

Tích hợp mô hình học máy và hệ thống địa lý 3
chiều trong dự đoán xu hướng đô thị hóa của
TP.HCM

Lê Hải Nam^{1*}, Trần Nguyễn Yên Nhi^{2*}, Nguyễn Gia Tuấn Anh^{3†},
Lưu Thanh Sơn^{4†}, Phạm Vĩ^{5†}

^{1,2*}Khoa học và kĩ thuật thông tin, Trường đại học Công nghệ thông tin.
^{3,4,5}Khoa học và kĩ thuật thông tin, Trường đại học Công nghệ thông tin.

*Corresponding author(s). E-mail(s): 21522357@gm.uit.edu.vn;
21522429@gm.uit.edu.vn;

Contributing authors: anhngt@uit.edu.vn; sonlt@uit.edu.vn;
19521101@gm.uit.edu;

† Đồng hướng dẫn.

Bài báo này trình bày việc tích hợp viễn thám, GIS (Hệ thống thông tin địa lý) và mô hình học máy để phân tích xu thế và đưa ra giải pháp cho sự đô thị hóa và quy hoạch tại thành phố Hồ Chí Minh. Dữ liệu thu thập được từ năm 2016 đến 2021 kết hợp với dữ liệu Sentinel-2 đa thời gian chồng lấp thông qua bề mặt không thâm để theo dõi và phân tích quá trình quy hoạch đất trong 5 năm này. Bài báo giúp giải quyết các vấn đề xác định xu thế phát triển của thành phố Hồ Chí Minh, đồng thời, giám sát và phân tích mặt không thâm trong việc nghiên cứu xu thế mở rộng không gian đô thị.

Từ khóa: Google Earth Engine[1], Đất phủ/đất sử dụng (LC/LU), dân số trung bình, NIR.

1 GIỚI THIỆU

Ở Việt Nam, trong hai thập kỷ gần đây, sự phát triển kinh tế và tăng trưởng dân số đã tác động lớn đến việc sử dụng đất, và theo xu thế đó, việc sử dụng và quy hoạch sẽ còn diễn ra nhiều và nhanh trong tương lai. Nhận thấy thành phố Hồ Chí Minh là khu vực có khả năng phát triển kinh tế lớn và cần lập kế hoạch quy hoạch để hỗ trợ sự phát triển, cùng với đó là sự khó khăn trong tái quy hoạch do vấn đề đô thị

hóa tự phát còn diễn ra nhiều dẫn đến hậu quả là du nhập dân cư ồ ạt, khó quản lý, nhóm chúng tôi đã chọn nghiên cứu về thành phố này để xây dựng kế hoạch đô thị hóa trong thời gian tới.

Để định hướng sự phát triển đó, cần có thông tin dữ liệu về phạm vi và tính chất của diện tích đất, tuy nhiên, ở đất nước đang phát triển như Việt Nam - một quốc gia thiếu nguồn lực và cơ sở hạ tầng để tạo dữ liệu tin cậy - rất khó để tính toán mức độ đô thị hóa. Với các dữ liệu viễn thám công khai có sẵn để lập bản đồ đất phủ / đất sử dụng (LC/LU) cho thành phố Hồ Chí Minh, nhóm chúng tôi đã lấy dữ liệu công khai từ Google Earth Engine và dữ liệu quốc gia từ cục thống kê Hồ Chí Minh, sau đó phân tích dựa trên 3 nhóm yếu tố ảnh hưởng tích cực đến quy hoạch sử dụng đất thành phố Hồ Chí Minh theo góc nhìn cán bộ quản lý gồm: nhóm yếu tố xã hội, nhóm yếu tố kinh tế, nhóm yếu tố môi trường.^[2]

Nhóm chúng tôi áp dụng công nghệ viễn thám đa phổ, đa thời gian với khả năng giám sát biến động của các đối tượng mặt đất kết hợp GIS để phân tích quá trình đô thị hóa tại thành phố Hồ Chí Minh. Quá trình đô thị hóa chính là cơ hội để Nhà nước tổ chức, quy hoạch lại cách thức hoạt động của đô thị, dân cư. Theo đó, các khu vực có tiềm năng phát triển kinh tế, văn hóa - xã hội sẽ được quy hoạch theo hệ thống hiện đại. Các khu vực chưa có điều kiện kinh tế xã hội hoặc mật độ dân số thấp sẽ được điều chỉnh các ngành nghề, quy hoạch phù hợp để tăng cơ hội phát triển trong tương lai.^[3]

