https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

Biogeosciences
Discussions
Open Access
EGU

Improving non-representative-sample prediction of forest aboveground biomass maps: A combined machine learning and spatial statistical approach

Shaoqing Dai [1,2], Xiaoman Zheng [1,2], Lei Gao [3], Shudi Zuo [1,2,4,] Qi Chen [5], Xiaohua Wei [6], Yin Ren [1,4]

[1] Key Laboratory of Urban Environment and Health, Key Laboratory of Urban Metabolism of Xiamen, Institute of Urban Environment, Chinese Academy of Sciences, CN 361021, China
[2] University of Chinese Academy of Sciences, CN 100049, China
[3] CSIRO, Waite Campus, Urrbrae, SA 5064, Australia
[4] Ningbo Urban Environment Observation and Research Station-NUEORS, Chinese Academy of Sciences, CN 315800, China
[5] Department of Geography, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA
[6] Department of Earth and Environmental Sciences, University of British Columbia, Kelowna, BC V1V 1V7, Canada

Correspondence to: Yin Ren(yren@iue.ac.cn)

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

Biogeosciences Discussions
Open Access
EGU

Abstract: High-precision prediction of large-scale forest aboveground biomass (AGB) is important but challenging on account of the uncertainty involved in the prediction process from various sources, especially the uncertainty due to non-representative sample units. Usually caused by inadequate sampling, non-representative sample units are common and can lead to geographic clusters of localities. But they cannot fully capture complex and spatially heterogeneous patterns, in which multiple environmental covariates (such as longitude, latitude, and forest structures) affect the spatial distribution of AGB. To address this challenge, we propose herein a low-cost approach that combines machine learning with spatial statistics to construct a regional AGB map from non-representative sample units. The experimental results demonstrate that the combined methods can improve the accuracy of AGB mapping in regions where only non-representative sample units are available. This work provides a useful reference for AGB remote-sensing mapping and ecological modelling in various regions of the world.

Keywords: Aboveground biomass map, Non-representative sample units, Machine learning, Spatial statistical model, small-size samples

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

1 Introduction

Accurate mapping of aboveground biomass (AGB) can provide a precise scientific basis for decision-making in sustainable forest management, involving reducing deforestation, forest degradation, and greenhouse-gas emissions (Bustamante et al., 2016;Houghton et al., 2009;Mendoza-Ponce and Galicia, 2010). AGB maps are usually constructed based on biomass data obtained from small-size samples or geographically limited locations. The uncertainty in such maps can be attributed to two primary sources: (1) inadequate sampling designs used to obtain the data for constructing prediction models, especially geographically limited designs that do not capture the entire range of conditions; and (2) model-dependent uncertainty, including unreasonable model parameter assumptions as well as improper model structure(Chen et al., 2015;Gao et al., 2016;McRoberts et al., 2016).

An estimated 20%–50% of the uncertainty in AGB mapping can be attributed to the inadequate sampling design (Pelletier et al., 2011). To alleviate the uncertainty derived from inadequate sampling and its consequences (i.e., non-representative samples), one type of approaches focuses on processing model input samples (front-end processing), in the form of spatial filtering of existing sample units, quantification of sampling uncertainty, and acquisition of representative sample units (Boria et al., 2014;Galante et al., 2017;Marvin et al., 2014). Although the front-end processing approaches are widely used to reveal the distribution of biological populations, they are rarely used in AGB mapping because they make it difficult to quantify sampling uncertainty and produce large samples for spatial filtering methods. Another type of approaches, in the case of non-representative samples, increase the prediction accuracy by screening or building an optimal adaptive model (back-end processing) (Boria et al., 2014). These approaches may substantially increase the accuracy of AGB maps.

A sizable group of prediction models has been applied to constructing accurate AGB maps, including linear models (Andersen et al., 2014;Morel et al., 2012), machine learning models(Chen, 2015;Gleason and Im, 2012), and spatial statistical models (Benitez et al., 2016;Propastin, 2012;Van der Laan et al., 2014). With the development of computer-science techniques and advances in nonlinear biomass modeling, machine learning methods have become prevalent. Compared to traditional parametric methods (these methods summarize data with a fixed number of parameters with respect to the sample size, such as logistic regression and perceptron)(Gao and Hailu, 2012), which have difficulty in characterizing nonlinear relationships between AGB and multiple environmental covariates,

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

nonparametric machine learning algorithms (the number of parameters in a nonparametric method is dependent of the number of training examples, e.g., K-nearest neighbor, support vector machine, and random forest) are advantageous because they are more elastic and have neither restrictions on variable types nor strict requirements regarding the distributions of predictor variables as well as the relationship between response and predictor variables (Lu et al., 2007). In addition, nonparametric machine learning algorithms may offer higher prediction accuracy (Frey et al., 2019;Gleason and Im, 2012).

Another frequently-used group of models for estimating relationships between forest AGB and multiple environmental covariates is based on spatial statistical approaches, such as geographically weighted regression and Kriging (Du et al., 2010;Van der Laan et al., 2014;Viana et al., 2012). Spatial statistical methods are based on the analysis of attribute information that includes spatial locations (Schabenberger and Gotway, 2005). Compared with traditional statistical methods, spatial methods integrate spatial factors affecting model responses, thus remove the constraint of traditional statistical methods that assume sample independence (Rangel and Bini, 2010) and improve the understanding of spatial autocorrelation and heterogeneity (He et al., 2011;Rosenberg and Anderson, 2011).

Although many studies have integrated plot data, multi-source remote-sensing data (e.g., lidar and Landsat), and machine learning or spatial statistical methods, the prediction accuracy of current AGB spatial mapping still suffers from uncertainty (Asner et al., 2012;Chen et al., 2016;Gregoire et al., 2016;McRoberts et al., 2018;Paul et al., 2016;Saatchi et al., 2011;Zheng et al., 2004) for two reasons. First, the existing studies with machine learning methods do not consider the spatial heterogeneity of multiple environmental covariates (such as longitude, latitude, and forest structures) that affect the spatial distribution of AGB (Babcock et al., 2015;Fassnacht et al., 2014). Uncertainty can be further magnified by applying regional area models to small-size samples or geographically limited samples. The second reason lies in the assumptions of the spatial statistical method (e.g., spatial autocorrelation and stability of the second steps), which may not always be valid in forest AGB.

The objective of this study is to develop and evaluate a method for improving the prediction accuracy of large-scale AGB spatial mapping given small-size, non-representative, and local geographically clustered samples. The method integrates the nonlinear mapping capabilities of machine learning algorithms (artificial neural network, support vector machine, and random forest) with the spatial autocorrelation and stratified heterogeneous advantages of a spatial statistical model (the Point Estimation Model of Biased Sentinel Hospitals-based Area Disease Estimation, P-BSHADE model)(Xu

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

et al., 2013). Our aim is tantamount to answer two specific questions: (1) What are the differences in prediction accuracy of AGB maps for different machine learning methods and between machine learning and spatial statistical methods? (2) Can the integration of spatial statistical and machine learning methods improve the accuracy of AGB maps based on small-size, non-representative samples in the form of local geographic clusters of forest inventory data? We explore these two questions by considering an empirical case of predicting an AGB map for *Eucalyptus* plantation in Nanjing County, China.

2 Materials and Methods

2.1 Site description

Nanjing County (117°00'–117°36'E, 24°26'–25°00'N, Figure 1b) is located in the upper upstream area of the Jiulong River, Fujian Province, China. Seventy-four percent (146,130 ha) of the county is covered with forests, where 79,346 ha are plantations. The region is affected by the South Asian tropical monsoon climate. In 2014, the average annual temperature in Nanjing County was 21.1°C, with an annual precipitation of 1,700 mm and 340 frost-free days. Red soil is its major soil type.

