Elsevier Editorial System(tm) for ISPRS Journal of Photogrammetry and Remote Sensing
Manuscript Draft

Manuscript Number:

Title: 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

Article Type: Full Article

Keywords: Vegetation; Mapping; Classification; LIDAR; Hyper spectral

Corresponding Author: Mr Laven Naidoo, MSc

Corresponding Author's Institution: Council for Scientific and Industrial Research (CSIR)

First Author: Laven Naidoo, MSc

Order of Authors: Laven Naidoo, MSc; Moses A Cho, PhD; Renaud Mathieu, PhD; Gregory P Asner, PhD

Laven Naidoo
CSIR – NRE
P.O.Box 395
Pretoria
0001
South Africa

Prof. M.G. Vosselman
Editor-in-Chief
Faculty of Geoinformation and Earth Observation,
University of Twente,
P.O. Box 6,
7500 AE Enschede,
Netherlands

Re: Submission of Manuscript for peer-review

Dear Sir

I would like to submit a full paper entitled "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" to the *ISPRS Journal of Photogrammetry and Remote Sensing*. The 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 classification of savanna tree species at high accuracies can benefit both communal and protected land management by providing accurate means of monitoring economically useful tree species and problematic alien species. Due to the limited literature available on savanna tree species classification and due to its potential benefits, it is the author's view that this study would benefit the remote sensing research and managerial communities. The results indicated which particular spectral and structural predictors used in the Random Forest models contributed most to and which model, itself, yielded the highest classification accuracies in classifying the eight target tree species.

I hope you find that this paper meets the required standards for publication.

Contact details of the corresponding Author, Laven Naidoo:
Email lnaidoo@csir.co.za
Address: *same as above*
Tel: +27 12 841 2233 or
Fax: +27 12 841 3909

Kind regards
Laven Naidoo

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

NAIDOO L.[1], CHO M.A.[1], MATHIEU R.[1] & ASNER G.[2]

[1]Council for Scientific and Industrial Research (CSIR), Natural Resources and the Environment, P.O. Box 395, Pretoria, 0001, South Africa.
lnaidoo@csir.co.za, mcho@csir.co.za, rmathieu@csir.co.za

[2]Carnegie Institution for Science, Stanford, CA, USA
gpa@stanford.edu

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 (466nm) 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

[1] The corresponding author for this journal article is L. Naidoo - ph: (012) 841 2233, fax: (012) 841 3909

prediction success for the eight savanna tree species with an overall classification accuracy of 87.68% and KHAT value of 0.843.

Key words: *savanna tree species, spectral variability, tree height, Random Forest, predictor datasets*

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, In. press) or mangrove types (Yingchin *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 & Cho *et al.*, 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 as non-timber forest products (NTFPs) for the local community enterprises in the communal rangelands who 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 and alien invasive species 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 intraspecies spectral variability resulting from genetic patrimony and various environmental and physical factors (weather, seasonality, geology and edaphic conditions; such as the influence of gabbro versus granite substrates on savanna vegetation; and natural phenological changes such as deciduous versus evergreen species during the savanna dry season – Hestir *et al.*, 2008, Lees & Ritman, 1991 and Tong *et al.*, 2004), the spectral resolution of a sensor must be high with numerous, contiguous bands along with a high canopy-scale spatial resolution. 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 using the spectral band configuration of Worldview-2 (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).

Although the use of spectra alone provided good results in these studies, it is evident from various structural remote sensing studies (Kim, 2007; Bork & Su, 2007; Geerling *et al.*, In. press and 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 & 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.* (In press) combined image spectroscopy and LiDAR data, by data fusion at the pixel level, to improve the classification of floodplain vegetation types. 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 simplest structural variable, the tree height, can be clearly illustrated in the histogram graph in figure 1 which describes the distribution of various savanna tree species height values (from sampled field data). The trend illustrates that certain species share distinct height ranges to that of other species. *Acacia gerrardi* and *Dichrostachys cinerea*, for example, both possessed a tree height range between 0 and 6m while *Berchemia discolor* possessed a distinctly taller range between 8 and 12m. This distinguishable difference in the different species' height ranges could clearly help potential classification opportunities.

