Dự báo AQI một/nhiều ngày tiếp theo sử dụng dữ liệu khí tượng, địa hình tại khu vực miền Bắc Việt Nam

Tóm tắt bài báo “The impact of meteorological conditions on Air Quality Index under different urbanization gradients: a case from Taipei”

1) Sơ lược

Nghiên cứu này tập trung vào việc đánh giá mối quan hệ giữa các điều kiện khí tượng (nhiệt độ, độ ẩm, tốc độ gió, lượng mưa) và mức độ ô nhiễm không khí (PM2.5, PM10, SO_2, CO, O_3, NO_2) tại Đài Bắc, Đài Loan. Bằng cách xem xét tác động của đô thị hóa đến sự khuếch tán ô nhiễm, nghiên cứu cung cấp thông tin hữu ích giúp cải thiện chất lượng môi trường đô thị.

2) Cách tiếp cận và phân chia theo mức độ đô thị hóa:

Nghiên cứu đã chọn thành phố Đài Bắc làm đối tượng nghiên cứu do đây là một thành phố toàn cầu hóa với mức độ đô thị hóa cao và cấu trúc đô thị rõ ràng.

Để phân tích theo mức độ đô thị hóa, các nhà nghiên cứu đã chọn ba khu vực nghiên cứu đại diện cho các mức độ đô thị hóa khác nhau:

- Zhongshan (Trung Sơn): khu vực trung tâm (downtown).
- Shilin (Sĩ Lâm): khu vực ngoại ô (suburbs).
- Yangmingshan (Dương Minh Sơn): khu vực vùng ven (outskirts).

Nghiên cứu đã thu thập dữ liệu về các yếu tố khí tượng (nhiệt độ trung bình, độ ẩm tương đối, tốc độ gió trung bình và lượng mưa) và các chất ô nhiễm không khí (PM2.5, PM10, SO2, CO, O3 và NO2) trong năm 2018 từ Cơ quan Bảo vệ Môi trường Hành chính Viện R.O.C.T (Đài Loan).

3) Cách sử dụng các phương pháp:

Chỉ số Chất lượng Không khí (AQI): Nghiên cứu đã sử dụng AQI để đơn giản hóa nồng độ của sáu chất ô nhiễm không khí thành một chỉ số duy nhất, giúp đánh giá điều kiện và xu hướng chất lượng không khí ngắn hạn ở các thành phố. AQI của Đài Bắc được tính toán dựa trên tiêu chuẩn US National Ambient Air Quality Standards (NAAQS).

Mô hình VAR (Vector Autoregression): Mô hình này đã được sử dụng để phân tích các hệ thống chuỗi thời gian có liên kết với nhau và tác động động của các nhiễu loạn ngẫu nhiên lên các biến số.

Kiểm định nhân quả Granger (Granger causality test): Đây là phương pháp chính để phân tích mối quan hệ nhân quả giữa các yếu tố khí tượng và AQI. Kiểm định này được sử dụng để xác định xem các giá trị trễ của một biến số có ảnh hưởng đến các giá trị hiện tại của một hoặc nhiều biến số khác hay không.

Hàm phản ứng xung (Impulse response function): Phương pháp này mô tả tác động của một biến nội sinh trong mô hình VAR lên các biến nội sinh khác. Nghiên cứu đã sử dụng nó để hiển thị các tác động cụ thể của các yếu tố khí tượng khác nhau lên AQI.

Phân tích phương sai (Variance decomposition): Phương pháp này phân tách sự thay đổi của các biến nội sinh thành các thành phần tác động của mô hình VAR, cung cấp thông tin về tầm quan trọng tương đối của mỗi nhiễu loạn ngẫu nhiên ảnh hưởng đến các biến số.

4) Áp dụng cách tiếp cận về sự gia tăng đô thị hóa:

Nghiên cứu đã so sánh kết quả phân tích (Granger causality, impulse response, variance decomposition) giữa ba khu vực có mức độ đô thị hóa khác nhau (trung tâm, ngoại ô, vùng ven) để hiểu rõ cách mức độ đô thị hóa khác nhau có thể điều chỉnh mối quan hệ giữa các yếu tố khí tượng và AQI. Kết quả cho thấy có sự khác biệt về các yếu tố khí tượng có tác động nhân quả Granger đến AQI ở mỗi khu vực.

Ví dụ, ở khu vực trung tâm Zhongshan, nhiệt độ và độ ẩm là nguyên nhân Granger của AQI, trong khi ở ngoại ô Shilin, độ ẩm và tốc độ gió là nguyên nhân Granger, và ở vùng ven Yangmingshan, chỉ có độ ẩm là nguyên nhân Granger.

5) AQI thay đổi theo từng mùa do các ngày lễ và điều kiện khuếch tán của không khí ảnh hưởng đến AQI:

Nghiên cứu chỉ ra rằng AQI ở Đài Bắc có sự thay đổi đáng kể theo mùa trong năm 2018, với giá trị cao nhất vào mùa đông và thấp nhất vào mùa hè. Điều này có thể liên quan đến các điều kiện khí tượng ở Đài Bắc.

Một lý do quan trọng khác được đề cập là sự tồn tại của nhiều lễ hội truyền thống của Trung Quốc vào mùa xuân, đặc biệt là Tết Nguyên Đán. Nghiên cứu của Huang và cộng sự (2012) đã xác nhận rằng việc đốt pháo hoa và pháo nổ trong các hoạt động truyền thống này làm tăng đáng kể lượng khí thải ô nhiễm không khí, dẫn đến sự gia tăng ô nhiễm không khí. Nguồn phát thải ô nhiễm không khí thay đổi và dao động lớn trước, trong và sau lễ hội, dẫn đến sự gia tăng AQI.

Nghiên cứu cũng chỉ ra rằng ô nhiễm không khí ở Đài Bắc chủ yếu là một quá trình tự tích tụ và tự khuếch tán. Hiệu ứng tự tích tụ của AQI chiếm hơn 70%. Khi điều kiện khuếch tán ô nhiễm không khí xấu đi, ô nhiễm không khí sẽ hình thành. Các yếu tố khí tượng như tốc độ gió có vai trò quan trọng trong việc khuếch tán và làm loãng các chất ô nhiễm. Ở khu vực trung tâm, do mật độ đô thị cao và cấu trúc đô thị, sự khuếch tán PM2.5 bằng tốc độ gió trở nên khó khăn hơn do thông gió kém. Ngược lại, ở vùng ngoại ô với mật độ đô thị thấp hơn, thông gió tốt hơn, giúp tốc độ gió có tác động rõ rệt hơn đến AQI.

6) Dữ liệu áp dụng

Dữ liệu cho nghiên cứu này đã được thu thập từ Cơ quan Bảo vệ Môi trường Hành chính Viện R.O.C.T (Đài Loan), có thể truy cập tại trang web https ://taqm.epa.gov.tw/taqm/tw/defau lt.aspx.
Dữ liệu được thu thập bao gồm:
Các yếu tố khí tượng:

- Nhiệt độ trung bình (AT)
- Độ ẩm tương đối (RH)
- Tốc độ gió trung bình (WS)
- Lượng mưa (RF)

Các chất ô nhiễm không khí:

- PM2.5 (Quan trọng nhất)
- PM10
- SO2
- CO
- O3
- NO2

Dữ liệu này được thu thập trong năm 2018 từ ba khu vực nghiên cứu ở thành phố Đài Bắc:

- Zhongshan (Trung Sơn): khu vực trung tâm (downtown)
- Shilin (Sĩ Lâm): khu vực ngoại ô (suburbs)
- Yangmingshan (Dương Minh Sơn): khu vực vùng ven (outskirts)

Dữ liệu đã được sử dụng như sau:

- Tính toán Chỉ số Chất lượng Không khí (AQI): Nồng độ của sáu chất ô nhiễm không khí (PM2.5, PM10, SO2, CO, O3 và NO2) đã được sử dụng để tính toán AQI theo công thức (1) dựa trên tiêu chuẩn US National Ambient Air Quality Standards (NAAQS). AQI được sử dụng để đại diện cho điều kiện chất lượng không khí ngắn hạn ở các khu vực khác nhau.

