Identification of Optimal Field Spectral Measurements Wavebands for Discriminating among Spatial Features in support of Mapping Using Hyper-spectral Imagery

Mwitwa Chilufya[1], Mulemwa Akombelwa[2] , Derek Stretch[3]

[1,2] University of KwaZulu-Natal, School of Engineering, Land Surveying Programme, Durban, South Africa; [1]Chilufyam@ukzn.ac.za; [2] akombelwa@ukzn.ac.za

[3]University of KwaZulu-Natal, School of Engineering, Civil Engineering Programme, Durban, South Africa

This paper has been through a process of double-blind peer review

Abstract

Field spectral measurements serve as ground truthing data necessary for classification of hyper-spectral imagery. For this purpose, optimal wavelengths capable of discriminating all or the majority of features sampled in the field need to be identified prior to hyper-spectral imagery classification. This paper discusses the identification of optimal wavebands from field spectral measurements for the purpose of mapping spatial features using hyper-spectral imagery. Spectral Measurements of several vegetation assemblages in the Mfabeni Wetland of the Isimangaliso Wetland Park were collected at three different sites using an Analytical Spectral Devices (ASD) spectralradiometer, then pre-processed to a format suitable for processing using Random Forest (RF), an open-source machine learning software package that runs on the statistical package R. The pre-processed data were normalized to Hyperion imagery spatial resolution, and then used to develop an optimal model that identifies the most suitable wavebands for discriminating pixels representing the vegetation assemblages collected at one of the three sites. The resultant model was then used to predict pixels on the other two sites with similar spectral characteristics. Results show that RF can be used to develop a model for discriminating optimal wavebands for mapping vegetation assemblages at a pixel level and that the developed model can be used to predict pixels with similar spectral characteristics in an independent dataset of spectral measurements. This approach can, by implication, be used for any given data set of spectral measurements, thus allowing for mapping of any combination of spatial features using hyper-spectral imagery.

1. Introduction

Mapping of terrestrial spatial features has conventionally been accomplished using ground based survey techniques, inclusive of Global Positioning Systems (GPS). These techniques however, have been found to be laborious, and in some instances less accurate in comparison to non-ground based techniques. In a study to map stream microhabitats using high spatial resolution hyper-spectral imagery, Macus (2001), established that airborne high spatial resolution hyper-spectral imagery

appeared to have mapped the stream habitats with greater accuracy than the ground-based surveys, in many instances, thus challenging classical mapping approaches.

Various approaches have therefore been devised over time to address the shortcomings of classical mapping techniques. One recent development is in the design of hyper-spectral sensors capable of capturing images in hundreds of spectral bands resulting in hyper-spectral images. Interpretation and analysis of hyper-spectral imagery, however, requires specialized techniques to facilitate extraction of details from such imagery. In this paper one such techniques is discussed, namely; the identification of optimal wavebands from field spectral measurements.

In order to use hyper-spectral imagery for mapping spatial features, ground truthing data in form of field spectral measurements, among others, is required. Once collected, spectral measurements need to be processed in order to identify suitable wavelength for discriminating sampled spatial features. The identified wavelengths can then subsequently be used to map the sampled spatial features by classifying the corresponding hyper-spectral imagery. Identifying such suitable wavelengths using variable selection methods without losing any important information is therefore, a pre-requisite in hyper-spectral remote sensing application (Adam and Mutanga, 2009; Bajcsy and Groves, 2004; Vaiphasa et al., 2007). A methodology for identifying such suitable wavelengths from a sample of wetland vegetation assemblages spectral reflectance measurements, using the machine learning software package 'Random Forest' (RF) running on R, is therefore presented in this paper.

Different statistical techniques inclusive of discriminant analysis, canonical variate analysis, classification trees, support vector machines and principal component analysis have been used (Adam and Mutanga, 2009; Mutanga and Skidmore, 2004) to identify optimal wavelengths for discriminating plant species and vegetation assemblages with varying success. RF algorithm has become a latest success in variable selection and classification algorithm for hyperspectral data (Breiman, 2001; Lawrence et al., 2006). RF provides both accuracy and variable importance information with the results. In a research conducted to discriminate indicator grass species for rangeland degradation assessment using hyperspectral data resampled to AISA Eagle resolution, Mansour et al. (2012) acknowledge the robustness and accuracy of RF, both for variables selection and classification of hyperspectral data.

RF is a tree ensemble algorithm that uses bagging, i.e., bootstrap aggregation, ensemble procedure to build multiple individual decision trees provided to be diverse by the use of random samples derived from the training data set (Mansour et al., 2012; Breiman, 2001). It describes an approach to building models where the actual model builder could be a decision tree algorithm, a regression algorithm, or any one of many other kinds of model building algorithms. Williams (2001) compares the performance of an ensemble of trees to bringing together panels of experts from various institutions such as government, industry and universities to ponder over an issue so as to come up with a consensus decision which none of these panels working in isolation, cannot

achieve. The RF algorithm, builds hundreds of decision trees to their maximum depth, without pruning, then combines them into a single model that reduces the instability observable when single decision trees are built (Williams, 2001). The final decision of the ensemble will be the decision of the majority of the constituent trees. In regression, the result is the average value over the ensemble of regression trees. Random forest is well suited to handling a diversity of data sets and often only requires little pre-processing (Williams, 2011).