Insert Figure 1

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 and Cho *et al.*, 2010), while taking advantage of the significant inter-species structural differences. These requirements can be met by the implementation of a Decision Tree approach, with the most commonly used approach being the Classification and Regression Trees (CART). 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 (Cartijo & 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 feat especially in highly, spectrally variable environments.

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 the 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 customisation that is required. It is also difficult to determine the optimal K value for the KNN classifier (Joseph, 2005).

The emergence of the Random Forest (RF) approach was seen as an improvement over the CART approach as concepts such as multiple (100's) 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), the mapping of 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 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 *Acacia gerrardii / Dichrostachys cinerea (AG/DC), Acacia nigrescens (AN)*, *Berchemia discolor (BD)*, *Combretum species (COM)*, *Pterocarpus rotundifolius (PR)*, *Spirostachys africana (SA)*, *Sclerocarya birrea (SB)* and *Terminalia sericea (TS).* Based on the assumption that tree height is an important addition to the classification dogma of savanna tree species, the objective of this study was to investigate the influence of tree height on savanna tree species level classification beyond the impact of spectra alone. 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 savanna tree species classification success?
- Which Random Forest model yielded the highest accuracy results for classifying the 8 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 & Westfall, 1986). Regionally, savannas have a long dry winter and a wet summer with an annual precipitation varying between 235 and 1000mm. 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 & Westfall, 1986).

The study area under investigation (figure 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 by the livestock ranching, harvesting and farming activities of neighbouring informal communities. The Sabi Sands Wildtuin is approximately 54 000 hectares 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 hectares in area 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 & Rutherford (Eds.)

(2006). The terrain in both study regions is gently undulating with catena geomorphological 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 & Rutherford, Eds. 2006). In the west, near the Drakensberg escarpment, the mean annual rainfall is approximately 1200mm and decreases to 550mm 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.

Insert Figure 2

2.2 Hyperspectral, LiDAR, and field datasets

At the end of May 2008 an integrated hyperspectral and LiDAR dataset was acquired for 35000ha over the study area (figure 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 nm to 1054.3 nm, bandwidth) and ii) raw LiDAR point clouds consisting of up to four ranges or returns per laser shot (at least one 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 5LiBatch (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 3-D location of the laser returns. 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 also encountered in the field 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 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 8 common savanna tree species found in the L456 study area. These species were *Acacia gerrardii / Dichrostachys cinerea (AG/DC), Acacia nigrescens (AN)*, *Berchemia discolor (BD)*, *Combretum species (COM)*, *Pterocarpus rotundifolius (PR)*, *Spirostachys africana (SA)*, *Sclerocarya birrea (SB)* and *Terminalia sericea (TS)*. Species such as *Combretum apiculatum*, *Combretum collinum* and *Combretum hereroense* were grouped together in the *Combretum* species class while *Acacia gerrardii* and *Dichrostachys 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 8 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 the table 2. In table 2, it is important to note a particular anomalous value for the mean height of *Pterocarpus rotundifolius* (0.126m) which was attributed to the limitation of the LiDAR sensor in detecting this small tree species.

