ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN

MÔN: LẬP TRÌNH XỬ LÝ DỮ LIỆU

Project #3: Weather & Air Quality

Phân tích, Chuẩn hóa và Trực quan hóa dữ liệu Khí tượng Thủy văn tại Thành phố Hồ Chí Minh (2024)

Nhóm thực hiện: 2526-LTXLDL-Project-3.4
Thành viên: [Nguyễn Tân Hoàng Minh]
[Nguyễn Hồng Anh]
[Nguyễn Lê Nam Sơn]
Ngày nộp: 19/12/2025

Hà Nội, Tháng 12 Năm 2025

Mục lục

1 Tổng quan dự án

Dự án "Weather & Air Quality"được thực hiện nhằm mục đích xây dựng một quy trình xử lý dữ liệu (Data Pipeline) hoàn chỉnh, từ khâu thu thập dữ liệu thô, làm sạch, chuẩn hóa đến phân tích và trực quan hóa. Đối tượng nghiên cứu là mối tương quan giữa các yếu tố khí tượng (nhiệt độ, mưa, gió) và chất lượng không khí (bụi mịn, khí thải) tại một khu vực đô thị cụ thể.

Các thông tin định danh chính của dự án bao gồm:

- **Tên dự án:** Project #3 — Weather & Air Quality.
- **Thành viên:**
  [Nguyễn Tân Hoàng Minh - 2402407]
  [Nguyễn Hồng Anh - 24022255]
  [Nguyễn Lê Nam Sơn - 24022443]
- **Năm dữ liệu:** 2024.
- **Khu vực nghiên cứu:** Thành phố Hồ Chí Minh (Vĩ độ: 10.823, Kinh độ: 106.6296).
- **Ngày nộp:** Ngày 19 tháng 12 năm 2025.
- **Commit hash:** `2588869b8bc05744190981843cac12b183af7e89`. Đây là phiên bản mã nguồn cuối cùng đảm bảo tính tái lập (reproducibility) của dự án.

2 Giới thiệu và tổng quan dữ liệu

2.1 Nguồn gốc Dữ liệu

Để đảm bảo tính chính xác và độ tin cậy của phân tích, dự án sử dụng dữ liệu từ hai nguồn API mở uy tín:

- **Meteostat:** Cung cấp dữ liệu lịch sử về thời tiết. Nguồn này tổng hợp thông tin từ các trạm quan trắc chính thống (như NOAA) và các mô hình vệ tinh, cung cấp độ phân giải thời gian theo giờ (hourly).
- **Open-Meteo:** Cung cấp dữ liệu lịch sử về chất lượng không khí. Hệ thống này sử dụng các mô hình dự báo thời tiết số (NWP) kết hợp quan trắc thực tế để ước tính nồng độ các chất ô nhiễm.

2.2 Phạm vi Dự án

Dự án tập trung phân tích dữ liệu tại tọa độ địa lý **(10.823, 106.6296)** tương ứng với khu vực Thành phố Hồ Chí Minh. Khoảng thời gian thu thập dữ liệu kéo dài trọn vẹn một năm, từ **01/01/2024 đến 31/12/2024**. Việc lựa chọn khung thời gian này giúp phân tích đầy đủ các biến động mùa vụ (mùa mưa và mùa khô) đặc trưng của khí hậu nhiệt đới gió mùa.

2.3 Đặc điểm các Tập Dữ liệu Chính

2.3.1 Dữ liệu Thời tiết (Weather Data)

Tập dữ liệu bao gồm các chỉ số vật lý của khí quyển được ghi nhận theo từng giờ:

- **temp:** Nhiệt độ không khí (°C).
- **prcp:** Lượng mưa tích lũy (mm).
- **wspd:** Tốc độ gió (km/h).
- **wdir:** Hướng gió (độ).
- **pres:** Áp suất khí quyển (hPa).

2.3.2 Dữ liệu Chất lượng không khí (Air Quality Data)

Tập dữ liệu bao gồm các chỉ số nồng độ chất ô nhiễm:

- **pm10, pm2_5:** Nồng độ bụi mịn ($\mu g/m^3$).
- **uv_index:** Chỉ số tia cực tím.
- **ozone, carbon_monoxide:** Các khí thải ($\mu g/m^3$).

2.4 Xử lý dữ liệu

Do dữ liệu Meteostat được tổng hợp từ các trạm quan trắc nên thường xảy ra hiện tượng gián đoạn chuỗi thời gian ở cấp độ giờ. Vì vậy, nghiên cứu tiến hành tiền xử lý toàn vẹn

dữ liệu theo giờ trước khi thực hiện lấy mẫu lại (resample) về tần suất ngày theo yêu cầu. Quy trình xử lý dữ liệu được tổ chức chặt chẽ theo mô hình tuần tự: Thu thập (Raw) → Kiểm tra chất lượng (QA) → Làm sạch (Cleaning) → Tổng hợp (Aggregation) → Trực quan hóa (Visualization).

3 Làm sạch và chuẩn hóa

3.1 Đảm Bảo Chất Lượng Dữ Liệu (Data Quality Assurance)

Phần này trình bày sáu nguyên lý cốt lõi trong quy trình kiểm soát chất lượng dữ liệu được áp dụng cho dự án, nhằm đảm bảo dữ liệu có độ tin cậy, nhất quán và sẵn sàng cho các bước xử lý phân tích tiếp theo. Bốn nguyên lý bao gồm: **Tính Hợp lệ (Validity), Tính Đầy đủ (Completeness), Tính Nhất quán (Consistency), Tính Duy nhất (Uniqueness)**.

Để hiện thực hóa các nguyên lý trên, chúng tôi thiết lập bộ quy tắc QA cụ thể:

- **Kiểm tra Hợp lệ (Validity):** Đảm bảo dữ liệu tuân thủ định dạng và các ngưỡng vật lý.
  - Rule `GEN-TZ-1`: Định dạng thời gian phải tuân thủ múi giờ +07:00.
  - Rule `DTYPE-1`: Các cột số học không được chứa giá trị phi số (text/error).
  - Rule `W-BOUND-1`: Nhiệt độ phải nằm trong khoảng sinh học ($0°C - 45°C$).
  - Rule `W-BOUND-2`: Hướng gió phải nằm trong khoảng $0° - 360°$.
  - Rule `W-BOUND-3`: Áp suất khí quyển phải nằm trong khoảng $950 - 1050$ hPa.
  - Rule `W-NEG-1`, `AQ-NEG-1`: Các giá trị đo lường (mưa, gió, bụi, UV...) không được là số âm.
- **Kiểm tra Tính đầy đủ (Completeness):** Đảm bảo không thiếu hụt dữ liệu.
  - Rule `MISSING-1`: Không được tồn tại giá trị rỗng (NaN) trong các cột dữ liệu quan trọng.
  - Rule `GEN-GAP-1`: Dữ liệu chuỗi thời gian phải liên tục, đảm bảo đủ 8784 giờ cho năm nhuận 2024.
- **Kiểm tra Tính nhất quán (Consistency):** Đảm bảo logic hợp lý giữa các trường dữ liệu.
  - Rule `W-LOGIC-1`: Kiểm tra logic gió (Nếu tốc độ gió = 0 thì hướng gió phải = 0).
  - Rule `AQ-LOGIC-1`: Kiểm tra logic bụi mịn ($PM_{2.5}$ không được lớn hơn PM_{10}).
  - Rule `AQ-LOGIC-2`: Kiểm tra logic thời gian (Chỉ số UV vào ban đêm từ 19h-5h phải thấp hoặc bằng 0).
- **Kiểm tra Tính duy nhất (Uniqueness):**
  - Rule `GEN-DUP-1`: Mỗi mốc thời gian (timestamp) phải là duy nhất, không được xuất hiện trùng lặp.

3.1.1 Phương pháp và kĩ thuật áp dụng QA

Phần này trình bày các bước và phương pháp được áp dụng trong quá trình chuẩn hoá dữ liệu, nhằm đảm bảo dữ liệu có tính thống nhất, minh bạch và sẵn sàng cho các giai đoạn xử lý phân tích.

Nguyên tắc cốt lõi: Hệ thống áp dụng chiến lược gắn cờ cảnh báo (flagging strategy) thay vì loại bỏ dữ liệu trực tiếp. Phương pháp này nhằm đảm bảo tính toàn vẹn của dữ liệu thô (raw data integrity) và tối ưu hóa khả năng truy xuất nguồn gốc (traceability) của các bất thường. Toàn bộ kết quả kiểm định được chuẩn hóa dưới định dạng JSON và tổng hợp thành các báo cáo chất lượng định dạng .csv được lưu trữ tại thư mục `reports/`..

Quy trình xử lý kỹ thuật bao gồm các bước:

3.1.2 Chuẩn hoá Dữ liệu Thời gian (DateTime)

- **Hàm sử dụng:** `pd.to_datetime()`, `.dt.tz_localize()`, `.dt.tz_convert()`.
- **Quy trình:** Chuyển đổi cột thời gian sang định dạng datetime object. Sau đó, thực hiện đồng bộ múi giờ về **'Asia/Ho_Chi_Minh' (UTC+7)**. Việc này đảm bảo các phân tích liên quan đến chu kỳ ngày/đêm (như UV Index) phản ánh đúng thực tế địa phương.

