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

Onisimo Mutanga
University of KwaZulu-Natal
421 PUBLICATIONS 14,679 CITATIONS
SEE PROFILE

Elhadi M.I Adam
University of the Witwatersrand
124 PUBLICATIONS 4,584 CITATIONS
SEE PROFILE

International Journal of Applied Earth Observation and Geoinformation 18 (2012) 399–406

Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm

Onisimo Mutanga[a], Elhadi Adam[a,b,*], Moses Azong Cho[a]

[a] University of KwaZulu-Natal, Discipline of Geography, P. Bag X01, Scottsville 3209, Pietermaritzburg, South Africa
[b] Elfashir University, Geography Department, P. Bag 125, Elfashir, Sudan

ARTICLE INFO

Article history:
Received 29 September 2011
Accepted 26 March 2012

Keywords:
Above ground biomass
Red edge bands
Savanna wetlands
Random forest regression algorithm
Variable importance

ABSTRACT

The saturation problem associated with the use of NDVI for biomass estimation in high canopy density vegetation is a well known phenomenon. Recent field spectroscopy experiments have shown that narrow band vegetation indices computed from the red edge and the NIR shoulder can improve the estimation of biomass in such situations. However, the wide scale unavailability of high spectral resolution satellite sensors with red edge bands has not seen the up-scaling of these techniques to spaceborne remote sensing of high density biomass. This paper explored the possibility of estimate biomass in a densely vegetated wetland area using normalized difference vegetation index (NDVI) computed from WorldView-2 imagery, which contains a red edge band centred at 725 nm. NDVI was calculated from all possible two band combinations of WorldView-2. Subsequently, we utilized the random forest regression algorithm as variable selection and a regression method for predicting wetland biomass. The performance of random forest regression in predicting biomass was then compared against the widely used stepwise multiple linear regression. Predicting biomass on an independent test data set using the random forest algorithm and 3 NDVIs computed from the red edge and NIR bands yielded a root mean square error of prediction (RMSEP) of 0.441 kg/m^2 (12.9% of observed mean biomass) as compared to the stepwise multiple linear regression that produced an RMSEP of 0.5465 kg/m^2 (15.9% of observed mean biomass). The results demonstrate the utility of WorldView-2 imagery and random forest regression in estimating and ultimately mapping vegetation biomass at high density – a previously challenging task with broad band satellite sensors.

1. Introduction

An understanding of the distribution and characteristics of wetland vegetation is critical for sustainable ecosystem management and in preserving biological diversity. Wetland vegetation plays a vital role in providing habitats for wildlife and livestock (Maclean et al., 2006; Mafabi, 2000; Owino and Ryan, 2007) as well as in influencing their grazing distribution patterns, especially during the dry season (Muthuri and Kinyamario, 1989). Therefore, the estimation of aboveground biomass of wetlands provides useful information to spatially and temporally monitor the stability and productivity of wetland ecosystems (Adam et al., 2010; Chen et al., 2009b).

Above ground biomass (AGB) of wetlands can be estimated from remotely sensed data acquired from satellite, airborne or field sensors (Chen et al., 2009a; Mutanga and Skidmore, 2004a, 2004b; Todd et al., 1998; Tucker, 1977). This has been mainly achieved by the use of vegetation indices such as the normalized difference vegetation index (NDVI) computed from the red and near infrared bands (Hoffer, 1978; Tucker, 1977). Such indices respond to variation in strong chlorophyll absorption in the red and high reflectance due to multiple scattering effects in the near infrared (Mutanga and Skidmore, 2004a; Wiegand et al., 1991). The main problem associated with indices computed from multispectral sensors is that they reach a saturation level on high density biomass estimation (Chen et al., 2009a; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000; Tucker, 1977). NDVI calculated from broad band sensors, asymptotically approach a saturation level after a certain AGB of about 0.3 g cm^{-1} (Hurcom and Harrison, 1998) or vegetation age of about 15 years in tropical forests (Lu and Batistella, 2005; Steininger, 2000). Therefore, NDVI yields poor estimates during peak growing seasons and in more densely vegetated areas (Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). Some wetland areas in southern Africa are characterized by grass species such as papyrus (*Cyperus papyrus* L.) with very high biomass production (Adam and Mutanga, 2009; Muthuri and Kinyamario, 1989) and the use of traditional indices to estimate biomass in such areas have had limited success (Adam et al., 2010).

[*] Corresponding author at: University of KwaZulu-Natal, Discipline of Geography, P. Bag X01, Scottsville 3209, Pietermaritzburg, South Africa. Tel.: +27 332605779; fax: +27 332605344.
E-mail addresses: Adam@ukzn.ac.za, emiadam2006@yahoo.com (E. Adam).

0303-2434/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jag.2012.03.012

Recent efforts have been geared towards the use of narrow band vegetation indices computed from hyperspectral data to estimate high canopy density biomass (Chen et al., 2009a; Mutanga and Skidmore, 2004a). Results have shown that modified vegetation indices calculated from the red edge and near infrared shoulder domains can estimate biomass at full canopy cover with a high accuracy (Chen et al., 2009a; Mutanga and Skidmore, 2004a) as compared to the standard red/near infrared based indices. However, the use of hyperspectral data comes with its own limitations in terms of cost, availability, processing and high dimensionality.

The advent of new generation satellites with moderate resolution is seen as a tradeoff between the advantages of multispectral resolution satellite data and hyperspectral data. WorldView-2 is one such sensor which contains a reasonable number of spectral bands that are configured in unique portions of the electromagnetic spectrum, including the red edge. In remote sensing, "red-edge" is the region of abrupt change in the leaf reflectance between 680 and 780 nm due to the combined effects of strong chlorophyll absorption in red wavelengths and high reflectance of in the NIR wavelengths due to leaf internal scattering (Horler et al., 1983). WorldView-2 offers more wavebands (8 bands) and higher spatial resolution (2 m) than the traditional broadband satellite images such as SOPT and Landsat TM while reducing unnecessary redundancy as contained in hyperspectral data (Omar, 2010; Sridharan, 2010a). WorldView-2 imagery has recently been used for urban forest classification (Sridharan, 2010b). Successful application of spectral information contained in WorldView-2 imagery for estimating biomass in high density canopies will be usefull not only for wetland biomass mapping, but also for global mapping and monitoring of densely vegetated areas.