Insert Table 1

Insert Table 2

2.4 Random Forest Predictor Datasets

An ensemble of seven datasets of predictors, which incorporated the tree species' spectral and/or structural data, was investigated individually in the Random Forest modelling procedure to ascertain which variable(s) drive or enhance the classification and differentiability of the target tree species. These seven main predictor datasets; including their descriptions, wavelengths and associated references are summarised in table 3. The various predictors were chosen for various reasons. Apart from investigating the importance of tree species height in this study, two particular predictor datasets (Indices and Nutrient and Leaf Mass) were considered, which made use of particular spectral vegetation indices and spectral bands (table 3) to best exploit the primary and secondary plant chemical compound differences in the savanna vegetation. Since the CAO hyperspectral imagery were taken during a dry rainfall period of May 2008, these differences could be significant both within and between different tree species. Selected bands from a Spectral Angle Mapper (SAM) approach, previously applied by Cho *et al.* (2010), were also modelled as a separate predictor dataset. 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 maximises inter-species SAM and starts off by selecting the two bands which have the highest average SAM, among all pair-wise combinations. It then adds the next consecutive important bands until no significant bands are left (Cho et al, 2010). These selected bands were found to improve savanna tree species discrimination 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 raw dataset was then subjected to a continuum removed transformation to create a new predictor dataset. As for the SAM approach 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 & Okada, In. press). This transformation would also contribute to minimize any effects arising from any possible Bi-directional Reflectance Distribution Function (BRDF) effect in the imagery. Finally, the most important predictors were identified within each datasets of predictors and combined in a hybrid approach in an attempt to improve overall classification results.

Insert Table 3

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 *Acacia nigrescens* and *Sclerocarya birrea* contain much larger sample sizes than for instances *Berchemia 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, 2006).

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 randomly selected samples, which would be excluded from each bootstrapped sample in the random forest construction, would be reserved as an internal test dataset for the Random Forest model validation (Ismail *et al.*, 2010). 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.

Under the Random Forest summary reports 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 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 score of 80 and greater was chosen as the authors' 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. A confusion matrix was created while overall, and species specific user's and producer's accuracies were computed. The producer's accuracy indicates the percentage of spectra for each species class that have been correctly classified while the user's accuracy indicates the probability that a spectra classified into a given species class actually represents that class on the ground (Baldi and Paruelo, 2008). The confusion matrix was created by comparing the modelled data against the internal test OOB sample data.

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:

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

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).

Finally, a hybrid dataset of predictors was 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 types separately. These important predictors which created the hybrid dataset are displayed in table 4 in the results section.

3. Results

3.1 Predictor Importance

Insert Table 4

From the Gini Index Score results in table 4, 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 contributed the most to the classification prediction success when all the different Random Forest models were executed. From the modelled results of the hybrid dataset, the tree species height predictor was by far the most valuable predictor (Gini Index score of 100) in contributing to the classification prediction success. The second most significant predictor was the continuum removed transformed band 30 (658.8nm) which only achieved a Gini Index score of 65.84.

3.2 Modelled Prediction Success

Insert Table 5

The summary results in table 5 illustrate the classification accuracies for the Random Forest models of the different predictor datasets and the encompassing hybrid dataset. In this table, the overall classification accuracies and the KHAT statistics remain mostly comparable to one another for the different models. Amongst the seven separate 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) than 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 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 classification accuracy results with an overall classification of 87.68%, a KHAT of 0.843 and only 495 misclassified pixels.

Insert Table 6

The confusion matrix resulting from the hybrid dataset modelling is presented in table 6. All 8 tree species classes were classified at a very high producer's accuracies with the lowest being 78.27% for *Combretum species*. *Terminalia 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. *Berchemia discolor* was the most problematic species, in the dataset, with the lowest user's accuracy of 35.29% (confusion with *Spirostachys africana* and *Sclerocarya birrea*) which starkly contrasted with its 94.74% producer's accuracy. *Sclerocarya birrea* was the most largely represented species class on the ground (97.91%). The remaining species displayed moderate (> 60%) to high (> 80%) user's accuracies.

4. Discussion

From the Gini Index variable importance results (table 4), the significant 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 to 1054.3nm), 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 conclusive that NDVI was scored as the most important vegetation index by the Gini value (100). However, in the context of the spectral indices and height dataset results, it was the tree height predictor which was considered more important, in the classification model, than any of the spectral indices. This trend is further supported in the hybrid dataset, which is a combination of all the significant predictors from all the modelled dataset results. In the hybrid dataset, the tree height predictor also was the most important predictor (100 Gini Index score) in the study, which was followed by the continuum removed transformed band 30 with a Gini Index score of 65.84. The significant difference in Gini Index score between the highest and the second highest scoring predictors could illustrate a sense of dominance of the tree height predictor over other spectral predictors in the Random Forest classification process. However, the inclusion of these spectral predictors (particularly CRT band 30), although low in Gini Index score and significance, largely contributed to the overall success of the model and prevented the model from obtaining much poorer results as was the case when tree height alone was implemented as a single predictor dataset (31.90% overall accuracy).

