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

Authors: Valbuena R[1,2*], Hernando A[3], Manzanera JA[3], Görgens EB[4], Almeida DRA[5], Mauro F[3], García-Abril A[3], Coomes DA[1].

1: University of Cambridge, Department of Plant Sciences, Forest Ecology and Conservation. Downing Street, CB2 3EA Cambridge, UK

2: University of Eastern Finland, Faculty of Forest Sciences, P.O. Box 111, Joensuu, Finland

3: Universidad Politecnica de Madrid, College of Forestry and Natural Environment, Research Group SILVANET, Ciudad Universitaria, 28040 Madrid, Spain.

4: Universidade Federal dos Vales do Jequitinhonha e Mucuri, Departament of Forestry, Campus JK, CEP 39100-000, Diamantina, Brazil.

5: University of São Paulo, Luiz de Queiroz College of Agriculture, Department of Forest Sciences, Av. Pádua Dias, 11. CEP 13418-900 Piracicaba, Brazil.

* Corresponding author: rv314@cam.ac.uk

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.

Key words: model assessment; overfitting; Theil's partial inequality coefficients; LIDAR.

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 & 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 & Williams, 1981), distribution hypothesis testing (Freese, 1960), and regression of predicted versus observed (Theil, 1958; Graybil, 1976; Reynolds & Chung, 1986) or *vice versa* (Piñeiro et al 2008). The advantages and disadvantages 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 & 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.

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 & Moskal 2010; McInerney et al., 2010; d'Oliveira et al 2012; Chen & Zhu, 2013; Straub et al., 2013; Asner & 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 & 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 & 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 & 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 & Zhu 2013; Valbuena et al., 2014). 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 & 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's (1958) partial inequality coefficients (Smith & 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 & Skovsgaard, 1997; Hurvich & Tsai, 1989; Rencher & 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 & 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.

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.

Material and Methods

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 sub-plot 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):

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

These tree-level agb estimations were aggregated to plot-level totals (AGB, Mg·ha^{-1}), after referring each of them to per-hectare equivalents according to the differing size of the sub-plot from which each tree was sampled (inner or outer). In this study, we used AGB as a response variable to be predicted throughout the target forest by using the remote sensing predictor variables.

The predictor variables were statistical metrics describing the distributions of signals received at those same field plots from both active LIDAR and passive multispectral sensors. This remotely sensed information was acquired on September 10, 2006, from a laser scanner ALS50-II (Leica Geosystems, Switzerland) and a digital mapping camera system (Zeiss-Intergraph, Germany). Simultaneously operating onboard a plane flying at a height of 1500 m, the LIDAR dataset was obtained with an average scan density of 1.15 pulses·m^{-2}, whereas images had spatial resolutions of 15 cm from panchromatic and 60 cm for multispectral. A back-projection data fusion algorithm using information from on-flight GNSS and inertial navigation systems assured a nearly perfect fit of all the sensor and field information (Valbuena, 2014). Back-projecting consists in mathematically rendering the position of each LIDAR return onto the camera at the time of exposure, retrieving back its radiometric information and effectively colouring the LIDAR return with an accuracy close to pixel size (Valbuena et al., 2011). Returns obtained from the LIDAR sensor, considered to represent the ground – by means of Axelsson's

(2000) classification algorithms –, were interpolated into a digital terrain model, which was used as a reference from which to calculate the heights above ground (h, m) for every single LIDAR return. The radiometric information acquired from the digital camera system was employed to calculate a value of normalised difference vegetation index (*NDVI*; Rouse et al., 1974) correspondent to each LIDAR first return. Using FUSION software (USDA Forest Service; McGaughey, 2012), the returns backscattered from each field plot were extracted, and several metrics describing the distributions and relative proportions of h and *NDVI* with each plot were computed (Manzanera et al., 2016). All these metrics were employed as initial dataset of predictors in all the predictive procedures.

Modelling biomass from airborne remote sensing data

Three prediction methods commonly employed for forest AGB predictions from remote sensing were compared within the R statistical environment (version 3.3.1; R Development Core Team, 2016):

Non-parametric modelling based on the most similar neighbour (MSN) method to obtain AGB predictions (Moeur & Stage, 1995) was applied using the "yaImpute" package of R (version 1.0-18; Crookston & Finley, 2007). MSN belongs to a type of non-parametric imputation approaches known as nearest neighbour methods and commonly abbreviated as k-NN (Franco-Lopez et al., 2001; McRoberts et al., 2002; McInerney et al., 2010), k being the number of neighbours used in the algorithm. In the particular case of MSN, the feature space – where distances to neighbours are measured – is modified according to canonical correlation projectors (Hudak et al., 2008; Packalén & Maltamo, 2008). The nearest neighbour algorithm was set for $k = 3$ and averaging by inverse distance weighting, also including a prior variable selection based on variance-weighted canonical correlation analysis (CCA). The value of k was kept low due to the small n available since, although a higher k may improve the precision of the estimation, it can also have an averaging effect (i.e., bias extreme values toward the average) (Eskelson et al., 2009; Almeida et al., 2016). The selection was done by recursively restricting the number of predictors (p) from $p = 30$ to $p = 1$, on the grounds of the absolute values of their coefficients in the canonical regression (Cohen et al., 2003; Manzanera et al., 2016). The highest p was intentionally left unrealistically large, given the subsequent low n/p ratio, to test the results that accuracy measures could provide in such an extreme case. An optimal p was selected according to a combination of accuracy measures, which restricted the p on the basis of a hypothesis test (Piñeiro et al., 2008) and avoiding model overfitting (Weisberg, 1985; Hawkins, 2004), as explained below. This same approach for restricting p (see "restricted" alternatives below) was also incorporated to optimize the *best-subset* and *step-wise* variable selection procedures typically used in parametric modelling for remote sensing prediction of AGB.

Parametric modelling based on variable selection via step-wise regression (Weisberg, 1985). A linear model was fitted using a natural logarithm transform of the response variable, as it is typically done in remote sensing predictions of forest attributes (e.g., Naesset, 2002; Hudak, 2005; Asner & Mascaro, 2014). Baskerville's (1972) correction for bias in log-transformed responses was applied taking into account the number of fitted parameters when calculating the standard error of the estimate (Sprugel, 1983). Function "stepAIC" of R was used for applying a backward selection of independent variables in linear regression models (Venables & Ripley, 2002). The final p was limited on the basis of the delta parameter (Δ; Burnham & Anderson, 2002), which measured the relative increase in Sugiura's (1978) corrected AIC (Akaike Information Criterion) at each step (Valbuena et al., 2013b) (hereafter denominated "step-wise"). The result was compared to an alternative incorporating the above-mentioned restrictions – hypothesis test plus avoided overfitting –, which modified the p derived from the step-wise procedure (hereafter denominated "step-wise restricted overfitting").

Parametric modelling based on variable selection via best-subset regression (Miller, 1984; Hudak et al., 2006). This approach also consisted of a linear model with log-transformed responses and bias correction (Baskerville, 1972; Sprugel, 1983). In this other case, package “leaps” of R (Lumley & Miller, 2009) was employed for this modelling approach. This approach exhaustively searches for all variable combinations. The limiting criterion for p was set to be based on minimization of Mallows’ Cp (Mallows, 1973) (hereafter denominated “best-subset”). Its result was also compared to a version incorporating the novel restrictions – hypothesis test plus avoided overfitting – to the best subset procedure for predictor variable selection (hereafter denominated “best-subset restricted overfitting”).

Statistical measures for accuracy assessment of AGB predictions

Leave-one-out cross-validation was carried out to assess all the prediction methods considered. Thus, after removing one case (i) from the total n, the remaining were used to calculate a new AGB prediction of the response for that given case (pre_i^{cv}). Hereafter, the superscript/subscript cv is used to distinguish measures calculated after the cross-validation procedure, as opposed to the superscript/subscript fit which will denote non-cross-validated measures, for instance the predictions that yield model residuals (pre_i^{fit}). The result was evaluated with observed versus leave-one-out predicted plots, from which we evaluated:

(1) The *mean difference* (MD) between the predicted minus the observed values:

$$MD = \sum_{i=1}^{n}(pre_i^{cv} - obs_i)/n\ , \tag{2}$$

which evaluates the degree of under- or over-prediction of the method employed. Eq. (2) is equivalent to the difference between the means of the observed and predicted (e.g., McInerney et al., 2010; Wing et al. 2012). MD was expressed in AGB units, whereas relative mean difference ($MD\%$) was calculated by dividing MD by the observed mean AGB ($\overline{obs}$).