The elevation in the study area varies significantly (0–1,566 m), with complex topography. There is also major spatiotemporal heterogeneity in forest composition, structure, and biomass. The main types of trees are *Eucalyptus*, *Pinus massoniana*, and *Cunninghamia lanceolata*. Recently the plantation area of *Eucalyptus* has increased rapidly, reaching 13,305 ha and increasing by 10,862 ha in one decade.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

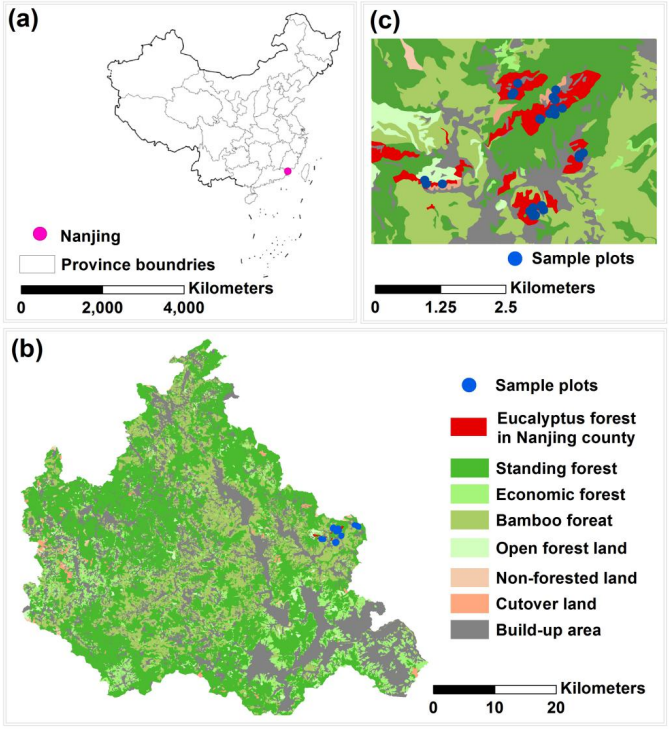


Figure 1. The study area is a typical example of a non-representative-sample problem. (a) Geographical location of the study area. (b) Spatial distributions of Eucalyptus plantations (red) and other major forests. (c) Spatial distributions of 30 sample plots (blue).

2.2 Datasets

The datasets included Forest Management and Planning Inventory (FMPI) data, sample plot data, and analytic-tree (destructive measurement) data.

2.2.1 Forest Management and Planning Inventory (FMPI) data

The FMPI data for the whole study area were provided by the Forestry Department of Fujian Province, China. By using large-scale sampling methods, this forest resource inventory collected detailed information about the characteristics and conditions of each type of forest. We selected the FMPI data of *Eucalyptus* plantation forest in this study.

The FMPI data were stored by patch and all trees with a diameter at breast height (DBH) greater than 8 cm were measured. The data contained (1) stand data (patch area, tree age which is the same for all trees in a given patch because they were planted at the same time, plantation density, mean DBH, mean tree

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

height, and total volume of each patch), (2) soil data (soil depth, humus depth, and site index of each patch), and (3) topographical data (elevation, slope degree, slope direction, and slope position of each patch). All variables were measured within each forest patch, with the average value being used as the factor value for each patch. The accuracy of forest patch attributes was tested based on differences in volume using a combined method of systematic and stratified samplings. A 95% sampling precision was required. Table B.1 lists the statistical description of the forest patch data.

2.2.2 Sample plot data

A total of 30 fixed sample plots were selected in the Yongfeng forest farm. The plots were located in the eastern part of the study area (Figure 1). The sample area accounted for 0.007% of the total area and featured local geographic plot clusters. Thus, the sample size was small and the sample units were not representative of the entire area. The 30 sampling plots with 10 age groups were built for *Eucalyptus* plantation patches. In each plot, tree height (H) and DBH of each tree were measured. In addition, mean plot-level variables were measured, including stand age, density, soil variables, and topographical variables.

2.2.3 Analytic-tree data

The analytic-tree data were derived from standard wood in 30 fixed sample plots. Three trees were cut in each of the plots, totaling 90 trees for 30 plots. We then calculated the biomass of each organ (foliage, stems, and roots) for each tree. In addition, DBH and H were measured. Table B.2 presents the data of 90 parse trees. Details of the selection of standard wood and the cutting process are provided in S1 of Supplementary Material.

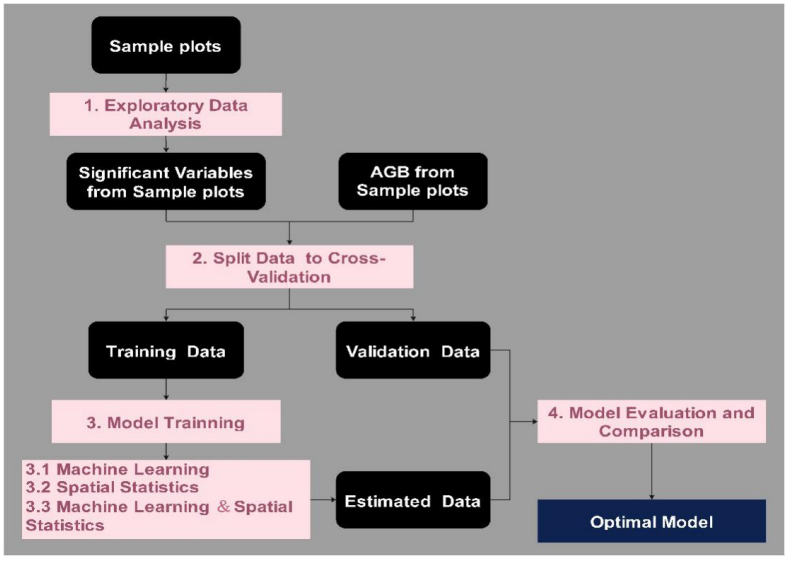
2.3 Construction of tree-level allometric models

All analyses were based on the underlying assumption that the relationship between the response and predictor variables in the sample data used to construct models was the same as the relationship in the entire population. Using 90 analytic-tree data, three age groups (age 1-2, age 3-5, age 6-10) of allometric models were constructed. Allometric models were then applied to each tree in each sample plot according to their ages, hence producing a reference AGB of sample plots.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

152 **2.4 Construct plot-level models to alleviate non-representative sample uncertainty**

153 The sample plots in this study were located in the east of the case study area and presented as non-
154 representative samples (Figure 1). Processing based on model screening was applied to alleviate the
155 uncertainty caused by non-representative samples and consisted of the following four steps (Figure 2).



156

157 Figure 2. Structure of the optimal model screening scheme.

158 **2.4.1 Exploratory data analysis**

159 We first identified predictor variables used for creating the plot-level model. Based on our previous work
160 (Ren et al., 2017), we selected plot-level environmental covariates including longitude and altitude, and
161 forest attribute variables including forest distribution density, DBH, H, tree stem volume, timber volume,
162 and forest age. The Pearson's correlation coefficient was used to investigate the correlation between
163 these variables and the reference AGB of sample plots.

164 We then analyzed the spatial autocorrelation and spatial heterogeneity of AGB data from the selected
165 non-representative sample plots. We used Moran's I (Cliff and Ord, 1981), a commonly used global
166 spatial autocorrelation index, to evaluate spatial autocorrelation among the reference AGB of sample
167 plots. The spatial stratified heterogeneity of the reference AGB of sample plots was evaluated using a
168 geographic detector, as proposed by Wang et al. (2010).

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019


Biogeosciences Discussions Open Access EGU

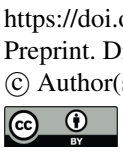
169 **2.4.2 Split data sets**

170 We used the leave-one-out cross-validation method to split the 30 sample plots into 30 sets with each set
171 including two groups of data: validation data (one plot AGB) and training data (AGB and predictor
172 variables of another 29 plots), see Table B.3. The leave-one-out cross-validation method supposes that,
173 in an N-sample dataset, each sample is taken as a test sample, and the other N-1 samples are taken as
174 training samples. Thus, there are N iterations and we can obtain N datasets and N cross-validation results.

175 **2.4.3 Model training**

176 Seven models including three machine learning models (a, b, and c in Figure 3), one spatial statistical
177 model (d+e in Figure 3), and three combined machine learning and spatial statistical models (a+e, b+e,
178 and c+e in Figure 3) were developed and trained to simulate the reference AGB of sample plots (Figure
179 3). As shown in Figure 3, the three machine learning models are support vector machine (SVM, a), radial
180 basis function-artificial neural network (RBF-ANN, b), and random forest (RF, c) models. The spatial
181 statistical model, named P-BSHADE, required reference plot AGB data, which was obtained from the
182 localization biomass model (d). Thus, the single spatial statistical model (P-BSHADE, d+e) was
183 comprised of "d" combined with "e" in Figure 3. For the combined machine learning and spatial
184 statistical models, the reference plot AGB data in P-BSHADE was obtained from "a", "b" or "c". The
185 three combined models are represented as RBF-ANN&P-BSHADE (a+e), RF&P-BSHADE (b+e), and
186 SVM&P-BSHADE (c+e). Every model was trained based on each of 30 datasets, yielding a total of 30
187 simulated AGB datasets for 30 sample plots (see Table B.3).

