Ecological Modelling 366 (2017) 15–26

Contents lists available at ScienceDirect

Ecological Modelling

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

Enhancing of accuracy assessment for forest above-ground biomass estimates obtained from remote sensing via hypothesis testing and overfitting evaluation

R. Valbuena[a,b,*], A. Hernando[c], J.A. Manzanera[c], E.B. Görgens[d], D.R.A. Almeida[e], F. Mauro[c], A. García-Abril[c], D.A. Coomes[a]

[a] University of Cambridge, Department of Plant Sciences, Forest Ecology and Conservation, Downing Street, CB2 3EA Cambridge, UK
[b] University of Eastern Finland, Faculty of Forest Sciences, P.O. Box 111, Joensuu, Finland
[c] Universidad Politecnica de Madrid, College of Forestry and Natural Environment, Research Group SILVANET, Ciudad Universitaria, 28040 Madrid, Spain
[d] Universidade Federal dos Vales do Jequitinhonha e Mucuri, Departament of Forestry, Campus JK, CEP 39100-000, Diamantina, Brazil
[e] University of São Paulo, Luiz de Queiroz College of Agriculture, Department of Forest Sciences, Av. Pádua Dias, 11, CEP 13418-900 Piracicaba, Brazil

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form
29 September 2017
Accepted 17 October 2017

Keywords:
Model assessment
Overfitting
Theil's partial inequality coefficients
LIDAR

ABSTRACT

The evaluation of accuracy is essential for assuring the reliability of ecological models. Usually, the accuracy of above-ground biomass (*AGB*) predictions obtained from remote sensing is assessed by the mean differences (*MD*), the root mean squared differences (*RMSD*), and the coefficient of determination (R^2) between observed and predicted values. In this article we propose a more thorough analysis of accuracy, including a hypothesis test to evaluate the agreement between observed and predicted values, and an assessment of the degree of overfitting to the sample employed for model training. Using the estimation of forest *AGB* from LIDAR and spectral sensors as a case study, we compared alternative prediction and variable selection methods using several statistical measures to evaluate their accuracy. We showed that the hypothesis tests provide an objective method to infer the statistical significance of agreement. We also observed that overfitting can be assessed by comparing the inflation in residual sums of squares experienced when carrying out a cross-validation. Our results suggest that this method may be more effective than analysing the deflation in R^2. We proved that overfitting needs to be specifically addressed since, in light of *MD*, *RMSD* and R^2 alone, predictions may apparently seem reliable even in clearly unrealistic circumstances, for instance when including too many predictor variables. Moreover, Theil's partial inequality coefficients, which are employed to resolve the proportions of the total errors due to the unexplained variance, the slope and the bias, may become useful to detect averaging effects common in remote sensing predictions of *AGB*. We concluded that statistical measures of accuracy, precision and agreement are necessary but insufficient for model evaluation. We therefore advocate for incorporating evaluation measures specifically devoted to testing observed-versus-predicted fit, and to assessing the degree of overfitting.

1. Introduction

The evaluation of accuracy is an essential step indicating the reliability of a given prediction method, thereby informing researchers about the level of confidence they should place in their predictions and allowing them to compare alternatives (Tedeschi, 2006). Accuracy assessment must be supported by rigorous statistical inference, with the ultimate target of evaluating the ability to generalize from the sample data to the population of interest (Särndal et al., 1992; Naesset, 2002; McRoberts et al., 2013; Asner and Mascaro, 2014; Chen et al., 2015; Mauro et al., 2016). Several quantitative techniques can be used to verify if the predicted values differ significantly from the observed, including squared sums of prediction errors (Wallach and Goffinet, 1989), coefficient of determination (R^2) or other correlation-like measures (Willmott, 1981), a reliability index (Leggett and Williams, 1981), distribution hypothesis testing (Freese, 1960), and regression of predicted versus observed (Theil, 1958; Graybill, 1976; Reynolds and Chung, 1986) or *vice versa* (Piñeiro et al., 2008). The advantages and dis-

* Corresponding author at: University of Cambridge, Department of Plant Sciences, Forest Ecology and Conservation, Downing Street, CB2 3EA Cambridge, UK.
E-mail address: rv314@cam.ac.uk (R. Valbuena).

https://doi.org/10.1016/j.ecolmodel.2017.10.009
0304-3800/

advantages of these approaches have been evaluated (e.g., Fox, 1981; Willmott, 1982). Since each scientific application has its own particularities, it is recognised that no single measure of model performance is appropriate in all circumstances (Smith and Rose, 1995). This article explores open questions on accuracy assessment in the context of predicting forest above-ground biomass (*AGB*) from remote sensing sources. The accuracy assessment measures proposed here can nonetheless be generalizable to many other contexts where predictions of ecological variables from different sources of auxiliary information are sought.

1.1. Common measures for accuracy assessment and aspects needing revision

When assessing the performance of their methods, remote sensing researchers usually report: (1) mean difference between observed and predicted values, which evaluates the degree of under- or over-prediction of the dependent variable, *AGB* in this case; (2) the precision of the prediction, often reporting the root mean squared differences (*RMSD*); and (3) the level of agreement between observed and predicted values, commonly considered by indicating their R^2 (e.g., Zhao et al., 2009; Erdody and Moskal, 2010; McInerney et al., 2010; d'Oliveira et al., 2012; Chen and Zhu, 2013; Straub et al., 2013; Asner and Mascaro, 2014; Valbuena et al., 2014). There is, however, no strong consensus, and it is not uncommon to find studies reporting alternative or complementary measures, for instance analysing the regression of predicted versus observed (Bright et al., 2012; Wing et al., 2012) or alternatives to R^2 (Yebra and Chuvieco, 2009; García et al., 2010; Almeida et al., 2016). Some studies (e.g., d'Oliveira et al., 2012; Estornell et al., 2014) perform hypothesis tests comparing distributions, similar to those suggested by Freese (1960). Moreover, the degree of overfitting is rarely accounted for (Valbuena et al., 2013a; Latifi et al., 2015a; Almeida et al., 2016), despite of being a common pitfall in predictive modelling (Weisberg, 1985; Hurvich and Tsai, 1989; Hawkins, 2004). In the context of remote sensing prediction of forest *AGB*, we detected two key aspects of accuracy lacking consensus (plus a third additional one, see Valbuena et al., 2018):