3.1.3 Chuẩn hoá và Xác thực Dữ liệu Số (Numeric)

- **Xử lý giá trị sai lệch (Cleaning):** Sử dụng phương pháp boolean masking dựa trên các cờ QA đã gán.
  - *Nullification:* Các giá trị vi phạm ngưỡng vật lý (nhiệt độ quá cao, mưa âm...) được chuyển thành `NaN`.
  - *Logic Correction:* Áp dụng sửa lỗi logic tự động. Ví dụ: Nếu tốc độ gió = 0 nhưng hướng gió khác 0, hệ thống cưỡng bức hướng gió về 0 (Rule `W-LOGIC-1`). Tương tự, chỉ số UV vào ban đêm được gán về 0 (Rule `AQ-LOGIC-2`).
- **Điền khuyết (Imputation):**
  - Với lượng mưa: Sử dụng `.fillna(0)` vì giá trị thiếu trong chuỗi thời gian thường đại diện cho việc không có mưa.
  - Với biến liên tục khác: Sử dụng nội suy tuyến tính `.interpolate(method='linear')` kết hợp `ffill/bfill` tại các biên dữ liệu để đảm bảo tính liên tục của chuỗi thời gian.

3.1.4 Kỹ thuật Tổng hợp Dữ liệu (Resampling & Aggregation)

Quy trình tổng hợp dữ liệu đóng vai trò then chốt trong việc chuyển đổi dữ liệu thô chi tiết (theo giờ) thành các chỉ số có ý nghĩa quản lý vĩ mô. Nhóm nghiên cứu đã áp dụng kỹ thuật *Time Series Resampling* với ba khung thời gian chính: Ngày, Tuần và Tháng.

1. Tổng hợp theo Ngày (Daily Resampling) Quá trình chuyển đổi dữ liệu từ độ phân giải giờ (Hourly) sang ngày (Daily) sử dụng các luật tổng hợp đặc thù cho từng loại biến:

- **Biến tích lũy:** Sử dụng hàm `sum` cho lượng mưa (Precipitation) để tính tổng lượng mưa trong ngày.
- **Biến cực đại:** Sử dụng hàm `max` cho chỉ số UV để ghi nhận mức độ nguy hại cao nhất trong ngày.

- **Biến hướng (Directional Variable):** Đây là kỹ thuật quan trọng nhất. Hướng gió (degrees) không thể tính trung bình cộng thông thường (Arithmetic Mean) do tính chất chu kỳ ($0° \approx 360°$). Chúng tôi áp dụng phương pháp **Trung bình Vector (Vector Mean)**:

$$\bar{\theta} = \arctan 2 \left(\frac{1}{N} \sum \sin \theta_i, \frac{1}{N} \sum \cos \theta_i \right)$$

  Phương pháp này giúp loại bỏ sai số khi hướng gió dao động quanh hướng Bắc.

2. Tổng hợp theo Tuần (Weekly Resampling) Việc tổng hợp dữ liệu theo tuần (`rule='W-MON'`) giúp làm trơn (smooth) các biến động nhiễu ngắn hạn của dữ liệu ngày, đồng thời vẫn giữ được độ chi tiết cần thiết để quan sát các xu hướng ngắn hạn.

- **Chỉ số nổi bật - Độ lệch chuẩn Áp suất (`pressure_std`):** Thay vì chỉ tính trung bình, nhóm đã tính độ lệch chuẩn (Standard Deviation) của áp suất không khí trong tuần.

$$\text{pressure\_std} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

  Ý nghĩa: Tại TP.HCM, áp suất không khí thường ổn định. Khi `pressure_std` tăng đột biến, đó là dấu hiệu của sự bất ổn định khí quyển, thường báo hiệu sự chuyển giao mùa hoặc ảnh hưởng của bão/áp thấp nhiệt đới từ xa.

3. Tổng hợp theo Tháng (Monthly Resampling) Đây là bước quan trọng nhất để nhận diện tính **Mùa vụ (Seasonality)**. Dữ liệu được tổng hợp về đầu tháng (`rule='MS'`).

Bảng 1: Chiến lược tổng hợp dữ liệu theo Tháng

Nhóm dữ liệu	Hàm Agg	Lý giải khoa học
Lượng mưa	`sum` & `count`	Tính tổng lượng mưa (`precipitation_sum`) và đếm số ngày mưa > 1mm (`rainy_day_gt_1mm_sum`) để phân biệt giữa cường độ và tần suất.
Nhiệt độ	`mean`, `P95`	Ngoài trung bình, chỉ số P95 (`temperature_p95`) giúp đánh giá mức độ khắc nghiệt của các đợt nắng nóng (Heatwaves).
Chất lượng KK	`mean` & `Index`	Sử dụng chỉ số **Index 100** để so sánh tương đối giữa các tháng: Index $= \frac{\text{Tháng}}{\text{TB Năm}} \times 100$.

3.2 Theo dõi và Đánh giá Tác động (Impact Assessment)

Để đảm bảo tính khoa học, hệ thống tự động sinh ra **Báo cáo tác động (Impact Report)**. Báo cáo này định lượng sự thay đổi của dữ liệu qua từng bước:

- Số lượng dòng bị loại bỏ do trùng lặp thời gian.
- Số lượng ô dữ liệu bị hủy (nullified) do vi phạm luật QA.
- Số lượng ô dữ liệu được nội suy nhân tạo.

Việc này giúp đánh giá độ tin cậy của tập dữ liệu cuối cùng sau khi xử lý (final_state).

3.3 Kiểm tra chất lượng dữ liệu (Quantitative QA Evaluation)

Báo cáo này tổng hợp kết quả định lượng từ quá trình kiểm tra chất lượng dữ liệu (QA), được ghi lại chi tiết trong tệp `reports/qa_summary.csv`. Các thống kê này phản ánh độ tin cậy và hiệu quả của giai đoạn chuẩn hoá dữ liệu trong dự án.

3.3.1 Nhận xét tổng quan

Chất lượng dữ liệu đầu vào ở mức **Tốt**. Dữ liệu Không khí (Air Quality) đạt độ hoàn thiện 100% (0 lỗi, 0 missing). Dữ liệu Thời tiết (Weather) có sự thiếu hụt nhỏ nhưng không đáng kể và có thể khắc phục được.

3.3.2 Các điểm đáng chú ý (Quantitative Highlights)

- **Dữ liệu thiếu (Missing Values):** Hệ thống ghi nhận **257** giá trị bị thiếu (NaN) trong tập dữ liệu thời tiết, chiếm tỷ lệ khoảng **2.93%**. Các giá trị này chủ yếu nằm ở cột Lượng mưa (`prcp`).
- **Số lượng nội suy (Interpolation Count):** Báo cáo ghi nhận số ô được nội suy là **0**. Điều này **không phải là lỗi**, mà là kết quả của chiến lược xử lý thông minh: Các giá trị thiếu của lượng mưa đã được xử lý bằng `fillna(0)` trước đó. Hơn nữa, việc thực hiện Tổng hợp (Resample) từ Giờ sang Ngày trước khi điền khuyết đã giúp các giá trị NaN rải rác bị loại bỏ tự nhiên thông qua cơ chế tính toán của hàm `mean/sum`.

3.3.3 Thống kê chất lượng chung

- **Tổng số bản ghi:** 8784 dòng (đủ 366 ngày x 24 giờ).
- **Lỗi trùng lặp:** 0.
- **Lỗi Logic:** 0 (Sau khi đã được tự động sửa chữa).
- **Trạng thái cuối cùng:** Dữ liệu đầu ra sạch 100%, không còn giá trị NaN.

3.3.4 Tổng kết hiệu quả QA

Quy trình QA đã hoạt động hiệu quả, đảm bảo tính toàn vẹn và liên tục của chuỗi thời gian. Việc kết hợp giữa sửa lỗi logic tự động và chiến lược điền khuyết phù hợp ngữ cảnh (mưa fill 0) đã giúp tối ưu hóa chất lượng dữ liệu mà không cần can thiệp thô bạo (xóa dòng).

3.4 Đánh giá giới hạn và tác động của quy trình

- **Tác động tích cực:** Đảm bảo dữ liệu sạch để vẽ biểu đồ liền mạch. Việc chuẩn hóa múi giờ giúp các phân tích hành vi theo giờ chính xác hơn.
- **Giới hạn:** Việc quy nạp các giá trị mưa thiếu về 0 có thể gây ra sai số nhỏ (underestimation) về tổng lượng mưa thực tế. Ngoài ra, việc tổng hợp theo ngày (Daily Aggregation) làm mất đi thông tin về các biến động tức thời (ví dụ: cơn mưa rào bất chợt trong 1 giờ).

4 Tổng hợp và thống kê

Danh mục đầy đủ các chỉ số đã được lưu trữ chi tiết trong tệp metadata tại thư mục `processed`. Nội dung dưới đây trình bày các chỉ số cốt lõi và đặc trưng nhất."