Multiple linear regression (MLR) methods based on more than two bands have been applied in estimating AGB (Lu, 2006). However, identifying suitable variables for developing a multiple regression model is often critical because some variables are weakly correlated with AGB or are highly correlated to each other (Lu, 2006). Given this problem, a powerful method for identifying the most useful vegetation indices to improve the prediction of AGB is essentially required (Lu, 2006).

Ensemble methods like random forest (Breiman, 2001) have successfully been used to enhance the prediction accuracy in the field of ecology (Grimm et al., 2008; Prasad et al., 2006). In the field of remote sensing, random forest has been widely applied in different fields as a classification algorithm (Adam et al., 2009; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Pal, 2005). To the best of our knowledge, however, only a few studies have investigated the use of random forest in regression type of remote sensing applications (e.g. Ismail and Mutanga, 2010; Abdel-Rahman et al., 2009).The RF algorithm is a non-parametric statistical technique that is capable of synthesizing regression or classification functions based on discrete or continuous datasets. RF also has a capability to deal with complex relationships between predictors due to the noise and large amounts of data (Ismail et al., 2010; Vincenzi et al., 2011).

In this study, we evaluated the performance of data extracted from WorldView-2 imagery to estimate biomass in a papyrus dominated wetland area of Northern KwaZulu-Natal, South Africa. We calculated NDVIs involving all possible band combinations from the WorldView-2 imagery and predicted AGB wetland biomass using field data and the random forest regression algorithm. The random forest algorithm was adopted in this study becuase of its capability to select and rank important variables for biomass prediction, in this case all possible NDVI combinations computed from the WorldView-2 image ($n = 64$). Specificallly the objectives of this study were: (i) To explore the use of Worldview-2 in solving the problem of estimating densely wetland biomass and, (ii) To test the performance and the strength of the random forest regression as variable selection and prediction method.

2. Materials and methodology

2.1. Study area

The study sites are located in the ISimangaliso Wetland Park in the eastern coast of KwaZulu-Natal Province, South Africa. The Park covers about 332 000 ha between longitudes 32°21′E and 32°34′E and latitudes 27°34′S and 28°24′S, and is considered to be the largest estuarine system in Africa (Taylor, 1995). The climate is subtropical with the mean annual rainfall varying from 1500 mm on the eastern shore to 700 mm on the western shore of the lake St Lucia (Taylor, 1995). ISimangaliso Wetland Park, which is recognized as a UNESCO World Heritage Site and a Ramsar wetland of global significance, is characterized by a high diversity of ecosystems including marine, inland lake, estuarine waters, forested dunes, mangrove, coastal and swamp forest ecosystems. This study focuses on approximately 7000 ha of wetland vegetation located on three sites, i.e. Futululu Park, Mfabeni and Mkuzi swamps (Fig. 1). These sites, characterized by a high-density of vegetation cover occur in large areas between forested dunes and plantation forest on organic and alluvial soil. The dominant vegetation species include *Cyperus papyrus* L., *Phragmites australis*, *Echinochloa pyramidalis*, and *Thelypteris interrupta.*

2.2. Field data collection

The field campaign was carried out between 12 December and 19 December 2010. This period is characterized by high rainfall and high biomass productivity. Random sampling was adopted in this study. Hawth's Analysis tool was used to generate 82 random points on a land cover map of the park obtained from ISimangaliso Wetland Park management. The sample points were subsequently uploaded into a GPS that was used to navigate to the field sites. Leica Geosystems GS20 GPS Sensor with multiple-bounce filtering and post-differential correction was used to measure the position of vegetation plots with an accuracy of 0–0.25 m after the post-processing differential correction (Leica Geosystems, 2004).

Once the sample site was located, a 20 m × 20 m vegetation plot was created to cover a homogenous area of the grass/herb. 3–5 subplots (1 m × 1 m) were then randomly selected within each plot to measure the AGB. AGB was clipped within the subplots (1 m × 1 m). All dry material was removed from the clipped plants and fresh biomass was then measured immediately using a digital weighing scale. Average fresh AGB per plot was then calculated from the subplot measurements ($n = 3–5$) (Cho et al., 2007; Mutanga and Skidmore, 2004a).

2.3. Image acquisition and pre-processing

WorldView-2 imagery covered the study sites were obtained in the first of December 2010 from DigitalGlobe. WorldView-2 image comprised eight multispectral bands with spatial resolution of 2 m and swath width of 16.4 km at nadir. The spectral ranges of the eight bands are 400–450 nm (B1-coastal), 450–510 nm (B2-blue), 510–581 nm (B3-green), 585–625 nm (B4-yellow), 630–690 nm (B5-red), 705–745 nm (B6-red edge), 770–895 nm (B7-near infrared-1), and 860–1040 nm (B8-near infrared-2). The images were orthorectified and geometrically corrected by DigitalGlobe (Updike and Comp, 2010). Radiance images were atmospherically corrected and transformed to canopy reflectance using the Fast Line-of-Sight Atmospheric Analysis of

