PHƯƠNG PHÁP THẨM ĐỊNH CHÉO MÔ HÌNH SINH KHỐI CÂY RỪNG TRÊN MẶT ĐẤT

Bảo Huy[1]

TÓM TẮT

Các mô hình được sử dụng để ước tính sinh khối và báo cáo CO_2 tương đương từ rừng trong khuôn khổ chương trình REDD+ (Giảm phát thải từ mất rừng và suy thoái rừng) cần chỉ ra độ tin cậy và sai số, vì vậy các mô hình sinh khối lựa chọn được đánh giá khả năng dự đoán trên cơ sở sử dụng 110 cây mẫu chặt hạ ở rừng lá rộng thường xanh vùng duyên hải Nam Trung bộ. Các hàm mũ khác nhau sử dụng các biến độc lập như đường kính ngang ngực (DBH), chiều cao cây (H), khối lượng thể tích gỗ (WD) và diện tích tán lá (CA) để dự đoán sinh khối cây rừng phần trên mặt đất (AGB) được đánh giá. Bốn phương pháp thẩm định chéo được áp dụng là sử dụng dữ liệu độc lập, Leave-One-Out (LOOCV), k fold và Monte Carlo. Trong những phương pháp này, Monte Carlo là thích hợp để cung cấp chỉ tiêu thống kê của mô hình, sai số qua thẩm định chéo ổn định, chính xác và có phân bố chuẩn. Các chỉ tiêu thống kê thẩm định chéo Monte Carlo như chênh lệch (Bias%), sai số trung phương (RMPE%) và sai số tuyệt đối % (MAPE) được tính bằng cách phân chia ngẫu nhiên bộ dữ liệu 200 lần, mỗi lần có 80% dữ liệu dùng lập mô hình và 20% dữ liệu để thẩm định chéo, sau đó các chỉ tiêu thống kê của mô hình và sai số được tính trung bình từ 200 lần rút mẫu. Mô hình tốt nhất được lựa chọn dựa vào hệ số xác định (R^2), chỉ tiêu AIC (Akaike information criterion). AGB có quan hệ chặt chẽ với bốn biến đầu vào là DBH, H, WD và CA theo mô hình tốt nhất là $AGB = a \times (DBH^2 HWD)^b \times CA^c$ với sai số MAPE thấp nhất là 17,9%.

Từ khóa: K-fold, LOOCV, Monte Carlo, mô hình sinh khối AGB, thẩm định chéo.

1. ĐẶT VẤN ĐỀ

Khi áp dụng hệ thống các mô hình sinh khối cây rừng để ước tính và báo cáo hấp thụ hoặc phát thải CO_2 tương đương từ rừng thông qua hệ thống "Đo lường – Báo cáo – Thẩm định (MRV)" trong chương trình "Giảm phát thải từ mất rừng và suy thoái rừng" của Liên hiệp quốc (UN-REDD+), cần chỉ ra sai số của các mô hình này. Các phương pháp thẩm định chéo (Cross Validation) là cơ sở để thẩm định và báo cáo sai số của các mô hình sinh khối.

Trong thiết lập các mô hình sinh khối-các bon rừng, việc lựa chọn mô hình tối ưu và cung cấp thông tin sai số của mô hình một cách chính xác là một nội dung quan trọng. Từ đây đã hình thành một lĩnh vực trong khoa học sinh trắc là "Thẩm định chéo – Cross Validation". Moore (2017), Zhang (1997) đã chỉ ra rằng thẩm định chéo các mô hình giúp tránh lựa chọn các mô hình có sai lệch lớn so với thực tế (overfitting). Picard và Cook (1984) cũng cho thấy thẩm định chéo ngoài việc xác định sai số, còn tránh cho mô hình dự đoán sai lệch với thực tế thì nó còn giúp cho việc lựa chọn các biến số thích hợp cho mô hình.

Nghiên cứu này giới thiệu kết quả thử nghiệm để lựa chọn phương pháp thẩm định chéo thích hợp cho các mô hình ước tính sinh khối cây rừng trên mặt đất (AGB). Bốn phương pháp thẩm định mô hình được thử nghiệm là: i) Sử dụng dữ liệu độc lập, ii) Leave-One-Out Cross Validation (LOOCV), iii) k-fold và iv) Monte Carlo.

2.VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Vùng sinh thái và kiểu rừng nghiên cứu

Khu vực tiến hành thu thập số liệu để lập và thẩm định các mô hình sinh khối là các khu rừng lá rộng thường xanh của vùng duyên hải Nam Trung bộ. Các ô mẫu được đặt tại tỉnh Quảng Nam (có tọa độ địa lý: 15°28'13.3" N đến 15°28'16.1" N và 107°48'56.6" E đến 107°48'59.6" E), ở độ cao 574-624 m so với mặt nước biển, độ dốc từ 10 - 40 độ. Đất có màu vàng nâu, phát triển trên phù sa cổ, với giá trị pH = 6,0-6,3 và tầng đất sâu hơn 100 cm.

Lượng mưa trung bình hàng năm là 3150-3500 mm với mức tối thiểu và lượng mưa tối đa 1.857 mm và 5.337 mm tương ứng.

Nhiệt độ trung bình hàng năm là 21,8°C, biến động giữa 16,0°C và 39,4°C. Có hai mùa rõ rệt: mùa khô từ tháng hai đến tháng tám và mùa mưa từ tháng chín đến tháng giêng.

[1] Trường Đại học Tây Nguyên

Độ ẩm không khí trung bình là 90% và có lượng thoát hơi nước trung bình năm là 800 mm và sương mù thường xảy ra từ tháng mười một - tháng hai (Nguồn: Trung tâm Khí tượng Thủy văn ở miền Trung Việt Nam, 2012)

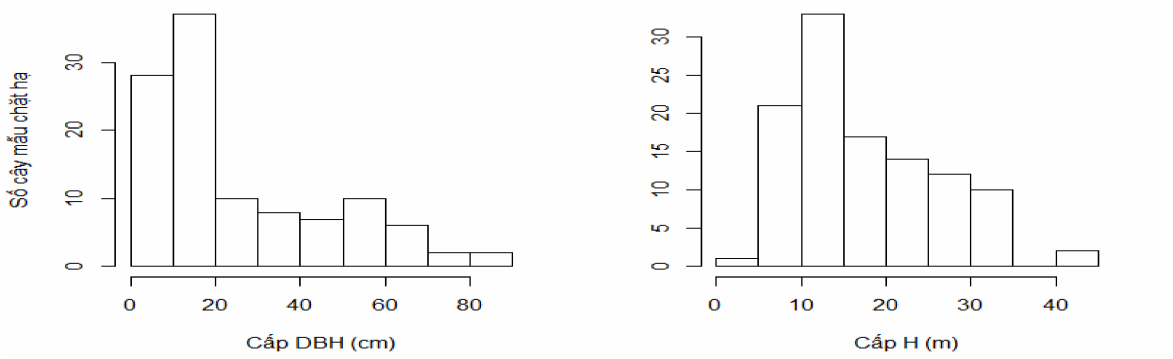
2.2. Lập ô mẫu, chọn cây mẫu, thu thập và xử lý số liệu