(2) The *precision* of the prediction, considered as the mean of absolute differences (MAD):

$$MAD = \sum_{i=1}^{n} |pre_i^{cv} - obs_i|/n, \tag{3}$$

and also the root mean squared differences ($RMSD$) of predicted values with respect to the observed ones:

$$RMSD = \sqrt{SS^{cv}/n}, \tag{4}$$

where SS^{cv} was the sum of the squared differences between the observed values and the predicted values obtained by cross-validation (a.k.a. predicted sum of squares $PRESS$; Allen, 1974; Geisser & Eddy, 1979; Weisberg, 1985: 217; e.g., Valbuena et al., 2013a):

$$SS^{cv} = \sum_{i=1}^{n} (pre_i^{cv} - obs_i)^2. \tag{5}$$

Both MAD and $RMSD$ represent the error in AGB units, the latter being more prone to the presence of outliers (e.g., García et al., 2010). Their respective relative counterparts, MAD% and $RMSD$% (a.k.a the coefficient of variation of $RMSD$; e.g. Valbuena et al., 2014), were also calculated by dividing them by $\overline{obs}$.

(3) A *hypothesis test* testing whether observed and predicted values follow the 1:1 correspondence line (Graybill, 1976; Leite & Oliveira, 2002), was assessed from the intercept (α) and slope (β) of the linear regression model between the observed and predicted (Piñeiro et al., 2008):

$$obs_i = \alpha + \beta pre_i^{cv}, \tag{6}$$

which is proven by not rejecting the null hypotheses that H_0: $\alpha = 0$ and H_0: $\beta = 1$ for $pre_i^{cv} - obs_i = \alpha + \beta pre_i^{cv}$ (Eq. 9 in Piñeiro et al., 2008). Hence, this is a means for assessing the residual distribution analytically, instead of evaluating it visually from a residuals versus predicted scatterplot (e.g., Mauro et al., 2016: Fig. 2).

(4) The *proportions of the total errors* which are due to the unexplained variance (U_{error}), the slope (U_{slope}), and the bias (U_{bias}), which were evaluated from Theil's (1958) partial inequality coefficients (Paruelo et al., 1998):

$$U_{error} = \sum_{i=1}^{n}(est_i^{cv} - obs_i)^2 / SS^{cv} , \tag{7}$$

where $est_i^{cv} = \hat{\alpha} + \hat{\beta} \cdot pre_i^{cv}$ were the values estimated by the regression model (Eq. 6);

$$U_{slope} = [(\beta - 1)^2 \sum_{i=1}^{n}(pre_i^{cv} - \overline{pre^{cv}})^2] / SS^{cv} ; \tag{8}$$

and

$$U_{bias} = [n \cdot MD^2] / SS^{cv}. \tag{9}$$

We multiplied the values of Theil's (1958) partial inequality coefficients by 100, to make it straightforward to the reader that they express the percentage of the total error which is due to either an overall bias of the model (U_{bias}), the presence of trends in the residuals (U_{slope}) or just the residual variance of the model (U_{error}).

(5) The *degree of overfitting* to the sample, which we assessed using a replication method comparing cross-validation results against model residuals (Allen, 1974; Snee, 1977; Vanclay & Skovsgaard, 1997; Geisser & Eddy, 1979; Hawkins, 2004). Most studies assume that overfitting is avoided if over-paramaterization of the prediction model is prevented by using condition number (κ; Weisberg, 1985; e.g., Naesset, 2002), variance inflaction factor (VIF; Fox & Monette, 1992; e.g., García et al., 2010), Mallows' (1973) Cp statistic (e.g., Hudak et al., 2006), or information criterion indices: Akaike (1992) (AIC; e.g., Bright et al., 2012), Bayesian (BIC; e.g., Wing et al., 2012) or deviance (DIC; e.g., Spriggs et al, 2015). Many authors deem these insufficient, however, advocating for methods dealing with overfitting directly (Allen, 1974; Snee, 1977; Hurvich & Tsai, 1989; Rencher & Pun, 1993).

Moreover, Hawkins (2004) argued in favour of using replication methods for non-parametric machine learning approaches like MSN, which may lack the theoretical basis on which κ, VIP, Cp or AIC are grounded. For this reason, we alternatively assessed overfitting directly from the sums of squares ratio (SSR) and R^2 ratio ($R2R$) (Ehrenberg, 1982; Weisberg, 1985: 68-69, 217; Lipovetsky, 2013), both obtained by comparison of a same measure acquired by model fit against cross-validation.

The ratio between the square root of the sums of squares attained in the cross-validation (SS^{cv}) (Eq. 5) and that using the whole dataset (SS^{fit}) (Snee, 1977; e.g., Valbuena et al., 2013a) yielded the SSR:

$$SSR = \sqrt{SS^{cv}}/\sqrt{SS^{fit}}, \tag{10}$$

where SS^{fit} was the sum of squares of the model residuals (j), i.e. the values fitted without cross-validation (Hawkins, 2004):

$$SS^{fit} = \sum_{j=1}^{n}(pre_j^{fit} - obs_j)^2. \tag{11}$$

On the other hand, a similar measure was obtained using the R^2 of the regression of observed versus predicted values (Piñeiro et al., 2008). This was the ratio between the one obtained by cross-validation and that from model residuals: the R^2 ratio ($R2R$). Equation (5) derives:

$$R_{cv}^2 = 1 - SS^{cv}/SS_{tot}, \tag{12}$$

where SS_{tot} was the sum of squared differences of each observation from the overall mean:

$$SS_{tot} = \sum_{i=1}^{n}(obs_i - \overline{obs})^2. \tag{13}$$

Whereas from model residuals the coefficient of determination obtained is derived from Eq. (11) instead:

$$R_{fit}^2 = 1 - SS^{fit}/SS_{tot}. \quad (14)$$

Then, the deflation observed by the cross-validation in the coefficient of determination can be then assessed as (Rencher & Pun, 1993; e.g., Latifi et al., 2015a):

$$R2R = R_{fit}^2/R_{cv}^2 = \left(1 - \frac{SS^{fit}}{SS_{tot}}\right) \bigg/ \left(1 - \frac{SS^{cv}}{SS_{tot}}\right), \quad (15)$$

Comparing these two functions, Eqs. (10) and (15), it can be seen that SSR and $R2R$ do, in essence, very similar tasks. While $R2R$ is a ratio of decrease in explained variance experienced when cross-validating, SSR is a ratio of increase in unexplained variance (square-rooted, in this case). These two measures can therefore be employed to adjust the inflation of the unexplained variance (SSR) or deflation of explained variance ($R2R$) in the cross-validation to a desirable limit, for example 5% or 10% (Lipovetsky, 2013) (i.e., SSR or $R2R$ would be lower than e.g. 1.05 or 1.10 – numerator and denominator in Eq. (15) have been swapped compared to Eq. (10), so that both SSR and $R2R$ rise for increasing overfitting). It may be worthwhile to mention that although in the univariate case the cross-validation necessarily leads to an increase in the sums of squares and a decrease in the R^2 (Ehrenberg, 1982; Weisberg, 1985), Lipovetsky (2013) showed that this property does not necessarily always hold in the multivariate case.

Comparing alternatives

The relative merits of each of the proposed statistical measures – MD, $MD\%$, MAD, $MAD\%$, $RMSD$, $RMSD\%$, α, β, U_{error}, U_{slope}, U_{bias}, SSR and $R2R$ – were evaluated by analysing the results provided when applying different alternative prediction methods to the same dataset, and also by comparing their corresponding scatterplots of observed versus predicted values. Firstly, we compared results obtained while increasing the number of predictors in MSN. We purposely included unrealistically low n/p ratios, with the intention to realize which statistical measures would flag up their unreliability. Additionally, we observed the correlations between pairs of statistical measures to prove whether they are simply redundant or provide additional information, using Spearman's rank correlation coefficient (ρ) because it could prove that two methods would rank alternatives in a similar manner. Secondly, we compared automatic variable selection procedures commonly employed in the assessment of remote sensing assisted AGB estimations: step-wise and best subset. The additional statistical measures were incorporated into these algorithms, showing that improvements in overfitting and avoiding systematic errors may be achieved without excessively compromising the overall precision of the estimates.