According to William (2011) deployment of Random forest involves the following steps: (1) selection of random different subset of training data (known as a'bag)' to train each tree making up the ensemble, (2) use of two-thirds of the training data to train each tree with the remaining one-third (known as 'Out Of Bag' [OOB] being used to estimate error and variable importance, and (3) assigning the number of votes from all the trees to classes or for regression, the average of the results.

The objective of this study was to identify the optimal wavelengths suitable for discriminating spatial features from a set of field spectral measurements for the purpose of spatial mapping by classifying hyper-spectral imagery of the area from which the field measurements are taken using the random forest algorithm.

2. Materials and Methods

2.1. Study Area

The study area for this research is the Mfabeni Wetland located in the Isimangaliso Wetland Park, one of the World Heritage Sites situated in the north of the Province KwaZulu-Natal in South Africa (Fig. 1). Three different sub-wetlands of the wetlands park were identified with the help of Google Earth as sites for data collection (Fig. 1). It was, however, established while on site that each wetland had different combinations of vegetation assemblages.

2.2. Data Collection

Spectral reflectance, spatial locations and digital images of randomly selected vegetation assemblages were taken in a single transect across each sub-wetland. Spectral reflectance measurements were taken using an Analytical Spectral Devices (ASD) FieldSpec ® 3 Spectralradiometer. The spectral range of the FieldSpec ® 3 is 350–2500nm with a resolution of 1.4nm in the 350–1000nm range and 2.0nm for the spectral region 1000–2500nm (Analytical Spectral Devices (ASD), 2005). Spectral reflectance measurements were taken at nadir using 1°, 8° and 25° fields of view in different situations at a height of about half a meter. The short height was due to the grass being high and the absence of a hoisting facility. The instrument was configured to make 20 repeated measurements, then record an average for each measurement of the reflectance spectra stored. Repeat measurements ranging from 2 to 4 where made for every identifiable vegetation assemblage.

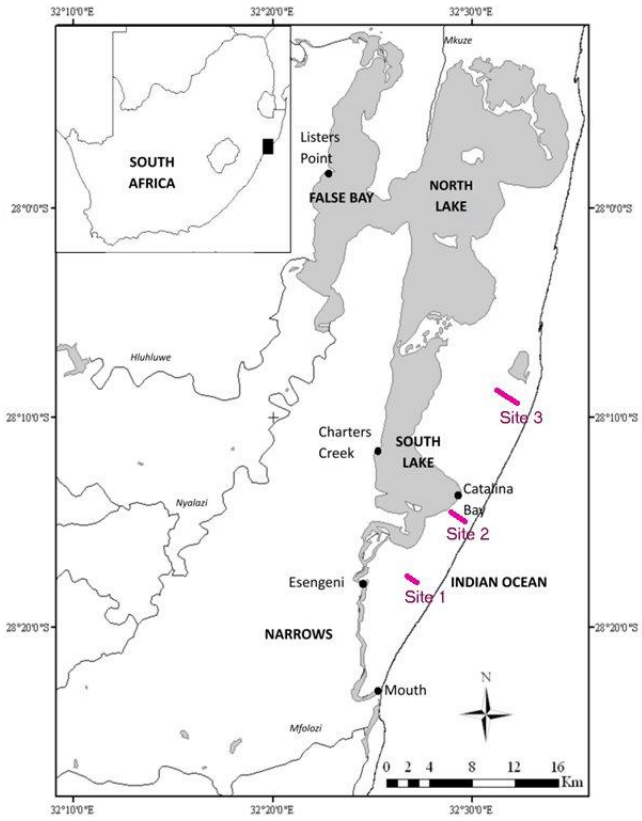

Figure 1: Location of the Study Area in the Isimangaliso Wetland Park showing data collection sites 1 to 3 in the Mfabeni Wetland (Map modified with permission of Carrasco)