Đã tiến hành lập hai ô mẫu mỗi ô 1 ha (100 m × 100 m), được chia thành 100 ô phụ 10 m × 10 m. Trong các ô, các chỉ tiêu được thu thập: (i) vị trí tọa độ ô mẫu; (ii) thông tin lâm phần: kiểu rừng và trạng thái, độ tàn che, số tầng rừng và tiết diện ngang (BA); (iii) địa hình: độ dốc và vị trí địa hình; (iv) đặc điểm của đất: pH, độ sâu và màu sắc; và (v) đo cây đứng: tên loài (địa phương và khoa học), đường kính ngang ngực (DBH, cm), và chiều cao cây (H, m) của tất cả các cây có DBH ≥ 5 cm. Trong ô mẫu, số lượng cây lấy mẫu được xác định bằng tỷ lệ số cây theo từng cấp đường kính, với cự ly cấp kính là 10 cm, riêng đối với các cấp đường kính lớn (ví dụ, đường kính ≥ 45 cm) thì có ít nhất ba cây được lấy mẫu. Các cây mẫu cũng đã được lựa chọn cũng dựa trên tỷ lệ loài ưu thế trong lâm phần, gồm các loài chính như dẻ (*Lithocarpus annamensis* (Hickel & A.Camus) Barnett), trâm (*Syzygium levinei* (Merr.) Merr.), trám (*Canarium littorale* Blume), sổ (*Dillenia indica var. aurea* (Sm.) Kuntze), nhọc (*Polyalthia nemoralis* Aug.DC.), thị rừng (*Diospyros decandra* Lour.), giổi (*Aglaia roxburghiana* (Wight & Arn.) Miq.).

Tổng cộng có 110 cây mẫu (55 cây ở mỗi ô) đã được chặt hạ để đo sinh khối và lấy mẫu. Phân bố DBH và H của cây mẫu thể hiện trong hình 1 cho thấy cây mẫu chặt hạ có phân bố đồng dạng với phân bố của lâm phần. Bảng 1 trình bày tóm tắt thông tin về các biến số của các cây mẫu chặt hạ, cho thấy dữ liệu thu thập bao gồm một dải rộng giá trị DBH và H.

Bảng 1. Tóm tắt thông tin về biến số độc lập và phụ thuộc trên các cây mẫu chặt hạ

Thông tin	DBH (cm)	H (m)	WD (g/cm^3)	CA (m^2)	AGB (kg)
Nhỏ nhất	4,9	4,7	0,430	0,79	5,9
Trung bình	25,7	17,5	0,586	24,53	804,4
Lớn nhất	87,7	41,4	0,712	201,06	8.633,0
Sai tiêu chuẩn	21,2	8,6	0,052	31,6	1.482,4



Hình 1. Phân bố DBH và H của số cây mẫu chặt hạ

Trước khi chặt hạ cây mẫu, tiến hành đo DBH (cm), H (m), đo đường kính tán (CD, m) theo hai hướng bắc nam và đông tây; xác định loài của mỗi cây mẫu. Chiều cao cây được đo lại sau khi cây mẫu đã được chặt. Khối lượng sinh khối tươi của các bộ phận cây như lá, cành và thân cây có vỏ được cân và ghi chép. Thân cây mẫu được phân thành năm đoạn có chiều dài bằng nhau và đường kính có và không có vỏ của cây tại vị trí 5 đoạn được đo đạc. Các mẫu của bốn thành phần sinh khối của cây đã được đưa đến phòng thí nghiệm để tính tỷ lệ khối lượng khô/tươi, và khối lượng thể tích gỗ (WD, g/cm^3). Mẫu gỗ và vỏ cây được lấy 500 g và 300 g và được lấy 5 mẫu ở 5 đoạn trên thân cây. Mẫu của cành là 500 g và thu thập 3 mẫu ở ba vị trí trên cành (lớn, trung bình và nhỏ). Mẫu lá là 300 g bao gồm lá già và non.

Trong phòng thí nghiệm, thể tích tươi của mẫu gỗ và vỏ cây được xác định bằng phương pháp nước

chuyển chỗ trong ống nghiệm. Tất cả các mẫu đều được chẻ nhỏ và sấy khô ở 105°C cho đến khi khối lượng không đổi. WD (g/cm^3) của mẫu được lấy bằng tỷ số giữa khối lượng khô và thể tích tươi của mỗi mẫu. WD của cây mẫu gỗ sau cùng là trung bình lấy từ năm phân đoạn. Khối lượng thể tích của vỏ cây cũng được tính tương tự. Sinh khối khô của mỗi thành phần cây đã được tính toán theo khối lượng tươi của nó nhân với tỷ lệ tươi/khô. Sinh khối của cây rừng trên mặt đất (AGB, kg) là tổng sinh khối của thân cây (Bst), sinh khối của cành nhánh (Bbr), sinh khối lá (Bl), sinh khối của vỏ cây (Bba) và sinh khối của gốc (Bstu). Biến diện tích tán lá được tính CA (m^2) = $\pi \frac{cd^2}{8}$

2.3. Phương pháp thiết lập, lựa chọn các mô hình sinh khối

Các mô hình sinh khối cây rừng được thiết lập trong nghiên cứu này sử dụng hàm power dựa theo Brown (1997), Chave *et al.* (2005, 2014).

Áp dụng phương pháp phi tuyến Maximum Likelihood có trọng số để ước lượng mô hình power. Sử dụng chương trình nlme chạy trong phần mềm mã nguồn mở R (Bates *et al.,* 2010; Pinheiro *et al.,* 2014) và chẩn đoán qua sơ đồ sử dụng code ggplot2 (Wickham *et al.,* 2013).