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019



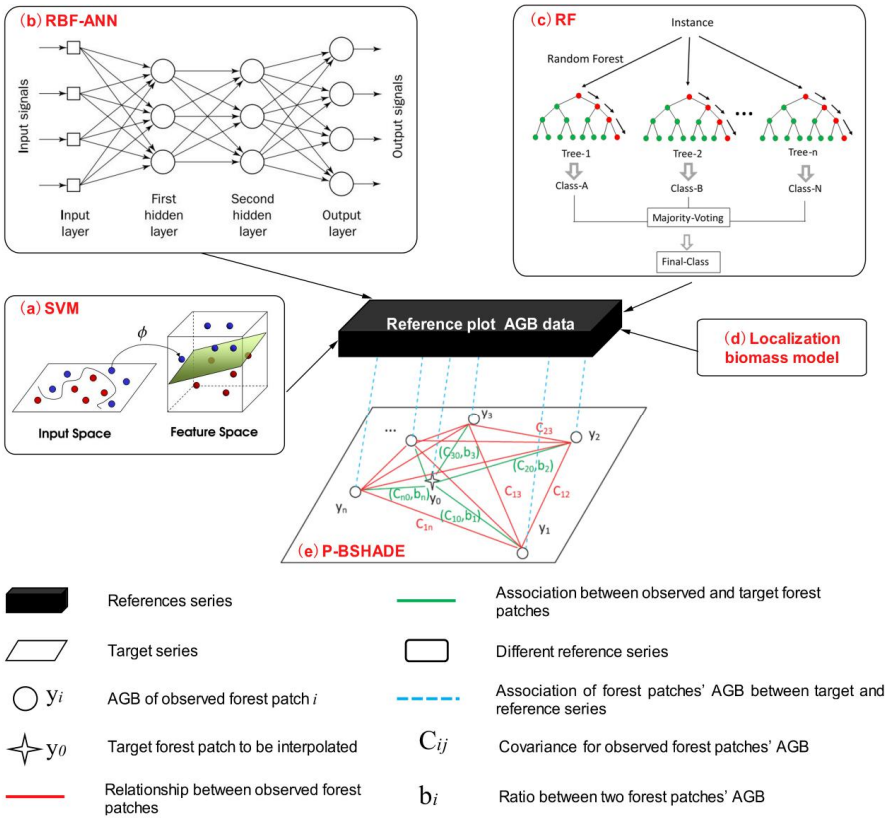


Figure 3. Framework for three machine learning (a, b, c), the PBSHADE (d+e), and three combined machine learning and PBSHADE (a+e, b+e, c+e) models for AGB estimation.

(1) Machine learning

The SVM is a type of categorized algorithms that improves the generalized machine learning ability by minimizing structural risks (so as to minimize the empirical risk and confidence intervals). In this way, the SVM can achieve adequate statistical trends from a sample set of limited size(Drucker et al., 1996). The basic components of the RBF-ANN include an input layer, a hidden layer, and an output layer, which are able to provide the best approximation for nonlinear functions and optimal global performance (Elanayar and Shin, 1994). The change from the input layer space to the hidden layer space is nonlinear, whereas the spatial transformation from the hidden layer to output layer space is linear. The RBF network not only has good generalizability, but also requires less calculation. In general, its learning speed is faster than that of other machine learning algorithms, therefore, the lengthy process of iterative

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

calculations found in the learning algorithm of back propagation neural networks and the possibility of falling into a local extremum can be avoided.

The RF is a relatively new machine learning technique. As one of modern classification and regression methods, it is a combination of self-learning technologies (Breiman, 2001). The idea of combinatorial learning is to integrate several individual classifiers when classifying new instances and to determine the final classification of the instances by combining the classification results of multiple classifiers, so as to achieve better performance than that achieved by each individual classifier.

The schematic function of machine learning is as follows

$$y_j = f(x_{j,1}, x_{j,2}, x_{j,3}, x_{j,4}) \quad (1)$$

where y_j is AGB of the j-th sample plot simulated by a machine learning model, $f(\dots)$ is a machine learning model represented by a function of $x_{j,k}$ $(k = 1, \dots 4)$, $x_{j,1}$, $x_{j,2}$, $x_{j,3}$, and $x_{j,4}$ are the longitude, the DBH, the tree height, and the forest age of the j-th sample plot, respectively. A specific description of the three machine learning models is given in S1 of Supplementary Material.

(2) Spatial statistical model: P-BSHADE

A spatial statistic model, P-BSHADE, was also used to estimate sample plot AGB. In essence, the P-BSHADE uses the reference AGB of sample plots and the weights of target sample plots AGB against reference AGB of each sample plot to obtain the AGB of the target sample plot. The P-BSHADE assumption requires knowledge of the spatial autocorrelation and spatial stratified heterogeneity of the reference AGB of sample plots. The specific mathematical expression of a P-BSHADE is as follows (Hu et al., 2013;Xu et al., 2013):

$$\hat{y}_j = \Sigma_{i=1}^{n} w_{ij} y_i \quad (2)$$

where $\hat{y}_j$ is the estimated AGB of the j-th sample plot by the P-BSHADE $(j = 1{\sim}30, n = 30)$, y_i is the reference AGB of the ith sample plot $(i = 1{\sim}30, n = 30)$, w_{ij} is the weight (contribution) of reference AGB of i-th sample plot to the AGB to be interpolated of j-th sample plot (when $j = 1, i = 2{\sim}30$; when $j = 1, i = 1,\ 3{\sim}30$). A specific description of the P-BSHADE and the corresponding algorithm formulas are presented in S1 of Supplementary Material.

(3) Combination of machine learning and spatial statistical models

P-BSHADE was separately integrated with three machine learning methods (SVM, RBF-ANN, and RF) to form three combined models (SVM&P-BSHADE, RBF-ANN&P-BSHADE, and RF&P-BSHADE). The reference AGB of 30 sample plots were replaced by the estimates produced from machine learning

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

models. A combined model can be represented as follows

$$\hat{y}_j = \Sigma_{i=1}^{n} w_{ij} y_i \quad (3)$$

where $\hat{y}_j$ is the estimated AGB of the j-th sample plot using the combined model $(j = 1{\sim}30, n = 30)$, y_i is AGB estimated by machine learning based on the i-th sample plot $(i = 1{\sim}30, n = 30)$, w_{ij} is the weight (contribution) of ith machine learning estimation AGB of the sample plot to j-th sample plot AGB to be interpolated (when $j = 1, i = 2{\sim}30$; when $j = 1, i = 1,\ 3{\sim}30$). A specific description of the combined models and the algorithm formulas are presented in S1 of Supplementary Material.

2.4.4 Model evaluation and comparison

To evaluate the prediction performance of the seven models (SVM, RBF-ANN, RF, P-BSHADE, SVM&P-BSHADE, RBF-ANN&P-BSHADE, and RF&P-BSHADE), the AGB results simulated by the seven models were compared to the reference AGB of sample plot groups (AGB group M in Table B.3) in terms of three performance indicators: mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMSE), as shown in Eq. (4)-(6).

$$\text{MAE} = \left(\sum_{i=1}^{n} |y_i^p - y_i|\right)/n \quad (4)$$

$$\text{MRE} = \left(\sum_{i=1}^{n} |y_i^P - y_i|\right)/(y_i \times n) \quad (5)$$

$$\text{RMSE} = \sqrt{\left(\sum_{i=1}^{n} \left(y_i^p - y_i\right)^2\right)/n} \quad (6)$$

where y_t^p is the predictive value of the different models, y_i is the AGB of the ith sample plot, and n is the number of training datasets.

Then, in terms of the calculated MAE, MRE, and RMSE, we identified the optimal model.

2.5 Model application

We applied the optimal model to each *Eucalyptus* forest patch and estimated the total AGB over all patches in the study area. In short, the relationship between the non-representative AGB data from the sample plots and their covariates were applied to each *Eucalyptus* forest patch in regional forests to estimate the AGB of the area.

