4th URSI AT-RASC, Gran Canaria, 19-24 May 2024

# JOINT MODELING OF SPACEBORNE RADAR AND LIDAR DATA WITH ENSEMBLE LEARNING FOR FOREST ABOVEGROUND BIOMASS ESTIMATION

*Fu-Gen Jiang, Ming-Dian-Li, Si-Wei Chen**

College of Electronic Science and Technology, National University of Defense Technology, Changsha, China
*chenswnudt@163.com

## ABSTRACT

Accurate estimation of forest above-ground biomass (AGB) is fundamental to natural resource management and ecosystem monitoring, which can improve understanding of energy exchange and carbon cycling. Synthetic Aperture Radar (SAR), which is weather-independent and can operate for all-weather, all-day ground observation, is uniquely suited for forest AGB estimation at high elevations in harsh environments. In the study, Sentinel-1 and ALOS-2 data were used as independent data sources for forest AGB modeling, and ICESat-2 was considered as an additional source of data used in conjunction to further improve forest AGB estimation. The XGBoost model in ensemble learning with multiple parametric and nonparametric models was built to estimate natural forest AGB in the eastern Tibetan Plateau. The results showed that the forest AGB models constructed by ALOS-2 were overall better than Sentinel-1, and the model errors were further greatly reduced after joining ICESat-2. The XGBoost achieved the best prediction accuracy among all models with a coefficient of determination of 0.75 (RMSE=18.57 Mg/ha).

***Index Terms*—**Aboveground biomass, natural forests, dual-polarized SAR, ICESat-2, Ensemble learning,

## 1 Introduction

Accurate estimation of AGB is a key task in the dynamic monitoring of forest ecosystems, which can deepen the understanding of biosphere interactions and the global carbon cycle. Modelling by combining remotely sensed data sources with a small amount of ground survey data has become a popular method for AGB estimation [1].

Synthetic Aperture Radar (SAR) is capable of penetrating clouds and fog for all-weather, all-day ground observation, and has been effectively used for terrain mapping and disaster monitoring. The ignorance of weather, especially in areas with harsh climates, makes SAR very promising for resource surveys [2]. In addition, the ICESat-2 of spaceborne LiDAR can obtain large-scale three-dimensional structural information of the Earth's surface. Combining ICESat-2 with optical data is effective in obtaining regional parameters such as forest canopy height and biomass, but the results of forest AGB estimation may be subject to large uncertainties due to the limitation of the saturation phenomenon [3].

Machine learning algorithms, such as multiple linear regression (MLR), Backpropagation (BP), neural network, k-nearest neighbors (kNN), extreme learning machine (ELM), and support vector machine (SVM), have been successfully applied to the prediction of forest parameters [4]. However, due to the complexity of forest ecosystems, the prediction results often lead to large uncertainties. Ensemble learning algorithms can reduce the error of a single model by combining the prediction results of multiple base learners, thus significantly improving the accuracy and stability of the prediction. As a representative of ensemble learning algorithms, the Extreme Gradient Boosting (XGBoost) is an end-to-end optimization algorithm based on boosted trees, which gradually optimizes each base learner in it through a forward-stepping algorithm. XGBoost has been proven to be able to achieve the estimation of grassland biomass, soil moisture, etc. [5], but the effectiveness of the AGB estimation in natural forests needs to be further verified.

This study aims to develop an XGBoost model for the estimation of natural forest AGB in the eastern Tibetan Plateau. To validate the effectiveness of the method, commonly used machine learning algorithms were also established for comparison. In particular, Sentinel-1 and ALOS-2 were used as independent SAR data sources to explore the effect of polarized SAR in different wavelengths on the estimation of natural forest AGB. ICESat-2 data was used for joint modeling to further improve the accuracy of AGB estimation.

## 2 Materials and methods

### 2.1 Study Area and Measured forest AGB

The Natural Forest Resources Protection Project (NFRPP) area is located in the eastern part of the Tibet Autonomous Region, with a geographic location of 97°20′~99°06′E, 28°41′~32°32′N. The NFRPP area has a temperate sub-humid climate, with an average annual temperature of 7.6°C, and an average annual precipitation of about 477.7 mm. The main dominant tree species in NFRPP area include spruce (*Picea asperata*), fir (*Abies fabri*), larch (*Larix gmelinii*) and alpine pine (*Pinus densata*).