Measurements were taken according to the procedure described by Lillesand and Kiefer (2000): (1) an ASD Spectralradiometer was aimed at a calibration panel assumed to reflect 100% of the electro-magnetic energy incident on it; (2) the instrument was then aimed at each of the vegetation assemblage for which spectral reflectance was being measured with periodic re-pointing at the calibration panel every so often; (3) spectral reflectance of each vegetation assemblage was then computed by ratioing the reflected energy measurement in each band of observation measured in step (2) to the incoming radiation measured in each band measured in step (1). Step 3 was done automatically by the software RS3 running on a controller lap top supplied with the ASD FieldSpec® 3 Spectralradiometer and the resultant file saved as a binary file of spectral reflectance. Spectral measurements were taken over a period of three days, with each day spent on a single site. The days were generally sunny while measurements were taken between 12:00pm and 1:30pm on site 1, 12:30pm and 2:30pm on site 2 and 11:30am and 2:30pm on day 3. Geographic location of each measured site sample was taken as Latitude and Longitude coordinates using Trimble GPS R4 Receivers in differential mode with the reference station set up near the start of the transect using 'here' (autonomous) global coordinates. A digital image of each measured site sample was taken using a hand-held digital camera. The images were necessary for identification of dominant plant species in each of the measured vegetation assemblage which subsequently would be essential for interpretation, classification and analysis of imagery.

2.3. Data Pre-processing

2.3.1. Spectral Reflectance

Spectral reflectance saved as binary files could only display as reflectance curves using SpecView Software supplied with the ASD FieldSpec® 3 Spectralradiometer. Each binary file of spectral measurements, therefore had to be converted to text format to be able to display and processed using other software packages. The resultant text reflectance spectra were then combined into a single file using MS- Excel spreadsheet for subsequent statistical analysis.

2.3.2. Pixel Identification

Pixels on the Hyperion imagery coinciding with the vegetation assemblage samples taken in the field were identified on the image with the help of field measured GPS spatial locations. This was necessary to allow for analysis to be performed using a pixel as the smallest logical spatial element under the assumption that the reflectance recorded over a pixel by the satellite sensor was equivalent to an average of the field spectral reflectance measured over a pixel once it has been corrected for the effect of atmospheric and other sources of error. The outlines of the pixels were on-screen digitized using ArcGIS software and assigned arbitrary identification numbers (Fig. 2). The pixel numbers were then associated with each of vegetation assemblage spectral measurements taken in the field, added as column to the file of spectral reflectance measurements and later used as a target for RF model building and related data analysis.

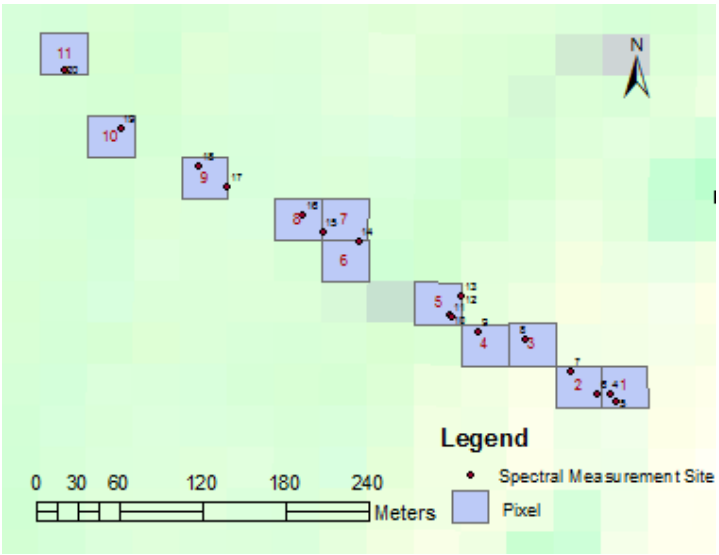

Figure 2: Outline of sample 30 by 30sqm pixels corresponding to field spectral measurement taken at site 3 superimposed over part of Hyperion image

2.4. Data Analysis

Spectral reflectance data were organized with wavelengths as variables and converted to Comma Separated Values (CSV) format prior to loading into RF. RF was implemented using the 'R Analytical Tool To Learn Easily' (Rattle) package, a graphical data mining application written in and serves as pathway into R. Rattle Graphical User Interface (GUI) brings together a multitude of R packages essential for the data miner, but often not easy for the novice to use (Williams, 2009). RF implemented on Rattle requires very few inputs. The user supplies the number of trees (mTree) to be created and the number of split variables to be permutated per node (mTry). If necessary the user may have to indicate the number of samples. This was not necessary in this study as all the variables contained the same number of entries.

The results from the RF algorithm include three 'importance' independent variable measures: the permutation accuracy importance measure, the Gini importance and the number of times each variable is selected (Breiman, 2001). The permutation accuracy importance measure, is considered to be the best measure because of its capability to assess the variable importance that relies on mean decreases in accuracy as measured using the 'out-of-bag' (OOB) samples (Breiman, 2001). The OOB error rate is computed by putting each OOB observation down the corresponding classification tree from which it was excluded. The error estimate is then computed as the misclassified proportion of that OOB observation. The OOB error produces a measure of the importance of the variables by comparing how much the OOB error of estimate increases when a variable is permutated whilst all other variables are left unchanged (Archer and Kimes, 2008; Peters et al., 2007). The variables permutation is the most reliable measure which computes variable importance as the mean decrease in accuracy based on the OOB observations (Breiman, 2001). Permutation of variables (mean decrease in accuracy) to measure the importance of wavelengths in discriminating vegetation assemblages was used (Breiman 2001) as a ranking index to measure the importance of the variables. To obtain the highest accuracy, the RF model was optimized based on OOB estimate of error rate using different number of trees (nTree) and various mTry values (Breiman, 2001). The model with lowest OOB estimate of error was adopted as the best under the circumstances and its associated wavelengths with relatively large importance as given by the mean decreasing accuracy adopted as the optimal wavebands for discriminating vegetation assembly (Archer and Kimes, 2008).

