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