TOWARDS GLOBAL FOREST BIOMASS ESTIMATORS FROM TREE HEIGHT DATA

Qian Song [1], *Conrad M Albrecht* [1], *Zhitong Xiong* [2], *Xiao Xiang Zhu* [1,2]

[1] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[2] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany

ABSTRACT

In order to estimate tree biomass, allometric equations take tree parameters such as tree height, wood density, circumference of trunk, and crown diameter as input parameters. Given that most of these quantities are challenging to be extracted from remote sensing data, we evaluate the option to approximate biomass by tree height only. We study our approach by evaluating linear regression, random forest, and Gaussian process regressor models when applied to the 2016 Jucker dataset. Results indicate that linear models fail to properly capture the relationship between biomass and tree height, but the Gaussian process regressor outperms the other two candidate models.

Index Terms— Tree biomass estimation, allometric equation, random forest models, Gaussian process regression

1. INTRODUCTION

Forest covers nearly a third of the land on Earth. As a biophysical parameter, biomass is key to tracking tree growth, and quantifying the health of corresponding ecosystems. Forest above-ground biomass—AGB, also simply referred to as biomass here—is defined as the total dry weight of the above-ground parts of trees in a forest. For accurate quantification of forest biomass, it is required to harvest, dry, and weight all trees [1]. Obviously, such an approach is cost inefficient, and contradicts effort of preserving the plant's biosphere.

As an alternative, tree parameters such as wood density, diameter at breast height, height, etc. may serve as estimators of tree biomass through input to allometric equations [2, 3, 4, 5]. In trees, the trunk accounts for a major fraction of the total biomass. Approximating the stem as a cone, biomass is related to average wood density, tree height, and tree diameter (at breast height). Consequently, the AGB of a tree is considered as a function of such parameters. This relation is specified empirically. For example, linear log–log regression models are commonly employed [4, 6] with model parameters estimated by training data. Allometric equations may vary in space, per tree species, tree age, etc. The *GlobAllomeTree* platform [7] makes available some of such results. Luo et al. listed the state-of-the-art allometric equations for China [8], specifically.

To evaluate biomass on a large scale, manually measuring the tree's parameters is out of reach in practice. High-resolution remote sensing data may serve to automatically estimate tree parameters [9, 10, 6]. Subsequently, biomass is derived from given allometric equations. However, some parameters such as stem diameter and wood density are difficult to be estimated from remotely sensed data. Nevertheless, LiDAR surveys have the ability to accuratly determine tree height. Instead of applying separate allometric equations for different tree species or geographic location, global biomass quantification calls for generic models. Jucker et al. [6] demonstrated estimation of crown diameter and tree height from LiDAR surveys to quantify above-ground biomass by two equations. Since estimation of crown diameter is a challenge for dense forests [11], this paper considers biomass quantification from tree height approximation.

In the following we employ three regression models, namely: linear regression (LR), random forest (RF), and a Gaussian process regressor (GPR) to compute biomass from tree height information. For model training, we utilize the 2016 Jucker [6] dataset building a global model applicaple to all biome types. Our focus targets on model error quantification to determine which generalizes best.

2. MODELS

2.1. Linear Regression

Linear regression models are widely employed to capture the relationship of tree biomass and tree parameters on a logarithmic scale (base 10):

$$\log B = a \log H + b + \epsilon, \tag{1}$$

with a and b are coefficient and bias specified by training data. B and H denote biomass and tree height, respectively, and ϵ the model residuals.

2.2. Random Forest

Random forest (RF) is one of the most widely used regressors for biomass estimation [12, 13]. It is an ensemble leaning technique training several decision tree models in parallel. RF randomly selects training data into sub-samples with replacement. Each decision tree is trained for regression on one of those subsets. In inference, RF outputs an average of predictions provided by the set of decision trees. Random forest outperforms any of the individual decision trees, and it minimizes model over-fitting.

2.3. Gaussian Process Regressor

A Gaussian processes [14] is defined as

$$f(\boldsymbol{x}) \sim \mathcal{N}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}_*)) \tag{2}$$

with $\boldsymbol{x}$ input data. Mean and covariance is given by

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})], \tag{3}$$

$$k(\boldsymbol{x}, \boldsymbol{x}_*) = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}_*) - m(\boldsymbol{x}_*))] \quad . \tag{4}$$

