Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Trái đất và Môi trường 5(SI3):40-53

Open Access Full Text Article

Bài nghiên cứu

Phân loại lớp đất phủ sử dụng mô hình Random Forest kết hợp chỉ số thực vật NDVI và yếu tố địa hình: một nghiên cứu điển hình tại tỉnh Thanh Hóa, Việt Nam

Lê Trọng Diệu Hiền[1,*], Phạm Hồng Luân[2], Hoàng Thị Tuyết[1], Đinh Quang Toàn[3]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Lập bản đồ lớp phủ/sử dụng đất (LULC) trong khu vực che phủ phức tạp là một công việc đầy thách thức đặc biệt là đối với khu vực với kiểu thảm thực vật hỗn tạp, đồi núi trập trùng và có các sông chảy xiết. Do đó, một kỹ thuật mới nên được áp dụng để cải thiện việc phân loại chính xác hơn các loại đất phủ phức tạp. Trong nghiên cứu này, nhóm nghiên cứu đã áp dụng phương pháp máy học có giám sát để thành lập bản đồ lớp đất phủ bằng dữ liệu chỉ số thực vật (Normalized Difference Vegetation Index, NDVI) từ ảnh vệ tinh MODIS (Moderate Resolution Imaging Spectroradiometer). Nhóm nghiên cứu đã sử dụng lớp đất phủ năm 2015 làm các biến phụ thuộc và 15 biến độc lập để đưa vào mô hình phân loại tự động hóa Random Forest (RF). Mô hình phân loại RF được huấn luyện và tiến hành đánh giá nhiều lần xuyên suốt quá trình huấn luyện để xác định mô hình tối ưu với độ chính xác cao nhất trên dữ liệu huấn luyện. Phân loại RF đạt độ chính xác tổng là 91% và hệ số Kappa (K) là 0.89 trên 8 lớp đất phủ khác nhau bao gồm: đất trống, đất rừng, đất trồng hoa màu, đất trồng lúa nước, đất ngập nước, đất thành thị và xây dựng, đồng cỏ, và mặt nước. Bên cạnh đó, kết quả cho thấy không chỉ chỉ số thực vật NDVI mà địa hình cũng là các biến số tương đối quan trọng kiểm soát việc phân loại lớp đất phủ.

Từ khoá: MODIS, phân loại đất, Random Forest, viễn thám

[1]Chương trình Khoa học Môi trường, Đại học Thủ Dầu Một, 06 Trần Văn Ơn, Thành phố Thủ Dầu Một, Bình Dương, Việt Nam

[2]Trung tâm Quản lý Nước và Biến đổi Khí hậu, Viện Môi trường và Tài nguyên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam, 01 Marie Curie, Phường Linh Trung, Quận Thủ Đức, Thành phố Hồ Chí Minh, Việt Nam

[3]Sở Tài nguyên và Môi trường Thanh Hóa, Thanh Hóa 400570, Việt Nam

Liên hệ

Lê Trọng Diệu Hiền, Chương trình Khoa học Môi trường, Đại học Thủ Dầu Một, 06 Trần Văn Ơn, Thành phố Thủ Dầu Một, Bình Dương, Việt Nam

Email: hienltd@tdmu.edu.vn

Lịch sử

- Ngày nhận: 12-12-2021
- Ngày chấp nhận: 06-3-2022
- Ngày đăng: 30-6-2022

DOI : 10.32508/stdjsee.v5iSI2.681

Check for updates

GIỚI THIỆU

Lớp đất phủ là một trong các yếu tố quan trọng trong các phân tích khác nhau như: nghiên cứu địa lý, phân tích môi trường, quản lý đất đai (xu hướng rừng bị tàn phá, quy hoạch nông nghiệp và tăng trưởng đô thị [1]). Lớp đất phủ thay đổi theo thời gian do sự tương tác giữa các hoạt động kinh tế - xã hội và sự thay đổi môi trường trong khu vực, do đó, độ che đất phủ cần được cập nhật thường xuyên [2]. Các bản đồ hiện có về lớp phủ được xác định bằng cách lấy mẫu thực địa và ảnh hàng không, nhưng phương pháp thực hiện này có nhiều hạn chế trong trường hợp khu vực nghiên cứu có diện tích lớn, là những vùng núi hiểm trở có sông chảy xiết; và tốn kém, mất thời gian. Các mô hình dựa trên các yếu tố đã biết về đặc điểm chỉ số thực vật, địa hình, và kết cấu địa chất có thể giúp dự đoán tự động lớp đất phủ. Những mô hình như vậy có thể nâng cao độ chính xác của việc phân loại lớp đất phủ và đánh giá sự thay đổi. Random Forest (RF) không nhạy cảm với sự thay đổi với kích thước dữ liệu và nhìn chung có hiệu suất phân loại cao so với các phương pháp máy học khác [3–5]; và cho phép xác định tầm quan trọng của các biến trong mô hình phân loại lớp đất phủ.

Viễn thám với ảnh vệ tinh là dữ liệu đầu vào đáng tin cậy cho những nghiên cứu về phân loại lớp đất phủ do có phạm vi địa lý rộng lớn và độ phủ ảnh theo thời gian cao. Bên cạnh đó, viễn thám cũng hỗ trợ điều tra lớp phủ trong quá khứ và cung cấp dữ liệu ở các khu vực không thể tiếp cận (ví dụ như các khu vực núi hiểm trở). Dữ liệu viễn thám có các đặc điểm vĩ mô và đồng bộ được sử dụng rộng rãi trong các nghiên cứu phân loại lớp đất phủ [2,6,7]. Phương pháp máy học kết hợp với hình ảnh vệ tinh để phân loại lớp đất phủ đã được ứng dụng nhiều trong các nghiên cứu trên thế giới chẳng hạn như vùng đô thị Atlanta, Georgia [8]; Đông Bắc Latvia [9]; Prahova Subcarpathians, Romania [2]; bốn tiểu bang bao gồm Brandenburg, Lower Saxony, North RhineWestphalia, và Rhineland Palatinate, ở Đức [10]; và ở quận Uppsala ở miền trung nam Thụy Điển nơi có diện tích nhỏ nhưng lớp đất phủ phức tạp [11].

Tuy nhiên, ở các khu vực có lớp đất phủ phức tạp, các bản đồ che đất phủ hiện vẫn còn tồn tại một số vấn đề. Ví dụ, khu vực đồng cỏ, rừng và đất trồng trọt không thể dễ dàng phân biệt nếu chỉ với thông tin hình thái thực vật từ một hình ảnh vệ tinh duy nhất. Do đó, dữ liệu chuỗi thời gian NDVI, trích xuất từ Landsat, MODIS và Sentinel-2, là dữ liệu có tầm quan trọng lớn đối với việc lập bản đồ lớp đất phủ [9,10,12]. Để tạo ra độ chính xác tốt nhất cho việc phân loại lớp đất phủ, bên

Trích dẫn bài báo này: Hiền L T D, Luân P H, Tuyết H T, Toàn D Q. **Phân loại lớp đất phủ sử dụng mô hình Random Forest kết hợp chỉ số thực vật NDVI và yếu tố địa hình: một nghiên cứu điển hình tại tỉnh Thanh Hóa, Việt Nam**. *Sci. Tech. Dev. J. - Sci. Earth Environ.;* 5(SI3):40-53.

cạnh chỉ số hình thái thực vật, các đặc tính của đất và đặc điểm địa hình cũng cần được xem xét trong phân loại lớp đất phủ [13].

Trong nghiên cứu này, mục tiêu của nhóm nghiên cứu là phân loại lớp đất phủ ở tỉnh Thanh Hóa, Việt Nam năm 2015 bằng cách sử dụng mô hình phân loại RF dựa vào các yếu tố dự báo kiểm soát việc phân loại sử dụng đất bao gồm thông tin thảm thực vật từ ảnh vệ tinh MODIS và địa hình. Ảnh MODIS được thu nhận từ hai hệ thống vệ tinh chính, bao gồm: bao gồm: MODIS Terra và MODIS Aqua. Với tầm quan sát lên đến hơn 2.330 km, vệ tinh này có thể quan trắc gần như toàn bộ Trái Đất. Ảnh MODIS có 36 băng phổ, với 3 độ phân giải: 250, 500 và 1000 mét. Trong nghiên cứu này, nhóm tác giả sử dụng ảnh MODIS (MOD13Q1) được tạo ra cứ 16 ngày một lần ở độ phân giải không gian 250 m thu từ vệ tinh Terra do những ảnh này tương đối ít bị ảnh hưởng bởi mây [14].