The performance of the adopted model was subsequently evaluated for accuracy using the testing data (i.e. 1/3 of the data used to train the model) and a full data set used to develop the model. Spectra data collected from the other two sites were used to assess the model's ability to predict (identify) similar data (pixels) in other datasets based on similar spectral characteristics.

3. Results

3.1. Optimized RF Model

Table 1 shows the OOB estimates of error results of the various RF models created in an attempt to obtain an optimized model. Number of trees (nTree) from 100 to 10,000 (maximum allowable by the software) with variable intervals of 50, 100 and 500 where used while the number of variables tried at each split or node (mTry) used, ranged from 1 to 100 with varying intervals of 5 and 10. The optimal model was identified as one with the lowest OOB estimate of error rate as this conversely gives highest accuracy obtainable from a given dataset (Williams, 2011). William (2011) recommends use of 100 – 500 nTree and square root of the variables number for mTry except when data has noise in which case the mTry could be increased. Mansour et al. (2012) used 500 to 10000 nTree in 500 intervals with mTry of 1 to 20. In this study a wider range was used for both nTree and mTry (Table 1).

Table 1: RF models generated for different nTree and mTry values to identify the optimal Model

	mTry													
nTree	1	5	10	15	20	30	40	46	50	60	70	80	90	100
100	78.26	72.83	72.83	76.09	75	78.26	82.61	78.26	79.35	79.35	76.09	76.09	80.43	84.78
150	76.09	70.65	73.91	77.17	76.09	79.35	82.61	77.17	78.26	77.17	79.35	78.26	78.26	82.61
200	73.91	71.74	73.91	71.74	72.83	75	82.61	76.09	76.09	77.17	77.17	79.35	75	82.61
300	75.00	75.00	75.00	72.83	75.00	77.17	79.35	76.09	73.91	73.91	78.26	77.17	79.35	82.61
400	77.17	75.00	76.09	71.74	76.09	78.26	79.35	76.09	73.91	76.09	77.17	75.00	78.26	80.43
500	75.00	75.00	76.09	72.83	75.00	77.17	77.17	76.09	70.65	76.09	75.00	75.00	78.26	83.70
1000	78.26	72.83	76.09	73.91	75.00	77.17	77.17	75.00	73.91	72.83	75.00	75.00	77.17	77.17
1500	78.26	73.91	75.00	73.91	72.83	73.91	78.26	75.00	72.83	72.83	73.91	77.17	76.09	78.26
2000	76.09	72.83	76.09	72.83	75.00	75.00	77.17	75.00	73.91	73.91	75.00	75.00	78.26	78.26
2500	73.91	72.83	76.09	73.91	72.83	72.83	73.91	75.00	73.91	73.91	75.00	75.00	76.09	78.26
3000	73.91	72.83	75.00	73.91	73.91	75.00	76.09	72.83	72.83	73.91	75.00	75.00	76.09	78.26
3500	73.91	72.83	75.00	75.00	72.83	73.91	73.91	73.91	72.83	73.91	75.00	75.00	77.17	78.26
4000	73.91	72.83	75.00	73.91	71.74	73.91	73.91	73.91	71.74	73.91	73.91	75.00	76.09	77.17
4500	72.83	72.83	75.00	73.91	72.83	73.91	72.83	73.91	71.74	75.00	76.09	76.09	73.91	77.71
5000	72.83	72.83	75.00	73.91	72.85	75.00	72.83	73.91	72.83	75.00	76.09	75.00	75.00	78.26
5500	72.83	72.83	75.00	73.91	73.91	73.91	72.83	73.91	71.74	75.00	76.09	75.00	75.00	77.17
6000	72.83	72.83	75.00	75.00	72.83	73.91	72.83	73.91	71.74	76.09	76.09	75.00	75.00	77.17
6500	72.83	72.83	75.00	75.00	72.83	73.91	72.83	73.91	71.74	77.17	76.09	73.91	76.09	78.26
7000	75.00	72.83	75.00	73.91	72.83	73.91	73.91	73.91	71.74	77.71	76.09	73.91	76.09	77.17
7500	72.83	72.83	75.00	75.00	72.83	73.91	73.91	73.91	71.74	77.17	75.00	75.00	76.09	78.26
8000	73.91	72.83	75.00	73.91	72.83	73.91	73.91	73.91	71.74	77.17	75.00	73.91	76.09	77.17
8500	73.91	72.83	75.00	75.00	72.83	73.91	73.91	73.91	71.74	77.17	75.00	75.00	76.09	77.17
9000	75.00	72.83	75.00	75.00	72.83	73.91	73.91	73.91	71.74	76.09	76.09	75.00	75.00	77.17
9500	73.91	72.83	75.00	73.91	72.83	73.91	73.91	73.91	71.74	73.91	75.00	75.00	75.00	78.26
10000	72.83	72.83	75.00	73.91	72.83	73.91	73.91	73.91	71.74	73.91	75.00	75.00	76.09	77.17