- Phân tích Thống kê (Mô hình VAR): Dữ liệu về các yếu tố khí tượng (nhiệt độ, độ ẩm, tốc độ gió, lượng mưa) và AQI đã được sử dụng để xây dựng mô hình VAR (Vector Autoregression). Mô hình VAR là một mô hình đa phương trình được sử dụng để phân tích mối quan hệ động giữa các chuỗi thời gian. Trong nghiên cứu này, mô hình VAR đã được áp dụng để phân tích tác động qua lại giữa các yếu tố khí tượng và AQI.

- Kiểm định Nhân quả Granger: Dựa trên kết quả của mô hình VAR, kiểm định nhân quả Granger đã được thực hiện để xác định xem liệu các yếu tố khí tượng có gây ra sự thay đổi trong AQI ở các khu vực đô thị hóa khác nhau hay không. Kiểm định này giúp xác định mối quan hệ nhân quả thống kê giữa các biến số. Kết quả kiểm định nhân quả Granger cho thấy các yếu tố khí tượng khác nhau có tác động nhân quả đến AQI ở từng khu vực (Bảng 1). Ví dụ, ở Zhongshan, nhiệt độ và độ ẩm là nguyên nhân Granger của AQI.

- Hàm Phản ứng Xung (Impulse Response Function): Phương pháp này đã được sử dụng để mô tả tác động của một cú sốc (một độ lệch chuẩn) trong một yếu tố khí tượng lên AQI theo thời gian. Các hình 5, 6 và 7 hiển thị phản ứng của AQI đối với những thay đổi trong nhiệt độ

(X1), độ ẩm (X2), tốc độ gió (X3) và lượng mưa (X4) ở ba khu vực khác nhau.

- Phân tích Phân rã Phương sai (Variance Decomposition): Phân tích này đã được sử dụng để xác định tỷ lệ phần trăm sự thay đổi trong AQI được giải thích bởi các yếu tố khí tượng khác nhau và chính AQI theo thời gian. Các bảng 2, 3 và 4 cho thấy mức độ ảnh hưởng của từng yếu tố khí tượng lên AQI ở mỗi khu vực qua 12 kỳ. Kết quả cho thấy AQI chủ yếu bị ảnh hưởng bởi chính nó (hiệu ứng tự tích tụ) và sau đó là các yếu tố khí tượng khác nhau tùy thuộc vào mức độ đô thị hóa. Ví dụ, tốc độ gió là yếu tố khí tượng chính ảnh hưởng đến AQI ở khu vực trung tâm và ngoại ô, trong khi độ ẩm lại quan trọng nhất ở vùng ven.

Áp dụng với bài toán đặt ra

1) Tóm tắt bài toán

Bài toán nhằm xây dựng hệ thống dự báo chỉ số chất lượng không khí (AQI) cho miền Bắc Việt Nam trong một hoặc nhiều ngày tiếp theo, dựa trên dữ liệu khí tượng và địa hình.

2) Phân tích và đánh giá dữ liệu

a) Dữ liệu đầu vào

Dữ liệu nồng độ PM2.5 đo được tại các trạm trong khoảng thời gian 2020-2021 được cung cấp dưới dạng bảng, trong đó mỗi trạm được đặc trưng bởi các thông tin sau:

- **ID trạm:** Mã định danh của từng trạm đo
- **Kinh độ, vĩ độ:** Vị trí địa lý của trạm đo.
- **Khoảng cách tới biển**: Khoảng cách từ trạm đo tới biển

Các biến khí tượng được sử dụng trong mô hình dự báo bao gồm:

- **WSPD** (Tốc độ gió): Đo lường tốc độ gió tại một điểm, có thể ảnh hưởng đến sự phân tán và di chuyển của các hạt bụi trong không khí.
- **WDIR** (Hướng gió): Chỉ ra hướng gió, giúp xác định hướng di chuyển của các chất ô nhiễm.
- **TX** (Nhiệt độ cao nhất): Nhiệt độ cao nhất trong ngày, ảnh hưởng đến sự bay hơi và quá trình trao đổi nhiệt trong khí quyển.

- **TP** (Tổng lượng mưa): Lượng mưa trong ngày, có thể làm sạch không khí nhưng cũng có thể ảnh hưởng đến sự di chuyển của các chất ô nhiễm.
- **TN** (Nhiệt độ thấp nhất): Nhiệt độ thấp nhất trong ngày, ảnh hưởng đến sự phát tán và ổn định của không khí.
- **TMP** (Nhiệt độ trung bình): Nhiệt độ trung bình trong ngày, giúp đánh giá các điều kiện nhiệt độ chung.
- **RH** (Độ ẩm tương đối): Mức độ ẩm trong không khí, có thể ảnh hưởng đến sự kết tụ của bụi và chất ô nhiễm.
- **PRES2M** (Áp suất): Đo lường áp suất khí quyển, ảnh hưởng đến sự chuyển động của không khí và điều kiện thời tiết.
- **HPBL** (Độ cao lớp tầng biên hành tinh): Mức độ của lớp không khí gần bề mặt trái đất, ảnh hưởng đến khả năng khuếch tán của chất ô nhiễm và bụi mịn PM2.5.

Các yếu tố khí tượng này đóng vai trò trong việc mô phỏng và dự báo chất lượng không khí, từ đó tính toán được chỉ số AQI tại miền Bắc Việt Nam.

b) Dữ liệu đầu ra

Dữ liệu đầu ra là nồng độ PM2.5 (bụi mịn có đường kính nhỏ hơn hoặc bằng 2.5 micromet) đo được tại các trạm trong khoảng thời gian 2020-2021.
Chỉ số PM2.5 này sau đó có thể được sử dụng để tính chỉ số AQI theo công thức dưới đây:

$$AQI_x = \frac{I_{i+1} - I_i}{BP_{i+1} - BP_i}(C_x - BP_i) + I_i \quad (Công\ thức\ 1)$$

Trong đó:

AQIx: Giá trị AQI

BPi: Nồng độ giới hạn dưới của giá trị thông số quan trắc được quy định trong Bảng 1 tương ứng với mức i

BPi+1: Nồng độ giới hạn trên của giá trị thông số quan trắc được quy định tương ứng với mức i+1

Ii: Giá trị AQI ở mức i đã cho trong bảng tương ứng với giá trị *BP*i

Ii+1: Giá trị AQI ở mức i+1 cho trong bảng tương ứng với giá trị BPi+1

Cx: là nồng độ PM2.5 đo được.