From the classification results (table 5); the vegetation indices and tree height combined dataset (Ht + Indices) yielded the highest classification results (82.38%; KHAT of 0.776) when compared to the remaining 6 separate predictor datasets. When the tree height predictor was combined with the most significant spectral predictors from the separate predictor datasets (NDVI, chlorophyll *b* wavelength and selected raw, CRT and SAM wavelengths) into a hybrid dataset, the highest classification accuracy and prediction success results (87.68%; KHAT of 0.843) were achieved in this study. It is clear from both these results that the incorporation of spectral information and structural information proved to be more useful in species level classification than the use of 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. In fact, it should be noted that if the Gini values indicate that the most important predictor is the tree height (this predictor always has the highest index when used in one dataset) the classification results show that the spectral data host the most important information, but these are significantly improved by the addition of this structural parameter. From corresponding confusion matrix results (table 6), majority of the species obtained producer's and user's accuracies that ranged from reasonable (>60%) to excellent (>90%) with *Berchemia discolor* being the only exception. Although achieving a producer's accuracy greater than 90%, the user's accuracy was dismally low (approximately 35%). The plausible reason for the poor representation of this species class at ground level could be due to the lack of a sufficient number of sampled tree canopies and related pixels (only 3 canopies containing 57 pixels were sampled in the field) needed for the Random Forest classification. An increase in the sampled data for *Berchemia discolor* most likely would improve the species' currently low user's accuracy. Besides the *Berchemia discolor* species class, the remaining seven savanna tree species would produce very reliable and accurate species distribution maps which would prove invaluable for both communal and protected savanna rangeland management practices. Overall, these results exceeded the authors' expectations with overall classification accuracies exceeding those achieved in previous tree species classification efforts in South African savannas (Cho *et al.*, 2010 and Cho *et al.*, 2011) and in other related ecosystems such as the shrubby American rangelands (Lawrence et al., 2006). The limited existence of other savannas tree species mapping studies, in the academic literature, makes it difficult to place these results in suitable context but will, hopefully, encourage the emergence of other future studies.

Despite the success of the modelled 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 to the minimal number of user-defined model settings (Prasad *et al.*, 2006). As a result, implementing the optimal decision tree design in remote sensing mapping software (e.g. ENVI) would be very challenging for the user wishing in putting this classification model into practice. Investigating alternative scripting and programming related approaches could circumvent this issue but this is beyond the scope of this study. Classification accuracies, although very good for the hybrid dataset, could be improved by implementing the probability cut-off adaptation (bias adjustment) approach which improves the cross-validated error rate for unbalanced datasets, as implemented and proven successful in Dahinden (2006) and Grossmann *et al.* (2010). Also instead of the traditional Gini Index variable importance measure, other successfully implemented 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 implemented for possibly improved results. These alternative variable selection methods could prove effective especially when dealing with datasets which have many explanatory variables that have very similar importance measures (Jiang et al., 2004 cited in Ismail et al., 2010). The incorporation of other more complex LiDAR-derived structural parameters in the modelling process, such as for instance canopy volume, canopy height and tree fractional cover obtained by a higher resolution waveform footprint, can be investigated further.