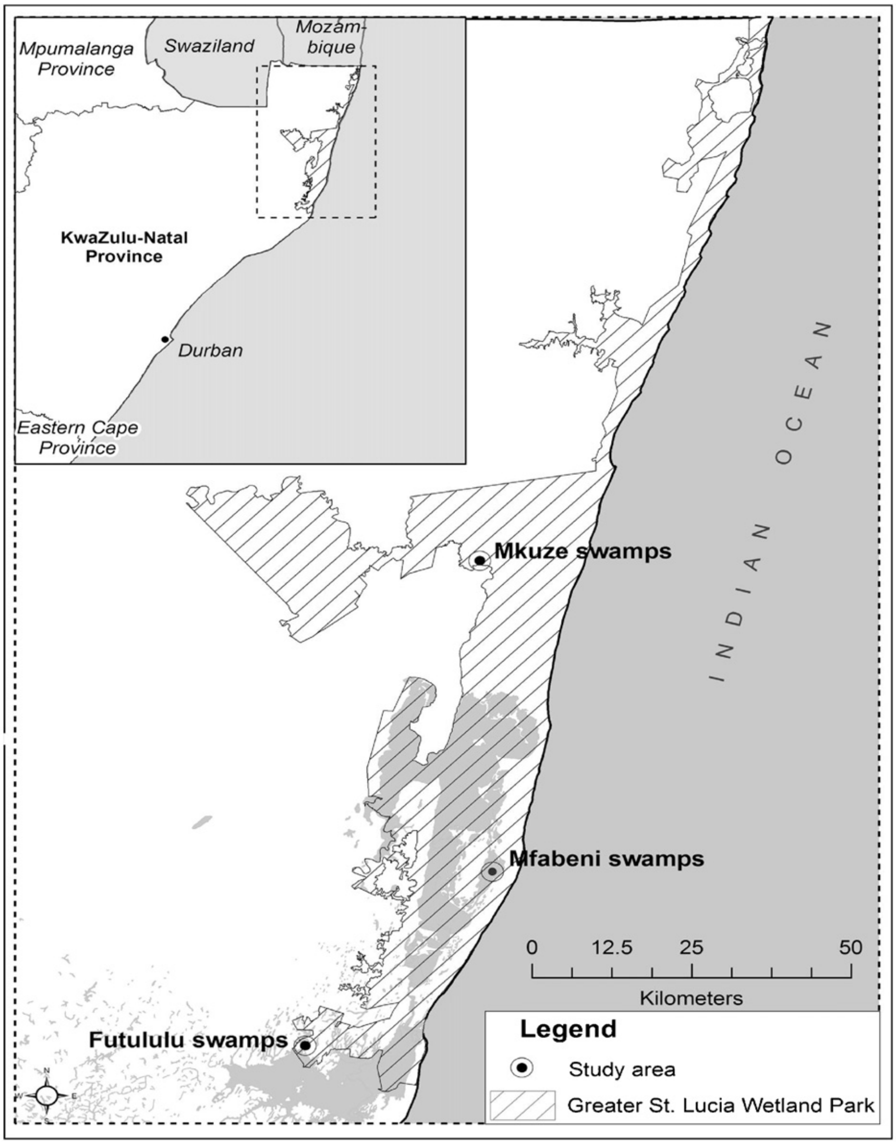

Fig. 1. Map of Greater St Lucia Wetlands Park and the three study areas.

Spectral Hypercubes (FLAASH) algorithm built in ENVI 4.7 software (Environment for Visualising Images: ENVI, 2009).

2.4. Extracting image spectra

A point map of the vegetation plots was developed using the field data and GPS readings. This map was then overlaid on the WorldVeiw-2 images to create a vegetation plot region-of-interest (ROI) map using the central GPS point for each plot ($n = 82$). An 8×8 pixels window (i.e. 16 m × 16 m) was used to collect vegetation image spectra from each band ($n = 8$) using ArcGIS 10 software. An 8×8 pixels window was used in order to avoid including pixels located outside the plot (20 m × 20 m) (Cho et al., 2007). Hence, only pixels that fall entirely within the ROIs were included in the spectral dataset, while the pixels that partially fall inside the ROIs were discarded (Cho et al., 2007; Wang et al., 2007). The spectra were collected and averaged for each vegetation plot.

2.5. The calibration and validation dataset

The dataset ($n = 82$) was randomly split in to 70% and 30% for a calibration dataset ($n = 57$) and a validation independent dataset ($n = 25$) respectively (Ismail et al., 2006). The calibration dataset was used to optimize the random forest regression and to train the prediction model, while the validation dataset was used to test the quality and reliability of the prediction model.

2.6. Statistical analysis

2.6.1. Vegetation indices

The NDVI-based vegetation indices were computed in this study from all possible two band combinations of WorldVeiw-2 bands ($n = 8$). NDVI was selected because it is the most commonly used in estimating biomass and crop yield (Cho et al., 2007; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). The discrete 8 bands allowed a computation of $N \times N = 64$ NDVI indices as follows:

$$\mathrm{NDVI} = \frac{R_{(i,n)} - R_{(j,n)}}{R_{(i,n)} + R_{(j,n)}}$$

where $R_{(i,n)}$ and $R_{(j,n)}$ are the reflectance of any two bands from the selected bands for spectral sample (n).

The NDVIs ($n = 56$) were then input into random forest regression to measure the importance of each NDVI in predicting AGB.

2.6.2. Random forest regression

To model the relationship between NDVIs and wetland vegetation biomass, we used random forest (Breiman, 2001) implemented in the "RandomForest" package (Liaw and Wiener, 2002) within R environment software (R Development Core Team, 2009). Random forest (RF) is an ensemble learning technique developed by Breiman (2001) to improve the classification and regression trees (CART) method by combining a large set of decision trees. In random forest regression, each tree is built using a deterministic algorithm by selecting a random set of variables and a random sample from the training dataset (i.e. the calibration data set). Three parameters need to be optimized in RF: *ntree*, the number of regression trees grown based on a bootstrap sample of the observations (the default value is 500 trees); *mtry*, the number of different predictors tested at each node (the default value is 1/3 of the total number of the variables) and; *nodesize*, the minimal size of the terminal nodes of the trees (the default value is one). To find *ntree* and *mtry* values that can best predict the wetland biomass, the two parameters (*mtry* and *ntree*) were optimized based on the root mean square error of calibration (RMSEC). The *ntree* values were tested from 500 to 9500 with 1000 interval (Prasad et al., 2006), while *mtry* was tested from 1 to 25 using a single interval. The default *nodesize* was accepted throughout the analysis (Vincenzi et al., 2011). RF regression model performs as follows (for full details see Breiman, 2001):

(1) *ntree* bootstrap sample X_i (i = bootstrap iteration) are randomly drawn with replacement from the original dataset (calibration dataset), each containing approximately one third of the elements of the calibration dataset X ($n = 57$). The elements not included in X_i are referred to as out-of-bag data (OOB) for that bootstrap sample.