The mean and variance at test point can be derived as

$$\mu_{y_*} = m(\boldsymbol{x}_*) + \boldsymbol{K}_*^{\mathrm{T}}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - m(\boldsymbol{X})), \tag{5}$$

$$\sigma^2_{y_*} = \boldsymbol{K}_{**} - \boldsymbol{K}_*{}^{\mathrm{T}}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}_*, \tag{6}$$

with $\boldsymbol{K}_* = k(\boldsymbol{x}, \boldsymbol{x}^*)$, $\boldsymbol{K} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{K}_{**} = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$, σ is the noise level of input data, and $\boldsymbol{I}$ is an identity matrix. The covariance matrix $\boldsymbol{K}$ can be defined by a learnable kernel function $k(\cdot, \cdot)$. The kernel function adds prior assumptions on the smoothness of f, *i.e.,* nearby data points would have highly correlated targets. Usually, radial basis function (RBF) kernel is exploited:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{l_i^2}\right\}, \tag{7}$$

where $d = 1$ in our case, and l_i is a hyper-parameter to be optimized. In our case, given the the height and biomass pairs in training set $\boldsymbol{X}$ and $\boldsymbol{y}$, the mean and variance of the estimated biomass corresponding to heights $\boldsymbol{x}_*$ are calculated using equations in Eq. (5)(6) respectively.

3. RESULTS

3.1. Datasets

The dataset assembled by Jucker et al. [6] serves to train and validate the three models. 2,395 globally collected measurements is available in the dataset covering all dominant biome types. For each data point, the tree was harvested to measure its height (in meter), crown diameter (in meter), trunk diameter (in centimeter), and AGB (in kilogram). We exclude data with trunk diameter less than 5cm. In addition we randomly pick 90% of the samples in order to estimate model parameters, leaving the remaining 10% for model testing.

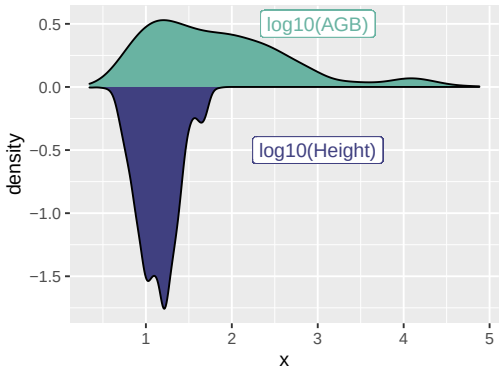

Fig. 1. Histograms of log-scaled biomass (in kilogram) and tree height (in meter) distribution of the 2016 Jucker dataset.

3.2. Evaluation Methods

Three indices, namely: the coefficient of determination (*R squared*, R^2), the *root mean square error* (RMSE), and the bias are used to quantify the accuracy of our candidate models:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{8}$$

$$RMSE^2(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{9}$$

$$Bias(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{y}_i - y_i}{y_i}. \tag{10}$$

where n is the number of the samples, y_i and $\hat{y}_i$ are the ith ground truth and predicted values, respectively. $\bar{y}$ denotes the mean $\sum_{i=1}^{n} y_i/n$.

In our experiments we modified the R-squared computation to exclude outliers: those have significant impact on the score. We define outliers as points whose absolute errors are three times larger than the overall mean absolute error.

3.3. Experimental Results

Table 1 lists R-squared, RMSE, and bias for linear regression, random forest, and Gaussian process regressor. Apparently, the linear model performs worst. The two non-linear models significantly improve performance: with a margin in R^2 of more than 25%. The corresponding RMSE adjusts by a total of about 320 to 350 kilograms. Both, random forest and Gaussian process regressor, exhibit comparable performance.

Figure 2 presents fitted curves and related prediction errors (left column), the scatter plots of predicted vs. observed AGB (column in center), and the histograms of absolute errors (right column) for all three candidate models. We observe that the error interval (blue-shaded area in Figure 2): (a) does not align well with the fitted curve, suggesting that the linear model is not able to properly fit the data. In fact, the LR model

Table 1. Model performance of linear (LR) and non-linear (RF, GPR) models to correlate tree height to tree biomass.

	R2	RMSE (kg)	Bias
LR	0.5326	1466.55	0.2931
RF	0.8039	1147.07	0.2091
GPR	0.8377	1117.9	0.2187

tends to over-estimate the biomass for tree height above 1.7 meters or below 3.0 meters (in dB). The error interval shoots above the fitted curve, and it under-estimates otherwise.

The qualitative improvement of model performance of the non-linear approaches over the linear one reflects in plotting absolute error histograms: We observe a decreased density for close-to-zero errors of fig. 2 (f) relative to (c). The random forest (f) and Gaussian process regressor (i) perform equally well. Indeed, the error intervals of fig. 2 (d) and (g) align well with the fitted curves. However, the fitted curve of the Gaussian process regressor is more smooth—an indication of model robustness.