PHƯƠNG PHÁP NGHIÊN CỨU

Khu vực nghiên cứu

Tỉnh Thanh Hóa nằm ở Bắc Trung Bộ, Việt Nam (Vĩ độ: 19 ° 18'N - 20 ° 40'N, Kinh độ: 104 ° 22'E - 106 ° 05'E) (Hình 1), và là một trong những tỉnh lớn nhất Việt Nam với diện tích 11,106 km^2. Thanh Hóa có địa hình đồi núi, mạng lưới sông ngòi dày đặc và tiếp giáp biển Đông. Ngoài ra, diện tích ven sông Mã tạo nên đồng bằng sông Mã (còn gọi là đồng bằng Thanh Hóa) lớn thứ ba ở Việt Nam. Mặc dù diện tích nông nghiệp và đất trồng trọt trong lưu vực không đáng kể, trong đó trồng lúa chiếm 65 % (Ngân hàng Thế giới, 2007), nhưng hoạt động nông nghiệp vẫn đáng kể và chiếm 35% GDP trong vùng.

Dữ liệu và lựa chọn mẫu

a. Chỉ số thực vật (NDVI)

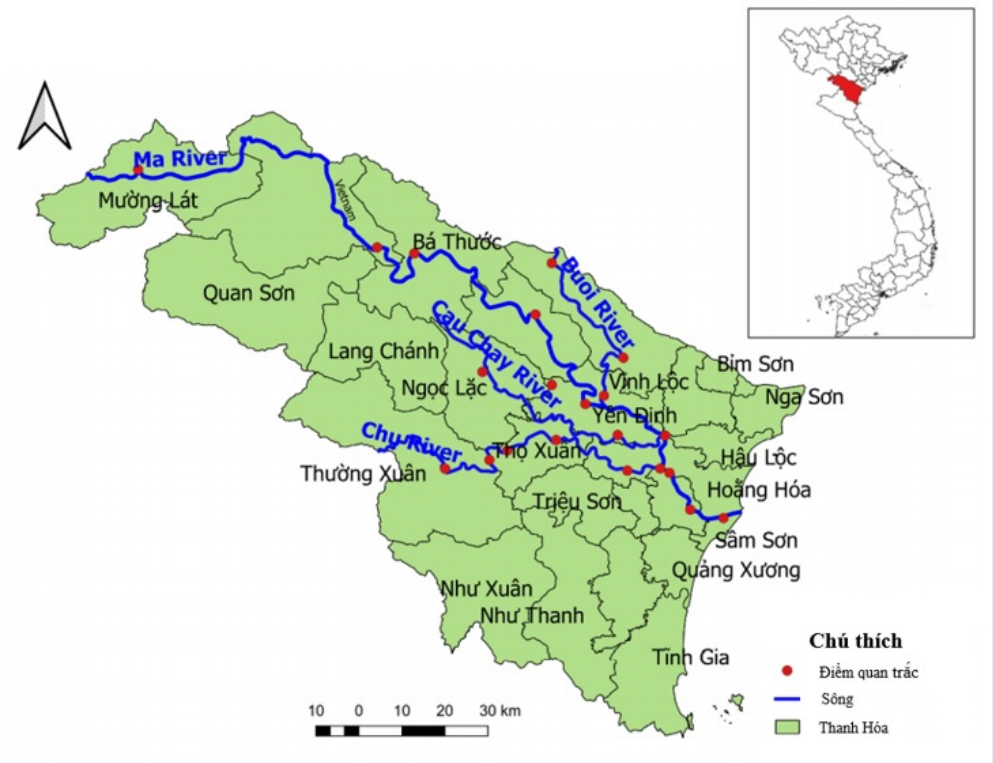
Tổng số 23 ảnh vệ tinh NDVI trong năm 2015 của tỉnh Thanh Hóa được tải xuống từ trang web https://lpdaacsvc.cr.usgs.gov/appeears/ từ tháng 1 đến tháng 12 năm 2015 dựa trên ảnh MOD13-Q1 16 ngày, độ phân giải 250 m. NDVI là nguồn dữ liệu phổ biến để phân loại lớp đất phủ, có thể được sử dụng để thu thập các đặc điểm của thảm thực vật [15]. Giá trị này dao động trong khoảng -1 đến 1; với giá trị cao là + 1 cho biết các đặc điểm bề mặt có thảm thực vật dày đặc và các giá trị gần bằng 0 hoặc nhỏ hơn cho biết bề mặt không có thảm thực vật. Trước khi trích xuất các giá trị NDVI cho mỗi vị trí tham chiếu, ảnh vệ tinh được làm sạch để loại bỏ các ô lưới bị che phủ bởi mây và sau đó áp dụng phép nội suy tuyến tính để nội suy các giá trị bị thiếu. Tiếp theo, nhóm tác giả tính toán NDVI trung bình, NDVI tối đa, NDVI tối thiểu, phương sai NDVI, tổng NDVI, ngày trong năm có NDVI tối đa và ngày trong năm có NDVI tối thiểu (Bảng 1).

b. Dữ liệu địa hình

Mô hình số độ cao (DEM) của khu vực nghiên cứu được tải xuống tại https://earthexplorer.usgs.gov/. Sau đó, các biến số liên quan đến địa hình như hướng sườn, độ dốc, độ cong tiếp tuyến và độ cong biên được tính từ DEM (Bảng 1). Để có được độ chính xác tốt nhất cho việc phân loại lớp đất phủ, bên cạnh đặc điểm hình thái thực vật thì các đặc điểm địa hình của mỗi loại đất phủ đều liên quan đến quá trình phân loại. Ví dụ, đất trồng trọt, rừng và đồng cỏ không dễ dàng phân biệt nếu chỉ sử dụng chỉ số thực vật [13,16].

c. Chọn mẫu

Dữ liệu tham chiếu là thành phần quan trọng trong phương pháp máy học, và hầu hết các mô hình đều yêu cầu hàng nghìn mẫu dữ liệu tham chiếu. Tuy nhiên, việc xác định và thu thập dữ liệu tham khảo trên các khu vực rộng lớn, xa xôi hoặc hẻo lánh cho quá trình huấn luyện là một nhiệm vụ khó khăn [17]. Việc sử dụng các bản đồ hiện có làm dữ liệu phụ trợ để phân loại lớp đất phủ trong quá khứ là cách làm phổ biến trong phân loại sử dụng đất bằng máy học. Ví dụ, Tran và cộng sự (2015) [18] đã sử dụng bản đồ chuyên đề từ những năm 1970 làm đầu vào để phân loại dữ liệu Landsat-1 từ năm 1973 nhằm điều tra sự thay đổi sử dụng đất ở tỉnh Cà Mau, Việt Nam. Việc sử dụng bản đồ hiện có làm dữ liệu huấn luyện trong phân loại độ che phủ có thể tạo ra sai sót. Tuy nhiên, các bản đồ lớp phủ hiện có với độ chính xác tổng thể cao, hợp lý có thể tạo ra một số lượng lớn các mẫu huấn luyện với hiệu quả cao, cho phép thể hiện một loạt các đối tượng địa lý [19]. Hơn nữa, một bản đồ hiện có cũng được sử dụng để kiểm tra ngẫu nhiên độ chính xác của mô hình nhằm giảm thiểu phân loại sai. Trong nghiên cứu này, nhóm nghiên cứu lựa chọn mẫu huấn luyện ban đầu bằng cách sử dụng bản đồ độ đất phủ hiện có của năm 2015 được cung cấp bởi Sở Tài nguyên và Môi trường, tỉnh Thanh Hóa. Sau đó, các mẫu này được so sánh với giải đoán từ ảnh Landsat 8 đa thời gian để xem có bất kỳ sự khác biệt nào; ví dụ như đất trống hoặc các khu vực đã xây dựng. Các mẫu được phân loại giống nhau trong cả hai phương pháp được chọn làm mẫu tham chiếu. Tổng số mẫu và số lượng mẫu cho từng loại lớp đất đất phủ được nêu trong Bảng 1. Lớp đất phủ được phân thành 8 loại: đất trống, hoa màu, rừng, đồng cỏ, rừng ngập mặn, lúa nước, đô thị và đất xây dựng, và mặt nước