Mô hình sinh khối tổng quát như sau (Huy *et al.,* 2016a,b):

$$Y_i = \alpha \times X_i^{\beta} + \varepsilon_i \quad (1)$$

$$\varepsilon_i \sim iid\ N(0,\sigma^2) \quad (2)$$

Trong đó Y_i là AGB (kg) ứng với cây thứ *i*; α và β là tham số của mô hình; X_i là các biến số DBH (cm), H (m), WD (g/cm^3), CA (m^2) của cây thứ *i* hoặc tổ hợp biến DBH^{2}H (m^3) = (DBH/100)^2×H đại diện cho thể tích cây; hoặc tổ hợp biến DBH^{2}HWD (kg) = DBH^{2}H×WD×1000 đại diện cho sinh khối thân cây gỗ.

Một hàm phương sai theo trọng số đã được áp dụng để điều chỉnh các tham số của mô hình nhằm giảm biến động sai số của các mô hình sinh khối. Hàm phương sai có dạng như sau (Huy *et al.,* 2016a):

$$Var(\varepsilon_i) = \sigma^2 (v_i)^{2k} \quad (3)$$

Trong đó ε_i là sai số ngẫu nhiên; σ^2 là sai số bình phương; v_i là biến trọng số Weight (*1/DBH, 1/DBH^{2}HWD*) tương ứng với cây thứ i; và k là hệ số của hàm phương sai.

Mô hình tốt nhất được lựa chọn dựa vào chỉ tiêu AIC (Akaike Information Criterion), AIC càng bé thì mô hình càng có độ tin cậy cao hơn.

$$AIC = -2\ln(L) + 2p \quad (4)$$

Trong đó L là Likelihood của mô hình, p là tổng số tham số của mô hình. Ngoài ra hệ số xác định R^2_{adj} cũng được sử dụng phối hợp với AIC để lựa chọn mô hình.

2.4. Phương pháp thẩm định chéo (Cross Validation) các mô hình

2.4.1. Phương pháp sử dụng dữ liệu độc lập

Đây là phương pháp truyền thống, sử dụng một bộ dữ liệu độc lập để đánh giá sai số của mô hình đã thiết lập. Phân chia dữ liệu ngẫu nhiên làm hai phần: 80% cho lập mô hình và 20% cho đánh giá sai số của các mô hình.

Các sai số của các mô hình được tính toán bao gồm % sai lệch giữa quan sát và dự báo qua mô hình (Bias %), sai số trung phương trung bình % (Root Mean Square Error - RMSE %), và sai số tuyệt đối trung bình % (Mean Absolute Percent Error - MAPE) (Mayer *et al.,* 1993; Chave *et al.*, 2005; Basuki *et al.,* 2009; Swanson *et al.,* 2011; Huy *et al.,* 2016a,b):

$$Bias\ \% = \frac{100}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)}{(y_i)} \quad (5)$$

$$RMSE\ \% = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (6)$$

$$MAPE\ \% = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (7)$$

Trong đó n là số cây mẫu độc lập dùng đánh giá; và y_i và $\hat{y}_i$ là giá trị quan sát và ước tính qua mô hình.

2.4.2. Phương pháp thẩm định chéo Leave-One-Out Cross Validation (LOOCV)

Từ n dữ liệu cây mẫu, phương pháp LOOCV sử dụng n-1 dữ liệu lập mô hình và 1 dữ liệu độc lập dùng để đánh giá sai số, lặp lại như vậy với n lần lập mô hình và đánh giá sai số, với sai số mỗi lần được tính từ một dữ liệu độc lập không tham gia lập mô hình, sau đó lấy trung bình (Moore, 2017).

Cách tính các sai số tương đối khi áp dụng LOOCV:

$$Bias\ (\%) = \frac{100}{n}\sum_{i=1}^{L}\frac{y_i - \hat{y}_i}{y_i} \quad (8)$$

$$RMSE\ (\%) = 100\sqrt{\frac{1}{n}\sum_{i=1}^{L}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (9)$$

$$MAPE\ (\%) = \frac{100}{n}\sum_{i=1}^{L}\frac{|y_i - \hat{y}_i|}{y_i} \quad (10)$$

Trong đó, L là số lần lặp lại tính sai số, mỗi lần sử dụng một dữ liệu độc lập để tính sai số mô hình (L=n dữ liệu); y_i và $\hat{y}_i$ là giá trị quan sát và dự đoán qua mô hình.

2.4.3. Phương pháp thẩm định chéo k-fold

Phương pháp này phân chia dữ liệu thành k phần bằng nhau (k-fold), một phần dữ liệu không tham gia lập mô hình dùng để đánh giá sai số, trong khi đó k-1 phần dữ liệu dùng lập mô hình. Tiến hành lập lại như vậy k = 10 lần, mỗi lần lấy một phần dữ liệu khác nhau để thẩm định mô hình và tính sai số trung bình từ k lần lặp (Moore, 2017). Cách tính các sai số tương đối theo phương pháp k-fold như sau:

$$Bias\ (\%) = \frac{1}{k}\sum_{k=1}^{k}\frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \quad (11)$$

$$RMSE\ (\%) = \frac{1}{k}\sum_{k=1}^{k}100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (12)$$

$$MAPE\ (\%) = \frac{1}{k}\sum_{k=1}^{k}\frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (13)$$

Trong đó, k là số phần dữ liệu bằng nhau được phân chia (k-fold), với k = 10; n là số dữ liệu đánh giá của mỗi lần và y_i và $\hat{y}_i$ là giá trị quan sát và dự đoán qua mô hình

2.4.4. Phương pháp Monte Carlo

Phương pháp này dùng để thẩm định chéo các mô hình sinh khối được mô tả như sau: phân chia dữ liệu ngẫu nhiên làm 2 phần, một phần dùng để lập mô hình (80% dữ liệu) và một phần dùng để đánh giá sai số (20% dữ liệu). Mỗi lần như vậy tính toán các chỉ tiêu thống kê đánh giá, so sánh các mô hình như AIC, R^2 và các sai số như Bias%, RMSE%, MAPE%. Tiến hành lặp lại như vậy R lần để thẩm định các mô hình và đánh giá sai số, cuối cùng giá trị thống kê so sánh các mô hình và sai số được tính trung bình từ R = 200 lần (Temesgen *et al.*, 2014) và Huy *et al.*, 2016a,b). Ngoài ra cũng thử nghiệm R khác nhau (50, 100, 200, 300 và 500 lần) để chỉ ra số lần lặp cho sai số của mô hình ổn định và có phân bố tần số tiệm cận chuẩn. Các sai số áp dụng theo phương pháp thẩm định chéo Monte Carlo với R lần lặp lại ngẫu nhiên như sau (Swanson *et al.*, 2011; Huy *et al.*, 2016a,b):

$$Bias\ (\%) = \frac{1}{R}\sum_{r=1}^{R}\frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \quad (14)$$

$$RMSE\ (\%) = \frac{1}{R}\sum_{r=1}^{R}100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (15)$$

$$MAPE\ (\%) = \frac{1}{R}\sum_{r=1}^{R}\frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (16)$$

Trong đó, R là số lần phân chia dữ liệu ngẫu nhiên thành hai phần, n là số dữ liệu đánh giá của mỗi lần rút mẫu (20% mẫu rút ngẫu nhiên) và y_i và $\hat{y}_i$ là giá trị quan sát và dự đoán qua mô hình.

