

Mô hình dự đoán sinh khối cây cá lẻ của rừng trồng keo lai dựa trên thuật toán rừng cây

Nguyễn Văn Quý¹, Bùi Mạnh Hưng², Nguyễn Hồng Hải², Phùng Văn Khang³, Nguyễn Văn Thịnh^{3*}

¹Trung tâm nhiệt đới Việt – Nga, Chi nhánh Phía Nam

²Trường Đại học Lâm nghiệp

³Viện Khoa học Lâm nghiệp Việt Nam

Single tree biomass prediction model of *acacia* hybrid plantations based on random forest algorithm

Nguyen Van Quy¹, Bui Manh Hung², Nguyen Hong Hai², Phung Van Khang³, Nguyen Van Thinh^{3*}

¹Southern Branch of Joint Vietnam-Russia Tropical Science and Technology Research Center

²Vietnam National University of Forestry

³Vietnamese Academy of Forest Sciences

*Corresponding author: nguyenthinhfsiv@gmail.com

<https://doi.org/10.55250/jo.vnuf.14.5.2025.040-049>

Thông tin chung:

Ngày nhận bài: 15/05/2025

Ngày phản biện: 15/06/2025

Ngày quyết định đăng: 21/07/2025

Từ khóa:

Dữ liệu huấn luyện, Đồng Nai, sinh khối rừng, trí tuệ nhân tạo, trữ lượng cacbon.

Keywords:

Artificial intelligence, carbon stocks, Dong Nai, forest biomass, training data.

TÓM TẮT

Các mô hình sinh khối cây cá lẻ đóng vai trò quan trọng trong việc ước tính sinh khối rừng. Nghiên cứu này phát triển các mô hình sinh khối dựa trên dữ liệu thu thập từ 90 cây keo lai tuổi từ 2-10 năm trong các lâm phần rừng trồng thuần loài tại Khu Dự trữ Sinh quyển Đồng Nai. Phương pháp cắt hạ và sấy khô được sử dụng để xác định sinh khối trên mặt đất (AGB) của cây. Trong đó, 60 cây được chọn ngẫu nhiên để xây dựng mô hình và 30 cây còn lại được dùng để kiểm định. Ba phương pháp mô hình hóa được áp dụng bao gồm hồi quy Ridge (RR), tương quan sinh trưởng (AG) và cây rừng (RF), với các biến đầu vào là tuổi (A), chiều cao vút ngọn (H) và đường kính ngang ngực (dbh), trong khi AGB đóng vai trò là biến phụ thuộc. Hiệu suất mô hình được đánh giá thông qua các chỉ số hệ số xác định (R^2), sai số trung phương (RMSE) và sai số tuyệt đối trung bình (MAE). Kết quả cho thấy mô hình RF vượt trội hơn so với RR và AG, với R^2 cao hơn và các giá trị RMSE và MAE thấp hơn. Phân tích tầm quan trọng của các biến trong mô hình RF cho thấy dbh là yếu tố có ảnh hưởng lớn nhất đến kết quả dự đoán AGB. Thêm vào đó, việc tích hợp A vào RF đã cải thiện độ chính xác trong ước lượng AGB của cây keo lai. Kết quả nghiên cứu cung cấp cơ sở khoa học quan trọng cho việc ước tính nhanh và chính xác AGB của rừng trồng keo lai tại khu vực nghiên cứu.

ABSTRACT

Single-tree biomass models play a critical role in the accurate estimation of forest biomass. This study involved collecting data from 90 individual *Acacia* hybrid trees, aged 2 to 10 years, obtained through harvesting from plantations in Dong Nai Biosphere Reserve, Dong Nai province. Among these, 60 trees were randomly selected for model development, while the remaining 30 were designated for model validation. We employed Ridge Regression, Allometric Growth, and Random Forest algorithms, using independent variables including tree age, total height, and diameter at breast height (dbh), with aboveground biomass (AGB) as the dependent variable. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The results demonstrated that the Random Forest model significantly outperformed

both Ridge Regression and Allometric Growth, showing superior values for R^2 , RMSE, and MAE. Notably, the variable importance analysis in the Random Forest model identified dbh as the most significant predictor of AGB. Furthermore, including tree age in the Random Forest model enhanced the accuracy of AGB predictions for *Acacia* hybrids. These findings provide valuable insights for rapid and precise estimation of biomass and carbon stocks in *Acacia* hybrid plantations within the study area.

1. ĐẶT VẤN ĐỀ

Sự tăng trưởng liên tục của nền kinh tế toàn cầu chủ yếu phụ thuộc vào năng lượng từ nhiên liệu hóa thạch [1]. Tuy nhiên, việc khai thác quá mức nguồn năng lượng này đã dẫn đến nhiều tác động tiêu cực đối với môi trường, bao gồm phát thải khí nhà kính [2]. Trong bối cảnh đó, sinh khối thực vật đóng vai trò then chốt trong các hệ thống cung cấp năng lượng bền vững, chiếm khoảng 70% tổng sản lượng năng lượng tái tạo toàn cầu [3]. Sinh khối thực vật không chỉ là giải pháp thay thế quan trọng cho nhiên liệu hóa thạch mà còn có khả năng chuyển đổi thành các dạng nhiên liệu sinh học khác nhau [4]. Để đáp ứng nhu cầu giảm phát thải khí nhà kính và tiêu thụ năng lượng bền vững, trữ lượng cacbon trong các khu rừng trồng, đặc biệt là các loài cây gỗ phát triển nhanh, đang ngày càng được các nhà hoạch định chính sách quan tâm [5, 6]. Keo lai (*Acacia auriculiformis* A. Cunn. ex Benth. × *Acacia mangium* Willd) là một trong những loài cây trồng rừng chủ lực tại Việt Nam, có thời gian thành thực kinh tế ngắn, chỉ từ 7-12 năm [7, 8]. Vì vậy, việc ước tính nhanh chóng và chính xác sinh khối của rừng trồng keo lai phục vụ mục đích thương mại và sản xuất sinh khối ở quy mô công nghiệp trở nên vô cùng cần thiết.