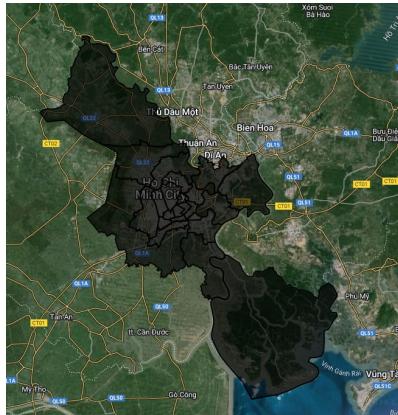
Do sự không đồng nhất của môi trường, việc phân loại các loại lớp phủ đô thị từ dữ liệu viễn thám thường gặp nhiều thách thức. Môi trường đô thị rất phức tạp bởi sự kết hợp của nhiều loại vật liệu và các lớp phủ khác nhau, tạo ra cảnh quan đô thị với sự đa dạng của các lớp phủ có phỏ phản xạ khác nhau. Các bề mặt không thấm (mái nhà, lối đi bộ, đường giao thông, bãi đỗ xe,...) được phủ bởi các vật liệu như nhựa đường, bê tông, đá và các vật liệu xây dựng khác, có các đặc tính vật lý đặc biệt và được nhận biết như là các thực thể riêng biệt trên dải quang phổ điện tử. Chính vì thế, bài báo của chúng tôi sử dụng mặt không thấm để theo dõi quá trình đô thị hóa của thành phố Hồ Chí Minh.^[4]

Theo Wikipedia, **đô thị hóa** là sự mở rộng của **đô thị**, tính theo tỉ lệ phần trăm giữa số dân đô thị hay diện tích đô thị trên tổng số dân hay diện tích của một vùng hay khu vực. Ta cũng có thể tính theo tỉ lệ gia tăng của hai yếu tố đó theo thời gian. Nếu tính theo cách đầu thì nó còn được gọi là **mức độ đô thị hóa**; còn theo cách thứ hai, nó có tên là **tốc độ đô thị hóa**.^[5]

2 KHU VỰC NGHIÊN CỨU, DỮ LIỆU VÀ PHƯƠNG PHÁP

2.1 Khu vực nghiên cứu

Thành phố Hồ Chí Minh, thành phố lớn nhất Việt Nam và là một siêu đô thị trong tương lai gần. Thành phố Hồ Chí Minh là thành phố trực thuộc trung ương thuộc loại đô thị đặc biệt của Việt Nam với tổng diện tích là 2,095km². Theo kết quả điều tra dân số sơ bộ vào năm 2021 thì dân số thành phố là 9.166.800 người (chiếm 9,3% dân số Việt Nam), mật độ dân số trung bình 4.375 người/km² (cao nhất cả nước).



Hình 1: Bản đồ hành chính TP. HCM

Tuy nhiên, nếu tính những người cư trú không đăng ký hộ khẩu thì dân số thực tế của thành phố này năm 2018 là gần 14 triệu người. Tọa độ của thành phố này là $10^{\circ}46'10''B$, $106^{\circ}40'55''D$. Tuy nhiên, sự đô thị ở từng quận huyện của mỗi khu vực là không đồng đều, có những khu vực còn có tiềm năng phát triển đô thị và có những quận có cơ sở hạ tầng xuống cấp trầm trọng do đô thị hóa tự phát diễn ra, chính vì thế, nhóm chúng tôi quyết định chọn thành phố Hồ Chí Minh làm khu vực nghiên cứu.

2.2 Dữ liệu

Dữ liệu sử dụng cho việc nghiên cứu này bao gồm: (1) Dữ liệu từ Google Earth Engine. (2) Dữ liệu dạng bảng từ Cục Thống kê TP. Hồ Chí Minh từ năm 2016 đến 2021. Hai nguồn dữ liệu này đáng tin cậy và phù hợp với QGIS trong việc tính toán và truy xuất thông tin. Ngoài ra, chúng tôi còn sử dụng dữ liệu độ cao từ (3) Nasa SRTM Digital Elevation 30m.