Tất cả các phương pháp thẩm định chéo các mô hình sinh khối được viết Code và chạy trong phần mềm mã nguồn mở R. Cuối cùng, sau khi lựa chọn dạng mô hình, thẩm định chéo và xác định được các sai số của mô hình lựa chọn, tham số của mô hình lựa chọn được thiết lập dựa vào toàn bộ dữ liệu.

3. KẾT QUẢ VÀ THẢO LUẬN

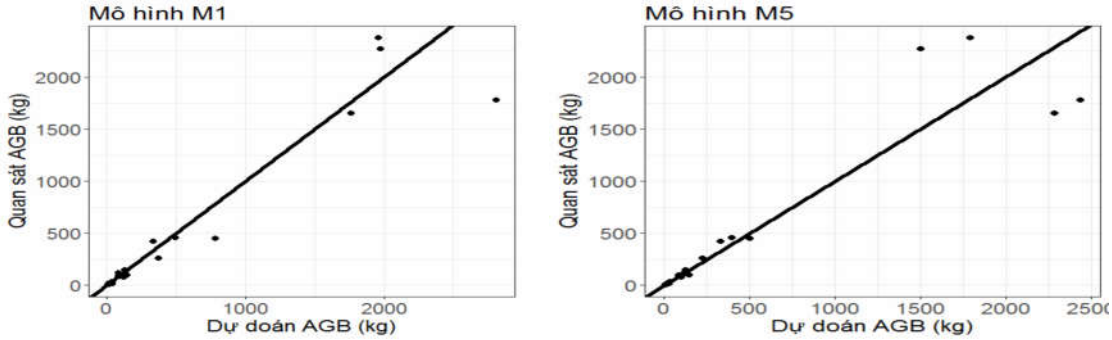
3.1. Kết quả so sánh và thẩm định sai số các mô hình sinh khối theo phương pháp sử dụng dữ liệu độc lập

Kết quả áp dụng phương pháp sử dụng dữ liệu độc lập để lập và thẩm định sai số các mô hình ước tính AGB với các biến độc lập khác nhau trình bày trong bảng 2.

Bảng 2. So sánh và thẩm định sai số của các mô hình sinh khối theo phương pháp sử dụng dữ liệu độc lập

Mã mô hình	Dạng mô hình	AIC	R^2_{adj}	Bias (%)	RMSE (%)	MAPE (%)
M1	$AGB = a \times DBH^b$	895	0,937	-14,7	44,7	30,8
M2	$AGB = a \times (DBH^2H)^b$	892	0,943	-5,43	31,2	26,2
M3	$AGB = a \times DBH^b WD$	878	0,947	-11,8	40,0	24,8
M4	$AGB = a \times (DBH^2HWD)^b$	887	0,957	-6,0	26,0	21,9
M5	$AGB = a \times (DBH^2HWD)^b \times CA^c$	877	0,965	-2,0	23,7	18,7

Ghi chú: R^2_{adj}, AIC được tính từ 80% dữ liệu độc lập để lập mô hình; các sai số Bias, RMSE, MAPE được tính từ 20% dữ liệu đánh giá được rút ngẫu nhiên và độc lập với dữ liệu lập mô hình.



Hình 2. Đồ thị quan hệ giữa giá trị AGB dự đoán qua mô hình với AGB quan sát của 20% dữ liệu rút ngẫu nhiên thẩm định độc lập. Trái: $AGB = a \times DBH^b$; phải: $AGB = a \times (DBH^2 HWD)^b CA^c$

Các giá trị dự đoán AGB từ các mô hình so với AGB quan sát của 20% dữ liệu độc lập dùng đánh giá sai số thể hiện ở Hình 2.

Kết quả này cho thấy khi tăng số biến số từ một biến DBH lên lần lượt đến bốn biến số DBH, H, WD và CA thì độ tin cậy của mô hình càng cao, AIC và các sai số đều giảm dần. Sai số MAPE giảm 12 % từ một biến lên bốn biến.

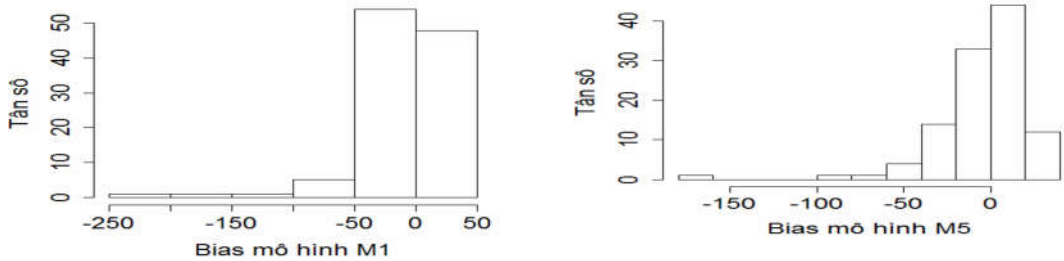
Phương pháp thẩm định sai số truyền thống sử dụng dữ liệu độc lập để so sánh và thẩm định sai số mô hình có hạn chế là sai số chỉ được xác định một lần cho một bộ dữ liệu độc lập nhất định, vì vậy sai số có thể khác đi nếu áp dụng theo một bộ dữ liệu độc lập khác. Do đó nó thường không cung cấp chính xác sai số trong mọi trường hợp ứng dụng. Vì vậy các phương pháp thẩm định chéo các mô hình cần được xem xét áp dụng để cung cấp thông tin sai số ổn định của các mô hình ước tính sinh khối.