Bảng giá trị BP_i và I_i lấy trong bảng dưới đây:

i	I_i	Giá trị BP_i quy định đối với từng thông số (*Đơn vị: $\mu g/m^3$*)						
		O_3(1h)	O_3(8h)	CO	SO_2	NO_2	PM_{10}	$PM_{2.5}$
1	0	0	0	0	0	0	0	0
2	50	160	100	10.000	125	100	50	25
3	100	200	120	30.000	350	200	150	50
4	150	300	170	45.000	550	700	250	80
5	200	400	210	60.000	800	1.200	350	150
6	300	800	400	90.000	1.600	2.350	420	250
7	400	1.000	-	120.000	2.100	3.100	500	350
8	500	≥1.200	-	≥150.000	≥2.630	≥3.850	≥600	≥500

Bảng 1: Các giá trị BPi đối với các thông số

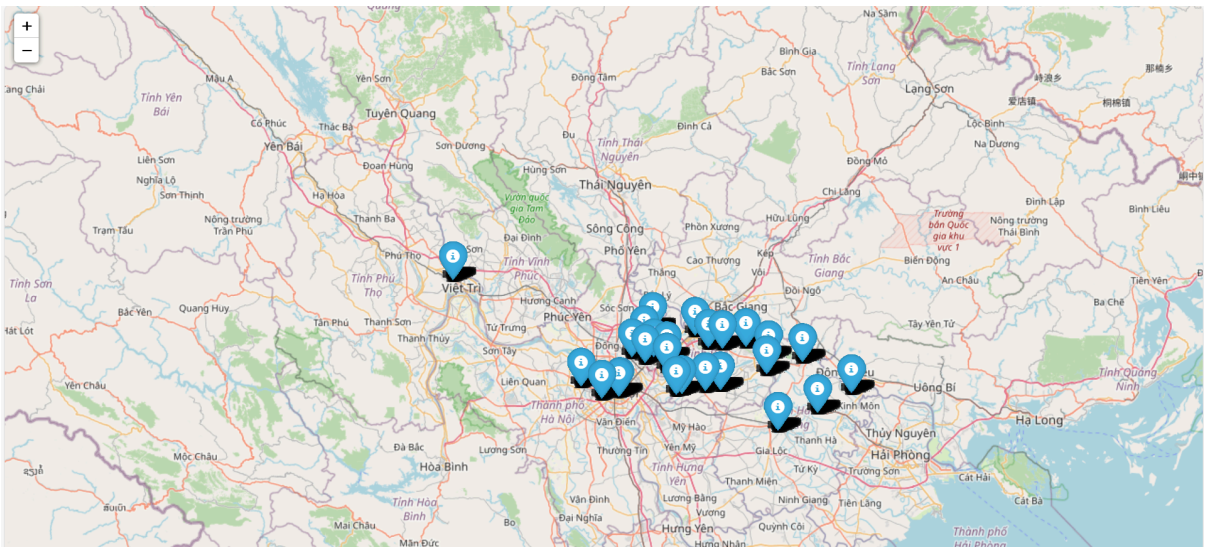
Sau khi tính được giá trị AQI, phân lớp thành 6 nhóm dựa trên giá trị AQI để làm labels cho bài toán dự báo AQI theo bảng sau:

Giá trị AQI	Đánh giá	Ảnh hưởng sức khỏe con người
0-50	Tốt	Chất lượng không khí tốt, không ảnh hưởng tới sức khỏe.
51-100	Trung bình	Chất lượng không khí ở mức chấp nhận được. Tuy nhiên, đối với những người nhạy cảm (người già, trẻ em, người mắc các bệnh hô hấp, tim mạch...) có thể chịu những tác động nhất định tới sức khỏe.
101-150	Kém	Những người nhạy cảm gặp phải các vấn đề về sức khỏe, những người bình thường ít ảnh hưởng.
151-200	Xấu	Những người bình thường bắt đầu có các ảnh hưởng tới sức khỏe, nhóm người nhạy cảm có thể gặp những vấn đề sức khỏe nghiêm trọng hơn.
201-300	Rất xấu	Cảnh báo ảnh hưởng tới sức khỏe: mọi người bị ảnh hưởng tới sức khỏe nghiêm trọng hơn.
300+	Nguy hại	Cảnh báo khẩn cấp về sức khỏe: Toàn bộ dân số bị ảnh hưởng tới sức khỏe tới mức nghiêm trọng.

Bảng 2: Bảng quy đổi giá trị AQI

c) Phân tích dữ liệu

Dữ liệu bao gồm tổng cộng 26 trạm đo khác nhau, với 11508 mẫu dữ liệu được thu thập trong khoảng thời gian từ năm 2020 đến 2021. Các mẫu dữ liệu này không được phân phối đều giữa các trạm. Tuy nhiên, tất cả các mẫu dữ liệu đều đầy đủ các biến khí tượng, mặc dù có một số ngày không có dữ liệu từ một số trạm nhất định. Vị trí của các trạm được đánh dấu trên bản đồ dưới đây dựa trên kinh độ và vĩ độ:



Dựa vào correlation matrix dưới đây, bài toán dự báo nồng độ PM2.5 cho thấy một mức độ phức tạp lớn nhưng có những yếu tố quan trọng cần được xem xét kỹ lưỡng.

	pm25	lat	lon	SQRT_SEA_DEM_LAT	WSPD	WDIR	TMP	TX	TN	TP	RH	PRES2M
pm25	1.000	-0.016	-0.073	0.122	-0.238	-0.046	-0.334	-0.280	-0.378	-0.242	-0.237	0.384
lat	-0.016	1.000	-0.442	0.389	-0.082	0.007	0.018	0.023	0.008	0.045	0.067	-0.147
lon	-0.073	-0.442	1.000	-0.163	0.155	-0.087	-0.058	-0.062	-0.046	-0.037	-0.028	0.123
SQRT_SEA_DEM_LAT	0.122	0.389	-0.163	1.000	0.065	-0.025	0.026	0.024	0.025	0.014	0.046	-0.009
WSPD	-0.238	-0.082	0.155	0.065	1.000	-0.112	0.022	0.002	0.067	0.023	0.026	-0.007
WDIR	-0.046	0.007	-0.087	-0.025	-0.112	1.000	0.303	0.286	0.299	0.080	0.199	-0.306
TMP	-0.334	0.018	-0.058	0.026	0.022	0.303	1.000	0.977	0.977	0.149	0.257	-0.903
TX	-0.280	0.023	-0.062	0.024	0.002	0.286	0.977	1.000	0.918	0.071	0.154	-0.870
TN	-0.378	0.008	-0.046	0.025	0.067	0.299	0.977	0.918	1.000	0.216	0.358	-0.893
TP	-0.242	0.045	-0.037	0.014	0.023	0.080	0.149	0.071	0.216	1.000	0.396	-0.250
RH	-0.237	0.067	-0.028	0.046	0.026	0.199	0.257	0.154	0.358	0.396	1.000	-0.319
PRES2M	0.384	-0.147	0.123	-0.009	-0.007	-0.306	-0.903	-0.870	-0.893	-0.250	-0.319	1.000

Một số yếu tố như nhiệt độ (TMP, TX, TN) và áp suất (PRES2M) có mối quan hệ mạnh mẽ nhất với PM2.5, điều này giúp việc dự báo trở nên dễ dàng hơn dù chúng vẫn có giá trị dưới mức trung bình.

Các yếu tố khác như tốc độ gió (WSPD), hướng gió (WDIR), độ ẩm (RH), và lượng mưa (TP) lại có mối quan hệ của chúng với PM2.5 khá yếu.

Các yếu tố địa lý như vĩ độ và kinh độ không có mối quan hệ mạnh với PM2.5, cho thấy rằng các yếu tố này không đóng vai trò quyết định trong việc dự báo chất lượng không khí, mặc dù chúng có thể tác động gián tiếp thông qua các yếu tố khác.

Nhìn chung với giá trị tương quan lớn nhất chỉ là 0.384, ta có thể xác định rằng bài toán này khá phức tạp và yêu cầu những mô hình học máy có độ phức tạp cao hoặc mô hình mạng nơ-ron sâu.

3) Phương pháp dự đoán

a) Tiền xử lý dữ liệu

Bài toán này nhằm dự đoán chỉ số chất lượng không khí (AQI) cho các ngày sắp tới. Để làm được điều này, nhóm đã xây dựng một bộ dữ liệu mới với các yếu tố đầu vào là dữ liệu khí tượng của 5 ngày gần nhất và tọa độ (kinh độ, vĩ độ) của các trạm đo. Kết quả đầu ra sẽ là các mức AQI được phân loại từ 1 đến 6, thể hiện mức độ ô nhiễm không khí trong 10 ngày tiếp theo.