To validate the estimated AGB map, we compared it with the AGB map obtained by an allometric model, and 95% credible interval width (CIW) was calculated and mapped for AGB. The allometric model was expressed as the formula $AGB = a(D^2H)^b$, where D is the breast height (m), H is the tree height (m),

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

and a and b are constants. This model is acknowledged as a fast, simple, and basic method to calculate regional AGB. In our study, we used the AGB, mean H, and mean D of 30 sample plots to constitute the allometric model.

3 Results

3.1 Reference AGB of sample plots

The range of reference AGB of these 30 sample plots was calculated as 1.02~135.79 $Mg \cdot ha^{-1} \cdot plot^{-1}$, with an average value of 47.34 $Mg \cdot ha^{-1} \cdot plot^{-1}$ and a standard deviation of 34.46 $Mg \cdot ha^{-1}$ $plot^{-1}$. The coefficients of variation of the AGB for all the sample plots and for the 10 age categories were calculated as 0.73 and 0.07~0.37, respectively.

3.2 Exploratory data analysis

3.2.1 Selection of variables

Figure 4 shows the correlation-coefficient matrix of variables. The following variables are strongly correlated with AGB: longitude $(r = -0.56)$, diameter at breast height $(r = 0.79)$, tree height $(r = 0.84)$, trunk volume $(r = 0.86)$, timber volume $(r = 0.98)$, and forest age $(r = 0.82)$. The AGB map for the *Eucalyptus* forest in Nanjing is based on the data from the forest resource inventory; therefore, the selected covariates should be accessible from the forest resource inventory dataset. Because the timber volume and stem volume were both estimated based on tree height and diameter at breast height, they were excluded as covariates for the AGB mapping. To summarize, four variables (longitude, diameter at breast height, tree height, and forest age) were selected as covariates for the AGB mapping of the *Eucalyptus* forest in the Nanjing region. Table B.4 lists the statistical descriptions of these covariates and the AGB statistics for the 30 sample plots.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

Biogeosciences Discussions Open Access EGU

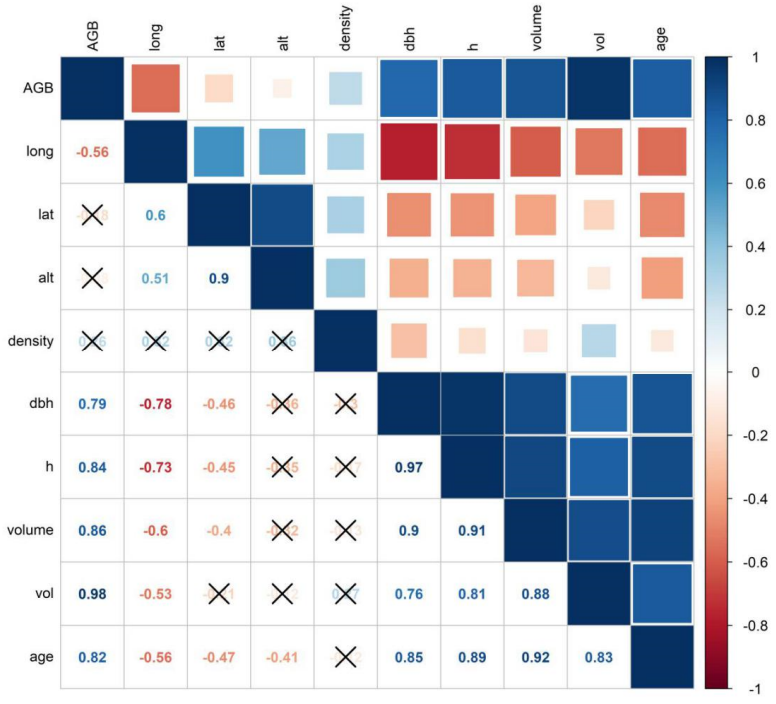


Figure 4. Pearson's correlation coefficients between AGB and other variables represented by numbers and squares. Negative numbers are negatively correlated and are colored in red, while positive blue numbers represent positive correlations. Larger absolute numbers, darker colors, and larger squares all indicate stronger correlation, while × indicates the variables were uncorrelated.

3.2.2 Spatial autocorrelation test

The spatial distribution of the reference AGB of the 30 sample plots shows a pattern of aggregation (see red part in Figure C.1 in the supplementary material and Table 1). In addition, because less than 1% of the AGB data is randomly distributed (see blue part in Figure C.1 and Table 1), the possibility of aggregation distribution is greater than that of random distribution. Furthermore, the null hypothesis is significantly rejected ($p < 0.01$). These results show that the spatial distribution of the AGB data displays aggregation and a pattern of strong spatial autocorrelation.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

294

295 Table 1. Spatial autocorrelation and heterogeneity test

Spatial autocorrelation		**Spatial heterogeneity**		
Items	**Values**	**Factors**	**q-value**	***p*-value**
Moran I	0.36	AGB	0.87	<0.01
		Longitude, long	0.38	<0.01
z-score	4.78	Diameter at breast height, dbh	0.54	<0.01
		Tree height, h	0.63	<0.01
p-value	0.00	Age	0.92	<0.01

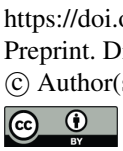
296 **3.2.3 Spatial heterogeneity test**

297 As shown in Table 1, the reference AGB of sample plots can be divided into three strata using K means
298 clustering with a q value of 0.87 and a p value less than 0.01. These results indicate that the within-
299 layer variance is far less than the sum of variances among different strata. The results also show that the
300 reference AGB of 30 sample plots is associated with obvious spatial differentiation.

301 **3.3 Performance of models**

302 We developed seven models for AGB estimation: three machine learning models (SVM, RBF-ANN, and
303 RF), one spatial statistical model (P-BSHADE), and three combined models that integrated each machine
304 learning method with the spatial statistical method (SVM & P-BSHADE, RBF-ANN & P-BSHADE, and
305 RF & P-BSHADE). Furthermore, we used the leave-one-out cross-validation method to split the datasets
306 and evaluated the prediction performance of these seven methods in terms of the indicators of MAE
307 (Figure 5a), MRE (Figure 5b) and RMSE (Figure 5c).

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019



Biogeosciences
Discussions
Open Access
EGU

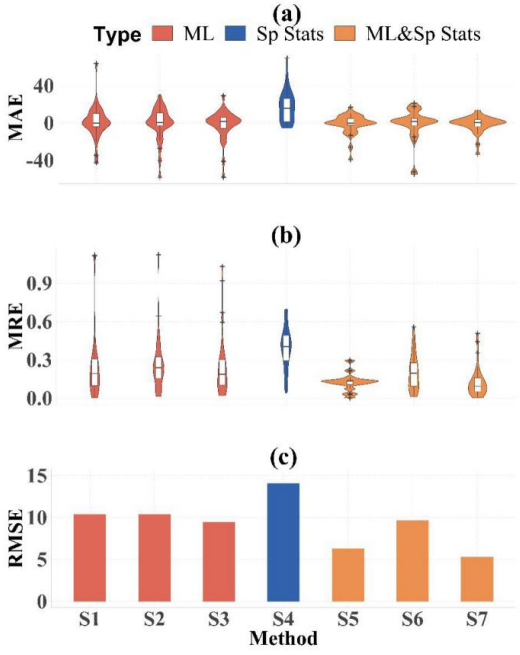


Figure 5. Results of prediction performance of the seven different models. The MAE (a) and MRE (b) presented by boxplots for each prediction method (S1=SVM, S2=RBF-ANN , S3=RF, S4=P-BSHDE , S5=SVM & P-BSHDE, S6=RBF-ANN & P-BSHDE, S7=RF & P-BSHDE, ML=machine learning, Sp Stats=Spatial statistic), with the median (black line in the box), inter-quartile range (25%-75% in the box), the range 5%-95% (whiskers), and outliers (asteroids) labeled. The histogram distributions of RMSE for each prediction method are presented in Figure 5 (c).

Compared with the calculated indicators by the P-BSHADE (MAE=18.37 Mg·ha^{-1}, MRE=39.13%, and RMSE=14.08 Mg·ha^{-1}), the forest AGB estimate obtained by the three machine learning methods has a MAE of 10.16~12.15 Mg·ha^{-1}, a MRE of 24.79~26.69%, and a RMSE of 9.43~10.39 Mg·ha^{-1}, which are substantially smaller than those obtained by the spatial statistical method.