3.2. Kết quả so sánh và thẩm định chéo sai số các mô hình sinh khối theo phương pháp LOOCV

Bảng 3. So sánh và thẩm định chéo LOOCV các mô hình sinh khối

Mã mô hình	Dạng mô hình	AIC	R^2_{adj}	Bias (%)	RMSE (%)	MAPE (%)
M1	$AGB = a \times DBH^b$	1109	0,934	-7,8	22,9	22,9
M2	$AGB = a \times (DBH^2 H)^b$	1080	0,952	-3,1	19,1	19,1
M3	$AGB = a \times DBH^b WD$	1084	0,946	-6,9	20,0	20,0
M4	$AGB = a \times (DBH^2 HWD)^b$	1089	0,953	-5,9	19,7	19,7
M5	$AGB = a \times (DBH^2 HWD)^b \times CA^c$	1074	0,960	-4,8	17,7	17,7

Ghi chú: R^2, AIC được tính từ n-1 dữ liệu độc lập; các sai số Bias, RMSE, MAPE được tính trung bình n lần từ một dữ liệu rút độc lập.



Hình 3. Phân bố tần số Bias của hai mô hình AGB theo phương pháp LOOCV

Các kết quả minh họa cho lập và thẩm định các mô hình AGB theo phương pháp LOOCV được tổng hợp trong bảng 3. Kết quả này cho thấy mô hình bốn biến (DBH^2HWD và CA) có độ tin cậy cao nhất (AIC bé nhất hơn và R^2 cao nhất) và các sai số đều nhỏ hơn các mô hình AGB chỉ với một biến số DBH hoặc hai biến số DBH^2H hoặc DBH^bWD. Với phương pháp này MAPE chỉ giảm 5% khi đi từ một biến số lên 4 biến số độc lập và có sự khác biệt với phương pháp đánh giá sử dụng dữ liệu độc lập ở trên đây.

Hình 3 cho thấy phân bố Bias của hai mô hình M1 và M5 được thẩm định theo phương pháp LOOCV có xu hướng lệch phải và chưa tiệm cận chuẩn. Đây là nhược điểm của phương pháp thẩm định chéo LOOCV, do chỉ tính sai số của từng cá thể trong một lần thẩm định, trong khi đó trong thực tế để sai số tiệm cận chuẩn thì mỗi lần rút mẫu đánh giá cần có số mẫu đủ lớn. Điều này cũng là hạn chế của phương pháp LOOCV trong ứng dụng, vì trong thực tế sai số không tính cho từng cá thể mà cho một ô mẫu, hoặc lâm phần.

3.3. Kết quả so sánh và thẩm định chéo sai số các mô hình sinh khối theo phương pháp k-fold

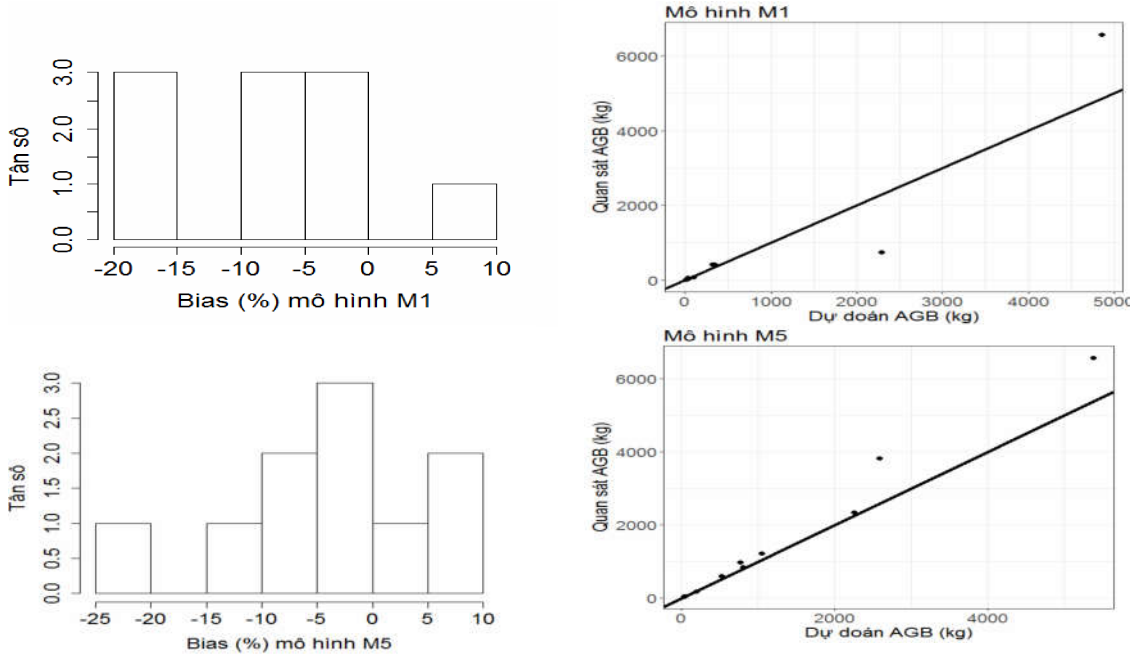
Các kết quả minh họa cho lập và thẩm định chéo các mô hình AGB theo phương pháp k-fold được tổng hợp trong bảng 4.

Kết quả áp dụng phương pháp thẩm định chéo k-fold là đồng nhất với LOOCV ở trên, có nghĩa mô hình có bốn biến số có độ tin cậy cao nhất và sai số bé nhất, biến động MAPE cũng tương đồng như khi áp dụng phương pháp LOOCV.

Bảng 4. So sánh và thẩm định chéo k-fold các mô hình sinh khối

Mã mô hình	Dạng mô hình	AIC	R^2_{adj}	Bias (%)	RMSE (%)	MAPE (%)
M1	$AGB = a \times DBH^b$	1008	0,933	-7,9	33,1	22,9
M2	$AGB = a \times (DBH^2 H)^b$	1001	0,934	-4,0	28,0	21,5
M3	$AGB = a \times DBH^b WD$	985	0,945	-6,8	29,3	20,0
M4	$AGB = a \times (DBH^2 HWD)^b$	990	0,953	-5,7	26,6	19,7
M5	$AGB = a \times (DBH^2 HWD)^b ăCA^c$	978	0,960	-4,7	24,4	17,6

Ghi chú: R^2, AIC được tính từ k-1 phần dữ liệu độc lập; các sai số Bias, RMSE, MAPE được tính trung bình k = 10 lần.



Hình 4. Phân bố tần số Bias (trái) và giá trị dự đoán qua mô hình so với dữ liệu đánh giá độc lập (phải) của phương pháp k-fold (k=10) cho hai mô hình M1 và M5

Hình 4 chi ra phân bố Bias của hai mô hình được thẩm định chéo theo phương pháp k-fold có nhấp nhô nhiều đỉnh và chưa tiệm cận chuẩn.

Đây là nhược điểm của phương pháp thẩm định chéo k-fold, do số lần lại khá nhỏ (k = 10).