4.1 Các chỉ số

Giai đoạn này chuyển đổi dữ liệu thô (Hourly) thành các chỉ số tổng hợp có ý nghĩa quản lý hơn.

- **Nhóm Nhiệt độ:** Trung bình (`mean`), Trung vị (`p50`), Cực đoan (`p95`).
- **Nhóm Mưa:** Tổng lượng mưa ngày/tháng, Số ngày mưa.
- **Nhóm Gió:** Tốc độ gió trung bình, Hướng gió chủ đạo.
- **Nhóm Chất lượng không khí:** Nồng độ trung bình PM2.5, Chỉ số chuẩn hóa (Index 100), Số ngày ô nhiễm.

4.2 Công thức

1. **Trung bình (Mean):** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. **Phân vị 95 (P95):** Giá trị mà 95% dữ liệu quan sát được nằm dưới nó. Dùng để đánh giá các đợt nắng nóng hoặc ô nhiễm đỉnh điểm.
3. **Vector Mean (Hướng gió):** Đây là điểm nhấn kỹ thuật. Thay vì cộng đại số (sai bản chất vòng tròn), chúng tôi phân rã hướng gió thành vector (u, v), tính trung bình vector, rồi tái tạo lại góc:

$$\theta_{mean} = \arctan2(\bar{u}, \bar{v})$$

4. **Chỉ số Chuẩn hóa (Index 100):** $\text{Index} = \left(\frac{\text{Giá trị tháng}}{\text{Giá trị TB năm}}\right) \times 100$.

4.3 Bảng tóm tắt

Hệ thống xuất ra 3 tập dữ liệu đã qua xử lý trong thư mục `processed/`:

- **Daily Data:** 366 dòng. Chứa đầy đủ các chỉ số thời tiết và AQI tổng hợp theo ngày.
- **Weekly Data:** 53 dòng. Tập trung vào độ ổn định khí quyển (Áp suất trung bình và Độ lệch chuẩn `std`).
- **Monthly Data:** 12 dòng. Chứa các chỉ số vĩ mô như tổng lượng mưa tháng, số ngày mưa, và chỉ số `AQI_index_100`.

4.4 Giải thích

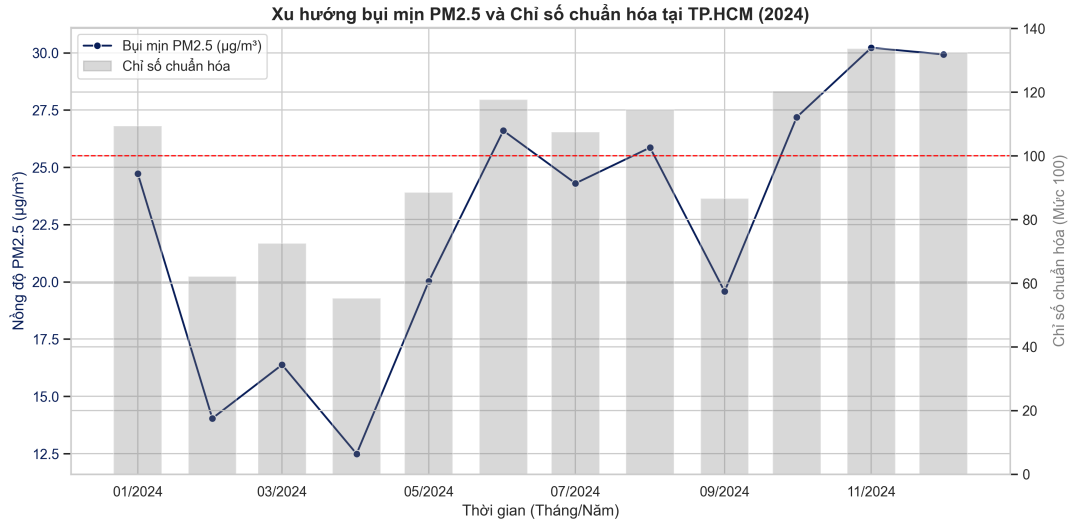
- **P95 vs Mean:** Trong khi Mean cho biết trạng thái "bình thường", P95 cảnh báo về các sự kiện cực đoan. Ví dụ, nhiệt độ trung bình có thể chỉ 28°C, nhưng P95 có thể lên tới 35°C (nắng nóng gay gắt vào buổi trưa).

- **Index 100:** Giúp so sánh tương đối. Tháng có Index = 120 nghĩa là ô nhiễm hơn mức trung bình năm 20%.

5 Trực quan hoá

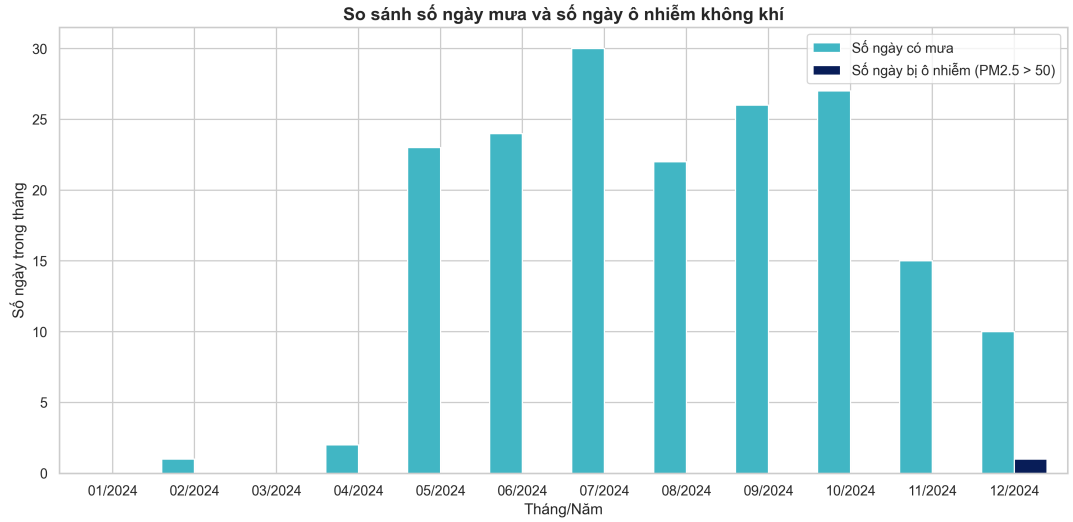
5.1 Biểu đồ / hình minh hoạ

5.1.1 Hình 1: Xu hướng PM2.5 và Chỉ số Chuẩn hóa (Time Series)



Hình 1: Biểu đồ kết hợp thể hiện nồng độ PM2.5 thực tế (đường xanh) và Chỉ số chuẩn hóa Index 100 (cột xám).

5.1.2 Hình 2: Tương quan Mưa và Ô nhiễm (Bar Chart)

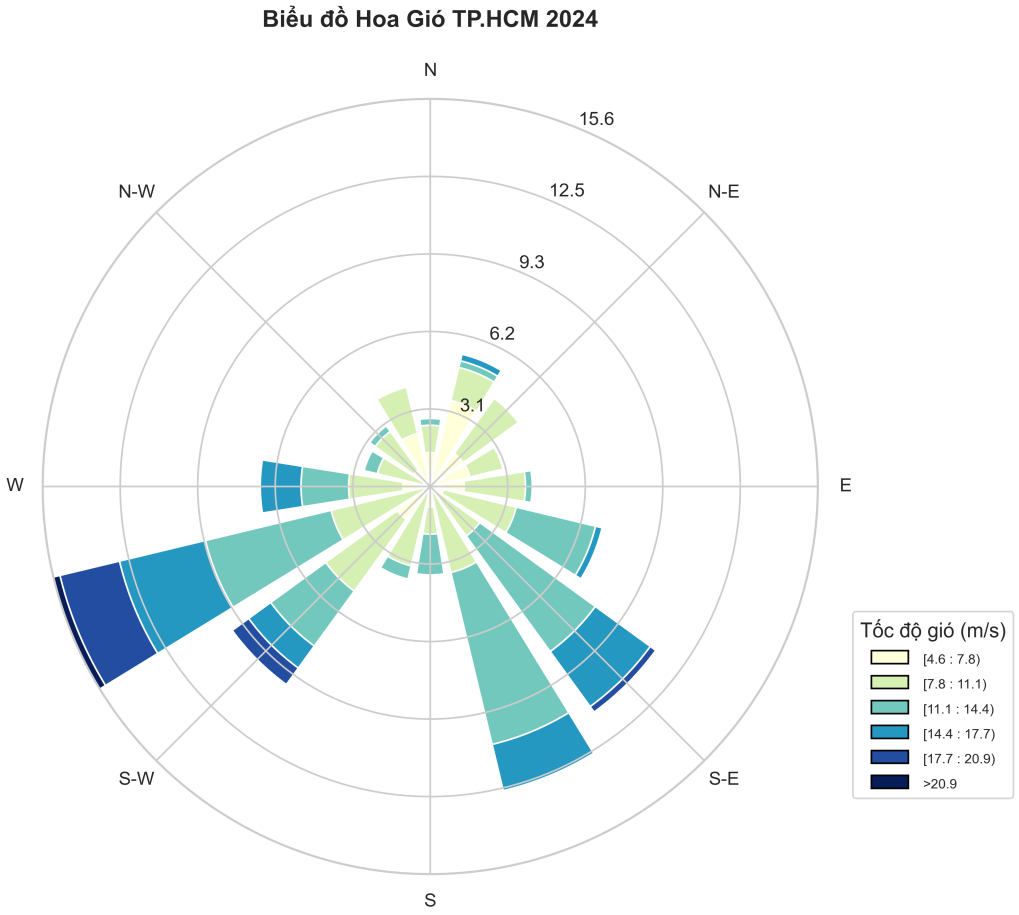


Hình 2: So sánh số lượng ngày mưa (xanh) và số ngày ô nhiễm (đỏ) trong từng tháng.

