Computers and Electronics in Agriculture 116 (2015) 221–227

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics

Eric Bastos Görgens [a,*], Alessandro Montaghi [b], Luiz Carlos Estraviz Rodriguez [a]

[a] Research Group in LiDAR Technology, Forest Science Department, Luiz de Queiroz College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, 13418-900. Piracicaba, SP, Brazil
[b] Foothills Facility for Remote Sensing and GIScience, Department of Geography, University of Calgary, ES 356, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 22 April 2015
Accepted 6 July 2015
Available online 17 July 2015

Keywords:
Forest quantification
LiDAR
Artificial intelligence
Neural network
Random forest
Support vector regression

ABSTRACT

Machine learning models appear to be an attractive route towards tackling high-dimensional problems, particularly in areas where a lack of knowledge exists regarding the development of effective algorithms, and where programs must dynamically adapt to changing conditions. The objective of this study was to evaluate the performance of three machine learning tools for predicting stand volume of fast-growing forest plantations, based on statistical vegetation metrics extracted from an Airborne Laser Scanning (ALS) survey. The forests used in this study were composed of 1138 ha of commercial plantations that consisted of hybrids of *Eucalyptus grandis* and *Eucalyptus urophylla*, managed for pulp production. Three machine learning tools were implemented: neural network (NN), random forest (RF) and support vector regression (SV); and their performance was compared to a regression model (RM). The RF and the RM presented an RMSE in the leave-one-out cross-validation of 31.80 and 30.56 $m^3\ ha^{-1}$ respectively. The NN and SV presented a higher RMSE than the others, equal to 64.44 and 65.30 $m^3\ ha^{-1}$. The coefficient of determination and bias were similar to all modeling techniques. The ranking of ALS metrics based on their relative importance for the estimation of stand volume showed some differences. Rather than being limited to a subset of predictor variables, machine learning techniques explored the complete metrics set, looking for patterns between them and the dependent variable.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Remote Sensing (RS) has been used as an efficient assessment tool to monitor large forest areas. RS techniques allow the retrieval of spatial data from the environment as trees, roads, stream flow, and other objects located over the ground surface (Zhou et al., 2013). The available expertise in multi-spectral image acquisition, processing, interpretation, and its relatively lower cost have resulted in the high use of this method within forest monitoring activities (Prasad et al., 2011). However, multi-spectral RS encounters problems when assessing vertical information directly (i.e. incorporating a third dimension) since it performs less impressively in sensing structure under medium to high leaf area conditions. The radiometric interference from the surface, the weather conditions, the atmospheric turbidities, and the angles of solar incidence also present problems to multi-spectral RS (Pflugmacher et al., 2012; Proy et al., 1989; Stojanova et al., 2010).

Airborne Laser Scanning (ALS) has been employed to generate Digital Elevation Models (DEM) throughout the last 20 years (Montaghi et al., 2013). Due to its ability to penetrate the forest canopy, ALS technology has become the primary data source for characterizing vertical forest structure (White et al., 2013), and its use has expanded towards new applications such as monitoring vegetation. Based on Light Detection and Ranging (LiDAR) technology, this sensor provides horizontal and vertical information at high spatial resolution and high vertical accuracy that significantly increase our understanding of the real 3D structure of the forest (Næsset, 2004; Patenaude et al., 2004).

The large amount of data constrains the direct use of ALS as an input to the modeling of forest parameters. After collection, the raw ALS data must subsequently be reduced and represented numerically within the calculation of several spatial ALS-metrics that can then be used to create predictive equations for forestry inventory attributes. The number of metrics can easily reach hundreds of variables, and the selection of these metrics remains an empirical process highly dependent on human intervention. After eliminating all but the most descriptive metrics, forest attributes are then estimated through statistical regression analyses that explore the correlation between field measurements and ALS

* Corresponding author. Tel.: +55 19 2105 8676; fax: +55 19 2105 8601.
E-mail addresses: gorgens@usp.br (E.B. Görgens), alessandro.montaghi@gmail.com (A. Montaghi), lcer@usp.br (L.C.E. Rodriguez).

http://dx.doi.org/10.1016/j.compag.2015.07.004
0168-1699/© 2015 Elsevier B.V. All rights reserved.

metrics (Gleason and Im, 2012; Lefsky et al., 1999; Næsset, 1997; Næsset and Bjerknes, 2001; Nelson et al., 1988; Reutebuch et al., 2005; Zhao et al., 2009). Linking ALS-metrics to field data is an effective method for estimating several forest attributes (e.g., stem volume, basal area, biomass, etc.) at the stand or regional level, but there remains a large set of assumptions and site-specific considerations that must be made (Zhao et al., 2011). In fact, a large number of variables could theoretically improve the precision of the models, but models with fewer variables are much easier to interpret (Murphy et al., 2010). It is thus important to develop parsimonious models, mainly because prediction models should be valid for general conditions, and degrees of freedom should not be unreasonably discarded. Finally, large sets of predictor variables often bear strong inter-correlations, which can lead to unstable predictions (Latifi et al., 2010).

Three approaches have been used to select metrics and develop regression models with ALS data. One is to adjust models based on empirical pre-established relationships between field data and ALS metrics established by other studies (Zhao et al., 2011). Another is to determine the best relationship between ALS metrics and field data through optimizing a certain statistical measure (stepwise or exhaustive search) (Næsset, 1997; Patenaude et al., 2004). A third approach involves the use of multivariate statistical analysis, based on the assumption that multicorrelated ALS metrics should be used to estimate forest parameters (Valbuena et al., 2013).