Mô hình sinh khối cây cá lẻ là công cụ quan trọng trong việc ước tính sinh khối rừng [9]. Nhiều nghiên cứu về sinh khối và trữ lượng cacbon của cây gỗ và rừng tại Việt Nam đã được thực hiện, từ đó các mô hình sinh khối cho các loài cây gỗ khác nhau đã được thiết lập, hỗ trợ đáng kể cho công tác đánh giá tiềm

năng lưu trữ cacbon và phân tích biến đổi trong khả năng tích lũy cacbon của rừng [10-12]. Các phương pháp ước tính sinh khối thường sử dụng mô hình sinh khối cây cá lẻ kết hợp với các yếu tố dễ đo lường như đường kính và chiều cao của cây, cho thấy sự kết hợp này mang lại độ tin cậy cao hơn [13]. Tuy nhiên, sai số trong đo chiều cao cây có thể gây ra sai lệch trong ước tính sinh khối, điều này nhấn mạnh nhu cầu tích hợp yếu tố tuổi cây để giảm thiểu lỗi [14].

Thuật toán rừng cây (Random Forest - RF), được phát triển bởi Breiman, là một phương pháp khai thác dữ liệu trong trí tuệ nhân tạo và khoa học máy tính [15]. RF nổi bật với khả năng lựa chọn ngẫu nhiên các biến đặc trưng và khả năng thích ứng tốt với các mô hình phi tuyến [16]. Thuật toán này thể hiện tính ổn định và độ chính xác cao hơn so với các mô hình cây quyết định đơn lẻ, đồng thời hiệu quả trong việc xử lý các vấn đề đa cộng tuyến giữa các biến trong mô hình [17]. Chính vì vậy, RF đã được ứng dụng rộng rãi trong nghiên cứu sinh thái học và các lĩnh vực liên quan [18].

Cây keo lai không chỉ mang lại giá trị kinh tế cao mà còn có tiềm năng lớn trong việc lưu trữ cacbon, góp phần giảm thiểu biến đổi khí hậu toàn cầu [19]. Mặc dù đã có nhiều nghiên cứu ước tính sinh khối của rừng trồng keo lai, việc áp dụng các phương pháp thống kê mô hình hóa bằng máy (machine learning) kết hợp với yếu tố tuổi cây trong mô hình dự đoán sinh khối vẫn còn khá hạn chế [20-23]. Nghiên cứu này thu thập dữ liệu thực nghiệm từ 90 cây keo lai trong các lâm phần rừng trồng thuần loài tại

Khu Dự trữ Sinh quyển Đồng Nai và áp dụng mô hình RF để ước tính sinh khối trên mặt đất (AGB) của cây cá lẻ, so sánh với các mô hình hồi quy Ridge (Ridge Regression - RR) và tương quan sinh trưởng (Allometric Growth - AG). Các câu hỏi nghiên cứu được đặt ra: (i) Trong các yếu tố được sử dụng để xây dựng mô hình ước tính AGB của cây keo lai, bao gồm tuổi cây (A), đường kính ngang ngực (dbh) và chiều cao vút ngọn (H), yếu tố nào có ảnh hưởng nhiều nhất đến kết quả dự đoán của mô hình? (ii) Có sự khác biệt nào giữa các mô hình RF, RR và AG trong ước tính AGB của cây keo lai không? (iii) Mô hình nào trong ba mô hình trên cho hiệu suất dự đoán tốt nhất? Kết quả của nghiên cứu này sẽ cung cấp cơ sở lý thuyết nhằm nâng cao độ chính xác trong ước tính sinh khối rừng trồng keo lai, hỗ trợ hiệu quả cho chiến lược quản lý rừng bền vững và ứng phó với biến đổi khí hậu toàn cầu.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Khu vực nghiên cứu

Nghiên cứu này tập trung vào AGB của cây keo lai trong các lâm phần rừng trồng thuần loài ở nhiều độ tuổi khác nhau tại Khu Dự trữ Sinh quyển Đồng Nai. Các lâm phần keo lai trong khu vực nghiên cứu chủ yếu bao gồm các dòng vô tính AH7, BV10, BV16 và BV32, với mật độ trồng ban đầu dao động từ 1.600-2.500 cây/ha. Ở giai đoạn khai thác (6–10 năm), mật độ cây còn lại giảm xuống còn khoảng 800–1.200 cây/ha. Các biện pháp kỹ thuật lâm sinh được áp dụng bao gồm tuyển chọn giống, bón phân và chăm sóc định kỳ trong ba năm đầu sau khi trồng [23].

Về điều kiện thổ nhưỡng, Khu Dự trữ Sinh quyển Đồng Nai chủ yếu là đất Feralit đỏ vàng phát triển trên đá mẹ phiến thạch sét và granit, có tầng đất dày và độ phì nhiêu cao. Khí hậu tại khu vực nghiên cứu thuộc kiểu gió mùa cận xích đạo, đặc trưng bởi nền nhiệt độ ổn định và lượng mưa nhiều. Nhiệt độ trung bình năm dao

động từ 25-27°C, với biên độ nhiệt ngày đêm và giữa các tháng tương đối thấp. Độ ẩm không khí trung bình duy trì ở mức 80–83%, trong khi tổng lượng mưa hàng năm đạt từ 2.500-2.800 mm [24]. Những điều kiện tự nhiên này tạo thuận lợi cho sự sinh trưởng và phát triển của các loài cây lâm nghiệp, đặc biệt là nhóm cây công nghiệp ngắn ngày.

2.2. Thu thập dữ liệu

Dữ liệu được thu thập trong quá trình thực hiện Đề tài Khoa học và Công nghệ độc lập cấp Quốc gia “*Nghiên cứu khả năng lưu trữ cacbon của các kiểu rừng tại Khu Dự trữ Sinh quyển Đồng Nai*”, mã số ĐTĐL.XH-06/20. Nghiên cứu đã thực hiện cắt hạ và thu các bộ phận trên mặt đất bao gồm thân, cành và lá của 90 cây keo lai ở 9 lâm phần từ 2-10 tuổi. Trong mỗi lâm phần, 10 cây đại diện cho các cấp đường kính khác nhau đã được lựa chọn để cắt hạ. Các cây cắt hạ được tách riêng từng bộ phận và cân ngay tại hiện trường bằng cân có độ chính xác 0,1 gram để xác định sinh khối tươi của từng bộ phận. Để xác định sinh khối khô trên mặt đất (AGB) của từng cây, nghiên cứu sử dụng phương pháp mẫu đại diện, trong đó 6 mẫu có khối lượng từ 0,3-1,0 kg được lấy từ mỗi bộ phận của cây [20]. Những mẫu này sau đó được vận chuyển đến phòng thí nghiệm và sấy khô ở nhiệt độ 80°C đối với lá (6-8 giờ sấy) và 103°C đối với thân và cành (72 giờ sấy). Sau khi sấy khô, các mẫu được cân lại để xác định tỉ lệ giữa sinh khối tươi và khô, phục vụ cho việc tính toán sinh khối khô trên mặt đất của từng bộ phận và toàn bộ cây mẫu.