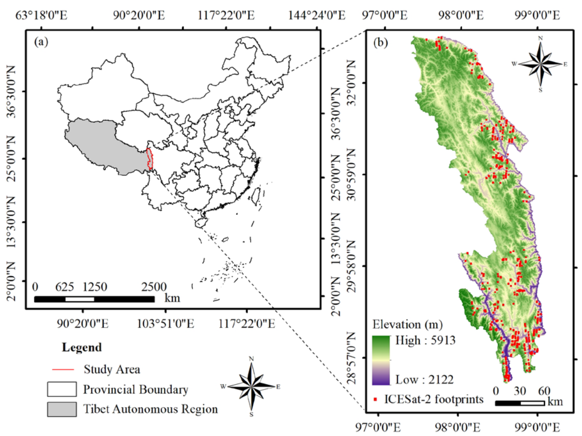


**Figure 1**. Location of the study area and the ICESat-2 footprints distribution.

The Forest Management Inventory (FMI) database of NFRPP area in 2019 was used for calculating forest AGB values. The boundaries of forests, the dominant tree species, and GSV values were obtained. The biomass expansion factor (BEF) method was used to calculate the forest AGB values per unit area. The forest AGB values of all the 841 sample plots ranged from 91.84 Mg/ha to 346.07 Mg/ ha, with a mean value of 143.06 Mg/ha and a standard deviation of 37.91 Mg/ha.

### 2.2 Data Processing

Sentinel-1 C-band SAR and ALOS-2 L-band SAR data were considered as independent data sources for comparison in AGB estimation modeling. The Sentinel-1 backscatter coefficient dataset covering NFRPP area in 2019 was acquired through the Google Earth Engine (GEE) platform [6], which contains both VV and VH polarized information with a resolution of 10 m. The dataset is provided in Ground Range Detection (GRD) products that have been radiometrically calibrated and orthogonally corrected with the Sentinel-1 toolbox. The ALOS-2 data are derived from the PALSAR-2 global mosaic datasets [7], which are free and open annual datasets generated by the Japan Aerospace Exploration Agency (JAXA). The dataset is available in HH and HV polarizations, given in linear amplitude Gamma-0 backscatter, with a resolution of about 25 m.

The ATL08 product of ICESat-2 covering the NFRPP area in 2019 was downloaded from the National Ice and Snow Data Center (NSIDC). All footprints were cropped using forest boundaries within the study area, and footprints marked as valid in the ATL08 field were retained (Figure 1).

### 2.3 Feature extraction and selection

Feature variables can significantly influence the modelling process. To make full use of the polarization information, texture features were extracted from the SAR data in addition to the original backscattering coefficients, and the gray level co-occurrence matrix (GLCM) tool in ENVI 5.3 was used to extract texture features. A total of eight textures were extracted in the study, which were Mean Value, Homogeneity, Dissimilarity, Angular Two Moment, Variance, Contrast Ratio, Entropy, and Relevance, respectively. The window sizes were set to 3×3, 5×5, 7×7, 9×9 and 11×11 to account for the effect of different windows on feature extraction. A total of 15 LiDAR variables, including ASR and canopy height metrics, were extracted from ICESat-2 ATL08 in the NFRPP region.

Too many variables can lead to model redundancy, reducing predictive efficiency and leading to additional errors. For forest ecosystems, remotely sensed features and forest parameters often show complex nonlinear relationships, which leads to the limited effectiveness of the traditional linear approach to feature variable selection. The stepwise random forest (SRF) method [8] was used in the study to screen different sets of feature variables and identify features highly correlated with forest AGB in nonlinear space.

### 2.4 Model Development and Assessment

Parametric models represented by Multiple Linear Regression (MLR) can directly describe the linear relationship between multiple independent variables and the dependent variable and have been widely used for modeling and prediction of vegetation parameters. However, the feature variables and forest parameters usually exhibit nonlinear relationships, resulting in limited predictive effectiveness of MLR.

Among the non-parametric models, SVM, kNN, ELM, and BP algorithms have been widely used for forest parameter estimation due to their strong fitting ability and applicability [9]. The SVM algorithm has several available kernel functions which are suitable for different sample distributions, respectively. However, SVM is particularly sensitive to model parameters and noise, leading to unstable prediction results. The kNN algorithm is easy to implement, but is too computationally intensive for high-dimensional data, resulting in limited efficiency. The BP algorithm has strong fitting ability and wide applicability, but the model training speed is slow and easily falls into the local optimal solution. ELM is a single hidden layer feed-forward neural network that improves generalization performance and computational efficiency compared to BP. The XGBoost is an improved integration algorithm based on Gradient Boosting Decision Tree (GBDT), which is widely used in the field of classification and regression due to its advantages of fast computation, high efficiency and accuracy, and strong generalization ability [5].

In the NFRPP area, an XGBoost model for natural forest AGB estimation was constructed, and hyper-parameter optimization of the XGBoost model was carried out to further improve the model estimation accuracy. The maximum depth of each tree, learning rate, and number of trees were iterated and the parameter combination that can achieve the lowest error was determined. In addition, multiple parametric and non-parametric models were constructed for AGB modeling comparison to verify the effectiveness and applicability of the optimized XGBoost model.