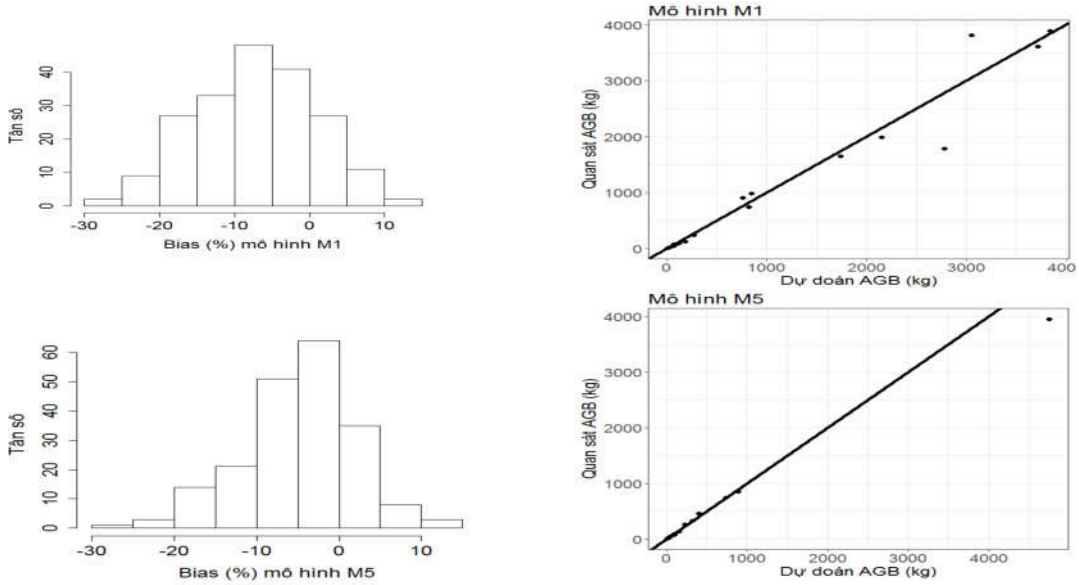
3.4. Kết quả so sánh và thẩm định chéo sai số các mô hình sinh khối theo phương pháp Monte Carlo

Các kết quả minh họa cho lập và thẩm định chéo các mô hình AGB theo phương pháp Monte Carlo được tổng hợp trong bảng 5.

Bảng 5. So sánh và thẩm định chéo các mô hình sinh khối theo phương pháp Monte Carlo

Mã mô hình	Dạng mô hình	AIC	R^2_{adj}	Bias (%)	RMSE (%)	MAPE (%)
M1	$AGB = a \times DBH^b$	899	0,933	-7,0	33,1	22,1
M2	$AGB = a \times (DBH^2H)^b$	892	0,936	-4,1	28,7	21,1
M3	$AGB = a \times DBH^b WD$	876	0,944	-7,1	30,9	19,9
M4	$AGB = a \times (DBH^2HWD)^b$	881	0,952	-6,0	27,4	19,7
M5	$AGB = a \times (DBH^2HWD)^b \ăCA^c$	872	0,959	-4,8	26,0	17,9

Ghi chú: R^2, AIC được tính từ 80% dữ liệu rút ngẫu nhiên; các sai số Bias, RMSE, MAPE được tính từ 20% dữ liệu đánh giá được rút ngẫu nhiên, độc lập và tính trung bình từ 200 lần lặp lại



Hình 5. Phân bố tần số Bias (trái) và giá trị dự đoán qua mô hình so với dữ liệu đánh giá độc lập (phải) theo phương pháp Monte Carlo của hai mô hình M1 và M5

Tương tự các phương pháp trên, mô hình tổ hợp ba biến DBH^2HWD cùng với biến CA có độ tin cậy cao nhất (AIC bé nhất và R^2 cao nhất) và các sai số đều nhỏ hơn so với mô hình ít biến số độc lập hơn. Kết quả sai số khi áp dụng phương pháp thẩm định chéo Monte Carlo là khá đồng nhất với các phương pháp LOOCV và k-fold đã giới thiệu ở trên. Hình 5 cho thấy phân bố Bias của hai mô hình được thẩm định chéo theo phương pháp Monte Carlo với 200 lần lặp lại đã tiệm cận chuẩn; đặc biệt là mô hình có ba biến số tổ hợp DBH^2HWD cùng với biến CA. Vì vậy phương pháp Monte Carlo có thể xem đã cung cấp sai số ổn định và khách quan của mô hình ước tính sinh khối so với các phương pháp thẩm định chéo khác nói ở trên.

Tiến hành tổng hợp kết quả thẩm định chéo sai số mô hình sinh khối tốt nhất $AGB = a \times (DBH^2HWD)^b \times CA^c$ theo bốn phương pháp khác nhau ở bảng 6.

Bảng 6. Tổng hợp kết quả thẩm định chéo mô hình lựa chọn $AGB = a \times (DBH^2HWD)^b \times CA^c$ theo các phương pháp khác nhau

	Phương pháp thẩm định mô hình			
Chỉ tiêu thống kê, sai số	Dữ liệu độc lập	LOOCV	k-fold	Monte Carlo
AIC	877	1074	978	872
R^2_{adj}	0,965	0,960	0,960	0,959
Bias (%)	-2,0	-4,8	-4,7	-4,8
RMSE (%)	23,7	17,7	24,4	26,0
MAPE (%)	18,7	17,7	17,6	17,9

Từ bảng 6 cho thấy nếu lấy kết quả theo Monte Carlo làm chuẩn (vì có sai số ổn định và phân bố chuẩn), thì sai số cung cấp theo phương pháp k-fold là khá tương đồng, tuy nhiên k-fold cho sai số chưa có phân bố chuẩn. Trong khi đó hai phương pháp

dùng dữ liệu độc lập hoặc LOOCV có sai lệch sai số RMSE khá lớn so với phương pháp Monte Carlo và có phân bố sai số sai lệch chuẩn. Vì vậy phương pháp Monte Carlo dùng thẩm định chéo các mô hình sẽ cung cấp sai số ổn định, khách quan khi số lần lặp đủ lớn là 200 lần. Đã thử nghiệm thay đổi số lần lặp lại trong phương pháp Monte Carlo để thẩm định chéo mô hình tốt nhất $AGB = a \times (DBH^2HWD)^b \times CA^c$ được lập theo phương pháp phi tuyến Maximum Likelihood có trọng số; số lần lặp R thay đổi từ 50, 100, 200, 300 và 500. Kết quả ở bảng 7 cho thấy với R = 50 trở lên thì các chỉ tiêu thống kê của mô hình (AIC, $R^2_{adj.}$) và các sai số Bias, RMSE, MAPE đã ổn định, không có sự khác biệt khi R tăng đến 500 lần.