(2) Unlike classic regression tree, for each bootstrap sample an un-pruned regression tree is grown with the modification that at each node, one third of the predictor variables is randomly selected and the best split from among those variables is chosen.

(3) At each bootstrap iteration, the response value for data not included in the bootstrap sample (OOB data) is predicted and averaged over all trees (*ntree*).

(4) The importance of each predictor is measured by calculating the *percent* increase in mean squared error (RMSEC) when OOB data for each variable are permuted, while all others are unchanged. These variable importance values are then used to rank the predictors in terms of the strength of their relationship to the response variables.

We used an independent dataset ($n = 25$) to validate the predictive performance of the random forest regression model (Ismail and Mutanga, 2010; Vincenzi et al., 2011). The performance of random forest regression in predicting ABG was then compared against the widely used stepwise multiple linear regression (Kokaly and Clark, 1999; Kumar et al., 2001; Tian et al., 2012; Xiaojun, 2012; Yang et al., 2009). The entire NDVIs ($n = 56$) calculated from WorldView-2 bands ($n = 8$) was used to train the multiple regression model. A regression equation developed from the training data set ($n = 57$) was then used to predict wetland biomass on an independent test data set ($n = 25$).

2.6.3. Variables selection

After ranking the predictor variables with RF, the challenge was to select the fewest number of predictors that offer the best predictive power and help in the interpretation of the final model (Ismail and Mutanga, 2010). In this regard, a backward feature elimination method (BFE) integrated with random forest regression as part of the evaluation process was implemented (Guyon and Elisseeff, 2003; Ismail and Mutanga, 2010). The method starts with the entire NDVIs ($n = 56$) and then progressively eliminates the least promising variable (NDVI). At each iteration ($n = 56$), the model is optimized by selecting best *mtry* and *ntree*, the least promising variable (NDVI) is eliminated and root mean square error of calibration is calculated. The smallest subset of variables with lowest RMSEC was then selected to predict the wetland biomass. We further evaluated the selection progress using 10-fold cross validation (Kohavi and John, 1997).

Table 1
Descriptive statistics of the measured above ground biomass (kg/m^2).

Sample No	Mean	Standard deviation	Minimum	Maximum	Range
82	3.4365	0.913376	1.524	4.995	3.471

A comprehensive analysis of the predictive performance of different subsets of DNVIs was implemented to explore the role of the new WorldVeiw-2 bands (red-edge and coastal blue) in predictive wetland biomass as well as to test if the variables selection method implemented in this study can enhance the predictive performance of random forest regression model.

3. Results

3.1. Biomass of wetland vegetation (kg/m^2)

Table 1 shows descriptive statistics of vegetation biomass. High biomass was observed due to predominance of papyrus (*Cyperus papyrus*) in the area. The average biomass obtained was 3.4365 kg/m^2 which is much higher than the standard saturation level biomass reported (Mutanga and Skidmore, 2004a).

3.2. Optimization of random forest regression models

The results of random forest parameters (*mtry* and *ntree*) are shown in Fig. 2. The optimization was done using the calibration dataset ($n = 57$) and RMSEC. Fig. 2 clearly indicates that RF random forest parameters (*ntree* and *mtry*) affect the error of prediction. The high number of *ntree* and the default setting of *mtry* (i.e., *mtry* = 1/3 of the total number of the variables) in our case $n = 19$, yielded the lowest RMSEC (0.445 kg/m^2). The high number of trees (6500–9500) produced the lowest prediction errors with different *mtry* values.

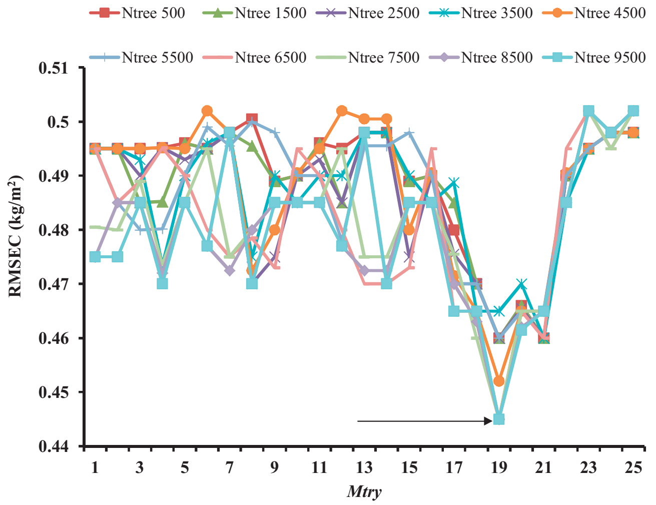

Fig. 2. Optimization of random forest parameters (*ntree* and *mtry*) using RMSEC. The optimal *ntree* and *mtry* that yielded the lowest RMSEC is shown with the black arrow.

