High density biomass estimation: Testing the utility of Vegetation Indices and the Random Forest Regression algorithm

O. Mutanga*[a], E. Adam [a]

[a] University of KwaZulu-Natal, Discipline of Geography, P. Bag X01, Scottsville 3209, Pietermaritzburg, South Africa

MutangaO@ukzn.ac.za, emiadam2006@yahoo.com

Abstract- Accurate estimates of wetland above ground biomass (AGB) have increasingly been identified as a critical component for an efficient wetland monitoring and management system. Multispectral remote sensing based indices have proven inadequate in estimating biomass especially at high canopy density. In this study we investigated the use of vegetation indices derived from field hyperspectral data to estimate papyrus (*Cyperus papyrus*) biomass. Spectral and above ground biomass measurements were collected at three different areas in the Greater St Lucia Wetland Park, South Africa. We evaluated the potential of narrow-band normalized difference vegetation index (NDVI) calculated from all possible two band combinations between 700 nm to 1000 nm. Subsequently, we utilized random forest (RF) as a modeling tool for predicting papyrus biomass. The results showed that papyrus biomass can be estimated at full canopy level under swamp wetland conditions (R^2 = 0.73, RMSEP = 276 g/m^2; 8.6 % of the mean). From our results, random forest has proved to be a robust feature selection method in identifying the minimum number (n = 4) of narrow-band NDVIs that offered the best overall predictive accuracy. The results can be scaled to spaceborne or airborne sensors such as Hyperion or HYMAP for predicting vegetation biomass in wetland areas using remotely sensed data.

Keywords: Environment, Vegetation, Hyperspectral, Ecosystem, saturation

1. INTRODUCTION

Papyrus (*Cyperus papyrus L.*) is increasingly being recognized as the most biomass productive plant species in the tropical wetlands in Africa (Muthuri and Kinyamario, 1989). However, the continued degradation in papyrus habitat represents a significant threat to biodiversity conservation particularly for papyrus-specialist birds and other papyrus-reliant species in many African countries (Owino and Ryan, 2007). Efficient techniques that can spatially and temporally monitor the stability of the productivity of papyrus ecosystems and whether significant changes are taking place in these swamp ecosystems are therefore required. Such techniques require up-to-date spatial information on the density of papyrus vegetation.

The measurement of vegetation quantity, leaf area index (LAI) and percent green vegetation cover has successfully been achieved using vegetation indices such as Normalised Difference Vegetation Index (NDVI), Simple Ratio (SR), Transformed Vegetation Index (TVI) and Transformed Soil Adjusted Vegetation Index (TSAVI) (Gao et al., 2000; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). Although these indices have been successfully used in areas with open canopy cover or sparsely vegetated regions, they have not been successful in estimating quantity at high canopy density. Specifically, the widely used vegetation indices particularly NDVI derived from broad band satellite images such as NOAA or Landsat TM tend to saturate after a certain biomass density (about 15 kgm^{-2}) or vegetation age (15 years in tropical forest) (Lu and Batistella, 2005,Gao et al., 2000; Tucker, 1977). Therefore, NDVI yields poor estimate during peak growing seasons and in more densely vegetated areas (Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). Papyrus biomass is usually dense since it grows in wetland areas with adequate water availability. Such a high density poses a major problem in measuring using the existing remote sensing indices thus there is need to improve techniques for better estimation of AGB in high diversity and densely vegetated areas such as wetlands where there is almost 100 % vegetation cover.

Ensemble methods like random forest (Breiman, 2001) have been used to enhance the prediction accuracy in the field of ecology (Grimm et al., 2008; Prasad et al., 2006). A combination of indices derived from hyperspectral data with ensemble techniques such as the random forest could improve the prediction of ABG in high canopy density areas because of its capability to simultaneously explore indices based on the whole region of the electromagnetic spectrum (Ismail and Mutanga, 2009).

This study sought to evaluate the utility of narrow-band NDVI derived from field spectrometry measurements for estimating papyrus AGB in complex and densely vegetated canopies, using the random forest algorithm. AGB and spectral data from papyrus vegetation were collected in the summer of 2009 at the Greater St Lucia Wetland Park, South Africa, which is characterized by mixed species composition.

