Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24521

Improved Methodology for Above Ground Biomass Estimation from GEDI - Drone Data Fusion

Bingwei Li

Southwest Jiaotong University, Chengdu, China
davideasden512@gmail.com

Abstract. Accurate assessment of the above-ground biomass density (AGBD) is still necessary for the assessment of carbon stock; however,the current space-borne lidar systems such the NASA Global Ecosystem Dynamics Survey (GEDI) are still subject to a lot of measurement uncertainty. We used the UAV airborne lidar and optical sensor data to obtain the amount of information, computed the volume using computational geometry and SFM, which was then obtained by building a random forest model, in order to evaluate the low accuracyof Gedi in measuring aboveground biomass because of the low precision of GEDI data collection. In comparison to Gedi and MAPE,which were reduced by 40%to 24.7%.

Keywords: GEDI(Global Ecosystem Dynamics Investigation), Random Forest Model, Drone Data Fusion, AGBD (Above Ground Biomass Density), LiDAR Data Processing

1. Background

Ecosystem Dynamics Survey(GEDI)mission attempts to properly estimate forest Above Ground Biomass Density(AGBD)by collecting space lidar data to better understand the function of forests in the carbon cycle and the impact of deforestation on atmospheric carbon dioxide concentration [1]. Currently,there are still significant doubts that can be used to calculate carbon emissions brought on by deforestation and changes in land use primarily attributable to the imprecise estimation of aboveground biomass(AGB).The research makes use of the airborne and spatial lidar data sets that are both spatially and temporally distributed throughout the world. Nonetheless,given the significance of precise RH index selection(which measures lidar energy distribution at different heights),RH98 and RH10 are usually selected as core predictive variables in waveform analysis,the selection of prediction variables is related to the model's performance and complexity [2]. Because of the various vegetation in different places,geographic zoning must be done in order to do so. To determine whether the current spatial stratification scheme is appropriate,we also need to consider the impact on the model performance and the performance in the data sparse region. Four PFTs(EBT,ENT,DBT,and GSW)were used in the investigation,and they divided the world into six regions: Asia,Europe,South America,North America,Africa,and Oceania. The volume and quality of training data exert a substantial influence on model performance, necessitating the refinement of data acquisition strategies to enhance effectiveness in these domains. Strengthening the model's generalization capabilities demands diverse approaches. Developing a more general model structure to reduce the dependence on geographical stratification and balance the simplicity of the model and

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24521

the accuracy of prediction is still a feasible path. Because the prediction results of the quantitative model have errors, it is necessary to evaluate and reduce the impact of errors from different sources on the performance of the model, especially in the high biomass area.

2. Literature review

Through the collection of space-borne lidar data,the NASA Global Ecosystem Dynamics Investigation(GEDI)initiative hopes to precisely quantify the amount of forest aboveground biomass(AGBD),and it hopes to increase our understanding of the contributions of forests to the carbon cycle and the consequences of deforestation on atmospheric carbon dioxide levels. At present, there is great uncertainty in the estimation of carbon emissions caused by land use change and forest loss, mainly due to the uncertainty of aboveground biomass (AGB) estimation.The research makes use of the airborne and spatial lidar data sets that are both spatially and temporally distributed in the world.

However, the selection of prediction variables is related to the performance and complexity of the model, so it is necessary to select the appropriate RH index (the percentage of lidar waveform energy accumulated from a certain height above the ground) as the prediction variables. RH98 and RH10 are usually selected as the prediction variables, RH10 is usually related to low vegetation or ground cover, while RH98 is close to the maximum height of vegetation. RH98 is used as the maximum height index in the study, because it is a more stable height index than RH100, especially when there are sparse canopy or isolated giant trees, RH98 can capture key structural information. The interaction term between these relative humidity indexes (such as RH50xRH98) is also considered as a potential predictor of AGBD. Although the candidate predictor suite is usually correlated, the candidate model with highly correlated predictors is deleted when fitting the model. Low RH index (such as RH10) is sensitive to canopy coverage and terrain, and the difference between simulated waveform and measured waveform affects its reliability, resulting in limited improvement of model performance. For the development of GEDI04_a model, the center of gravity of ALS points classified as ground echo is used to eliminate the uncertainty related to waveform ground search algorithm in AGBD model fitting. The research shows that the mandatory inclusion of RH98 can reduce the complexity of the model, but in some regions (such as DBT Europe and North America), it may increase the deviation (mre>10mg/ha), which has a limited impact on the performance as a whole. By exhausting all possible combinations of RH indicators (1-4 variables and their interactive items), and ranking them in combination with cross validated performance indicators (such as %RMSE and MRE), the optimal combination is finally selected [2].