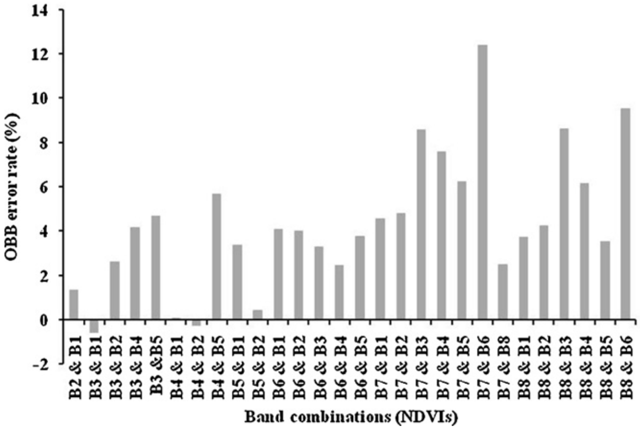

Fig. 3. Measuring the variables importance (NDVIs) in predicating AGB using RF regression. The model developed using 19 *mtry* and 9500 *ntree*. Higher % OOB error indicates greater variable importance.

3.3. Variables importance measures

The OOB estimates of error rate were used to measure the importance of NDVIs developed from all possible two band combinations. The random forest model was able to explore and rank the predictors by their importance in estimating AGB. Fig. 3 shows the variable importance measured in terms of the increase of OOB error, which represents the deterioration of the predictive performance of the model when each predictor is permuted. Few of the predictors (NDVIs) contributed noticeably to the estimation of AGB, namely the band combinations involving the red edge band (band 6) and NIR bands (band 7 and 8).

The variable selection method used in this study was able to identify the smallest number of NDVIs that would offer the best predictive ability of the random forest. It is worth to note that the RMSEC generally decreased while the least important variables were discarded progressively. The use of three variables (NDVIs) produced lowest RMSEC using OOB (0.466 kg/m^2) and 10-fold cross validation (0.472 kg/m^2), while the use of entire variables (n = 56) produced highest RMSEC for both OOB error method (0.5817 kg/m^2) and 10-fold cross validation (0.5905 kg/m^2) (Fig. 4). The best NDVIs that produced the lowest RMSEC are calculated from a combination of the red-edge band (band 6) and the NIR1, 2 (band 7 and 8) as well as NDVI derived from band 7 and band 3 (green). Three indices (n = 3) were then used as input predictive variables in random forest regression to predict the wetland biomass using the test dataset.

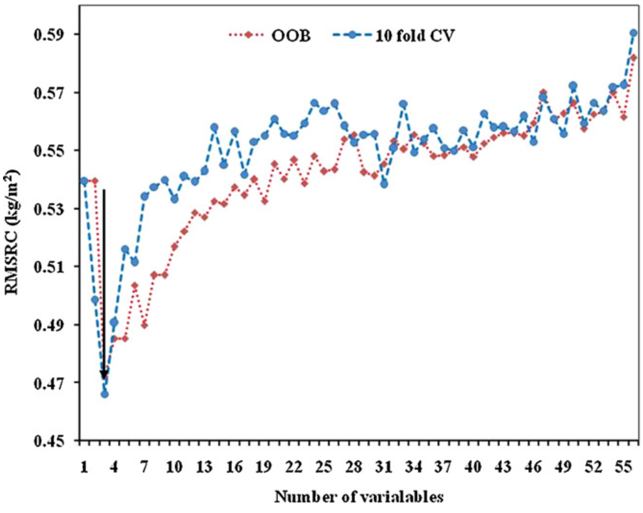

Fig. 4. Selecting the optimal number of variables (NDVIs) using backward elimination search function. The RMSEC is calculated from the calibration dataset (n = 56) using OOB method and 10-fold cross validation.

3.4. Predictive performance of the random forest regression

A comparative analysis of the predictive performance of the optimal number of NDVIs (n = 3), best NDVI (n = 1) calculated from red-edge and shoulder bands (band 6 and band 7), and the standard NDVI involving the red and NIR bands is presented in Table 2. The random forest regression produced the highest R^2 (0.76) and the lowest root mean square error of prediction (RMSEP) (0.441 kg/m^2) using the three NDVIs compared to the top NDVIs which produced R^2 0.63 and RMSEP of 0.505.1 kg/m^2 and the standard NDVI, which yielded an R^2 of 0.31 and RMSEP of 0.858.1 kg/m^2. Table 2 also indicates that the standard NDVI computed from the World View-2 near-infrared band (770–895 nm) and the red band (630–690 nm) resulted in poor prediction performance compared with the NDVI that involved the near-infrared band (770–895 nm) and the red-edge band (705–895 nm).

One to one relationship between actual and predicted biomass using RF regression models is shown in Fig. 5. For each model, R^2, RMSEC, and RMSRP were reported. The use of random forest and standard NDVI provided poor prediction for high biomass (more than 3.000 kg/m^2) compared to selected NDVIs which include the red-edge band (Fig. 5).

With regard to the performance of stepwise multiple regression, the results show that the use of the entire NDVIs (n = 56) resulted in selection of a model with two NDVIs involving a combination of the red-edge (band 6) with the NIR1 (band 7) and NIR2 (band 8). These NDVIs yielded an R^2 of 0.69 and RMSEP of 0.5465 kg/m^2. The performance of the non-linear random forest predictive model of the selected of NDVIs (n = 3) was better compared to stepwise multiple regression model (Table 2).

4. Discussion

The complexity of species composition and dense vegetation in tropical wetland areas introduces a challenge for remote sensing (Adam et al., 2010). Although the use of optical sensors with different types of spectral and spatial resolution (i.e. fine, medium and coarse) has achieved different degrees of success for AGB estimation, some drawbacks and limitations have been reported in previous studies (Lu, 2006). The use of fine spatial and spectral resolution sensors (less than 5 m and more than 100 bands) in estimating biomass is limited in terms of cost, availability, processing and high dimensional data involved. These limitations prohibit its application in large areas (Lu, 2006; Thenkabail et al., 2004), while the estimation of biomass has also been constrained by the asymptotic nature of the relationship between biomass and NDVI computed from medium spatial-resolution (10–100 m) multispectral sensors using the NIR and Red bands (Kumar et al., 2001; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). This paper aimed at investigating the use of a new generation multispectral sensor (WorldVeiw-2) in estimating biomass in densely vegetated environments.