Evaluating regression of observed versus predicted. Piñeiro et al. (2008) argued that the correct assessment is done by setting the predicted values as independent variable (in the x-axis) and the observed values as dependent variable (in the y-axis), to properly evaluate their regression coefficients (Reynolds and Chung, 1986). However, when evaluating remote sensing predictions of forest attributes, many authors have presented predicted (in the y-axis) vs. observed (in the x-axis) instead (e.g., McRoberts et al., 2002; Holmgren et al., 2008; Zhao et al., 2009; McInerney et al., 2010; Chen and Zhu, 2013; Valbuena et al., 2014; Latifi et al., 2015b). Furthermore, they usually lack reporting the regression of observed against predicted (e.g., Naesset, 2002; García et al., 2010; Straub et al., 2013). Although some report the coefficients (e.g., Yebra and Chuvieco, 2009; Bright et al., 2012; Wing et al., 2012), they may still miss the hypothesis test suggested by Piñeiro et al. (2008). There have therefore not been reports on the importance of carrying out these hypothesis tests in the context of remote sensing predictions of *AGB*. Complementary statistics may also be included in order to fully comprehend the source of prediction errors, such as Theil (1958) partial inequality coefficients (Smith and Rose, 1995). They disaggregate the total error into model variance (unsystematic error), bias (systematic error), and slope (averaging effects) (Paruelo et al., 1998). To our knowledge, these coefficients have not been employed in the context of remote sensing estimates of forest characteristics before.

The degree of overfitting to the sample. Franco-Lopez et al. (2001) argued that statistical measures to assess model overfitting should be included when reporting the accuracy assessment of remote sensing estimates. Those measures of overfitting have been, however, largely overlooked in remote sensing estimations of forest attributes (Latifi et al., 2015a). Overfitting is usually prevented beforehand by avoiding over-parameterization with variable selection methods (e.g., Naesset, 2002; Hudak et al., 2006; García et al., 2010; Wing et al., 2012; Spriggs et al., 2015). These methods, however, have been suspected of being insufficient to truly avoid model overfitting (Allen, 1974; Vanclay and Skovsgaard, 1997; Hurvich and Tsai, 1989; Rencher and Pun, 1993). As an alternative, some authors recommend preventing model overfitting using replication methods such as cross-validation, and compare their results against model residuals (Weisberg, 1985; Hawkins, 2004). These would also be particularly convenient for non-parametric machine learning methods, whose flexibility makes them especially prone to overfitting (Franco-Lopez et al., 2001; Hawkins, 2004), and which are of widespread use in remote sensing predictions of forest attributes (McRoberts et al., 2002; Hudak et al., 2008; Packalén and Maltamo, 2008; McInerney et al., 2010). However, overfitting is rarely addressed in the context of remote sensing predictions of forest variables (Franco-Lopez et al., 2001; Valbuena et al., 2013a; Latifi et al., 2015a; Almeida et al., 2016).

These alternative methods for testing the reliability of *AGB* predictions obtained by using remotely sensed sources may also be employed to minimise errors in the estimation of ecological variables in general. Results may therefore be relevant to other contexts too, for example studies on ecosystem management responses to climate change or habitat suitability for fauna, where the use of models to predict ecological attributes from auxiliary variables is common.

1.2. Objectives

The objective of this research is to call into question the sufficiency of statistical measures commonly used for accuracy assessment of predictions of ecological variables from auxiliary information, and suggest the convenience of incorporating additional ones, with a focus on remote sensing estimations of forest *AGB*. Our hypothesis is that the statistics usually reported in *AGB* assesments may be insufficient for accepting the degree of agreement between predicted and observed as reliable, and also that the fact that overfitting effects may remain undetected. This article therefore aspires to present a thorough analysis of accuracy that applies to ecological modelling in general, and to explain how to interpret the suggested statistical metrics for readers unfamiliar to them in the given context.

2. Material and methods

2.1. Field and remote sensing datasets

The field datasets consisted of $n = 37$ plots surveyed during summer 2006 in the Scots pine (*Pinus sylvestris* L.) dominated forests of Valsaín (Spain, approx. lat.: 41°04′ N, lon.: 4°09′ W; 1.3–1.5 km a.s.l.). These plots consisted of two concentric circles of radii 10 and 20 m. Diameters at breast height (*dbh*, cm) were measured for every tree located within the inner sub-plot, whereas at the outer subplot only those with *dbh* > 10 cm were measured (Valbuena et al., 2013b). Differentially-corrected global navigation satellite systems (GNSS) were used to obtain the positions of these plots with centimetre accuracy (Valbuena et al., 2012), enabling to link the field and remote sensing information.

Locally-adjusted tree allometry specific for *P. sylvestris* was employed to obtain the above-ground biomass (*agb*, kg) of each individual tree from the field measurements (Montero et al., 2005):

$$agb = 0.08439 \cdot dbh^{2.41194} \quad (1)$$