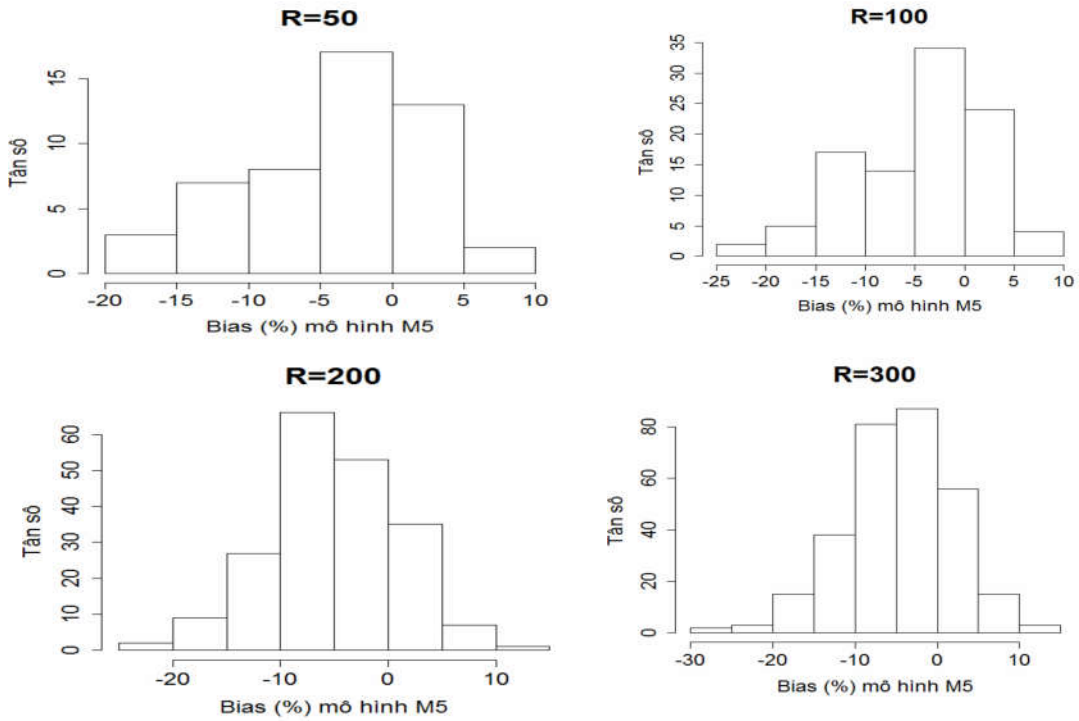
Tuy nhiên xét thêm phân bố của Bias ở Hình 6 thì với R=50 và 100 phân bố có nhiều đỉnh, chỉ khi R ≥ 200 lần dạng phân bố của Bias mới tiệm cận chuẩn.

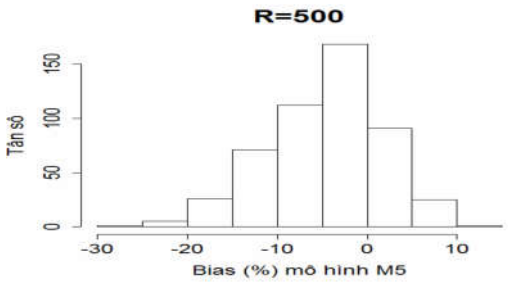
Vì vậy sử dụng thẩm định chéo theo Monte Carlo với R = 200 lần là hợp lý, cung cấp sai số ổn định và có phân bố chuẩn; kết quả này phù hợp với nghiên cứu của Temesgen *et al.,* (2014) và Huy *et al.,* (2016a,b). Không nhất thiết lặp lại quá lớn (R = 500 lần) như Zhang (1997) đề nghị.

Bảng 7. So sánh các chỉ tiêu thống kê, sai số thẩm định chéo mô hình lựa chọn $AGB = a \times (DBH^2HWD)^b \times CA^c$ theo phương pháp Monte Carlo với số lần lặp lại R khác nhau

Chỉ tiêu thống kê, sai số	Số lần lặp R của phương pháp thẩm định chéo Monte Carlo				
	50	100	200	300	500
AIC	873	871	872	870	870
R^2_{adj}	0,960	0,959	0,959	0,959	0,959
Bias (%)	-3,8	-4,5	-4,8	-4,7	-4,8
RMSE (%)	24,3	24,9	26,0	26,0	25,3
MAPE(%)	17,1	17,4	17,9	17,8	17,8

Ghi chú: R^2, AIC được tính từ 80% dữ liệu rút ngẫu nhiên; các sai số Bias, RMSE, MAPE được tính từ 20% dữ liệu đánh giá được rút ngẫu nhiên, độc lập và tính trung bình từ R lần lại





Hình 6. Phân bố Bias của mô hình $AGB = a \times (DBH^2 HWD)^b \times CA^c$ theo phương pháp thẩm định chéo Monte Carlo với số lần lặp R khác nhau

Sau khi thẩm định được sai số, tiến hành lập mô hình sinh khối với toàn bộ dữ liệu và các chỉ tiêu thống kê của mô hình (AIC, $R^2_{adj.}$); sai số MAPE của các mô hình được lựa chọn vì ổn định nhất và được lấy từ kết quả áp dụng phương pháp Monte Carlo với R =200 lần lặp (Bảng 8).

Bảng 8. Kết quả ước lượng các mô hình sinh khối từ toàn bộ dữ liệu và sai số từ thẩm định chéo theo phương pháp Monte Carlo

Mã mô hình	Mô hình	Trọng số (Weight)	AIC	R^2_{adj}	MAPE (%)
M1	$AGB = 0{,}10959 \times DBH^{2.47432}$	$1/DBH^k$	1119	0,934	22,1
M2	$AGB = 267.35155 \times (DBH^2 H)^{0.96377}$	$1/DBH^k$	1110	0,939	21,1
M3	$AGB = 0{,}19574 \times DBH^{2.45968} WD$	$1/DBH^k$	1093	0,948	199
M4	$AGB = 0{,}59164 \times (DBH^2 HWD)^{0.98655}$	$1/DBH^k$	1090	0,954	19,7
M5	$AGB = 0{,}613816 \times (DBH^2 HWD)^{0.86998} \times CA^{0.18783}$	$1/(DBH^2 HWD)^k$	1084	0,960	17,9

Ghi chú: Mô hình và các chỉ tiêu AIC, $R^2_{adj.}$ được thiết lập từ toàn bộ dữ liệu, sai số MAPE được lấy từ kết quả của phương pháp Monte Carlo với R = 200 lần; k là hệ số của hàm phương sai; P-value của các tham số <0,0001.

