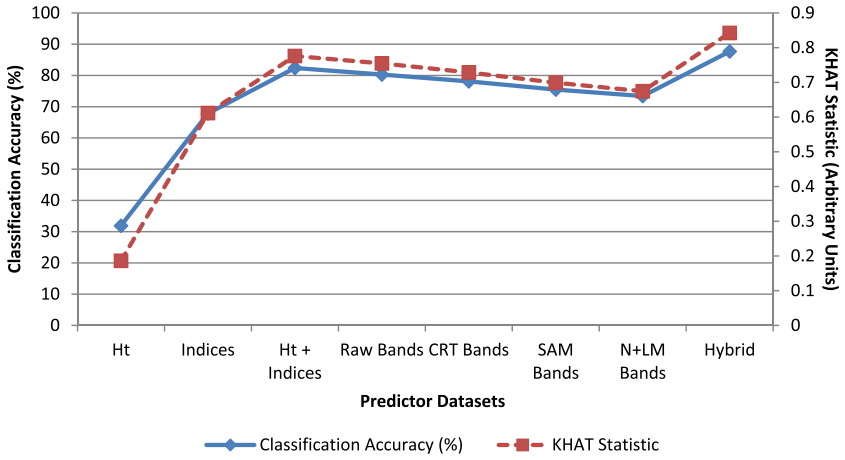

Fig. 4. Overall classification accuracies (percent) and KHAT statistic values for the Random Forest classification of eight common savanna tree species using eight predictor datasets.

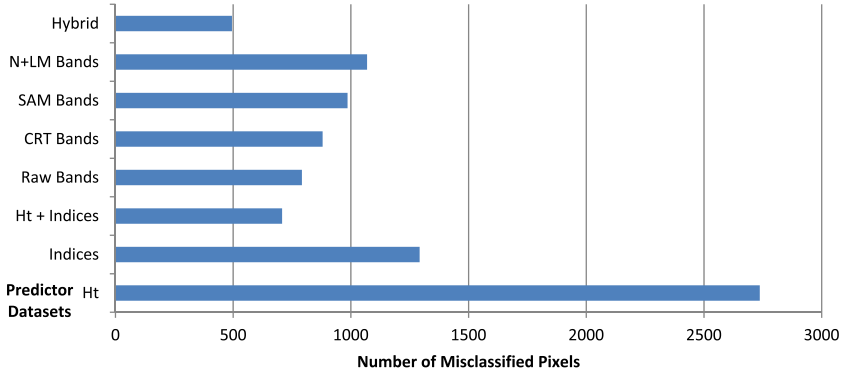

Fig. 5. Total number of misclassified pixels resulting from the Random Forest classification of eight common savanna tree species using eight predictor datasets.

Table 5
Confusion matrix displaying the producer's and user's classification accuracies, for all eight tree species, obtained from the Random Forest modelling of the raw band predictor dataset.

Classified → Field ↓	Producer's accuracy (%)	User's accuracy (%)	AG/DC	AN	BD	COM	PR	SA	SB	TS	Reference pixels
AG/DC	82.57	57.05	251	1	18	6	15	2	11	0	304
AN	96.46	89.88	7	764	0	5	1	0	11	4	792
BD	96.49	29.73	1	0	55	0	0	0	1	0	57
COM	77.83	78.00	6	19	12	351	41	13	5	4	451
PR	87.97	51.77	8	1	2	4	117	0	1	0	133
SA	89.01	95.33	7	0	52	7	1	551	1	0	619
SB	67.23	97.01	160	65	46	76	51	12	1069	111	1590
TS	94.52	36.70	0	0	0	1	0	0	3	69	73
Classified pixels			440	850	185	450	226	578	1102	188	4019

Table 6
Confusion matrix displaying the producer's and user's classification accuracies, for all eight tree species, obtained from the Random Forest modelling of the hybrid predictor dataset.

