THE CROP JOURNAL 4 (2016) 212–219

Available online at www.sciencedirect.com

ScienceDirect

Estimation of biomass in wheat using random forest regression algorithm and remote sensing data

Li'ai Wang[a], Xudong Zhou[b], Xinkai Zhu[a], Zhaodi Dong[a], Wenshan Guo[a,*]

[a]Key Laboratory of Crop Genetics and Physiology of Jiangsu Province, Yangzhou University, Yangzhou 225009, China
[b]Information Engineering College of Yangzhou University, Yangzhou 225009, China

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form
29 January 2016
Accepted 15 March 2016
Available online 30 March 2016

Keywords:
Above-ground dry biomass
Triticum aestivum
Vegetation indices
Wheat

ABSTRACT

Wheat biomass can be estimated using appropriate spectral vegetation indices. However, the accuracy of estimation should be further improved for on-farm crop management. Previous studies focused on developing vegetation indices, however limited research exists on modeling algorithms. The emerging Random Forest (RF) machine-learning algorithm is regarded as one of the most precise prediction methods for regression modeling. The objectives of this study were to (1) investigate the applicability of the RF regression algorithm for remotely estimating wheat biomass, (2) test the performance of the RF regression model, and (3) compare the performance of the RF algorithm with support vector regression (SVR) and artificial neural network (ANN) machine-learning algorithms for wheat biomass estimation. Single HJ-CCD images of wheat from test sites in Jiangsu province were obtained during the jointing, booting, and anthesis stages of growth. Fifteen vegetation indices were calculated based on these images. In-situ wheat above-ground dry biomass was measured during the HJ-CCD data acquisition. The results showed that the RF model produced more accurate estimates of wheat biomass than the SVR and ANN models at each stage, and its robustness is as good as SVR but better than ANN. The RF algorithm provides a useful exploratory and predictive tool for estimating wheat biomass on a large scale in Southern China.

1. Introduction

Biomass is one of the most useful indicators of crops vegetation development and health. Measuring biomass directly is a destructive and expensive procedure. More recent estimates are based on remotely sensed data, such as vegetation indices (VIs) [1–4]. Kross et al. [1] established relationships between corn biomass and VIs such as the NDVI (Normalized Difference Vegetation Index), Green-NDVI, RVI (Ratio Vegetation Index), and MTVI2 (Modified Triangular Vegetation Index 2) computed from the SPOT and Landsat images. Gnyp et al. [3] found that SAVI (Soil-Adjusted Vegetation Index), OSAVI (Optimized Soil-Adjusted Vegetation Index), and MTVI2 had stronger relationships with rice biomass at the jointing stage than that at booting. Gao et al. [4] proposed that maize biomass could be estimated by VIs

* Corresponding author.
E-mail addresses: wla001@163.com (L. Wang), guows@yzu.edu.cn (W. Guo).
Peer review under responsibility of Crop Science Society of China and Institute of Crop Science, CAAS.

http://dx.doi.org/10.1016/j.cj.2016.01.008

calculated using Chinese environmental satellite (HJ) images [e.g. NDVI, RVI, and the enhanced vegetation index (EVI)]. Jin et al. [5] reported that the estimation accuracy of wheat biomass was better using a combination of VIs and radar polarimetric parameters (RPPs) than using VIs or RPPs alone.

Remote estimation of biomass requires application of diverse methods and techniques. In recent years machine-learning algorithms were trialed for ability to perform flexible input–output nonlinear mappings between remotely sensed data and biomass [6–8]. Typically, artificial neural networks (ANNs) and support vector regressions (SVRs) were employed to couple with VIs to build monitoring models with improved prediction accuracy of remote estimation of biomass in crops. For instance, Wang et al. [9] provided an effective model for assessing the biomass of wheat with ANNs and VIs (i.e. RVI, NDVI, GNDVI, SAVI, OSAVI, RDVI) calculated based on ASD FieldSpec data. Clevers et al. [10] estimated grassland biomass using SVRs and VIs such as the RVI, NDVI, WDVI, SAVI, GEMI (Global Environmental Monitoring Index), and EVI (Enhanced Vegetation Index) calculated based on ASD FieldSpec data.