2.2.1 Dữ liệu từ Google Earth Engine.

Sentinel-2 thuộc chương trình Copernicus nhằm thu thập hình ảnh quang học ở độ phân giải không gian cao (10m đến 60m) trên đất liền và vùng nước ven biển. Chính vì thế dữ liệu của nhóm thu thập được có độ phân giải cao, ngoài ra, còn được trải qua các bước tiền xử lý để nâng cao chất lượng hình ảnh gồm nhóm các pixel, hiệu chỉnh hình học, tính toán các chỉ số NDVI, NDWI, NDBI, BSI. [6]

Nhóm chúng tôi dùng bản đồ lớp đất che phủ được phân loại lấy từ dữ liệu Sentinel-2 có thể được phân tích sâu hơn để trích xuất thông tin có giá trị về những thay đổi trong quy hoạch đô thị. Các bước xử lý bao gồm tổng hợp dữ liệu, phát hiện thay đổi, phân tích chuỗi thời gian hoặc lập mô hình không gian để hiểu rõ hơn về các kiểu che phủ đất được quan sát.

2.2.2 Dữ liệu bảng từ Tổng cục Thống kê.

Tổng cục Thống kê (General Statistics Office, viết tắt là **GSO**) thực hiện chức năng tham mưu, điều phối hoạt động thống kê,... Dữ liệu về dân số như dân số trung bình, mật độ dân số hay về đất như diện tích đất hằng năm được nhóm chúng tôi thu thập và phân tích theo các phân cụm từng khu vực để thực hiện nghiên cứu.[7]

2.2.3 Dữ liệu Nasa SRTM Digital Elevation 30m

Nasa SRTM Digital Elevation 30m (Dữ liệu độ cao kỹ thuật số của Nhiệm vụ Địa hình) là một nghiên cứu quốc tế nhằm thu được các mô hình độ cao kỹ thuật số ở quy mô gần như toàn cầu. Sản phẩm SRTM V3 (SRTM Plus) này được NASA JPL cung cấp ở độ phân giải 1 cung giây (khoảng 30m).

2.2.4 Phương trình tính toán.

Chúng tôi đã áp dụng các phương trình tính toán được trình bày như sau:

Normalized Difference Vegetation Index (NDVI) là chỉ số thực vật khác biệt được chuẩn hóa, đây là thước đo được sử dụng rộng rãi để định lượng tình trạng và mật độ thực vật bằng dữ liệu cảm biến. Nó được tính toán từ dữ liệu quang phổ ở hai dải cụ thể: đỏ (RED) và cận hồng ngoại (Near-Infrared viết tắt là NIR). Một số tùy chọn được xác định cho NDVI, tùy vào băng tần NIR và RED được tìm thấy:

NDVI (Sentinel 2 MSI): NIR=B8, RED=B4

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$$

Normalized Difference Water Index (NDWI) là chỉ số nước chênh lệch chuẩn hóa có thể đề cập đến một trong ít nhất hai chỉ số thu được từ viễn thám liên quan đến nước lỏng: Một được sử dụng để theo dõi sự thay đổi hàm lượng nước của lá, sử dụng bước sóng hồng ngoại gần và sóng ngắn (Short Wave Infrared, viết tắt là SWIR) , được lấy từ các kênh Cận hồng ngoại (NIR) và xanh (GREEN, viết tắt là G). Công thức này chỉ rõ lượng nước trong các vùng nước. NDWI (Sentinel 2 MSI): G=B3, NIR=B8.

$$\text{NDWI} = \frac{\text{G} - \text{NIR}}{\text{G} + \text{NIR}}$$

Phương pháp khác: McFeeters 2007 (MNDWI)

Tuy nhiên, hạn chế lớn của NDWI là nó không triệt tiêu nhiễu tín hiệu đến từ đặc điểm che phủ đất. MNDWI được tạo ra bằng cách sử dụng hai dải sóng phổ khác nhau: dải màu xanh (GREEN) và sóng ngắn (SWIR). Khi áp dụng MNDWI, nước thường có độ phản xạ trong dải xanh lục (GREEN) cao hơn so với dải SWIR, trong khi đất khô thường có độ phản xạ tương đối thấp hơn trong cả hai dải sóng phổ.

$$\text{MNDWI} = \frac{\text{GREEN} - \text{SWIR}}{\text{GREEN} + \text{SWIR}}$$

Công thức của MNDWI tính toán sự khác biệt đối với độ phản xạ giữa dải màu xanh lục (GREEN) và dải SWIR, được chuẩn hóa bằng tổng của cả hai độ phản xạ. Kết

quả của MNDWI thường dương khi ánh sáng được phản xạ từ nước, và âm khi ánh sáng được phản xạ từ đất khô.