Hình 1: Vị trí địa lý tỉnh Thanh Hóa

Mô hình phân loại Random Forest (RF)

a. Mô hình

Random Forest (RF) là một phương pháp máy học được giới thiệu bởi Breiman[3] cho phép cải thiện độ chính xác của dự đoán và phân loại mà không cần trang bị quá nhiều dữ liệu. RF dựa trên cây phân loại và hồi quy (CART) (Hình 7) và kỹ thuật đánh giá chéo (cross-validation). Nhóm tác giả sử dụng kỹ thuật đánh giá chéo với số nhóm là 10 (kfold=10) và số lần lặp lại là 3 (reapeated=3) để tối ưu hóa số lượng cây quyết định cần xây dựng (ntree) và số lượng biến tại mỗi lần chia nút cây quyết định (mtry). Nói một cách chi tiết, kỹ thuật đánh giá chéo lặp lại bao gồm việc chia ngẫu nhiên dữ liệu tham chiếu ban đầu thành 10 nhóm với nhóm 1 là nhóm dùng để kiểm định và nhóm từ 2-10 là nhóm huấn luyện. Quy trình tương tự được lặp lại 3 lần với mỗi lần sử dụng các nhóm khác nhau làm nhóm kiểm định và phần còn lại của nhóm làm nhóm huấn luyện. Rodriguez-Galiano và cộng sự (2012)[20] quan sát thấy rằng giảm mtry làm giảm mối tương quan giữa các cây riêng lẻ, điều này làm tăng hiệu suất dự đoán của mô hình. Tuy nhiên, Oliveira S và cộng sự (2012)[21] lại cho rằng sự gia tăng giá trị của mtry sẽ dẫn đến độ chính xác cao hơn cho mô hình và phân bổ từ biến quan trọng cao hơn cho đến biến ít quan trọng hơn. Nhóm nghiên cứu tối ưu hóa mô hình RF cuối cùng dựa trên độ chính xác của từng thuật toán được đánh giá bằng một số chỉ số (độ chính xác nhà sản xuất, độ chính xác người sử dụng) trích từ ma trận sai số.

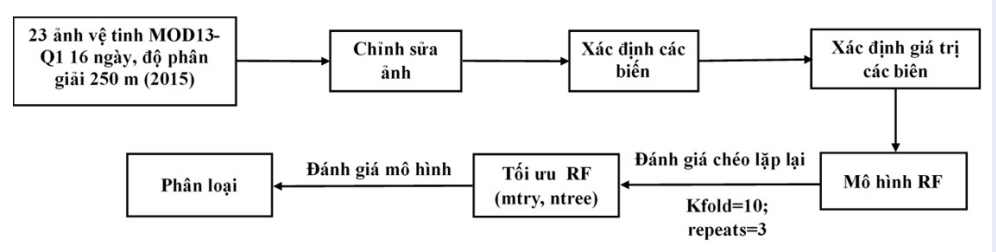
Trong nghiên cứu này, bộ dữ liệu tham chiếu được áp dụng trong RF bao gồm các hạng mục sử dụng đất như biến phụ thuộc và 15 biến dự báo có tỷ lệ thông tin cao để phân loại độ che đất phủ: 8 mô tả đặc điểm chỉ số thực vật NDVI và 5 mô tả đặc điểm địa hình (Bảng 1). Tổng số 1114 mẫu và số lượng mẫu tham chiếu cho từng loại hình sử dụng đất được cung cấp trong Bảng 2.

b. Đánh giá hiệu quả của mô hình phân loại RF

Mô hình huấn luyện cuối cùng được xác nhận với dữ liệu huấn luyện trong quá trình đánh giá độ chính xác nội bộ của mô hình. Các nội dung phải được báo cáo để đánh giá độ chính xác của mô hình phân loại bao gồm ma trận lỗi, độ chính xác tổng thể (OA), độ chính xác của nhà sản xuất (PA) và độ chính xác của người sử dụng (UA) và hệ số Kappa[22]. Ma trận sai số so sánh định lượng sự giống nhau giữa các mẫu được phân

Bảng 1: Mô tả các biến đáp ứng (phụ thuộc) và dự báo (độc lập) sử dụng trong mô hình Random Forest

Biến đáp ứng/dự báo	Danh mục biến	Biến	Mô tả
Đáp ứng	Lớp đất phủ	Danh mục lớp đất phủ	đất trống, hoa màu, rừng, đồng cỏ, rừng ngập mặn, lúa nước, đô thị và xây dựng, và mặt nước
Dự báo	Chỉ số thực vật (NDVI)	NDVI nhỏ nhất	NDVI nhỏ nhất trong năm
		NDVI lớn nhất	NDVI lớn nhất trong năm
		NDVI trung bình	NDVI trung bình trong năm
		NDVI tổng	NDVI tổng trong năm
		NDVI phương sai	NDVI phương sai trong năm
		Ngày có NDVI lớn nhất	Ngày trong năm có NDVI tối đa
		Ngày có NDVI thấp nhất	Ngày trong năm có NDVI tối thiểu
		Số lần NDVI thay đổi	Số lần NDVI thay đổi trong năm
		Ngày thay đổi NDVI đầu tiên	Ngày trong năm có sự thay đổi NDVI đầu tiên
		Độ dốc xu hướng NDVI	Độ dốc xu hướng NDVI
	Địa hình	Độ dốc	Độ dốc địa hình
		Hướng sườn	Hướng sườn địa hình
		Độ cong tiếp tuyến	Độ cong tiếp tuyến
		Độ cong biên	Độ cong biên
		Độ cao	Độ cao



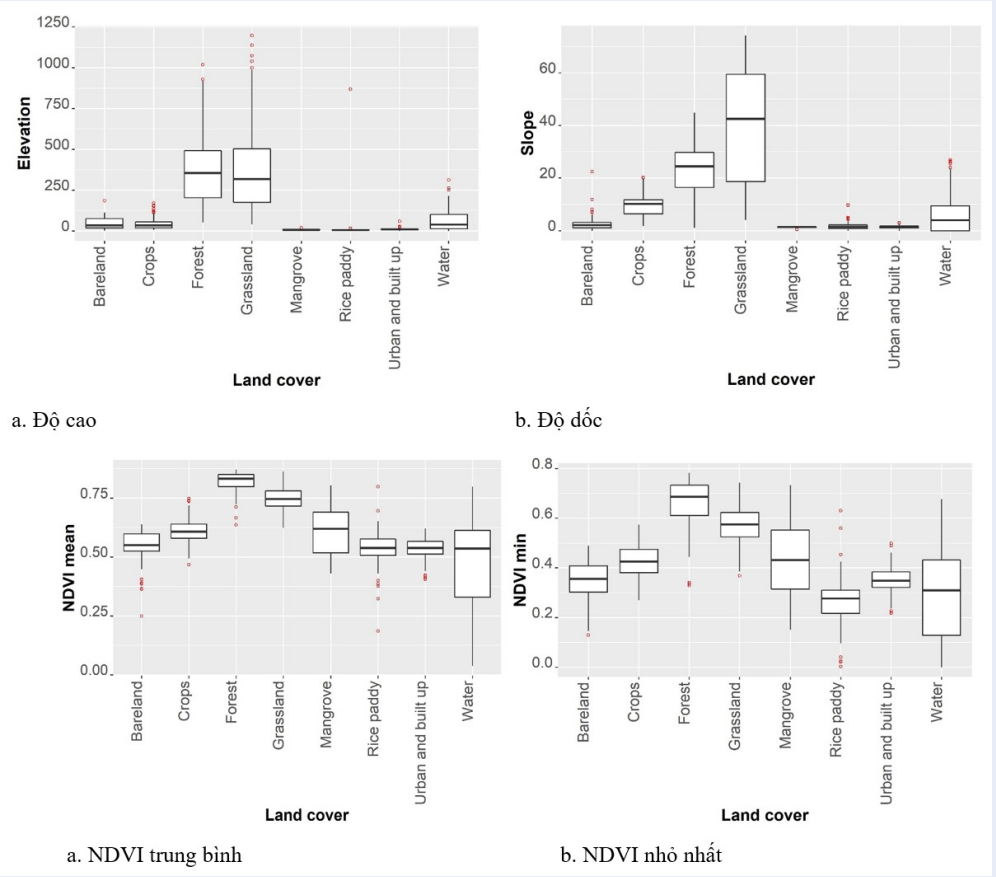
Hình 2: Sơ đồ quy trình phân loại lớp đất phủ