2. MATERIAL AND METHODS

2.1 Study area

The study sites are located in the Greater St Lucia Wetlands Park (GSWP) on the eastern coast of South Africa. The park covers about three million hectares between longitudes 32°21' E and 32°34' E and latitudes 27°34' S and 28° 24' S, and it is considered to be the largest estuarine system in Africa. This study focuses on approximately 7000 ha of wetland vegetation located on three sites i.e. Futululu Park, and the Mfabeni and Mkuzi swamps .At these sites, papyrus (Cyperus papyrus) occurs in large areas between forested dunes and plantation forest on organic and alluvial soil.

2.2 Field spectral measurements and biomass harvesting

Random sampling was adopted in this study. A 30 m by 30 m vegetation plot was created to cover an area of papyrus. Three subplots (1 m × 1 m) were then randomly selected to cover a homogenous area of papyrus within each plot (30 m × 30 m) to measure the spectral reflectance. The spectral reflectance measurements were performed in the spectral range from 350 nm to 2500 nm under sunny and cloudless conditions using the Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer. From each subplot (1 m × 1 m) approximately 5 to 10 field spectrometer measurements were randomly taken at nadir from 1 m using a 5^{o} field of view. These spectral measurements were then averaged to obtain the final spectral measurement for each vegetation plot. After spectral measurements, fresh papyrus biomass was clipped within the subplots (1 m × 1 m). All dry material was removed from the clipped plants and fresh above ground biomass (AGB) was then measured immediately using a digital weighing scale.

2.3. Data analysis

2.3.1. Narrow band indices

The narrow NDVI-based vegetation indices were computed in this study from all possible two band combinations using all the red, red edge, and NIR bands (i.e. 600 nm to 1000 nm). These indices and spectral regions were selected because they are the most commonly used in estimating biomass and crop yield (Cho et al., 2007; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). The discrete 401 narrow bands allowed a computation of N*N = 160,801 narrow band indices.

2.3.2 Random forest regression ensemble

The random forest (RF) algorithm (Breiman, 2001) was used in this study to predict the AGB of papyrus (g m^{-1}). The algorithm generates multiple bootstrap samples from the original training data set with replacement to create multiple regression trees (*ntree*). Each tree is grown to maximum size without pruning with a randomized subset of predictors (*mtry*) to determine the best split at each node of the tree (Breiman, 2001). The results from each aggregation are then averaged to get the overall prediction accuracy. When a bootstrap sample is drawn, about 37 % of the dataset is excluded from the sample and the remaining data are replicated to bring the dataset to full size. This dataset is defined as "in bag" data, while the excluded dataset (approximately 37 %) is known as the "out-of-bag" data (OOB) (Breiman, 1996). For each tree in the ensemble, the RF algorithm also calculates the mean square error as the difference between predictions (i.e. mean square error) made using the OOB data and the "in bag" data, known as the OOB error. The OOB estimate of error is considered to be a reliable assessment and cross validation of predictive accuracy since the OOB data were not used to build or prune any regression trees in the ensemble (Breiman, 1996, 2001; Grimm et al., 2008; Ismail and Mutanga, 2009; Prasad et al., 2006). Therefore, it may not be necessary to have an independent validating dataset (Lawrence *et al.*, 2006). Additionally, the OOB data allow for the evaluation of the importance of each variable in the prediction by determining how much the prediction error would increase if the OOB data of that variable were permuted (Prasad *et al.*, 2006). The number of trees (*ntree*) in the forest and the randomly selected number of variables tried at each node (*mtry*) have been optimized and selected based on the lowest RMSEC (Breiman, 2001).

To validate the performance of the random forest algorithm (Lawrence *et al.*, 2006), the data were randomly divided into 70 % training or calibration and 30 % test data samples (n = 32 and 14 respectively). Regression analyses were performed on the calibration dataset using the OOB estimates of error. The test data set was used to validate the predictive performance of the random forest (Ismail and Mutanga, 2009; Lawrence et al., 2006).

2.3.3. Selection of the predictive variables

The narrow-band indices NDVIs (n = 160,801) were ranked based on the correlation coefficient = r (R^{2} = coefficient of determination). The top 20 NDVIs that yielded the highest R^{2} were then selected for further analysis in order to simplify the modelling process (Mutanga and Skidmore, 2004a).The combination of random forest and backward elimination function was then used to identify the sequence in which to discard the least important variables (NDVI) (Ismail and Mutanga, 2009). The backward variable selection process iteratively builds multiple random forests for regression. At each iteration (n = 20), a new forest was developed after gradually eliminating one of the least promising narrow-band NDVI and RMSEC was calculated. We compared the performance of OOB with both the hold out test dataset and the 10 fold cross validation (Ismail and Mutanga, 2009). The nested subset of variables (NDVI) that yielded the lowest RMSEC was then selected as the optimal variable for biomass prediction.