Công thức này cho phép MNDWI phân biệt nước và đất khô một cách hiệu quả trên hình ảnh viễn thám, với kết quả thường được sử dụng để tạo ra các bản đồ phân loại nước và đất khô, đặc biệt trong các ứng dụng như giám sát tài nguyên nước và quản lý môi trường.

Normalized Difference Built-Up Index (NDBI) là chỉ số chuẩn hóa phân biệt khu vực xây dựng, đây là một chỉ số phổ biến đặc biệt thường được sử dụng trong viễn thám để phân biệt các khu vực xây dựng từ các loại bề mặt đất khác.

$$\text{NDBI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}}$$

The Bare Soil Index (BSI) là chỉ số đất trắn. BSI kết hợp các dải màu xanh lam, đỏ, hồng ngoại gần (NIR) và hồng ngoại sóng ngắn (SWIR) để ghi lại các biến thể của đất. Các dải này được sử dụng theo cách chuẩn hóa, trong đó dải SWIR và dải màu đỏ được sử dụng để định lượng thành phần khoáng chất trong đất, trong khi các dải màu xanh lam và NIR được sử dụng để tăng cường sự hiện diện của thực vật. Chỉ báo số này có thể được sử dụng để lập bản đồ đất và xác định cây trồng.

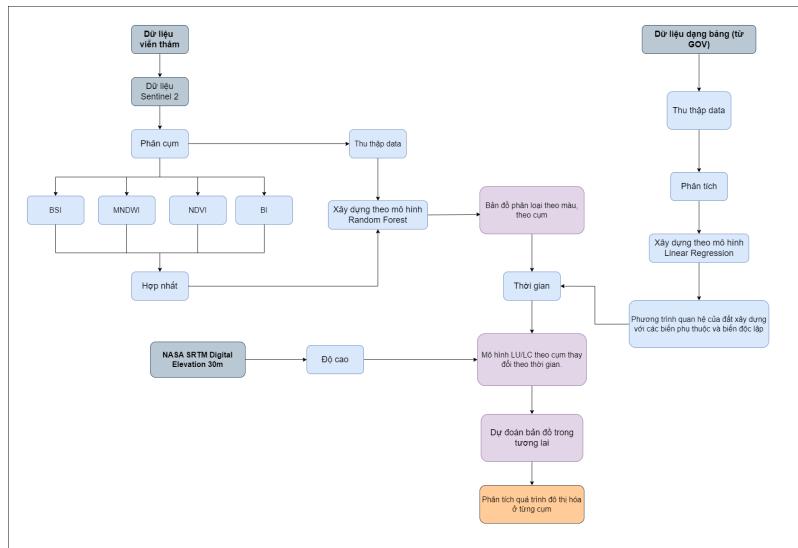
$$\text{BSI} = \frac{(\text{Red} + \text{SWIR}) - (\text{NIR} + \text{Blue})}{(\text{Red} + \text{SWIR}) + (\text{NIR} + \text{Blue})}$$

2.3 Phương pháp

2.3.1 Phương pháp

Trong nghiên cứu này chúng tôi sử dụng 3 dữ liệu đầu vào là dữ liệu viễn thám, dữ liệu dạng bảng từ HoChiMinhGOV ngoài ra còn dùng dữ liệu từ NASA SRTM Digital Elevation 30m. Các dữ liệu viễn thám sẽ được chúng tôi phân cụm, sau đó xử lý màu và hợp nhất. Ngoài ra, khi phân cụm chúng tôi sẽ tiếp tục tiến hành thu thập data và xây dựng theo mô hình Random Forest. Từ đó, chúng tôi tiếp tục xây dựng bản đồ phân loại theo màu và cụm, xây dựng theo khung thời gian (time series). (1)