Bộ dữ liệu phục vụ xây dựng các mô hình bao gồm ba biến độc lập: tuổi cây (A), chiều cao vút ngọn (H) và đường kính ngang ngực (dbh), cùng một biến phụ thuộc là AGB của cây. Các dữ liệu thống kê phục vụ cho mô hình hóa (tập dữ liệu huấn luyện) và xác thực mô hình dự đoán AGB (dữ liệu kiểm tra) được trình bày trong Bảng 1.

Bảng 1. Thông tin của tập dữ liệu để xây dựng các mô hình ước tính sinh khối cây keo lai

Dữ liệu	Các biến	N	Max	Min	Mean	SE	SD	VR
Huấn luyện	A	60	10	2	5,5	0,265	2,72	7,39
	dbh	60	24,8	4,21	12,70	5,38	5,38	23,90
	H	60	27,8	5,81	15,97	5,30	5,30	26,13
	AGB	60	744,134	8,846	156,268	161,603	161,603	115,581
Kiểm tra	A	30	10	2	5,87	0,367	2,46	6,07
	dbh	30	24,7	4,20	12,68	5,37	5,37	28,89
	H	30	27,4	5,29	15,94	5,34	5,29	28,06
	AGB	30	689,353	8,763	158,269	167,032	165,502	132,587

Ghi chú: A – tuổi cây; dbh – đường kính ngang ngực của cây (cm), H – chiều cao vút ngọn của cây (m); AGB – sinh khối khô trên mặt đất (kg); N – số lượng cây lấy mẫu (cây); Max – giá trị lớn nhất; Min – giá trị nhỏ nhất; Mean – giá trị trung bình, SE – sai số chuẩn; SD – độ lệch chuẩn; VR – phương sai.

2.3. Phương pháp nghiên cứu

Mô hình rừng cây (Random Forest) là một thuật toán trong lĩnh vực thống kê mô hình hóa bằng máy, dựa trên nguyên tắc huấn luyện và tự học từ dữ liệu tổng hợp [15]. Thuật toán này không chỉ ước lượng và xử lý hiệu quả các giá trị bị thiếu trong tập dữ liệu mà còn đảm bảo độ chính xác cao cho mô hình [16]. RF xây dựng một tập hợp các cây quyết định từ dữ liệu huấn luyện bằng cách chọn ngẫu nhiên các nút (câu hỏi để giải quyết vấn đề) trong từng cây quyết định đơn lẻ. Kết quả cuối cùng của mô hình dự đoán là giá trị trung bình của tất cả các cây quyết định [25].

Trong nghiên cứu này, mô hình RF được phát triển dựa trên hai phần ba dữ liệu khảo sát, với các biến độc lập A, dbh và H tương ứng, được sử dụng làm tập dữ liệu huấn luyện (training dataset). Một phần ba còn lại của dữ liệu khảo sát được sử dụng làm tập dữ liệu xác thực (validation dataset), nhằm đánh giá độ chính xác trong dự đoán của mô hình. Để so sánh hiệu suất của mô hình RF, hai mô hình khác là hồi quy Ridge (sử dụng 3 biến độc lập A, dbh và H) và tương quan sinh trưởng (chỉ sử dụng 2 biến dbh và H làm biến độc lập) cũng được đưa vào trong nghiên cứu [26, 27]. Cả ba mô hình này được đánh giá hiệu suất hoạt động thông qua các chỉ số gồm hệ số quan hệ (R^2), sai số trung phương (RMSE) và sai số tuyệt đối

trung bình (MAE) với các công thức tính toán cụ thể như sau [13]:

$$R^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_{i=0}^n |y_i - \hat{y}_i|}{n} \quad (3)$$

Trong công thức 1, hệ số quan hệ R^2 đo lường mức độ giải thích của mô hình đối với biến phụ thuộc; giá trị R^2 dao động từ 0-1, với giá trị gần 1 cho thấy mô hình phù hợp hơn. Công thức 2 giới thiệu hệ số RMSE, phản ánh sai số trung bình giữa giá trị dự đoán của mô hình và giá trị thực tế; giá trị RMSE nhỏ hơn biểu thị cho hiệu suất hoạt động cao hơn của mô hình. Tương tự, công thức 3 trình bày hệ số MAE để đo lường sai số trung bình, nhưng MAE tính toán giá trị tuyệt đối của sai số; giá trị MAE càng nhỏ thì mô hình càng được đánh giá cao. Trong các công thức trên, y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự đoán từ mô hình, $\bar{y}$ là giá trị trung bình của biến quan sát trong tập dữ liệu, và n là kích thước mẫu.

Để đánh giá tầm quan trọng của các biến trong mô hình RF, nghiên cứu áp dụng chỉ số quan trọng hoán vị (Permutation importance index - PI) [28]. Các thuật toán thống kê mô hình hóa dựa trên cây quyết định được sử dụng nhằm xác định tầm quan trọng tương đối của

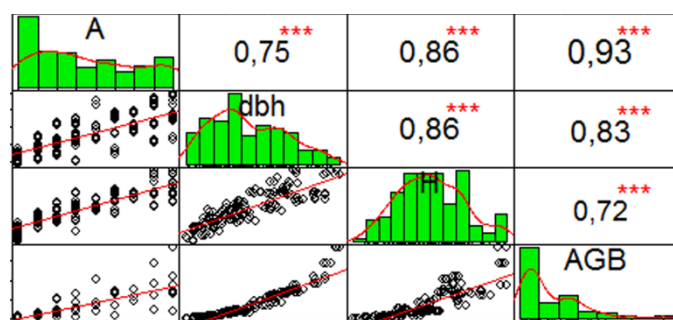
từng biến thông qua việc đo lường độ phức tạp của dữ liệu. Tuy nhiên, phương pháp này có thể dẫn đến việc phóng đại tầm quan trọng của những biến có dung lượng lớn và tính liên tục cao. Do đó, chỉ số PI đã được lựa chọn để kiểm tra tầm quan trọng của các biến, dựa trên tác động của việc sắp xếp lại một cách ngẫu nhiên của chúng. Chỉ số PI được đánh giá là có nhiều ưu điểm, bao gồm khả năng tính toán dễ dàng, độ chính xác cao và khả năng diễn giải tốt [29].