Graphs of nTree against mTry (Fig. 3) were used to observe how the model changed for a given nTree as mTry was adjusted, based on the OOB estimate of error, and conversely when nTree changed for a given mTry. Models that exhibited a downward or near downward trend around maximum mtry of 100 were extended further to check their performance beyond mTry 100. This was done to ensure that none represented a model with OOB estimate error less than the minimum obtained already.

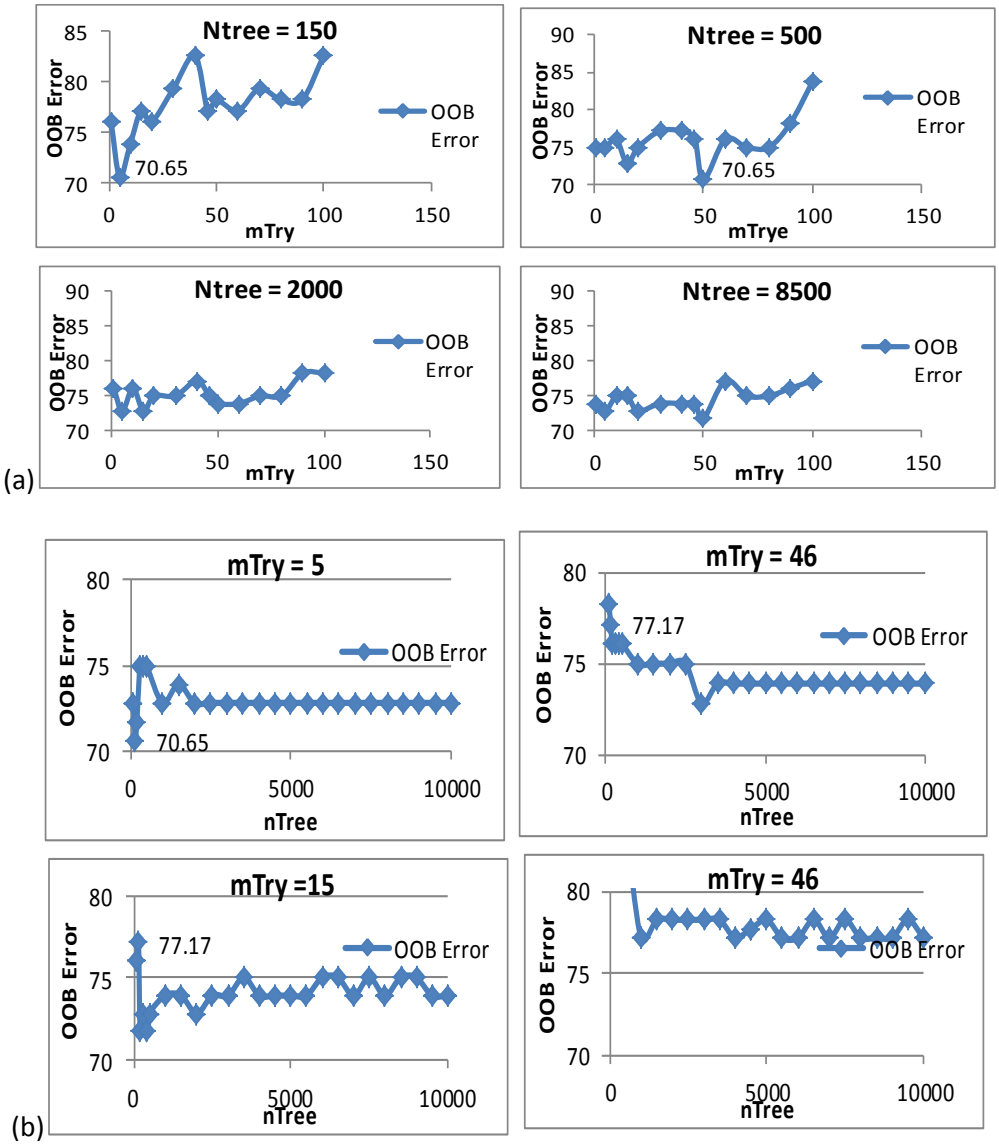

Figure 3: RF Optimization Graphs showing OOB estimate of error: (a) indicated number of trees (nTree) against different number of split variables per node (mTry) (b) indicated number of split variables per node (mTry) against different number of trees (nTree)

Two models with OOB estimate of error value of 70.65% were identified as being optimal. One model was attained at nTree, mTry of 150, 5 and the other at 500, 50. This implies that when each of these models is applied to new set of observations, the prediction accuracy will be in error 70.63% of the time. The models therefore were 29.35% accurate.