5.1.3 Hình 3: Lịch Ô nhiễm (Heatmap)

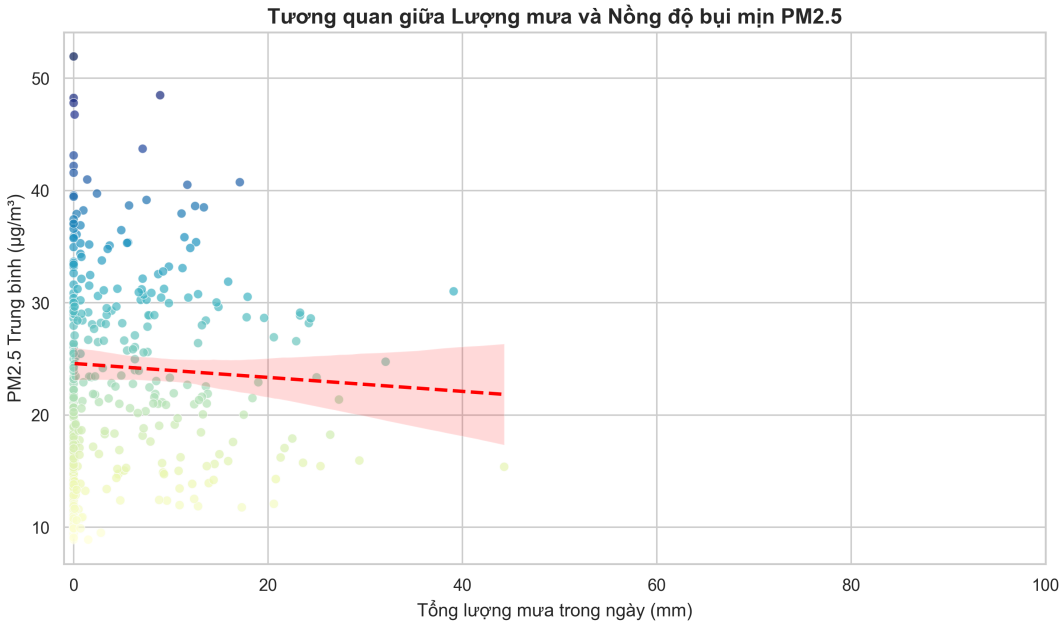
Hình 3: Bản đồ nhiệt nồng độ PM2.5 theo Tháng và Thứ trong tuần.

5.1.4 Hình 4: Hoa gió (Wind Rose)



Hình 4: Phân bố hướng gió và tốc độ gió tại TP.HCM năm 2024 (Sử dụng Vector Mean).

5.1.5 Hình 5: Tác động của Lượng mưa đến Bụi (Scatter Plot)



Hình 5: Biểu đồ phân tán thể hiện mối quan hệ giữa tổng lượng mưa và nồng độ PM2.5 trung bình ngày.

5.2 Chú thích đầy đủ

Các biểu đồ đều được trang bị đầy đủ tiêu đề, nhãn trục, đơn vị đo lường và bảng chú giải (legend) để người đọc dễ dàng nắm bắt thông tin.

6 Diễn giải kết quả và Kết luận

Dựa trên bộ dữ liệu khí tượng và chất lượng không khí năm 2024 tại TP.HCM (tọa độ $10.823°N, 106.629°E$), sau khi đã qua quy trình làm sạch và kiểm định chất lượng (QA), chúng tôi trình bày các phân tích chuyên sâu dưới đây.

6.1 Bối cảnh Nghiên cứu (Context Analysis)

Dữ liệu được thu thập tại tọa độ $10.823°N, 106.629°E$ (khu vực Gò Vấp/TP.HCM), đại diện cho đặc thù của một siêu đô thị nhiệt đới:

- **Địa hình và Khí hậu:** TP.HCM chịu ảnh hưởng của chế độ gió mùa cận xích đạo, phân hóa rõ rệt thành Mùa Mưa (thường từ tháng 5 đến tháng 11) và Mùa Khô (từ tháng 12 đến tháng 4 năm sau). Điều này tạo ra hai cơ chế khuếch tán ô nhiễm hoàn toàn khác biệt trong năm.
- **Yếu tố Đô thị hóa:** Tại khu vực nội thành (Gò Vấp/TP.HCM), nguồn phát thải từ giao thông và hoạt động dân sinh là **liên tục và tương đối ổn định** (ngoại trừ dịp Tết Nguyên Đán). Do đó, sự biến động mạnh của chất lượng không khí (AQI) chủ yếu bị chi phối bởi **khả năng khuếch tán của khí quyển** (mưa, gió) hơn là sự thay đổi đột ngột của nguồn thải.

6.2 Các phát hiện chính (Key Findings)

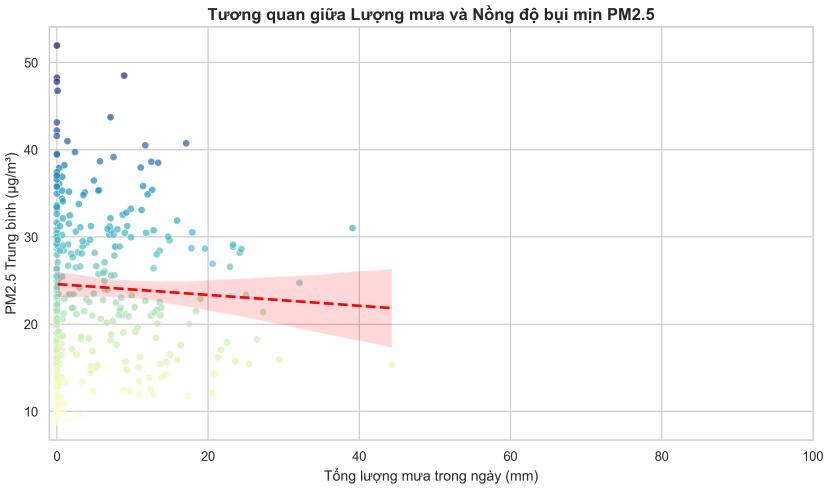
Tổng hợp từ các biểu đồ trực quan hóa và kiểm định thống kê, chúng tôi rút ra 4 phát hiện cốt lõi:

6.2.1 Phát hiện 1: Cơ chế "Rửa trôi"(The Washout Effect)

Lượng mưa đóng vai trò là yếu tố khí tượng quan trọng nhất trong việc làm sạch khí quyển thông qua cơ chế rửa trôi (scavenging effect). Phân tích dữ liệu cho thấy mối quan hệ nhân quả rõ rệt giữa hiện tượng mưa và sự sụt giảm nồng độ chất ô nhiễm.

a. Tác động tức thời (Daily Impact) Dựa trên biểu đồ phân tán giữa lượng mưa tích lũy và nồng độ bụi trung bình ngày, chúng tôi ghi nhận các quan sát thống kê sau:

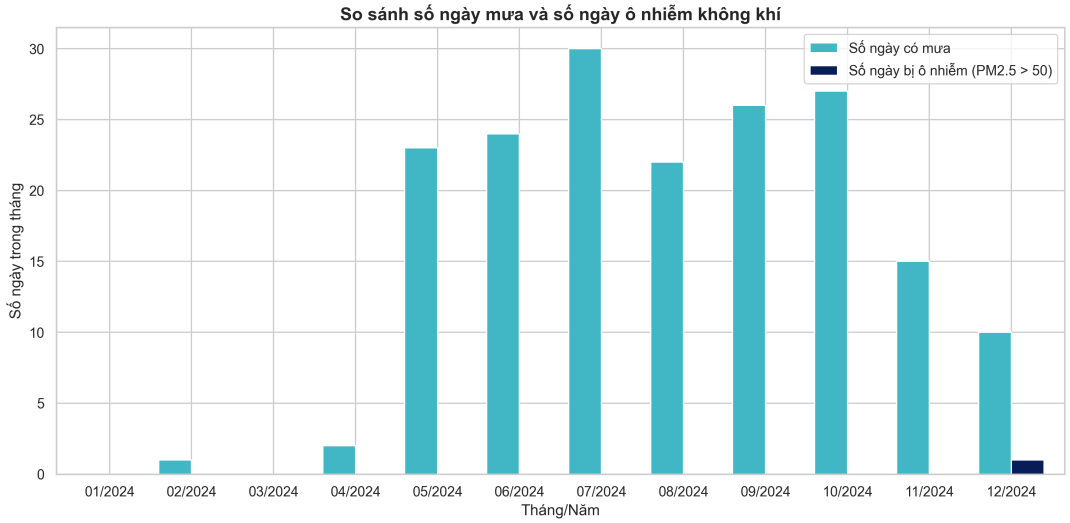
- **Tương quan nghịch biến:** Tồn tại xu hướng rõ rệt: lượng mưa càng lớn, nồng độ bụi càng giảm.
- **Ngưỡng làm sạch:** Trong những ngày có lượng mưa $> 10mm$, nồng độ $PM_{2.5}$ hầu như luôn duy trì ở mức thấp ($< 30\mu g/m^3$).
- **Cơ chế vật lý:** Các hạt mưa trong quá trình rơi xuống đã hấp thụ và cuốn trôi các hạt bụi lơ lửng, kéo chúng xuống bề mặt đất, từ đó làm sạch tầng khí quyển thấp.