3. RESULTS

3.1. Hyperspectral indices (NDVI) and biomass

The results of the correlation coefficients, R^{2} between the entire possible two narrow-band NDVIs (n = 160,801) and papyrus fresh biomass are shown in figure 1. The band combinations involving the far red edge bands located from 720 nm to 850 nm range yielded the strongest correlations (0.73 to 0.83).

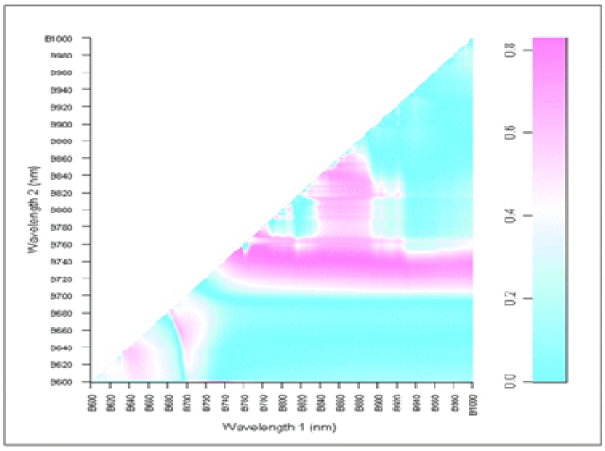

Figure 1. The correlation of reflectance indices involving all the possible two band combinations and papyrus biomass.

The NDVIs were then ranked based on their correlation coefficients, and the top 20 two band combinations that yielded the highest R^{2} values were then selected for further analysis.

3.3. Parameters optimization of the random forest regression

The results of optimizing random forest parameters (*ntree* and *mtry*) are shown in Figure 2. The lowest RMSEC was obtained with the default value of *mtry* and high *ntree* (5500).

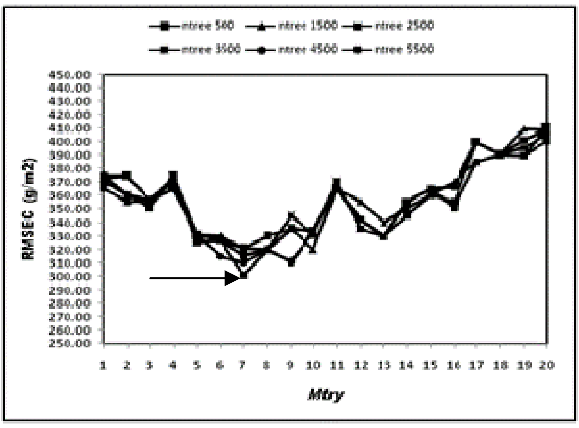

Figure 2. Optimizing the random forest parameters (*mtry* and *ntree*) using RMSEP. The black arrow shows the lowest RMSEC value

3.4. Determinations of predictor variables

Figure 3 shows the results of the variables selection using the combination of random forest and backward elimination function. Four NDVIs achieved the lowest RMSEC using the OOB sample (269 g/m^2), 10 fold cross validation (271 g/m^2) and hold out test dataset (276 g/m^2). These four NDVIs involve a combination of wavelengths located in the NIR (853 nm, 853 nm, 847 nm, and 776 nm) and shorter wavelengths of the red edge (741 nm, 740 nm, 741 nm, and 749 nm) respectively.

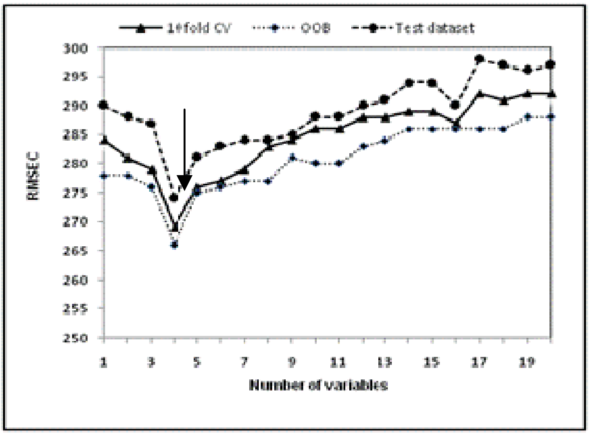

Figure 3. The optimal predictive variables selection using the backward elimination process using OOB method, 10 fold cross validation, and the test dataset). The lowest RMSEC obtained is shown by the black arrow