Tất cả các phân tích trong nghiên cứu được thực hiện trên phần mềm R phiên bản 4.4.1 với sự hỗ trợ của các gói 'psych' cho kiểm tra tương quan giữa các biến, 'randomForest' và 'caTools' cho xây dựng mô hình RF, cũng như 'glmnet' và 'Metrics' cho xây dựng các mô hình RR và AG.

3. KẾT QUẢ NGHIÊN CỨU

3.1. Các nhân tố điều tra ảnh hưởng đến sinh khối cây cá lẻ của loài keo lai

Mối tương quan giữa AGB và các nhân tố điều tra A, dbh và H của các cá thể cây keo lai trong khu vực nghiên cứu được thể hiện trong Hình 1. Kết quả phân tích cho thấy AGB có mối tương quan dương với A, H và dbh và tất cả đều đạt mức ý nghĩa cao ($p < 0,001$). Hệ số tương quan giữa các biến với AGB giảm dần theo thứ tự: A (0,93), dbh (0,83) và H (0,72). Đặc biệt, trong ba nhân tố này, cả ba đều cho thấy mối tương quan có ý nghĩa từ chặt đến rất chặt (hệ số tương quan dao động từ 0,72-0,93), trong đó A có mối tương quan với AGB chặt hơn ($r = 0,93$) so với dbh (0,83) và H (0,72), đồng thời cho thấy sự đa cộng tuyến giữa chúng.



Hình 1. Sự tương quan giữa các nhân tố điều tra và AGB của các cá thể cây keo lai
(Mức ý nghĩa thông kê: *** biểu thị giá trị $p < 0,001$)

3.2. Mô hình hồi quy Ridge cho ước tính sinh khối cây cá lẻ của loài keo lai

Các hệ số không chuẩn hóa từ mô hình hồi quy Ridge (RR) cho biết mức độ ảnh hưởng của từng biến độc lập đến biến phụ thuộc trong mô hình. Như thể hiện trong Bảng 2, ảnh hưởng

của các nhân tố điều tra đến AGB của cây cá lẻ được xếp hạng theo thứ tự: A (18,88), dbh (6,64), H (4,17), phù hợp với kết quả phân tích sự tương quan giữa các nhân tố điều tra được trình bày trong Hình 1.

Bảng 2. Các thông số trong mô hình hồi quy Ridge cho ước tính AGB cây cá lẻ của keo lai

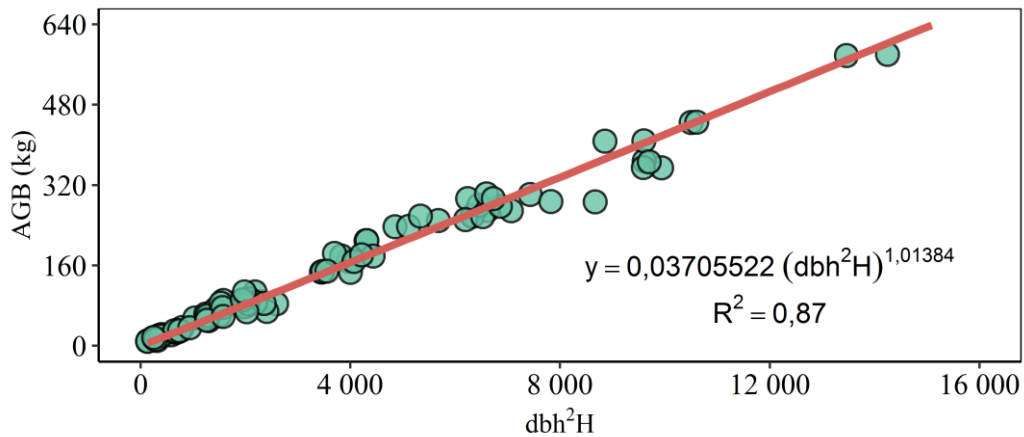
Các biến	B	SE	Giá trị t	Mức ý nghĩa	R^2	Giá trị F
A	18,88	2,231	8,463	$p < 0,001$	0,817	234,236
dbh	6,64	2,724	2,438	$p < 0,001$		
H	4,17	4,099	1,017	$p < 0,01$		

Ghi chú: B - Hệ số không chuẩn hóa cho biết mức độ ảnh hưởng của mỗi biến đến kết quả dự đoán AGB của mô hình; SE - Sai số chuẩn của hệ số cho biết độ chính xác của ước tính hệ số; Giá trị t được tính bằng cách lấy hệ số chia cho sai số chuẩn của nó. Giá trị t lớn hơn cho thấy rằng biến đó có ý nghĩa thống kê cao hơn trong mô hình. Hệ số quan hệ R^2 cho biết tỷ lệ biến thiên của AGB được giải thích bởi các biến trong mô hình. Giá trị F được sử dụng để kiểm tra tính có ý nghĩa của toàn bộ mô hình.

3.3. Mô hình tương quan sinh trưởng cho ước tính sinh khối cây cá lẻ của loài keo lai

Hình 2 cho thấy, mô hình tương quan sinh trưởng (AG), dù không bao gồm yếu tố tuổi cây, vẫn có khả năng giải thích tốt hơn sự biến đổi của AGB ($R^2 = 0,87$) so với mô hình RR ($R^2 = 0,817$). Điều này cho thấy rằng mối quan hệ

giữa AGB và các yếu tố A, H và dbh là rất phức tạp và không thể được mô tả đầy đủ chỉ bằng một mô hình tuyến tính đơn giản chỉ gồm 2 nhân tố điều tra là dbh và H, cần phải sử dụng thêm những nhân tố khác để thu được kết quả dự đoán tốt hơn.