3.2. Model Performance Evaluation

Each of the models was accompanied by a confusion matrix detailing the disagreements between the model's predictions and the actual outcomes of the training observations. The resultant class errors were generally similar except for a few minor differences in pixels 5, 11, 14, 16, 19, and 30 The overall prediction accuracy for both models, however, was the same at 29.35% corresponding to the inverse percentage of OOB estimate of error value of 70.65%.

The performance of the two models was further evaluated using the testing data and, full data set. The models yielded quite close, but not exact overall prediction accuracy for the testing data. The 150 by 5 model produced 28.57% while 500 by 50 produced 30.95%. Both values were in the range of the 29.36 produced by the model. The testing data errors of omission and commission as well as the producer's and user's accuracy for both models, however, were quite different for certain variables. This presumably can be attributed to a smaller number of variables (pixels) involved in the model development.

Model performance evaluation based on the full data set yielded the same overall prediction accuracy of 78% with errors of omission and commission as well as the producer's and user's accuracy for both models being exactly the same. This however, is significantly different from the expected 70.65% produced by the model.

3.3. Optimal Wavebands for Discriminating Vegetation Assemblages

The importance of each waveband in determining their role to the model development is provided in form of a table by RF as part of the overall results allowing for each waveband's contribution to be evaluated. However, two other measures of variable importance are also provided by RF in form of (1) the scaled average of the prediction accuracy of each variable and (2) the total decrease in a decision tree node's impurity (splitting criterion) when splitting on a variable. The former is reported as the 'mean decrease in the accuracy' of the model and calculated by randomly permuting the values of each variable across the observations and measuring the impact on the predictive accuracy of the resulting tree. The latter is based on the 'Gini index' splitting criterion measured for a variable over all trees giving a measure of the mean decrease in the Gini index of diversity relating to the variable (William, 2001).

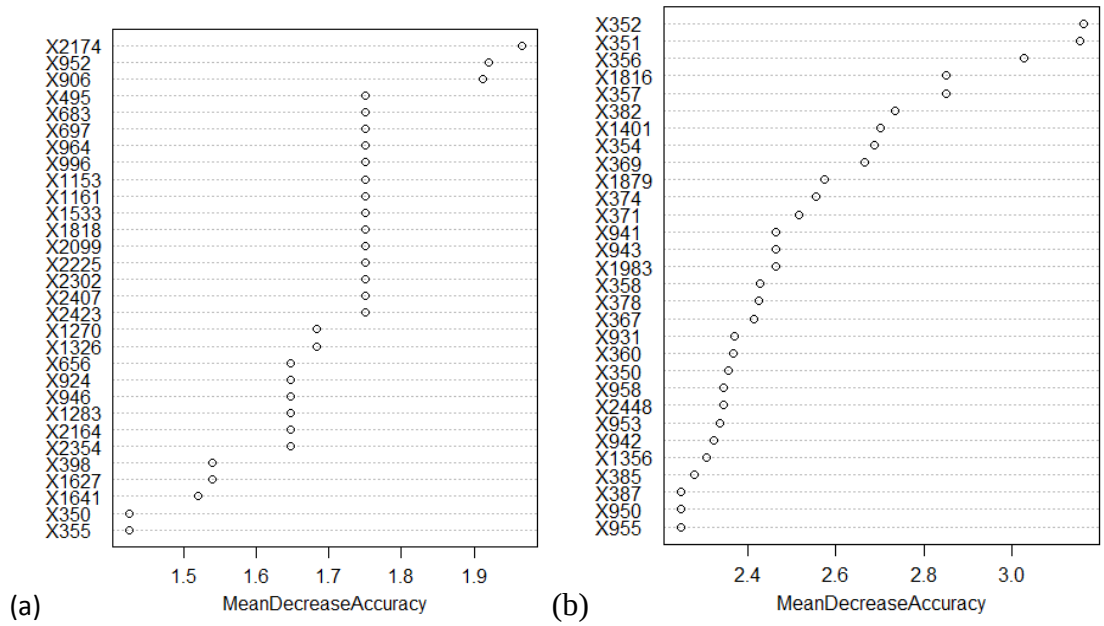

Figure 4: Ranking of Pixel Importance in discriminating among vegetation assemblages (a) 150 by 5 model (b) 500 by 50 model

The order of importance of variables given by the 'mean decrease in accuracy' was used in this study as a ranking for identifying the wavebands critical to the discrimination of pixels representing vegetation assemblages. Figures 4 and 5 show the first thirty (30) waveband of importance and their distribution respectively based on the 150by5 and 500by50 models.