3.5. Development of the prediction model

Table 1 shows the RF prediction performance of the best selected NDVIs (n = 4) compared to those obtained by the standard NDVI calculated from a near infrared (833 nm) and red band (680 nm) (Tucker, 1977), the best NDVI computed in this study (853 nm and 741 nm), and the top 20 NDVIs. The R^2 values and root mean squared error for calibration (n = 32) and test (n = 14) datasets indicate the best predictive performance of the RF model obtained when using the selected four NDVIs: NDVI (853 nm, 741nm), NDVI (853 nm, 740 nm), NDVI (847 nm, 741 nm), and NDVI (776 nm, 749 nm).

Table 1. The performance of random forest model for prediction of papyrus biomass using different subsets of NDVIs

NDVIs	Calibration (n = 33)			Independent validation (n = 14)		
	R2 actual vs. Predicted	RMSEC g/m2	Mean %	R2 actual vs. predicted	RMSEP g/m	Mean %
Standard NDVI (833nm and 680 nm)	0.026	539	16.7	0.015	694	21.5
Best NDVI (741 nm and 853 nm)	0.72	295	9.2	0.66	306	9.5
Selected NDVIs (n = 4)	0.77	266	8.2	0.73	276	8.6

4. DISCUSSION

The use of remote sensing techniques in estimating biomass from dense vegetation or high leaf area index (LAI) such as wetland environments has been constrained by the asymptotic saturation of vegetation indices such as NDVI (Kumar et al., 2001; Mutanga and Skidmore, 2004a; Tucker, 1977). This study showed that papyrus biomass can be estimated with high accuracy in areas of high dense vegetation using random forest regression algorithm and narrow band NDVI calculated from the red edge and NIR regions of electromagnetic spectrum.

4.1. Relationship between the narrow band NDVIs and papyrus biomass

The model developed in this study indicated that considerable information on the status of papyrus biomass is contained in the red edge and near infrared wavelengths. However, the high correlation between AGB and NDVIs obtained in this study consisted of narrow band NDVI calculated from shorter wavelengths of the near red edge portion of the electromagnetic spectrum (700 nm to 750 nm), and the longer wavelengths of the red edge (750 nm to 800 nm). This result is consistent with the findings of previous studies (Cho et al., 2007; Mutanga and Skidmore, 2004a).. Additionally, the wavelengths used to develop the best NDVIs (n = 20) in this study are within ± 10 nm of the known wavelengths that have strong relationships with biomass prediction as reported in other studies. These are 740 nm, (Cho *et al.*, 2007), 746 nm (Mutanga and Skidmore, 2004a), and 775 nm (Kawamura *et al.*, 2008).

4.3 Variable selection

It has been noted that the use of the standard NDVI might not be able to explore the strength of the large number of hyperspectral bands because only two bands from red and NIR are used to formulate the NDVI (Mutanga and Skidmore, 2004a). In this study, the results of calculating the narrow band NDVIs from all possible two band combinations between red and NIR and then correlating it with AGB (g/m^2) improved an understanding of the relationship between the wavelength regions and biomass estimation at full canopy cover, as well as presenting a possibility to explore the rich information content in the hyperspectral wavelengths (Mutanga and Skidmore, 2004a). This study demonstrates the validity and significance of NDVI in estimating AGB. However, selection of the best wavelengths is an important task to formulate the NDVI. Our results as shown in Figure 1 explored and ranked all the possible wavelength combinations, then the best combination of wavelengths (n = 20) was selected based on the strong correlation with AGB for further analysis. Besides ranking and selecting the best narrow band combinations (n = 20) that yielded the highest correlation with biomass, using random forest with backward elimination search function facilitated the selection of the fewest most important predictive variables (n = 4) for a simple modeling process and best predictive accuracy. The consistency of the three methods (OOB, 10 fold cross validation, and the test dataset) proposed in this study to identify the optimal number of the predictive variables (n = 4), demonstrates the reliability of OOB as an internal estimate of error rate in random forest algorithm. Our finding in this regard is identical to those of other studies that tested the reliability of OOB estimate error in the classification model (Lawrence *et al.*, 2006), and the regression model (Ismail and Mutanga, 2009).