Classified → Field ↓	Producer's accuracy (%)	User's accuracy (%)	AG/DC	AN	BD	COM	PR	SA	SB	TS	Reference pixels
AG/DC	90.79	77.75	276	0	2	8	16	1	1	0	304
AN	96.59	88.75	5	765	2	2	0	0	15	3	792
BD	94.74	35.29	1	0	54	0	0	1	1	0	57
COM	78.27	83.45	22	16	3	353	31	10	9	7	451
PR	87.97	70.48	12	0	0	4	117	0	0	0	133
SA	93.54	95.39	6	0	22	8	0	579	2	2	619
SB	82.33	97.91	32	81	70	47	2	16	1309	33	1590
TS	97.26	61.21	1	0	0	1	0	0	0	71	73
Classified pixels			355	862	153	423	166	607	1337	116	4019

was scored as the most important vegetation index by the Gini index (100). However, in the context of the spectral indices and height dataset results, it was the tree height predictor which scored as more important in the classification model than the NDVI or any of the other spectral indices. This was also found in the hybrid dataset, which was a combination of all the significant predictors from the separately modelled results. In the hybrid dataset, the tree height predictor also was the most important predictor (Gini Index score of 100), followed by the continuum removed transformed bands 30, 10, 32, and 11 with a Gini Index score varying between 65.84 and 57.90. The large difference in Gini Index score between the highest and the second highest scoring predictors could illustrate the dominance of the tree height predictor over the spectral predictors in the Random Forest classification process. However, the inclusion of several spectral predictors (particularly CRT bands, but possibly others too), although appearing to be less influential according to the Gini Index score, seem to have largely contributed to the overall success of the hybrid model, because much poorer accuracy was obtained when tree height was used as a single predictor dataset (31.90% overall accuracy).

Considering the success of the seven original RF models, the classification results in Figs. 4 and 5 show that the dataset which combined vegetation indices and tree height (Ht + Indices) yielded the highest classification accuracy (82.38%; KHAT of 0.776) compared to the other six predictor datasets. When the tree height predictor was combined with the most influential spectral predictors from each of the separate predictor datasets (NDVI, chlorophyll *b* wavelength and selected raw, CRT and SAM wavelengths) to create the final hybrid dataset, the highest classification accuracy and prediction success (87.68%; KHAT of 0.843) were achieved. We conclude from these results that the incorporation of spectral information together with structural information proved to be more useful for species level classification than the use of either spectral (highest accuracies achieved by raw bands predictor dataset – 80.29%; KHAT of 0.755) or structural information (31.90%; KHAT of 0.186) alone. It could be further argued the Gini values indicate that in this study the most important predictor is the tree height (this predictor always has the highest index when used in one dataset) while the classification accuracy results show that the spectral data contribute by far most to class accuracy, but classification is further improved by the addition of the structural tree height parameter.

When comparing the confusion matrix results in Tables 5 and 6, it was evident that height mainly, together with the supporting spectral predictors in the continuum removed visible spectrum (refer to Table 4), significantly improved the species specific classification accuracies (both producer's and user's) with a particular reduction in the errors of commission in the short tree species. For example, in the case of *P. rotundifolius*, 51 classified pixels were misclassified as *S.* birrea in the raw band RF model but this figure reduced to only 2 misclassified pixels in the hybrid RF model. In a clearer case regarding *A. gerrardii*/*D. cinerea*, 160 pixels were also misclassified as *S.* birrea in the raw band RF model but this figure dropped to 32 misclassified pixels in the hybrid RF model. In addition to the improved classification results in these short tree species, the producer's accuracy for *S. birrea*, a tall tree species, also improved greatly from 67.23% to 82.33%. This improvement was likely due to the inclusion of the height predictor which partly resolves the spectral confusion between species – a claim which was well supported by the Gini Index results of the hybrid model in Table 4. *B. discolor*, however, remained a problematic species to classify even under the hybrid RF model. Although achieving a producer's accuracy greater than 90%, the user's accuracy was low (approximately 35%). A plausible reason for the poor mapping accuracy for this species class could be that an insufficient number of tree canopies were sampled and related to image pixels (only three canopies containing 57 pixels were sampled in the field) than were needed for an effective Random Forest classification. An increase in the sampled data for *B.* discolor might improve the currently low user accuracy. Although not being considered as a problematic species, *Combretum* species yielded the lowest producer's accuracy (78.27%) in hybrid RF model. This accuracy could be attributed to the high variability in the tree species' height range (evident in Fig. 1) which meant that the relative advantage of using height for discriminating *Combretum* species from other species was limited. This observation was further supported by their very similar producer's accuracies in Tables 5 and 6.