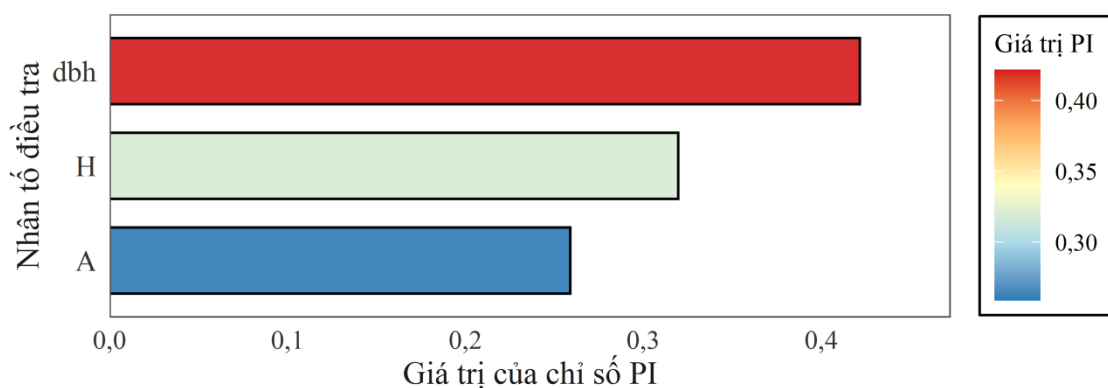


Hình 2. Mô hình tương quan sinh trưởng cho dự đoán sinh khối cây cá lẻ của loài keo lai

3.4. Mô hình rừng cây cho ước tính sinh khối cây cá lẻ của loài keo lai

Mô hình RF bao gồm ba tham số chính: (1) số lượng cây quyết định trong mô hình, mặc định là 500; (2) số biến được xem xét để tách tại mỗi nút, mặc định là 3; và (3) kích thước mẫu tối thiểu cho các nút đầu cuối, mặc định là

5. Như thể hiện trong Hình 3, tầm quan trọng của các nhân tố ảnh hưởng được xếp hạng như sau: dbh (0,42) > H (0,32) > A (0,26). Điều này chỉ ra rằng dbh có ảnh hưởng đáng kể nhất đến mô hình ước tính AGB của cây keo lai trong khu vực nghiên cứu.



Hình 3. Tầm quan trọng của các biến độc lập trong mô hình rừng cây

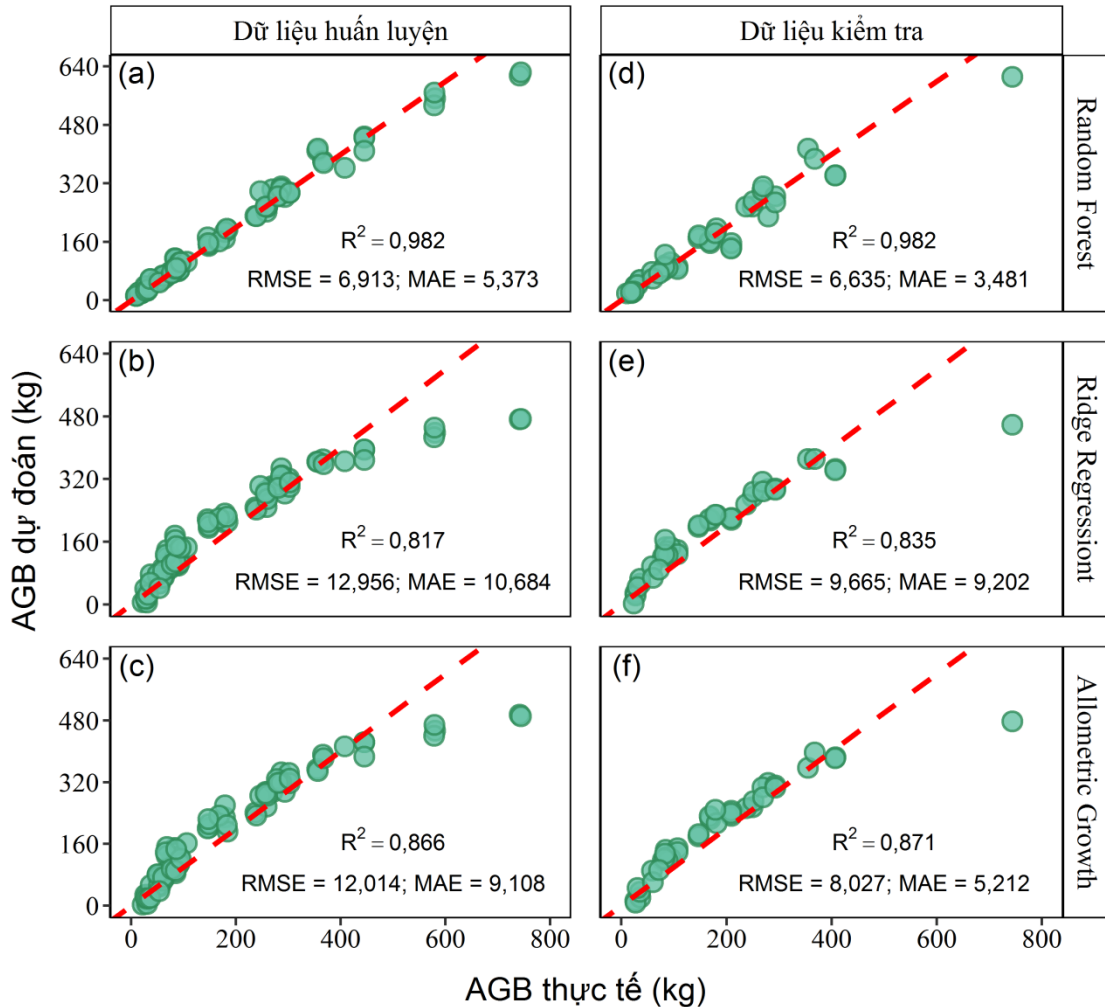
3.5. Đánh giá hiệu suất hoạt động của các mô hình

Thể hiện trong Hình 4, mô hình RF cho thấy sự cải thiện đáng kể trong ước tính AGB của cây

cá lẻ loài keo lai ở khu vực nghiên cứu, so với các mô hình RR và AG. Cụ thể, giá trị R^2 của mô hình RF tăng lần lượt 0,165 và 0,116 so với RR và AG trong tập dữ liệu huấn luyện. Đồng thời,

giá trị RMSE của mô hình RF giảm 6,043 so với RR và 5,101 so với AG, trong khi MAE giảm lần lượt 5,311 và 3,735 (Hình 4a-c). Trong tập dữ liệu kiểm tra, R^2 của mô hình RF cũng cải thiện 0,147 và 0,111 so với RR và AG, với RMSE giảm

3,030 và 1,392, và MAE giảm lần lượt 5,721 và 1,731 (Hình 4d-f). Những kết quả này chỉ ra rằng mô hình RF thể hiện hiệu suất dự đoán vượt trội so với hai mô hình RR và AG.