Among the three machine learning methods, the accuracy of RF is the highest and its three evaluation indexes are MAE=10.16 Mg·ha^{-1}, MRE=25.93%, and RMSE=9.43 Mg·ha^{-1}, which are not only substantially smaller than those for P-BSHADE, with MAE=18.37 Mg·ha^{-1}, MRE=39.13%, and RMSE=14.08 Mg·ha^{-1}, but are also smaller than most of those obtained by the other two machine learning methods with MAE=11.17~12.15 Mg·ha^{-1}, MRE=24.79~26.69%, and RMSE=10.39~10.39

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

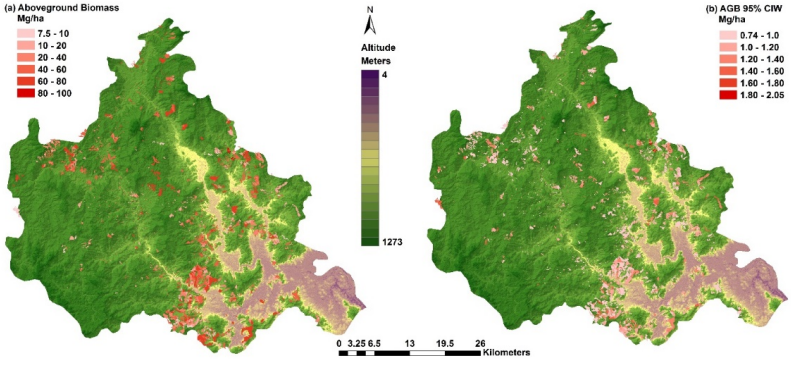
325 Mg·ha^{-1}.

326 Finally, compared with single machine learning methods, the combination of machine learning and
327 spatial statistical models produced smaller MAE (5.68~10.14 Mg·ha^{-1}), MRE (12.47~20.49%), and
328 RMSE (5.30~9.63 Mg·ha^{-1}). In addition, among the three combined methods, the combination of random
329 forest and the spatial statistical model (RF&P-BSHADE) produced a higher accuracy with the smallest
330 MAE (5.68 Mg·ha^{-1}), modest MRE (12.97%), and smallest RMSE (5.30 Mg·ha^{-1}). In contrast, the MAE
331 (10.14 Mg·ha^{-1}), MRE (20.49%), and RMSE (9.63 Mg·ha^{-1}) of RBF-ANN&P-BSHADE were the
332 highest among the three combined methods. Furthermore, compared with the RBF-ANN&P-BSHADE
333 model, the RF&P-BSHADE model achieved a reduction of the cross-validated prediction error of 36.73–
334 44.99% (43.97% for MAE, 36.73% for MRE, and 44.99% for RMSE).

335 **3.4 Model application and mapping of AGB**

336 Figure 6(a) shows the spatial distribution of the AGB predicted by the RF&P-BSHADE. The AGB
337 simulated by RF&P-BSHADE is 7.54~89.93 Mg·ha^{-1}, with an average of 41.21 Mg·ha^{-1}, a median of
338 43.53 Mg·ha^{-1}, a standard deviation of 18.83 Mg·ha^{-1}, and a coefficient of variation of 45.69%. The 95%
339 predictive distribution credible interval width (CIW) was calculated and is mapped for AGB in Figure
340 6(b). Wide CIWs are distributed not only in the high-altitude areas, but also in the low-altitude areas
341 which are easier to access.

342

343 

344 Figure 6. Map of AGB (a) and associated 95% credible interval width (CIW) (b) using RF&P-
345 BSHADE. This map shows two main areas: (1) red, study area of *Eucalyptus* plantations, and (2)
346 green, outside of study area.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

347

348 The total AGB of the Nanjing area (2,980 forest patches) estimated by RF&P-BSHADE is 122,812.1
349 Mg, and that estimated by the allometric model is 123,021.5 Mg. The relative percent difference in total
350 AGB between the two methods is 0.17%.

351 **4 Discussion**

352 **4.1 The significance of the AGB map at the regional scale**

353 In the past, ecologists would often assume that a limited number of sample plots could be used to
354 represent a large range of landscapes, and such sample plots have long served as the main source of
355 information for understanding the spatial distribution of AGB from the sample-plot scale to the regional
356 scale. However, the research by Marvin et al. (2014) confirmed that the distribution of most AGB is non-
357 Gaussian, skewed, or multi-modal, especially in tropical and subtropical regions. Marvin et al. (2014)
358 asserted that the most influential source of uncertainty is the non-representativeness in the sample design
359 in the form of local geographic clusters of sample units. Therefore, AGB maps based on non-
360 representative samples introduce greater uncertainty. For example, in the Amazon basin, fewer than
361 500 geographically concentrated sample plots were used to represent more than 109 hectares of forest,
362 thus undoubtedly contributing to relatively large uncertainty (Mitchard et al., 2014). However,
363 reducing the uncertainty in AGB maps to levels corresponding to high precision would require unrealistic
364 sample sizes; for example, 44 low-lying 1 ha sample plots or more than 85 mountain 1 ha plots are
365 required for every 100 ha on an AGB map (Mitchard et al., 2014). Inevitably, the area represented by
366 these geographically concentrated plots is much less than the total area of the tropical forest represented
367 in the final map. Provided that the limited sample size cannot represent the spatial heterogeneity of the
368 large-scale area, subsequently, the AGB map cannot lead to reliable quantitative conclusions (Duncanson
369 et al., 2015).
370 To overcome the small sample size and non-representative sample problems which lead to
371 geographically concentrated local plot clusters, we integrated the advantages of machine learning and
372 spatial statistics at a regional scale (the key region linking the sample plots to the landscape scale) to
373 construct an AGB map for a subtropical region. The approach provides not only a low-cost, high-
374 precision map of AGB whose estimates can be compared with those obtained from remote sensing,

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

ground observation, and model simulation, but also a scientific basis to assist forest-management decisions (e.g., the quantitative evaluation of carbon emissions from deforestation). Combining the advantages of machine-learning-based quantification of AGB and the complex nonlinear relationship between multiple environmental covariates, in conjunction with the proposed P-BSHADE model, the spatial correlation and heterogeneity of multiple environmental covariates are incorporated into the model, and the sample points are subsequently rectified, thus leading to the best linear unbiased estimate (BLUE) of the target site. Given that current multi-source databases cannot provide high-precision accuracy of mapping affected by the variations of AGB in subtropical areas, especially in regions with large variability, current studies mainly use fusion maps composed of different and independent data sets (Avitabile et al., 2015). Therefore, we provide the most accurate AGB map by data fusion of single analytic trees and forest resource inventory data which may be used to extrapolate AGB from the tree scale to the field and regional scales.

4.2 Benefits of random forest in predicting an AGB map

This study shows that among the three machine learning methods, the prediction accuracy of random forest in AGB mapping is the highest. This is consistent with the results from Gleason and Im (2012) and Fassnacht et al. (2014). For example, Fassnacht et al. (2014) combined lidar with multiple remote-sensing data, such as airborne hyperspectral data from Karlsruhe, Germany, to compare the AGB prediction accuracy of five machine learning methods: stepwise regression, support vector machine, random forest, Gaussian processes, and K-nearest neighbor. The evaluation indexes for leave-one-out cross validation (i.e., R^2 and RMSE) showed that the random forest method was associated with the highest prediction accuracy due to self-learning techniques of the random forest method. The random forest method clearly differs from the other machine learning methods in the flexibility of its conceptual design and method. In detail, the following advantages of random forest method may help improve the precision of predicting an AGB map (Breiman, 2001): (1) The random forest method can generate highly accurate classifiers, detect the interaction between variables, and also detect outliers and monitor data; (2) For unbalanced and categorized data sets, the random forest method can balance the deviations; (3) The random forest method can be extended to unlabeled data, which usually use unsupervised clustering; (4) In the construction of a forest, the random forest method can internally produce unbiased estimates for generalized deviations; (5) The random forest method contains a good way to estimate missing data.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

In addition, if a large part of the data is missing, the random forest method can still maintain accuracy.

4.3 Machine learning outperforms the spatial statistical model in prediction performance

Regarding the AGB mapping of non-representative sample units, the machine learning methods outperformed the spatial statistical method (P-BSHADE) in the prediction accuracy. This may be because machine learning offers an array of supervised learning models capable of relating forest AGB to multi-variables including forest variables and environmental variables via complex, potentially nonlinear functional relationships. Machine learning models appear to be good at 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 (Görgens et al., 2015;Latifi et al., 2010;Stojanova et al., 2010). In addition, the P-BSHADE model yielded negative weights between a small number of patches which might introduce slight uncertainty into the result (Xu et al., 2013). Our results were consistent with the study of Povak et al. (2014) and Li et al. (2011), who found that a machine learning method (RF) outperformed the spatial statistical method (e.g., Geographically Weighted Regression, Inverse Distance Weighting) in terms of prediction accuracy.