4. KẾT LUẬN

Các phương pháp thẩm định chéo đã hỗ trợ cho lựa chọn mô hình và xác định đúng các sai số so với phương pháp truyền thống là sử dụng một bộ dữ liệu độc lập để đánh giá mô hình.

Trong đó phương pháp thẩm định chéo của Monter Carlo phân chia dữ liệu ngẫu nhiên thành hai phần: 80% dữ liệu để lập mô hình và 20% dữ liệu để thẩm định sai số, được lặp lại 200 lần là thích hợp nhất, cung cấp sai số các mô hình ổn định và có phân bố chuẩn.

Trong vùng sinh thái Nam Trung bộ, mô hình ước tính sinh khối cây rừng trên mặt đất được lựa chọn gồm bốn biến độc lập theo dạng $AGB = a \times (DBH^2 HWD)^b \times CA^c$ với sai số MAPE trung bình của 200 lần thẩm định chéo theo phương pháp Monte Carlo là 17,9%.

TÀI LIỆU THAM KHẢO

1. Basuki, T. M.; Van Lake, P. E.; Skidmore, A. K.; Hussin, Y. A. 2009. Allometric equations for estimating the above-ground biomass in the tropical lowland Dipterocarp forests. For. Ecol. and Manag. 257(2009): 1684-1694. DOI 10.1016/j.foreco.2009.01.027.

2. Bates, D.M,2010.lme 4; Mixe –efects modeling with R.Springer, 131p

3. Brown S. 1997. Estimating biomass and biomass change of tropical forests: A Primer. FAO Forestry paper – 134. ISBN 92-5-103955-0. Available on-line: http://www.fao.org/docrep/w4095e/w4095e00.htm

4. Chave J, Andalo C, Brown S, Cairns MA, Chambers JQ, Eamus D, Folster H, Fromard F, Higuchi N, Kira T, Lescure JP, Nelson BW, Ogawa H, Puig H, Rier B, Yamakura T. 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145 (2005): 87-99. DOI 10.1007/s00442-005-0100-x.

5. Chave, J., Mechain, M. R., Burquez, A., Chidumayo, E., Colgan, M. S., Delitti, W. B. C., Duque, A., Eid, T., Fearnside, P. M., Goodman, R. C., Henry, M., Yrizar, A. M., Mugasha, W. A., Mullerlandau, H. C., Mencuccini, M., Nelson, B. W., Ngomanda, A., Nogueira, E. M., Ortiz-Malavassi, E.,

Pelissier, R., Ploton, P., Ryan, C. M., Saldarriaga, J. G., and Vieilledent, G. 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Global change biology, 20(2014): 3177-3190. Doi: 10.1111/gcb.12629.

6. Huy, B., Kralicek, K., Poudel, K. P., Phuong, V. T., Khoa, P. V., Hung, N. D., Temesgen, H. 2016a. Allometric Equations for Estimating Tree Aboveground Biomass in Evergreen Broadleaf Forests of Viet Nam. For. Ecol. and Mgmt. 382: 193-205.

7. Huy, B., Poudel K. P., Temesgen, H. 2016b. Aboveground biomass equations for evergreen broadleaf forests in South Central Coastal ecoregion of Viet Nam: Selection of eco-regional or pantropical models. For. Ecol. and Mgmt. 376: 276-282.

8. Mayer, D. G., Butler, D. G., 1993. Statistical validation. Ecological Modelling, 68, 21-32.

9. Moore, A. W. 2017. Cross-validation for detecting and preventing overfitting. School of Computer Science. Carnegie Mellon University. Available on-line: https://www.autonlab.org/_media/tutorials/overfit10.pdf on February 02, 2017.

10. Picard, R., Cook, D. 1984. Cross-Validation of Regression Models. Journal of the American Statistical Association. 79 (387): 575–583. doi:10.2307/2288403. JSTOR 2288403.

11. Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. & Team, R. C. 2014. nlme: Linear and nonlinear mixed effects models. R package version 3.1-117.

12. Swanson, D. A., Tayman, J., Bryan, T. M., 2011. MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecastts. J Pop Research 28(2011):225-243. DOI 10.1007/s12546-011-9054-5.

13. Temesgen, H., Zhang, C. H., Zhao, X. H. 2014. Modelling tree height-diameter relationships in multi-species and multi-layered forests: A large observational study from Northeast China. Journal of Forest Ecology and Management, 316(2014): 78-89

14. Wickham, H. & Chang, W. 2013. Package 'ggplot2': an implementation of the Grammar of Graphics.

15. Zhang, L. 1997. Cross-validation of Non-linear Growth Functions for Modelling Tree Height–Diameter Relationships. Annals of Botany 79(1997): 251–257.

CROSS VALIDATION METHODS OF ABOVEGROUND BIOMASS EQUATIONS

Bao Huy

Summary

The models used to estimate biomass and report CO_2 equivalent from forests under in REDD+ (reducing emissions from deforestation and forest degradation programme) should indicate the reliability and their uncertainty, therefore selected biomass equations were validated for their predictive abilities using data collected from destructively sampled 110 trees of the evergreen broadleaf forests of the South Central Coastal region of Viet Nam. Different power models that used diameter at breast height (DBH), tree height (H), wood density (WD), and crown area (CA) as covariates to predict tree aboveground biomass (AGB) were evaluated. Four methods of cross validation were performed: one round of conventional validation, Leave-One-Out (LOOCV), k-fold and Monte Carlo. In these methods, Monte Carlo was most appropriate to provide stable cross-validation statistics and normal error distribution. Monte Carlo cross-validation statistics of percent bias, root mean square percentage error (RMPE %), and mean absolute percent error (MAPE) were computed by randomly splitting data 200 times into model development (80%) and validation (20%) datasets and averaging over the 200 realizations. Best model was selected based on the coefficient of determination (R^2), the Akaike information criterion (AIC). AGB was strongly related to four covariates - DBH, H, WD, and CA. Accuracy of the selected model ($AGB = a \times (DBH^2HWD)^b \times CA^c$) had the lowest MAPE of 17.9 percent.

Keywords: *AGB, biomass equation, cross validation, k-fold, LOOCV, Monte Carlo.*

Người phản biện: GS.TS. Võ Đại Hải

Ngày nhận bài: 02/12/2016

Ngày thông qua phản biện: 3/01/2017

Ngày duyệt đăng: 10/01/2017