Hình 4. Hiệu suất hoạt động của các mô hình dự đoán AGB cho cây cá lẻ loài keo lai

4. THẢO LUẬN

Nghiên cứu này đã tiến hành đánh giá và so sánh hiệu quả của ba phương pháp ước tính AGB của cây cá lẻ loài keo lai tại Khu Dự trữ Sinh quyển Đồng Nai bao gồm RR, AG và RF. Kết quả nghiên cứu cho thấy RF thể hiện ưu thế rõ rệt so với hai phương pháp còn lại, phù hợp với nhận định từ các nghiên cứu trước đây về ứng dụng học máy trong mô hình hóa sinh khối rừng [30-32]. Trong số các phương pháp được đánh giá (RF, RR và AG), RF nổi bật nhờ khả năng xử lý hiệu quả các mối quan hệ phi tuyến và đa biến trong dữ liệu. Khác với các mô hình thống

kê truyền thống, thuật toán học máy này không yêu cầu giả định về phân phối dữ liệu hay mối quan hệ tuyến tính [15]. Cơ chế gom nhóm tự sinh (bootstrap aggregation - mỗi cây trong mô hình được huấn luyện dựa trên một tập dữ liệu con được lấy mẫu ngẫu nhiên trong tập dữ liệu gốc) giúp RF giảm thiểu hiện tượng quá khớp (overfitting) đồng thời xử lý tốt các tương tác biến và đa cộng tuyến [33]. Thực tế đã cho thấy trong nghiên cứu này, RF đạt hệ số $R^2 > 0,9$ với các chỉ số RMSE và MAE thấp hơn đáng kể so với RR và AG. Kết quả này tương đồng với các nghiên cứu về ước lượng sinh khối rừng nhiệt

đới [34, 35]. Một điểm mạnh đáng chú ý của RF là khả năng đánh giá mức độ quan trọng của các biến đầu vào [36]. Phân tích của chúng tôi xác nhận dbh là yếu tố ảnh hưởng mạnh nhất đến kết quả dự đoán AGB, phản ánh mối tương quan chặt chẽ giữa kích thước cây và sinh khối, phù hợp với các nghiên cứu trước đây về các loài cây sinh trưởng nhanh như keo và bạch đàn [37, 38]. Việc bổ sung biến tuổi cây vào mô hình giúp cải thiện độ chính xác, chứng tỏ các yếu tố sinh trưởng theo thời gian đóng vai trò quan trọng trong dự đoán AGB của cây keo lai. Tuy nhiên, RF cũng tồn tại một số hạn chế như tính chất "hộp đen" khiến việc diễn giải các mối quan hệ sinh học trở nên khó khăn hơn so với mô hình hồi quy truyền thống, đồng thời yêu cầu lượng dữ liệu huấn luyện lớn để đạt hiệu suất tối ưu [39].

Trong khi đó, RR là một phương pháp hồi quy tuyến tính sử dụng kỹ thuật chính quy hóa (regularization), tuy có ưu điểm trong việc giảm hiện tượng đa cộng tuyến và quá khớp, nhưng do bản chất tuyến tính nên không thể mô hình hóa đầy đủ các mối quan hệ phi tuyến giữa AGB và các biến đầu vào [40]. Điều này giải thích tại sao RR trong nghiên cứu này lại xác định tuổi cây là yếu tố quan trọng nhất (Bảng 2), khác với kết quả từ RF nhấn mạnh vai trò của dbh - phù hợp hơn với các nghiên cứu về sinh khối cây gỗ [13]. Hơn nữa, RR còn phụ thuộc vào giả định phân phối chuẩn và tính đồng nhất của phương sai (những yêu cầu khó đáp ứng với dữ liệu sinh khối thường có phân phối lệch và phương sai không đồng nhất) [41].

Về phương pháp AG, dù được sử dụng phổ biến nhờ tính đơn giản và dễ áp dụng, mô hình này chỉ phản ánh mối quan hệ đơn giản và cố định giữa các biến dbh, H và AGB mà bỏ qua các yếu tố tác động như tuổi cây và điều kiện môi trường. Kết quả nghiên cứu của nhóm tác giả cho thấy AG có hiệu suất thấp hơn RF, đặc biệt khi không tích hợp biến tuổi cây vào mô hình, phù hợp với nhận định từ các nghiên cứu của Yang và cộng sự và Vanclay, cho rằng mô hình allometric truyền thống thường đánh giá thấp sinh khối ở cây non (sapling) hoặc già

(overmature tree) [42, 43].

Với mục tiêu nâng cao độ chính xác của các mô hình ước tính sinh khối trong tương lai, nghiên cứu này đề xuất một số hướng nghiên cứu quan trọng. Thứ nhất, cần mở rộng tập dữ liệu để bao phủ đa dạng độ tuổi, điều kiện lập địa và kết hợp với dữ liệu viễn thám như LiDAR hoặc ảnh vệ tinh độ phân giải cao. Thứ hai, việc bổ sung các biến sinh thái như tỷ trọng gỗ, đường kính tán và chỉ số diện tích lá (LAI) sẽ giúp cải thiện đáng kể hiệu suất của mô hình. Cuối cùng, cần đánh giá thêm hiệu quả của các thuật toán học máy khác như Gradient Boosting (XGBoost) hoặc Support Vector Machine (SVM) để có sự so sánh toàn diện với RF.

5. KẾT LUẬN

Trong nghiên cứu này, thuật toán RF thể hiện hiệu suất vượt trội so với RR và AG trong việc mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các biến điều tra và AGB của cây keo lai cá lẻ. Ưu thế này bắt nguồn từ khả năng của RF trong việc loại bỏ yêu cầu về các giả định thống kê tiên nghiệm hoặc dạng hàm cụ thể, đồng thời duy trì độ chính xác cao ngay cả với dữ liệu đa chiều. Đặc biệt, việc tích hợp biến tuổi cây vào mô hình RF không chỉ tăng cường khả năng khái quát hóa mà còn nâng cao độ tin cậy thống kê của các ước lượng, biến RF thành công cụ tối ưu cho dự đoán AGB cây cá lẻ, phân tích các yếu tố ảnh hưởng và hỗ trợ ước lượng nhanh trữ lượng sinh khối rừng trồng keo lai.