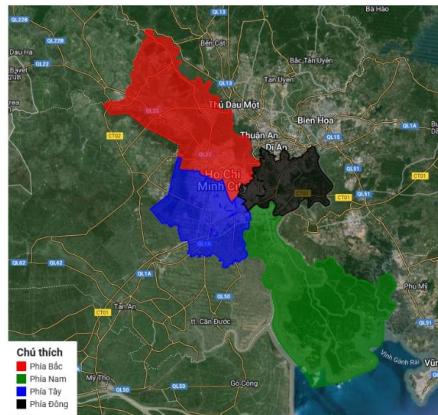
Từ dữ liệu dạng bảng của cục Thống kê, chúng tôi tiến hành thu thập, phân tích, xây dựng mô hình Linear Regression, từ đó, xây dựng phương trình quan hệ của đất xây dựng với các biến phụ thuộc và biến độc lập. (2) Từ (1) và (2), chúng tôi gộp lại theo mô hình time series để dựng mô hình LU/LC theo cụm thay đổi theo thời gian. Đồng thời lấy độ cao từ NASA SRTM Digital Elevation để dựng mô hình trên, từ đó, dự đoán được bản đồ trong tương lai và phân tích quá trình đô thị hóa ở từng cụm.



Hình 2: Phương pháp thực hiện.

2.3.2 Phương pháp phân cụm

Chúng tôi chia bản đồ làm 4 cụm gồm: phía Đông, phía Tây, phía Nam, phía Bắc như hình bên dưới.



Hình 3: Mô tả 4 cụm và chú thích.

1. Phía Đông:

Gồm 7 quận huyện, thành phố là Thành phố/Quận Thủ Đức, Quận 1, Quận 2, Quận 3, Quận 9, Phú Nhuận, Bình Thạnh. Hiện nay, Quận 2, Quận 9 và Quận Thủ Đức đã được gộp lại thành Thành phố Thủ Đức, tuy nhiên do lấy số liệu từ

2016 nên chúng tôi vẫn để 3 quận này riêng biệt (trừ năm 2020 và 2021, thành phố Thủ Đức được sáp nhập).

2. *Phía Tây:*

Gồm 8 quận huyện: Quận 5, Quận 6, Quận 8, Quận 10, Quận 11, Bình Tân, Bình Chánh, Tân Phú.

3. *Phía Nam:*

Gồm 4 quận huyện: Quận 4, Quận 7, Nhà Bè, Cần Giờ.

4. *Phía Bắc:*

Gồm 5 quận huyện: Quận 12, Tân Bình, Hóc Môn, Gò Vấp, Củ Chi.

2.3.3 Phân loại lớp phủ mặt đất

Các loại lớp phủ mặt đất bao gồm: Vùng nước - water (0), vùng thực vật - vegetation (1), vùng xây dựng - buildup (2), vùng đất trống - bare (3).

2.4 Mô hình

2.4.1 Mô hình Logistic Regression

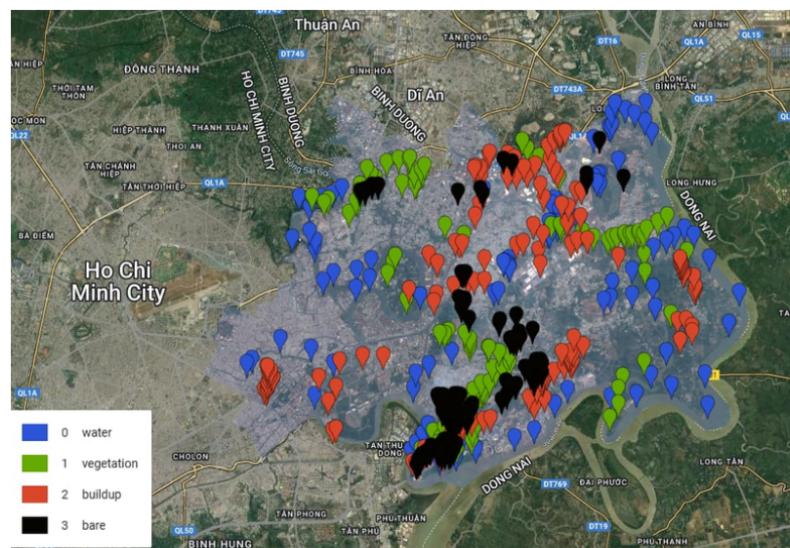
Sử dụng mô hình này để tính hệ số tương quan giữa đất xây dựng (đất ở, đất chuyên dùng) với các biến khác như: cơ sở kinh tế, mật độ dân số, dân số trung bình, đất chưa dùng và hệ số chặn (const). Từ đó có thể dự đoán được diện tích và tỉ lệ gia tăng của đất xây trong tương lai.