4.1. The relationship between WorldView-2 vegetation indices and biomass

Results of two possible WorldVeiw-2 band combinations presented in this study have shown that the standard NDVI computed from the red absorption maximum (630–690 nm) and the NIR (770–895 nm) correlate poorly with biomass. The poor correlations between the standard NDVI indices and biomass are a reflection of the saturation level reached on dense vegetation (Chen et al., 2009a; Mutanga and Skidmore, 2004a; Thenkabail et al., 2000). When

Table 2
Performance of the random forest model for prediction of wetland vegetation biomass using different subsets of NDVIs.

NDVIs	Calibration ($n = 57$)			Independent validation ($n = 25$)		
	R^2 actual vs. predicted	RMSEC (kg/m^2)	Mean %	R^2 actual vs. predicted	RMSEP (kg/m^2)	Mean %
Standard NDVI (red and NIR1 nm)	0.39	0.656	19.1	0.31	0.858	25.0
Most important NDVI (red-edge and NIR1)	0.67	0.528	15.4	0.63	0.505	14.7
Selected NDVIs ($n = 3$)	0.79	0.439	12.8	0.76	0.442	12.9

canopy cover reaches 100%, the amount of red light (630–690 nm) that can be absorbed by vegetation reaches a peak while NIR reflectance continue to increase due to multiple scattering effects. This mismatch results in poor relationship between biomass and the NDVI (Mutanga and Skidmore, 2004b; Tucker, 1977). Table 1 indicates that the minimum biomass recorded in this study exceeds the biomass threshold (1.5 kg/m^2) of saturation.

On the other hand, results from this study indicate that the wetland biomass is significantly correlated with the NDVIs involving the new red-edge band of WorldView-2. The red-edge is one of the additional spectral bands of WorldView-2 that detects energy on a narrow band between 705 and 745 nm at 1.84 m spatial resolution. This band, which has not been in any commercial moderate-resolution satellite before, provides very sensitive measurements

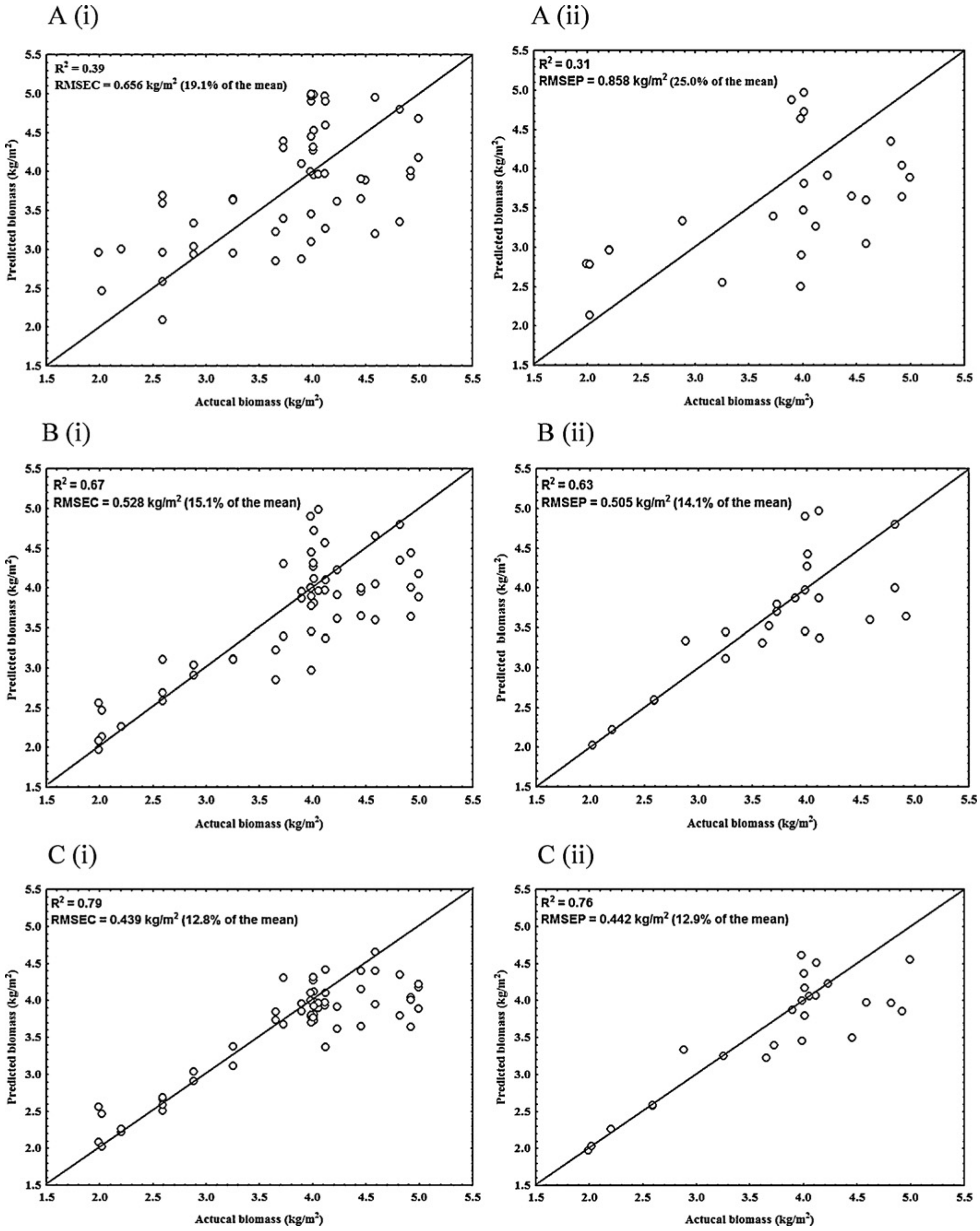

Fig. 5. Relationship between actual and predicted biomass of wetland vegetation for (i) calibration ($n = 57$) and (ii) validation ($n = 25$) analyses using random forest regression model. The regression model was developed using (a) standard NDVI computed from red and NIR1 bands, (b) the most important NDVI developed in this study, and (c) the best three NDVIs selected in present study.

of red-edge reflectance and therefore strengthens the performance of WorldView-2 in measuring the vegetation parameters such as biomass (Ozdemir and Karnieli, 2011). Our study in this regard supports the assertion that WorldView-2 multispectral imagery contains more spectral biomass information as compared to other multispectral satellite data.