Besides the *B.* discolor species class, the evidence from this study is that the datasets used for the hybrid RF model could produce reliable and accurate species distribution maps for the seven savanna tree species in the Kruger National Park Region, which would be valuable for assisting management of both the communal and protected savanna lands. Specific uses of these potential tree species distribution maps could include bush encroachment (*Terminalia* serica and *A. gerrardii*/*D. cinerea*), fuel wood stock and NTFP (*S. birrea*) monitoring. These tree species maps could also be implemented in future ecological studies conducted in the study area, which tackle the understanding of the ecological processes and effects of environmental factors on species distribution and density (e.g. elephant impact on preferentially targeted species). The overall classification accuracies achieved exceeded those achieved in previous tree species classification efforts in South African savannas (Cho et al., 2010, 2011) and in other related ecosystems such as the shrubby American rangelands (Lawrence et al., 2006). The limited number of earlier studies mapping savannas tree species to date makes it difficult to place these results in suitable context but will, hopefully, encourage the emergence of other future studies.

Despite the general success of the classification results and the robustness of the Random Forest approach displayed in this study, Random Forest is still considered to be a 'black-box' approach due mainly to the fact that the user cannot separately analyse and view the individual decision trees created in the 'forest' and has only a small number of settings with which to define or adjust the model (Prasad et al., 2006). As a result, implementing the optimal decision tree as a classifier in popular remote sensing mapping software (e.g. ENVI) would be very challenging for the user wishing to put this classification model into practice, for example to produce maps over larger image extents. Investigating alternative scripting and programming related approaches could circumvent this issue but this is beyond the scope of this study. Classification accuracies, although yet generally high for the hybrid dataset, might be improved further by implementing the probability cut-off adaptation (bias adjustment) approach which can improve the cross-validated error rate for unbalanced datasets, as implemented and proven successful in Dahinden (2011) and Grossmann et al. (2010). Also, instead of using the traditional Gini Index to measure the relative importance of predictor datasets, predictor data sets selected by other techniques, such as the sequential reverse and forward variable selection method (Grossmann et al., 2010) or the backward and recursive variable selection method (Ismail et al., 2010), could be used and classification results compared. These alternative methods for selecting input variables could prove to be effective especially when dealing with large datasets which have many explanatory variables that have similar importance measures (Jiang et al., 2004 cited in Ismail et al. (2010)). The incorporation in the modelling process of other more complex structural parameters, such as for instance canopy volume, canopy height and tree fractional cover that can be obtained by a higher resolution waveform LIDAR, could also be investigated further.

In closing, the repeatability of this study, especially involving the acquisition of multiple temporal datasets, would depend highly

on the logistics and funds available for initiating future flight plans of this integrated CAO sensor. This point bring into light an important limitation in that the CAO specialized integrated datasets are usually very expensive to obtain and repeat on a meaningful temporal interval (seasonally for inter-seasonal predictor datasets construction or annually for change detection analysis, etc.), especially in an African landscape, which ultimately limits widespread application of this study for larger geographical study sites. This not withstanding, the sensor's affordability would be limited to small to medium sized protected areas, such as the Sabi Sands Wildtuin, in which it would serve as a new toolbox for research between very detailed field studies and coarse satellite data.