Tham số	coef
const	0.0086
Số cơ sở kinh tế	-0.0032
Mật độ dân số	0.0158
Đất chưa dùng	-1.2747
Dân số trung bình	-0.0005

Bảng 1: Bảng hệ số tương quan của đất xây dựng khu vực phía Đông

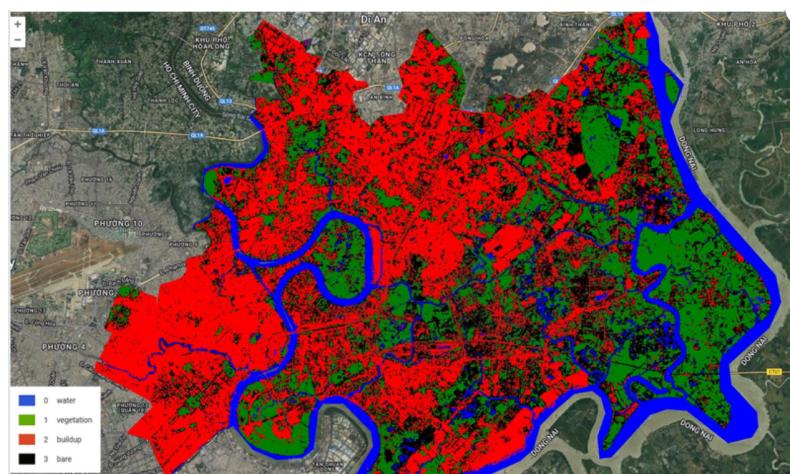
2.4.2 Mô hình Random Forest

Đầu tiên đánh nhãn cho các lớp phủ mặt đất dựa trên ảnh vệ tinh Sentinel-2 bằng các point (diểm) các point này sẽ ghi lại các giá trị quang phổ của các band màu: NDVI, MNDWI, NDBI, BSI tạo thành bộ dữ liệu mẫu. Bộ dữ liệu mẫu gồm 4 nhãn tương ứng với 4 lớp phủ mặt đất và mỗi nhãn sẽ có khoảng 100 point được đánh dấu đều trên khu vực.



Hình 4: Dán nhãn cho các lớp phủ mặt đất khu vực phía Đông

Từ bộ dữ liệu này dùng mô hình Random Forest để phân loại lớp phủ mặt đất cho toàn bộ khu vực theo từng năm.