Some possible explanations for the better performance of the red edge band as compared to the traditional NDVI is that the indices calculated from red-edge are more sensitive to properties of vegetation such as canopy biomass and chlorophyll content as compared to other regions of the electromagnetic spectrum (Mutanga and Skidmore, 2004a,b). A slight change in these vegetation properties results in a shift in the red edge curve. Furthermore, vegetation indices computed from the red-edge and NIR can minimize the influence of the atmospheric and water absorption and soil background (Kokaly and Clark, 1999). From this background, the additional spectral bands of WorldView-2 were able to estimate biomass at high canopy density as compared to traditional multispectral indices computed using the red and NIR bands.

The results from this study are consistent with those by Cho et al. (2007) who found better predictive performance of the red-edge extracted from airborne HyMap imagery in estimating grass/herb biomass in the Majella National Park. Mutanga and Skidmore (2004a,b) in their study on characterizing pasture biomass in densely vegetated areas also indicated that the red edge (700–750 nm) and longer wavelengths of the red edge (750–780 nm), yielded higher correlations with biomass ($R^2 = 0.77$) than the standard NDVI. By using WorldView-2 multispectral imagery, results from this study go a long way in overcoming the difficulties of using multispectral data associated with the lack of strategically positioned bands for biomass estimation. Such results could also provide tremendous saving of financial resources as well as time spent in hyperspectral data acquisition and processing. We caution, however, that such conclusion based on empirical models might be site or sensor specific and unsuitable for application to large areas or different seasons (Gobron et al., 1997).

4.2. Prediction performance of RF using different subset of NDVIs

Random forest has increasingly been used as a classifier in remote sensing applications (Adam et al., 2012; Chan and Paelinckx, 2008; Gislason et al., 2006; Lawrence et al., 2006; Pal, 2005; Stumpf and Kerle, 2011). However, only a few studies have investigated the use of random forest in regression type of remote sensing applications (e.g. Abdel-Rahman et al., 2009; Ismail and Mutanga, 2010). The present study showed that the utility of random forest regression model based on WorldVeiw-2 imagery provides an effective methodology for variable selection and predicting biomass in wetlands environment.

Results from the present study demonstrated that random forest regression is a useful and a robust method for remote sensing applications. The ability of automatically producing accuracy assessments and measuring the variable importance make random forest effective algorithms. Despite the advantages of random forest regression model as detailed in Breiman (2001), the RF regression has a limitation, especially on the way in which regression trees are constructed. In this study, the random forest tended to underestimate the high biomass values that fall beyond the range (Fig. 5) of the training data set, a subject for further studies using additional data sets.

5. Conclusion

The major conclusion of our work is that the NDVIs extracted from WorldVeiw-2 can be used to model and predict wetland biomass in a high density and vegetated wetland. The NDVIs involving the additional spectral bands of WorldView-2, such as the red-edge and near infrared regions of the electromagnetic spectrum, improve the prediction accuracy compared with the traditional NDVIs.

Random forest regression was able to provide a small subset of variables and reasonable prediction accuracies. We recommend further experiments are conducted using different data sets to assess the performance of the random forest algorithm in predicting values that are beyond the range of the training data set.

Acknowledgments

The authors are thankful to DigitalGlobe for accepting our proposal and providing such a high-quality dataset of WorldView-2 at no charge. We thank the University of KwaZulu-Natal for providing the optimal conditions for the success of the study as well as the NRF for partly funding this study. Gratitude is extended to the R Development Team for the open source package for the statistical analysis. We also thank the anonymous reviewers for their valuable comments and suggestions that greatly improved the manuscript.

References

Abdel-Rahman, E.M., van den Berg, M., Way, M.J., Ahmed, F.B., 2009. Hand-held spectrometry for estimating thrips (*Fulmekiola serrata*) incidence in sugarcane. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Cape Town, South Africa, pp. IV-268–IV-271.

Adam, E., Mutanga, O., 2009. Spectral discrimination of papyrus vegetation (*Cyperus papyrus* L.) in swamp wetlands using field spectrometry. ISPRS Journal of Photogrammetry and Remote Sensing 64, 612–620.

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

Adam, E., Mutanga, O., Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. Wetlands Ecology and Management 18, 281–296.

Adam, E.M., Mutanga, O., Rugege, D., Ismail, R., 2012. Discriminating the papyrus vegetation (*Cyperus papyrus* L.) and its co-existent species using random forest and hyperspectral data resampled to HYMAP. International Journal of Remote Sensing 33, 552–569.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Chan, J.C.-W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 112, 2999–3011.

Chen, J., Gu, S., Shen, M., Tang, Y., Matsushita, B., 2009a. Estimating aboveground biomass of grassland having a high canopy cover: an exploratory analysis of in situ hyperspectral data. International Journal of Remote Sensing 30, 6497–6517.

Chen, J., Song, G., Shen, M., Tang, Y., Matsushita, B., 2009b. Estimating aboveground biomass of grassland having a high canopy cover: an explanatory analysis of in situ hyperspectral data. International Journal of Remote Sensing 24, 6497–6517.

Cho, M., Skidmore, A., Corsi, F., van Wieren, S., 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.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognition Letters 27, 294–300.