Within the context of remote sensing forest structures via ALS data, studies have shown that the majority of estimation models are not only site- or species-specific, but also scale-dependent, which indicates that the models should be applied to a scale or pixel (cell of the grid) size commensurate to the plot size used in the model fitting (Næsset, 2002). Zhao et al. (2009) demonstrated that it was possible to reduce the effects of plot scale by using machine learning tools.

These machine learning tools are effective when producing software, since they have the ability to tackle high-dimensional problems, poorly understood domains where there is lack of knowledge needed to develop effective algorithms, or domains where programs must dynamically adapt to changing conditions (Mitchell, 1997). An additional advantage is that machine learning allows the user to implement a continuous learning process. Previous remote sensing studies have shown a superior or promising level of performance by artificial intelligence techniques over more classical methods (Atzberger, 2004; Durbha et al., 2007; Fang et al., 2003; Zhao et al., 2011, 2008).

The primary objective of this study is to evaluate the performance of two machine learning tools (NN, RF, and SV) stand volume estimation. The secondary objectives are: (1) to compare the performance of artificial intelligence (AI) tools to the usual regression model and (2) to assess the relative importance of ALS metrics through AI.

2. Material and methods

The study area is located in the State of São Paulo, characterized by a mountainous topography, ranging from 589 to 1294 m above the sea level. The area covers approximately 1138 ha. The forest consists of a commercial plantation, the stock being hybrids of *Eucalyptus grandis* W. Hill ex Maiden and *Eucalyptus urophylla* S.T. Blake managed by the Fibria SA to supply a pulp production company. The trees spacing is 3×2 m, resulting in 1666 plants per hectare, with an age range of 2–8 years old.

The ground data were collected using 112 georeferenced circular plots of 400 square meters. The plot volume was determined from the sum of individual trees, based on their diameter, height and specific volume equation linked to the total area. The statistical summary of stand volume is presented in Table 1.

Table 1
Statistical summary of the stand volume ($m^3 \, ha^{-1}$).

Min.	1st quartile	Median	Mean	3rd quartile	Max.
30.87	196.90	260.10	242.80	308.00	452.80

Table 2
Parameters of ALS campaign used in this work.

Parameter	Unit	Values
Average points density	pts m^{-2}	5
Flight height	m	792
Flight speed	km h^{-1}	148
Pulse rate	kHz	200
Datum	UTM	SIRGAS2000
Year of flight	Year	2012
Overlapping	%	30

The LiDAR data were acquired over the 2012 summer by an RIEGL LMS Q860I sensor combined with an Applanix 510 IMU/GPS installed on a Piper Seneca II aircraft. The mean posting density was 5 points per square meters. The main ALS survey parameters are reported in Table 2.

In order to retrieve canopy height in reference to the ground, the 1-m resolution digital elevation model derived from the ground returns was used to normalize the point cloud. The understory influence was then eliminated by excluding all echoes below 1-m from the calculation of ALS-metrics (Nilsson, 1996).

The ALS metrics for the corresponding field plots were initially extracted from just the first returns (single and first of many) and then later from all returns. A total of 104 ALS-metrics, divided into either height metrics or canopy density metrics, were extracted from each plot. The ALS-metrics considered in this study were widely used in the existing literature, and have been established as effective for stem volume estimation (Hudak et al., 2006; Lim and Treitz, 2004; Næsset, 2004; Næsset and Gobakken, 2008; Parker and Russ, 2004).

The metrics used were: mean of heights (HMEAN), quadratic mean (HSQ), cubic mean (HCUB), mode (HMODE), median (HMEDIAN), median of the absolute deviations from the overall median (HMADIAN), median of the absolute deviations from the overall mode (HMADODE), percentiles of height (HP##), maximum (HMAX), interquartile distance (HIQ), standard deviation (HSD), variance (HVAR), average absolute deviation (HAAD), kurtosis (HKUR), skewness (HSKW), L-moments (HL1, HL2, HL3, HL4, HLSKW, HLKUR), canopy relief ratio (CRR), percentage of all returns above a particular height (ARA2FR, PARA2, ARAMOTFR, PARAM, ARAMO, ARAMTFR, PARAMO, ARAM, TARF), percentage of first returns above a particular height (PFRAMO, PFRA2, PFRAM, FRAM, FRA2, TFRF).

We used three machine learning, all based on supervised learning: neural network (NN), random forest (RF) and support vector regression (SV). The underlying goal of artificial intelligence is to perform acceptably (or even optimally) well at a specific task in a particular environment (White, 1989). The learning algorithm used as the target the stand volume from the field plots to adjust the AI tool while receiving the ALS metrics as inputs. Usually, AI tools are not sensible to collinearity, normality or linearity. RF is not sensitive to the unit dimension as well. Despite RF, NN and SV required scaled data to avoid weights saturation in case of NN and to avoid that attributes in greater numeric ranges dominating those in smaller numeric ranges in case of SV. We scaled the data shifting the variable unity to match 0–1 range. The models were implemented in R software and language (R Core Team, 2013).

Download English Version:

https://daneshyari.com/en/article/84154

Download Persian Version:

https://daneshyari.com/article/84154

Daneshyari.com