Remote Sensing of Environment 125 (2012) 80–91

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

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

Forest biomass estimation from airborne LiDAR data using machine learning approaches

Colin J. Gleason [a], Jungho Im [b,c,*]

[a] Department of Geography, University of California, Los Angeles, CA, USA
[b] School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689-798, South Korea
[c] Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, Syracuse, NY 13210, USA

ARTICLE INFO

Article history:
Received 17 August 2011
Received in revised form 9 July 2012
Accepted 11 July 2012
Available online 5 August 2012

Keywords:
Biomass estimation
Lidar remote sensing
Machine learning
Random forest
Cubist
Support vector regression
Linear mixed effects regression
Tree crown delineation

ABSTRACT

During the past decade, procedures for forest biomass quantification from light detection and ranging (LiDAR) data have been improved at a rapid pace. The scope of these methods ranges from simple regression between LiDAR-derived height metrics and biomass to methods including automated tree crown delineation, stochastic simulation, and machine learning approaches. This study compared the effectiveness of four modeling techniques—linear mixed-effects (LME) regression, random forest (RF), support vector regression (SVR), and Cubist—for estimating biomass in moderately dense forest (40–60% canopy closure) at both tree and plot levels. Tree crowns were delineated to provide model estimates of individual tree biomass and investigate the effects of delineation accuracy on biomass modeling. We used our previously developed method (COTH) to delineate tree crowns. Results indicate that biomass estimation accuracy improves when modeled at the plot level and that SVR produced the most accurate biomass model (671 kg RMSE per 380 m^2 plot when forest plots were modeled as a collection of trees). All models provided similar results when estimating biomass at the individual tree level (505, 506, 457, and 502 kg RMSE per tree). We assessed the effect of crown delineation accuracy on biomass estimation by repeating the modeling procedures with manually delineated crowns as inputs. Results indicated that manually delineated crowns did not always produce superior biomass models and that the relationship between crown delineation accuracy and biomass estimation accuracy is complex and needs to be further investigated.

1. Introduction

Quantifying the amount of biomass within a forest stand is necessary for property managers to make informed decisions about the value and use of their forested land. Biomass quantification procedures developed for remotely sensed data, especially from LiDAR data, have been published at a rapid pace with an increasing complexity and variety of techniques (Gleason & Im, 2011). LiDAR data are well suited to biomass estimation, as point clouds generated from forest canopies can accurately depict the physical characteristics of the canopy surface. These physical characteristics are correlated with biomass, and may be regressed against either diameter at breast height (dbh) or biomass to obtain general LiDAR-biomass models (Salas et al., 2010; Zhao et al., 2009). More recently, biomass quantification procedures have moved away from the regression between LiDAR-derived height metrics and biomass and increasingly include methods for automated tree crown delineation, stochastic simulation, and machine learning (Breidenbach et al., 2010; Muinonen et al., 2001; Salas et al., 2010; Vauhkonen et al., 2010).

Biomass quantification models (regardless of method) must model forests in practical units, which is accomplished either by estimating biomass for individual trees or for semi-arbitrary areas of forest (i.e., plots). Identifying treetops is often the first step in locating individual trees, as biomass is strongly correlated with crown width and other crown dimensions that can be derived from treetop position (Popescu, 2007). This requires accurate field data describing the position and height of each tree within a study plot as well as measurements of crown dimensions. Without such field data, treetops may be identified from LiDAR data, often through the process of local maxima filtering. There are numerous methods of local maxima detection involving varying search window sizes based on tree height (Bunting & Lucas, 2006; Chen et al., 2006, 2007; Jang et al., 2008; Kwak et al., 2010; Persson et al., 2002; Popescu, 2007; Popescu & Wynne, 2004; Zhao et al., 2009). Crown dimensions should be measured in the field to assess the accuracy of local maxima filtering. If such field data are not available, crown dimensions may be obtained through interpretation of image data or a LiDAR-generated

* Corresponding author at: School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689–798, South Korea. Tel.: +82 52 217 2824; Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, Syracuse, NY 13210, USA. Tel.: +1 315 470 4709.
E-mail addresses: ersgis@unist.ac.kr, imj@esf.edu (J. Im).

0034-4257/$ – see front matter
doi:10.1016/j.rse.2012.07.006

canopy surface (Chen et al., 2006; Ke et al., 2010; Wang et al., 2004). Accurately obtaining these crown dimensions is critical as these measurements constitute the parameters from which biomass is derived.

The delineation of tree crowns is itself a robust and growing field of study, and crown delineation methods are explained in greater detail in Gleason and Im (2012). Of particular interest are those methods that use the concept of tree crowns as geometric volumes, sometimes called crown/canopy geometric volume (CGV) (Kato et al., 2009). These methods associate LiDAR heights contained within a crown footprint to determine CGV, and produce more accurate representations of volume when forest density is sparse or trees are isolated (Brandtberg, 2007; Breidenbach et al., 2010; Chen et al., 2007; Kato et al., 2009; Kwak et al., 2010). The CGV can extend downward from the canopy surface to either the forest floor or height thresholds that are either species-specific or LiDAR-derived. The concept of a CGV is intuitive in its correlation to biomass, but has not yet been proven a sufficient metric to act as a sole model predictor variable in dense forests. Dense forest conditions often require multiple LiDAR derived variables to estimate biomass, and the accuracy of crown delineation/biomass estimation in such forests tends to be lower for complex canopy surfaces than homogenous forests (Chen et al., 2007).