Dữ liệu đầu vào: nhóm sử dụng dữ liệu khí tượng và chỉ số PM2.5 của 5 ngày trước đó (có thể loại bỏ chỉ số PM2.5 nếu không cần thiết). Đầu ra của mô hình sẽ là các mức AQI, được phân loại theo từng ngày. Ví dụ: các đặc trưng đầu vào của yếu tố nhiệt độ (TMP) của các ngày gần nhất là TMP_lag_1, TMP_lag_2,...

Dữ liệu đầu ra: sẽ là các mức AQI tương ứng cho các ngày sau, ví dụ AQI_category_day_1, AQI_category_day_2, ...

Lý do nhóm chọn dữ liệu của 5 ngày gần nhất là vì khi thử nghiệm với các khoảng thời gian từ 3 đến 10 ngày, các chỉ số từ ngày thứ 6 trở đi không còn ảnh hưởng đáng kể đến kết quả, và việc sử dụng dữ liệu của 5 ngày gần nhất sẽ giúp giảm nhiễu và nâng cao hiệu quả dự đoán.

b) Phân chia dữ liệu

Dữ liệu được chia thành ba tập: huấn luyện (training), xác thực (validation) và kiểm tra (testing) theo tỷ lệ xấp xỉ 60% – 20% – 20%.

Để đạt được tỉ lệ này, nhóm chia testing data là 6 tháng cuối cùng của 2021 (tổng 2000 samples chiếm 22.43%) và trong số còn lại, 25% là validation và 75% training chia ngẫu nhiên. Điều này nhằm đảm bảo quá trình huấn luyện

mô hình hiệu quả và đánh giá kết quả dự đoán một cách chính xác với yêu cầu bài toán.

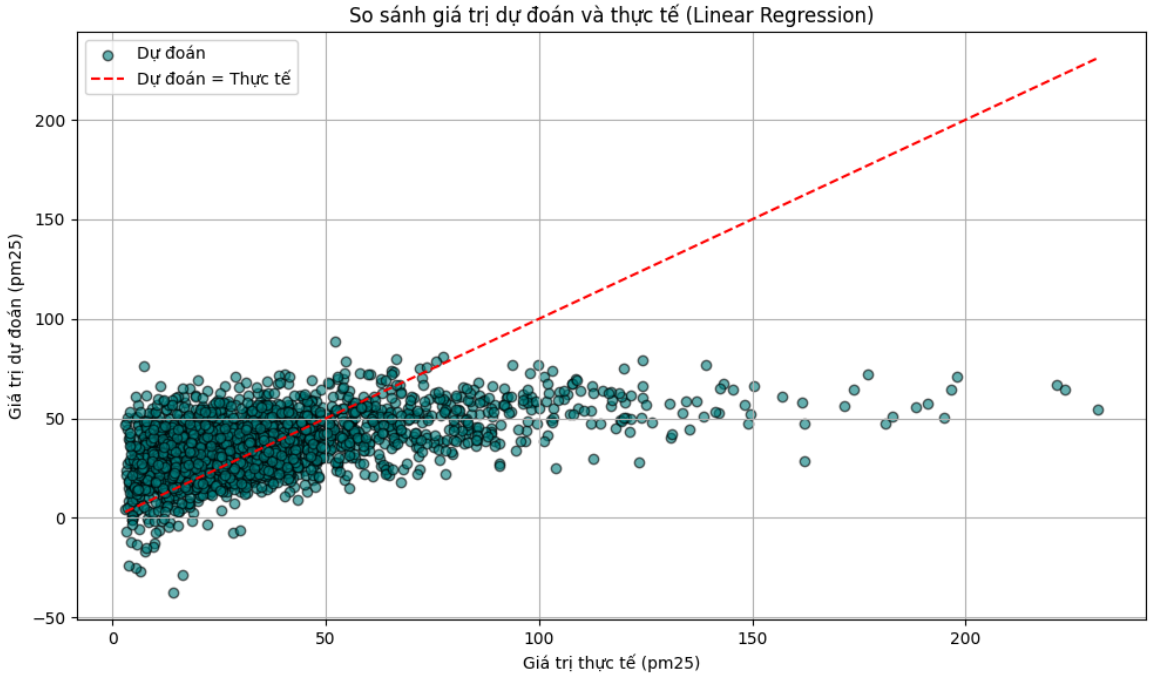
c) Mô hình học máy

i) Hồi quy tuyến tính (Linear Regression)

Ban đầu nhóm sử dụng mô hình Hồi quy tuyến tính (Linear Regression). Mô hình đơn giản, huấn luyện nhanh. Lý do nhóm quyết định sử dụng mô hình hồi quy thay vì mô hình phân loại ngay từ đầu là vì nhóm muốn thử nghiệm dự đoán chỉ số PM2.5 trước. Sau khi có dự đoán PM2.5, nhóm sẽ tiến hành phân loại các lớp AQI từ giá trị này.
Sau khi huấn luyện và tiến hành đánh giá, nhóm có những số liệu sau:

```
MAE: 17.415813898338598
RMSE: 24.801787190820445
R²: 0.2678837708988081
Pearson R: 0.517952078217668
```

Từ kết quả đánh giá, ta có thể thấy sai số lớn, R-squared thấp. Từ đây nhóm quyết định thử những mô hình phức tạp hơn và thay vì dự đoán PM2.5, phân loại thẳng AQI sẽ có thời gian huấn luyện nhanh hơn.



ii) Mô hình cây quyết định (Decision Tree)

Để tăng độ phức tạp cho mô hình, nhóm sử dụng ba mô hình cây quyết định gồm XGBoost, CatBoost và LightGBM để phân loại lớp AQI cho từng ngày trong 10 ngày tiếp theo. Kết quả dự đoán từ mỗi mô hình cho từng ngày sẽ

được nhân với một hệ số trọng số tương ứng theo từng ngày, sau đó cộng dồn lại rồi làm tròn để tạo ra dự đoán cuối cùng.

Hệ số trọng số được lựa chọn bằng cách sinh ngẫu nhiên 10.000 tổ hợp trọng số khác nhau, sau đó đánh giá hiệu suất của từng tổ hợp trên tập validation. Tổ hợp có chỉ số độ chính xác (accuracy) cao nhất được chọn làm hệ trọng số tối ưu.

Nhóm quyết định sử dụng hệ trọng số sinh ngẫu nhiên thay vì voting regressor hay stacking regressor vì trong quá trình huấn luyện, nhóm nhận thấy rằng từng ngày, các mô hình khác nhau sẽ có hiệu quả tốt hơn. Việc chọn một trọng số cố định như trong voting regressor không thể tối ưu hóa điều này. Mặc dù stacking regressor có thể cải thiện kết quả, nhưng thời gian huấn luyện lại tăng gấp 10 lần, khiến một lần huấn luyện mất tới 10 tiếng, điều này không khả thi cho nhóm.

Thay vào đó, việc tạo ra nhiều hệ trọng số ngẫu nhiên và chọn hệ tốt nhất dựa trên tập validation cho kết quả tương đương với stacking regressor trên tập test, nhưng thời gian huấn luyện lại bằng nhau khi so với voting regressor, mặc dù có rủi ro overfitting trên tập validation.

d) Phương pháp chọn tham số

Ban đầu nhóm quyết định sử dụng RandomSearch để tìm kiếm tham số thay vì GridSearch vì dự tính cho thấy việc áp dụng GridSearch để tìm tham số tối ưu cho dự đoán 10 ngày có thể mất tới 50 tiếng. Trong khi đó, RandomSearch giúp rút ngắn thời gian tìm kiếm tham số xuống còn chỉ 10 tiếng mà vẫn mang lại kết quả khá khả quan. Điều này giúp nhóm tiết kiệm thời gian mà không làm giảm quá nhiều chất lượng mô hình.