In view of the different vegetation in different regions, it is necessary to conduct geographical zoning.To evaluate whether the current geographical stratification scheme is optimal, we also need to consider the impact on the model performance and the performance in the data sparse region.The four PFTs(EBT,ENT,DBT,and GSW)were used in the study,which also separated the entire world into six regions(Asia,Europe,South America,North America,Africa,and Oceania).Through geographical cross validation(the training data is divided into subsets according to geographical regions,and the prediction performance in region B is tested after training the model in region a to avoid over fitting caused by spatial autocorrelation),the %RMSE and MRE indexes of different hierarchical models are compared,and the differences of model performance of different hierarchical schemes are evaluated. The performance of the model is significantly increased by layering(PFT x regional area). However, in regions with sparse data (such as Asia), due to insufficient samples (only 326 training points), the cross continental PFT model was forced to be adopted, sacrificing regional specificity to improve generalization ability.

DOI: 10.54254/2755-2721/2025.PO24521

When it comes to how well the model works, there are major differences in the quantity and quality of training data.Therefore,in order to increase the model's performance in these areas is important. There were only 326 samples in EBT Asia and very few AGBD samples(961.8 mg/ha at most)make the model unable to capture the features of high AGBD regions. The RMSE was 78.94%.There is no evidence to say that the model is of high quality. Simultaneously,the AGB estimating error of a few allometric growth models is considerable. Data sparse regions (such as EBT Asia) and high AGBD regions have poor representativeness and performance due to insufficient samples. Due to sparse data, complex canopy structure (meaning that a single RH index is difficult to characterize), high biomass heterogeneity, and large estimation error of allometric growth model for big trees, the model performance is poor. Therefore, it is necessary to increase the field survey and ALS data collection in high AGBD and sparse data areas, and improve the allometric growth model.

Our goal is for the model to have a high level of generalisation ability. Improving the general ability of the model requires a variety of ways,which might improve the geographical migration ability. The research forced the model to be verified in unprecedented geographical regions (geographical cross validation), and screened out the cross regional robust variable combination (such as RH98+RH50). And select the indicators with clear biophysical relationship with AGBD (such as RH98 for height, RH50 for canopy density) to reduce regional dependence. In order to achieve a general model structure, random effects (such as region or PFT as random intercept) can be introduced to hide hierarchical information in the global model, reducing the need for explicit layering. Machine learning methods,such as convolutional neural networks,are used to process waveforms directly [3]. We still need to consider the balance between conciseness and accuracy and choose a reduced model (1-3 variables) to avoid over fitting. At the same time, when the %RMSE increases by<5%, we still choose a reduced model (for example, the %RMSE of GSW Oceania in Table 4 rose from 37.45 to 44.53).

Because the prediction results of the quantitative model have errors, it is necessary to evaluate and reduce the impact of errors from different sources on the performance of the model, especially in the high biomass area.The GEDI04_a product provides the covariance matrix of the model coefficient,which is used to calculate the standard error of the predicted value,and mark whether the predicted value exceeds the range of training data(for example,when AGBD>1578mg/ha,it is marked as high risk). The ratio method is applied to the back transform bias correction [4] to reduce the system bias introduced by the logarithmic or square root transform.

The error sources are wide, including ground data error, waveform simulation error and model error, while the ground data error is from allometric model error (± 20-30%) to training label error and model parameter deviation; The waveform simulation error is from the waveform difference between ALS and GEDI (such as noise and pulse shape) to the RH index deviation and then to the predicted value deviation. The OLS model is written as a random forest

$$h(AGBD) = \sum_{j=1}^{p} B_j f(x_j) + \epsilon \tag{1}$$

$$\widehat{AGBD} = h^{*}\left(\sum_{j=1}^{p} B_j f(x_j) + \epsilon\right) \tag{2}$$

The regression coefficients which is B_j are associated with predictor variables(x_j),with representing $f()$, a transformation function(identity,logarithmic,or square root)and $h()$ denoting a back transformation function(identity,exponential,or quadratic). With a mean of zero,the error term follows a normal distribution. GEDI's hybrid and GHMB estimators posit that Eqn.1 is used to make