4.4 Why a combined model outperforms a single machine learning or spatial statistical model

As expected, the prediction accuracy of the combined methods is higher than that of any single method (either a machine learning or a spatial statistical). In the previous sections, we described how the advantages of the P-BSHADE model can compensate for the inherent defects of machine learning. Virtually, the P-BSHADE model is also handicapped by the fact that the founding assumption does not conform to reality. The assumption is that the AGB is accurate in all other sampling plots except at this target sampling plot. In reality, each sampling plot has a varying degree of AGB uncertainty. In other words, the premise behind only using the P-BSHADE model is that the reference AGB data is accurate. Since the P-BSHADE model combined with machine learning uses the results optimized by machine-learning as the reference values; therefore, it further improves the accuracy of AGB mapping. Machine learning methods or the P-BSHADE model have been adopted to model the uncertainty of temperature observation obtained by weather stations (Fassnacht et al., 2014;Paul et al., 2016;Xu et al., 2013). However, methods in these studies were adopted independently. Conversely, the combination of machine learning and spatial statistics can improve the prediction accuracy of AGB maps, which in turn can be

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

used as criteria for improving the accuracy of lidar remote-sensing technology and the results of ecological-process models. Eventually, these achievements can promote process-oriented projects of dynamic AGB predictions for large-scale forests in different forest-management scenarios.

In addition, we compared the prediction accuracy of AGB mapping obtained by the combined spatial statistical and machine learning models with that reported by recent local and international research into AGB mapping. In the current literature on remote-sensing estimation of forest AGB, RMSE and R^2 were commonly used as indexes for evaluating prediction performance when these studies looked at the importance of research sample size, data types, and forecasting methods (Fassnacht et al., 2014). In contrast, our study uses three conventional indexes for evaluating prediction performance: RMSE, MAE, and MRE. Because the main goal of this work is to predict regional forest AGB based on a small number of non-representative sample units, the criterion of model selection is to choose indexes summarized from sample prediction (such as RMSE), rather than choosing the goodness-of-fit R^2 (Babcock et al., 2015). Based on calculated RMSE indexes, the AGB prediction accuracy of the combined random forest and P-BSHADE method (5.30 $Mg \cdot ha^{-1}$) is higher than that obtained by Babcock et al. (2015)(34.21 $Mg \cdot ha^{-1}$) in Colorado, USA, where the authors used a combination of airborne lidar, forest inventory database, and a Bayesian spatial hierarchical framework model and introduced spatial random effects to compensate for the residual spatial dependence and nonstationarity of model covariates. In addition, prediction accuracy of AGB in this work is also higher than that obtained by (Ene et al., 2016)(RMSE=15.92 $Mg \cdot ha^{-1}$) in southeast Norway using a general linear regression model with airborne lidar and ground survey. Furthermore, the prediction accuracy of AGB in this work is also higher than those obtained by (Avitabile et al., 2015) in the tropics (Central America: 22.8±0.3 $Mg \cdot ha^{-1}$; Africa: 83.7±2.5 $Mg \cdot ha^{-1}$) using fusion maps of multi-source databases combined with the random forest method. Our prediction performance is close to that obtained by Marvin et al. (2014) (6 Mg $C \cdot ha^{-1}$) who studied the Amazon tropical forests using a Monte Carlo method based on airborne lidar in conjunction with on-site monitoring. Because RMSE is an absolute measure of the deviation between the predicted and the observed data, a large range of reference values may cause large deviations. With our combined methods, the calculated RMSE for the prediction accuracy of AGB is relatively small, which we attribute to the following reasons: (1) The reference AGB of 30 sample plots were calculated from each tree by the allometric model constructed with 90 most accurate analytic trees. There were no differences in the range of reference values. (2) Machine learning methods were used to quantify the complex nonlinear

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

462 relationship between AGB and multiple environmental covariates. (3) We applied a statistical method
463 based on the hypothesis of spatial heterogeneity. Although the RMSE index was calculated by different
464 studies using different datasets and prediction methods in different locations, most studies deemed that
465 RMSE was the most commonly used indicator for measuring the prediction errors of remote-sensing
466 AGB models and calculating the real AGB of forest sample plots. In contrast to other studies, our work
467 reflects not only our attention to subtropical forests, but also the methodological differences in
468 uncertainty mitigation, especially in comprehensively addressing the sources of uncertainty caused by
469 multiple spatial and environmental covariates.

470 **4.5 Comparison of RF&P-BSHADE with the allometric growth model**

471 Because the allometric growth model can offer a fast and simple calculation method, it has been used as
472 the basis for determining the benchmark map in quite a few studies. Nevertheless, spatial heterogeneity
473 caused by multiple environmental covariates is not considered in the allometric model, as there may be
474 errors in the AGB estimate and the errors may be propagated to affect the accuracy of the regional AGB
475 benchmark map. This study shows that the relative percent difference in total AGB between RF&P-
476 BSHADE and the allometric method is 0.17%. Meanwhile, the MRE of AGB between the two methods
477 ranged from 0.04% to 99.8% with an average of 19.93%. These results confirm that the RF&P-BSHADE
478 estimates can be used as the main reference for regional-scale forest AGB maps. Furthermore, it also
479 shows that the two methods are roughly the same in terms of overall estimates of AGB, but the local
480 spatial distribution of AGB is different. The differences in AGB spatial distribution have been reported
481 in many studies of AGB maps. Babcock et al. (2015) asserted that the main reasons for the differences
482 in the spatial distribution of AGB maps between different methods include the following: (1) The
483 structural framework of different research methods and schemes cannot truly reflect the actual situation
484 of the forest growth. (2) The model is usually a simplification of an ecological process and ignores the
485 spatial heterogeneity at the regional scale. (3) The model does not consider the influence of multiple
486 environmental covariates (vegetation, topography, etc.) on forest growth in the region.

487 **4.6 Implications for AGB mapping and future research directions**

488 Based on the results of this study, we have the following two implications. First, to enhance the prediction
489 accuracy of large-scale AGB mapping, we should not only reduce the effect of sampling uncertainty by

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

improving the sampling method (by data treatment such as quantification of sampling errors and spatial filtering of existing data sets), but also solve the problems of nonlinearity, complexity, and spatial heterogeneity from the perspective of both model and algorithm. Second, in all probability, the sampling plots for the real values on the ground are only accessible in small sampling areas within non-representative locations. Therefore, the combined use of spatial-differentiation-based statistical analysis and machine learning with nonlinear fitting should improve the prediction accuracy of AGB mapping. Additionally, more machine learning methods (such as KNN algorithms) can be tried and combined with P-BSHADE in future research to explore the best AGB mapping methods for large-scale forests. The case we present herein is only for a pure *Eucalyptus* forest, and further research can create separate databases for different forest types in a complex tropical forest system to create a hierarchical mapping. If the identification of plant species is also included in field plot-based AGB assessment and monitoring, such identification information can also provide important information about changes in species composition. Overall, forest AGB mapping should not be static. Instead, it should be generated based on time sequences using an ecological-process model, so as to capture the changes in the AGB map database over time(Bustamante et al., 2016). In addition, more environmental and socio-economic datasets (for example, the meteorological variables that are missing in the present study) should be included and the correlation between them should be taken into account in the future work.

5 Conclusion

Currently, extrapolations and predictions based on sparse and/or non-randomly distributed forest plots cannot solve the problem of regional carbon balance in tropical forests. With the continuous development of remote sensing, ground observation, and methods of ecological-process modeling, the number of global and regional AGB datasets is continuously increasing. As criteria to judge the differences between different estimates of biomass, an AGB map not only provides a decision-making basis for forest managers to mitigate the negative impact of climate change, but also helps different countries evaluate and implement the policies and programs that aim at reducing regional-scale deforestation and forest degradation, so as to avoid more carbon emissions.