Results

Estimation with different number of predictors

Let us first analyse the results observed when modifying the number of predictors p during the variable selection procedure for MSN imputation. **Figure 1** shows the evolution of the statistical measures for increasing p, grouped by the characteristics they describe: mean difference and precision of predictions, their 1:1 correspondence with the observed values, and the degree of overfitting. **Table 1** summarizes the numerical results attained for a relevant selection of these: $p = 2,3,5,8,10,15,20$ and 30. Their corresponding observed versus predicted plots are shown in **Fig. 2**. Results obtained from the hypothesis tests applied to the fit of observed versus predicted rejected the reliability of accepting the options using either $p =$ 1-4, $8, 12, 29$ or 30 (denoted with asterisks in **Figs. 1c**), whereas every other option passed the test successfully.

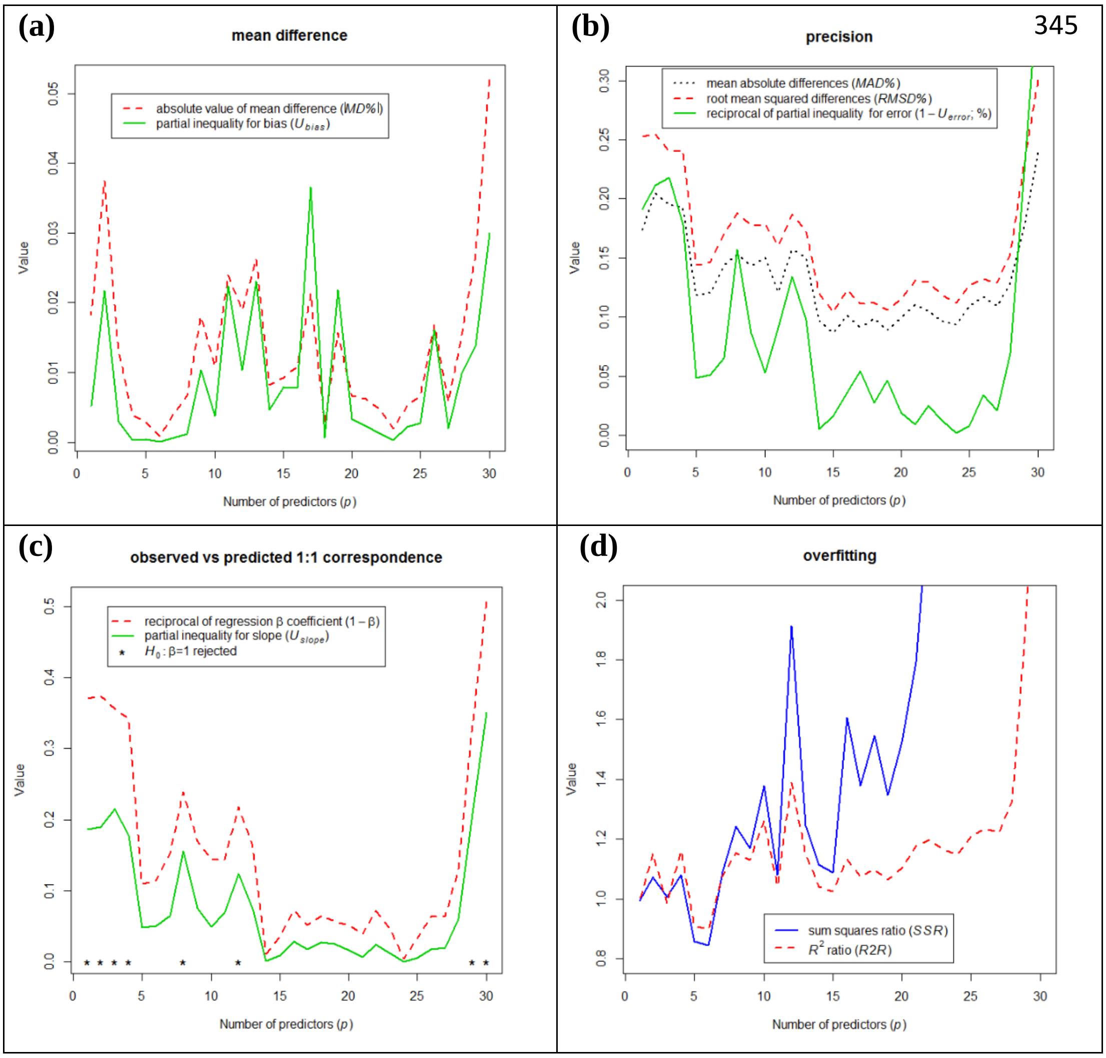
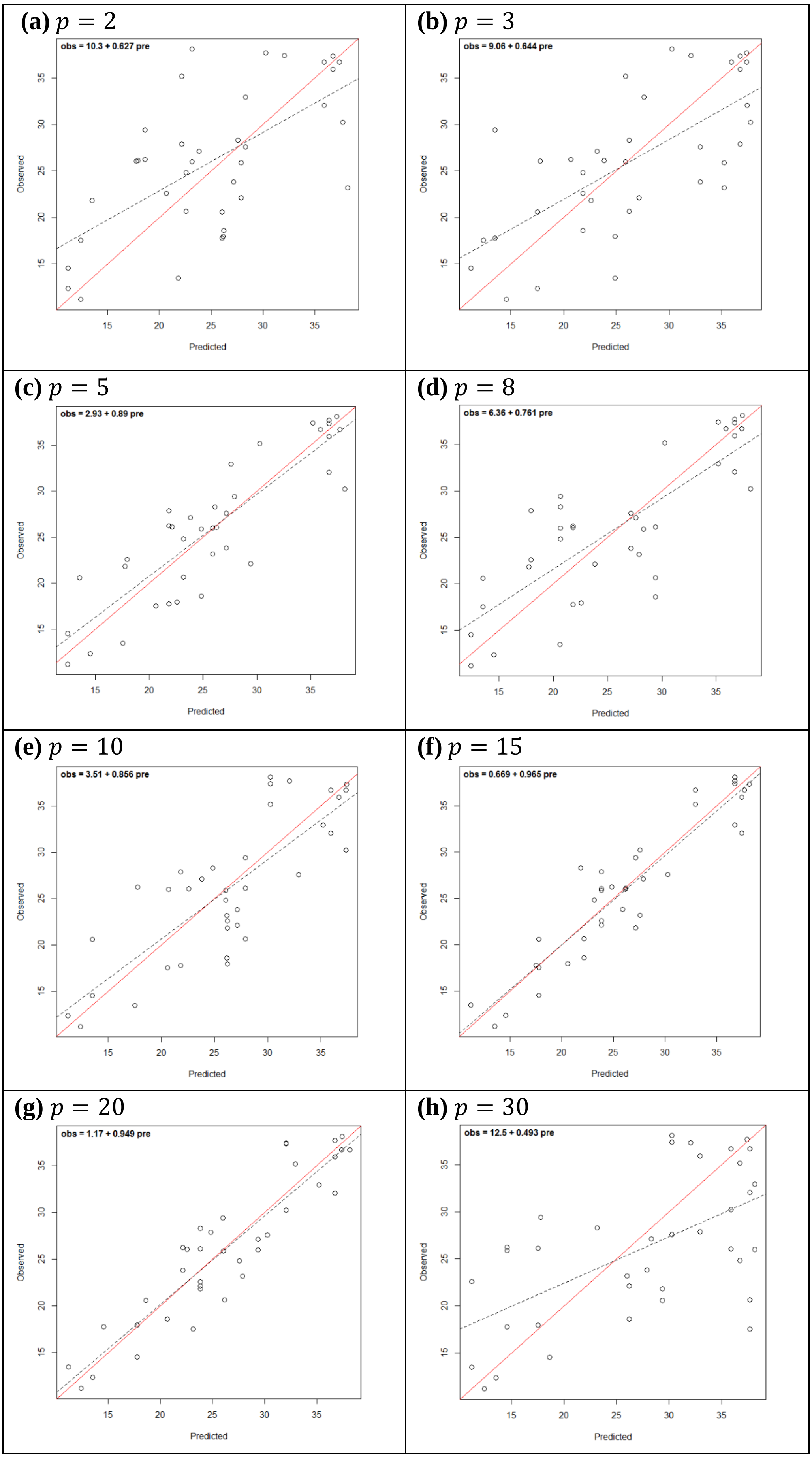


Figure 1. Statistical evaluation of MSN predictive method for increasing the number of predictors (p), grouped according to whether they define (a) the mean difference or (b) precision of predictions, (c) their 1:1 correspondence or (d) the degree of overfitting.