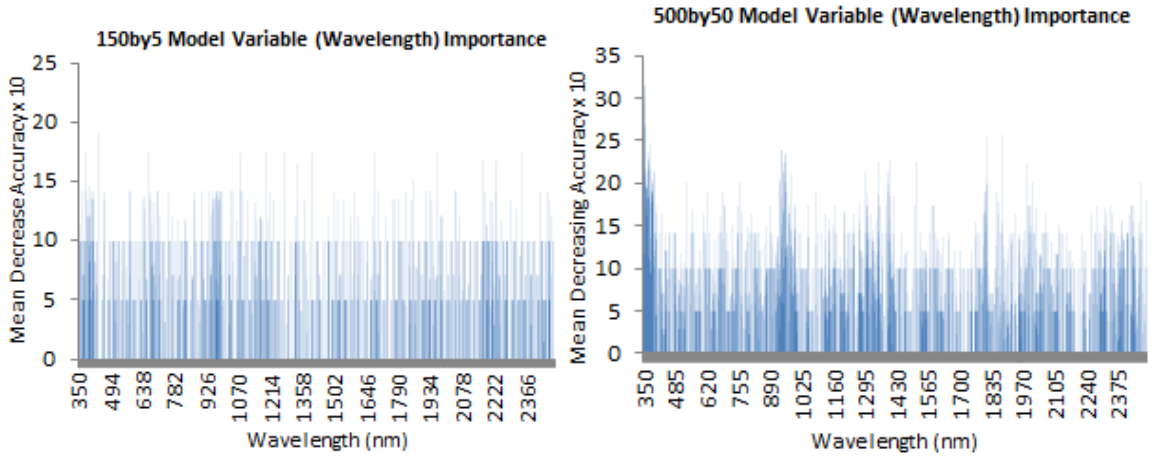

Figure: 5: Distribution of the Waveband importance over the wavelength region 350 -2500 nm

It can be seen from figures 4 and 5 that despite the OOB estimate of error being the same for the two models, the identified wavebands are not the same. The performance of each of these data sets in discriminating corresponding vegetation assemblages on the relevant imagery will therefore, need to be investigated as part of the continued research.

3.4. Data Prediction Using the Developed Model

Where an independent set of data similar to that used to develop the model is available, such data can be used to score the model performance. In this study, the available independent data set were collected from different sites (sites 1 and 2 on Fig. 1). Despite this being the case, the model was used to predict the pixels in the new data set that had similar spectral characteristics to that of the pixels used to build the model. The results of the prediction of each model when applied to the field spectral measurements collected from site 2 (Fig. 1) are given in table 2. This can be interpreted for column 5 and row 35 (table 2(b)) to imply that 3 of the 5 five pixels with pixel value 5 predicted from the data used to build a model have spectral characteristics similar to that of pixel 35 in the new data set while the remaining 2 pixels resemble pixel 41 in the new data set.

Table 2: Predicted pixel values of spectral data collected from site 2 by applying (a) 150 by 5 model (b) 500 by 50 model

(a)

150by5 Model		Predicting Pixels (Pixels from data set use to create model)																														
		1	2	3	5	6	7	8	9	10	11	12	14	15	16	17	18	19	21	22	23	24	25	26	27	28	29	30	31	32	33	34	**Total**
Predicted Pixels (Pixels from the new dataset)	35	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**
	36	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	**3**
	37	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**3**
	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	**3**
	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	**2**
	40	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
	41	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**3**
	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**1**
	Total	**0**	**1**	**1**	**4**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**6**	**0**	**0**	**4**	**0**	**0**	**0**	**0**	**0**	**1**	**21**

(b)

500by50 Model		Predicting Pixels (Pixels from data set use to create model)																														
		1	2	3	5	6	7	8	9	10	11	12	14	15	16	17	18	19	21	22	23	24	25	26	27	28	29	30	31	32	33	34	**Total**
Predicted Pixels (Pixels from the new dataset)	35	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**
	36	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	**3**
	37	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**3**
	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	**3**
	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**2**
	40	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
	41	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**
	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**1**
	Total	**0**	**1**	**1**	**5**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**6**	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**21**

4. Discussion

The aim of this study was to identify optimal wavebands suitable for discriminating among spatial features based on their field spectral measurements using the random forest algorithm. The identified wavebands could then be used to map spatial coverage of each feature by classifying hyper-spectral imagery of the area from which the field measurements are taken. To achieve this, an optimal RF model was developed. The wavebands identified as exhibiting highest 'variable importance' for the developed model were adopted as being the optimal for subsequent classification of the hyper-spectral imagery in order to map the corresponding spatial features.