Among various machine-learning algorithms, the emerging Random Forest (RF) algorithm proposed by Leo Breiman and Cutler Adele in 2001 has been regarded as one of the most precise prediction methods for classification and regression, as it can model complex interactions among input variables and is relatively robust in regard to outliers. The RF algorithm presents several advantages; it runs efficiently on large datasets, it is not sensitive to noise or over-fitting [11], it can handle thousands of input variables without variable deletion, and it has fewer parameters compared with that of other machine-learning algorithms (e.g. ANN or SVR). The RF classification algorithm has been applied to many remote sensing domains such as land cover classification [12–14] and other fields related to the environment and water resources [15–16]. To our knowledge, only a few studies have reported the use of the RF regression algorithm in remote sensing applications, including monitoring of forest growth, wetland vegetation, and water resources [6,17–18]. Furthermore, few studies have employed the RF regression algorithm based on VIs for estimating the biomass of winter wheat.

The major objectives of this study were to: (i) investigate the applicability of the RF regression algorithm in combination with VIs to remotely estimate wheat biomass, (ii) test the performance of RF regression for estimating biomass, and (iii) compare the performance of RF with that of other machine-learning algorithms for the estimation of wheat biomass. Specifically, based on VIs calculated from China's environmental satellite (HJ) charge-coupled device (CCD) images, we employed the RF algorithm to construct models to estimate wheat biomass, and then, the RF algorithm was compared with the SVR and ANN machine-learning algorithms in terms of accuracy, goodness of fit, and robustness for estimating wheat biomass.

2. Data source

2.1. Experimental design and data acquisition

Experiments were carried out in four counties (YiZheng, JiangYan, GaoYou and TaiXing) of Jiangsu province during the winter wheat growing seasons of 2010, 2011, 2012 and 2014. The local wheat cultivars were Yangmai 13, Yangmai 15, Yangmai 16, and Yangfumai 2. For each year's experiment, fifteen sample sites were established in each county and a plot of 30 × 30 m was randomly demarcated at each site. Within each plot, five subplots of 0.4 m × 0.4 m were established at least 10 m from each other. During three growth stages (jointing, booting and anthesis) wheat plants from each subplot (positions determined with a Global Positioning System GPS, Trimble GeoExplorer 2008 Series GeoXH, Trimble Navigation Limited, USA) at each site were collected, sealed in plastic bags, and sent to a laboratory for analysis. In the laboratory, the wheat plants from each subplot were dried in an oven at 80 °C for 48 h, after which the dry weight was determined. The dry weight was divided by the surface area of the subplot, and then the weight was converted to kg ha^{-1}. The biomass values of plants from the five subplots within each plot were averaged to represent the biomass of the entire plot.

For each stage, the pooled data from 2010, 2011, 2012 and 2014 were randomly divided into a training dataset and an independent test dataset (75% and 25% of the pooled data, respectively). For the training dataset, the number of samples was 174 at jointing, 174 at booting, and 147 at anthesis. For the test dataset, the number of samples was 58 at jointing, 58 at booting, and 49 at anthesis. The training dataset was used to establish models to predict biomass during each growth stage, and the test dataset was used to test the quality and reliability of each prediction model.

2.2. Remote sensing data and preprocessing

Remotely sensed data (HJ satellite charge-coupled device) of wheat from the three stages were retrieved online from the China Centre for Resources Satellite Data and Application (CRESDA). The HJ satellite charge-coupled device (HJ-CCD) satellite system is China's environmental disaster and environmental monitoring satellite system. It includes two optical satellites, HJ-1A and HJ-1B, which are symmetrically equipped with two CCD cameras. They comprise four multispectral bands with a 30-m resolution and a 720-km swath. The spectral ranges of the four bands are 430–520 nm (B_1-blue), 520–600 nm (B_2-green), 630–690 nm (B_3-red) and 760–900 nm (B_4-near infrared).

All HJ-CCD image data used in this study were completely corrected using ENVI4.7 remote sensing image processing software. Ground control points were located with a differential GPS unit during the field experiments. The map projection used a geographic coordinate system (Lat/Lon) as the projection type (WGS84) and a pixel size of 30 m × 30 m. A radiometric calibration was conducted using the HJ satellite calibration coefficients (e.g. gains and offsets). Atmospheric corrections were conducted using the MOTRAN 4 model embedded in the ENVI/FLAASH module of ENVI 4.7 software, and the input parameters were set based on the location, sensor type and ground weather conditions observed on the day each image was acquired. To improve the accuracy of pixel registration to within one pixel, coarse geometric corrections were made based on the 1:10,000 digitized raster map, after which, precise geometric corrections were made based on the GPS ground control points.