350 **Table 1.** Summary diagnosis of most similar neighbour (MSN) predictions for above-ground
351 biomass (AGB, Mg·ha^{-1}) using an increasing number of predictors (p).

		Number of predictors (p)							
		2	3	5	8	10	15	20	30
Prediction bias	MD	-.98	.34	-.08	-.17	.28	.24	.17	1.35
	$MD\%$	-3.75	1.31	-.29	-.66	1.09	.93	.66	5.20
Prediction precision	MAD	5.33	5.08	3.09	4.00	3.93	2.26	2.60	6.23
	$MAD\%$	20.5	19.5	11.8	15.5	15.1	8.7	9.60	23.9
	$RMSD$	6.63	6.26	3.75	4.90	4.65	2.72	3.01	7.68
	$RMSD\%$	25.4	24.0	14.4	18.8	17.9	10.4	11.5	30.0
Hypothesis test	α	10.3**	9.06**	2.93^{NS}	6.36*	3.51^{NS}	.66^{NS}	1.17^{NS}	12.5***
	β	.63**	.64***	.89^{NS}	.76**	.86^{NS}	.97^{NS}	.95^{NS}	.49***
Partial inequality coefficients	U_{error} (%)	78.8	78.2	95.1	84.3	94.7	98.3	98.0	61.9
	U_{slope} (%)	18.9	21.5	4.84	15.6	4.9	.88	1.59	35.1
	U_{bias} (%)	2.16	.38	.01	.01	.02	.79	.40	3.00
Agreement	R^2_{cv} (%)	38.7	45.8	76.3	64.2	63.7	87.1	84.3	33.1
Overfitting	SSR	1.07	1.01	.89	1.24	1.38	1.09	1.52	7.69
	$R2R$	1.15	.99	.91	1.15	1.26	1.02	1.11	2.97

MD: mean differences (Eq. 2). $MD\%$: relative MD. MAD: mean absolute differences (Eq. 3). $MAD\%$: relative MAD. $RMSD$: root mean squared differences (Eq. 4). $RMSD\%$: relative $RMSD$. α/β : intercept/slope of observed versus predicted regression (Eq. 6) (levels of significance for rejecting H_0: *:.05; **:.01; ***:.001; ^{NS}: non-significant). $U_{error}/U_{slope}/U_{bias}$: Theil's (1958) partial inequality coefficients for error variance/slope/bias (Eqs. 7-9). R^2_{cv}: cross-validated coefficient of determination (Eq. 12). SSR: sum of squares ratio (Eq. 10). $R2R$: R^2 ratio (Eq. 15). Relative figures and agreement/inequality coefficients have been multiplied by hundred to yield percentage units.



(a) $p = 2$
obs = 10.3 + 0.627 pre
(b) $p = 3$
obs = 9.06 + 0.644 pre
(c) $p = 5$
obs = 2.93 + 0.89 pre
(d) $p = 8$
obs = 6.36 + 0.761 pre
(e) $p = 10$
obs = 3.51 + 0.856 pre
(f) $p = 15$
obs = 0.669 + 0.965 pre
(g) $p = 20$
obs = 1.17 + 0.949 pre
(h) $p = 30$
obs = 12.5 + 0.493 pre
Observed
Predicted
15
20
25
30
35

Figure 2. Observed versus predicted plots of most similar neighbour (MSN) imputation models for above-ground biomass (AGB, Mg·ha^{-1}) using an increasing number of predictors (p). The solid red line represents the 1:1 correspondence. Dashed line is the linear regression fit between observed and predicted $obs_i = \alpha + \beta \cdot pre_i$.

Mean differences, i.e. over- or under-prediction, were negligible in most cases (**Table 1**; **Fig. 1a**), usually below $|MD| = 2\%$ (which in practice implies an approximate deviation of 0.5 Mg·ha^{-1}). Therefore, in almost every case the prediction methods would yield an unbiased estimation of the mean AGB for the population. The absolute value of MD has been depicted in **Fig. 1a** in order to express its magnitude regardless of whether it implies under- or over-prediction. The results obtained by $|MD|$ were also corroborated by the low proportions of error due to bias, as shown by its corresponding Theil's partial inequality coefficient (U_{bias}). These two measures were very highly correlated $\rho_{(|MD|,U_{bias})} = 0.94$, and hence reiterative. The largest over-predictions resulted from the MSN model with $p = 30$, which showed a $MD = 5.20\%$ with a proportion of the total error due to bias reaching $U_{bias} = 3.00\%$. Any other alternative $p =$ 1-29 could have been deemed as providing a reliable AGB prediction. However, scatterplots in **Fig. 2a-b** show examples of some cases were the unreliability of predictions could also have been perceived visually. Alternatively to visual assessment, lack of reliability may also be automatically detected via significance of hypothesis tests (denoted by asterisks in **Fig. 1c**).

With regards to the precision of predictions, results were also reasonably acceptable, ranging $RMSD =$ 10.4-18.8% for $p =$ 5-28. Higher ($p =$ 29-30) or lower ($p =$ 1-4) number of predictors reached larger $RMSD =$ 24.0-30.0% (**Fig. 1b**). $RMSD$ and MAD changed very similarly for different p, MAD being systematically lower than $RMSD$, as it could be expected from Eqs. 3-5. As a result, MSN imputations using $p = 1$-4 seemed apparently better when evaluated by their $MAD = 17.4$-20.5%, as compared to observing their higher $RMSD =$ 24.0-25.4%. In fact $\rho_{(MAD,RMSD)} = 0.99$, and hence there is no need to report both measures. Moreover, Theil's partial inequality for error (U_{error}) and the slope of the regression β also showed similar patterns as $RMSD$, being $\rho_{(RMSD,U_{error})} = -0.92$ and $\rho_{(RMSD,\beta)} = -0.85$. Significances in the test of lack of fit to the 1:1 correspondence were therefore closely associated to low precisions in the AGB prediction (**Fig. 2**). The use of β, however, provided the added value of incorporating a significance test that can be used as an objective threshold for rejecting excessively low precision in prediction error (denoted with asterisks in **Fig. 1c**).

For assessing the degree of overfitting to the sample, the suggested statistical measures – SSR and $R2R$ – yielded diverging results for high values of p (**Fig. 1d**). Results in **Table 1** and **Fig. 1d** revealed that, for many of the alternatives, the 'real' (cross-validated) precision exceeded 10% of model residual variance (denoted by values of SSR or $R2R < 1.1$). Among all the alternatives considered, only those MSN imputations using $p = 1$-$7, 11$ and 15 obtained values of $SSR < 1.1$. Being 10% a fairly acceptable level of divergence, if such criterion is set in conjunction with the hypothesis tests for rejecting a given AGB estimation, then only the MSN predictions using $p = 5$-$7, 11$ and 15 would be acceptable options. On the other hand, $R2R$ was generally less sensitive to overfitting than SSR (**Table 1**). **Fig. 1d** shows that $R2R$ was critically low at elevated values of p, which is in disagreement with what would have intuitively be assumed by the subsequent low n/p ratios, whereas SSR unveiled a dramatical increase in the overfit for most alternatives above $p = 7$. In fact, SSR correlated to the p itself – $\rho_{(p,SSR)} = 0.93$ –, while $R2R$ has a weaker relationship to the number of predictors used – $\rho_{(p,R2R)} = 0.62$ –, which indicates that comparing the deflation in R^2 may be not useful to avoid over-parameterization.

In light of our results, Theil's partial inequality coefficients can be useful for a detailed evaluation of results. U_{bias} may detect systematic differences between observed and predicted values. Additionally, large values for U_{slope}, such as those obtained for $p = 3$ or $p = 8$-10, indicated a tendency for predicting towards the average AGB (**Fig. 2**) (i.e., over-predicting low AGB areas and under-predicting large ones). Hence, even if the overall population mean may be assumed unbiased in light of MD or U_{bias}, there is still a chance for the values shown at the scale of the estimation units (the pixels in the remote sensing case) to be selectively under- or over-predicted for certain values within the range of observed AGB. Our results showed that this was indeed the case, since large values of $U_{slope} = 10.4$-11.3% were associated to significant test results for either the α or β coefficient, or both (**Table 1**). On the other hand, the alternatives for which the null hypotheses were not rejected by the tests (signified by non-significances for the coefficients) obtained much lower values, such as $U_{slope} = 7.09\%$ for $p = 5$ and $U_{slope} = 0.51$-4.22% for $p = 15$-25. For instance, Theil's partial inequality coefficients were particularly relevant for $p = 3$, (**Fig. 2b**), since its $U_{slope} = 21.5\%$ revealed and averaging effect which remained concealed by its low $MD = 1.31\%$ (**Table 1**).