5. Conclusions

Readdressing the questions, posed in the introduction of this study, it can be concluded that using the hybrid dataset consisting of 23 spectral and structural parameters as input to a Random Forest model yielded the highest classification accuracy and prediction success for the eight savanna tree species in the study region, with an overall classification accuracy of 87.68% and KHAT value of 0.843. The most important predictors, which played an important role in the initial classification models and also contributed to the success of the hybrid dataset model when they were combined together, were species tree height; NDVI; the chlorophyll *b* wavelength (466 nm) and a selection of raw, continuum removed and SAM bands (see Table 4 for the entire list of significant predictors). According to the Gini Index measures of variable importance and also from the classification results, it was clear that tree species height derived from the LiDAR was the predictor most influential to overall classification success. However, when tree height was used on its own it yielded the lowest overall classification results. It can only be concluded from this that tree height significantly improves the classification accuracy for savanna tree species only when it is combined with other spectral predictors. The incorporation in the modelling process of other more complex structural parameters, such as canopy volume, canopy height and tree fractional cover, which can be obtained by a higher resolution waveform LiDAR, could further test this claim. The overall classification accuracies achieved exceeded those achieved in previous tree species classification efforts in South African savannas (Cho et al., 2010, 2011) and in other related ecosystems such as the shrubby American rangelands (Lawrence et al., 2006). The main limitation of this study, however, would be its current scope in that it would be limited to small to medium protected areas and that its affordability would ultimately be limited by the expensive logistical costs and challenges associated with the deploying of such an airborne sensor system in an African landscape.

Acknowledgements

The authors would like to graciously thank the Council for Scientific and Industrial Research (CSIR), South Africa, for the providing the necessary financial support for this study. Thanks also go to the Andrew Mellon Foundation for the funding of the airborne remote sensing with the CAO. The CAO was made possible by the W.M. Keck Foundation, the Gordon and Betty Moore Foundation and William Hearst III. The hyperspectral and LiDAR pre-processed data products used in this study was made possible by Dr. Greg Asner and his CAO team of T. Kennedy-Bowdoin, D. Knapp, J. Jacobson and R. Emerson. The authors would finally like to thank the colleagues involved, at the Ecosystems Earth Observation Unit in the CSIR, for their assistance in field work and other contributions to this study. Special thanks also go to two anonymous reviewers for their valuable contributions.

References

Affendi, S., Ainuddin, N.A., Shafri, H.Z.M., 2009. A rule based approach for the mapping of tropical forest canopy from airborne hyperspectral data. <http://www.geospatialworld.net> (accessed 30.06.09.).

Asner, G.P., Knapp, D.E., Kennedy-Bowdoin, T., Jones, M.O., Martin, R.E., Boardman, J., Field, C.B., 2007. Carnegie Airborne Observatory: in-flight fusion of hyperspectral imaging and waveform LiDAR for 3D studies of ecosystems. Journal of Applied Remote Sensing 1 (1), 1–27.

Asner, G.P., Knapp, D.E., Kennedy-Bowdoin, T., Jones, M.O., Martin, R.E., Boardman, J., Hughes, R.F., 2008. Invasive species detection in Hawaiian Rainforests using Airborne Imaging Spectroscopy and LiDAR. Remote Sensing of Environment 112 (5), 1942–1955.

Ben-Shahar, R., 1991. Abundance of trees and grasses in a woodland savanna in relation to environmental factors. Journal of Vegetation Science 2 (3), 345–350.

Bork, E.W., Su, J.G., 2007. Integrating LiDAR data and multispectral imagery for enhanced classification of rangeland vegetation: a meta analysis. Remote Sensing of Environment 111 (1), 11–24.

Breiman, L., 2003. Manual on setting up, using, and understanding Random Forests v4.0, <http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf> (accessed 08.02.11.).

Carletta, J., 1996. Assessing agreement on classification tasks: the Kappa statistic. Computational Linguistics 22 (2), 1–6.

Cho, M.A., Skidmore, A., Corsi, F., van Wieren, S.E., Sobhan, I., 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. International Journal of Applied Earth Observation and Geoinformation 9 (4), 414–424.

Cho, M.A., Debba, P., Mathieu, R., van Aardt, J., Asner, G.P., Naidoo, L., Main, R., Ramoelo, A., Majeke, B., 2009. Spectral variability within species and its effects on savanna tree species discrimination. In: Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Cape Town, South, Africa, 12–19 July (on CDROM).

Cho, M.A., Debba, P., Mathieu, R., Naidoo, L., van Aardt, J., Asner, G.P., 2010. Improving discrimination of savanna tree species through a multiple endmember spectral angle mapper (SAM) approach: canopy level analysis. IEEE International Journal of Geoscience and Remote Sensing 48 (11), 4133–4142.