loại theo mô hình và dữ liệu tham chiếu. OA là tỷ lệ của các mẫu được phân loại chính xác theo phân loại đầu vào của mẫu. Tuy nhiên, OA có xu hướng đánh giá quá cao hiệu suất thực tế, hệ số Kappa được dùng để đánh giá hiệu suất phân loại của mô hình. PA là tỷ lệ các mẫu được phân loại chính xác của một danh mục cụ thể trong khi UA là tỷ lệ các mẫu được phân loại chính xác và tổng số mẫu được phân loại vào danh mục đó[23]. Tầm quan trọng của biến dự báo là một số liệu thống kê cho thấy trọng số của mỗi yếu tố dự báo được sử dụng để xác định và phân loại lớp đất phủ tự động trong mô hình RF tối ưu. Nhóm nghiên cứu đã sử dụng gói *caret* trong R để lắp chạy mô hình RF. Hình 2 là tóm tắt về phân loại lớp đất phủ bằng RF. Tất cả các tính toán và hình ảnh trong nghiên cứu được thực hiện trên R phiên bản 4.1.2[24].

KẾT QUẢ VÀ THẢO LUẬN

Các thông số kỹ thuật

Sự phân tán các thông số kỹ thuật của các lớp đất phủ đối với các biến đã chọn được mô tả trong Hình 3. Các đối tượng có lớp phủ thực vật bao gồm rừng (0.82), hoa màu (0.61), đồng cỏ (0.74) và rừng ngập mặn (0.60) có sự phân bố giá trị NDVI cao hơn nhiều so



a. Độ cao b. Độ dốc

a. NDVI trung bình b. NDVI nhỏ nhất

Hình 3: Các biến phụ trợ trong RF. Những biến khác được mô tả trong tài liệu bổ sung Hình 7.

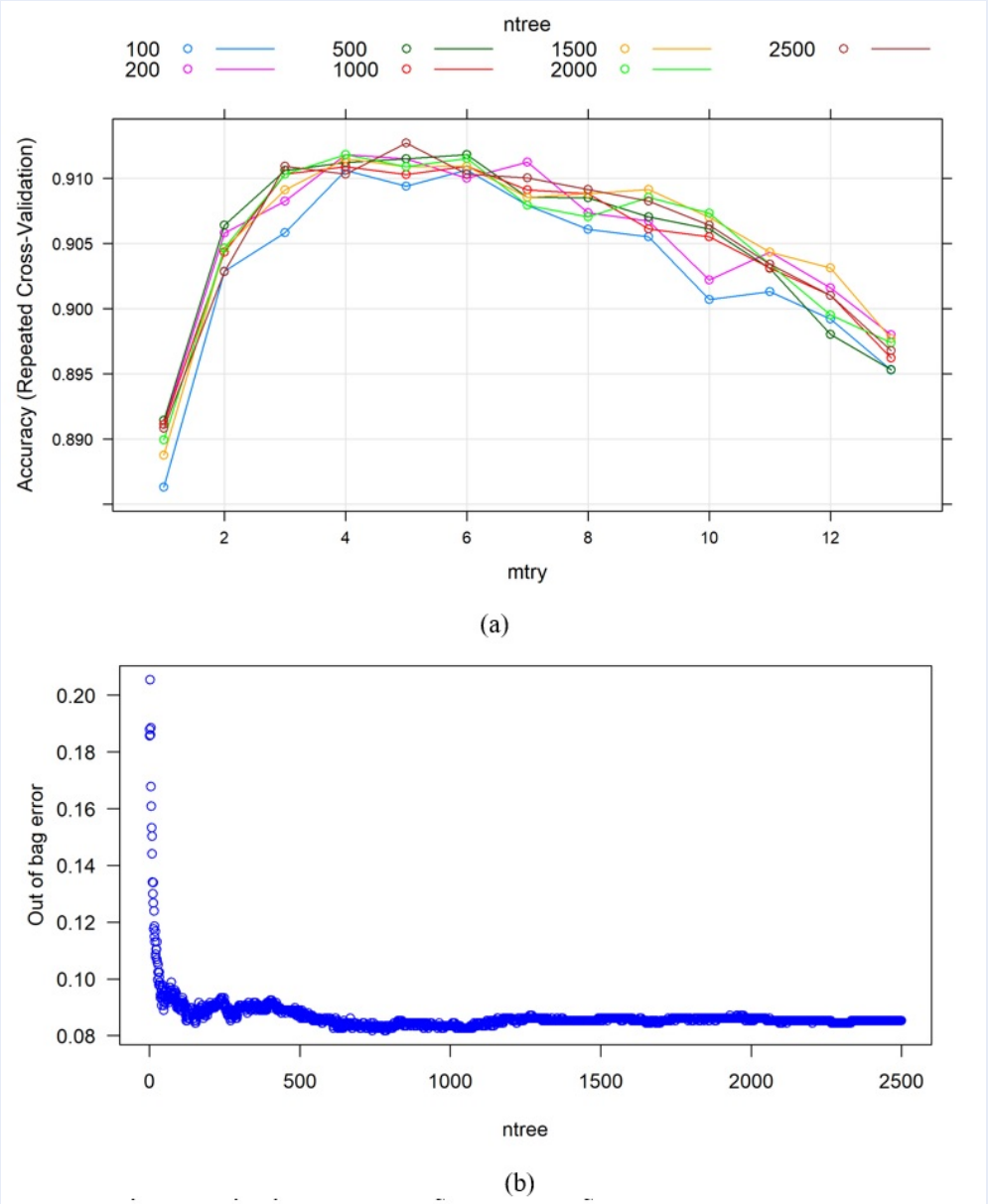
Bảng 2: Số lượng mẫu được sử dụng để huấn luyện

Danh mục lớp đất phủ	Số lượng mẫu
Đất trống	120
Hoa màu	181
Rừng	122
Đồng cỏ	274
Đất ngập nước	48
Lúa nước	102
Đô thị và đất xây dựng	147
Mặt nước	120
Tổng	1114

với mặt nước (0.47), đất trống (0.55) và đất xây dựng (0.54). Nhóm nghiên cứu cũng đã tìm thấy xu hướng tổng thể với đối tượng là lúa nước. Hình 8, NDVI theo thời gian của lúa nước dao động theo mùa đặc trưng với ba chu kỳ kéo dài khoảng bốn tháng. Trong khi đó NDVI của đối tượng mặt nước là ổn định theo thời gian (Hình 8). Đối với các đặc điểm địa hình, rừng có các giá trị về địa hình cao hơn hẳn so với các loại đất phủ khác có thể do rừng ở vùng đồi núi.

Tham số điều chỉnh và các biến quan trọng

Trong đánh giá chéo lặp lại với kfold=10 và số lần lặp lại là 3, mẫu tham chiếu ban đầu được chia một cách ngẫu nhiên thành 10 nhóm xây dựng mỗi cây trong tiến trình huấn luyện. Các mẫu huấn luyện được dự đoán từ mô hình để đánh giá độ chính xác của sự phân loại và out of bag error (OBB). Số lượng cây cây quyết định cần xây dựng tối ưu (ntree = 2500) và và số lượng biến tại mỗi lần chia nút cây quyết định (mtry = 5) tạo ra độ chính xác cao nhất đã được xác định (Hình 4a, Bảng 5 trong tài liệu bổ sung). RF có độ chính xác phân loại thấp hơn (OOB cao hơn) khi số lượng cây nhỏ. Tuy nhiên, sai số tổng quát không tăng hoặc giảm nhiều nếu chúng ta thêm vô số cây; và số lượng cây lớn hơn thực hiện phân loại ổn định hơn (Hình 4b) [13,25].