Comparison of alternative modelling methods

We also wanted to use the proposed measurements of accuracy to compare the results obtained by the MSN imputation with two other modelling alternatives commonly employed in remote sensing-assisted predictions of AGB: best-subset and step-wise regression (**Table 2**). Based on the results detailed on the previous sub-section, we decided to incorporate two additional constraints on variable selection (called 'restricted' in **Table 2**) on top of their original limitation criteria (i.e., Cp for best-subset and Δ for step-wise). These were the hypothesis tests and the degree of overfitting, i.e. a model would be declined if either of the null hypotheses H_0: $\alpha = 0$ or H_0: $\beta = 1$ were rejected, or $SSR > 1.1$. **Table 2** compares all these versions against the previously-selected MSN imputation model for $p = 5$, which was selected as the optimal MSN predictions under the same criteria. **Figure 3** shows the observed versus predicted plots corresponding to each of these alternatives.

In a similar manner to the previous comparison of MSN imputation, all the alternatives resulted in unbiased predictions of population mean ($|MD| =$ 0.10-1.32% and $U_{bias} =$ 0.01-0.79%) performing a reasonable error variance ($RMSD =$ 9.67-15.3%) and good agreement between observed and predicted (see Valbuena et al., 2018). In this case they all passed the hypothesis tests, as most of the overall errors were simply due to unsystematic sources affecting the error variance of the model itself ($U_{errors} =$ 94.7-98.9%). None of the models therefore had to be declined due to failing the hypothesis test on the correspondence between observed and predicted. However, we detected an overfitting effect at both the best-subset model selected on the grounds of Mallow's Cp ($SSR = 1.28$) and also at the step-wise regression selected via Δ's difference in Sugiura's corrected AIC ($SSR = 2.90$). It is noteworthy to point out that this contingency could have simply remained overlooked if overfitting had been analysed according to the deflation in R^2, which was only $R2R = 1.04$ for best subset and $R2R = 1.17$ for the step-wise regression model. Accordingly, we imposed the criterion of $SSR \leq 1.1$ to further constrain the prediction dataset of these models. This resulted in unbiased models including just $p = 2$ independent variables, which avoided overfitting ($SSR = 1.08$) while not excessively compromising model precision ($RMSD = 15.3\%$ and $RMSD = 14.3\%$, respectivelly).

Table 2. Comparison of diagnoses for different prediction method and variable selection alternatives to obtain above-ground biomass (AGB, Mg·ha^{-1}) predictions.

		Best-subset	Best-subset restricted	Step-wise	Step-wise restricted	MSN restricted
Number of predictors (p)		8	2	23	2	5
Prediction bias	MD	-.06	-.02	.34	-.16	-.08
	$MD\%$	-.24	-.10	1.32	-.63	-.29
Prediction precision	MAD	2.09	3.32	2.26	2.87	3.09
	$MAD\%$	8.01	12.7	8.69	11.0	11.8
	$RMSD$	2.52	3.99	3.07	3.73	3.75
	$RMSD\%$	9.67	15.3	11.8	14.3	14.4
Hypothesis test	α	.96^{NS}	3.21^{NS}	.13^{NS}	1.71^{NS}	2.93^{NS}
	β	.97^{NS}	.87^{NS}	.98^{NS}	.94^{NS}	.89^{NS}
Partial inequality coefficients	U_{error} (%)	98.9	94.7	98.6	98.5	95.1
	U_{slope} (%)	1.04	5.30	.17	1.30	4.84
	U_{bias} (%)	.06	.01	.01	.19	.01
Agreement	R^2_{cv} (%)	88.9	73.4	83.6	75.7	76.3
Overfitting	SSR	1.28	1.08	2.90	1.08	.89
	$R2R$	1.04	1.05	1.17	1.04	.91

MD: mean differences (Eq. 2). $MD\%$: relative MD. MAD: mean absolute differences (Eq. 3). $MAD\%$: relative MAD. $RMSD$: root mean squared differences (Eq. 4). $RMSD\%$: relative $RMSD$. α/β : intercept/slope of observed versus predicted regression (Eq. 6) (levels of significance for rejecting H_0: *:.05; **:.01; ***:.001; ^{NS}: non-significant). $U_{error}/U_{slope}/U_{bias}$: Theil's (1958) partial inequality coefficients for error variance/slope/bias (Eqs. 7-9). R^2_{cv}: cross-validated coefficient of determination (Eq. 12). SSR: sum of squares ratio (Eq. 10). $R2R$: R^2 ratio (Eq. 15). Relative figures and agreement/inequality coefficients have been multiplied by hundred to yield percentage units.

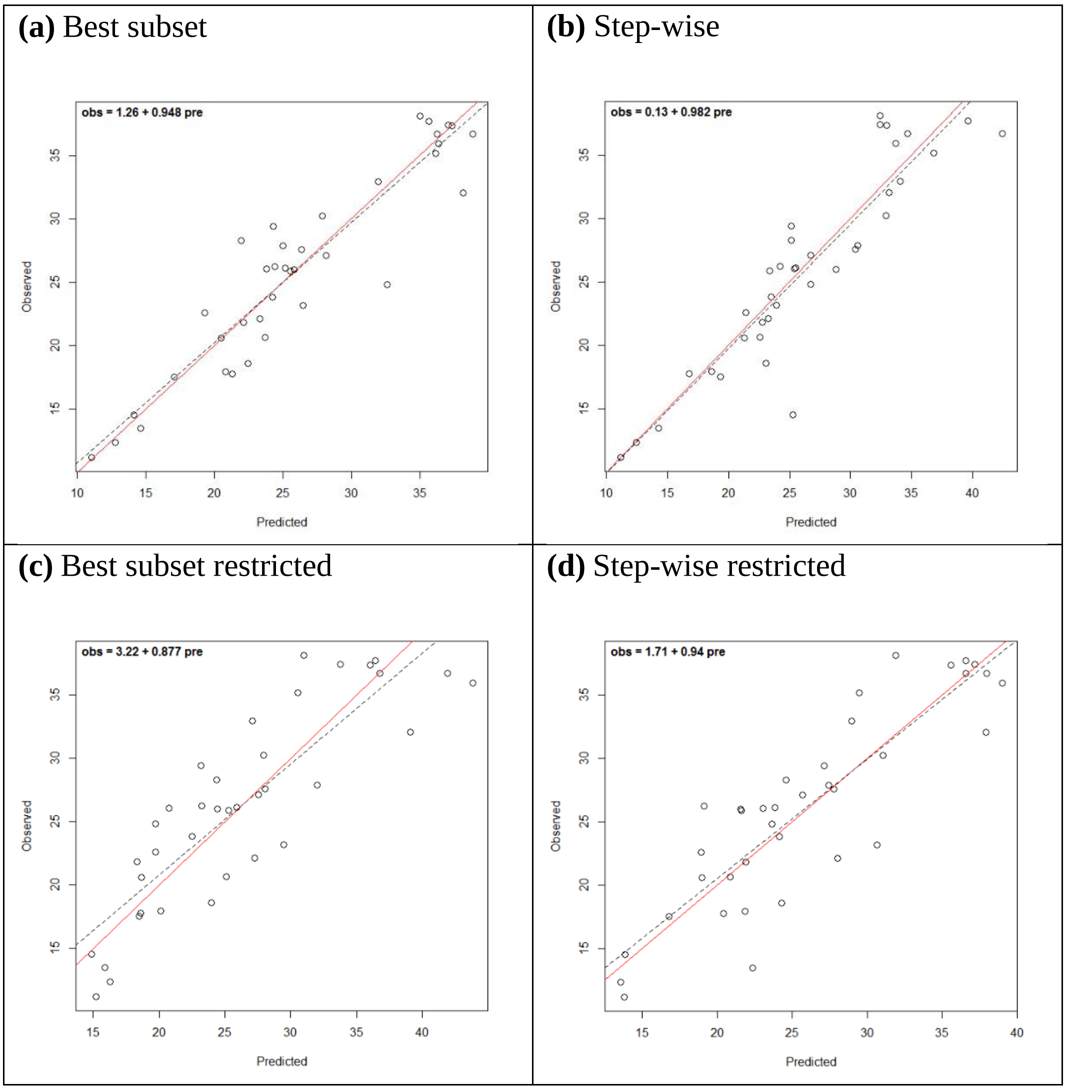


Figure 3. Observed versus predicted plots of different modelling and variable selection alternatives to obtain above-ground biomass (AGB, Mg·ha^{-1}) predictions. The solid red line represents the 1:1 correspondence. Dashed line is the linear regression fit between observed and predicted $obs_i = \alpha + \beta \cdot pre_i$.