Given the conditions of insufficient sample size and non-representative sample units that lead to geographic clusters of localities, we propose a method to integrate the advantages of machine learning

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

and spatial statistics, different datasets, and multiple environmental covariates, to solve the problem of uncertainty in regional AGB maps. Based on the most accurate data for single analytic trees and forest resource inventory data, we extrapolate the study from the single-tree to the regional scale. In this study, although the forest resource inventory data and the data of analytic-trees are solely available for *Eucalyptus* forests located in the Nanjing area of China, the proposed method and the findings can provide references for AGB remote sensing and simulation of ecological processes in different countries and in different types of tropical forests.

Acknowledgements

Shaoqing Dai and Xiaoman Zheng contributed equally to this work and should be considered as co-lead authors. This work was supported by National Science Foundation of China (31670645 and 31470578), the National Key Research Program of China (2016YFC0502704), National Social Science Fund (17ZDA058), Fujian Provincial Department of S&T Project (2016T3032, 2016T3037, 2016Y0083, 2018T3018), Key Laboratory of Urban Environment and Health of CAS (KLUEH-C-201701) and Key Program of the Chinese Academy of Sciences (KFZDSW-324). We are grateful to Professor Li Hu for his helpful suggestions.

References

Andersen, H.-E., Reutebuch, S. E., McGaughey, R. J., d'Oliveira, M. V. N., and Keller, M.: Monitoring selective logging in western Amazonia with repeat lidar flights, Remote Sensing of Environment, 151, 157-165, 10.1016/j.rse.2013.08.049, 2014.

Asner, G. P., Clark, J. K., Mascaro, J., and García, G. A. G.: High-resolution Mapping of Forest Carbon Stocks in the Colombian Amazon, Biogeosciences, 9, 2683, 2012.

Avitabile, V., Herold, M., Heuvelink, G. B., Lewis, S. L., Phillips, O. L., Asner, G. P., Armston, J., Asthon, P., Banin, L. F., and Bayol, N.: An integrated pan-tropical biomass map using multiple reference datasets, Global Change Biology, 22, 1406-1420, 2015.

Babcock, C., Finley, A. O., Bradford, J. B., Kolka, R., Birdsey, R., and Ryan, M. G.: LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients, Remote Sensing of Environment, 169, 113-127, 2015.

Benitez, F. L., Anderson, L. O., and Formaggio, A. R.: Evaluation of geostatistical techniques to estimate the spatial distribution of aboveground biomass in the Amazon rainforest using high-resolution remote sensing data, Acta Amazonica, 46, 151-160, 2016.

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P.: Spatial filtering to reduce sampling bias can improve the performance of ecological niche models, Ecological Modelling, 275, 73-77, 2014.

Breiman, L.: Random forests, Machine Learning, 45, 5-32, 2001.

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

551 Bustamante, M. M., Roitman, I., Aide, T. M., Alencar, A., Anderson, L., Aragão, L., Asner, G. P.,
552 Barlow, J., Berenguer, E., and Chambers, J.: Towards an integrated monitoring framework to assess the
553 effects of tropical forest degradation and recovery on carbon stocks and biodiversity, Global Change
554 Biology, 22, 92-109, 2016.
555 Chen, Q.: Modeling aboveground tree woody biomass using national-scale allometric methods and
556 airborne lidar, ISPRS Journal of Photogrammetry and Remote Sensing, 106, 95-106,
557 10.1016/j.isprsjprs.2015.05.007, 2015.
558 Chen, Q., Laurin, G. V., and Valentini, R.: Uncertainty of remotely sensed aboveground biomass
559 over an African tropical forest: Propagating errors from trees to plots to pixels, Remote Sensing of
560 Environment, 160, 134-143, 2015.
561 Chen, Q., McRoberts, R. E., Wang, C., and Radtke, P. J.: Forest aboveground biomass mapping and
562 estimation across multiple spatial scales using model-based inference, Remote Sensing of Environment,
563 184, 350-360, 10.1016/j.rse.2016.07.023, 2016.
564 Cliff, A., and Ord, V. J.: Spatial processes: model and applications, Pion Ltd, London, 1981.
565 Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V.: Support vector regression
566 machines, Proceedings of the 9th International Conference on Neural Information Processing Systems,
567 Denver, Colorado, 1996.
568 Du, H., Zhou, G., Fan, W., Ge, H., Xu, X., Shi, Y., and Fan, W.: Spatial heterogeneity and carbon
569 contribution of aboveground biomass of moso bamboo by using geostatistical theory, Plant Ecology, 207,
570 131-139, 2010.
571 Duncanson, L., Rourke, O., and Dubayah, R.: Small sample sizes yield biased allometric equations
572 in temperate forests, Scientific Reports, 5, 13, 2015.
573 Elanayar, V. T. S., and Shin, Y. C.: Radial basis function neural network for approximation and
574 estimation of nonlinear stochastic dynamic systems, IEEE Transactions on Neural Networks, 5, 594-603,
575 1994.
576 Ene, L. T., Næsset, E., and Gobakken, T.: Simulation-based assessment of sampling strategies for
577 large-area biomass estimation using wall-to-wall and partial coverage airborne laser scanning surveys,
578 Remote Sensing of Environment, 176, 328-340, 2016.
579 Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., and Koch, B.:
580 Importance of sample size, data type and prediction method for remote sensing-based estimations of
581 aboveground forest biomass, Remote Sensing of Environment, 154, 102-114, 2014.
582 Frey, U. J., Klein, M., and Deissenroth, M.: Modelling complex investment decisions in Germany
583 for renewables with different machine learning algorithms, Environmental Modelling & Software, 118,
584 61-75, https://doi.org/10.1016/j.envsoft.2019.03.006, 2019.
585 Galante, P. J., Alade, B., Muscarella, R., Jansa, S. A., Goodman, S. M., and Anderson, R. P.: The
586 challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model
587 complexity, Ecography, 001-010, 2017.
588 Gao, L., and Hailu, A.: Ranking management strategies with complex outcomes: An AHP-fuzzy
589 evaluation of recreational fishing using an integrated agent-based model of a coral reef ecosystem,
590 Environmental Modelling & Software, 31, 3-18, https://doi.org/10.1016/j.envsoft.2011.12.002, 2012.
591 Gao, L., Bryan, B. A., Nolan, M., Connor, J. D., Song, X., and Zhao, G.: Robust global sensitivity
592 analysis under deep uncertainty via scenario analysis, Environmental modelling & software, 76, 154-166,
593 2016.
594 Gleason, C. J., and Im, J.: Forest biomass estimation from airborne LiDAR data using machine