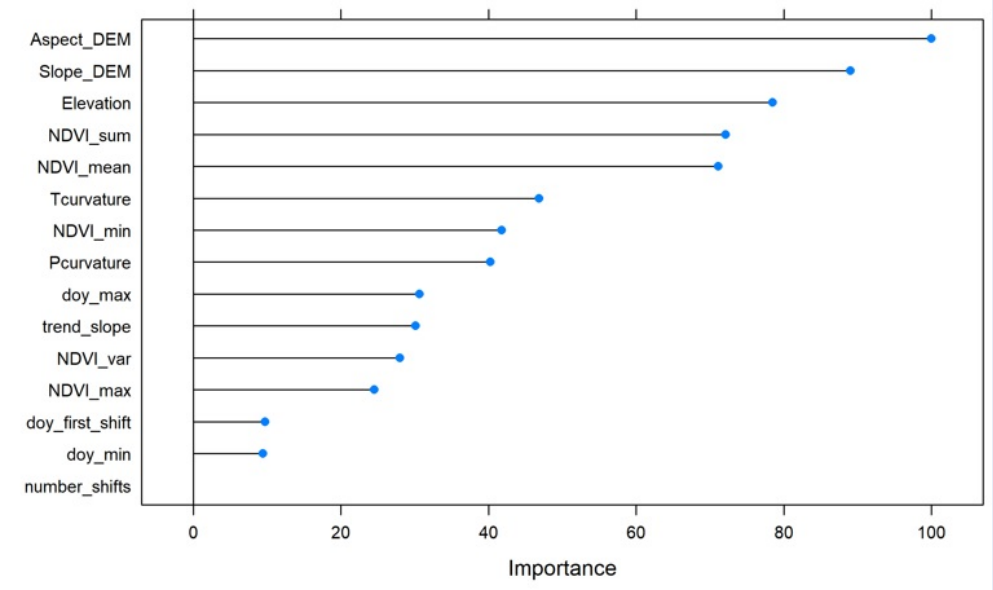


Hình 4: Số cây và số biến được chia ngẫu nhiên tại mỗi nút của RF (a). Out of bag error (OOB) so với ntree trong RF

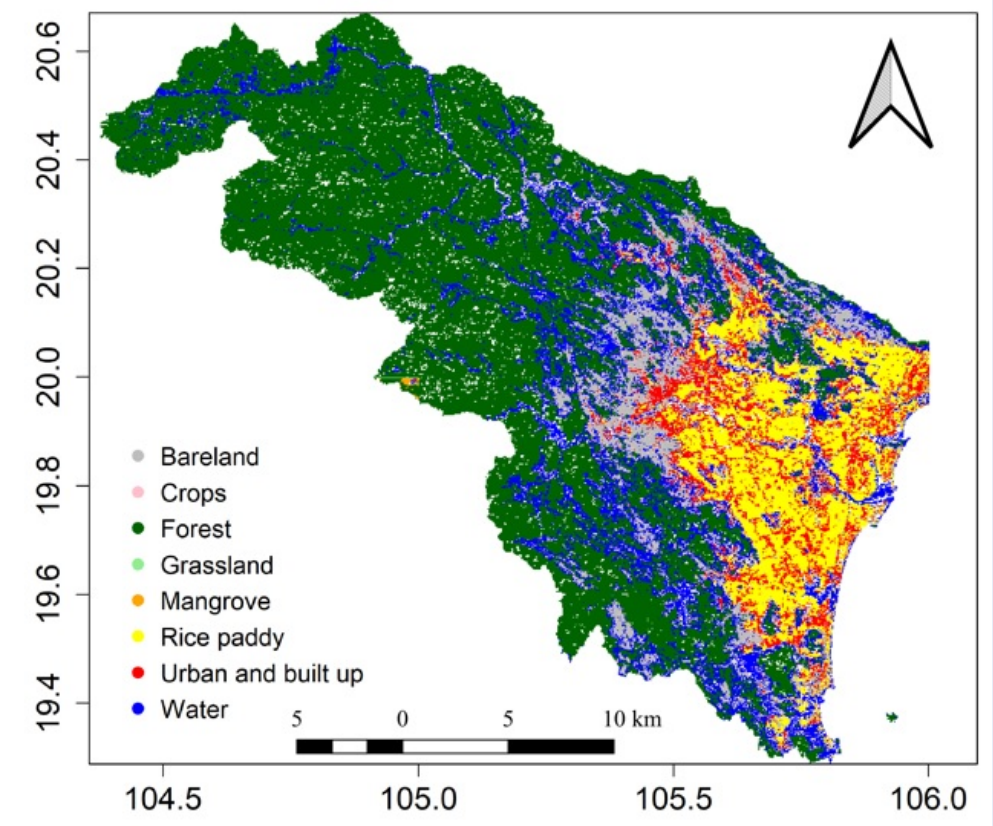
Mặc dù tất cả 15 biến đều hữu ích cho việc phân loại lớp phủ, nhưng một số đặc điểm rất quan trọng đối với việc phân loại (Hình 5). Từ tầm quan trọng của các biến, nhóm nghiên cứu rút ra ba kết luận. Thứ nhất, hầu hết các biến đều có tầm quan trọng trung bình cao hơn 25 % trong dữ liệu huấn luyện, đây là điều cần thiết để RF hoạt động ổn định. Thứ 2, các đặc điểm liên quan đến địa hình là các biến quan trọng để phân loại chính xác lớp đất phủ. Thứ 3, các biến số: số lần NDVI thay đổi trong năm và ngày có sự thay đổi NDVI đầu tiên, phản ánh ngày thu hoạch mùa hoa màu/ lúa và ngày vụ trồng hoa màu/lúa, có tầm quan trọng thấp bất ngờ so với mong đợi đối của nhóm nghiên cứu. Kết quả tương tự cũng được tìm thấy bởi Schulz C và cộng sự (2021)[10].

Hiệu suất phân loại

Hình 6 trình bày kết quả phân loại lớp đất phủ thu được bằng mô hình RF. Đánh giá trực quan cho thấy kết quả phân loại có hiệu quả tốt, và cho thấy cấu trúc



Hình 5: Tầm quan trọng của các biến



Hình 6: Bản đồ kết quả phân loại lớp đất phủ

lớp đất phủ chung của khu vực nghiên cứu.

Ma trận sai số được sử dụng để đánh giá độ chính xác cho kết quả phân loại lớp phủ bằng mô hình RF. Bảng 3 mô tả độ chính xác cho việc lập bản đồ độ đất phủ thu được bằng mô hình RF tối ưu sử dụng dữ liệu huấn luyện là 1114 mẫu. Các kết quả đánh giá độ chính xác thông qua mô hình RF xác nhận hiệu suất phân loại nói chung là tốt cho các tổ hợp 8 loại lớp đất phủ được xem xét trong nghiên cứu. Độ chính xác tổng thể là 91 %, với hệ số kappa là 0.89 cho dữ liệu tham chiếu. Những kết quả này cho thấy một sự trùng khớp gần như hoàn toàn theo Landis (1977). Giá trị độ chính xác của người dùng và nhà sản xuất trung bình của các lớp riêng lẻ xấp xỉ là 90 %. Tuy nhiên, độ chính xác của người sử dụng thấp hơn đối với mặt nước (77.5 %), lúa (80.4 %) và đất trống (83.3 %) so với các lớp phủ khác. Nói cách khác, mô hình phân loại đã bỏ sót 22.5 %, 19.6 % và 16.7 % (lỗi phân loại) của mặt nước, lúa nước và đất trống, cho thấy xu hướng mô hình phân loại nhầm nước là đất trống (10/120) và rừng (11/120); lúa nước là thành thị và đất xây dựng (13/102); và đất trống là lúa (13/120). Sai số này chủ yếu là do các đặc điểm hình thái học phức tạp của lớp đất phủ. Mặt khác, diện tích đất được ước tính là chính xác nhất đối với rừng; đô thị và đất xây dựng do các đặc điểm hình thái học đặc biệt của chúng. Bên cạnh đó, RF đã đánh giá quá cao diện tích đất trống vì các loại đất che phủ khác có thể bị phân loại nhầm thành "đất trống" (Bảng 3).