Discussion

Importance of adding complementary analyses for assessing the accuracy of models

The most important implication of the present results is that most of the alternatives contrasted could have been reasonably judged as reliable when observing only statistical descriptors for mean difference, precision and agreement. These three types of statistics are the ones most commonly employed for assessing accuracy in this field (e.g., Zhao et al., 2009; Erdody &

Moskal 2010; McInerney et al., 2010; d'Oliveira et al 2012; Chen & Zhu, 2013; Straub et al., 2013; Valbuena et al., 2014). In our analysis, by looking only at $MD\%$, $RMSD\%$ and R^2_{cv}, and also most scatterplots in **Figs. 2-3**, it could be rationally deduced that any option including a MSN imputation with $p =$ 1-28 would yield reliable accuracies, including the best-subset and step-wise models as well. The suggested complementary analyses however, showed that many more of the presented alternatives for AGB prediction should in fact be discarded.

Significances in the hypothesis tests suggested by Piñeiro et al. (2008) demonstrated that MSN imputations using $p = 3$ or $p = 8$ gave an insufficient fit between observed and predicted values. This diagnosis may have been difficult to make by merely observing the scatterplots (**Figs. 2b,d**). Although testing the regression of observed versus predicted values is a well-settled practice in ecological modelling (Graybill, 1976; Reynolds & Chung, 1986; Leite & Oliveira, 2002; Piñeiro et al., 2008), to our knowledge, hypothesis tests have never before been included in the evaluation of forest AGB using remote sensing, and they have seemingly been simply overlooked. The results presented in this article suggest that there may be a need to include them in future accuracy assessment procedures in this field as well. Furthermore, we also wish to seek consensus and promote the arguments advanced by Piñeiro et al. (2008) in favour using observed (on the y-axis) versus predicted (on the x-axis) – and not predicted versus observed (e.g., McRoberts et al., 2002; Holmgren et al., 2008; Zhao et al., 2009; McInerney et al., 2010; Chen & Zhu 2013; Valbuena et al., 2014) – for reporting the accuracy of remote sensing-assisted AGB estimates. Piñeiro et al. (2008) showed that such distinction matters since it may change the result and conclusions of model evaluation.

Regarding the overfitting tests based on cross-validation (Allen, 1974; Snee, 1977; Geisser & Eddy, 1979; Weisberg, 1985; Hawkins, 2004), we wish to emphasize that SSR succeeded in revealing both the best-subset and the step-wise models initially considered, and also any MSN imputation using $p \geq 8$, as being unreliably overfitted to the sample and therefore hardly generalizable. The described step-wise and best subset approaches to variable selection are very frequently employed in remote sensing-assisted estimations of forest attributes (e.g., Naesset, 2002; Hudak et al., 2006; Wing et al., 2012; Straub et al., 2013; Estornell, et al. 2014). We therefore suggest that accuracy assessment procedures for AGB predictions obtained from remote sensing should be improved by using hypothesis testing and overfitting evaluation.

Unveiling averaging effects: unbiased means, and yet over/under-predicting

Even having an unbiased prediction method and a robust sampling design, the outcome is still susceptible to under- and over-prediction within specific ranges of AGB values. Sometimes the discrepancy between observed and predicted is due to an averaging effect, which in practice translates into an underestimation of large AGB values and an overestimation at areas of lesser AGB, which in turn may remain concealed if only observing the bias of the population mean. The averaging effect is a typical and intrinsic weakness of nearest neighbours methods (Franco-Lopez et al., 2001; McInerney et al., 2010). It is caused by the lack of available neighbours beyond the limits of the observed AGB range, hence tending to shift the predictions towards the average for values located in the borderline of that range. This effect therefore becomes more accentuated as the n/p ratio decreases (McRoberts et al., 2002). Our results indicate this shortcoming may be detected with the assistance of hypothesis tests suggested by Piñeiro et al. (2008) and Theil's (1958) partial inequality coefficients (Smith & Rose, 1995; Paruelo et al., 1998). Taking our results and as a rule of thumb, we would suggest that the proportions of error due to causes other than the residual variance must not exceed the thresholds $U_{slope} \leq 10\%$ and $U_{bias} \leq 1\%$, and in general the model error itself should be no lesser than $U_{error} \geq 90\%$.

Under-prediction in areas of large AGB is a common problem in remote sensing assessments (e.g., Bright et al., 2012; Asner & Mascaro, 2014), and these areas are of very high importance for the purposes of the inventory. To our knowledge, however, these coefficients have not been employed in the context of remote sensing estimates of forest characteristics before, and only García et al. (2010) resolved the $RMSD$ into systematic and unsystematic portions. U_{slope} could still be useful for identifying these averaging effects, as it was revealed in our results for MSN imputations using $p = 3$ or $p = 8$ (**Table 2**), where averaging effects were indeed undergoing (**Fig. 2b,d**). This flaw was also detected by significant results in the hypothesis tests. Therefore, averaging effects may be detected by either large values of U_{slope}, or via interpretation of α or β coefficients. When statistical significance proves $\alpha \neq 0$ but cannot reject $\beta = 1$, it is an indication for a source of systematic under- or over-prediction along the full AGB range. If $\alpha = 0$ cannot be rejected but $\beta \neq 1$ significantly, the under-prediction is concentrated in values of large AGB only, for instance due to saturation of the remote sensor. A combination of $\alpha = 0$ and $\beta \neq 1$ may as well indicate an over-prediction for small AGB values. If both null hypotheses are rejected and we accept $\alpha \neq 0$ and $\beta \neq 1$, then we are detecting an averaging effect whenever $\beta < 1$, as was the case in many of the results presented in this study.

Overfitting to the field sample training the prediction method

We also detected potential problems of overfitting in some of the alternatives proposed. Such contingency would in practice have a harmful effect when applying the resulting fit to the predictor variables to obtain AGB maps. It is noteworthy that the added value of remote sensing, compared to traditional design-based sampling using field plots only, is on the capacity to provide AGB predictions throughout large inaccessible forest areas (Naesset, 2002; McRoberts et al., 2013; Asner & Mascaro, 2014; Chen et al., 2015; Mauro et al., 2016). This advantage is therefore lost if overfitting to the sample renders AGB predictions unreliable at the pixel scale, even if the population mean estimate is unbiased. We therefore suggest the inclusion of overfitting measures in addition to those already widespread: mean difference, precision and agreement.

The true degree of overfitting will always remain elusive unless an external validation using an independent field AGB dataset is carried out (Allen, 1974; Snee, 1977; Geisser & Eddy, 1979; Hawkins, 2004). However, even in the event of having the opportunity to acquire a large enough number of plots from the field, modellers would face trade-offs between the advantages separating a subset for validation of the main dataset and the gain in incorporating them to the model for increasing its degrees of freedom, strengthening the certainty of the relationships found, and the power of their statistical inference (Cohen et al., 2003). As an alternative, the cross-validation approach seems to provide a good indicative proxy for assessing overfitting (Weisberg, 1985; Rencher & Pun, 1993; Vanclay & Skovsgaard, 1997; Hawkins, 2004). SSR succeeded in identifying risk of overfitting for some of the alternatives in **Tables 1 and 2** that could have otherwise remained undetected. For this reason, we suggest that SSR may provide a useful indication that a given predictive method may undergo overfitting effects. For predictor variable selection purposes, a desirable limit for model rejection may be chosen, as for instance we suggested to limit $SSR \leq 1.1$. It is worth emphasizing that such limit should also be employed in combination with the suggested hypothesis test, since otherwise MSN imputations using $p = 2$-3 would have been deemed reliable if judged on the basis of SSR only (**Table 1**). Surprisingly, decreasing p did not univocally lead to a decrease in SSR and $R2R$, and hence it may be as detrimental to have either too few or too many predictors. The key question is possibly to include in the model only non-collinear predictors which truly add separate portions of explained variance in the observed AGB (Ehrenberg, 1982; Weisberg, 1985).