Tham số mà nhóm chọn để áp dụng phương pháp RandomSearch là số lần lặp (iterations) cho cả ba mô hình. Mỗi mô hình sẽ có các tham số khác nhau, nhưng có một tham số cố định là learning rate (tốc độ học) là 0.01. Vì learning rate đầu vào thấp, nhóm quyết định bắt đầu với số lần lặp (iterations) cao, dao động từ 1000 đến tối đa là 2500. Khoảng cách giữa mỗi lần thử nghiệm sẽ là 50 iterations.

Nhưng khi có các cải tiến từ bài báo cũng như đặt ra nhiều giả định khác nhau của mô hình, nhóm nhận thấy việc tốn đến 40+ tiếng để tìm parameters cho từng loại mô hình một là quá dài, chưa kể sau đó nhóm có thể đề xuất các phương án phát triển khác nên việc cố định một bộ parameters từ

RandomSearch là không linh hoạt nên nhóm quyết định sử dụng early stopping với evaluation metric là Accuracy trên tập validation. Điều này khiến cho mô hình có thể trở nên overfit trên tập valid, và kết quả cũng thể hiện điều này nhưng kết quả trên tập test cũng cải thiện đáng kể trên tập testing.

e) Phương pháp đánh giá

Nhóm lựa chọn mô hình cuối cùng dựa trên mô hình có chỉ số độ chính xác (accuracy) tốt nhất trên tập validation. Sau khi lựa chọn, mô hình này được đánh giá lại trên tập test để đưa ra kết quả cuối cùng một cách khách quan và chính xác nhất.

4) Áp dụng từ bài báo

a) Mức độ đô thị hoá

Bài báo đã chỉ ra rằng các yếu tố khí tượng có mối quan hệ nhân quả Granger đối với chỉ số AQI, và mối quan hệ này khác nhau tùy theo mức độ đô thị hóa của từng khu vực. Cụ thể, tại khu vực trung tâm, nhiệt độ và độ ẩm là nguyên nhân Granger của AQI; ở khu vực ngoại ô, đó là độ ẩm và tốc độ gió; trong khi ở vùng ven, chỉ có độ ẩm đóng vai trò là nguyên nhân Granger.

Tuy nhiên, do chưa có cách phân loại chính xác khu vực của các trạm đo trong tập dữ liệu, nhóm quyết định bổ sung các yếu tố phản ánh mức độ đô thị hóa vào dữ liệu đầu vào, cụ thể là mật độ dân cư và tỷ lệ đô thị hóa. Hai biến này được lấy từ dữ liệu năm 2022 do Bộ Xây dựng cung cấp. Những biến mới này sẽ được thêm vào từng mẫu dữ liệu (data sample) dựa trên thành phố mà dữ liệu được lấy từ.

Lý do nhóm chọn theo tỉnh chứ không theo khu vực gần các trạm vì đơn giản là không phải trạm nào cũng có đủ chỉ số, những chỉ số tại các khu vực lân cận thường lẻ tẻ ở nhiều nguồn và một số còn không tồn tại, thế nên để đảm bảo cân bằng giữa các trạm, nhóm chọn chỉ số của tỉnh mà trạm được đặt.

Sau đó, các biến này được nhân với các đặc trưng khí tượng tương ứng nhằm làm nổi bật mối liên hệ giữa yếu tố khí tượng và mức độ đô thị hóa của từng khu vực.

b) AQI thay đổi theo từng mùa

Dựa trên các nghiên cứu cho thấy các mùa trong năm có thể ảnh hưởng khác nhau đến chỉ số AQI, nguyên nhân có thể đến từ yếu tố khí hậu đặc trưng theo mùa hoặc sự thay đổi trong hoạt động xã hội như các dịp lễ, nhóm đã bổ

sung đặc trưng về mùa vào bộ dữ liệu đầu vào. Việc này nhằm giúp mô hình học được ảnh hưởng tiềm ẩn của mùa trong năm đến chất lượng không khí.

c) Điều kiện khuếch tán của không khí ảnh hưởng đến AQI

Các nghiên cứu cũng chỉ ra rằng ô nhiễm không khí chủ yếu là một quá trình tự tích tụ và tự khuếch tán. Khi điều kiện khuếch tán trở nên kém, các chất ô nhiễm có xu hướng tích tụ lại trong không khí, dẫn đến mức độ ô nhiễm gia tăng.

Cụ thể:

- **Gió**: Tốc độ gió cao có tác dụng phân tán các chất ô nhiễm, từ đó giúp cải thiện chất lượng không khí. Ngược lại, khi tốc độ gió giảm, điều kiện khuếch tán trở nên kém hiệu quả, làm tăng khả năng tích tụ ô nhiễm. Do đó, gió có mối tương quan âm với AQI.

- **Mưa**: Lượng mưa giúp rửa trôi các chất ô nhiễm khỏi khí quyển, góp phần làm giảm chỉ số AQI. Tuy nhiên, mức độ ảnh hưởng có thể khác nhau tùy vào khu vực địa lý. Khi lượng mưa thấp, khả năng làm sạch không khí cũng suy giảm, dẫn đến nguy cơ ô nhiễm cao hơn.

Nhằm phản ánh điều kiện khuếch tán ô nhiễm tại các trạm đo, nhóm đã nhân hai chỉ số **lượng mưa** và **tốc độ gió** để biểu thị mức độ khuếch tán tại từng thời điểm.

5) Kết quả

Nhóm tiến hành đánh giá hiệu quả dự đoán bằng cách chia mô hình thành 4 loại chính, dựa trên hai tiêu chí:

- Có sử dụng hay không sử dụng dữ liệu PM2.5 của các ngày trước đó.
- Có áp dụng hay không áp dụng các kỹ thuật và đặc trưng bổ sung được rút ra từ bài báo (như yếu tố đô thị hóa, mùa, điều kiện khuếch tán, v.v.).

Cụ thể:

- Loại 1: Mô hình không sử dụng PM2.5 quá khứ và không áp dụng các cải tiến từ bài báo.
- Loại 2: Mô hình không sử dụng PM2.5 quá khứ, nhưng có áp dụng các cải tiến từ bài báo.

- Loại 3: Mô hình có sử dụng PM2.5 quá khứ, nhưng không áp dụng các cải tiến từ bài báo.
- Loại 4: Mô hình có sử dụng PM2.5 quá khứ và có áp dụng đầy đủ các cải tiến từ bài báo.

Cách phân loại này giúp nhóm đánh giá rõ ràng ảnh hưởng của từng yếu tố (dữ liệu lịch sử PM2.5 và các đặc trưng bổ sung) đến hiệu suất dự đoán trên tập test. Kết quả như sau:

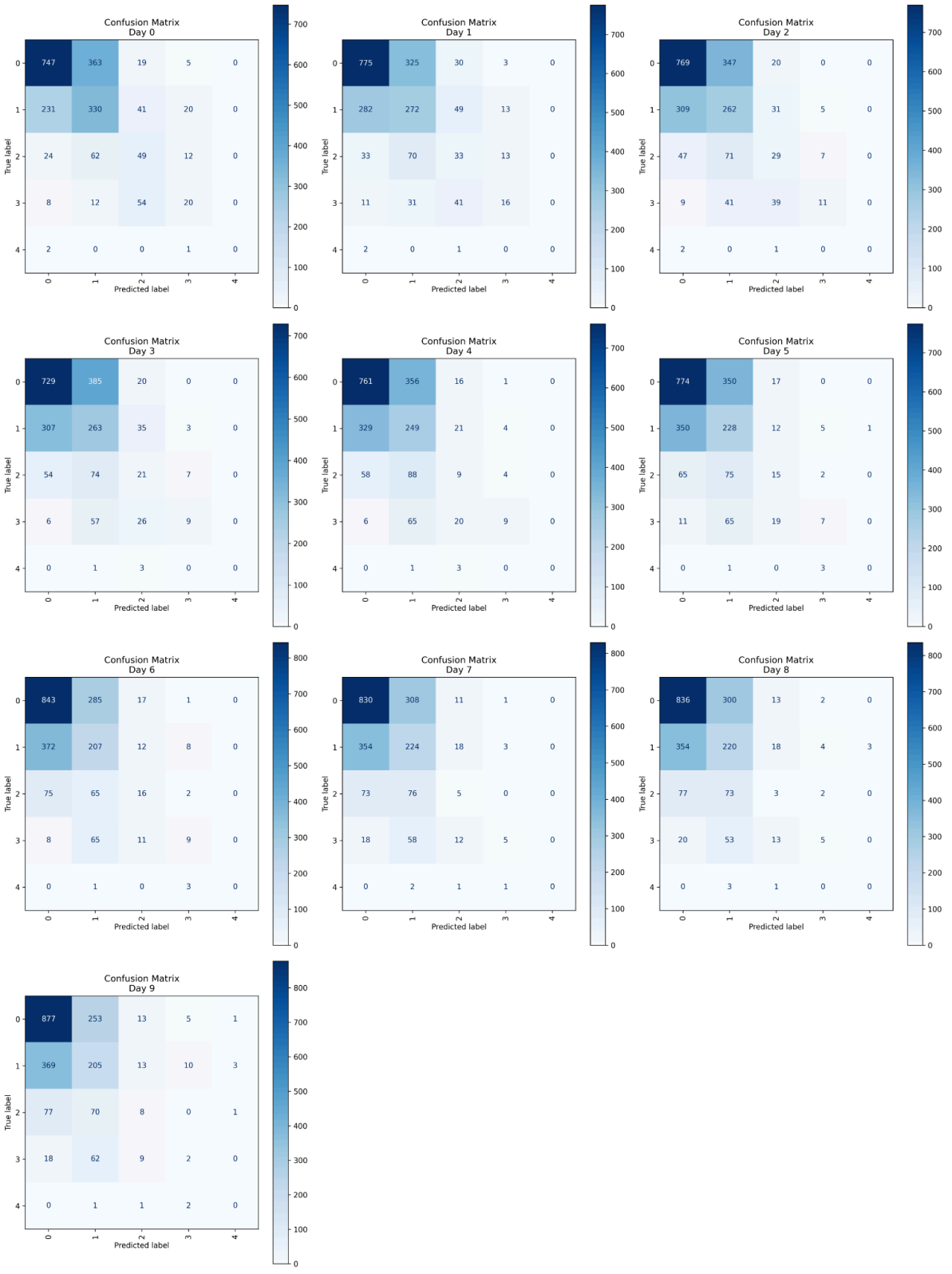
	Có sử dụng pm25 quá khứ	Không sử dụng pm25 quá khứ
Có áp dụng các cải tiến từ bài báo	**Accuracy**: 0.5605 – 0.7325 **Precision**: 0.5222 – 0.7284 **Recall**: 0.5605 – 0.7325	**Accuracy**: 0.5165 – 0.5830 **Precision**: 0.4941 – 0.6018 **Recall**: 0.5165 – 0.5830
không áp dụng các cải tiến từ bài báo	**Accuracy**: 0.5600 – 0.7320 **Precision**: 0.5212 – 0.7284 **Recall**: 0.5600 – 0.7320	**Accuracy**: 0.5160 – 0.5750 **Precision**: 0.4850 – 0.5789 **Recall**: 0.5160 – 0.5750

Kết quả cho thấy rằng việc áp dụng các cải tiến rút ra từ nghiên cứu bao gồm các đặc trưng như yếu tố đô thị hóa, mùa và điều kiện khuếch tán mang lại hiệu quả rõ rệt trong việc nâng cao chất lượng dự đoán khi accuracy tối đa tăng nhưng tối thiểu lại giảm nếu không sử dụng dữ liệu pm25.

Việc có sử dụng pm25 của những ngày trước giúp cho mô hình phán đoán chuẩn xác hơn đáng kể, và áp dụng cả 2 hai giúp mô hình dự đoán tốt nhất vào ngày 1 với 0.7265 accuracy.

Nhóm chọn mô hình không sử dụng PM2.5 quá khứ và có áp dụng các cải tiến từ bài báo để tiếp tục cải tiến vì mô hình này sẽ không yêu cầu phải có sẵn trạm đo, phù hợp cho việc dự đoán ở nhiều nơi khó xây dựng trạm.

Confusion matrices của mô hình không sử dụng PM2.5 quá khứ và có áp dụng các cải tiến từ bài báo cho từng ngày :



Dựa trên confusion matrices, nhóm nhận thấy mô hình thường xuyên nhầm từ nhãn 0 sang nhãn 1 nên đã điều chỉnh ngưỡng phân lớp: thay vì làm tròn thành 1 khi tổng điểm nằm trong [0.5, 1.5), giờ chỉ gán nhãn 1 nếu tổng điểm ≥ 1.0 và < 1.5; tương tự, các ngưỡng khác cũng được hiệu chỉnh thành [1.5, 2.25) cho nhãn 2, [2.25, 3.25) cho nhãn 3, [3.25, 4.5) cho nhãn 4, còn x ≥ 4.5 là nhãn 5, và x < 1.0 là nhãn 0. Việc này giúp giảm thiểu đáng kể lỗi dự đoán sai tại các mốc nhạy cảm.

Ta có kết quả của mô hình sau khi thay đổi cách làm tròn:

Day 0: Accuracy=0.5765, Precision=0.5724, Recall=0.5765
Day 1: Accuracy=0.5735, Precision=0.5655, Recall=0.5735
Day 2: Accuracy=0.5555, Precision=0.5406, Recall=0.5555
Day 3: Accuracy=0.5405, Precision=0.5234, Recall=0.5405
Day 4: Accuracy=0.5370, Precision=0.5104, Recall=0.5370
Day 5: Accuracy=0.5375, Precision=0.5147, Recall=0.5375
Day 6: Accuracy=0.5545, Precision=0.5136, Recall=0.5545
Day 7: Accuracy=0.5530, Precision=0.5059, Recall=0.5530
Day 8: Accuracy=0.5580, Precision=0.4961, Recall=0.5580
Day 9: Accuracy=0.5730, Precision=0.5039, Recall=0.5730

Accuracy: 0.5370 – 0.5765
Precision: 0.4987 – 0.5720
Recall: 0.5375 – 0.5760

Nhìn chung thay đổi đã giúp mô hình dự đoán tốt hơn, cụ thể là khả năng dự đoán tối đa ở ngày đầu giảm đi nhưng bù lại thì ở các ngày còn lại, khả năng dự đoán trở nên đều và chính xác hơn.

6) Tạo bản đồ dự báo

Dữ liệu dạng .tif được cho bao gồm nhiều ngày khác nhau liên tục một phần. Ảnh có kích thước 591x 345, độ phân giải thời gian là 1 ngày và từ cuối năm 2021 đến giữa năm 2022.

Vì nhóm sử dụng bộ 5 ngày để dự đoán AQI trong 10 ngày kế tiếp nên nhóm sẽ chia những ngày này vào những bộ để dự đoán, cụ thể các nhóm như sau:

Nhóm 5 ngày	10 ngày được dự đoán
22/12/2021-26/12/2021	27/12/2021-05/01/2022
27/12/2021-31/12/2021	01/01/2022-10/01/2022
10/01/2022-14/01/2022	15/01/2022-24/01/2022
11/04/2022-15/04/2022	16/04/2022-25/04/2022
01/05/2022-05/05/2022	06/05/2022-15/05/2022
06/05/2022-10/05/2022	11/05/2022-20/05/2022

Các ngày còn lại hoặc đứng riêng lẻ hoặc không đủ tạo một bộ liên tục 5 ngày nên nhóm quyết định bỏ qua.