Ảnh hưởng của giảm số lượng mẫu trong dữ liệu huấn luyện

Thu thập dữ liệu huấn luyện quy mô lớn cho quá trình đào tạo RF là một công việc tốn nhiều thời gian trong việc phân loại các khu vực phức tạp với nhiều loại lớp đất phủ. Nhiều nghiên cứu trước đây đã quan sát thấy rằng càng nhiều mẫu dữ liệu đào tạo thì càng chứa nhiều điều kiện thay đổi trong mỗi loại [3,13,25] dẫn đến việc tăng độ chính xác của kết quả phân loại. Tuy nhiên, trong trường hợp dữ liệu huấn luyện không mang tính đại diện thì độ chính xác có thể bị giảm xuống [26]. Do đó, một kế hoạch chọn số lượng mẫu huấn luyện cần được thiết kế để đáp ứng tính khả thi cả về thời gian, điều kiện kinh tế và đạt được độ chính xác có thể chấp nhận được [27].

Bảng 4 cho thấy độ nhạy của hiệu suất RF do giảm kích thước của dữ liệu huấn luyện. Dữ liệu huấn luyện đã giảm từ 5 % xuống 70 %, trong khi độ chính xác phân loại tổng thể chỉ giảm dưới 5 %. Sau đó, ở ngưỡng giảm 70 % số lượng mẫu, độ chính xác bị giảm đột ngột hơn để đạt được độ chính xác tổng thể xấp xỉ 70 %, khi dữ liệu huấn luyện đã giảm 95%.

Kết quả cho thấy độ chính xác phân loại của RF giảm cùng với việc giảm kích thước tập dữ liệu huấn luyện, nhưng không theo mô hình tuyến tính. Kết quả này cũng được tìm thấy trong nghiên cứu của Jin và cộng sự (2018) [13]. Vì độ chính xác tổng thể bị giảm đột ngột sau khi đạt đến ngưỡng 70 % trong nghiên cứu này, nhóm nghiên cứu kết luận rằng phân loại RF không bị ảnh hưởng nhiều đến độ nhạy khi giảm dữ liệu huấn luyện.

KẾT LUẬN

Nhóm nghiên cứu đã phát triển một phương pháp hiệu quả và dễ dàng áp dụng dựa trên phương pháp máy học kết hợp chỉ số thực vật và yếu tố địa hình để phân loại lớp đất phủ cho toàn tỉnh Thanh Hóa năm 2015. Kết quả phân loại của mô hình Random Forest là tốt với độ chính xác tổng thể là 91 % và hệ số Kappa là 0.89. Mô hình phân loại này có thể được áp dụng ở các khu vực nghiên cứu khác để phân loại lớp đất phủ, sau đó dữ liệu này có thể được sử dụng cho những nghiên cứu, phân tích tiếp theo như nghiên cứu địa lý, phân tích môi trường, và quản lý đất đai. Bên cạnh đó, thông qua nghiên cứu này nhóm nghiên cứu còn nhận thấy rằng trong mô hình RF, việc giảm dữ liệu huấn luyện dẫn đến mức tăng sai số tương đối về tổng thể của kết quả phân loại, nhưng không phải tuyến tính. Giảm quy mô số lượng mẫu huấn luyện không có ảnh hưởng đáng kể đến độ chính xác của trình phân loại trước ngưỡng 70 %, cho thấy rằng phân loại lớp đất phủ bằng RF đa biến không nhạy cảm với giảm số lượng mẫu huấn luyện.

DANH MỤC TỪ VIẾT TẮT

RF: Random Forest

MODIS: Moderate Resolution Imaging Spectroradiometer

OA: Overal accurarcy, độ chính xác tổng thể

PA: Producer accuracy, độ chính xác của nhà sản xuất

UA: User accuary, độ chính xác của người sử dụng

KINH PHÍ

Nghiên cứu này được tài trợ bởi Trường Đại học Thủ Dầu Một trong đề tài mã số DT.21.2-006.

LỜI CẢM ƠN

Nhóm nghiên cứu xin chân thành cảm ơn trường đại học Thủ Dầu Một đã cấp tài trợ cho nghiên cứu này. Nhóm nghiên cứu cũng đánh giá cao Sở môi trường và tài nguyên tỉnh Thanh Hóa đã cung cấp dữ liệu tham chiếu trong quá trình nghiên cứu.

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả không có xung đột lợi ích với cá nhân hay tổ chức nào liên quan đến bài nghiên cứu

Bảng 3: Ma trận lỗi và các phép đo độ chính xác dựa trên mô hình tối ưu RF

Phân loại	Tham chiếu									
	Đất trống	Hoa màu	Rừng	Đồng cỏ	Đất ngập nước	Lúa nước	Đô thị và đất xây dựng	Mặt nước	Tổng	Độ chính xác người sử dụng
Đất trống	100	0	0	0	1	13	4	2	120	83.3
Hoa màu	1	173	0	5	0	0	2	0	181	95.6
Rừng	0	0	118	0	0	0	0	4	122	96.7
Đồng cỏ	0	6	0	268	0	0	0	0	274	97.8
Đất ngập nước	1	1	0	0	44	0	2	0	48	91.7
Lúa nước	4	0	0	0	1	82	13	2	102	80.4
Đô thị và đất xây dựng	2	0	0	0	0	2	143	0	147	97.3
Mặt nước	10	4	11	0	1	1	0	93	120	77.5
Tổng	118	184	129	273	47	98	164	101	1114	NA
Độ chính xác nhà sản xuất	84.7	94	91.5	98.2	93.6	83.7	87.2	92.1	NA	91.7

Bảng 4: Tác động của kích thước dữ liệu huấn luyện đến độ chính xác của phân loại RF

Tỷ lệ giảm (%)	Độ chính xác tổng thể	Tỷ lệ giảm (%)	Độ chính xác tổng thể
5	91.33	55	90.09
10	91.98	60	88.79
15	91.86	65	88.53
20	91.43	70	86.09
25	91.77	75	85.81
30	91.15	80	85.27
35	91.16	85	85.23
40	89.81	90	79.99
45	90.61	95	73.41
50	91.03		

ĐÓNG GÓP CỦA NHÓM TÁC GIẢ

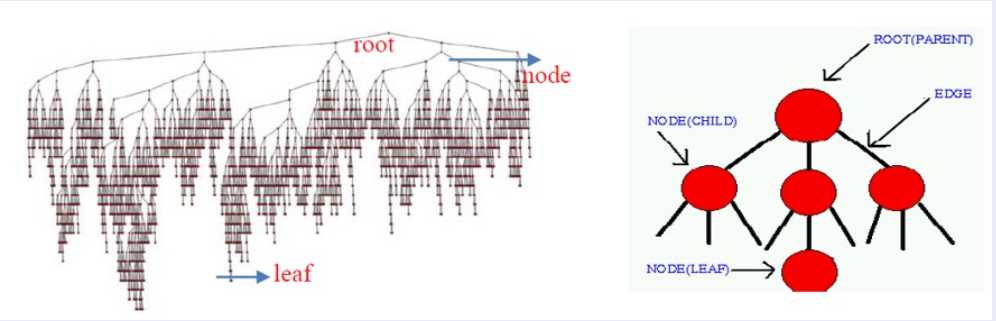
T.D.H.L: đóng góp vào việc thu thập dữ liệu, trích xuất dữ liệu; phân tích và giải thích dữ liệu, soạn thảo bài báo và sửa đổi bản thảo. L.H.P: đã giúp soạn thảo bản thảo và chỉnh sửa. Q.T.D: đã giúp soạn thảo bản thảo và chỉnh sửa. H.T.T: tham gia phân tích dữ liệu địa lý. Tất cả các tác giả đã phê duyệt cuối cùng để xuất bản.