DOI: 10.54254/2755-2721/2025.PO24521

the estimation of the true AGBD values, and Eqn. 2 provides an unbiased estimation of transformed AGBD. Consequently ,a bias adjustment factor is included throughout the back transformation process [4] . It is also assumed that the modeling framework is appropriately specified[5], with the error distribution potentially exhibiting skewness in high AGBD regions. Therefore, it is necessary to reduce the uncertainty of high AGBD areas, increase the samples of high AGBD areas, arrange large plots (>1ha) for Sequoia forests, tropical rain forests and other areas, and cooperate with UAV lidar scanning. At the same time, the allometric equation is improved to replace the traditional DBH formula. Fusion of multi-source data to compensate for the penetration limitation of lidar in high-density canopy [2].

3. Data collection

200 sets of data were collected in Sichuan and Hebei Province of China by using satellite lidar acquisition+CMOS sensor photography, 200 sets of data were collected in Southampton, England, and 300 sets of data were collected in inland near Pont de SOR, Portugal. The vegetation mainly investigated includes permanent forests, shrubs and grasslands near forests, perennial nut trees, shrubs and grasslands near permanent agricultural areas.(TLS)The three main categories of equipment are the following, unmanned aerial vehicle(UAV)scanners, aerial systems(ALS),and ground-based systems(TLS)in which the equipment is used for data collection.

In view of the large population differences of the investigated vegetation, for low vegetation such as grassland and shrub, we randomly selected 10 collection points in an area, each of which was 50cm × 50cm, sampled, dried, and measured the biochar content. For the forest, we measured its diameter and height, calculated the volume of trees, calculated the mass according to the tree density, and then calculated the relative weight of branches and leaves and trunk according to the fixed formula.

DOI: 10.54254/2755-2721/2025.PO24521



Figure 1. Sample collection

4. Vegetation classification

In the study area, vegetation classification necessitates the application of parameters derived from specific data sources. Under the assumption that satellite-derived parameters are integrated into the multi temporal dataset with maximum precision, assess if the overall data quality improves. Principal components(PCA)are the first kind of variable that is produced by PCA bands created by the use of statistical methods for linear data transformation. The coordinate system's new principal axis and maximum data variance are calculated by projecting variable values across multidimensional space, and then the new independent variables called principal components are computed using the ENVI5.3 program(Harris Geospatial Solutions,Bedfield,CO,USA) [2]. The information-rich spectral channel was chosen as the initial principal component band by analysing the correlation matrix for variables in the optimal dataset and evaluating the eigenvalues. The second category of variables involves widely applied vegetation indices, which serve to assess plant cover conditions or canopy moisture levels. By referencing data availability and existing research, 18 vegetation indices were derived from the dataset's spectral resolution and processed using ENVI software [6].

DOI: 10.54254/2755-2721/2025.PO24521

4.1. Removing interference factors

Due to the presence of cloud cover and topographic shadows in the imagery, an exclusion mask is required to ensure accurate analysis. The spectra of the blue(band 2)and the near-infrared(band)are used to create this mask 8) channels. The water vapor map(WVM)product is used as one of the auxiliary materials in the development of shadow mask.Because the range of deep shadow is similar to that of water surface,the lake may be covered up. Furthermore, the spatial extent of the research area is defined as regions above 1250 meters elevation, which is obtained from a 1-meter resolution digital elevation model. Using the LAStools software(Rapidlasso,G GmbH,Gilching,Germany),this model was created by processing the ALS-derived georeferenced point clouds [6].Using the built-in features needed to generate and use the mask,the ENVI5.3 platform was used to perform all of the masking procedures(Harris Geospatial Solutions,Bedfield,CO,USA).

5. Data processing

5.1. Lidar

Use LIDAR360 for data preprocessing of ULS [7], denoising and filtering of single tree point clouds.(DEM)the ground points were classified using the improved Progressive TIN Encryption(IPTD)algorithm and the Digital Elevation Model(DEM),and an irregular triangular interpolation was used to produce a resolution of 0.5 m [7]. Finally,in order to eliminate the terrain in the data,the point cloud data was changed. Point clouds over 2 meters are extracted as canopy point clouds, and three sets of indicators (height related, density related, and canopy related) are calculated

5.2. Volume calculation

5.2.1. Convex shell algorithm

Find a set of minimum points that form convex edges on each face, which can include all points in a plane or space. Project the point cloud of a single tree crown onto a plane, calculate the two-dimensional convex hull, and then calculate the area. To calculate the volume of the convex shell and rebuild the surface of the tree crown,use the quickhull algorithm.