Rutherford đã phân loại khoa học thành hai nhóm: khoa học đích thực (ám chỉ các ngành như vật lý với nền tảng lý thuyết sâu sắc) và "sưu tập tem" (các nghiên cứu đơn thuần mang tính mô tả, phân loại) [44]. Dựa trên quan điểm đó có thể thấy, nghiên cứu này không chỉ đóng góp vào bộ sưu tập giá trị về sinh khối rừng trồng, mà còn vượt qua giới hạn của một dạng sưu tập thông thường. Cụ thể, kết quả thu được có ý nghĩa học thuật quan trọng nhờ việc phát triển các mô hình ước tính sinh khối tối ưu, đồng thời mang lại ứng dụng thực tiễn cao: (i) hỗ trợ công tác quản lý rừng trồng bền vững tại Việt Nam thông qua dữ liệu định lượng chính xác; (ii) cung cấp công cụ ước tính nhanh sinh

khối rừng, góp phần vào các chiến lược giảm phát thải quốc gia. Như vậy, nghiên cứu của nhóm tác giả đã kết hợp hài hòa giữa giá trị học thuật và tính ứng dụng, đáp ứng tiêu chí của một công bố khoa học đích thực.

Lời cảm ơn

Nghiên cứu này sử dụng dữ liệu từ Đề tài KH&CN độc lập cấp Quốc gia "*Nghiên cứu khả năng lưu trữ cacbon của các kiểu rừng tại Khu Dự trữ Sinh quyển Đồng Nai*" (mã số ĐTDL.XH-06/20). Các tác giả trân trọng cảm ơn TS. Nguyễn Văn Thịnh (Viện Nghiên cứu Lâm sinh, Viện Khoa học Lâm nghiệp Việt Nam) đã cung cấp dữ liệu và đóng góp ý kiến chuyên môn quan trọng cho nghiên cứu. Nhóm tác giả cũng xin ghi nhận sự hỗ trợ từ các thành viên khác trong Đề tài.

TÀI LIỆU THAM KHẢO

- [1]. Nel, W. P. & Cooper, C. J. (2009). Implications of fossil fuel constraints on economic growth and global warming. *Energy Policy*. 37(1): 166-180.
- [2]. Wang, J. & Azam, W. (2024). Natural resource scarcity, fossil fuel energy consumption, and total greenhouse gas emissions in top emitting countries. *Geoscience frontiers*. 15(2): 101757.
- [3]. Demirbas, M. F., Balat, M. & Balat, H. (2009). Potential contribution of biomass to the sustainable energy development. *Energy Conversion and Management*. 50(7): 1746-1760.
- [4]. Ioelovich, M. (2015). Recent findings and the energetic potential of plant biomass as a renewable source of biofuels—a review. *Bioresources*. 10(1): 1879-1914.
- [5]. Psistaki, K., Tsantopoulos, G. & Paschalidou, A. K. (2024). An overview of the role of forests in climate change mitigation. *Sustainability*. 16(14): 6089.
- [6]. Sedjo, R. A. (2001). Forest carbon sequestration: some issues for forest investments.
- [7]. Dương Hồng Quân, Nguyễn Đức Kiên & Trần Việt Hà (2021). Sinh trưởng các dòng Keo lai (Acacia hybrid) trong khảo nghiệm dòng vô tính tại Ba Vì, Hà Nội và Cam Lộ, Quảng Trị. *Tạp chí Khoa học và Công nghệ Lâm nghiệp*. 4: 58-65.
- [8]. Phạm Thị Luyện, Hoàng Liên Sơn & Đỗ Văn Bản (2019). Xác định tuổi thành thực kinh tế của loài cây Keo lai (Acacia mangium x Acacia auriculiformis) tại vùng Đông Bắc Bộ. *Tạp chí Nông nghiệp và Phát triển nông thôn*. 10: 88-94.
- [9]. Kankare, V., Rätty, M., Yu, X., Holopainen, M., Vastaranta, M., Kantola, T., Hyypä, J., Hyypä, H., Alho, P. & Viitala, R. (2013). Single tree biomass modelling using airborne laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*. 85: 66-73.
- [10]. Dang, A. T. N., Nandy, S., Srinet, R., Luong, N. V., Ghosh, S. & Kumar, A. S. (2019). Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. *Ecological Informatics*. 50: 24-32.
- [11]. Van Con, T., Thang, N. T., Khiem, C. C., Quy, T. H., Lam, V. T., Van Do, T. & Sato, T. (2013). Relationship between aboveground biomass and measures of structure and species diversity in tropical forests of Vietnam. *Forest Ecology and management*. 310: 213-218.
- [12]. Nam, V. T., Van Kuijk, M. & Anten, N. P. (2016). Allometric equations for aboveground and belowground biomass estimations in an evergreen forest in Vietnam. *PloS one*. 11(6): e0156827.
- [13]. Jie, S., Zhao, J., He, P. L., Cheng, Y. J., Huang, R. X. & Zhu, W. K. (2024). Single Tree Biomass Prediction Model of Eucalyptus Plantations Based on Random Forest Algorithm. *Eucalypt Science & Technology*. 41(2): 11-16.
- [14]. Bi, H., Long, Y., Turner, J., Lei, Y., Snowdon, P., Li, Y., Harper, R., Zerihun, A. & Ximenes, F. (2010). Additive prediction of aboveground biomass for Pinus radiata (D. Don) plantations. *Forest Ecology and Management*. 259(12): 2301-2314.
- [15]. Breiman, L. (2001). Random forests. *Machine learning*. 45: 5-32.
- [16]. Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*. 13(1): 1063-1095.
- [17]. Hu, J. & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in bioinformatics*. 24(2): bbad002.
- [18]. Miranda, E. N., Barbosa, B. H. G., Silva, S. H. G., Monti, C. A. U., Tng, D. Y. P. & Gomide, L. R. (2022). Variable selection for estimating individual tree height using genetic algorithm and random forest. *Forest Ecology and Management*. 504: 119828.
- [19]. Nguyễn Văn Bích, Hà Thị Mai & Võ Đại Hải (2022). Sinh khối rừng trồng Keo lai theo tuổi và cấp đất tại Yên Bái. *Tạp chí Khoa học Lâm nghiệp*. 4: 70-78.
- [20]. Trần Quang Bảo & Võ Thành Phúc (2019). Nghiên cứu sinh khối và khả năng hấp thụ CO₂ của rừng trồng keo lai tại tỉnh Bà Rịa Vũng Tàu. *Tạp chí Khoa học và Công nghệ Lâm nghiệp*. 2: 69-75.
- [21]. Trần Thị Ngoan & Nguyễn Tấn Chung (2018). Sinh khối trên mặt đất đối với rừng trồng Keo lai (Acacia auriculiformis* Acacia mangium) tại tỉnh Đồng Nai. *Tạp chí Khoa học và Công nghệ Lâm nghiệp*. 6: 61-68.
- [22]. Nguyễn Hải Hòa & Nguyễn Hữu An (2016). Ứng dụng ảnh viễn thám Landsat 8 & GIS xây dựng bản đồ sinh khối và trữ lượng cacbon rừng trồng Keo lai (Acacia hybrid) tại huyện Yên Lập, tỉnh Phú Thọ. *Tạp chí Khoa học và Công nghệ Lâm nghiệp*. 4: 70-78.