Gobron, N., Pinty, B., Verstraete, M.M., 1997. Theoretical limits to the estimation of the leaf area index on the basis of visible and near-infrared remote sensing data. IEEE Transactions on Geoscience and Remote Sensing 35, 1438–1445.

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

Guyon, I., Elisseeff, A., 2003. An introduction to varaible and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Ham, J., Chen, Y., Crawford, M., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43, 492–501.

Hoffer, R.M., 1978. Biological and physical considerations in applying computer-aided analysis techniques to remote sensor data. In: Swain, P.H., Davis, S.M. (Eds.), Remote Sensing: The Quantitative Approach. McGraw-Hill International Book Company, New York, pp. 227–289.

Horler, D.N.H., Dockray, M., Barber, J., 1983. The red edge of plant leaf reflectance. International Journal of Remote Sensing 4, 273–288.

Hurcom, S.J., Harrison, A.R., 1998. The NDVI and spectral decomposition for semi-arid vegetation abundance estimation. International Journal of Remote Sensing 19, 3109–3125.

Ismail, R., Mutanga, O., 2010. 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 12, S45–S51.

Ismail, R., Mutanga, O., Bob, U., 2006. The use of high resolution airborne imagery for the detection of forest canopy damage by *Sirex noctilio*. In: Langin, P.A., Antonides, M.C. (Eds.), Precision Forestry in Plantations, Semi-Natural Areas and Natural Forest: Proceedings of the International Precision Forestry Symposium. Stellenbosch University, Stellenbosch University, South Africa, pp. 119–134.

Ismail, R., Mutanga, O., Kumar, L., 2010. Modeling the potential distribution of pine forests susceptible to sirex noctilio infestations in Mpumalanga, South Africa. Transactions in GIS 14, 709–726.

Kohavi, R., John, G., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Kokaly, R.F., Clark, R.N., 1999. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. Remote Sensing of Environment 67, 267–287.

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

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

Leica Geosystems, 2004. Leica Geosystems GS20 Field Guide-1.1.0en. Leica Geosystems AG CH-9435 Heerbrugg, Switzerland.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Lu, D., 2006. The potential and challenge of remote sensing-based biomass estimation. International Journal of Remote Sensing 27, 1297–1328.

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

Maclean, I., Hassall, M., Boar, R., Lake, I., 2006. Effects of disturbance and habitat loss on papyrus-dwelling passerines. Biological Conservation 131, 349–358.

Mafabi, P., 2000. The role of wetland policies in the conservation of waterbirds: the case of Uganda. Ostrich 71.

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

Mutanga, O., Skidmore, A.K., 2004b. Hyperspectral band depth analysis for a better estimation of pasture biomass. International Journal of Applied Earth Observation and Geoinformation 5, 87–96.

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

Omar, H., 2010. Commercial Timber Tree Species Identification Using Multispectral Worldview2 Data. Digital Globe® 8Bands Research Challenge.

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

Ozdemir, I., Karnieli, A., 2011. Predicting forest structural parameters using the image texture derived from WorldView-2 multispectral imagery in a dryland forest, Israel. International Journal of Applied Earth Observation and Geoinformation 13, 701–710.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26, 217–222.

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

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sridharan, H., 2010a. Multi-level urban forest classification using the WorldView-2 8-band hyperspectral imagery. Digital Globe® 8Bands Research Challenge.

Sridharan, H., 2010b. Multi-level urban forest classification using the WorldView-2 8-band hyperspectral imagery. Digital Globe® 8Bands Research Challenge.

Steininger, M., 2000. Satellite estimation of tropical secondary forest above-ground biomass: data from Brazil and Bolivia. International Journal of Remote Sensing 21, 1139–1157.

Stumpf, A., Kerle, N., 2011. Combining Random Forests and object-oriented analysis for landslide mapping from very high resolution imagery. Procedia Environmental Sciences 3, 123–129.

Taylor, R.H., 1995. St-Lucia Wetland Park. Struik Publishers, Cape Town, South Africa.

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

Thenkabail, P.S., Stucky, N., Griscom, B.W., Ashton, M.S., Diels, J., Van Der Meer, B., Enclona, E., 2004. Biomass estimations and carbon stock calculations in the oil palm plantations of African derived savannas using IKONOS data. International Journal of Remote Sensing 25, 5447–5472.

Tian, X., Su, Z., Chen, E., Li, Z., van der Tol, C., Guo, J., He, Q., 2012. Estimation of forest above-ground biomass using multi-parameter remote sensing data over a cold and arid area. International Journal of Applied Earth Observation and Geoinformation 14, 160–168.

Todd, S.W., Hoffer, R.M., Milchunas, D.G., 1998. Biomass estimation on grazed and ungrazed rangelands using spectral indices. International Journal of Remote sensing 19, 427–438.

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

Updike, T., Comp, C., 2010. Radiometric Use of WorldView-2 Imagery, DigitalGlobe. Technical Note. DigitalGlobe®, Colorado, USA.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A., Torricelli, P., 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecological Modelling 222, 1471–1478.

Wang, C., Menenti, M., Stoll, M., Belluco, E., Marani, M., 2007. Mapping mixed vegetation communities in salt marshes using airborne spectral data. Remote Sensing of Environment 107, 559–570.

Wiegand, C.L., Richardson, A.J., Escobar, A.J., Gerbermann, A.H., 1991. Vegetation indices in crop assessments. Remote Sensing of Environment 35, 105–119.

Xiaojun, Y., 2012. An assessment of landscape characteristics affecting estuarine nitrogen loading in an urban watershed. Journal of Environmental Management 94, 50–60.

Yang, X.-H., Wang, F.-M., Huang, J.-F., Wang, J.-W., Wang, R.-C., Shen, Z.-Q., Wang, X.-Z., 2009. Comparison between radial basis function neural network and regression model for estimation of rice biophysical parameters using remote sensing. Pedosphere 19, 176–188.