5.2.2. Concave shell algorithm

Concave hull algorithm is an algorithm in computational geometry, which aims to find a polygon formed by a group of points. This polygon is called concave hull, which is opposite to convex hull. Convex hull is a convex polygon that surrounds all points and has the smallest angle, while concave hull allows concave parts in the point set, which can describe the shape of the point set in more detail. The concaveman algorithm is applied. Based on the idea proposed by Jin SEO Park and Se-JongOh, the algorithm optimizes the generation of concave shells by introducing a fast k-nearest neighbours query.

5.2.3. Convex shell slicing algorithm

The core idea is integration,which involves slicing the tree crown into equally thick horizontal layers in the vertical direction,with the thickness minimized to treat each layer as a plane.In other

DOI: 10.54254/2755-2721/2025.PO24521

words,each slice of the tree crown can be considered a collection of multiple straight cylinders. For each layer,we compute the point cloud of the plane,project the aggregate of these point clouds onto the plane,and apply a two-dimensional convex hull algorithm to determine the projected area [8]. Subsequently,we calculate the volume of each layer and sum the volumes of all slices to derive the total three-dimensional volume. The formula is as follows:

$$V = \sum \frac{\left(S_n + S_{n+1} + \sqrt{S_n + S_{n+1}}\right)}{3} * \Delta h \tag{3}$$

where S_n is the projected area of the nth layer of point cloud calculated based on the 2D convex packet algorithm and Δh is the height of the table[7].

5.2.4. Voxel algorithm

Voxel algorithm is a computer method based on voxel data. The acronyms for volume pixels are voxel,which is used to depict gridded data in three dimensions [7]. Using rule-based 3D mesh segmentation to discretize canopy point clouds, determine the extremum of the initial mesh in XYZ coordinates, and sum up the space occupied by voxels to obtain the volume of the canopy. The best ratio can be calculated by computing the RMSE after first setting the voxel's edge length between 0.1 and 1 m and a step size of 0.1 m in order to find the ideal ratio.

5.2.5. Voxel coupled convex shell slicing algorithm

The convex shell algorithm considers the tree crown as a whole,ignoring the gaps within the crown and the boundaries of the convex shell are not the true contour of the crown,overestimating the volume of the crown. At the same time, the volume of gaps calculated by the concave shell algorithm is larger than the actual gaps between the crowns. As a result,the height of a single course tree crown point cloud is divided into two sections for calculation using the voxel coupled convex shell slicing algorithm. The upper half is computed using the voxel algorithm,and the lower half is calculated using the layer-by-layer convex shell algorithm. This can prevent errors in upper layer calculations caused by uniform thickness slicing and errors caused by missing point clouds in lower layers. Determine the optimal ratio by adjusting various layering ratios and calculating the RMSE.

5.3. Sensitivity analysis of ULS data density

In order to explore the effect of different LIDAR point cloud densities on 3D green volume extraction, the original density was reduced from 230 points/m² to 75% (172.5 points/m²), 50% (15 points/m²), 25% (57.5 points/m²), 10% (23 points/m²) and 5% (1.5 points/m²) [7]. The stratified sampling strategy is adopted, and the point cloud data is stratified according to elevation, and the secondary sampling is implemented in proportion in each layer. On the basis of preserving the spatial pattern of the original data set, this method ensures the uniformity of sampling intensity at each altitude level.

5.3.1. Structure from Motion (SFM)

In areas with simple vegetation structure, we use the Structure from Motion (SFM) to reduce costs and improve accuracy.