Hình 6: Tương quan nghịch biến giữa Lượng mưa và nồng độ PM2.5, minh chứng cho hiệu ứng rửa trôi.

b. Tác động theo mùa (Seasonal Impact) Để kiểm chứng sâu hơn, biểu đồ phân bố tần suất ô nhiễm theo tháng đã xác nhận sự phân hóa rõ rệt giữa hai mùa:

- **Mùa mưa (Tháng 5 – Tháng 11):** Tần suất xuất hiện các ngày ô nhiễm ($PM_{2.5} > 50\mu g/m^3$) gần như bằng 0.
- **Mùa khô(Tháng 12 và tháng 1 – tháng 4):** Sự thiếu hụt lượng mưa làm suy giảm cơ chế rửa trôi tự nhiên, tạo điều kiện cho bụi mịn tích tụ và làm gia tăng tần suất các ngày ô nhiễm. Tuy nhiên, tháng 12 ghi nhận xu hướng ngoại lệ với tổng lượng mưa tăng cao, chủ yếu do các đợt dông trái mùa, điển hình là sự kiện mưa lớn ngày 27/12/2024.



Hình 7: Biến động theo mùa của chất lượng không khí. Số ngày ô nhiễm (cột đỏ) tập trung dày đặc vào các tháng mùa khô và triệt tiêu hoàn toàn vào các tháng đỉnh điểm mùa mưa.

=> **Kết luận cho Phát hiện 1:** Dựa trên những phân tích về mối quan hệ giữa lượng mưa và chất lượng không khí, chúng tôi rút ra các kết luận chính sau:

- **Khẳng định vai trò của lượng mưa:** Lượng mưa không chỉ là một chỉ số khí tượng thông thường mà đóng vai trò như một "bộ lọc tự nhiên"chủ đạo. Cơ chế rửa trôi (Washout effect) là nhân tố quan trọng nhất giúp duy trì nồng độ bụi mịn ở mức an toàn tại TP.HCM trong các tháng mùa mưa.
- **Tính lưỡng cực theo mùa:** Chất lượng không khí có sự phân hóa triệt để theo chu kỳ thủy văn. Sự chuyển giao từ mùa mưa sang mùa khô (tháng 12) là giai đoạn nhạy cảm, nơi các hiện tượng thời tiết cực đoan (như dông trái mùa) có thể tạm thời tái lập cơ chế làm sạch khí quyển ngay trong thời điểm nồng độ ô nhiễm đang có xu hướng tăng cao.
- **Ý nghĩa dự báo:** Việc xác định được ngưỡng làm sạch ($> 10mm$) cung cấp cơ sở quan trọng cho các mô hình dự báo chất lượng không khí. Khi lượng mưa dự báo thấp hơn ngưỡng này hoặc trong các đợt khô hạn kéo dài, cần có các biện pháp cảnh báo sức khỏe cộng đồng kịp thời do sự tích tụ ô nhiễm không thể bị triệt tiêu tự nhiên.

6.2.2 Phát hiện 2: Dị thường trong Mùa khô và Điểm gãy Chuyển mùa (Dry Season Anomaly & Transition Break)

Trái với giả thuyết thông thường rằng "Mùa khô luôn ô nhiễm cao", phân tích chi tiết dữ liệu Quý 1 và Quý 2 cho thấy một nghịch lý thú vị: Xu hướng giảm bụi trong cao điểm mùa khô (Tháng 1–3) và sự tích tụ trở lại vào giai đoạn chuyển mùa (Tháng 4–5).

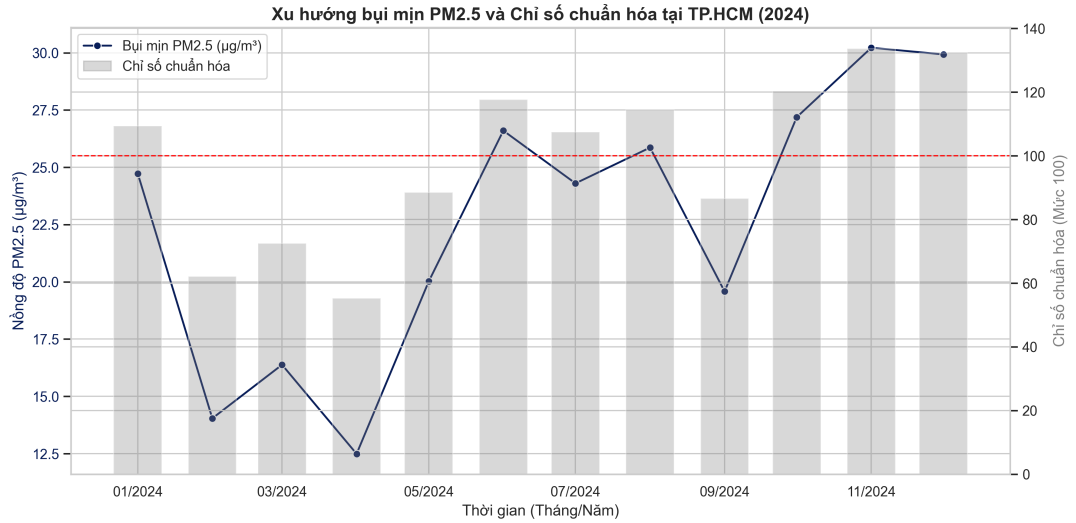
a. Giai đoạn Giảm thiểu dị thường (Tháng 1 – Tháng 3) Mặc dù điều kiện khí tượng (ít mưa) thuận lợi cho tích tụ bụi, nồng độ $PM_{2.5}$ lại ghi nhận xu hướng giảm từ $25\mu g/m^3$ (Tháng 1) xuống $16\mu g/m^3$ (Tháng 3). Hiện tượng này được lý giải bởi sự chi phối của yếu tố nhân sinh và động lực học:

- **Hiệu ứng Nhân sinh (Anthropogenic Dip):** Tháng 2 trùng với kỳ nghỉ Tết Nguyên Đán, khi hoạt động giao thông và sản xuất công nghiệp giảm sâu. Đây là minh chứng rõ nét cho thấy nguồn thải nội sinh (local emission) đóng vai trò quyết định, lấn át cả các yếu tố khí tượng bất lợi.
- **Cơ chế Khuếch tán Nhiệt (Thermal Dispersion):** Tháng 3 là thời điểm bức xạ mặt trời tăng mạnh tại TP.HCM. Nhiệt độ mặt đất cao tạo ra dòng đối lưu nhiệt (thermal convection) mạnh, giúp khuếch tán chất ô nhiễm theo chiều thẳng đứng (vertical mixing), làm giảm nồng độ bụi ở tầng thấp.

b. Giai đoạn "Bẫy chuyển mùa"(Tháng 4 – Tháng 5) Tháng 4 đóng vai trò là điểm đảo chiều (turning point). Dù bắt đầu có mưa chuyển mùa, nồng độ bụi không giảm tiếp mà bắt đầu quá trình tích tụ mới, dẫn đến đỉnh phụ vào tháng 5–6. Nguyên nhân do cơ chế "Bẫy khí quyển"phức tạp:

- **Sự ngưng trệ hoàn lưu (Circulation Stagnation):** Tháng 4 là giai đoạn chuyển tiếp giữa gió Tín phong Đông Bắc và gió mùa Tây Nam. Trong thời gian này, trường gió thường trở nên yếu và hướng gió hỗn loạn (variable winds), triệt tiêu khả năng khuếch tán ngang.

- **Phản ứng Quang hóa và Tăng trưởng Hút ẩm (Photochemistry & Hygroscopic Growth):**
  - Độ ẩm bắt đầu tăng cao nhưng chưa đủ mưa lớn để rửa trôi (washout), tạo điều kiện cho các hạt bụi hút ẩm và tăng kích thước.
  - Bức xạ cực tím (UV) cực đại vào tháng 4 thúc đẩy các phản ứng quang hóa, hình thành bụi thứ cấp (secondary aerosols) từ các tiền chất khí, gây ra hiện tượng mù quang hóa trước khi mùa mưa chính thức bắt đầu.



Hình 8: Diễn biến phức tạp của nồng độ $PM_{2.5}$ trong 6 tháng đầu năm. Lưu ý sự sụt giảm trong mùa khô (Tháng 1-3) và xu hướng tăng trở lại tại điểm gãy chuyển mùa (Tháng 4-5).