Regarding the choice of either SSR or $R2R$ for assessing overfitting, our results showed unexpected differences which may in practice be critical. **Fig. 1d** demonstrated that the values obtained by SSR or $R2R$ diverged from $p \geq 8$. As a result, $R2R$ was too low at high values of p, which in practice would imply insufficiently low n/p ratios, and therefore the reliability of $R2R$ as a measure of overfitting is questionable. We therefore suggest that evaluating the inflation in the sums of squares of errors (Weisberg, 1985; e.g., Valbuena et al., 2013; Almeida et al., 2016) may be a more sensible approach to assessing overfitting than analysing the deflation in R^2 (Rencher & Pun, 1993; e.g., Latifi et al., 2015a).

Most studies assume that avoiding over-parametrized models via κ, VIP, Cp or AIC is sufficient to avoid overfitting (e.g., Naesset, 2002; Hudak et al., 2006; Erdody & Moskal, 2010; García et al., 2010; Bright et al., 2012; Wing et al., 2012; Latifi et al., 2015a; Spriggs et al, 2015). Many of these indices, however, have been suspected in some occasions of being insufficient to avoid model overfitting (Hurvich & Tsai, 1989; Rencher & Pun, 1993; Vanclay & Skovsgaard, 1997). In the present research we also detected the need for incorporating further restrictions to Cp and AIC (**Table 2**). As an alternative, Weisberg (1985) and Hawkins (2004) recommended using cross-validation to prevent overfitting. Our results suggest that, while model precision was not excessively compromised, the assessment of overfitting presented an opportunity for increased reliability of remote sensing predictions of AGB (Franco-Lopez et al., 2001; Valbuena et al., 2013b; Latifi et al., 2015a). We therefore suggest that in addition to the use of κ, VIP, Cp or AIC, a specific measure devoted to evaluate the degree of overfitting, such as SSR, should become a general requirement

Conclusions

Given the results presented in these comparisons, we wish to put forward a suggestion to perform a more thorough analysis of accuracy in ecological models, which in particular we wish to address to authors carrying out remote sensing-assisted predictions of AGB. We may draw four main conclusions from the discussion of our results (plus an additional one, see

Valbuena et al., 2018). (1) By simply looking at the most common measures of accuracy assessment – mean difference, precision and agreement –, there is a risk of interpreting as reliable AGB predictions which are in fact unreliable. MD, $RMSD$ and R^2 are useful statistics for accuracy assessment, but perhaps not sufficient for truly evaluating the convenience of a given prediction alternative. (2) Piñeiro et al.'s (2008) hypothesis tests were clearly useful in providing objective means for inferring the statistical significance of the agreement between observed and predicted values, which would otherwise be difficult to grasp just by visual diagnosis of scatterplots. (3) Theil's partial inequality coefficients can be useful for diagnosis of the causes leading to disagreement, detecting averaging effects or other types of under- or over-predictions occurring at specific ranges of AGB. (4) We also observed that overfitting effects may remain concealed unless specifically addressed. When comparing the evaluation of inflation in sums of squares versus deflation of R^2, our results suggested the former to be a more advantageous approach. We therefore recommend researchers to incorporate the presented statistical measures for (2), (3) and (4) in their own accuracy assessment protocols. This recommendation may, of course, be extended to other fields of applied ecological modelling as well.

Acknowledgments

This work was partially supported by the Spanish Directorate General for Scientific and Technical Research under Grant CGL2013-46387-C2-2-R. We also thank the Valsain Forest Centre, of the National Park Body (Spain), for their valuable help. Dr Valbuena and Prof. Coomes work is supported by an EU Horizon 2020 Marie Sklodowska-Curie Action entitled "Classification of forest structural types with LIDAR remote sensing applied to study tree size-density scaling theories" (LORENZLIDAR-658180). Danilo Almeida acknowledges support from São Paulo Research Foundation (FAPESP) (grant 2016/05219-9). The authors are grateful for the comments received from the editor and reviewers.

References

Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16: 125-127

Almeida DRA, Nelson BW, Schietti J, Gorgens EB, Resende AF, Stark SC, Valbuena R. (2016) Contrasting fire damage and fire susceptibility between seasonally flooded forest and upland forest in the Central Amazon using portable profiling LiDAR. *Remote Sensing of Environment* 184: 153-160

Akaike H (1992) Information theory and an extension of the maximum likelihood principle. In: Kotz S, Johnson NL (Eds) *Breakthroughs in statistics* 1. Springer, London, pp 610–624

Asner GP, Mascaro J (2014). Mapping tropical forest carbon: Calibrating plot estimates to a simple LiDAR metric. *Remote Sensing of Environment* 140: 614-624

Axelsson P (2000) DEM generation from laser scanner data using adaptive TIN models. *International Archives of Photogrammetry and Remote Sensing* 33, Part B4: 110-117

Baskerville G (1972) Use of logarithmic regression in the estimation of plant biomass. *Canadian Journal of Forest Research* 2: 49-53

Bright BC, Hicke JA, Hudak AT (2012) Estimating aboveground carbon stocks of a forest affected by mountain pine beetle in Idaho using lidar and multispectral imagery. *Remote Sensing of Environment* 124: 270–281

Burnham KP, Anderson DR (2002) *Model selection and multimodel inference: a practical information-theoretic approach* (2nd Edition). Secaucus, NJ, USA: Springer

Chen Y, Zhu X (2013) An integrated GIS tool for automatic forest inventory estimates of Pinus radiata from LiDAR data. *GIScience & Remote Sensing* 50: 667-689

Chen Q, Vaglio Laurin G, Valentini R (2015) Uncertainty of remotely sensed aboveground biomass over an African tropical forest: Propagating errors from trees to plots to pixels. *Remote Sensing of Environment* 160: 134-143

Cohen WB, Maiersperger TK, Gower ST, Turner DP (2003) An improved strategy for regression of biophysical variables and Landsat ETM+. *Remote Sensing of Environment* 84: 561-571

Crookston NL, Finley AO (2007) yaImpute: An R package for κNN imputation. *Journal of Statistical Software* 23 (10): 1-16

d'Oliveira MVN, Reutebuch SE, McGaughey RJ, Andersen HE (2012) Estimating forest biomass and identifying low-intensity logging areas using airborne scanning lidar in Antimary State Forest, Acre State, Western Brazilian Amazon. *Remote Sensing of Environment* 124: 479-491

Ehrenberg ASC (1982). How good is best? *Journal of the Royal Statistical Society,* Series A 145: 364-366

Erdody TL, Moskal LM (2010) Fusion of LiDAR and imagery for estimating forest canopy fuels. *Remote Sensing of Environment* 114, 725−737

Estornell J, Velázquez-Martí B, López-Cortés I, Salazar D, Fernández-Sarría A (2014) Estimation of wood volume and height of olive tree plantations using airborne discrete-return LiDAR data. *GIScience & Remote Sensing* 51, 17-29.

Fox DG (1981) Judging air quality model performance. *Bulletin of the American Meteorological Society* 62: 599–609

Fox J, Monette G (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association* 87: 178-183.

Franco-Lopez H, Ek AR, Bauer ME (2001) Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment* 77(3): 251-274

Freese F (1960) Testing accuracy. *Forest Science* 6: 139-145.

Eskelson, B.N.I., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L. & Hudak, A.T. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research* 24: 235–246.

García M, Riaño D, Chuvieco E, Danson FM (2010) Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LiDAR height and intensity data. *Remote Sensing of Environment* 115: 1369-1379

Geisser S, Eddy W (1979) A Predictive Approach to Model Selection. *Journal of the American Statistical Association* 74 (365): 153-160.

Graybill FA (1976) *Theory and Application of the Linear Model*. Duxbury Press: Belmont, CA.

Hawkins DM (2004) The problem of overfitting. *Journal of chemical information and computer sciences* 44.1: 1-12.

Holmgren J, Persson A, Soderman U (2008). Species identification of individual trees by combining high resolution LIDAR data with multi-spectral images. *International Journal of Remote Sensing* 29: 1537−1552

Hudak AT, Crookston NL, Evans JS, Falkowski MJ, Smith AMS, Gessler PE, Paul E, Morgan P (2006) Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. *Canadian Journal of Remote Sensing* 32: 126-138

Hudak AT, Crookston NL, Evans JS, Hall DE, Falkowski MJ (2008) Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112 (5): 2232–2245

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. *Biometrika* 76: 297–307

Latifi H, Heurich M, Hartig F, Müller J, Krzystek P, Jehl H, Dech S (2015a) Estimating over- and understorey canopy density of temperate mixed stands by airborne LiDAR data. *Forestry* 89 (1): 69-81.