Previous studies have demonstrated that RF is sensitive to nTree and mTry (Adam et al, 2009; 2010; Mansour et al 2012). In this research the lowest OOB error rate was not obtainable when the number of nTree were necessarily large, nor was it with the mTry default setting given as square root of the number of variables in the data set. This is confirmed by the fact that two models with different combinations of nTree and mTry were able to deliver the lowest of the OOB obtainable from all attempted models. What was clearly observable from the graphs (Fig.3) though was that the OOB error got smaller and the range became predominantly smaller with large nTree. The later, though was not consistently so with large mTry.

The variable importance ranking provided by RF has enabled identification of optimal wavebands necessary for discriminating among different vegetation assemblages as determined by the representative pixels. Any reasonable number of the top most wavebands in the ranking can therefore be used to map spatial features using the relevant imagery by way of classification. These wavebands, however, were found to be different between the two models that delivered the lowest OOB estimate of error of 70.65%. Which one of these models produces the best results will need to be investigated.

The evaluation of the models that delivered the lowest OOB estimate of error exhibited similar results for the full data set, with differing results for the test data set. The use of the model to predict the results of the new set of observation demonstrated the power of RF as a machine learning ensemble capable of being trained to forecast result. How reliable the results are is subject to accuracy assessment which was not conducted in this study. It must be admitted though that the obtained OOB estimate of error for the given dataset was generally high. This may be attributed to the fact that the used data had a lot of noise.

5. Conclusion

The results of this study have demonstrated that RF is capable of identifying optimal wavebands for discriminating among various field spectral measurements for the purpose of using such identified wavebands to map coverage of vegetation assemblages at any imagery pixel level. This simplifies the tedious process of using classical non-parametric inferential approaches to determine optimal wavebands. Evaluation of the RF model is done quite simply using the mis-classification matrix and methods used in secular image classification accuracy assessment. The developed RF model has shown that it can be used to predict the results of a new set of data similar in structure to that used to develop the model, thus providing means for forecasting. Lastly this approach can, by implication, be deployed to any given data set of spectral measurements, thus allowing for mapping of any combination of spatial features using hyper-spectral imagery.

References

Adam, E. and Mutanga O 2009, 'Spectral discrimination of papyrus vegetation (Cyperus papyrus L) in swamp wetlands using field spectrometry', *ISPRS Journal of Photogrammetry and remote Sensing,* vol. 64, no. 6, pp. (612-620).

Analytical Spectral Devices (ASD) 2005, *Analytical Spectral Devices, Inc, Handheld Spectralradiometer: User's Guide*, Boulder.

Archer, K and Kimes R 2008, 'Empirical characterization of random forest variable importance measures', *Computational Statistics and Data Analysis, vol.* 52 no. 4, pp. 2249 – 2260.

Bajcsy, P and Groves P 2004, 'Methodology for hyperspectral band selection', *Photogrammetric Engineering and Remote Sensing, vol.* 70, no. 7, pp. 793 - 802.

Breiman, L 2001, 'Random Forests', *Machine Learning,* vol. 4, pp. 5-32.

Lawrence, R, Wood S and Sheley R 2006, 'Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest)', *Remote Sensing of Environment* , vol. 100 , no. 3, pp. 356 – 362.

Lillesand, T. M. and R. W. Kiefer, RW 2000, *Remote Sensing and Image Interpretation*, John Wiley & Sons, Inc. New York

Macus, W 2001, 'Mapping of stream microhabitats with high spatial resolution hyperspectral imagery', *Journal of Geographic Systems*, vol. 4, pp. 113-126.

Mansour, K., Mutanga O, Everson, T and Adam, E 2012, 'Discriminating indicator species for rangeland degradation assessment using hyperspecral data resampled to Eagle resolution', *ISPRS Journal of Photogrammetry and remote Sensing*, vol. 70, no. 2012, pp. 56-65.

Mutanga, O and Skidmore A 2004, 'Hyperspectral band depth analysis for a better estimation of grass biomass (Cenchrus ciliaris) measured under controlled laboratory conditions', *International Journal of Applied Earth Observation and Geoinformation*, vol. 5, no. 2, pp. 87-96.

Peters, JB, De Baets, B Samson, R and Verhoest N 2007, 'Modelling groundwater dependent vegetation patterns using ensemble learning', *Hydrology and Earth System Sciences Discussions*, vol., no. 4, pp. 3687–3717.

Vaiphasa, C., Skidmore A, de Boer W and Vaiphasa T 2007, 'A hyperspectral band selector for plant species discrimination', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 3, pp. 225 - 235.

Williams, G 2011, *Data Mining with Rattle and R - The Art of Excavating data for Knowledge Discovery*, Springer, New York

Williams, G J 2009, 'A Graphical User Interface for Data Mining in R' *R package Version 2.4.55. viewed* 29 Dec, 2012, <http://cran.r-project.org/package=rattle>.