PHỤ LỤC

Bảng 5, Hình 7, 8 và 9

TÀI LIỆU THAM KHẢO

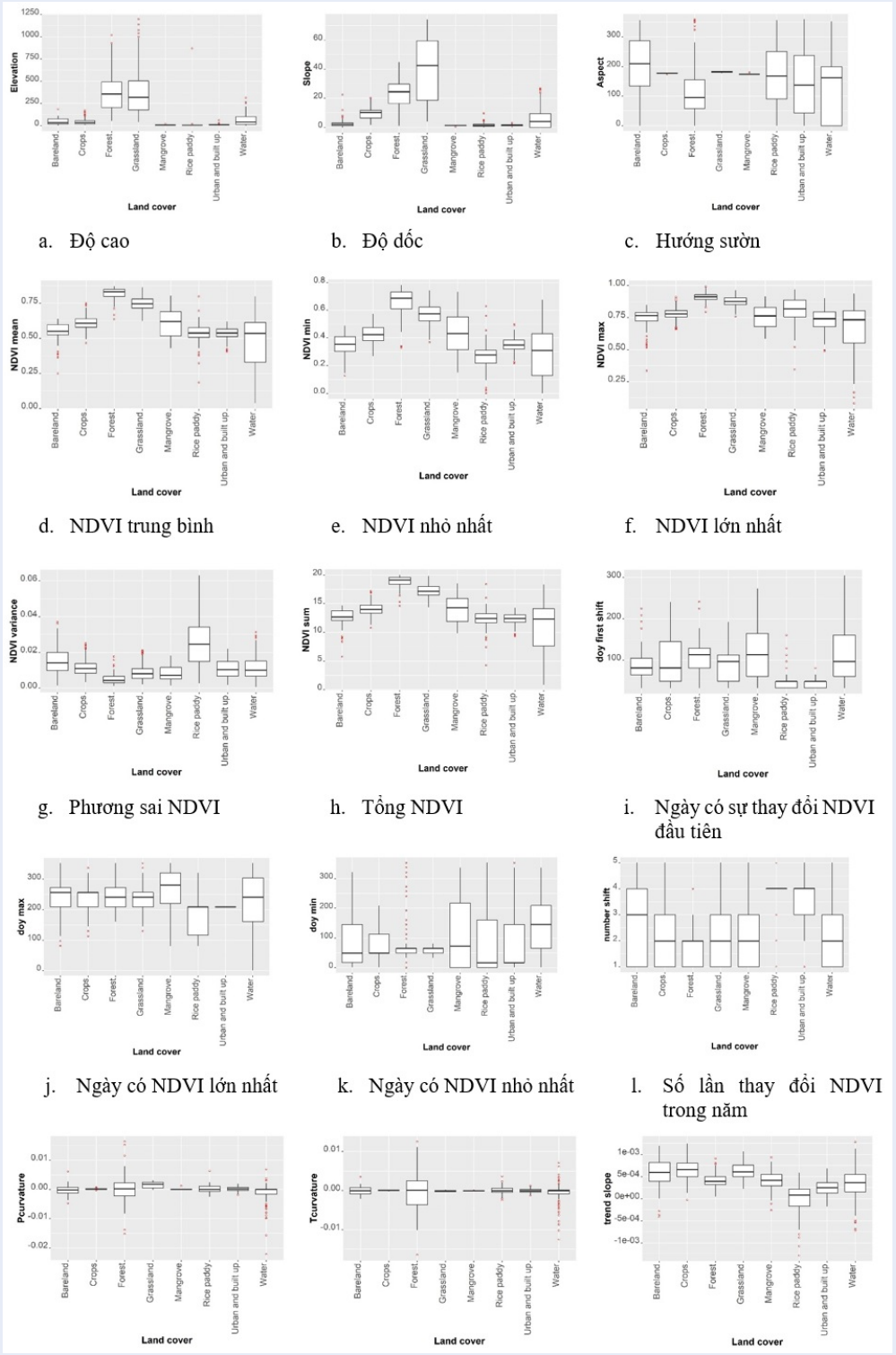
1. Foley JA et al. Solutions for a cultivated planet. Nature 2011; 478, 337-342;PMID: 21993620. Available from: https://doi.org/10.1038/nature10452.
2. Rujoiu-Mare M-R, Mihai B-A. Mapping Land Cover Using Remote Sensing Data and GIS Techniques: A Case Study of Prahova Subcarpathians. Procedia Environmental Sciences. 2016; 32, 244-255;Available from: https://doi.org/10.1016/j.proenv.2016.03.029.
3. Breiman L. Random Forests. Machine Learning 2001; 45, 5-32;Available from: https://doi.org/10.1023/A:1010933404324.
4. Liaw A, Wiener M. In press. Classification and Regression by Random Forest. R News. 18-22;.
5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction,. 2nd edn. New York: NY: Springer. 2009;Available from: https://doi.org/10.1007/978-0-387-84858-7.
6. Keshtkar H, Voigt W, Alizadeh E. Land-cover classification and analysis of change using machine-learning classifiers and multi-temporal remote sensing imagery. Arab J Geosci. 2017; 10, 154;Available from: https://doi.org/10.1007/s12517-017-2899-y.
7. Fichera CR, Modica G, Pollino M. Land Cover classification and change-detection analysis using multi-temporal remote sensed imagery and landscape metrics. European Journal of Remote Sensing. 2012; 45, 1-18;Available from: https://doi.org/10.5721/EuJRS20124501.
8. Yang X, Lo CP. Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. International Journal of Remote Sensing. 2002; 23, 1775-1798;Available from: https://doi.org/10.1080/01431160110075802.
9. Fonji SF, Taff GN. Using satellite data to monitor land-use land-cover change in North-eastern Latvia. SpringerPlus. 2014; 3, 61;PMID: 24567875. Available from: https://doi.org/10.1186/2193-1801-3-61.
10. Schulz C, Holtgrave A-K, Kleinschmit B. Large-scale winter catch crop monitoring with Sentinel-2 time series and machine learning-An alternative to on-site controls? Computers and Electronics in Agriculture. 2021; 186, 106173;Available from: https://doi.org/10.1016/j.compag.2021.106173.
11. Abdi AM. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GIScience & Remote Sensing. 2020; 57, 1-20;Available from: https://doi.org/10.1080/15481603.2019.1650447.
12. Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, Huang X. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment. 2010; 114, 168-182;Available from: https://doi.org/10.1016/j.rse.2009.08.016.
13. Jin Y, Liu X, Chen Y, Liang X. Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong. International Journal of Remote Sensing. 2018; 39, 8703-8723;Available from: https://doi.org/10.1080/01431161.2018.1490976.
14. In press. MOD13Q1 v061 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid. USGS, Earthdata;Available from: https://lpdaac.usgs.gov/products/mod13q1v061/.
15. Pittman K, Hansen MC, Becker-Reshef I, Potapov PV, Justice CO. Estimating Global Cropland Extent with Multi-year MODIS Data. Remote Sensing. 2010; 2, 1844-1863;Available from: https://doi.org/10.3390/rs2071844.
16. Zeferino LB, Souza LFT de, Amaral CH do, Fernandes Filho EI, Oliveira TS de. Does environmental data increase the accuracy of land use and land cover classification? International Journal of Applied Earth Observation and Geoinformation. 2020; 91, 102128;Available from: https://doi.org/10.1016/j.jag.2020.102128.
17. Chi M, Feng R, Bruzzone L. Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. Advances in Space Research. 2008; 41, 1793-1799;Available from: https://doi.org/10.1016/j.asr.2008.02.012.
18. Tran H, Tran T, Kervyn M. Dynamics of Land Cover/Land Use Changes in the Mekong Delta, 1973-2011: A Remote Sensing Analysis of the Tran Van Thoi District, Ca Mau Province, Vietnam. Remote Sensing. 2015; 7, 2899-2925;Available from: https://doi.org/10.3390/rs70302899.
19. Hermosilla T, Wulder MA, White JC, Coops NC, Hobart GW. Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series. Canadian Journal of Remote Sensing. 2018; 44, 67-87;Available from: https://doi.org/10.1080/07038992.2018.1437719.
20. Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F, Atkinson PM, Jeganathan C. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sensing of Environment. 2012; 121, 93-107;Available from: https://doi.org/10.1016/j.rse.2011.12.003.
21. Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira JMC. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management. 2012; 275, 117-129;Available from: https://doi.org/10.1016/j.foreco.2012.03.003.
22. Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment. 1991; 37, 35-46;Available from: https://doi.org/10.1016/0034-4257(91)90048-B.
23. Olofsson P, Foody GM, Stehman SV, Woodcock CE. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sensing of Environment. 2013; 129, 122-131;Available from: https://doi.org/10.1016/j.rse.2012.10.031.
24. R Core Team. R: The R Project for Statistical Computing. Foundation for Statistical Computing, Vienna, Austria. 2021;Available from: https://www.r-project.org/.
25. Tatsumi K, Yamashiki Y, Canales Torres MA, Taipe CLR. Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. Computers and Electronics in Agriculture. 2015; 115, 171-179;Available from: https://doi.org/10.1016/j.compag.2015.05.001.
26. Ghimire B, Rogan J, Miller J. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sensing Letters. 2010; 1, 45-54;Available from: https://doi.org/10.1080/01431160903252327.
27. Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sensing of Environment. 2008; 112, 2272-2283;Available from: https://doi.org/10.1016/j.rse.2007.10.004.