4. CONCLUSION

In this paper, we compared three models—namely: linear regression, random forest, and Gaussian process regressor—for the above ground biomass estimation from tree height. Results indicate that linear models are unable to properly capture the relationship of biomass and tree height, but a Gaussian process regressor reflect such a correlation well. It seems more robust compared with other non-linear models such as random forests. However, our study is limited to the Jucker dataset. It remains to confirm and test our findings with larger amounts of data in future work. In fact, preliminary results on a dataset covering a continuous patch of forest indicate: model bias significantly decreases on spatially aggregating single tree biomass estimates. Moreover, the trained model should be compared with other allometric equations that has trunk diameter involved as input.

5. REFERENCES

[1] L. Duncanson, J. Armston, M. Disney, et al., "Aboveground woody biomass product validation good practices protocol," 2021.

[2] Y. Malhi, D. Wood, T. R. Baker, et al., "The regional variation of aboveground live biomass in old-growth amazonian forests," *Global Change Biology*, vol. 12, no. 7, pp. 1107–1138, 2006.

[3] Yude Pan, Richard A Birdsey, Jingyun Fang, Richard Houghton, Pekka E Kauppi, Werner A Kurz, Oliver L Phillips, Anatoly Shvidenko, Simon L Lewis, Josep G Canadell, et al., "A large and persistent carbon sink in the world's forests," *Science*, vol. 333, no. 6045, pp. 988–993, 2011.

[4] Jérôme Chave, Maxime Réjou-Méchain, Alberto Búrquez, Emmanuel Chidumayo, Matthew S Colgan, Welington BC Delitti, Alvaro Duque, Tron Eid, Philip M Fearnside, Rosa C Goodman, et al., "Improved allometric models to estimate the aboveground biomass of tropical trees," *Global change biology*, vol. 20, no. 10, pp. 3177–3190, 2014.

[5] Kristina J Anderson-Teixeira, Stuart J Davies, Amy C Bennett, Erika B Gonzalez-Akre, Helene C Muller-Landau, S Joseph Wright, Kamariah Abu Salim, Angélica M Almeyda Zambrano, Alfonso Alonso, Jennifer L Baltzer, et al., "Ctfs-forest geo: a worldwide network monitoring forests in an era of global change," *Global change biology*, vol. 21, no. 2, pp. 528–549, 2015.

[6] Tommaso Jucker, John Caspersen, Jérôme Chave, Cécile Antin, Nicolas Barbier, Frans Bongers, Michele Dalponte, Karin Y van Ewijk, David I Forrester, Matthias Haeni, et al., "Allometric equations for integrating remote sensing imagery into forest monitoring programmes," *Global change biology*, vol. 23, no. 1, pp. 177–190, 2017.

[7] Matieu Henry, Antonio Bombelli, Carlo Trotta, Alfredo Alessandrini, Luca Birigazzi, Gael Sola, Ghislain Vieilledent, Philippe Santenoise, Fleur Longuetaud, Riccardo Valentini, et al., "Globallometree: international platform for tree allometric equations to support volume, biomass and carbon assessment," *Iforest-biogeosciences and forestry*, vol. 6, no. 6, pp. 326, 2013.

[8] Yunjian Luo, Xiaoke Wang, Zhiyun Ouyang, Fei Lu, Liguo Feng, and Jun Tao, "A review of biomass equations for china's tree species," *Earth System Science Data*, vol. 12, no. 1, pp. 21–40, 2020.

[9] Wei Yao, Peter Krzystek, and Marco Heurich, "Tree species classification and estimation of stem volume and dbh based on single tree extraction by exploiting airborne full-waveform lidar data," *Remote Sensing of Environment*, vol. 123, pp. 368–380, 2012.

[10] LI Duncanson, BD Cook, GC Hurtt, and RO Dubayah, "An efficient, multi-layered crown delineation algorithm for mapping individual tree structure across multiple ecosystems," *Remote Sensing of Environment*, vol. 154, pp. 378–386, 2014.

[11] Dimitrios Panagiotidis, Azadeh Abdollahnejad, Peter Surovỳ, and Vasco Chiteculo, "Determining tree height and crown diameter from high-resolution uav imagery,"