5. Conclusions

By readdressing the scientific questions, posed in the introduction of this study, it can be concluded that the hybrid dataset Random Forest model yielded the highest classification accuracy and prediction success for the 8 savanna tree species 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 different classification models and contributed to the success of the hybrid dataset model when combined, were species tree height; NDVI; the chlorophyll *b* wavelength (466nm) and a selection of raw, continuum removed and SAM bands (see table 4 for the entire list of significant predictors). However, according to the Gini Index variable importance and classification results, it was clear that LiDAR-derived tree species height was the most dominant and influential predictor in ensuring classification success but since on its own it yielded the lowest overall classification results, it can only be concluded that tree height significantly improves savanna tree species level classification accuracies only when combined with other significant spectral predictors.

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.

References

- Affendi. S, N.A Ainuddin & H.Z.M Shafri (2009). A rule based approach for the mapping of tropical forest canopy from airborne hyperspectral data. *GIS development.net Online journal*; pp 1-13

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

- Asner. G.P, D.E Knapp, T. Kennedy-Bowdoin, M.O Jones, R.E Martin, J. Boardman & R.F Hughes (2008). Invasive species detection in Hawaiian Rainforests using In-flight fusion of Airborne Imaging Spectroscopy & LiDAR. Carnegie Institution: Stanford University; pp 1-33

- Baldi. G & J.M Paruelo (2008). Land use and land cover dynamics in South American, temperate grasslands. *Ecology and Society*. Vol 13; Issue 2; Article 6; pp 1-20

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

- Bork. E.W & J.G Su (2007). Integrating LiDAR data and multispectral imagery for enhanced classification of rangeland vegetation: A meta analysis. *Remote Sensing of Environment*. Vol 111; pp 11-24

- Cho. M.A, P. Debba, R. Mathieu, J. van Aardt, G.P Asner, L. Naidoo, R. Main, A. Ramoelo & B. Majeke (2009). Spectral variability within species and its effects on savanna tree species discrimination. *IEEE International Geoscience & Remote Sensing Symposium (IGARSS) 2009 Conference Paper*, University of Cape Town, July 2009

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

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

- Cho. M.A, A. Skidmore, F. Corsi, S.E van Wieren & I. Sobhan (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*. Vol 9; pp 414-424

- Carletta. J (1996). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*. Vol 22; Issue 2; pp 1-6

- Congalton. R.G (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*. Vol 37; pp 35-46

- Dahinden. C (2006). An improved random forest approach with application to the performance prediction challenge datasets. *Seminar für Statistik, CH-8092, Zürich, Switzerland*; pp 1-6

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

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

- Geerling. G.W, M. Labrador-Garcia, J.G.P.W Clevers, A.M.J Ragas & A.J.M Smits (In. press). Classification of floodplain vegetation by data-fusion of Spectral (CASI) and LiDAR data. *International Journal of Remote Sensing*. Research Article; pp 1-24

- Gitelson. A.A & M.N. Merzlyak (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*. Vol 143; pp 286-292.

- Gitelson. A.A, Y. Zur, O.B. Chivkunova & M.N. Merzlyak (2002). Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy. *Photochemistry and Photobiology*. Vol 75; pp 272-281.

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

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

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

- Joubert. S (2007). The Kruger National Park – A history (Volumes I, II & III), 1st edition; *High Branching (Pty) Ltd*, Johannesburg: South Africa.

- 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, L. Sanchez-Prieto, H.M Bartholomeus, M.E Schaepman (In. Press). Regional mapping of plant functional types in river floodplain ecosystems using airborne imaging spectroscopy data. *Commission VII*; pp1-6

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

- Lees. B.G & K. Ritman (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*. Vol 15; Issue 6; pp 823-831

- Meyer. K.M, K. Wiegand, D. Ward & A. Moustakas (2007). The rhythm of savanna patch dynamics. *Journal of Ecology*. Issue 95; pp 1306-1315

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

- Mutanga. O & A.K Skidmore (2003). Continuum-removed absorption features estimate tropical savanna grass quality in situ. *3rd EARSEL Workshop on Imaging Spectroscopy, Herrsching, 13-16 May 2003*; pp 542-558

- Odagawa. S & K. Okada (In. press). Tree species discrimination using continuum removed airborne hyperspectral data. Earth Remote Sensing Data Analysis Center. Preliminary draft; pp 1-4

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

- Rouse. J.W., R.H. Haas, J.A. Schell, and D.W. Deering (1973). Monitoring Vegetation Systems in the Great Plains with ERTS. Third ERTS Symposium, NASA SP-351 I; pp 309-317.