DOI: 10.54254/2755-2721/2025.PO24521

5.3.2. Volume calculation

(CHM)by removing the digital surface models(DTMs)from the digital surface(DSMs),the flow of volume estimation through surface difference analysis entails creating a canopy height model(CHM),which also standardizes ground-level object heights. Forest volume calculations are achieved by multiplying canopy height values of raster cells with their spatial resolution, considering all contiguous pixels representing above-ground vegetation. Furthermore, the mean absolute error(MAE)between the volume estimates and the canopy height model is computed,and the phrase "raster data" is converted to point cloud format.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=n}^{n}\left|y_j - \widehat{y}_j\right| \tag{4}$$

Using the following formula,the volume difference is also given as a%[9]:

$$\text{Percentage difference} = 100^{*}|\mathrm{TLS} - \mathrm{SFM} - \mathrm{MVS}|/((\mathrm{TLS} + \mathrm{SFM} - \mathrm{MVS})/2) \tag{5}$$

In the end,CHM was categorized to only display vegetation that was higher than(to be determined) meters,offering a general approximation of canopy coverage. To analyse the geographical variation in forest volume discrepancies, generate a difference map by subtracting the SFM MVS volumetric data from the TLS-derived volume grid.

6. Building a random forest model

An ensemble learning system that consists of several decision trees,which are used in the whole random forest algorithm. The main part of this method is the individual tree,which is a basic part of the whole performance prediction. By aggregating results from numerous tree structures, the ensemble approach enhances model accuracy. Install the scikit ear package in python,import the Random Forest Classifier library,use the formerly processed data as the dataset,divide the dataset into training and predictionset,settest_size to 0.3,and set random_state to 42.After training with the training set,predict,and compare the prediction results with the predict set.

6.1. Model validation

Determine the coefficient (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE) to evaluate the model performance [10]. Generally speaking, the higher the R^2 value, the lower the RMSE and MAPE indicators. The higher the accuracy of the model, the better the prediction result of the model. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{n}} \tag{7}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \times 100\% \tag{8}$$

DOI: 10.54254/2755-2721/2025.PO24521

The samples that make up the average of actual value, expected value and actual value are $\overline{y_i}$, $\hat{y}_i$ and y_i.

7. Result analysis

The average percentage error (MAPE) of the improved model is set at 24.7%, which is much higher than the mape40% of NASA global ecosystem dynamics Survey (GEDI)

8. Conclusion

In this paper, a hybrid algorithm of Structure from Motion and Voxel coupled convex shell slicing is proposed by using lidar+CMOS data acquisition. To validate the algorithm, The green quantity of permanent forest was calculated by using this method, and the AGB of individual tree was estimated based on 3D green quantity, and compared with the existing methods. The results show that the MAPE of the hybrid algorithm is 24.7%, which is more accurate than that of GEDI. This algorithm can improve the quality of the evaluation results when evaluating aboveground biomass, while reducing the cost.

References

[1] Giles M. Foody, Gintautas Palubinskas, Richard M. Lucas, Paul J. Identifying terrestrial carbon sinks: Classification of successional stages in regenerating tropical forest from Landsat TM data
[2] Laura Duncanson, James R. Kellner .etc. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission
[3] Tong Liu, Yongle Zhang, Lin Qi, Junyu Dong, Mingdong Lv, Qi Wen. WaveNet: learning to predict wave height and period from accelerometer data using convolutional neural network
[4] Snowdon, P., 1991. A ratio estimator for bias correction in logarithmic regressions. Can. J. For. Res. 21, 720-724
[5] Patterson, P.L., Healey, S.P., Ståhl, G., Saarela, S., Holm, S., Andersen, H.-E., Dubayah, R.O., Duncanson, L., Hancock, S., Armston, J., Kellner, J.R., Cohen, W.B., Yang, Z., 2019. Statistical properties of hybrid estimators proposed for GEDI—NASA's globalecosystem dynamics investigation. Environ. Res. Lett.
[6] Martyna Wakulińska; Adriana Marcinkowska-Ochtyra. Multi-Temporal Sentinel-2 Data in Classification of Mountain Vegetation
[7] Lv Zhou, Xuejian Li, Bo Zhang, Jie Xuan, Yulin Gong, Cheng Tan, Huaguo Huang and Huaqiang Du. Estimating 3D Green Volume and Aboveground Biomass of Urban Forest Trees by UAV-Lidar
[8] Remco C. Veltkamp. Generic Geometric Programming in the Computational Geometry Algorithms Library
[9] Angus D. Warfield and Javier X. Leon. Estimating Mangrove Forest Volume Using Terrestrial Laser Scanning and UAV-Derived Structure-from-Motion
[10] Monstafa Mostafaei. Prediction of biodiesel fuel properties from its fatty acids composition using ANFIS approaches