In the study, 80% of the samples were randomly selected as training samples for model construction, and the remaining 20% of the samples are used as test samples for model accuracy validation. The Coefficient of Determination ($R^2$) and the Root Mean Square Error (RMSE) were calculated for model accuracy assessment [10].

## 3 Results

### 3.1 AGB estimation with SAR data

The fitting results of the forest AGB estimation models constructed using Sentinel-1 and ALOS-2 as independent data sources are shown in Figures 2 and 3, respectively. For Sentinel-1, the RMSEs of all models ranged from 29.95 to 37.08 Mg/ha. The MLR, BP and ELM models were less valid, with the coefficient of determination of the models being less than 0.1, while the XGBoost model achieved the highest $R^2$ and the lowest RMSE of all the models.

ALOS-2 maintained a similar prediction trend to Sentinel-1, with the overall estimation accuracy of the non-parametric model being better than the parametric model, but the XGBoost model still had the smallest prediction error compared to the other models. Compared to MLR, the RMSE of XGBoost decreased by 18.6%.

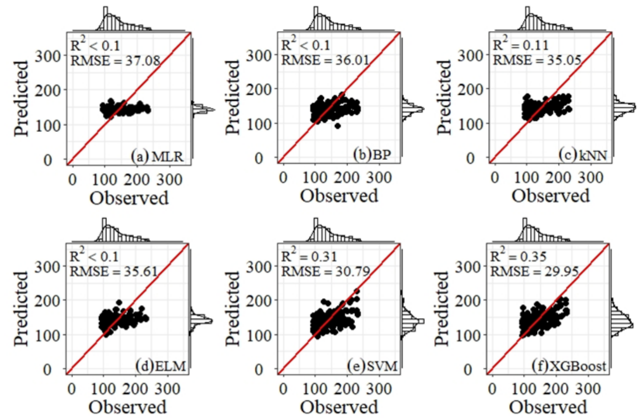


**Figure 2**. Forest AGB estimation results with Sentinel-1.

Overall, the accuracy of the forest AGB prediction models constructed with ALOS-2 as the data source was better than that of Sentinel-1, and the RMSEs of MLR, BP, kNN, ELM, SVM, and XGBoost models were reduced by 9.1%, 4%, 11.2%, 11.3%, 7.3% and 8.4%, respectively. Among all the forest AGB estimation models constructed with Sentinel-1 and ALOS-2 as data sources, the XGBoost model constructed with ALOS-2 finally achieved the best prediction accuracy with an $R^2$ of 0.47 (RMSE = 27.43 Mg/ha).

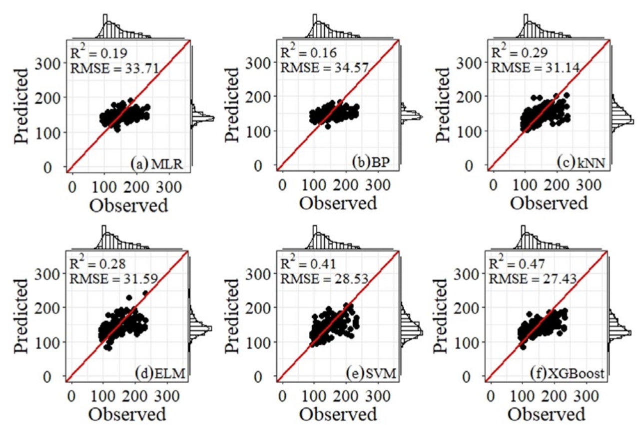


**Figure 3**. Forest AGB estimation results with ALOS-2.

### 3.2. AGB estimation with ICESat-2 and combined data

Compared with using feature variables extracted from ALOS-2 or Sentinel-1, the RMSEs of the forest AGB estimation models constructed using the canopy height metrics extracted from ICESat-2 were significantly reduced (Figure 4). Spaceborne LiDAR can obtain three-dimensional structural information of forests, and the extracted features can more accurately describe the vertical features of forests, which can effectively improve the accuracy of forest AGB estimation.

In addition, ALOS-2 and ICESat-2 were combined in this study to further improve forest AGB estimation. The RMSEs of all models were significantly reduced, with XGBoost still achieving the highest estimation accuracy with a coefficient of determination of 0.75 (RMSE = 18.57 Mg/ha), which is an average of 4.9% reduction in RMSE compared to ICESat-2 alone (Figure 5).