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

595 learning approaches, Remote Sensing of Environment, 125, 80-91, 2012.
596 Görgens, E. B., Montaghi, A., and Rodriguez, L. C. E.: A performance comparison of machine
597 learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics,
598 Computers and Electronics in Agriculture, 116, 221-227, https://doi.org/10.1016/j.compag.2015.07.004,
599 2015.
600 Gregoire, T. G., Næsset, E., Mcroberts, R. E., Ståhl, G., Andersen, H. E., Gobakken, T., Ene, L., and
601 Nelson, R.: Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass, Remote
602 Sensing of Environment, 173, 98-108, 2016.
603 He, C., Tian, J., Shi, P., and Hu, D.: Simulation of the spatial stress due to urban expansion on the
604 wetlands in Beijing, China using a GIS-based assessment model, Landscape and Urban Planning, 101,
605 269-277, https://doi.org/10.1016/j.landurbplan.2011.02.032, 2011.
606 Houghton, R. A., Hall, F., and Goetz, S. J.: Importance of biomass in the global carbon cycle, Journal
607 of Geophysical Research Biogeosciences, 114, G00E03, 2009.
608 Hu, M. G., Wang, J. F., Zhao, Y., and Jia, L.: A B-SHADE based best linear unbiased estimation
609 tool for biased samples, Environmental Modelling & Software, 48, 93-97, 2013.
610 Latifi, H., Nothdurft, A., and Koch, B.: Non-parametric prediction and mapping of standing timber
611 volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors,
612 Forestry: An International Journal of Forest Research, 83, 395-407, 10.1093/forestry/cpq022, 2010.
613 Li, J., Heap, A. D., Potter, A., and Daniell, J. J.: Application of machine learning methods to spatial
614 interpolation of environmental variables, Environmental Modelling & Software, 26, 1647-1659,
615 https://doi.org/10.1016/j.envsoft.2011.07.004, 2011.
616 Lu, Z., Lin, F., and Ying, H.: DESIGN OF DECISION TREE VIA KERNELIZED
617 HIERARCHICAL CLUSTERING FOR MULTICLASS SUPPORT VECTOR MACHINES,
618 Cybernetics and Systems, 38, 187-202, 10.1080/01969720601139058, 2007.
619 Marvin, D. C., Asner, G. P., Knapp, D. E., Anderson, C. B., Martin, R. E., Sinca, F., and Tupayachi,
620 R.: Amazonian landscapes and the bias in field studies of forest structure and biomass, Proceedings of
621 the National Academy of Sciences of the United States of America, 111, 5224-5232, 2014.
622 McRoberts, R. E., Chen, Q., Domke, G. M., Ståhl, G., Saarela, S., and Westfall, J. A.: Hybrid
623 estimators for mean aboveground carbon per unit area, Forest Ecology and Management, 378, 44-56,
624 10.1016/j.foreco.2016.07.007, 2016.
625 McRoberts, R. E., Chen, Q., Gormanson, D. D., and Walters, B. F.: The shelf-life of airborne laser
626 scanning data for enhancing forest inventory inferences, Remote Sensing of Environment, 206, 254-259,
627 10.1016/j.rse.2017.12.017, 2018.
628 Mendoza-Ponce, A., and Galicia, L.: Aboveground and belowground biomass and carbon pools in
629 highland temperate forest landscape in Central Mexico, Forestry: An International Journal of Forest
630 Research, 83, 497-506, 10.1093/forestry/cpq032, 2010.
631 Mitchard, E. T. A., Feldpausch, T. R., Brienen, R. J. W., Lopez-Gonzalez, G., Monteagudo, A., Baker,
632 T. R., Lewis, S. L., Lloyd, J., Quesada, C. A., Gloor, M., Ter Steege, H., Meir, P., Alvarez, E., Araujo-
633 Murakami, A., Aragão, L. E. O. C., Arroyo, L., Aymard, G., Banki, O., Bonal, D., Brown, S., Brown, F.
634 I., Cerón, C. E., Chama Moscoso, V., Chave, J., Comiskey, J. A., Cornejo, F., Corrales Medina, M., Da
635 Costa, L., Costa, F. R. C., Di Fiore, A., Domingues, T. F., Erwin, T. L., Frederickson, T., Higuchi, N.,
636 Honorio Coronado, E. N., Killeen, T. J., Laurance, W. F., Levis, C., Magnusson, W. E., Marimon, B. S.,
637 Marimon Junior, B. H., Mendoza Polo, I., Mishra, P., Nascimento, M. T., Neill, D., Núñez Vargas, M. P.,
638 Palacios, W. A., Parada, A., Pardo Molina, G., Peña-Claros, M., Pitman, N., Peres, C. A., Poorter, L.,

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

639 Prieto, A., Ramirez-Angulo, H., Restrepo Correa, Z., Roopsind, A., Roucoux, K. H., Rudas, A., Salomão,
640 R. P., Schietti, J., Silveira, M., de Souza, P. F., Steininger, M. K., Stropp, J., Terborgh, J., Thomas, R.,
641 Toledo, M., Torres-Lezama, A., van Andel, T. R., van der Heijden, G. M. F., Vieira, I. C. G., Vieira, S.,
642 Vilanova-Torre, E., Vos, V. A., Wang, O., Zartman, C. E., Malhi, Y., and Phillips, O. L.: Markedly
643 divergent estimates of Amazon forest carbon density from ground plots and satellites, Global Ecology
644 and Biogeography 23, 935-946, 10.1111/geb.12168, 2014.
645 Morel, A. C., Fisher, J. B., and Malhi, Y.: Evaluating the potential to monitor aboveground biomass
646 in forest and oil palm in Sabah, Malaysia, for 2000–2008 with Landsat ETM+ and ALOS-PALSAR,
647 International Journal of Remote Sensing, 33, 3614-3639, 2012.
648 Paul, K. I., Roxburgh, S. H., Chave, J., England, J. R., Zerihun, A., Specht, A., Lewis, T., Bennett,
649 L. T., Baker, T. G., Adams, M. A., Huxtable, D., Montagu, K. D., Falster, D. S., Feller, M., Sochacki, S.,
650 Ritson, P., Bastin, G., Bartle, J., Wildy, D., Hobbs, T., Larmour, J., Waterworth, R., Stewart, H. T., Jonson,
651 J., Forrester, D. I., Applegate, G., Mendham, D., Bradford, M., O'Grady, A., Green, D., Sudmeyer, R.,
652 Rance, S. J., Turner, J., Barton, C., Wenk, E. H., Grove, T., Attiwill, P. M., Pinkard, E., Butler, D.,
653 Brooksbank, K., Spencer, B., Snowdon, P., O'Brien, N., Battaglia, M., Cameron, D. M., Hamilton, S.,
654 McAuthur, G., and Sinclair, J.: Testing the generality of above-ground biomass allometry across plant
655 functional types at the continent scale, Global Chang Biology, 22, 2106-2124, 10.1111/gcb.13201, 2016.
656 Pelletier, J., Ramankutty, N., and Potvin, C.: Diagnosing the uncertainty and detectability of
657 emission reductions for REDD + under current capabilities: an example for Panama, Environmental
658 Research Letters, 6, 024005, 2011.
659 Povak, N. A., Hessburg, P. F., McDonnell, T. C., Reynolds, K. M., Sullivan, T. J., Salter, R. B., and
660 Cosby, B. J.: Machine learning and linear regression models to predict catchment-level base cation
661 weathering rates across the southern Appalachian Mountain region, USA, Water Resources Research, 50,
662 2798-2814, 10.1002/2013WR014203, 2014.
663 Propastin, P.: Modifying geographically weighted regression for estimating aboveground biomass
664 in tropical rainforests by multispectral remote sensing data, International Journal of Applied Earth
665 Observation and Geoinformation, 18, 82-90, 2012.
666 Rangel, T. F., and Bini, L. M.: SAM: A comprehensive application for Spatial Analysis in
667 Macroecology, Ecography, 33, 46-50, 2010.
668 Ren, Y., Zhang, C., Zuo, S., and Li, Z.: Scaling up of biomass simulation for *Eucalyptus* plantations
669 based on landsenses ecology, International Journal of Sustainable Development & World Ecology, 24,
670 135-148, 2017.
671 Rosenberg, M. S., and Anderson, C. D.: PASSaGE: Pattern Analysis, Spatial Statistics and
672 Geographic Exegesis. Version 2, Methods in Ecology and Evolution, 2, 229-232, 2011.
673 Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R.,
674 Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A.: Benchmark
675 map of forest carbon stocks in tropical regions across three continents, Proceedings of the National
676 Academy of Sciences of the United States of America, 108, 9899-9904, 10.1073/pnas.1019576108, 2011.
677 Schabenberger, O., and Gotway, C. A.: Statistical methods for spatial data analysis, Chapman &
678 Hall0CRC, Boca Raton, 2005.
679 Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A., and Džeroski, S.: Estimating vegetation height
680 and canopy cover from remotely sensed data with machine learning, Ecological Informatics, 5, 256-266,
681 https://doi.org/10.1016/j.ecoinf.2010.03.004, 2010.
682 Van der Laan, C., Verweij, P. A., Quiñones, M. J., and Faaij, A. P.: Analysis of biophysical and

https://doi.org/10.5194/bg-2019-202
Preprint. Discussion started: 11 June 2019

683 anthropogenic variables and their relation to the regional spatial variation of aboveground biomass
684 illustrated for North and East Kalimantan, Borneo, Carbon Balance and Management, 9, 8, 2014.
685 Viana, H., Aranha, J., Lopes, D., and Cohen, W. B.: Estimation of crown biomass of Pinus pinaster
686 stands and shrubland above-ground biomass using forest inventory data, remotely sensed imagery and
687 spatial prediction models, Ecological Modelling, 226, 22-35, 2012.
688 Wang, J. F., Li, X. H., Christakos, G., Liao, Y. L., Zhang, T., Gu, X., and Zheng, X. Y.: Geographical
689 Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the
690 Heshun Region, China, International Journal of Geographical Information Science, 24, 107-127, 2010.
691 Xu, C. D., Wang, J. F., Hu, M. G., and Li, Q. X.: Interpolation of Missing Temperature Data at
692 Meteorological Stations Using P-BSHADE*, Journal of Climate, 26, 7452-7463, 2013.
693 Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Moine, J. L., and Ryu, S. R.: Estimating
694 aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin,
695 USA, Remote Sensing of Environment, 93, 402-411, 2004.
696
697