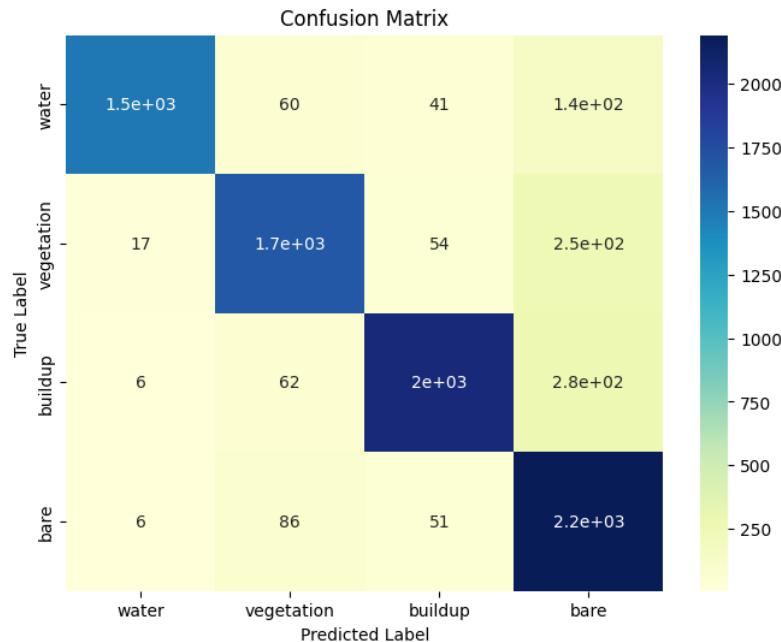


Hình 5: Phân loại đất phủ khu vực phía Đông năm 2018

2.4.3 Mô hình Time Series

Sử dụng các ảnh phân loại đất phủ từ năm 2018 đến năm 2021 để tạo ảnh sự thay đổi các nhãn qua các năm. Lấy các dữ liệu mẫu (3000 point mỗi nhãn) là dữ liệu các

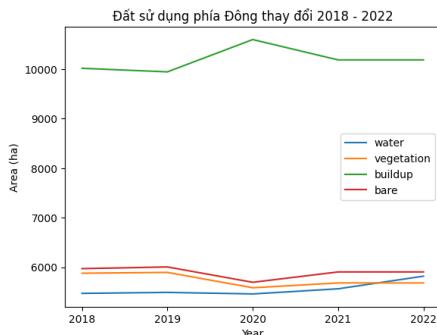
dữ liệu point ghi lại các giá trị quang phổ của từng nhãn từ các ảnh này đưa vào mô hình Random forest kết hợp với kết quả dự đoán tỉ lệ gia tăng đất xây dựng từ mô hình Logistic Regression để tạo mô hình dự đoán phân loại đất phủ trong tương lai.



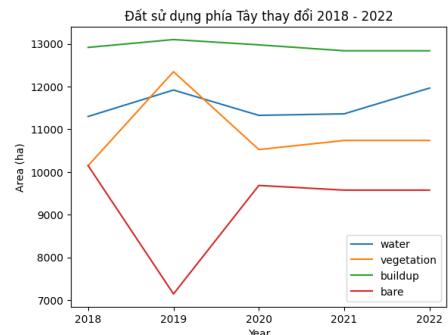
Hình 6: Confusion Matrix của mô hình dự đoán phân loại đất khu vực phía Đông

Có thể thấy được nhãn được dự đoán sai nhiều nhất là nhãn bare - khu đất trống, do giá trị quang phổ của các khu vực đất trống rất dễ nhầm lẫn sang khu vực xây dựng. Các khu vực đất trống thường sẽ thay đổi thành các khu vực thực vật thấp hay nước thấp do ảnh hưởng thời tiết ở năm trước, nhưng thật ra các khu vực này vẫn là các khu đất trống và sẽ có nhãn là đất trống ở năm kế tiếp chính vì sự thay đổi khá thất thường như vậy, nên nhãn này thường sẽ bị dữ đoán nhầm.

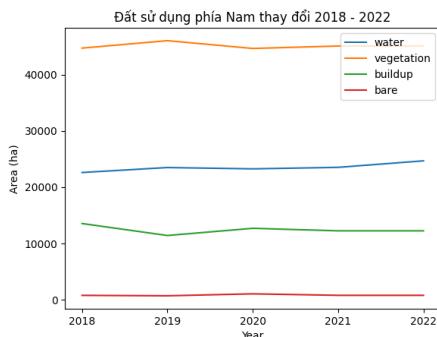
Ta có biểu đồ phân loại đất phủ qua thời gian theo từng khu:



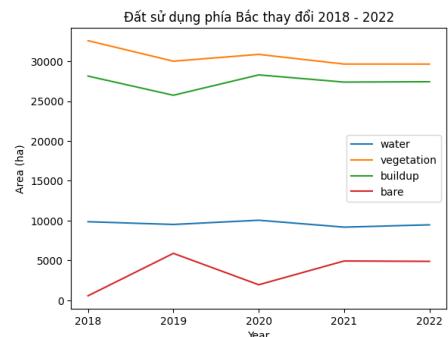
Hình 7: Biểu đồ sự thay đổi đất sử dụng ở phía Đông từ 2018-2022



Hình 8: Biểu đồ sự thay đổi đất sử dụng ở phía Tây từ 2018-2022



Hình 9: Biểu đồ sự thay đổi đất sử dụng ở phía Nam từ 2018-2022



Hình 10: Biểu đồ sự thay đổi đất sử dụng ở phía Bắc từ 2018-2022

Dựa vào biểu đồ trên có thể thấy xu hướng phát triển của thành phố Hồ Chí Minh đang lan tỏa dần về phía Đông Bắc thành.