Cho, M.A., Naidoo, L., Mathieu, R., Asner, G.P., 2011. Mapping savanna tree species using Carnegie Airborne Observatory hyperspectral data resampled to World View-2 multispectral configuration. In: Proc. 34th International Symposium on Remote Sensing of Environment, Sydney, Australia, 10–15 April (on CDROM).

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35–46.

Cortijo, F.J., de la Blanca, N.P., 1996. Image classification using non-parametric classifiers and contextual information. International Archives of Photogrammetry and Remote Sensing (Part B3) 31, 120–124.

Dahinden, C., 2011. An improved random forest approach with application to the performance prediction challenge datasets. In: Guyon, I., Crawley, G., Dror, G., Saffari, A. (Eds.), Hands-on Pattern Recognition, Challenges in Machine Learning, vol. 1. Microtome, MA, pp. 223–230.

Du, Y., Chang, C.-I., Ren, H., Chang, C.-Chi., Jensen, J.O., D'Amico, F.M., 2004. New hyperspectral discrimination measure for spectral characterization. Society of Photo-Optical Instrumentation Engineers 8 (43), 1777–1786.

Furlanello, C., Neteler, M., Merler, S., Menegon, S., Fontanari, S., Donini, A., Rizzoli, A., Chemini, C., 2003. GIS and the random forest predictor: integration in R for tick-borne disease risk assessment. In: Proc. 3rd International Workshop on Distributed Statistical Computing, March 20–22, Vienna, Austria, pp. 1–11.

Gamon, J., Penuelas, J., Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sensing of Environment 41 (1), 35–44.

Geerling, G.W., Labrador-Garcia, M., Clevers, J.G.P.W., Ragas, A.M.J., Smits, A.J.M., 2007. Classification of floodplain vegetation by data-fusion of spectral (CASI) and LiDAR data. International Journal of Remote Sensing 28 (19), 4263–4284.

Gitelson, A.A., Merzlyak, M.N., 1994. Spectral reflectance changes associated with autumn senescence of *Aesculus Hippocastanum* L. and *Acer Platanoides* L. leaves. Spectral features and relation to chlorophyll estimation. Journal of Plant Physiology 143 (3), 286–292.

Gitelson, A.A., Zur, Y., Chivkunova, O.B., Merzlyak, M.N., 2002. Assessing carotenoid content in plant leaves with reflectance spectroscopy. Photochemistry and Photobiology 75 (3), 272–281.

Grossmann, E., Ohmann, J., Kagan, J., May, H., Gregory, M., 2010. Mapping ecological systems with a random forest model: tradeoffs between errors and bias. Gap Analysis Bulletin 17 (1), 16–22.

Hestir, E.L., Khanna, S., Andrew, M.E., Santos, M.J., Viers, J.H., Greenberg, J.A., Rajapakse, S.S., Ustin, S.L., 2008. Identification of invasive vegetation using hyperspectral remote sensing in the California Delta ecosystem. Remote Sensing of Environment 112 (11), 4034–4047.

Hill, R.A., Thomson, A.G., 2005. Mapping woodland species composition and structure using airborne spectral and LiDAR data. International Journal of Remote Sensing 26 (17), 3763–3779.

Ismail, R., Mutanga, O., Kumar, L., 2010. Modelling the potential distribution of pine forests susceptible to *Sirex Noctilo* infestations in Mpumalanga, South Africa. Transactions in GIS 14 (5), 709–726.

Joubert, S., 2007. The Kruger National Park A history, vol. I–III, first ed. High Branching (Pty) Ltd., Johannesburg, South Africa.

Keshava, N. 2003. Angle-based band selection for material identification in hyperspectral processing. In: Proc. Algorithms and Technologies for Multispectral, Hyperspectral and Ultraspectral Imagery IX (SPIE), Orlando, FL, 21 April, pp. 440–451.

Kim, S., 2007. Individual tree species identification using LiDAR-derived crown structures and intensity data. Doctoral thesis. University of Washington, College of Forest Resources, pp. 1–137.