- Rutherford. M.C & R.H Westfall (1986). Biomes of Southern Africa – an objective categorization. *Memoirs of the Botanical Survey of South Africa*. Vol 54; pp 1-98

- Salford Systems (2004). Random Forests™ - An implementation of Leo Breiman's RF. *Random Forest software help document*; pp 1-161

- Schmidt. E, M. Lotter & W. McCleland (2007). Trees and shrubs of Mpumalanga and Kruger National Park (2nd Edition). *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.* Issue 94; pp 273-285.

- Shackleton. C.M & S. Shackleton (2003). Value of non-timber forest products and rural safety nets in South Africa. *Paper presented at the International Conference on Rural Livelihoods, Forests and Biodiversity*; 19-23 May, Bonn, Germany; pp 1-18

- Shackleton. C.M, G. Guthrie & R. Main (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 & Development.* Vol 16; pp 273-286

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

- Tong. Q, B. Zhang & L. Zheng (2004). Hyperspectral remote sensing technology and applications in China. *Proceedings of the 2nd CHRIS/Proba Workshop, ESA/ESRIN*, Frascati, Italy, 28-30 April; pp 1-10

- Yingchun. S, Z. Xianfeng, C. Xiuwan, H. Zhaoqiang & W. Caicong (2006). Mangrove type classification using airborne hyperspectral images at Futian reservation, Shenzhen, China.

Geoscience and Remote Sensing Symposium 2006, IGARSS 2006, IEEE International Conference; pp 3451-3454

Figure 1: Histogram of the common tree species' height distribution obtained from a selected field sample

Figure 2: Study area map of the Greater Kruger National Park with focus on the L456 study region