Hình 7: Sơ đồ cây quyết định

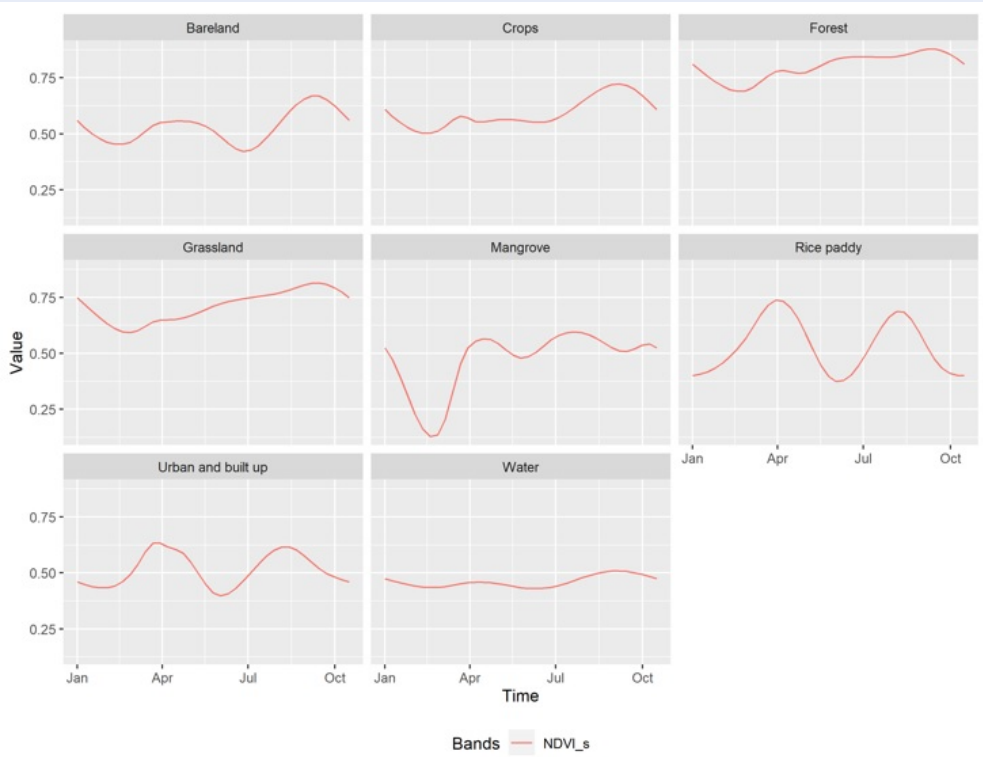
Bảng 5: Kết quả huấn luyện với các giá trị mtry khác nhau

mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.889365	0.869351	0.026683	0.031619
2	0.904025	0.886931	0.026592	0.031385
3	0.910307	0.894421	0.025408	0.02993
4	0.912087	0.896516	0.024173	0.028472
5	0.912387	0.896878	0.024967	0.029406
6	0.911798	0.896184	0.024431	0.02877
7	0.909409	0.893361	0.024423	0.028768
8	0.909698	0.893692	0.023392	0.027542
9	0.907918	0.891584	0.022731	0.026784
10	0.907617	0.891231	0.025119	0.029576
11	0.905518	0.88875	0.025639	0.030202
12	0.901294	0.88378	0.025884	0.030465
13	0.897699	0.879554	0.025567	0.030127
14	0.894704	0.876054	0.026976	0.031733
15	0.893794	0.874967	0.029433	0.034647



a. Độ cao
b. Độ dốc
c. Hướng sườn
d. NDVI trung bình
e. NDVI nhỏ nhất
f. NDVI lớn nhất
g. Phương sai NDVI
h. Tổng NDVI
i. Ngày có sự thay đổi NDVI đầu tiên
j. Ngày có NDVI lớn nhất
k. Ngày có NDVI nhỏ nhất
l. Số lần thay đổi NDVI trong năm

Hình 8: Các biến phụ trợ trong RF



Hình 9: Xu hướng NDVI theo thời gian của các đối tượng đất phủ

Science & Technology Development Journal – Science of The Earth & Environment 5(SI3):40-53

Open Access Full Text Article

Research Article

Land-cover classification using Random Forest and incorporating NDVI time-series and topography: a case study of Thanh Hoa province, Vietnam

Trong Dieu Hien Le[1,*], Luan Hong Pham[2], Hoang Thi Tuyet[1], Quang Toan Dinh[3]

ABSTRACT

Land cover/land use (LULC) mapping in the complex land cover area is a challenging task due to the mixed vegetation patterns, and rough mountains with fast-flowing rivers. Therefore, a new technique should be applied to improve the accurate classification of complex LULC. In this study, we applied a supervised machine learning approach to map land use in Thanh Hoa province, Vietnam utilizing multi-temporal Normalized Difference Vegetation Index (NDVI) data from MODIS, combined with topographic features. We used distinctive temporal features of land cover in 2015 as response variables and developed fifteen engineering features as predictors for automatic prediction. Then, we trained Random Forest classification (RFC) and conducted repeated cross-validation to identify the optimal RFC with the highest robustness on test data. RFC reached a total prediction accuracy of 91 % and Kappa coefficient (K) of 0.89 across eight different land covers including bareland, crops, rice paddy, forest, mangrove, urban and built up, grassland, and water. Besides, the results showed that the features extracted from time-series NDVI comprising the mean of yearly NDVI, the sum of NDVI, and the topography were the relative importance variables controlling the land cover classification.

Key words: MODIS, land cover classification, Random Forest, remote sensing

[1]*Faculty of Resources & Environment, University of Thu Dau Mot, 06 Tran Van On street, Thu Dau Mot City, Binh Duong, Viet Nam*

[2]*Center of Water Management and Climate Change, Institute of Environment and Resources, National University of Ho Chi Minh city, Vietnam, 01 Marie Curie, Linh Trung ward, Thu Duc district, Ho Chi Minh city, Vietnam.*

[3]*Department of Natural Resources and Environment of Thanh Hoa, Thanh Hoa 400570, Vietnam.*

Correspondence

Trong Dieu Hien Le, Faculty of Resources & Environment, University of Thu Dau Mot, 06 Tran Van On street, Thu Dau Mot City, Binh Duong, Viet Nam

Email: hienltd@tdmu.edu.vn

History

- Received: 12-12-2021
- Accepted: 06-3-2022
- Published: 30-6-2022

DOI : 10.32508/stdjsee.v5iSI2.681

Cite this article : Le T D H, Pham L H, Tuyet H T, Dinh Q T. **Land-cover classification using Random Forest and incorporating NDVI time-series and topography: a case study of Thanh Hoa province, Vietnam**. *Sci. Tech. Dev. J. - Sci. Earth Environ.;* 5(SI3):40-53.