- [23]. Nguyễn Văn Tuấn, Nguyễn Văn Thịnh, Phạm Tiến Dũng, Nguyễn Huy Hoàng, Nguyễn Việt Cường, Nguyễn Thị Thu Phương & Trần Hồng Vân (2024). Xác định sinh khối và khả năng tích lũy carbon của rừng trồng thuần loài Keo lai tại Khu Dự trữ Sinh quyển Đồng Nai. *Tạp chí Khoa học Lâm nghiệp*. 1: 46-53.
- [24]. Cục Thống kê tỉnh Đồng Nai (2023). Niên giám thống kê tỉnh Đồng Nai năm 2022. NXB Thống kê.
- [25]. Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R. & Friedman, J. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction*. 11: 587-604.
- [26]. Ohsowski, B. M., Dunfield, K. E., Klironomos, J. N. & Hart, M. M. (2016). Improving plant biomass estimation in the field using partial least squares regression and ridge regression. *Botany*. 94(7): 501-508.
- [27]. Cheng, Z., Gu, X., Wei, C., Zhou, Z., Zhao, T., Wang, Y., Li, W., Du, Y. & Cai, H. (2024). Monitoring aboveground organs biomass of wheat and maize: A novel model combining ensemble learning and allometric theory. *European Journal of Agronomy*. 161: 127338.
- [28]. Yu, H., Xie, S., Guo, L., Liu, P. & Zhang, P. (2022). Extremely Randomized Trees Estimation of Soil Heavy Metal Content by Fusing Spectra and Spatial Features. *Trans. Chin. Soc. Agric. Mach.* 53: 231-239.
- [29]. Yan, M., Xia, Y., Yang, X., Wu, X., Yang, M., Wang, C., Hou, Y. & Wang, D. (2023). Biomass estimation of subtropical arboreal forest at single tree scale based on feature fusion of airborne LiDAR data and aerial images. *Sustainability*. 15(2): 1676.
- [30]. Gleason, C. J. & Im, J. (2012). Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment*. 125: 80-91.
- [31]. Wongchai, W., Onsree, T., Sukkam, N., Promwungkwa, A. & Tippayawong, N. (2022). Machine learning models for estimating above ground biomass of fast growing trees. *Expert Systems with Applications*. 199: 117186.
- [32]. Luo, M., Anees, S. A., Huang, Q., Qin, X., Qin, Z., Fan, J., Han, G., Zhang, L. & Shafri, H. Z. M. (2024). Improving forest above-ground biomass estimation by integrating individual machine learning models. *Forests*. 15(6): 975.
- [33]. Nsiah, E. O. (2023), A Comparison of Variant Methods in Random Forest with Multicollinearity Data for Classification Prediction Modeling, Southeast Missouri State University.
- [34]. Yadav, S., Padalia, H., Sinha, S. K., Srinet, R. & Chauhan, P. (2021). Above-ground biomass estimation of Indian tropical forests using X band Pol-InSAR and Random Forest. *Remote Sensing Applications: Society and Environment*. 21: 100462.
- [35]. Singh, C., Karan, S. K., Sardar, P. & Samadder, S. R. (2022). Remote sensing-based biomass estimation of dry deciduous tropical forest using machine learning and ensemble analysis. *Journal of Environmental Management*. 308: 114639.
- [36]. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*. 47(1): 31-39.
- [37]. Lövenstein, H. M. & Berliner, P. R. (1993). Biometric relationships for non-destructive above ground biomass estimations in young plantations of *Acacia salicina* Lindl. and *Eucalyptus occidentalis* Endl. *New forests*. 7: 255-273.
- [38]. de Oliveira, B. R., da Silva, A. A. P., Teodoro, L. P. R., de Azevedo, G. B., Azevedo, G. T. d. O. S., Baio, F. H. R., Sobrinho, R. L., da Silva Junior, C. A. & Teodoro, P. E. (2021). Eucalyptus growth recognition using machine learning methods and spectral variables. *Forest Ecology and Management*. 497: 119496.
- [39]. Liu, Y. L., Wang, Y. R. & Zhang, J. (2012), New machine learning algorithm: Random forest, *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, Springer. 246-252.
- [40]. Roy, P. & Ravan, S. A. (1996). Biomass estimation using satellite remote sensing data—an investigation on possible approaches for natural forest. *Journal of biosciences*. 21: 535-561.
- [41]. Manly, B. F. (2018). Randomization, bootstrap and Monte Carlo methods in biology. ed. chapman and hall/CRC.
- [42]. Yang, M., Zhou, X., Liu, Z., Li, P., Liu, C., Huang, H., Tang, J., Zhang, C., Zou, Z. & Xie, B. (2024). Dynamic carbon allocation trade - off: A robust approach to model tree biomass allometry. *Methods in Ecology and Evolution*. 15(5): 886-899.
- [43]. Vanclay, J. K. (1995). Synthesis: Growth models for tropical forests: A synthesis of models and methods. *Forest science*. 41(1): 7-42.
- [44]. Satoo, T. & Madgwick, H. (2012). Forest biomass. *Springer Science & Business Media*. 6.