Latifi H, Fassnacht FE, Müller J, Tharani A, Dech S, Heurich M (2015b) Forest inventories by LiDAR data: A comparison of single tree segmentation and metric-based methods for inventories of a heterogeneous temperate forest. *International Journal of Applied Earth Observation and Geoinformation* 42, 162-174

Leggett RW, Williams LR (1981) A reliability index for models. *Ecological Modelling* 13: 303–312.

Leite HG, Oliveira FHT (2002) Statistical procedure to test identity between analytical methods. *Communications in Soil Science and Plant Analysis* 33 (7-8): 1105-1118

Lipovetsky S (2013) How good is best? Multivariate case of Ehrenberg-Weisberg analysis of residual errors in competing regressions. *Journal of Modern Applied Statistical Methods* 12 (2): 14

Lumley T, Miller A (2009) *leaps: regression subset selection*. R package version 2.9. https://CRAN.R-project.org/package=leaps

Mallows CL (1973) Some Comments on Cp. *Technometrics* 15 (4): 661–675

Manzanera JA, Garcia-Abril A, Pascual C, Tejera R, Martín Fernández S, Tokola T, Valbuena R (2016) Fusion of airborne LiDAR and multispectral sensors reveals synergic capabilities in forest structure characterization. *GIScience & Remote Sensing* 53: 723-738

Mauro F, Molina I, Garcia-Abril A, Valbuena R, Ayuga-Téllez E (2016) Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics* 27(4): 225-238

McGaughey RJ (2012). *FUSION/LDV: Software for LIDAR data analysis and visualization*. Version 3.10. USDA Forest Service. Seattle, Washington, USA.

McInerney DO, Suárez J, Valbuena R, Nieuwenhuis M (2010). Forest canopy height retrieval using Lidar data, medium-resolution satellite imagery and kNN estimation in Aberfoyle, Scotland. *Forestry* 83(2): 195-206

McRoberts RE, Nelson MD, Wendt DG (2002) Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. *Remote Sensing of Environment* 82(2-3): 457-468

McRoberts RE, Naesset E, Gobakken T (2013) Accuracy and precision for remote sensing applications of nonlinear model-based inference. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6 (1): 27-34

Miller A (1984) Selection of subsets of regression variables. *Journal of the Royal Statistical Society*, Series A 147: 389-425

Moeur M, & Stage AR (1995) Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science* 41 (2): 337-359.

Montero G, Ruiz-Peinado R, Muñoz M. (2005) *Producción de biomasa y fijación de CO2 por los bosques españoles*. Monografías Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Serie Forestal, Madrid, Spain (in Spanish).

Naesset E (2002) Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment* 801: 88-99

Packalén P, Maltamo M (2008) Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Canadian Journal of Forest Research* 38(7): 1750-1760

Paruelo JM, Jobbágy EG, Sala OE, Lauenroth WK, Burke I (1998) Functional and structural convergence of temperate grassland and shrubland ecosystems. *Ecological Applications* 8(1): 194–206

Piñeiro G, Perelman S, Guerschman JP, Paruelo JM (2008) How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecological Modelling* 216(3): 316−322

R Development Core Team (2016) *R: a language and environment for statistical computing*.

Rencher AC, Pun CP (1993) Inflation of R^2 in best subset regression. *Technometrics* 22: 49-53

Reynolds MR, Chung J (1986) Regression methodology for estimating model prediction error. *Canadian Journal of Forest Research*, 16 (5): 931-938

Rouse JW, Haas RH, Scheel JA, Deering DW (1974) Monitoring vegetation systems in the great plains with ERTS. *Proceedings, 3rd Earth Resource Technology Satellite (ERTS) Symposium* 1: 48-62

Särndal CE, Swensson B, Wretman J (1992) *Model assisted survey sampling*. Springer-Verlag, Inc. New York

Smith EP, Rose KR (1995) Model goodness of fit using regression and related techniques. *Ecological Modelling* 77: 49-64

Snee R (1977) Validation of Regression Models: Methods and Examples. *Technometrics* 19(4): 415-428.

Spriggs R, Vanderwel M, Jones T, Caspersen J, Coomes D (2015) A simple area-based model for predicting airborne LiDAR first returns from stem diameter distributions: an example study in an uneven aged, mixed temperate forest. *Canadian Journal of Forest Research* 45 (10): 1338-1350

Sprugel DG. (1983) Correcting for bias in log-transformed allometric equations. *Ecology* 64: 209-210

Straub C, Tian J, Seitz R, Reinartz P (2013) Assessment of Cartosat-1 and WorldView-2 stereo imagery in combination with a LiDAR-DTM for timber volume estimation in a highly structured forest in Germany. *Forestry* 86 (4): 463-473

Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* 7 (1): 13-26.

Tedeschi LO (2006) Assessment of the adequacy of mathematical models. *Agricultural Systems* 89 (2-3): 225-247

Theil H (1958) *Economic Forecasts and Policy*. Amsterdam: North Holland

Valbuena R (2014) Integrating airborne laser scanning with data from global navigation satellite systems and optical sensors. In: Maltamo M, Næsset E, Vauhkonen J (Eds.) *Forestry applications of airborne laser scanning. Concepts and case studies*. Managing Forest Ecosystems Series 27. Springer, Dordrecht. pp. 63-88

Valbuena R, Mauro F, Arjonilla F, Manzanera JA (2011) Comparing airborne laser scanning-imagery fusion methods based on geometric accuracy in forested areas. *Remote Sensing of Environment* 115: 1942–1954

Valbuena R, Mauro F, Rodriguez-Solano R, Manzanera JA (2012) Partial least squares for discriminating variance components in global navigation satellite systems accuracy obtained under Scots pine canopies. *Forest Science* 582: 139-153

Valbuena R, Maltamo M, Martín-Fernández S, Packalen P, Pascual C, Nabuurs GJ (2013a) Patterns of covariance between airborne laser scanning metrics and Lorenz curve descriptors of tree size inequality. *Canadian Journal of Remote Sensing* 39 (S1), S18–S31

Valbuena R, Packalen P, Mehtätalo L, Garcia-Abril A, Maltamo M (2013b) Characterizing forest structural types and shelterwood dynamics from Lorenz-based indicators predicted by airborne laser scanning. *Canadian Journal of Forest Research* 43: 1063–1074.

Valbuena R, Vauhkonen J, Packalén P, Pitkanen J, Maltamo M (2014) Comparison of airborne laser scanning methods for estimating forest structure indicators based on Lorenz curves. *ISPRS Journal of Photogrammetry and Remote Sensing* 95: 23-33

Valbuena R., Hernando A., Manzanera J.A., Görgens E.B., Almeida D.R.A., Mauro F., García-Abril A. and Coomes D.A. (2018) Evaluating observed versus predicted: R-squared, index of agreement or maximal information coefficient? (forthcoming).

Vanclay JK, Skovsgaard JP (1997) Evaluating forest growth models. *Ecological Modelling* 98: 1-12

Venables WN, Ripley B D (2002) *Modern Applied Statistics with S* (4^{th} Ed.). Springer, New York

Wallach D, Goffinet B (1989) Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling* 44: 299–306

Weisberg S (1985) *Applied linear regression* (2^{nd} Ed.) John wiley & Sons, New York

White JD, Coops NC, Scott NA (2000) Estimates of New Zealand forest and scrub biomass from the 3-PG model. *Ecological Modelling* 131: 175–190

Willmott CJ (1981) On the validation of models. *Physical Geography* 2: 184–194

Willmott CJ (1982) Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* 63 (11): 1309−1313

Wing BM, Ritchie MW, Boston K, Cohen WB, Gitelman A, Olsen MJ (2012) Prediction of understory vegetation cover with airborne lidar in an interior ponderosa pine forest. *Remote Sensing of Environment* 124: 730-741

Yebra M, Chuvieco E (2009) Linking ecological information and radiative transfer models to estimate fuel moisture content in the Mediterranean region of Spain: Solving the ill-posed inverse problem. *Remote Sensing of Environment* 113 (11): 2403-2411

Zhao K, Popescu S, Nelson R (2009) Lidar remote sensing of forest biomass: a scale-invariant estimation approach using airborne lasers. *Remote Sensing of Environment* 113 (1): 182-196