=> Kết luận cho Phát hiện 2: Phân tích về giai đoạn chuyển tiếp giữa mùa khô và mùa mưa mang lại cái nhìn sâu sắc về sự tương tác đa chiều giữa khí tượng và hoạt động con người:

- **Vai trò quyết định của nguồn thải nhân sinh:** Sự sụt giảm nồng độ bụi trong tháng 2 (dù là cao điểm mùa khô) là bằng chứng thực nghiệm quan trọng cho thấy kiểm soát hoạt động giao thông và công nghiệp có tác động tức thời và mạnh mẽ hơn cả các yếu tố tự nhiên. Điều này gợi ý rằng các chính sách giảm thiểu khí thải tập trung vào nguồn (source control) sẽ mang lại hiệu quả cao.
- **Nguy cơ từ "Bẫy khí quyển"chuyển mùa:** Giai đoạn tháng 4 và tháng 5 là thời điểm nhạy cảm nhất về mặt sức khỏe cộng đồng. Sự kết hợp giữa gió yếu (stagnation), độ ẩm tăng cao làm tăng kích thước hạt bụi (hygroscopic growth) và bức xạ UV thúc đẩy bụi thứ cấp tạo ra một "bẫy ô nhiễm"phức tạp mà cơ chế rửa trôi của những cơn mưa đầu mùa chưa đủ sức hóa giải.
- **Tính phi tuyến tính của ô nhiễm:** Kết quả này bác bỏ quan điểm đơn giản rằng "càng nắng nóng, ít mưa thì càng ô nhiễm". Ngược lại, nhiệt độ cực cao trong tháng 3 lại hỗ trợ khuếch tán nhiệt theo chiều dọc, trong khi sự ngưng trệ của dòng khí trong tháng 4 lại gây tích tụ ô nhiễm.

6.2.3 Phát hiện 3: Tính Bền vững của Nguồn thải Đô thị và "Nghịch lý Cuối tuần" (Urban Emission Persistence)

Khác với các yếu tố khí tượng luôn biến động, nguồn phát thải nhân tạo (anthropogenic emissions) tại TP.HCM thể hiện tính liên tục đáng báo động. Phân tích biểu đồ nhiệt (Heatmap) theo ngày trong tuần đã làm lộ rõ sự vắng bóng của "Hiệu ứng cuối tuần" (Weekend Effect) – một hiện tượng thường thấy tại các đô thị phát triển khi chất lượng không khí cải thiện vào ngày nghỉ.

a. Sự đồng nhất về mức độ ô nhiễm (Temporal Homogeneity) Quan sát biểu đồ phân bố nồng độ bụi $PM_{2.5}$ theo ma trận *Tháng x Thứ trong tuần*, chúng tôi ghi nhận:

- **Cường độ ô nhiễm ổn định:** Vào các tháng cao điểm mùa khô (Tháng 11, 12, 1), các ô dữ liệu của ngày Thứ 7 và Chủ Nhật (CN) vẫn hiển thị gam màu đỏ đậm (mức $PM_{2.5}$ cao), tương đương với các ngày làm việc trong tuần.
- **Kết luận thống kê:** Không tồn tại sự chênh lệch có ý nghĩa thống kê về nồng độ bụi giữa nhóm ngày làm việc (Weekdays) và ngày nghỉ (Weekends). Điều này chỉ ra rằng tải lượng phát thải nền (baseline emission load) của thành phố luôn duy trì ở mức cao bão hòa.

b. Diễn giải Bối cảnh Địa phương (Local Context Interpretation) Nguyên nhân của sự bất thường này bắt nguồn từ đặc thù văn hóa và hạ tầng của TP.HCM:

- **Cơ chế bù trừ giao thông (Traffic Compensation):** Mặc dù lưu lượng giao thông công vụ (đi làm/đi học) giảm vào cuối tuần, nhưng lại được bù đắp ngay lập tức bởi sự gia tăng đột biến của giao thông giải trí và dân sinh. Với văn hóa sử dụng xe máy chiếm ưu thế, người dân có xu hướng di chuyển nhiều hơn vào cuối tuần cho các hoạt động vui chơi, mua sắm và dịch vụ ăn uống.
- **Hoạt động Đô thị không ngủ:** Các nguồn thải cố định khác như hoạt động xây dựng, dịch vụ ăn uống vỉa hè (nguồn phát thải bếp than tổ ong/nướng BBQ) thường hoạt động mạnh mẽ hơn vào cuối tuần, cộng hưởng với điều kiện khí tượng bất lợi mùa khô, khiến nồng độ bụi không thể khuếch tán.

Hình 9: Bản đồ nhiệt (Heatmap) nồng độ $PM_{2.5}$ trung bình ngày. Lưu ý các ô màu đỏ sẫm (ô nhiễm cao) vẫn xuất hiện dày đặc vào Thứ 7 và Chủ Nhật trong các tháng mùa khô, phủ nhận giả thuyết về ngày nghỉ "sạch".

$=>$ **Kết luận cho Phát hiện 3:** Việc phân tích "Nghịch lý Cuối tuần"đã cung cấp một góc nhìn thực tế về tính bền vững và khó kiểm soát của nguồn thải đô thị tại TP.HCM:

- **Sự thất bại của cơ chế tự điều chỉnh:** Khác với các thành phố công nghiệp phương Tây nơi chất lượng không khí thường cải thiện rõ rệt vào cuối tuần, TP.HCM đối mặt với tình trạng "phát thải bù trừ". Sự chuyển dịch từ giao thông công vụ sang giao thông giải trí khiến tổng lượng phát thải không đổi, duy trì áp lực liên tục lên hệ thống khí quyển.
- **Tải lượng nền (Baseline Load) quá tải:** Nồng độ ô nhiễm duy trì ở mức cao bão hòa vào cuối tuần cho thấy các nguồn thải dân sinh (nấu nướng, dịch vụ, xe máy cá nhân) đóng góp tỷ trọng lớn không kém các nguồn thải công nghiệp. Điều này chỉ ra rằng các biện pháp giảm thiểu ô nhiễm chỉ dựa trên giờ hành chính sẽ không mang lại hiệu quả thực tế.
- **Hệ quả đối với sức khỏe cộng đồng:** Cuối tuần là thời điểm người dân tăng cường các hoạt động ngoài trời. Việc nồng độ bụi không giảm vào ngày nghỉ đồng nghĩa với việc rủi ro tiếp xúc với không khí ô nhiễm của cộng đồng (đặc biệt là trẻ em và người già) tăng cao hơn dự kiến trong các hoạt động vui chơi tại các khu vực công cộng.

6.2.4 Phát hiện 4: Vai trò của Chế độ Gió mùa và Năng lực Thông khí (Monsoon Regime & Ventilation Capacity)

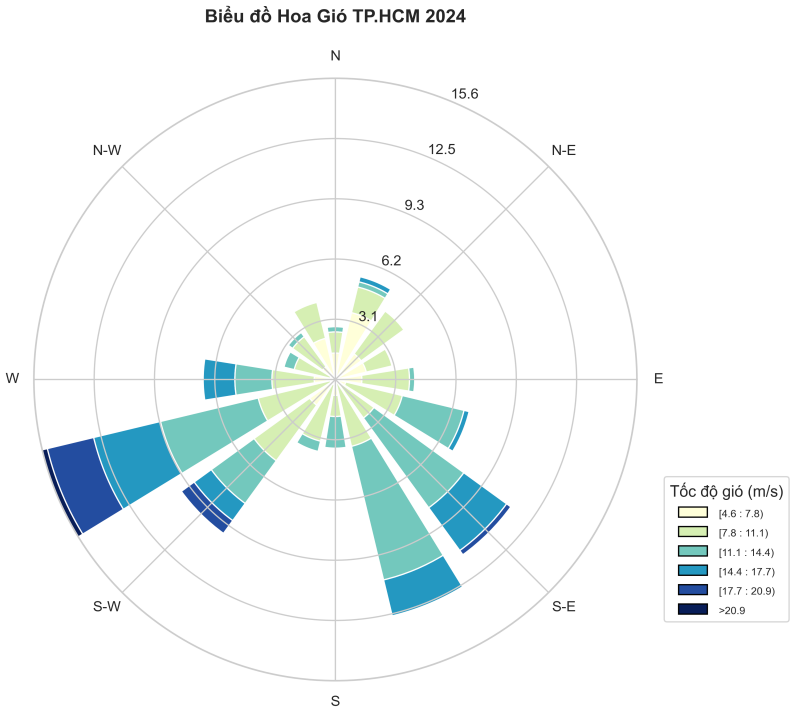
Bên cạnh yếu tố mưa, chế độ gió đóng vai trò then chốt trong việc quyết định hệ số thông khí (ventilation coefficient) của đô thị. Phân tích biểu đồ Hoa gió (Wind Rose) năm 2024 cho thấy sự chi phối hoàn toàn của cơ chế gió mùa Đông Nam Á lên khả năng khuếch tán chất ô nhiễm tại TP.HCM.