Kooistra, L., Sanchez-Prieto, L., Bartholomeus, H.M., Schaepman, M.E. (2007). Regional mapping of plant functional types in river floodplain ecosystems using airborne imaging spectroscopy data. In: Proc. 10th International Symposium on Physical Measurements and Spectral Signatures in Remote Sensing (ISPMSRS), Davos, 12–14 March, pp. 1–6.

Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sensing of Environment 100 (3), 356–362.

Lees, B.G., Ritman, K., 1991. Decision tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environmental Management 15 (6), 823–831.

Lucas, R., Lee, A., Armston, J., Breyer, J., Bunting, P., Carreiras, J., 2008. Advances in forest characterisation, mapping and monitoring through integration of LiDAR and other remote sensing datasets. SilviLaser 4 (1), 2–12.

Meyer, K.M., Wiegand, K., Ward, D., Moustakas, A., 2007. The rhythm of savanna patch dynamics. Journal of Ecology 95 (6), 1306–1315.

Mucina, L., Rutherford, M.C. (Eds.), 2006. The vegetation of South Africa, Lesotho and Swaziland. South African National Biodiversity Institute, Pretoria.

Mutanga, O., Skidmore, A.K., 2003. Continuum-removed absorption features estimate tropical savanna grass quality in situ. In: Proc. 3rd EARSEL Workshop on Imaging Spectroscopy, Herrsching, 13–16 May, pp. 542–558.

Odagawa, S., Okada, K., 2009. Tree species discrimination using continuum removed airborne hyperspectral data. In: Proc. First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Grenoble, 26–28 August, pp. 1–4.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9 (2), 181–199.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the great plains with ERTS. In: Proc. Third ERTS Symposium, NASA SP-351 I, pp. 309–317.

Rutherford, M.C., Westfall, R.H., 1986. Biomes of Southern Africa – an objective categorization. Memoirs of the Botanical Survey of South Africa 54 (1), 1–98.

Salford Systems, 2004. Random Forests™ – An implementation of Leo Breiman's RF. Random Forest software help document – Version 6.6, 1–161.

Schmidt, E., Lotter, M., McCleland, W., 2007. Trees and Shrubs of Mpumalanga and Kruger National Park, second ed. Jacana Media, Johannesburg, South Africa.

Shackleton, C.M., 2000. Comparison of plant diversity in protected and communal lands in Bushbuckridge lowveld savanna, South Africa. Biological Conservation 94 (3), 273–285.

Shackleton, C.M., Shackleton, S., 2003. Value of non-timber forest products and rural safety nets in South Africa. In: Proc. International Conference on Rural Livelihoods, Forests and Biodiversity, Bonn, Germany, 19–23 May, pp. 1–18.

Shackleton, C.M., Guthrie, G., Main, R., 2005. Estimating the potential role of commercial over-harvesting in resource viability: a case study of five useful tree species in South Africa. Land degradation and Development 16 (3), 273–286.

Sobhan, I., 2007. Species discrimination from a hyperspectral perspective. Mapping shrub and tree species richness from hyperspectral imagery using a matched filtering unmixing technique. Doctoral Thesis, International Institute for Geo-information Science and Earth Observation, Enschede, the Netherlands, pp. 103–124 (Chapter 6).

Tong, Q., Zhang, B., Zheng, L., 2004. Hyperspectral remote sensing technology and applications in China. In: Proc. 2nd CHRIS/Proba Workshop, ESA/ESRIN, Frascati, Italy, 28–30 April, pp. 1–10.

Van der Meer, F., 2006. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. International Journal of Applied Earth Observation and Geoinformation 8 (1), 3–17.

Wessels, K.J., Mathieu, R., Erasmus, B.F.N., Asner, G.P., Smit, I.P.J., van Aardt, J., Main, R., Fisher, J., Marais, W., Kennedy-Bowdoin, T., Knapp, D.E., Emerson, R., Jacobson, J., 2011. Impact of contrasting land use on woody vegetation structure in the Lowveld savannas of South Africa. Forest Ecology and Management 261 (1), 19–29.

Yingchun, S., Xianfeng, Z., Xiuwan, C., Zhaoqiang, H., Caicong, W., 2006. Mangrove type classification using airborne hyperspectral images at Futian reservation, Shenzhen, China. In: Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Denver, 31 July–4 August, pp. 3451–3454.