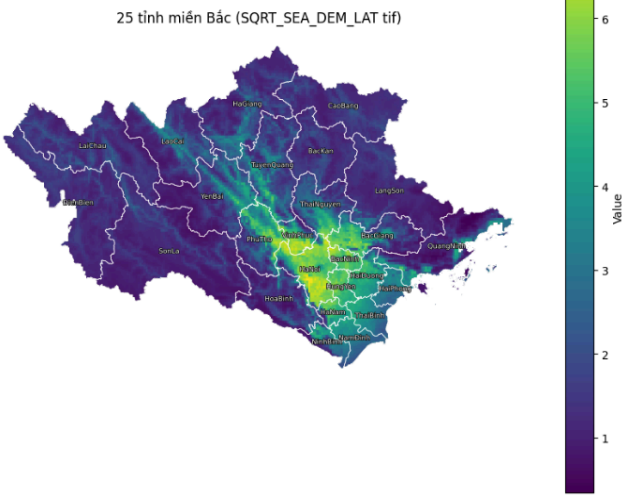
Sau khi đọc file, nhóm có thể chuẩn bị database bằng cách gán dữ liệu từng ngày vào các features tương ứng (Ví dụ 22/12-26/12 thì ngày 22 sẽ ở lag_5 và 26 sẽ ở lag_1 nhằm dự đoán cho 10 ngày kế tiếp)

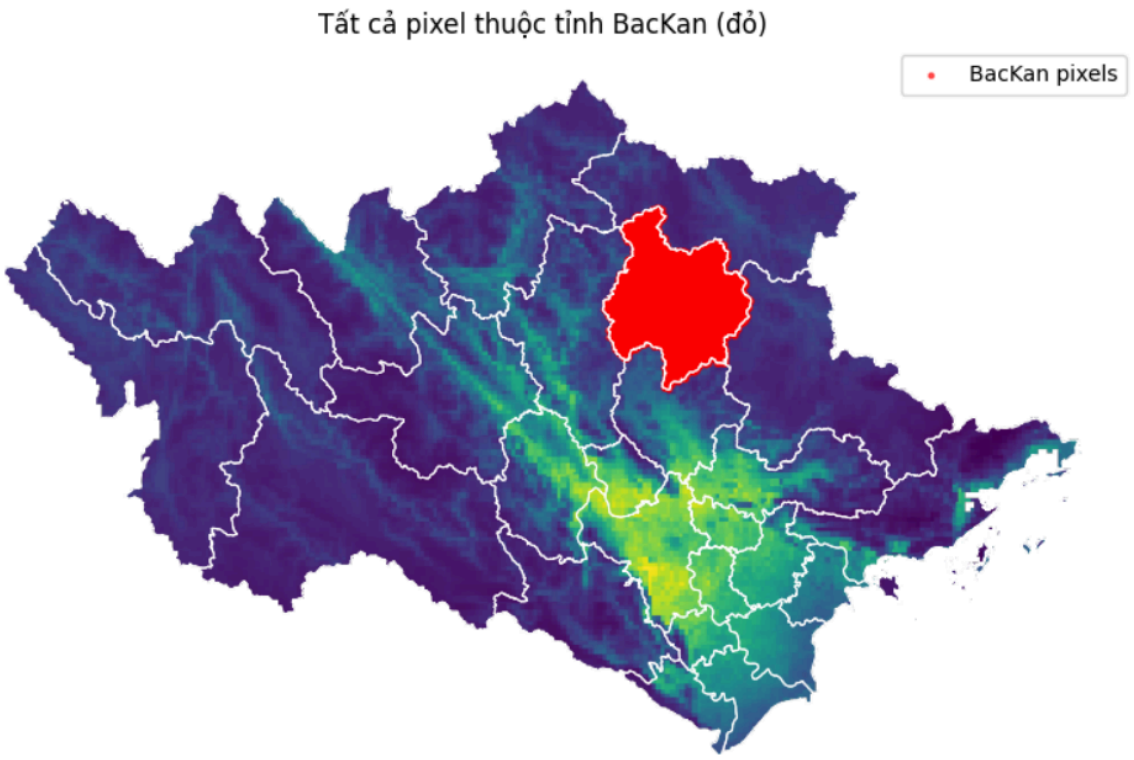
Vì nhóm có sử dụng các feature phụ thuộc vào địa lý của từng tỉnh, nhóm cần chia các ô vào đúng tỉnh mà nó thuộc về. Để đạt được điều này, nhóm sử dụng dữ liệu địa lý từ ShapeFile của GADM để phân loại địa giới hành chính các tỉnh thành tại Việt Nam, với cấp độ chia tương ứng là cấp tỉnh.

Từ đó, nhóm lọc ra các tỉnh thuộc khu vực phía Bắc Việt Nam để phục vụ cho mô hình phân tích, bao gồm:

Vĩnh Phúc, Quảng Ninh, Hải Phòng, Hưng Yên, Thái Bình, Hà Nam, Nam Định, Ninh Bình, Hà Giang, Cao Bằng, Bắc Kạn, Tuyên Quang, Lào Cai, Yên Bái, Thái Nguyên, Lạng Sơn, Điện Biên, Lai Châu, Sơn La, Hoà Bình, Bắc Giang, Bắc Ninh, Hải Dương, Hà Nội, Phú Thọ

Sau khi lọc được danh sách các tỉnh phía Bắc, nhóm tiến hành khớp dữ liệu từ ShapeFile với các file .tif nhằm gán chính xác từng ô.