4.4. The predictive performance of random forest model

Our results from optimizing RF model supports the assertions made in the other studies that the highest accuracy and stability of RF can be achieved by using a large number of trees (Adam et al., 2009; Díaz-Uriarte and de Andrés, 2006) and the default *mtry* values (Díaz-Uriarte and de Andrés, 2006; Grimm et al., 2008).The higher accuracy obtained in this study demonstrated the utility of random forest algorithm as a feature selection method (Adam et al., 2009; Lawrence et al., 2006) and its application as a regression model (Ismail and Mutanga, 2009). The relatively high R^2 and low RMSEC and RMSEP as shown in Table 1 indicates that the selected NDVIs (n = 4) improved the predictive performance of the model compared to the use of the entire top 20 NDVIs. Our results in this regard indicate that the variable selection method developed in this study was able to refine the performance of RF. The poor predictive performance of standard NDVI is consistent with the finding of Cho et al. (2007), involving grass/herb in the Majella National Park in Italy, and of Mutanga and Skidmore (2004a), involving blue buffalo grass (*Cenchrus Ciliaris*) grown under controlled conditions in a greenhouse. This could be explained by the saturation problem of the standard NDVI at the high biomass or leaf area index which has been reported in several studies (Mutanga and Skidmore, 2004a; Tucker, 1977).

Overall, this study has revealed that it is possible to predict dense papyrus biomass at canopy level using filed spectrometry measurements. Additionally, the developed model provides a better understanding of (i) those narrow band regions that are most sensitive for papyrus biomass estimation and (ii) the potential of random forest ensemble as a feature selection and regression type model in remote sensing applications. This permits the up scaling of the model to spaceborne or airborne sensors such as HYMAP and Hyperion

5. REFERENCE

Adam, E.M., Mutanga, O., Rugege, D. and Ismail, R., 2009. Field spectrometry of papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of St Lucia, South Africa. Geoscience and Remote Sensing Symposium,2009 IEEE International,IGARSS 2009, IV-260-IV-263.

Breiman, L., 1996. Bagging predictors. *Machine learning 24*, 123-140.

Breiman, L., 2001. Random forests. *Machine learning 45*, 5-32.

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

Díaz-Uriarte, R. and de Andrés, A., 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics 7*, 3.

Gao, X., Huete, A., Ni, W. and Miura, T., 2000. Optical-biophysical relationships of vegetation spectra without background contamination. *Remote Sensing of Environment 74*, 609-620.

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

Ismail, R. and Mutanga, O., 2009. A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. *International Journal of Applied Earth Observation and Geoinformation*, (In Press).

Kawamura, K., Watanabe, N., Sakanoue, S. and Inoue, Y., 2008. Estimating forage biomass and quality in a mixed sown pasture based on partial least squares regression with waveband selection. *Grassland Science 54*, 131-145.

Kumar, L., Schmidt, K.S., Dury, S. and Skidmore, A.K., 2001. Imaging spectrometry and vegetation science. *In*: van der Meer, F., de Jong, S.M. (ed.), Imaging Spectrometry. Kluwer Academic, Dordrecht, The Netherlands, 111–155.

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

Lu, D. and Batistella, M., 2005. Exploring TM image texture and its relationships with biomass estimation in Rondônia, Brazilian Amazon. *Acta Amazonica 35*, 249-257.

Mutanga, O. and Skidmore, A., 2004a. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing 25*, 3999-4014.

Muthuri, F. and Kinyamario, J., 1989. Nutritive value of papyrus (Cyperus papyrus, Cyperaceae), a tropical emergent macrophyte. *Economic Botany 43*, 23-30.

Owino, A. and Ryan, P., 2007. Recent papyrus swamp habitat loss and conservation implications in western Kenya. *Wetlands Ecology and Management 15*, 1-12.

Prasad, A., Iverson, L. and Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems 9*, 181-199.

Thenkabail, P., Smith, R. and De Pauw, E., 2000. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment 71*, 158-182.

Tucker, C., 1977. Asymptotic nature of grass canopy spectral reflectance. *Applied Optics 16*, 1151-1156.