Figure 1

Click here to download Figure(s): Figure1.docx

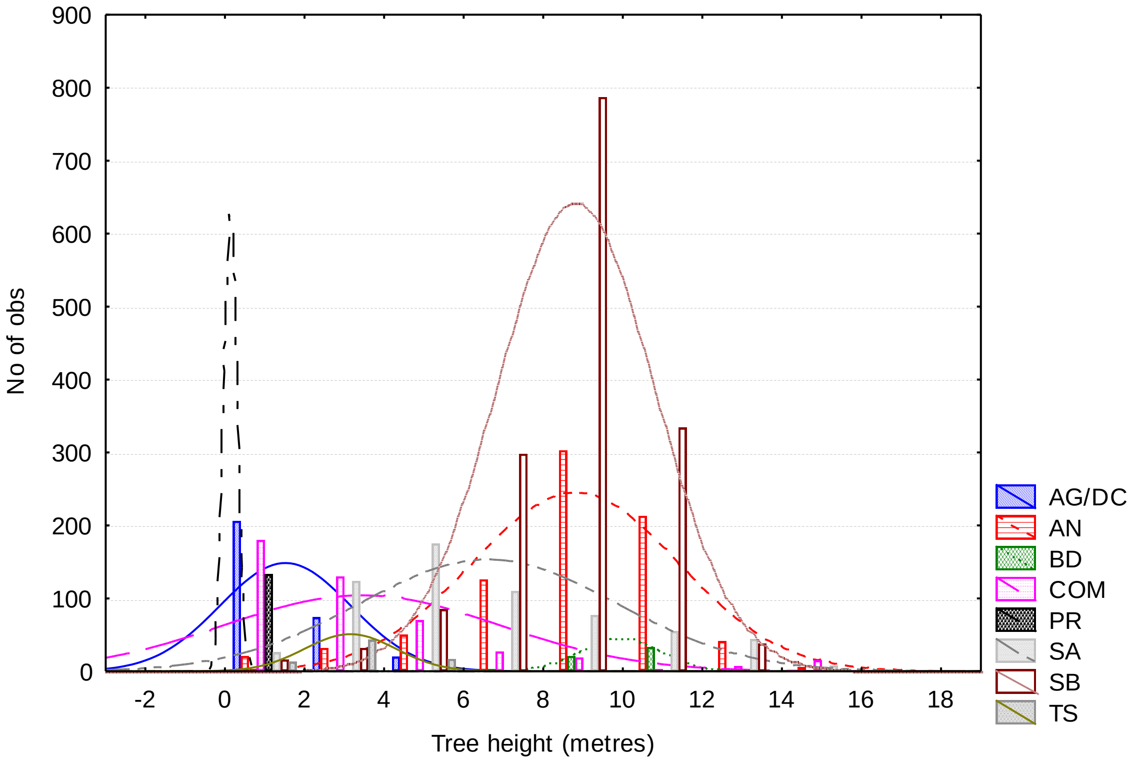

Figure 2

Click here to download Figure(s): Figure2.docx

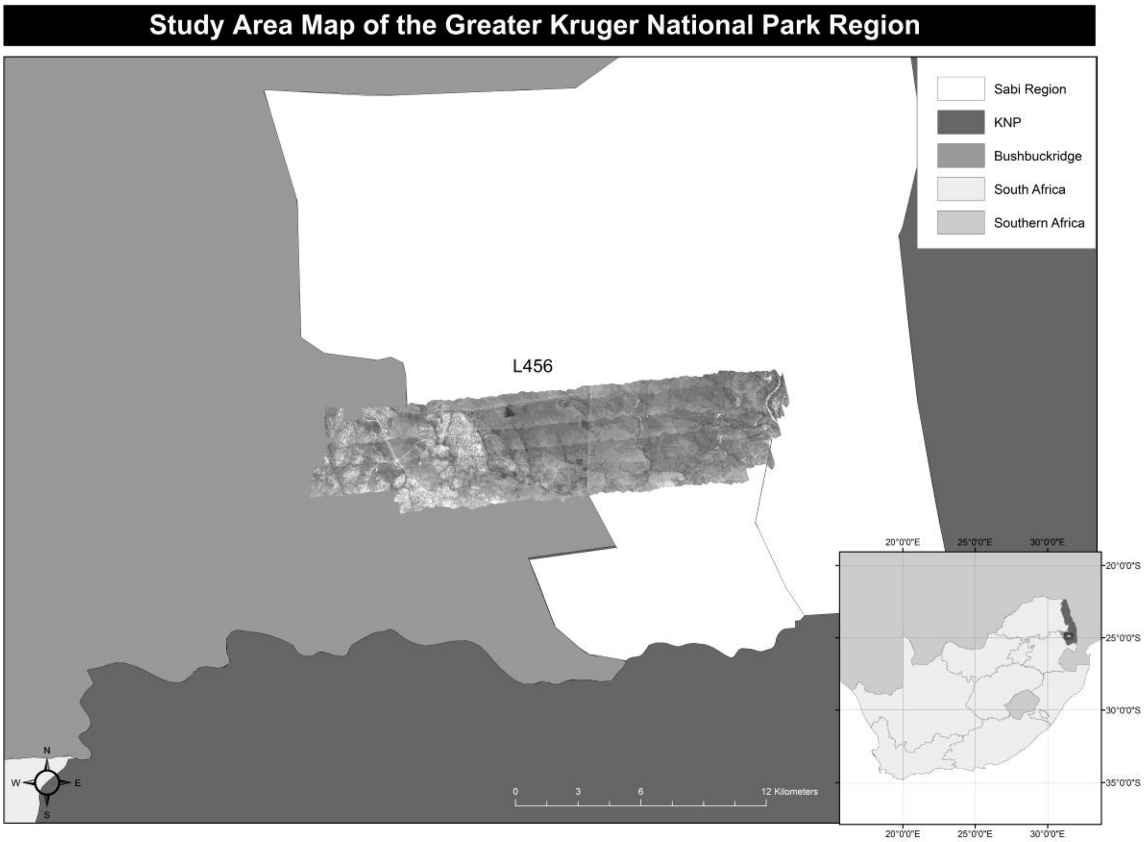

Table 1
Click here to download Table(s): Table1.docx

Table 1: Attribute information 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 30m. 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 20m. 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 20m). 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 & Shackleton (2003) and Shackleton et al. (2005)