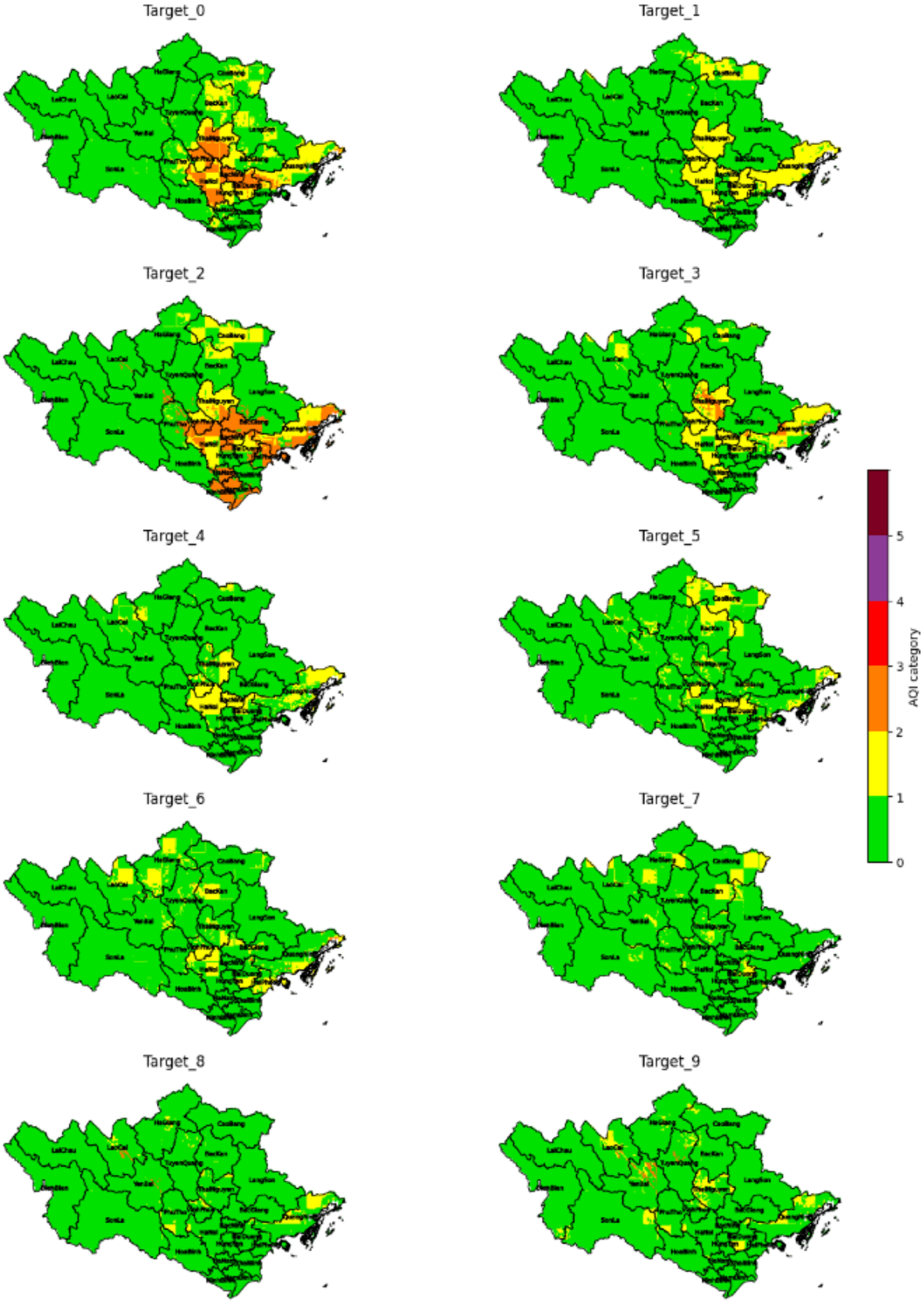




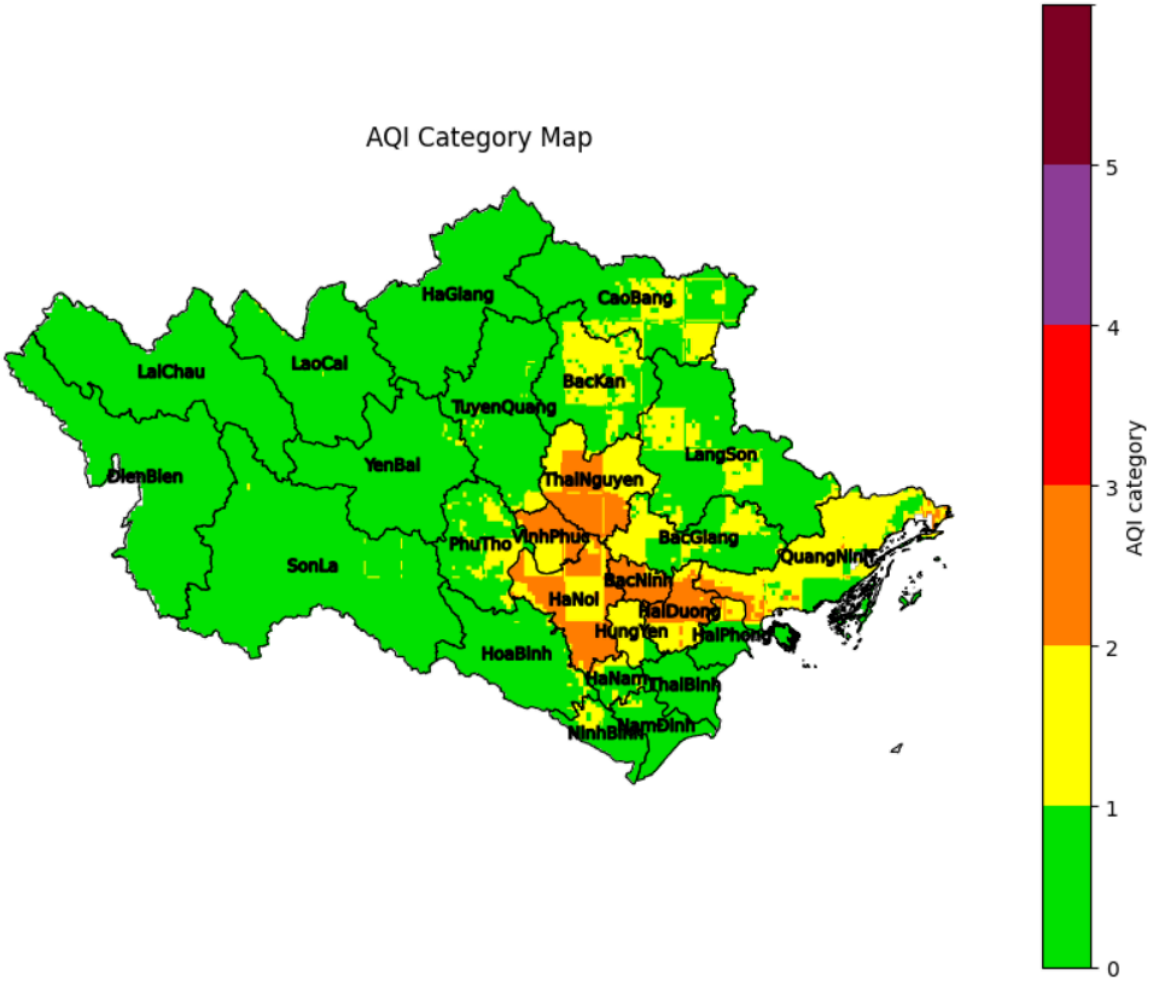
Từ đây nhóm đã có thể sử dụng các dữ liệu về mật độ dân cư và tỷ lệ đô thị hoá tương ứng của từng tỉnh.

Cuối cùng nhóm sử dụng mô hình hợp lý với bài toán là áp dụng những cải tiến mà nhóm đề xuất cùng với loại bỏ chỉ số pm2.5 để phân loại AQI cho từng ô, cuối cùng chuyển nó về dạng file .tif

Ví dụ về outputs 27/12/2021-05/01/2022:



Dự đoán ngày 27/12/2021:



7) Kết luận

Nhóm nhận thấy đây là một bài toán có độ phức tạp cao và mô hình tốt nhất mà nhóm xây dựng hiện chỉ đạt độ chính xác trong khoảng 0.5370 – 0.5765, đây vẫn là một kết quả tương đối khiêm tốn và cho thấy còn nhiều điều để cải thiện.

Việc tích hợp dữ liệu dân số theo từng tỉnh vào từng khu vực nhằm phản ánh tác động của đô thị hoá đến mức độ ô nhiễm không khí được nhóm cho là một hướng tiếp cận hợp lý khi nhóm gặp khó khăn trong việc tìm kiếm data. Tuy nhiên, nhóm nhận thấy cách phân bổ này vẫn còn hạn chế, do chưa phản ánh đầy đủ sự phân bố dân cư thực tế ở từng ô.

Bên cạnh đó, quá trình huấn luyện mô hình chưa được tối ưu tốt. Việc lạm dụng early stopping đã khiến mô hình overfit trên tập validation, làm giảm khả năng tổng quát hóa của mô hình. Ngoài ra, nhóm cũng chưa triển khai thử

nghiệm với các mô hình học sâu, vốn có thể phù hợp hơn với bài toán phức tạp như thế này.

STT	Họ và tên	Msv	Công việc	Đóng góp
1	Trần Tuấn Phong*	22028081	Nghiên cứu, phân tích, xử lý data, huấn luyện mô hình.	25%
2	Phạm Xuân Huy	22028271	Tìm dữ liệu, nghiên cứu.	25%
3	Nguyễn Mạnh Quỳnh	22028241	Tìm và phân tích, lọc dữ liệu	25%
4	Mai Quang Huy	22028223	Tìm dữ liệu, vẽ bản đồ, biểu đồ	25%