

## IMPORTANCE OF SAMPLE SIZE, DATA TYPE AND PREDICTION METHOD FOR REMOTE SENSING BASED ESTIMATIONS OF ABOVEGROUND FOREST BIOMASS

<sup>a,f,\*</sup> F.E. Fassnacht, <sup>b</sup> F. Hartig, <sup>c</sup> H. Latifi, <sup>d</sup> C. Berger, <sup>e</sup> J. Hernandez, <sup>e</sup> P. Corvolan, <sup>f</sup> B. Koch

<sup>a</sup> Karlsruhe Institute of Technology, Institute of Geography and Geoeontology, Kaiserstraße 12, 76131 Karlsruhe, Germany,  
Fabian.fassnacht@kit.edu

<sup>b</sup> University of Freiburg, Department of Biometry and Environmental System Analysis, Tennenbacherstrasse 4, 79106 Freiburg,  
Germany, florian.hartig@biom.uni-freiburg.de

<sup>c</sup> University of Wuerzburg, Department of Remote Sensing, Oswald-Kuelpe-Weg 86, 97074 Wuerzburg, Germany,  
hooman.latifi@uni-wuerzburg.de

<sup>d</sup> University of Jena, Department of Earth Observation, Loebdergraben 32, 07743 Jena, Germany, cberger@uni-jena.de

<sup>e</sup> Laboratorio de Geomática y Ecología del Paisaje, Universidad de Chile, Av. Santa Rosa 11315, Santiago de Chile, Chile,  
jhernandez@uchile.cl, pcorvolan@uchile.cl

<sup>f</sup> University of Freiburg, Chair of Remote Sensing and Landscape Information Systems, Tennenbacherstrasse 4, 79106 Freiburg,  
Barbara.koch@felis.uni-freiburg.de

### THEME: Forests, biodiversity and terrestrial ecosystems

**KEY WORDS:** Forest, biomass, hyperspectral, LiDAR, airborne, sample size, prediction method

### ABSTRACT:

Estimates of forest biomass are needed for various technical and scientific applications, ranging from carbon and bioenergy policies to sustainable forest management. As local measurements are costly, there is a great interest in obtaining reliable estimates over large areas from remote sensing data. Currently, such estimates are obtained with a variety of data sources, statistical methods and prediction standards, and there is no agreement on what are best practices for this task.

To improve our understanding of how these different methods affect prediction quality, we first conducted a systematic review of the available literature to identify the most common sensor types and prediction methods. Based on the review, we identified sample size of the reference points on the ground, prediction method (step-wise linear regression, support vector machines, random forest, Gaussian processes and k-nearest neighbor), and sensor type as the main differences that could potentially affect predictive quality. We then compared those factors in two case study areas in Germany and Chile, for which airborne discrete return Light Detection And Ranging (LiDAR) and airborne hyperspectral as well as airborne discrete return LiDAR and spaceborne hyperspectral data were available. For each factor combination, we calculated Pearson's coefficient of correlation between observations and predictions ( $r^2$ ) and root mean squared error (RMSE) for bootstrapped estimates using k-fold cross-validation with a varying number of folds. Finally, Analysis of Variance (ANOVA) was used to quantify the influence of the factors on the predictive error of the biomass models.

Our results confirm previous findings that predictor data (sensor) type is the most important factor for the accuracy of biomass estimates, with LiDAR being preferable to hyperspectral data. In contrast to some previous studies, complementing LiDAR with hyperspectral data did not improve predictive accuracy. Also the prediction method had a substantial effect on accuracy and was generally more important than the sample size. In most cases, random forest performed best and stepwise linear models worst, judging from  $r^2$  and RMSE under cross-validation. Additional results suggested that  $r^2$  may deliver unrealistically large values when the hold-out sample during the cross-validation is too small.

In conclusion, our literature review revealed that different methods for biomass estimation are currently used, with no general agreement on best practices. In our case studies, we found substantial accuracy differences between those methods, with LiDAR data, in combination with a random forest algorithm and a large number of reference sample units on the ground yielding the lowest error for biomass predictions. The comparatively high importance of the statistical prediction method seems particularly relevant, as they suggest that choosing the appropriate statistical method may be more effective than obtaining additional field data for obtaining good biomass estimates.

---

\* Corresponding author.