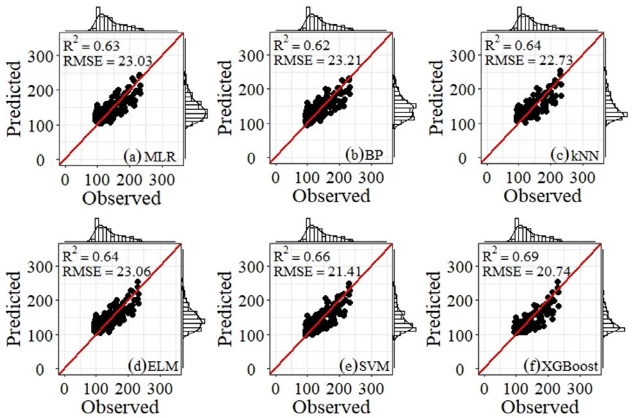


**Figure 4**. Forest AGB estimation results of ICESat-2.

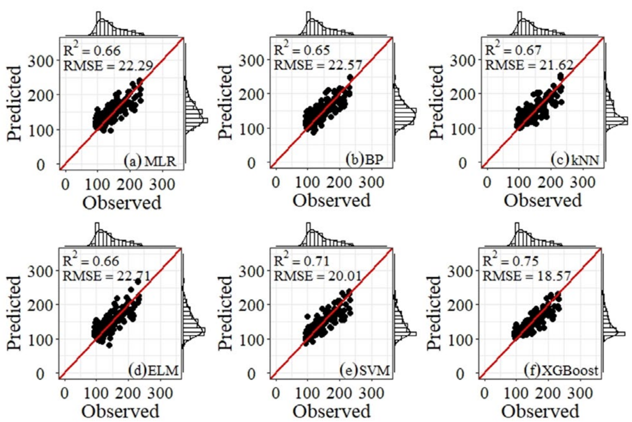


**Figure 5**. Forest AGB estimation results of combination of ALOS-2 and ICESat-2.

**4 Conclusions**

Compared with Sentinel-1 C-band SAR, ALOS-2 L-band SAR, which has a longer wavelength, has a stronger penetration capability and can obtain more accurate forest canopy parameters, thus realizing a higher accuracy of forest AGB estimation. The combination of ICESat-2 data further realized the complementarity of observations and greatly reduced the estimation error. The XGBoost achieved the best prediction accuracy among all the models, which is an average of 4.9% reduction in RMSE compared to ICESat-2 alone. High-precision AGB estimation of natural forests can be realized by combining SAR and spaceborne LiDAR, which provides a new reference for vegetation surveys in high-altitude areas with harsh climates.

**5 Acknowledgements**

This work was supported in part by the National Natural Science Foundation of China under Grant 62122091.

**References**

[1] J. Xiao, F. Chevallier, C. Gomez, et al. "Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years". *Remote Sensing of Environment*, vol. 233, no. 111383, 2019.
[2] M. Mahdianpari, B. Salehi, F. Mohammadimanesh, and M. Motagh. "Random forest wetland classification using ALOS-2 L-band, RADARSAT-2 C-band, and TerraSAR-X imagery". *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 13–31, 2019.
[3] F. Jiang, H. Sun, K. Ma, L. Fu, and J. Tang. "Improving aboveground biomass estimation of natural forests on the Tibetan Plateau using spaceborne LiDAR and machine learning algorithms". *Ecological indicators*, vol. 143, n. 109365, 2022.
[4] C. Gleason, J. Im. "A Review of Remote Sensing of Forest Biomass and Biofuel: Options for Small-Area Applications," *GIScience & Remote Sensing*, vol. 48, no.2, pp. 141–170, 2011.
[5] T. Chen T, C. Guestrin. "XGBoost: A Scalable Tree Boosting System," *ACM*, 2016.
[6] H. Tamiminia, B. Salehi, M. Mahdianpari, et al., "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 152–170, 2020.
[7] S. Masanobu, I. Takuya Itoh, M. Takeshi, et al., "New Global Forest/Non-forest Maps from ALOS PALSAR Data (2007-2010)," *Remote Sensing of Environment*, vol. 155, pp. 13-31, 2014.
[8] F. Jiang, M. Kutia, A. Sarkissian, et al., "Estimating the Growing Stem Volume of Coniferous Plantations Based on Random Forest Using an Optimized Variable Selection Method," *Sensors*, vol. 20, n. 7248, 2020.
[9] X. Li, Z. Liu, H. Lin, et al., "Estimating the Growing Stem Volume of Chinese Pine and Larch Plantations based on Fused Optical Data Using an Improved Variable Screening Method and Stacking Algorithm," *Remote Sensing*, vol. 12, no.5, 2020.
[10] C. Willmott, S. Ackleson, R. Davis, et al., "Statistics for the evaluation and comparison of models," *Journal of Geophysical Research-Oceans*, vol. 90, pp. 8995–9005, 1985.