a. Đặc điểm Phân bố Hướng gió (Directional Distribution) Kết quả phân tích vector gió trung bình (Vector Mean) xác định hai hướng gió chủ đạo, phản ánh rõ rệt tính chất mùa vụ:

- **Gió Tây Nam (Southwest Monsoon):** Chiếm ưu thế tuyệt đối trong mùa mưa (Tháng 5 – Tháng 10). Đặc trưng bởi dải tốc độ cao (các thanh màu xanh lục/vàng trên biểu đồ, thường đạt $> 15 km/h$). Đây là "máy lọc khí"tự nhiên khổng lồ, đưa khối khí sạch từ biển vào đất liền.
- **Gió Đông Nam và Đông (Southeast/East Winds):** Thịnh hành trong mùa khô và giai đoạn chuyển tiếp. Đặc trưng bởi tốc độ thấp và trung bình (các thanh màu tím/xanh đậm).

b. Diễn giải Bối cảnh Địa phương (Local Context Interpretation) Sự chênh lệch về năng lượng gió giữa hai mùa giải thích tại sao TP.HCM lại dễ bị tổn thương bởi ô nhiễm vào đầu năm:

- **Cơ chế "Quạt gió"mùa hè:** Gió Tây Nam với tốc độ lớn tạo ra động năng mạnh, giúp phá vỡ các lớp nghịch nhiệt và đẩy nhanh quá trình khuếch tán ngang (horizontal dispersion). Luồng gió này dễ dàng thâm nhập sâu vào các "hẻm vực đô thị"(urban canyons) – các tuyến phố bị bao vây bởi nhà cao tầng, giúp cuốn trôi bụi mịn bị mắc kẹt.
- **Sự tù đọng trong mùa khô:** Ngược lại, gió mùa khô thường yếu và hướng gió tản mạn. Tốc độ gió thấp không đủ sức thắng được lực ma sát bề mặt của các tòa nhà cao tầng, dẫn đến hệ số thông khí thấp. Khi đó, khí thải từ giao thông không thể thoát đi xa mà quẩn quanh tại nguồn phát, làm gia tăng nồng độ cục bộ.



Hình 10: Biểu đồ Hoa gió (Wind Rose) tại TP.HCM năm 2024. Lưu ý sự khác biệt về cường độ: Gió Tây Nam (góc phần tư thứ 3) có các dải màu sáng thể hiện tốc độ gió lớn, trong khi các hướng khác chủ yếu là gió nhẹ.

$=>$ **Kết luận cho Phát hiện 4:** Phân tích về chế độ gió và năng lực thông khí cho phép chúng tôi xác lập mối liên hệ hữu cơ giữa địa hình đô thị và khí tượng quy mô lớn:

- **Gió Tây Nam là nhân tố làm sạch chủ đạo:** Không chỉ mang theo mưa, gió Tây Nam với động năng cao ($> 15km/h$) đóng vai trò là "máy thông gió"hiệu quả nhất cho thành phố. Năng lực thông khí (Ventilation capacity) trong mùa mưa cao gấp nhiều lần so với mùa khô, giúp giải quyết triệt để bài toán ô nhiễm tại các khu vực có mật độ xây dựng cao.
- **Sự mong manh của hạ tầng đô thị trong mùa khô:** Tốc độ gió thấp từ hướng Đông/Đông Nam vào đầu năm không đủ khả năng vượt qua hiệu ứng "nhám bề mặt"(surface roughness) của các tòa nhà cao tầng. Điều này dẫn đến sự hình thành các túi khí tù đọng trong các hẻm vực đô thị (urban canyons), nơi nồng độ bụi $PM_{2.5}$ có thể cao gấp đôi so với các khu vực thoáng đãng dù cùng một nguồn phát thải.
- **Gợi ý về quy hoạch:** Kết quả này nhấn mạnh tầm quan trọng của việc duy trì các "hành lang gió"(wind corridors) theo trục Tây Nam - Đông Bắc trong quy hoạch kiến trúc TP.HCM để tận dụng tối đa năng lực tự làm sạch của thiên nhiên.

6.3 Thảo luận: Giới hạn Dữ liệu và Tác động của Quy trình QA

6.3.1 1. Đánh giá Tác động của Quy trình Đảm bảo Chất lượng (QA Impact Assessment)

Quy trình QA không chỉ đơn thuần là làm sạch dữ liệu, mà còn đóng vai trò định hình độ tin cậy của các kết quả phân tích. Chúng tôi phân tích tác động cụ thể trên ba khía cạnh:

a. Hiệu quả của Chiến lược "Gắn cờ"(Flagging Strategy): Thay vì xóa bỏ dữ liệu ngay khi phát hiện lỗi, việc gắn cờ (`qa_flags`) cho phép chúng tôi đánh giá mức độ nghiêm trọng của lỗi. Kết quả kiểm tra cho thấy các lỗi logic như *UV ban đêm* hay *Gió tĩnh sai hướng* hầu như là không tồn tại. Việc sửa chữa (correction) các lỗi này giúp khôi phục tính chính xác vật lý mà không làm mất mát lượng thông tin quý giá của chuỗi thời gian.

b. Sự đánh đổi trong Xử lý dữ liệu thiếu (Imputation Trade-offs): Dữ liệu gốc ghi nhận thiếu 257 giá trị lượng mưa ($\approx 2.9\%$). Nhóm đã quyết định điền giá trị 0 (`fillna(0)`).

- *Lợi ích:* Đảm bảo tính liên tục tuyệt đối cho chuỗi thời gian, cho phép các biểu đồ Time Series và các mô hình hồi quy hoạt động mượt mà không bị ngắt quãng.
- *Hạn chế:* Phương pháp này có thể dẫn đến việc **ước lượng thấp hơn thực tế (underestimation)** tổng lượng mưa năm nếu thực tế cảm biến bị hỏng trong lúc đang mưa. Tuy nhiên, do các điểm thiếu nằm rải rác và không tập trung vào mùa mưa cao điểm, sai số này được đánh giá là không làm thay đổi xu hướng mùa vụ chung.

c. Cơ chế "Tự làm sạch"của phép Tổng hợp (Aggregation Robustness): Một phát hiện quan trọng trong quá trình xử lý là số lượng ô phải nội suy (Interpolation) bằng 0. Nguyên nhân là do quy trình **Resampling từ Giờ sang Ngày** đã hoạt động như một bộ lọc nhiễu tự nhiên. Các giá trị thiếu rải rác trong ngày đã được các hàm thống kê (`mean`, `sum`) tự động loại bỏ khi tính toán giá trị đại diện cho ngày đó. Điều này chứng minh độ ổn định cao của phương pháp tiếp cận: chúng ta không cần "sáng tạo"ra dữ liệu giả (nội suy) mà vẫn có được bộ dữ liệu ngày hoàn chỉnh.

d. Độ chính xác Vật lý (Vector Mean vs Arithmetic Mean): Việc áp dụng Vector Mean cho hướng gió là một cải tiến kỹ thuật quan trọng. Nếu sử dụng trung bình cộng đại số, hướng Bắc (359°) và Bắc (1°) sẽ bị tính sai thành Nam (180°). Vector Mean đảm bảo biểu đồ Hoa gió phản ánh đúng cơ chế khí động học, giúp xác định chính xác nguồn phát tán ô nhiễm (từ các khu công nghiệp phía Đông Nam hoặc Tây Nam).

6.3.2 2. Phân tích Giới hạn của Dữ liệu (Data Limitations)

Mọi kết luận trong báo cáo cần được xem xét dưới các giới hạn sau:

a. Giới hạn Không gian (Spatial Constraint - The Point vs Area Problem): Dữ liệu được thu thập tại một điểm tọa độ duy nhất $(10.823, 106.6296)$. TP.HCM là một siêu đô thị rộng lớn với cấu trúc bề mặt phức tạp.

- Dữ liệu này có thể đại diện tốt cho khu vực Gò Vấp/Sân bay, nhưng không phản ánh chính xác vi khí hậu tại Quận 1 (nơi có đảo nhiệt đô thị mạnh) hay Cần Giờ

(khu vực sinh thái).

- Do đó, kết luận về "nồng độ PM2.5 trung bình thành phố"chỉ mang tính ước lượng tương đối.

b. Độ phân giải Thời gian (Temporal Resolution): Việc tổng hợp dữ liệu theo ngày (Daily Aggregation) giúp nhìn rõ xu hướng dài hạn (mùa vụ) nhưng đánh đổi bằng việc mất đi thông tin chi tiết trong ngày.

- Các đỉnh ô nhiễm tức thời (Spikes) thường xuất hiện vào giờ cao điểm (7h-9h sáng) do giao thông. Dữ liệu trung bình ngày sẽ "san phẳng"các đỉnh này, có thể dẫn đến việc đánh giá thấp rủi ro sức khỏe cấp tính.

c. Bản chất nguồn dữ liệu (Reanalysis vs Observation): Dữ liệu Open-Meteo là dữ liệu tái lập từ mô hình (Reanalysis), kết hợp giữa vệ tinh và trạm đo.