An alternative to biomass estimation at the individual tree level is biomass estimation at the plot level. Studies employing this technique frequently characterize LiDAR data through statistical descriptions of canopy height, number of LiDAR returns, and ratios of returns. These descriptors can also be used to estimate other forest biophysical parameters (Anderson et al., 2006; Donoghue et al., 2007; Hawbaker et al., 2010; Hyde et al., 2006; Ioki aet al., 2010; Næsset, 2004; Næsset & Økland, 2002; Popescu et al., 2002, 2004; Solberg et al., 2010). Modeling biomass using this approach requires reference biomass data measured at the plot level, which may introduce bias into the modeling procedure. If reference biomass was calculated including snags, woody debris, and understory vegetation, LiDAR first returns may not penetrate denser canopy to a sufficient degree to accurately describe these features (Næsset, 2005).

There are multiple methods used to estimate biomass/tree volume, which are varied in their assumptions and complexity. Sophisticated regression techniques take into account bias and the correlation of predictor variables (e.g., linear mixed effects regression, geographically weighted regression) rather than the somewhat rigid assumptions of ordinary least squares regression (Hudak et al., 2008; Powell et al., 2010; Salas et al., 2010; Yu et al., 2011; Zhao et al., 2009). When comparing statistical regression models, Salas et al. (2010) found that the linear mixed effects (LME) model significantly outperforms geographically weighted regression, ordinary least squares regression, and generalized least squares regression when estimating tree diameter from LiDAR data. Such studies provide evidence that regressing LiDAR-derived variables with field data is an effective method for estimating biomass, yet there is a large set of assumptions and site-specific considerations that must be made for each study. Zhao et al. (2009) also note that scale issues often affect the performance of biomass estimation regression procedure: i.e., models are built to output biomass at a specific plot size and changing this plot size may affect the accuracy of results. To reduce the effects of regression assumptions on plot scale biomass estimation (population assumptions that do not represent the heterogeneity of forest stands), machine learning techniques such as random forest (RF) and most similar neighbor (MSN) may be used (Breidenbach et al., 2010; Muinonen et al., 2001; Vauhkonen et al., 2010).

This study aims at evaluating machine learning approaches—RF, support vector regression (SVR), Cubist ® regression trees—of forest biomass estimation at both individual tree and plot levels using high posting density airborne LiDAR data. Unlike other biomass estimations that estimate tree diameter or volume from LiDAR data and then calculate biomass from this prediction, this study estimates biomass using field-measured biomass to inform the models. Such a choice allows for combined species modeling, as dbh dependent species-specific allometry is applied *a priori*. This modeling also provides more flexibility and may increase the accuracy of estimating deciduous biomass, which is traditionally more difficult to quantify than coniferous biomass. The objectives of this study are to (1) delineate individual trees from airborne LiDAR data, (2) assess impacts of this delineation on biomass estimation, (3) estimate biomass through four different models, LME, RF, SVR and Cubist, and (4) compare the output of these four models for estimating biomass of all trees at the individual tree and the plot level, paying particular attention to the effects of segregating trees on a coniferous/deciduous divide. The different modeling scenarios for plot/tree combinations are referred to as 'schemes,' and are described fully in the methods section.

2. Study area and reference data

2.1. Study area

The study was conducted within the 1700 ha Heiberg Memorial Forest, located in Tully, NY and managed by State University of New York College of Environmental Science and Forestry (SUNY ESF) (Fig. 1). The College maintains continuous forest inventory (CFI) plots within the forest, and these plots were inventoried in summer of 2010, seasonally coincident with an August 10th, 2010 LiDAR collection. Each CFI plot has a radius of 11 m and is located on a 14 chain grid throughout Heiberg Forest. The plots contain coniferous and deciduous trees common to Upstate New York. Table 1 presents a summary of the CFI plots and the field data which were recorded for each tree with a dbh greater than 9.14 cm.

2.2. LiDAR data

Discrete multiple-return LiDAR data for this study was acquired on August 10th, 2010 from an airborne ALS60 sensor (Table 2). Raw laser data was post-processed using the TerraSolid software suite with manual editing by the vendor (Kucera International, Inc.), and resulted in the creation of a canopy point cloud and a bare earth point cloud. Both point clouds were then converted to raster surface data (cell size 0.25 m) in ArcGIS 9.3 using inverse distance weighted (IDW) interpolation, which is a valid surface creation method for LiDAR data with high point density (Popescu et al., 2002). Point density for our study varied across the study site, with an average point density of 12.7 pts/m^2.

2.3. Reference data

Reference data for this study include ground inventory of tree species, dbh, and height. Tree height data was collected in August of 2010 using a Haglof Vertex III hypsometer, and these data were used to visually examine the LiDAR-derived canopy surface data. Leaf area index (LAI) was measured at the plot level using an ACCUPAR LP80 Ceptometer. Forty LAI measurements per plot were averaged to arrive at the LAI used for biomass estimation.

The field team relied on the allometric equations provided by the USDA Forest Service to provide reference biomass levels (Jenkins et al., 2004). These generalized equations were developed for all deciduous and coniferous species in the United States via a thorough canvas of published allometric equations and previous reviews on the subject conducted in 2003 by the same authors. The Jenkins et al. (2003) equations give two parameters that fit their biomass equation, as well as the number of data points used to generate such an equation and the maximum dbh for which the formulation is applicable. These national scale formulations were adopted for this research because the goal was to investigate a transferable model that requires minimal region-specific information, and because the reported root-mean-squared-errors (RMSEs) of the Jenkins et al. (2003) equations provide evidence that