2.5 Đánh giá và kết luận

Sau khi hoàn thành xây dựng mô hình, nhóm của chúng tôi đã đưa ra đánh giá và kết luận cho nghiên cứu này như sau:

Chúng tôi đã thành công trong việc kết hợp nhiều dải màu để xây dựng một mô hình phân loại đất phủ với độ chính xác lên đến 98.9%. Mô hình này được xây dựng dựa trên dữ liệu hình ảnh phân loại đất và dữ liệu xã hội theo thời gian, và đạt độ chính xác là 88%.

Ngoài ra, chúng tôi đã thành công trong việc giải quyết các vấn đề liên quan đến xác định xu hướng phát triển của thành phố Hồ Chí Minh.

2.6 Hướng phát triển

Dựa trên những đánh giá và kết luận từ nghiên cứu, chúng tôi đề xuất mở rộng và phát triển các hướng và tính năng sau đây trong nghiên cứu của chúng tôi. Đầu tiên, chúng tôi sẽ tập trung vào giải quyết vấn đề phân loại cho lớp đất trồng. Tiếp theo, chúng tôi đề xuất kết hợp dữ liệu địa chính để cải thiện hiệu suất dự đoán đất trồng trong tương lai. Chúng tôi cũng dự định áp dụng deep learning để tối ưu hóa mô hình học máy hiện tại, nhằm cải thiện khả năng dự đoán và phân loại. Cuối cùng, chúng tôi sẽ kết hợp các bài toán liên quan đến GIS và các bài toán xã hội để đề xuất các khu vực phù hợp cho việc xây dựng và tái xây dựng. Những nỗ lực này sẽ giúp chúng tôi không chỉ nghiên cứu mà còn áp dụng hiệu quả trong quản lý và phát triển đô thị.

Tài liệu

- [1] Google Earth Engine: Earth Engine Guides. <https://developers.google.com/earth-engine/guides>. Accessed: 2024-06-10 (2024)
- [2] Mu, L., Wang, L., Wang, Y., Chen, X., Han, W.: Urban land use and land cover change prediction via self-adaptive cellular based deep learning with multisourced data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(12), 5233–5247 (2019) <https://doi.org/10.1109/JSTARS.2019.2956318>
- [3] Goldblatt, R., Deininger, K., Hanson, G.: Utilizing publicly available satellite data for urban research: Mapping built-up land cover and land use in ho chi minh city, vietnam. Development Engineering **3**, 83–99 (2018) <https://doi.org/10.1016/j.deveng.2018.03.001>
- [4] Nguyễn, M., Trần, T., Phan, T.: Dánh giá kết quả thực hiện quy hoạch sử dụng đất đai thành phố cần thơ giai đoạn 2010-2020 từ góc nhìn cán bộ quản lý. Can Tho University Journal of Science **57**, 82–90 (2021) <https://doi.org/10.22144/ctu.jsi.2021.052>
- [5] Wikipedia contributors: Đô thị hóa. https://vi.wikipedia.org/wiki/%C4%90%C3%B4_th%E1%BB%8B_h%C3%B3a. Accessed: 2024-06-10 (2024)
- [6] Zhao, Z., Islam, F., Waseem, L.A., Tariq, A., Nawaz, M., Islam, I.U., Bibi, T., Rehman, N.U., Ahmad, W., Aslam, R.W., Raza, D., Hatamleh, W.A.: Comparison of three machine learning algorithms using google earth engine for land use land cover classification. Rangeland Ecology Management **92**, 129–137 (2024) <https://doi.org/10.1016/j.rama.2023.10.007>
- [7] Cổng thông tin điện tử Tổng cục Thống kê Thành phố Hồ Chí Minh: Tổng cục Thống kê Thành phố Hồ Chí Minh. <http://pso.hochiminhhcity.gov.vn/>. Accessed: 2024-06-10 (2024)