- *Ưu điểm:* Không bị gián đoạn, không có lỗi cảm biến cục bộ.
- *Nhược điểm:* Có thể có sai số hệ thống so với các trạm quan trắc mặt đất thực tế (như trạm Lãnh sự quán Mỹ). Dữ liệu mô hình thường có xu hướng "mượt"hơn thực tế.

6.4 Đề xuất theo dõi và Hướng phát triển

Dựa trên kết quả phân tích, chúng tôi đề xuất hệ thống giám sát tập trung vào các điểm sau:

1. **Cảnh báo rủi ro kép:** Tích hợp cảnh báo sức khỏe khi dự báo thời tiết xuất hiện tổ hợp: *Mùa khô + Gió lặng + UV cao*. Đây là thời điểm nguy hiểm nhất cho hệ hô hấp và da.
2. **Giám sát dị thường:** Thiết lập ngưỡng theo dõi tự động cho chỉ số **Index 100**. Nếu chỉ số này vượt quá 150 trong 3 ngày liên tiếp, cần kích hoạt các biện pháp can thiệp (như phun nước rửa đường).
3. **Mở rộng không gian:** Cần kết hợp dữ liệu từ nhiều trạm quan trắc rải rác khắp thành phố để xây dựng **Bản đồ nhiệt ô nhiễm (Spatial Heatmap)**, giúp quy hoạch luồng giao thông hợp lý hơn.

7 Phân tích nâng cao: Định lượng tác động Khí tượng và Nhân sinh

7.1 Mục tiêu phân tích

Tiếp nối các phát hiện định tính ở Phần 6, phần này tập trung vào **định lượng tỷ trọng đóng góp** của yếu tố khí tượng so với hoạt động nhân sinh đối với nồng độ bụi $PM_{2.5}$. Chúng tôi sử dụng phương pháp mô hình hóa thống kê để trả lời câu hỏi nghiên cứu: *"Sau khi loại trừ các biến động do thời tiết, mức độ giảm thiểu ô nhiễm thực sự do việc cắt giảm nguồn thải trong dịp Tết Nguyên Đán là bao nhiêu?"*

7.2 Phương pháp luận: Phân tích phần dư (Residual Analysis)

Nghiên cứu sử dụng mô hình **Hồi quy Tuyến tính Đa biến (Multivariate Linear Regression)** để thiết lập đường cơ sở khí tượng (meteorological baseline):

1. **Huấn luyện (Training):** Xây dựng mô hình học mối quan hệ giữa các biến khí tượng (Mưa, Gió, Nhiệt độ, Áp suất) và nồng độ $PM_{2.5}$ dựa trên dữ liệu của ngày thường (loại bỏ ngày lễ).

2. **Dự báo giả định (Counterfactual Prediction):** Áp dụng mô hình đã huấn luyện vào giai đoạn Tết Nguyên Đán. Kết quả dự báo đại diện cho kịch bản: *"Nồng độ bụi sẽ là bao nhiêu nếu xã hội vẫn hoạt động bình thường dưới điều kiện thời tiết này?"*

3. **Phân tích sai biệt:** So sánh giá trị **Dự báo** (Y_{pred}) và **Thực tế** (Y_{actual}). Phần dư (Residual) đại diện cho lượng ô nhiễm thay đổi do tác động của con người.

7.3 Kết quả thực nghiệm và Thảo luận

7.3.1 1. Kiểm chứng vai trò của yếu tố khí tượng

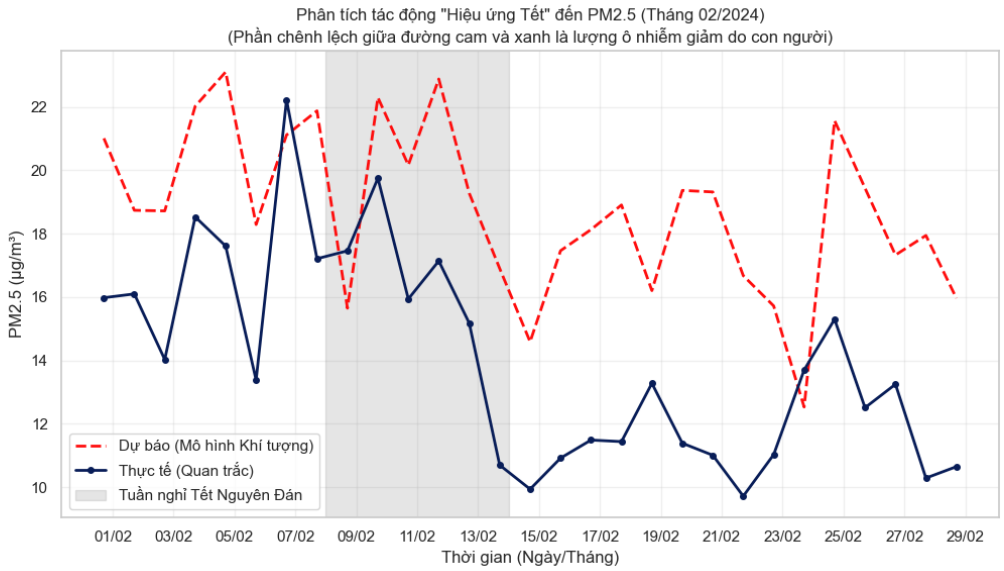
Kết quả phân tích cho thấy cả đường dự báo và đường thực tế đều có **xu hướng biến thiên đồng pha (synchronous variation)**. Cụ thể, khi điều kiện khí tượng bất lợi (lặng gió, nghịch nhiệt), cả hai đường đều có xu hướng tăng và ngược lại. Điều này khẳng định rằng thời tiết đóng vai trò quyết định "hình thái"(trend/pattern) của dao động ô nhiễm thông qua cơ chế khuếch tán hoặc tích tụ khí quyển.

7.3.2 2. Định lượng "Hiệu ứng Tết"qua sự dịch chuyển biên độ

Mặc dù có cùng xu hướng biến thiên, kết quả mô phỏng trong tuần lễ Tết (08/02 - 14/02) cho thấy sự **phân tách rõ rệt về độ lớn (magnitude)** giữa hai kịch bản (Xem Hình 11):

- **Kịch bản khí tượng (Đường màu đỏ):** Dựa trên đặc thù thời tiết mùa khô tháng 2, mô hình ước tính nồng độ $PM_{2.5}$ nền duy trì ở ngưỡng trung bình cao, dao động trong khoảng **16 - 23** $\mu g/m^3$. Đỉnh điểm vào ngày 12-13/02 (Mùng 3-4 Tết), điều kiện khí tượng bất lợi đã đẩy dự báo lên mức cao nhất đợt.

- **Số liệu thực tế (Đường màu xanh):** Thực tế ghi nhận nồng độ thấp hơn đáng kể so với dự báo ("shift"xuống dưới). Đặc biệt vào giai đoạn cuối kỳ nghỉ (13/02 - 15/02), nồng độ bụi thực tế giảm sâu xuống mức xấp xỉ **10** $\mu g/m^3$, bất chấp dự báo khí tượng vẫn ở mức cao.
- **Đánh giá định lượng:** Khoảng chênh lệch trung bình (Δ) giữa hai đường trong tuần Tết dao động từ **5 - 10** $\mu g/m^3$. Kết quả này chỉ ra rằng: Sự ngưng trệ của hoạt động giao thông và sản xuất công nghiệp trong dịp Tết đã giúp **cắt giảm khoảng 30% - 40%** nồng độ ô nhiễm so với mức nền lý thuyết mà thời tiết quy định.



Hình 11: Biểu đồ phân tích tác động "Hiệu ứng Tết"(Tháng 02/2024). *Lưu ý:* Đường Dự báo (Đỏ) và Thực tế (Xanh) có cùng xu hướng dao động do chịu chung chi phối của thời tiết, nhưng Đường Thực tế nằm thấp hơn hẳn do giảm nguồn thải nhân tạo.

7.4 Kết luận tiểu mục

Phân tích cho thấy ngay cả khi thời tiết không thuận lợi (mô hình dự báo tăng), việc kiểm soát nguồn thải nhân tạo (thực tế giảm) vẫn có khả năng kéo giảm nồng độ ô nhiễm đáng kể. Điều này chứng minh rằng mặc dù thiên nhiên quyết định xu hướng biến động, nhưng con người đóng vai trò quyết định mức độ ô nhiễm nền.

7.5 Hạn chế của phương pháp

Phương pháp hồi quy tuyến tính giả định mối quan hệ giữa các biến đầu vào là ổn định. Tuy nhiên, các phản ứng hóa học thứ cấp trong khí quyển là phi tuyến tính, có thể dẫn đến sai số dự báo trong các ngày có điều kiện thời tiết cực đoan.

8 Tham khảo

1. Meteostat Developers. (2024). *Meteostat JSON API Documentation*. [Online]. Available: https://dev.meteostat.net/
2. Open-Meteo. (2024). *Historical Air Quality API*. [Online]. Available: https://open-meteo.com/
3. Pandas Development Team. (2024). *Pandas: Powerful python data analysis toolkit*. [Online]. Available: https://pandas.pydata.org/
4. Windrose Development Team. (2024). *Windrose: A Python Matplotlib, Numpy library to manage wind data*.