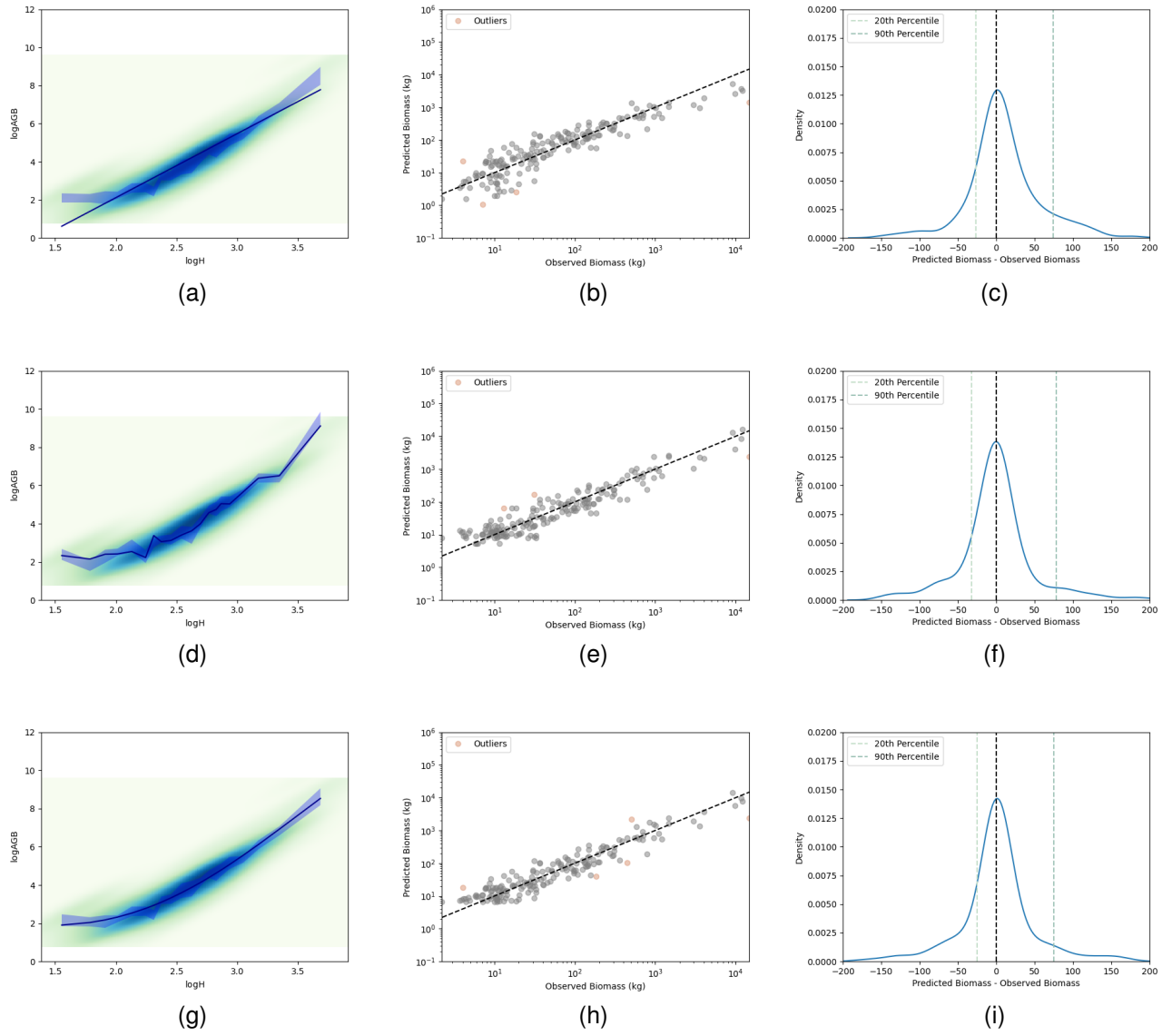

Fig. 2. Plots of tree height–AGB model fits including prediction errors (a,d,g), corresponding scatters of predicted vs. observed biomass (b,e,h), and the distributions errors (c,f,i) training on the 2016 Jucker dataset. Each row corresponds to one of the three models, namely: linear regression (a)–(c), random forest (d)–(f), and Gaussian process regression (g)–(i)). The dashed line in (b,e,h) is the 1:1 line.

International journal of remote sensing, vol. 38, no. 8-10, pp. 2392–2410, 2017.

[12] Masato Hayashi, Takeshi Motohka, and Yoshito Sawada, "Aboveground biomass mapping using alos-2/palsar-2 time-series images for borneo's forest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 5167–5177, 2019.

[13] Pedro Rodríguez-Veiga, Shaun Quegan, Joao Carreiras, Henrik J Persson, Johan ES Fransson, Agata Hoscilo, Dariusz Ziółkowski, Krzysztof Stereńczak, Sandra Lohberger, Matthias Stängel, et al., "Forest biomass retrieval approaches from earth observation in different biomes," *International Journal of Applied Earth Observation and Geoinformation*, vol. 77, pp. 53–68, 2019.

[14] Christopher KI Williams and Carl Edward Rasmussen, "Gaussian processes for regression," 1996.