Table 2: Total number of recorded canopies, tree pixels sampled, and tree heigh statictics (from LiDAR) of the tree sample

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

Table 3

Table 3: Seven Predictor datasets that were modelled in RF including their description, formulae or wavelengths used, and 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)* *Photochemical Reflectance Index (PRI)* *Normalized Difference Vegetation Index (NDVI)* *Red Edge NDVI (RE)*	 λ800(1/λ520 - 1/λ550) (λ531 - λ570)/(λ531 + λ570) (λ800 – λ678)/(λ800.5 + λ678) (λ750 - λ705)/(λ750 + λ705)	Gitelson *et al.* (2002) Gamon *et al.* (1992) Rouse *et al.* (1973) Gitelson *et al.* (1994)
Height + Indices	Tree species' height data and Vegetation Spectral Indices (CRI, PRI, NDVI & 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 & Skidmore (2003)
Spectral Angle	Spectrally significant bands (31 bands) selected from	706; 762.7; 696.6; 668.2; 677.7; 687.1	Cho *et al.* (2010)

Mapper (SAM) Selected Bands	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	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	
Nutrient & 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 & LAI)* 786 *(Leaf mass)* 846 *(Leaf mass, LAI, chlorophyll)*	Cho *et al.* (2007)

Table 4
Click here to download Table(s): Table4.docx

Table 4: The Gini Index Score summary table and the most significant predictors in each predictor dataset (score of >80*)

Predictor Dataset	Important Variables/Predictors	Gini Index Score
Height	Height	100
Indices	NDVI	100
Height + Indices	Height	100
	NDVI	84.06
Raw Bands	B8 (450.9nm)	100
	B35 (706nm)	97.35
	B9 (460.3nm)	91.93
	B10 (469.7nm)	90.82
	B11 (479.2nm)	89.3
	B7 (441.4nm)	87.36
	B14 (507.5nm)	86.54
	B6 (432nm)	82.1
CRT Bands	B30 (658.8nm)	100
	B32 (677.7nm)	99.95
	B31 (668.2nm)	96.66
	B10 (469.7nm)	94.29
	B33 (687.1nm)	92.92
	B12 (488.6nm)	89.07
	B39 (743.8nm)	88.61
	B29 (649.3nm)	86.15
	B11 (479.2m)	82.91
SAM Bands	B10 (706nm)	100
	B4 (413.1nm)	95.91
	B5 (422.6nm)	92.15
	B6 (668.2nm)	87.79
	B7 (677.7nm)	83.34
N+LM Bands	B1 (466nm)	100
Hybrid	**Height**	**100**
	CRT Band 30 (658.8nm)	65.84

*B = Bands; * With exception to the hybrid predictor dataset*

Table 5

Click here to download Table(s): Table5.docx

Table 5: Modelled prediction success summarized results for all predictor datasets

Predictor Dataset	Classification Accuracy (%)	KHAT Statistic*	Pixels Misclassified
Ht	31.90	0.1861	2737
Indices	67.85	0.6118	1292
Ht + Indices	82.38	0.776	708
Raw Bands	80.29	0.7547	792
CRT Bands	78.10	0.7287	880
SAM Bands	75.47	0.699	986
N+LM Bands	73.40	0.6746	1069
Hybrid	**87.68**	**0.8425**	**495**

** Cohen's Unweighted Kappa (Cohen, 1960 cited in Congalton, 1991)*

Table 6
Click here to download Table(s): Table6.docx

Table 6: Confusion matrix displaying the classification accuracies obtained by the hybrid dataset RF modelling

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